MASTERS THESIS

Application of Machine Learning to Star Formation

Author:
Joseph P. Mwatukange

Supervisor:
Prof. Gary. Fuller

A thesis submitted to the University of Manchester
for the degree of Masters of Science by Research

in the

Department of Physics and Astronomy in the School of Natural Sciences
Faculty of Science and Engineering

2023

Contents

List of Figures

List of Tables

List of Abbreviations

ALMA	**A**tacama **L**arge **M**illimeter **A**rray
CO	**C**arbon **M**onoxide
CART	**C**lassification **A**nd **R**egression **T**ree
CASSIS	**C**ombined **A**tlas of **S**ources with **S**pitzer **I**RS **S**pectra
CWT	**C**ontinuous **W**avelet **T**ransform
DWT	**D**iscrete **W**avelet **T**ransform
FFT	**F**ast **F**ourier **T**ransform
GMCs	**G**iant **M**olecular **C**louds
GSO	**G**rid **S**earch Hyper-parameter **O**ptimisation
ISM	**I**nterstellar **M**edium
LTE	**L**ocal **T**hermodynamic **E**quilibrium
MAD	**M**ean **A**bsolute **D**eviation
MAE	**M**ean **A**bsolute **E**rror
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
ML	**M**achine **L**earning
MSE	**M**ean **S**quared **E**rror
NN	**N**eural **N**etwork
OLS	**O**rdinary **L**east **S**quares
PE	**G**eneralisation **E**rror
R^2	Coefficient of Determination
RF	**R**andom **F**orest
RMSE	**R**oot **M**ean **S**quared **E**rror
WT	**W**avelet **T**ransform
xgboost	e**X**treme **G**radient **B**oosting
UV	**U**ltra **V**iolet

THE UNIVERSITY OF MANCHESTER

Abstract

Faculty of Science and Engineering
Department of Physics and Astronomy in the School of Natural Sciences

Masters of Science by Research

Application of Machine Learning to Star Formation

by Joseph P. Mwatukange

This dissertation applies machine learning to a dataset of spectral lines from the Atacama Large Millimetre/sub-millimetre Array (ALMA) telescope of the complex organic molecule, methyl cyanide (CH_3CN). In terms of finding or forecasting interesting events or patterns, big data analysis presents new challenges for astronomy. In this work, we design and implement a spectral data compression technique using discrete wave transform (DWT) and present ensemble machine learning (ML) models as a method for obtaining the physical parameters of CH_3CN line emissions such as excitation temperature, column density, source size, FWHM, and velocity gradients. A random forest regressor and an extreme gradient boosting (xgboost) regressor were both used as ML techniques.The ML models were trained using synthetic data, and they performed well, with the tuned xgboost having the highest accuracy of all models. After applying the ML models to the observational data, our analysis revealed that they performed poorly in reconstructing spectra that perfectly matched the observations. Finally, we then investigate possible reasons for this poor performance.

Declaration of Authorship

I, Joseph P. Mwatukange, declare that this thesis titled, "Application of Machine Learning to Star Formation" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Copyright Statement

Acknowledgements

I want to start by expressing my gratitude to Professor Gary Fuller, who oversaw my thesis at the University of Manchester's Department of Physics and Astronomy. Whenever I encountered a problem or had a query regarding my research or writing, Professor Gary Fuller's office door was always open. He continuously allowed me to write this thesis on my own, but he gave me guidance when he believed I needed it. In addition, I'd like to thank my office mates Dr. Mike Walmsley, Micah Bowels, Fiona Potter, and the entire AI research group at the University of Manchester's Department of Physics and Astronomy as well as the start formation research group for their advice and extremely helpful criticism of my work. Lastly, I must express my sincerest gratitude to DARA for providing financial support for my studies, as well as to my family and friends for their unwavering love and unceasing support as I went through the process of conducting research for and writing this thesis. Without them, this feat would not have been possible. I'm grateful.

Dedicated to myself…
"All men by nature desire knowledge" - Aristotle

Chapter 1

Introduction

1.1 Background

Astrophysical research now includes a sizeable portion devoted to the analysis of the formation and early evolution of stars. One of the first models developed in the early 1930s postulated that galaxies emerged from primordial gas, a process that could be followed by viscous hydrodynamic simulations following condensation (Dayal, 2019; Pacifici et al., 2016) and semi-analytic models (Pacifici et al., 2016). Giant molecular clouds (GMCs), where massive stars are born, and dark clouds, where low-mass stars are born, are the two types of star formation clouds (Carraro, 2021). It has been challenging to understand and build a hypothesis that describes the beginning conditions of massive stars' formation for observational studies (Henning et al., 2006) because they form in remote and extremely veiled locations (Lery et al., 2005). GMCs, where star clusters originate, are usually only visible at infrared wavelengths because they are severely shrouded by dust (Lada & Lada, 2003; Lery et al., 2005). The clouds, on the other hand, are visible due to the absorption of light (in the optical) by background stars, the emission of cold dust at millimeter and sub-millimeter wavelengths, and the emission of simple molecules such as carbon monoxide (CO).

1.1.1 The Interstellar Medium (ISM)

Everything in the Universe that has mass, such as protons and electrons, is dispersed between stars as opposed to within them, making up the vast majority of the universe's baryonic matter. Because stars are created in the ISM, it is crucial to understand how they develop because it affects how galaxies arise and evolve (Colombo et al., 2014). When supernovae explode, their remains impact and shock the surrounding medium, compressing ambient ISM into rapidly expanding shells that cool quickly because of their high densities and may turn into molecular gas after 10^6 years (Wooden et al., 2004). Most of the ISM's volume is made up of the ionised and atomic gas phases. This material has the ability to create new star generations under specific circumstances. Additionally, a variety of proposed hypotheses about the origin or evolution of GMCs have been made

(Chevance et al., 2022) such as (*a*) theories governing the GMC's process, (*b*) gravity-induced compression of Jeans mass, the junction of local turbulence-induced filamentary gas flows, (*c*) the compression of shock waves from supernova shocks,(*d*), the buildup of mass through cloud-cloud collisions, and (*e*), the compression of matter in massive galaxy mergers. These formation hypotheses all have a strong connection to one another.

1.1.2 Giant Molecular Clouds (GMCs)

The gravitational collapse of interstellar gas clouds, which contract under gravity, is how stars are created (Lery et al., 2005). The clouds with the greatest chance of collapsing are the coldest and densest. These are the densest regions in GMCs. Large gas clumps known as molecular clouds, which are primarily made of molecular hydrogen gas, often have masses between $10^2 - 10^5$ $M_\odot$ (to a first approximation, determined by the Jeans length) (Maoz, 2016; Dayal, 2019). Molecular line emission, together with the emission from dust, are the most popular techniques for determining GMC masses since lines are strong and easy to spot even in distant galaxies. On the galactic scale, the three most frequently employed species are ^{12}CO, ^{13}CO, and for the densest gas HCN (Krumholz, 2015).

The intensity of the emission, I_ν, from a cloud of temperature T with an optical depth of τ_ν at a frequency ν is given by (Eq. 1.1)

$$I_\nu = (1 - e^{-\tau_\nu}) B_\nu(T). \tag{1.1}$$

where $B_\nu(T)$ is the Planck blackbody function. The cloud is opaque and radiates like a blackbody at its physical temperature in an optically thick cloud at a local thermodynamic equilibrium (LTE), where the optical depth of the cloud is $\tau_\nu \gg 1$. In the case of an optically thin cloud, $\tau_\nu \ll 1$, photons can pass through and reach the observer since the cloud is transparent. In light of this, the intensity is simply proportional to the optical depth, which is proportional to the quantity of atoms or molecules in the line of sight.

By observing these species in a molecular cloud, we may determine the column density-the number of molecules per unit area in our line of sight (using Eq. 1.1). Surveys of ^{12}CO and ^{13}CO species are particularly challenging to analyse because they are vulnerable to blending of emission from unrelated clouds in the galactic plane (Dobbs et al., 2014). We are able to determine the masses of GMCs, but we can also make educated guesses about their temperatures, velocity widths, diameters, and surface densities.

GMCs, which have temperatures ranging from 10-20 K throughout most of their volume and above in the regions near protostars(Maoz, 2016). As a result, silicate and carbon grain cores are encircled by a cold blanket of ices. The protostar's inner core begins to warm up to temperatures of about 100 to 300 K as the protostar grows, which eventually causes the hot core to heat up (Herbst & van Dishoeck, 2009). Once enough material from the envelope has accumulated, the star will have an accretion disc that will gradually build up before dispersing and the star contracts onto the main sequence. These

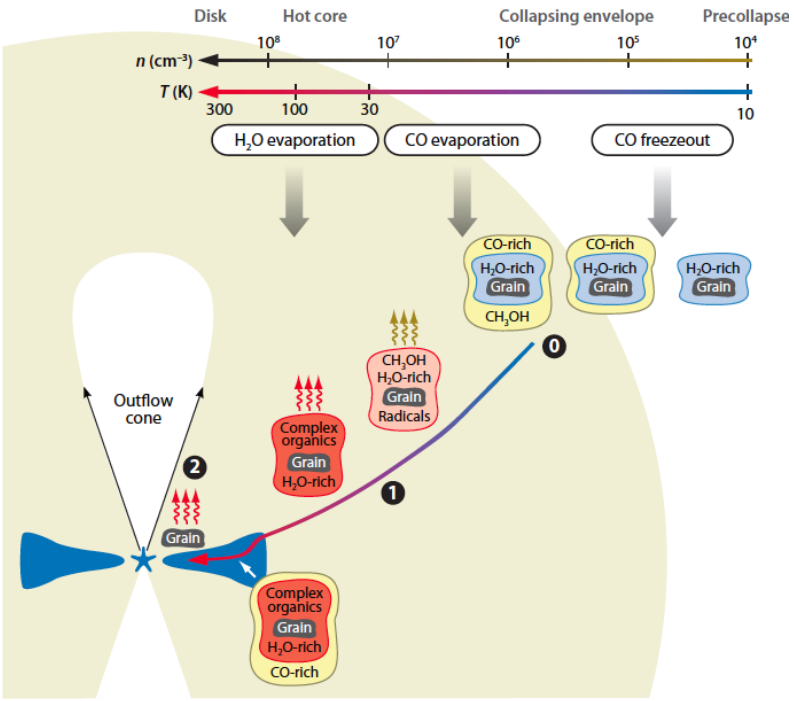


FIGURE 1.1: An illustration of the material's progression through a collapsing envelope from the prestellar core stage to a protoplanetary disc. The numbers (0) and (1), respectively, represent the creation of first- and zeroth-generation organic molecules in ices. When the envelope temperature hits 100 K and even strongly bound ices begin to evaporate, second-generation, (2) molecules begin to form in the hot-core area (Herbst & van Dishoeck, 2009).

steps lead to the zeroth, first, and second generations based on the chemistries of the organic molecules (see Fig. 1.1).

According to Herbst & van Dishoeck (2009)'s time-dependent gas-phase chemistry model, structures start off as a cold hydrostatic protostar sphere, which then undergoes an inside-out collapse in which materials start to flow inward to warmer and denser regions.

During the major accretion phase, when massive stars accumulate a lot of material during their birth process, their luminosities increase. The radiation pressure on the dust grains, however, cancels out the accretion and causes the gas to expand even faster as a result of enhanced ionisation. Large stars are unlikely to form when a spherically symmetric mass in-fall occurs. However, materials do gather through a circumstellar disc (Henning et al., 1990). Protostellar objects, which are cold objects, are created when molecular clouds break and collapse.

The ISM contains complex molecules, the majority of which are organic in origin since the heavy element Carbon, such as CO, predominates therein (Herbst & van Dishoeck, 2009). These molecules, which can be found in protoplanentary discs, were initially examined after the development of sub-millimeter astronomy, which made it possible to identify them. Because protoplanetary discs are cold and thick, and because most species are frozen out as ices on the grain surfaces and surface layer where both dissociation and

ionisation are dominant, the abundances of these species in the gas phase are significantly lower than those in the dark clouds (Ilee et al., 2021). The employment of interferometers has enabled the discovery of a few gas-phase complex compounds, such as methyl cyanide (CH_3CN), in protoplanetary discs (Öberg et al., 2015; Ilee et al., 2021).

The virial theorem asserts that for an equilibrium self-gravitating system,

$$2U + \Omega = 0 \quad \text{or} \quad U = -\frac{1}{2}\Omega \tag{1.2}$$

where U is the total thermal (kinetic) energy of the cloud and Ω is the total gravitational energy of the cloud. In other words, a classically bound star made of an ideal, non-relativistic gas has a negative total energy. All stars are destined to collapse at some point (Ω becomes more negative) since they all produce energy (and as a result, U grows increasingly negative).

Consequently, we need the gravitational term to outweigh the pressure in order for the cloud to contract, that is,

$$-\Omega > 2U \tag{1.3}$$

For the sake of brevity, let's assume a spherical gas cloud with constant density and temperature T, as well as particles with mean masses of $\bar{m}$. The gas is non-relativistic, ideal, and classical. The cloud's mass is M, its radius is r, and its gravitational energy is represented by

$$|\Omega| \approx \frac{GM^2}{r^2} \tag{1.4}$$

If the cloud experiences radial compression , dr, its gravitational energy will change (become more negative).

$$|d\Omega| = \frac{GM^2}{r^2}dr \tag{1.5}$$

This results in a volume reduction of the cloud by

$$dV = 4\pi r^2 dr, \tag{1.6}$$

The thermal energy will consequently rise by,

$$dU = PdV = nk_BT4\pi r^2 dr = \frac{M}{\frac{4}{3}\bar{m}\pi r^3}k_BT4\pi r^2 dr = \frac{3Mk_BT}{\bar{m}}\frac{dr}{r}, \tag{1.7}$$

The cloud will be prone to gravitational collapse if the change in gravitational energy is greater than the increase in thermal energy (and the pressure support it provides),

$$|d\Omega| > |dU| \tag{1.8}$$

As a result, it may be concluded that clouds will collapse if their mass exceeds the Jeans mass (M_J) for a given radius r and temperature T,

$$M_J = \frac{3k_B T}{G\bar{m}} r \tag{1.9}$$

The initial cold temperature that results from a cloud's collapse causes the Jeans mass to decrease while the density increases. Smaller mass components may thus become unstable, leading to the development of star clusters. A hydro-static core will eventually form when the particles heat up.

Star formation typically produces clusters of stars rather than single stars, where associations are born simultaneously before feedback inhibits star formation (Rieder et al., 2021). Additionally, groups of stars that are physically linked together are referred to as stellar clusters. Star clusters form as clumps, which are areas of excessively dense material, grow larger and are frequently gravitationally bound (gravity keeping them together). Open clusters and globular clusters are the two categories into which clusters fall. Open clusters have been found to be young due to the existence of massive stars. Most of them are not restrained by the gravitational pull of their own bodies. Globular clusters are bound systems with low metal content, which suggests that they were formerly relatively pure gas before forming some time ago.

Massive stars play an important role in the evolution of the Universe since they are the primary source of heavy elements and ultraviolet (UV) radiation Zinnecker & Yorke (2007). Given first order, this suggests that they form later than low-mass stars because of their extreme UV radiation and winds, which have an adverse effect on the gas reservoir around them after they are formed (Beuther, 2011). According to (Dopita & Stromlo, 1988), they are the outcome of molecular cloud collisions or crushing events. A dense sheet of shock-compressed materials is created by these events, which are primarily caused by supernova explosions and star winds (Woodward, 1978). Despite all of the theories on how major stars are born, we still don't fully grasp how they form and develop. To start, it is believed that clouds of dust created by cloud collisions are an effective mechanism that triggers cloud collapse (Wang et al., 2004) and produces stars, making them difficult to observe and giving insufficient details on their early formation. In addition, it is particularly challenging to explain the theory underlying the development process due to how swiftly it happens (Zinnecker & Yorke, 2007).

1.2 Aims and Objectives

This project aims to analyse observational data, create theoretical methods, and construct computational models to advance knowledge and collaborations in one of the following fields: pulsar astrophysics, radio astronomy technology, stellar astrophysics, solar plasma physics, cosmology and gravitational physics, galaxies and cluster formation, or astrochemistry. The subsequent goals will help achieve this goal:

- Gain knowledge about spectral data cubes from Atacama Large Millimeter Array (ALMA) telescope, their meaning, and information extraction techniques.

- Examine machine learning (ML) approaches to object identification, data compression, and categorization.
- Investigate novel ML approaches that may prove useful in the processing of spatial-spectral data.

Chapter 2

Star Formation and the Function of Spectral Lines

Introduction

When stars are born, they leave behind intricate precursor molecules to biological molecules that live in the protostars' gaseous envelopes (Calcutt et al., 2018). These complex organic molecules are the tracer materials left behind during the production of prestellar cores, massive hot corinos, and low hot corinos (Jiménez-Serra et al., 2016), which are crucial events in determining some of the early phases of star formation. Numerous star-forming regions have been extensively examined, including Sgr A and Sgr B (Solomon et al., 1971; Menten et al., 2010; Bonfand, M. et al., 2019; Matthews & Sears, 1983; Meng et al., 2022, 2019) and IRAS 20126+4104 (Cesaroni et al., 2014). Methyl cyanide traces reveal a wide range of ISM gas cloud physical characteristics, including temperature, velocity, column density, and ionisation. However, variations of the global and local pressure dictates the amount of cold, dense material free for star formation process (Heyer et al., 2019). This chapter will cover some of the key details on the methyl cyanide (CH_3CN)) species, its detection in high mass objects (Cesaroni et al., 2017), and its significance for understanding the chemistry and dynamics (Barrientos & Solar, 2019) of the cosmos.

2.1 Rotational Energy Levels for Symmetric-Top Molecules

Methyl cyanide belongs to the group of symmetric-top compounds with rotational energy levels. The labels for the levels are typically divided into four categories: J, which specifies the total angular momentum; K, which specifies the angular momentum about the top axis z; $+l$ or $-l$, labels resulting from a consideration of the Coriolis splittings in degenerate vibronic states; and Γ, which is a symmetry species of the rotational subgroup of the full molecular point group and is useful when considering the statistical weights of the rotational energy levels (Hougen, 1962).

The expectations values of the internal (i.e., vibronic) angular momentum operator about the symmetric top axis z have a significant impact on the rotational energy levels

of a molecule in a ^2E state (Brown, 1971). The doubly degenerate bending state $v_8 = 1$ has the lowest vibrationally excited state in CH_3CN (Müller, Holger S. P. et al., 2016). Radiative transitions can only take place within a *K*-ladder since, according to selection principles, they don't alter the value of *K* while collisional transitions are permitted across K-ladders (Remijan et al., 2004). As a result, collisions have a significant role in determining the population of one K-ladder in relation to another, which makes the kinetic temperature and density of the area a key factor (Remijan et al., 2004; Solomon et al., 1971).

CH_3CN *K* components are closely spaced in frequency but have a wide range of excitation energies above ground (Watson et al., 2002), allowing CH_3CN lines to be observed concurrently at similar sensitivities. This also helps to reduce the impact of calibration errors (Remijan et al., 2004). Overall, CH_3CN is an effective probe of the physical conditions in hot molecular cores.

2.2 Emission and Absorption of Spectral Lines

The observed astrophysical sources through spectral lines gives insight about some of the initial conditions of star formation. The physics behind is essential in understanding the intrinsic strength of the observed emission or absorption of a particular molecule. Circumstellar discs surrounding B-type protostars dominate gravity, and tracers like CH_3CN, which typically live in hot corinos, have been successfully identified in dense and hot gas (Cesaroni et al., 2017, 2014). Additionally, because of its abundance and the ability to simultaneously observe multiple line emissions (Remijan et al., 2004) that occur under various conditions, such as those first discovered and observed by (Solomon et al., 1971) (see Figure 2.1) in the Sagittarius A (Sgr A) and Sagittarius B (Sgr B) molecular clouds, CH_3CN is an ideal molecule for observation. It can be used as well to derive kinetic temperatures in star-forming regions (Müller, Holger S. P. et al., 2016). In order to better understand the chemistry that results in the production of more complex molecules in the hot regions, CH_3CN makes it feasible to get measurements on temperature and column density.

In order to understand the formation of the grains in molecular clouds throughout the phase chemistries it experiences (i.e., warm-gas phase and cold-gas phase), it is helpful to know the column density of complex molecules, such as CH_3CN (Remijan et al., 2004). Different techniques are employed in conjunction with assumptions to determine the column density and temperatures of CH_3CN regions. Remijan et al. (2004) used a logarithmic plot of the normalized column density against the upper energy state. If one knows the excitation temperature, the column density can be deduced for the lower state energy level at the observed optical depth τ, considering absorption spectroscopy (Menten et al., 2010).

$$N_l = \frac{h}{8\pi^2}\frac{g_l}{S\mu^2}\left[1 - e^{\frac{-h\nu}{kT}}\right]^{-1}\tau\Delta\nu \tag{2.1}$$

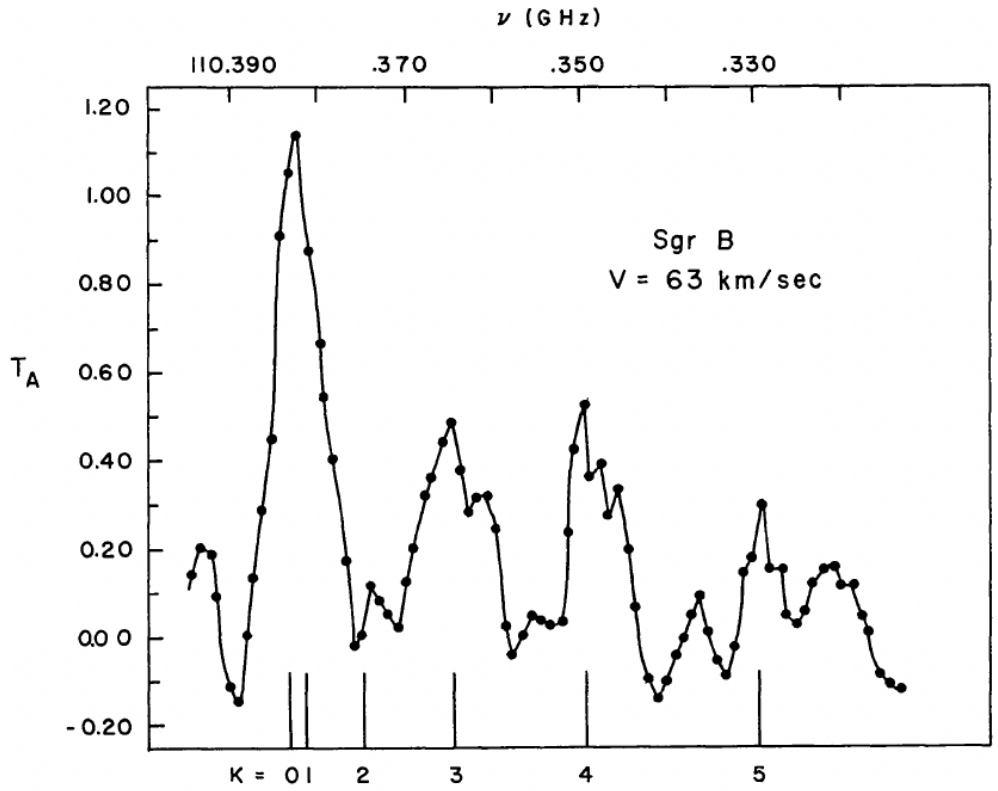


FIGURE 2.1: The first CH_3CN) emission lines detection in Sgr B for the $J = 5 \rightarrow 6$, (Solomon et al., 1971).

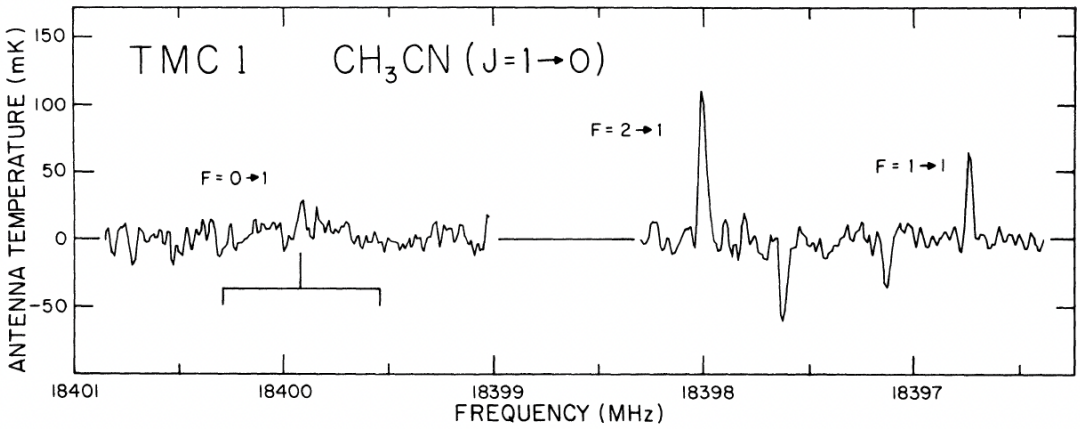


FIGURE 2.2: CH_3CN emission lines detected in the TCM-1 dark cloud for the $J = 0 \rightarrow 1$, $v = 0$ rotational transition at 18.4 GHz, (Matthews & Sears, 1983)

Where N_l is column density at the lower state energy level, τ is the optical depth, ν is the line width, the subscript l represents the lower state energy level, T_{ex} is the excitation temperature while h and k are the Planck and Boltzmann constants.

This is comparable to the formula (Equation 2.2) used by Remijan et al. (2004) to deduce the column densities in the upper state energy levels.

$$\frac{N_u}{g_u} = \frac{3c^2}{\Omega_s 16\pi^3\nu^3}\frac{\int \Delta I d\nu}{S_{ij}\nu^2} \tag{2.2}$$

From Equation 2.2, N_u is the column density at the upper energy level, g_u is the statistical weight of the upper level $(2J+1)$, c is the speed of light in vacuum, Ω_s is the solid angle subtended by the source, $\int \Delta I d\nu$ is the integrated line flux and $S_{ij}\nu^2$ is given by the product of total torsion-rotational line strength and square of the electric dipole moment.

The total column density is given by Equation 2.3 depending on the energy level state i.e., lower state energy level or upper state energy level. Q is the partition function for the rotational temperature T_{rot}.

$$N_{tot} = \frac{N}{g} e^{\frac{-E}{kT}} Q(T_{rot}) \tag{2.3}$$

There are a couple of assumptions that have to be made in order to calculate the total column density depending on the energy state level. (1), we assume there is a uniform physical distribution and the energy levels of the population are described by the Boltzmann distribution. (2) another assumption is optically thin conditions where $\tau_\nu \ll 1$ where we can see the radiation from the observed source. (3) also, assume we are able to measure the source size of the emitting region. (4) we can neglect the background radiation.

2.3 The Spectral Line Formation and Intensity

The line form of the spectral emission and absorption plays a significant role in comprehending some of the processes taking place at the detected source when the emission and absorption of molecules are observed. The spectral lines contain a tremendous amount of data that can be retrieved, including details about temperature, chemical composition, turbulence, and radial velocity. The spectral lines are primarily fitted to a Gaussian distribution, while there are some different ways that largely depend on whether the line is an isolated single line or an asymmetric blend (Trypsteen & Walker, 2017). For emission lines, we can tell whether the emission is from an optically thick or optically thin cloud (Fig. 2.3)

Using radiative energy transfer, the Planck function $B_v(T)$ as a function of temperature, T, can be used to express the intensity of a source (I_v) (Equation 1.1). Although this is not true at low temperatures and high frequencies, the radiation temperature will be equal to the brightness temperature in the Rayleigh-Jeans regime. Our LTE code script

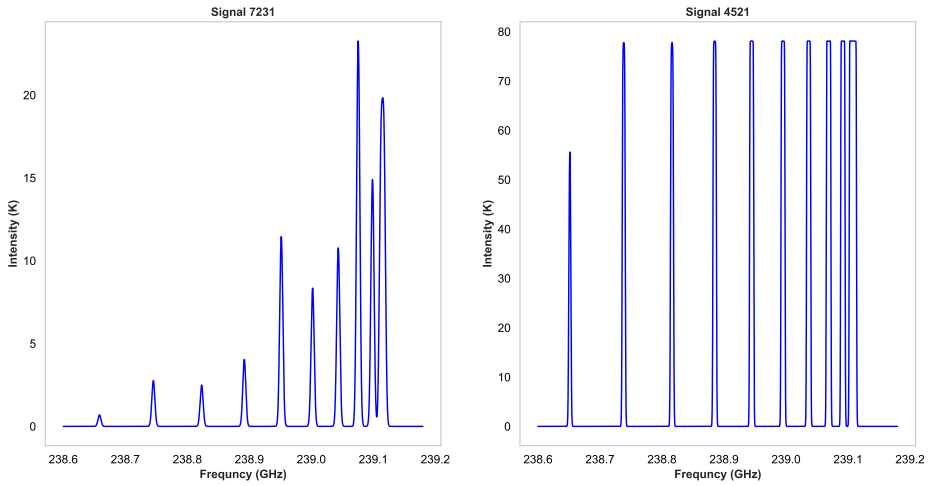


FIGURE 2.3: Spectra from randomly generated parameters (excitation temperature, column density, source size, velocity and line width (velocity dispersion)) showing CH_3CN spectra assuming a locally thermodynamic equilibrium (LTE) environment. The figure on the left shows an optically thin cloud since they have sharp tops while the figure on the right shows an optically thick cloud because of flattened tops.

assumes LTE conditions, meaning the energy levels are filled using the Boltzmann distribution and the gas temperature is the same as the dust temperature in high-density areas.

Some of the physical processes that take place within the detected source and its surroundings have an impact on the spectral lines. The full width at half maximum (FWHM) line profile shape is influenced by the temperature, pressure, density and turbulence effects in cosmic environments (Trypsteen & Walker, 2017). Despite all the figures in 2.3 being generated from a simplified model that mimics real cosmic environment conditions, it can be applied to real observational data of CH_3CN line profiles to build better robust ML models (Frasca, A. et al., 2016).

2.4 Summary

The purpose of this chapter is to provide background information and context on the function of spectral lines in stellar evolution and stellar dynamics. It is important to emphasise the atomic transition from the basic atom to the CH_3CN molecule. Furthermore, diverse synthetic line profiles of CH_3CN spectra were exhibited in Figure 2.3, accompanied by the implications that could be inferred from the emission lines' morphologies.

Chapter 3

Wavelet Transform and Modelling

Introduction

I employ the wavelet decomposition method to decompose CH_3CN spectra. The wavelet decomposition method's approximation and detail coefficients will be thresholded and decomposed while taking into account a number of wavelet families, and the parameter estimate model will then be developed using machine learning (ML).

3.1 The Wavelet Decomposition Method's Characteristics

Wavelet decomposition is a signal compression technique where data is compressed from the original domain by expanding the raw data while retaining much of the original information (Li et al., 2002). Wavelets are robust and provide both the scale (frequency) and time domain of the signal information (Rowe & Abbott, 1995), thus they still maintain the form of data. Furthermore, they are an extension of the Fourier theory (Eq. 3.1) where the Fourier transform $F(\omega)$ of a function $f(t)$ is given by

$$F(\omega) = \frac{\phi}{\sqrt{2\pi}} \int f(t) e^{-i\omega t} dt \tag{3.1}$$

Equation 3.2 is the wavelet transform (WT) equation.

$$W(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} h^* \left(\frac{t-b}{a} \right) s(t) dt \tag{3.2}$$

h^* is the complex conjugate of the wavelet $h(t)$, a represents the scaling term and b describe the a time-shift value. New coherent applications of wavelet transforms based on Fast Fourier Transforms (FFT) have been developed (Sava et al., 1997).

The LTE code script that we used to model the spectra created data that needed to be compressed, which is why we utilised the wavelet decomposition technique. The organic methyl cyanide's line parameters were adjusted to a random selection, which produced the LTE spectra. In addition, wavelets offer data compression while preserving a large portion of the original data. Before applying machine learning to forecast the line parameters, which should be an automatic process, it is helpful to understand the shape

of the spectrum and its parameters. All in all, this work was done to try and anticipate the line parameter values and infer the existence of methyl cyanide in a cluttered spectrum.

3.1.1 Properties of Wavelets

- Computation Complexity - The Wavelet Transform (WT) requires an $O(N)$ multiplication which is based on Mallat's pyramidal algorithm where the space complexity is linear (Li et al., 2002; Sava et al., 1997).
- Vanishing Moments - Wavelets satisfy Eq. 3.1 in a bounded region ω. In other words, the integral of the product of the function $f(x)$ and the low degree polynomial x^j are all zero.

$$\int_{\omega} f(x)x^j dx = 0, \qquad j = 0, 1, ..., n \tag{3.3}$$

  The noisy data can normally be approximated by a low-degree polynomial only if the data is smooth in most of the regions (Li et al., 2002).
- Orthonormal Basis - The standard basis vectors have a unit length and are orthogonal. Particularly, the wavelet transform does not alter the time-based distances between two objects.

3.1.2 Discrete Wavelet Transform (DWT)

The Haar and Daubechies wavelet families have received the most attention in wavelet research (Daubechies, 1992; Cárcamo et al., 2022). Since filter banks are efficient in dividing signals into equal-width frequency sub-bands, DWT is always configured as a filter bank, which means it is provided as a combination of the high-pass and low-pass filters. The detail coefficients and approximation coefficients are the sets of two coefficients that the DWT returns. The approximation coefficients represent the output of the low pass filter (high scale with a low frequency), whereas the detail coefficients represent the output of the high pass filter (low scale with a high frequency). Because of its simplicity and better relevancy of the returned data, the DWT is preferred and utilised more commonly than the continuous wavelet transform (CWT).

Each level of decomposition reduces the signal by half, compressing it while retaining the majority of its original characteristics (Figure 3.1). In addition, to get the majority of a signal's characteristics, this depends on the kind of wavelet family that was used to split the signal up. To separate signal data into its lower resolution components, in the case of spectral data, we used a Daubechies wavelet (db1), which is best suited for signal data. The mother wavelet equation used, Equation 3.2 of order 1, is comparable to the Haar wavelet (Figure 3.2). The plot for the ϕ and ψ functions and in Figure 3.2 is not particularly compelling because it is similar to the Haar wavelet. The Daubechies 2 scaling and wavelet functions are shown in Figure 3.3.

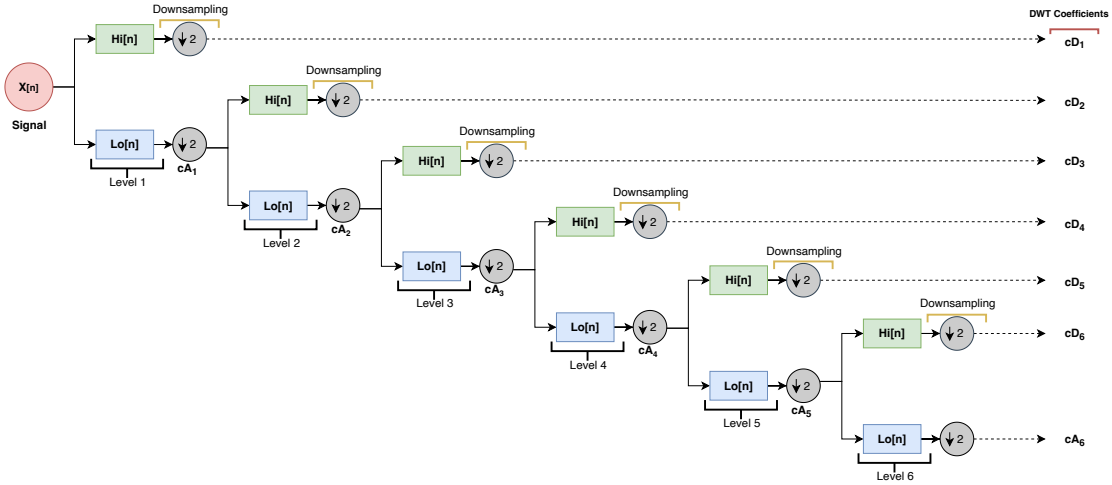


FIGURE 3.1: An illustration of the high-pass and low-pass filters in a multi-level decomposition tree of a signal into six levels. A signal X of length **n** is decomposed by passing through the high pass filter (Hi[n]) and low-pass filter (Lo[n]). $cA_1, ...cA_6$ are the approximation coefficients, while $cD_1, ...cD_6$ are the detail coefficients at level 1-6, respectively.

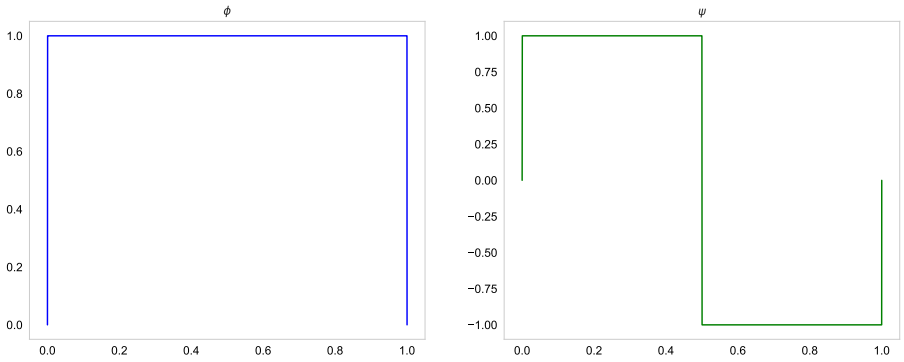


FIGURE 3.2: The Daubechies 1 wavelet [1], db1, is represented by two plots: a scaling function ϕ plot on the left and a wavelet function ψ plot on the right. Asymmetrical, biorthogonal, and orthogonal describe its characteristics. The term "orthogonal" refers to wavelet basis functions or filters that are mutually perpendicular to each other, whereas "biorthogonal" refers to a pair of wavelet bases or filters that are not necessarily orthogonal to each other but have different sets of properties that complement each other.

The pseudo-code for the DWT procedure is displayed in Algorithm 1. Here, we select the wavelet family, level of decomposition which specifies the depth of the decomposition, and signal extension mode that we will employ.

Algorithm 1 DWT pseudo-code for decomposing signals at level 6 using a specified wavelet and obtaining the approximation and detail coefficients.

Require: Data pre-processing
Require: Get feature matrix (c_A and c_D) shape at decomposition level 6
Ensure: c_A and c_D feature matrices of size $[X_n, 422]$
  for $i_x \leftarrow X_n$ **do**
    coefficients $\leftarrow$ decompose(i_x, wavelet, mode) ▷ Apply DWT to data X_n
    $coeff_{arr}, coeff_{slices} \leftarrow$ convert coefficients to array (coefficients)
    $c_A \leftarrow$ coefficients$[i_x :]$ ▷ Approximation coefficients
    $c_D \leftarrow$ coefficients$[i_x + 1 :]$ ▷ Detail coefficients
    reconstructedSignal $\leftarrow$ reconstructSignal(coefficients, wavelet, mode)
  end for
return $coeff_{arr}, coeff_{slices}$

The many discrete wavelet families shown in Fig. 3.3 have various characteristics. Asymmetric, orthogonal, and biorthogonal describe the Haar, Daubechies, and discrete Meyer wavelets. Nearly symmetric, orthogonal, and biorthogonal are Coiflets and Symlets. The Biorthogonal wavelets, however, are symmetric rather than orthogonal.

3.2 Signal Estimation via Thresholding

Since the generated synthetic signals are noiseless, the threshold estimation of the signals in our case is only applicable to observational data that contains noise. The formula $2\sqrt{\log(\mathrm{n})}$, where n is the sample size (Donoho & Johnstone, 1994), is used to estimate the universal threshold value λ, which was adopted from (Tomáš, 2018). The thresholding formula was only used on observational data, not synthetic data. Since only a small number of wavelet coefficients constitute a signal, Tomáš (2018) uses an evolutionary-based method to estimate the threshold and only keeps observations that are more than a multiple of the noise level (Donoho & Johnstone, 1994). The threshold value is applied in one of two ways: soft thresholding or hard thresholding described by the thresholding function $T(x)$ below

$$T_{hard}(x) = \begin{cases} x, & \text{if } |x| > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

$$T_{soft}(x) = \begin{cases} x - \lambda, & \text{if } x < \lambda \\ 0, & \text{if } |x| \leq \lambda \\ x + \lambda, & \text{if } x < -\lambda. \end{cases} \tag{3.5}$$

When using a hard threshold, a substitute zero is used in place of any data values whose absolute value is less than the thresholding value λ . Data values whose absolute value exceeds the thresholding are unaffected. If the data values are less than the

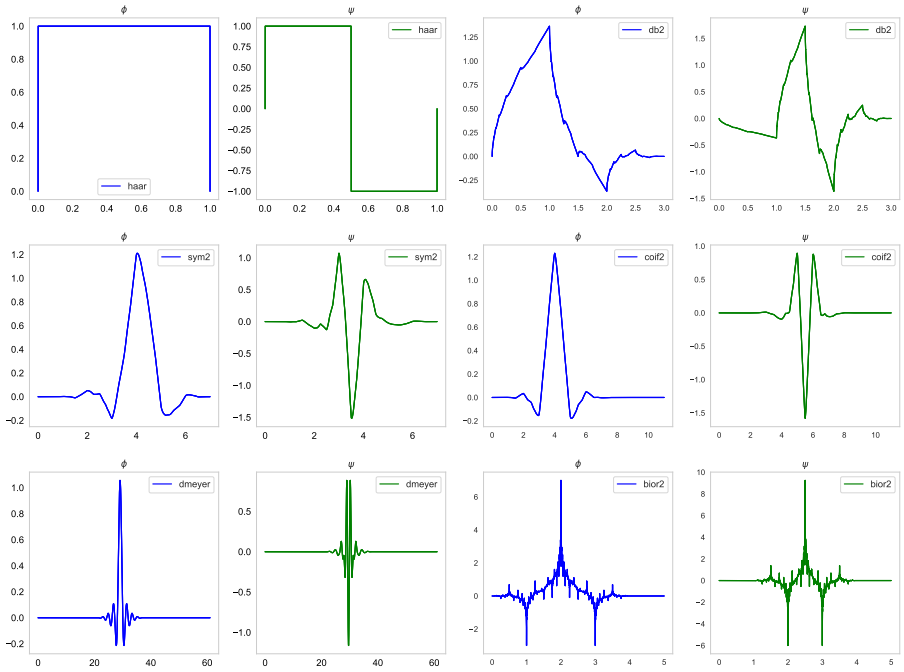


FIGURE 3.3: A schematic of some of the common discrete wavelets types Haar (haar), Daubechies (db2), Symlets (sym2), Coiflets (coif2), Discrete Meyer (dmey) and Biorthogonal (bior2). The blue plots show the scaling function ϕ and the green plots show the wavelet function ψ. The type of wavelet is labelled in the upper right of each panel.

thresholding value, soft thresholding subtracts the thresholding value from the values. A substitute zero is used in place of any data values whose absolute value is less than or equal to the thresholding value. Finally, the threshold value is added to the data values in cases where they are less than the negative threshold value. The following thresholding procedures were used to denoise the observational data from Tomáš (2018).

1. After preprocessing the data, use DWT to extract the wavelet coefficient—that is, the approximation coefficients (c_A) and detail coefficients (c_D)—from the signal.
2. Use Mean Absolute Deviation (MAD) to approximate the thresholds $T_{hard}(x)$ and $T_{soft}(x)$ at the selected level of decomposition.

$$\sigma = \frac{1}{0.6745}\text{MAD}(|c_A|) \tag{3.6}$$

   σ: This is the standard deviation of the coefficients at the selected level of decomposition.

$$T_A = \sigma\sqrt{2log(n)} \tag{3.7}$$

   T_A: This is the threshold value for the approximation coefficients c_A at the selected level of decomposition.

3. Apply thresholding to the DWT-derived approximation coefficients.
4. From the thresholded coefficients, recreate the original signal's denoised form..

The noisy observational data are then subjected to this after the thresholding method previously described, as illustrated in algorithm 2. The key objective of this is to have data that closely resembles the synthetic data that our ML models were trained on, as this will allow for improved parameter estimates.

3.3 Modelling of CH_3CN Spectra

The ALMA telescope Band 6 configuration was used to model the CH_3CN spectrum using an LTE code script that uses Centre d'Analyse Scientifique de Spectres Instrumentaux et Synthétiques (CASSIS) Vastel et al. (2015) database to generate synthetic data spectra at a frequency range of 238.6 MHz to 239.18 MHz. Synthetic data were used because there were insufficient amounts of observational spectral data from ALMA to use. The input physical parameters for these models include excitation temperature, column density, source size, velocity, and line width (velocity dispersion), were all uniformly generated at random and placed within the appropriate ranges consistent with the results from other observations shown in Table 3.1 (Pols et al., 2018; Andron et al., 2018).

Algorithm 2 This algorithm performs denoising and thresholding on feature matrices c_A and c_D obtained from wavelet decomposition at level 6. The algorithm calculates the sigma of c_A using the median absolute deviation (MAD) and applies thresholding to both c_A and c_D using T_A, a threshold value based on the calculated sigma and the size of the signal. The thresholding function used for c_A and c_D is different, with $T_{hard}(x)$ used for c_A and $T_{soft}(x)$ used for c_D. The resulting denoised approximation and detail coefficients are returned.

Require: Get feature matrix (c_A and c_D) shape at decomposition level 6
Ensure: $c_A = [X_n, 422]$ and $c_D = [X_n, 422]$
  for $i_x \leftarrow X_n$ **do**
    coefficients = $decompose[i_x,\ wavelet,\ mode] \leftarrow X_n$
    $coeff_{arr}, coeff_{slices} = coefficients[i_x] \leftarrow X_n$ ▷ convert the coefficients to an array
    $\sigma = (1/0.6745)\mathrm{MAD}(|c_A|)$ ▷ calculate sigma of the coefficients
    $T_A = \sigma\sqrt{(2 \times log(n))}$
    apply thresholding to detail and approximation coefficients
    $c_A \leftarrow coefficients[i_x :] = \mathrm{threshold}(coefficients[0], T_A, T_{hard}(x))$
    $c_D \leftarrow coefficients[i_x :] = \mathrm{threshold}(coefficients[1 :], T_A, T_{soft}(x))$
  end for

Parameter	Range	Units
Excitation temperature	10 - 400	K
Column density	10^{14} - 10^{18}	cm^{-2}
Source Size	0.2 - 1.1	arcsec
Velocity (V_{LSR})	-50 - +50	km s^{-1}
FWHM	1- 10	km s^{-1}

TABLE 3.1: Input physical parameters space

Figure 3.5 shows some of the shows the intensity-frequency profiles (spectra) of CH_3CN and the corresponding randomly generated parameters used to generate them (same as Figure 2.3). A high column density, temperature, and line width optically thin observed source is shown in the figure on the left. The molecular gas temperature, which can be discovered through spectral fitting, governs the line intensity ratios between various energy transition components (Hung et al., 2019). This indicates that CH_3CN lives in the cold to hot environment based on the temperature range. In contrast to the figure to the right, where the column density is higher, which is causing the spectral lines' tops to flatten, which are indicators that this is from an opaque cloud. Additionally, it should be noted that the intensity is higher for the optically thick plot, it is lower for the optically thin figure.

3.4 LTE Model Fits and Model Limitations

The synthetic CH_3CN (J=13–12) transition molecular spectra are produced by our LTE code script. When used with observational data, the LTE model is unfortunately constrained, thus we make assumptions regarding the emitting gas. The derived column

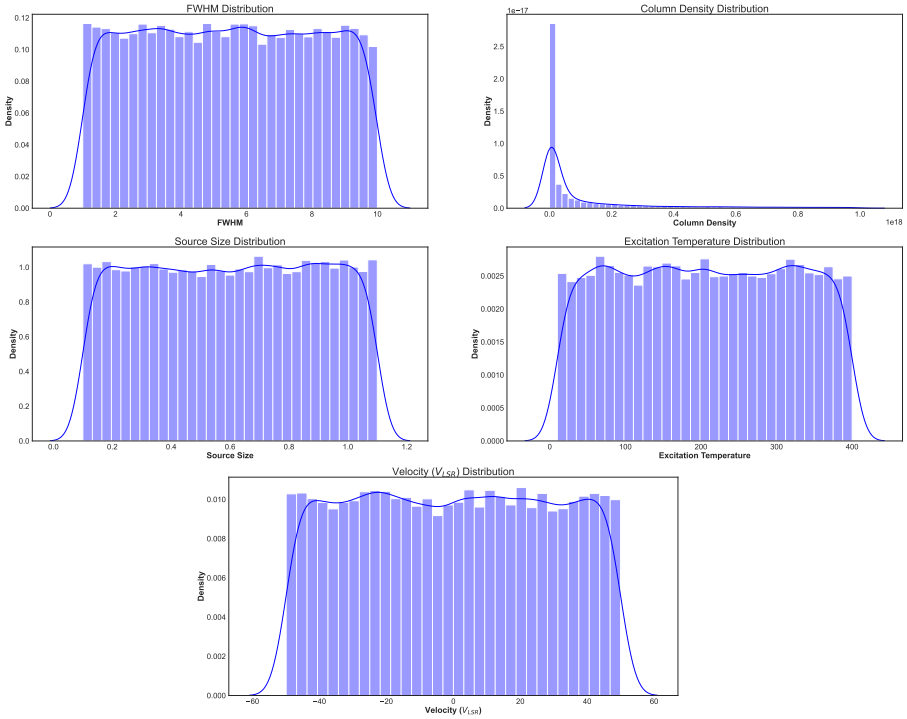


FIGURE 3.4: A representation of the distribution of all the parameters used to create the artificial frequency-intensity plots. All of the parameters have a uniform distribution, with the exception of the column density, which is most influenced by the gas's excitation temperature, which is frequently in charge of the line intensity ratios between various K components.

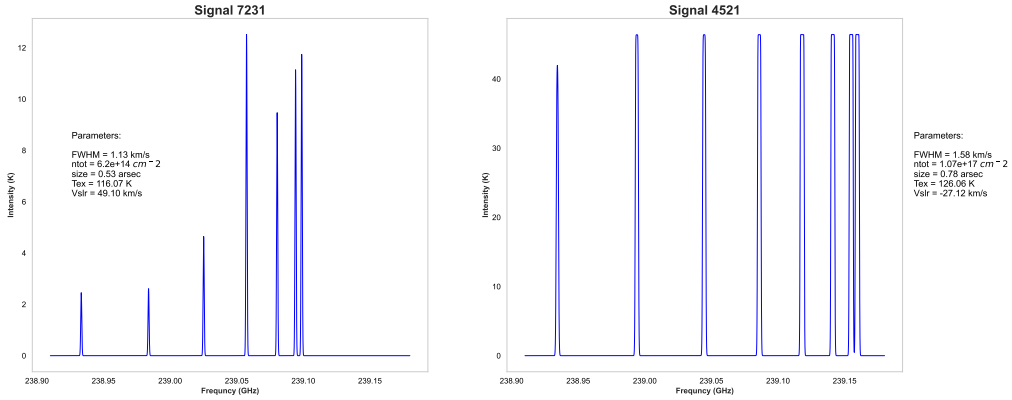


FIGURE 3.5: An illustration of the CH_3CN spectra's generated intensity-frequency profiles and the matching randomly generated parameters.

densities are beam-averaged values and the beam dilution of the source in the telescope beam is taken into account. The LTE approach also has a further drawback because it relies on the assumption that the emitting gas will always have the same temperature and abundance and that each transition line's emission would come from a source of the same size (Bell et al., 2014).

Equation 3.8, in which θ_b is the beam size in arcseconds, c is the speed of light, ν is the frequency, and B is the telescope's interferometer baseline in meters, provides the beam size from our LTE code script (for ALMA, this is the diameter of the interferometer)

$$\theta_b = \frac{1.22 \times c/\nu}{B \times 3600 \times \frac{180}{\pi}} \tag{3.8}$$

Equation 3.9 provides the dilution factor (f_b), assuming that the source size's (θ_s) geometry and the telescope's beam size's (θ_b) geometry are both represented by a 2D Gaussian function.

$$f_b = \frac{\theta_s^2}{\theta_s^2 + \theta_b^2} \tag{3.9}$$

Additionally, in the presence of a continuum source, the gas chunk will not only emit photons but also have the potential to absorb them, creating profiles for both absorption and emission lines (Martín et al., 2019). The LTE code script provided by the CASSIS team, (Vastel et al., 2015) requires knowledge of the molecular energy levels and transition parameters in order to calculate the line transitions using the LTE approximation, hence we used the CASSIS database together with an ALMA 400 m telescope configuration file.

Chapter 4

Application of Machine Learning in Astronomy

Introduction

Machine learning (ML), also known as statistical learning, is the process of deriving conclusions from data using statistical models. There are many different ML algorithms that can be used in astronomy, ranging from classification and regression issues applied to natural language processing, computer vision, and ensemble learning, which are the main topic of this chapter. With the increasing large volume of data in astronomy, the use of ML techniques has proliferated with the application of various algorithms to these domain-specific fields, like astronomy, physics, biology, etc. This has made it possible for astronomers to quickly analyse the ever-growing amount of data that is now at their disposal from instruments all over the world and beyond.

In this chapter, we will describe the machine learning methods that will be used to predict physical parameters. The ensemble learning algorithms with multi-regression analysis would be the focus. The metrics for evaluating the methods used will also be described.

4.1 Ensemble Learning Algorithms

Ensembling is the synthesis of various machine learning models and predictions. These are ML algorithms that use labelled datasets as features (columns) to train them in pattern recognition before applying the learned model to forecast the behaviour of new dataset features. The majority of ML algorithms simply multiply mathematical vector equations together to perform complex computations. Depending on the kind of issue that needs to be resolved, ensemble learning can be divided into classification and regression issues. In this study, we use ensemble learning algorithms for multi output regression, where we simultaneously predict several numerical targets, i.e., physical parameters (excitation temperature, source size, column density, and velocity gradients). One regression model

is fitted for each target that we want to make predictions on in this method. Additionally, there are numerous ways to use the ensemble learning technique; however, in this work, we'll focus on bagging and boosting. Bagging, a term coined by Breiman (2001), is a "bootstrap" ensemble strategy that cultivates individuals for its ensemble by training each classifier on a training set that has been randomly rearranged (Opitz & Maclin, 1999). While boosting is an algorithm that repeatedly executes a "weak" or "basic" learning method while feeding it a different subset of the training data or a different distribution or weighting over the training examples (Schapire, 2003). The Random Forest (RF) method is used for bagging (Figure 4.2), and the Extreme Gradient Boosting (xgboost) technique is used for boosting. Figure 4.1 depicts the layout of our solution, from data preprocessing through parameter estimations using the ensemble tree algorithms.

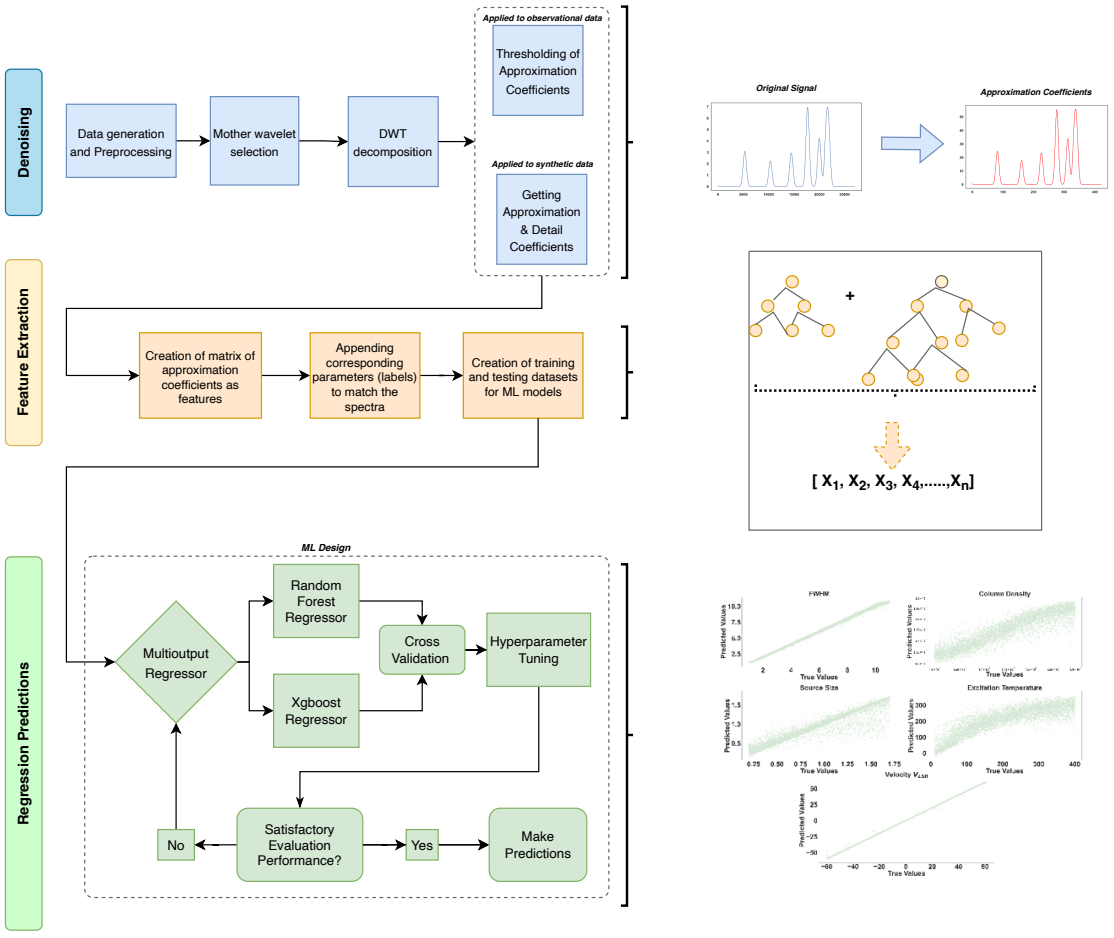


FIGURE 4.1: A visual representation of all DWT operations and ML model forecasts.

4.1.1 Random Forest (RF)

RF is based on ensemble trees by which the most popular class is selected as proposed by Breiman (2001). In other words, it's built on the application of an ensemble of classification and regression tree (CART) like classifiers in which their learning is performed

on the boosted-averaged observations (Tomáš, 2018). The mathematical representation of RF is expressed as;

Consider a set of tree-classifiers, i.e., $\{h_1(\mathbf{x}), h_2(\mathbf{x}), h_3(\mathbf{x}), ..., h_n(\mathbf{x})\}$ each of which casts a unit vote for the most popular class at input $\mathbf{x}$. The training set is randomly selected from a distribution of random vectors $\mathbf{U}_n$. The way $\mathbf{U}_n$ is used in the construction of trees determines its nature and dimensions.

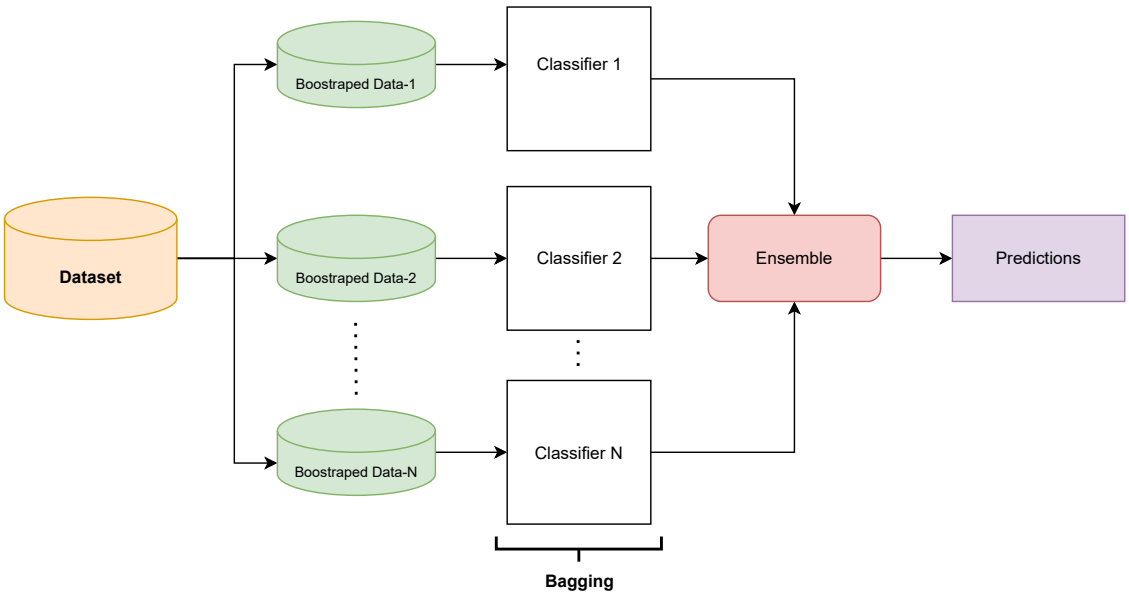


FIGURE 4.2: An illustration of the mechanism used by the bagging ensemble. The data is bootstrapped into smaller random samples of the population in N-dimension using replacement, and then average N-dimension independent classifiers are used to make predictions from the bootstrapped data (bagging).

The bootstrapping mechanism, also known as random sampling with replacement, has its roots in statistics (Tomáš, 2018; Efron & Tibshirani, 1994). When using random features, bagging appears to improve accuracy. Additionally, bagging can be used to continuously provide estimates for the strength and correlation of the ensemble of combined trees' generalisation error (PE) (Breiman, 2001). At each node split, the number of features to be considered can be tuned.

4.1.2 Extreme Gradient Boosting (xgboost)

The extreme gradient boosting (xgboost) algorithm is described as an optimised ML system for tree boosting (Chen & Guestrin, 2016). For xgboost, the models (classifiers) that follow the first model are trained on error residuals of misclassified data to reduce errors of the preceding models; as a result, they learn the data well and are frequently prone to overfitting. Given a dataset with n samples and m features, $D = \{(x_i, y_i)\}$ $(|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$. K additive functions are used by this ensemble tree model to forecast the results.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$
$$\text{where} \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{4.1}$$

Here, the difference between the prediction ($\hat{y}_i$)) and the true value (y_i) is measured by the differentiable convex loss function (l), T is the number of leaves in the tree, w is the leaf weight. Since the loss function is convex, we can use gradient descent, an optimisation algorithm, to identify the weights that reduce it. The second term Ω penalises the model's complexity (i.e., the regression tree functions) (Chen & Guestrin, 2016). Each f_k represents a separate tree structure q where ($q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T$). γ and λ are constant terms. In order to prevent overfitting, the additional regularisation term helps to smooth the final learned weights.

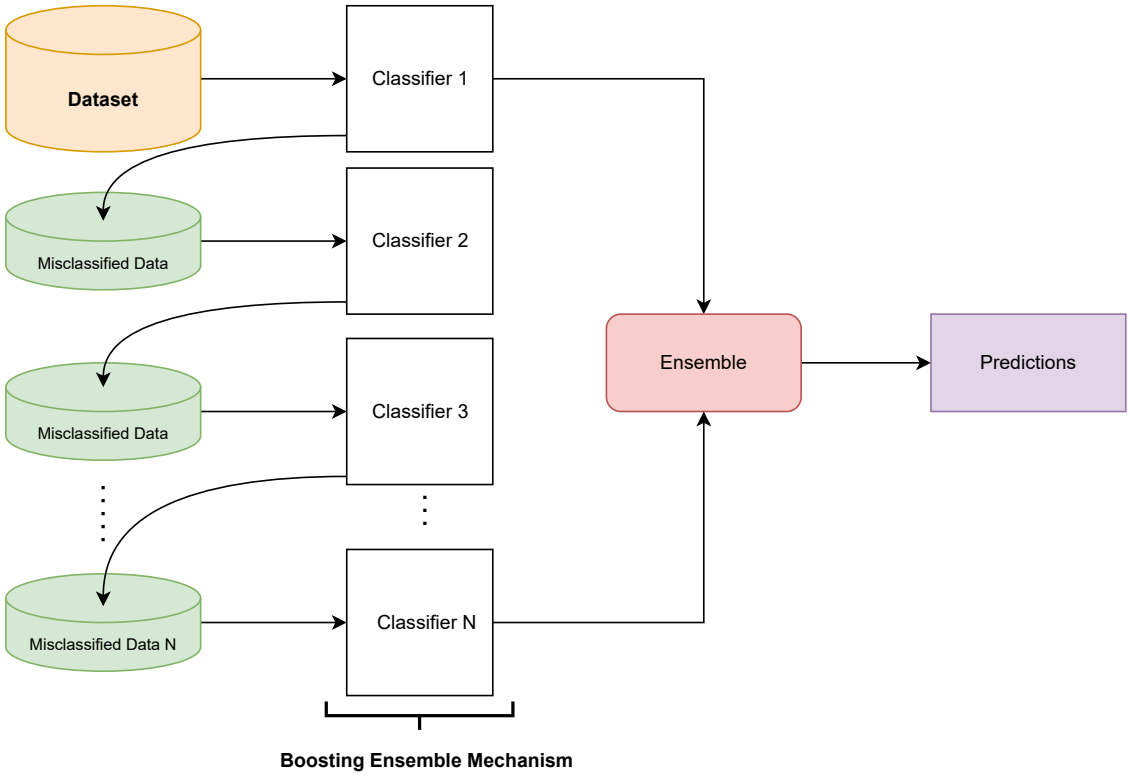


FIGURE 4.3: A representation of the bagging ensemble's mechanism. To make predictions on the dataset with increased sample weight, a number of weak learners are employed. The next decision tree receives the weighted data (misclassified data) as a result of this action by the model.

4.1.3 Cross Validation

Cross-validation is a technique of evaluating the performance of an ML model and testing its performance. It is a member of the family of Monte Carlo techniques (Berrar, 2018). Cross-validation is frequently used to calculate a model's prediction error (Bates

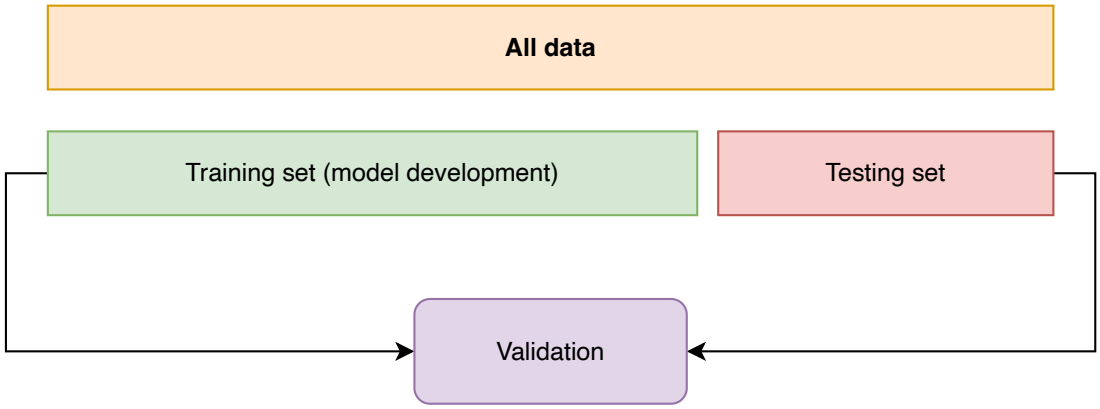


FIGURE 4.4: A diagram showing the division of the data into training and testing sets. The training dataset is used for model building, and both the training and testing sets are used for evaluation.

et al., 2021). It is frequently relatively simple to construct a model that precisely fits the available training data set but fails to generalise adequately to unseen data. Most often, this results in model *over-fitting* and *under-fitting*. Over-fitted models merely duplicate the training data instead of modelling the general pattern, which is what happens when a model mimics the training data set rather than illuminating the underlying pattern. Therefore, it is suggested to attempt to strike a balance between the two so that the model works well by generalising on unexplored data by assessing our model on new data from the same population as our ML model is trained. This provides a fair assessment of what the model will look like when it is used to make predictions in the real world. To test ML models, there are numerous validation, (e.g., including *train/test split* (Figure 4.4)) and cross-validation techniques, *k-fold cross validation* such as , *Repeated k-fold cross-validation*, *Leave P Out (LPO)*, etc. Whether a classification or regression method is used in a cross-validation process depends on the issue being addressed. In order to offer an estimated performance of the final model on a fresh, untested dataset, cross-validation is a crucial step in the development of ML models (Berrar, 2018).

The model is most likely over-fitted if the validation error is bigger and the training error is smaller. A model is probably under-fitted if it has a significant training and validation error. The model probably accurately depicts the relationship between the predictors and response if the validation error is modest. For tree-based models, the validation error set approach is used to locate smaller trees nested within the entire tree that outperform the bigger tree on the validation set in order to find a suitable tree model for the data set.

We used a *test/train split validation* and *k-fold cross-validation* to assess our models. In k-fold cross-validation, the training dataset is portioned out into smaller subsets of roughly equal sizes (Vabalas et al., 2019). The training data were used to repeatedly perform our *k*-fold cross-validation procedure ten times. Each time, a new one-tenth of the data was chosen to verify our model. The mean scores of all performances in each of the ten validation folds were used to calculate the model performance.

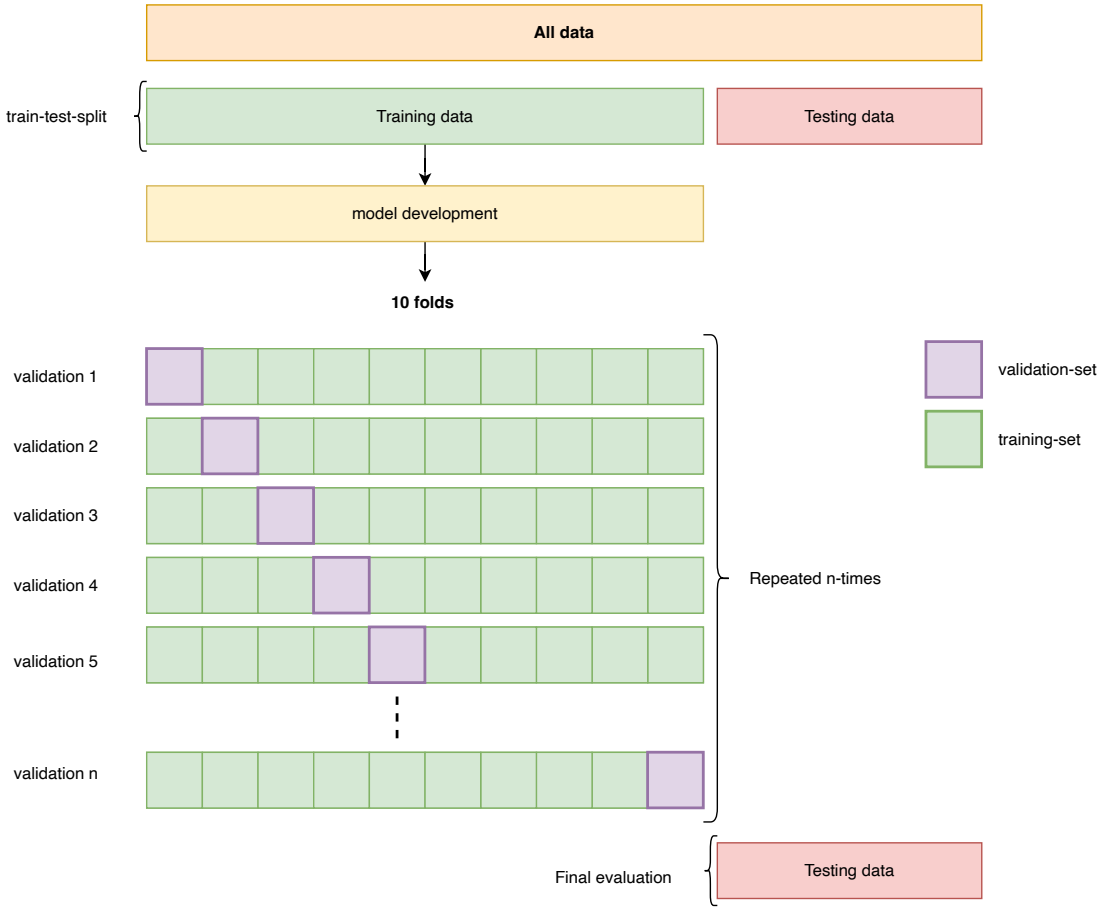


FIGURE 4.5: An illustration of a 10-fold cross-validation used to evaluate our ML models.

To represent the k-fold mathematically, let $\hat{f}_k$ denote the model trained on all of the k^{th} subset of the training set. The predicted value $\hat{y}_i = \hat{f}_k(x_i)$ of the true value y_i of $x_i \in k^{th}$ subset. The cross-validation error, $\hat{\epsilon}_{CV}$ which is the mean scores of all the splits used is given by;

$$\hat{\epsilon}_{CV} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\hat{f}_k(x_i), y_i) \tag{4.2}$$

The loss function used to measure the estimating error in this instance is $\mathcal{L}$. The training and testing errors are then computed using this.

4.1.4 Hyper-parameter Optimisation

The process of locating the optimum ideal model architecture in ML models is known as hyper-parameter optimisation, or hyper-parameter tuning, occasionally. During the ML training process, these parameters cannot be changed (Yu & Zhu, 2020). Grid search hyper-parameter optimisation (GSO), which seeks the ideal ML model parameter values, is the most often used automatic technique for hyper-parameter optimisation (Bergstra & Bengio, 2012; Tomáš, 2018). The majority of the time, a user specifies a set of parameters they wish to optimise, which largely depend on the kind of ML model they are using as

well as the kind of training data they have at their disposal in order to perform the optimisation. In our case, we concentrated on fine-tuning some of the xgboost algorithm's parameters, including *n estimators* (the number of trees), *max depth* (the maximum depth of trees), *colsample bytree* (the ratio of sub-sample columns when building each tree), *learning rate* (step size shrinkage to update weights to prevent overfitting), *min child weight* (in linear regression tasks, a child must have the bare minimum of instance weight (hessian), the term "hessian" refers to the second-order derivative of the loss function with respect to the model's parameters. This is the bare minimum number of instances that must reside in each node) and *objective* (specifies the function we want to minimise).

Various optimization strategies exist, including both *manual* and *automated searches*. Additionally, GSO is so easy to construct that parallelization is negligible, this means that utilising GSO, it is very easy to parallelize the optimization process with little to no additional processing resources needed. In other words, the optimization process doesn't incur much more overhead as a result of parallelization. Moreover, it frequently discovers better optimisation values faster than manual sequential optimisation when using a compute cluster, and it is trustworthy in low-dimensional spaces, such as 1-D and 2-D spaces (Bergstra & Bengio, 2012).

4.1.5 Performance Criterion

The effectiveness of ensemble learning algorithms like the supervised learning algorithms (RF and xgboost) used is no different from that of other ML models, which must all be evaluated in order to determine how well they solve the problem at hand. In this section, we introduce a few of the evaluation metrics that ML algorithms frequently use to evaluate regression problems. Both multiple linear regression and simple linear regression use the same evaluation metrics. *Error* is a typical regression metric (Eq. 4.3) that is straightforward and simple to comprehend. In light of this, this study suggests the following special performance standard:

$$\text{Error} = \text{True Value} - \text{Predicted Value} \tag{4.3}$$

Mean Absolute Error (MAE)

Mean Absolute Error calculates the average discrepancy between the predicted values and true values. For most regression problems in statistics, the MAE formula is given by;

$$\text{MAE} = \frac{1}{n}\sum_{i}^{n}|y_i - \hat{y}_i| \tag{4.4}$$

where y_i and $\hat{y}_i$ are the true and predicted values of the $i-th$ data. n is the number of predictions. This tells us how closely our predictions match the true values and how much of a deviation there is. MAE is not very sensitive to outliers since it does not punish

huge errors. In order words, it gives a linear value, which averages the weighed individual differences equally. The lower the MAE value, the better the model, depending on the scale of the values used in the training.

Mean Squared Error (MSE)

MSE is the average squared differences between the true values and the predicted values. It always gives a positive value and the more the value is is closer to zero or a lower value, the better the model.

$$\text{MSE} = \frac{1}{n}\sum_{i}^{n}(\hat{y}_i - y_i)^2 \tag{4.5}$$

Similar to MAE, the $i-th$ data's true and predicted values are denoted by y_i and $\hat{y}_i$, respectively, and n is the number of predictions. It is one of the most widely used metrics for regression issues, but it is least helpful when a single poor prediction would undermine the predictive power of the entire model, particularly when the dataset is noisy, i.e., it is more sensitive to outliers than MAE.

Root Mean Square Error (RMSE)

RMSE is the square root of the average squared differences between the true and predicted values. In other words, the squared errors are squared before averaging them. This basically means that the RMSE gives large errors a higher weight because they have a much bigger impact on the model's performance.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2} \tag{4.6}$$

where y_i and $\hat{y}_i$ are the true and predicted values of the $i-th$ data. n is the number of predictions. It quantifies the error between true and predicted values. The performance of the model is also inversely correlated with this metric's value.

Coefficient of Determination (R^2)

The coefficient of determination (R^2) introduced by Wright (1921) is the amount of the dependent variable's variance (predicted values) that can be predicted from the independent variables (true values) (Chicco et al., 2021).

$$\text{R}^2 = 1 - \frac{\sum_{i}^{n}(\hat{y}_i - y_i)^2}{\sum_{i}^{n}(\bar{y}_i - y_i)^2} \tag{4.7}$$

Here, y_i and $\hat{y}_i$ represent the $i-th$ value's true and predicted values, n denotes the number of predictions, and $\bar{y}_i$ represents the mean of the true values. When the value is negative, R^2's model performance suffers, and when it's in the positives, it performs

best. Therefore, there's a limitation when the R^2 is in the negative space, since it does not indicate the degree at which the model performs poorly.

Mean Absolute Percentage Error (MAPE)

MAPE, which has a very straightforward interpretation in terms of relative error, is another performance indicator for regression models. Its use is therefore suggested for tasks where sensitivity to relative variations rather than absolute variations is more important (Chicco et al., 2021). It utilizes most of the information pertaining to the error (Tayman & Swanson, 1999).

$$\text{MAPE} = \frac{1}{n}\sum_{i}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4.8}$$

Similar to other metrics, y_i and $\hat{y}_i$ stand for the true and predicted values for the $i-th$ value, respectively, while n stands for the number of predictions. In multivariate regression analyses, such as the one used in this study, MAPE provides the percentage error on each predicted parameter.

4.1.6 Residual Analysis

Residuals in multiple regressions analysis is a useful class of method for the assessing of the accuracy of a fitted model (Topp & Gómez, 2004). In conventional regression analysis, we frequently presume that a relationship exists for a certain data set (Zuo, 2022). Predicted values are plotted on the x-axis in residual plots, and residuals, $r_i = y_i - \bar{y}_i$ are plotted on the y-axis. The true and expected values for the $i-th$ value are represented by y_i and $\hat{y}_i$, respectively. The deviation from zero indicates how inaccurate the forecast was for that number; for example, positive residuals (on the y-axis) indicate low residuals, while negative residuals indicate high residuals, while zero indicates a perfect guess.

Based on the residuals plots of the highest value from the centre and the lowest value from the centre, we can determine the variance of our predictions. In other words, if the greatest point value and the lowest point value (along the y-axis) are significantly different, then the variance is higher; if the converse is true, and if all of the residuals are along the zero line along the y-axis, then the model is perfect and has no variance.

4.1.7 Summary

The purpose of this chapter was to discuss and describe machine learning (ML) methods which can be used in astronomy, with a particular emphasis on spectral analysis of organic molecules (methyl cyanide) from the ALMA telescope. The physical parameter predictions were done using a variety of ensemble ML algorithms. Random forest and extreme gradient boosting were the two ensemble learning algorithms used to generate the multivariate predictions of the five physical parameters. The topic of cross-validation, which is frequently used to evaluate the model's performance, was covered. It can be difficult to determine the ideal model architecture. The most popular hyper-parameter

approach, GSO, was explored. Different performance metrics were employed to rate the performance of our models. The most effective assessment metrics for regression algorithms appear to be RMSE and R^2.

Chapter 5

Results and Discussion

Introduction

The results of the study are discussed and presented in this chapter. A description of the ML algorithms used can be found in chapter 4. All algorithms were evaluated after cross-validation and hyper-parameter tuning using the evaluation metrics covered in section 4.1.5 of chapter 4. This chapter applies the study's training data, algorithms, and ML models to actual observational data and discusses the results. Afterwards, the constraints of our models are then discussed, along with the observational data.

5.1 Training Data

The results of the simulation runs are used to create the training and test data for our models due to the lack of sufficient observational data. To generate synthetic data for CH_3CN through the software package CASSIS database, we used a script written in Python with the ALMA 400 metres telescope configuration, assuming an LTE environment with a cosmic microwave background temperature (T_{CMB}) of 2.75 K. Furthermore, data was generated for 40,000 simulation models to replicate the CH_3CN emission physical conditions under various circumstances. The input physical parameters used to create the spectra plots for CH_3CN were listed in chapter 3's table 3.1. Figure 3.4 displays the parameter distribution for the range of input physical parameters that were used to create our spectra plot. Some of the CH_3CN spectra that were generated under the assumption of an LTE environment are depicted in Figure 5.2. Our data was divided into a 70% training set and a 30% testing set after preprocessing and compressing it using a DWT technique. We had set the features as the approximation coefficients from using DWT with a shape of 422 values of each spectra feed into the our ML regression models to make prediction on the five physical parameters.

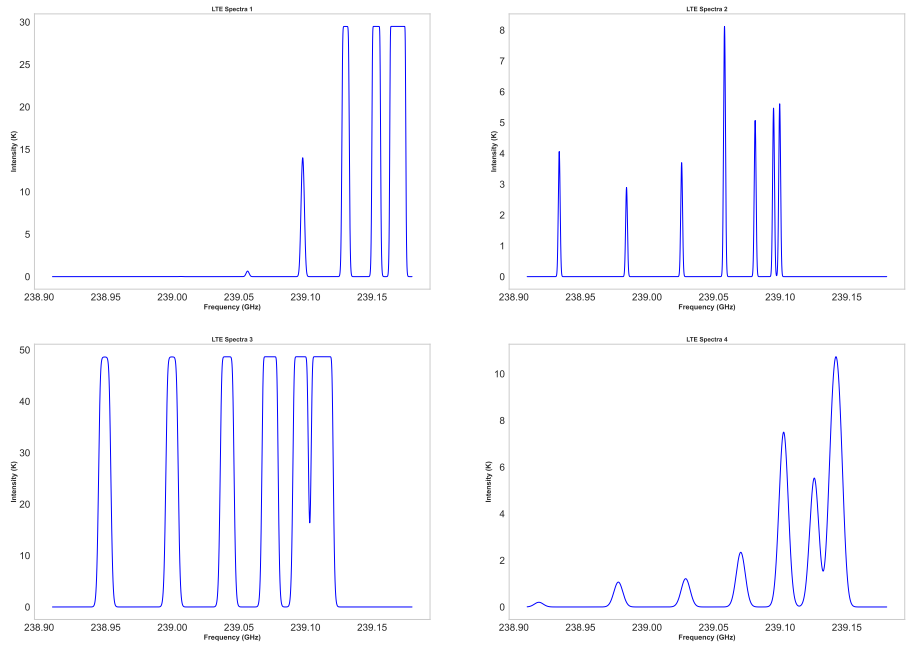


FIGURE 5.1: Several CH_3CN spectra produced by the LTE code script.

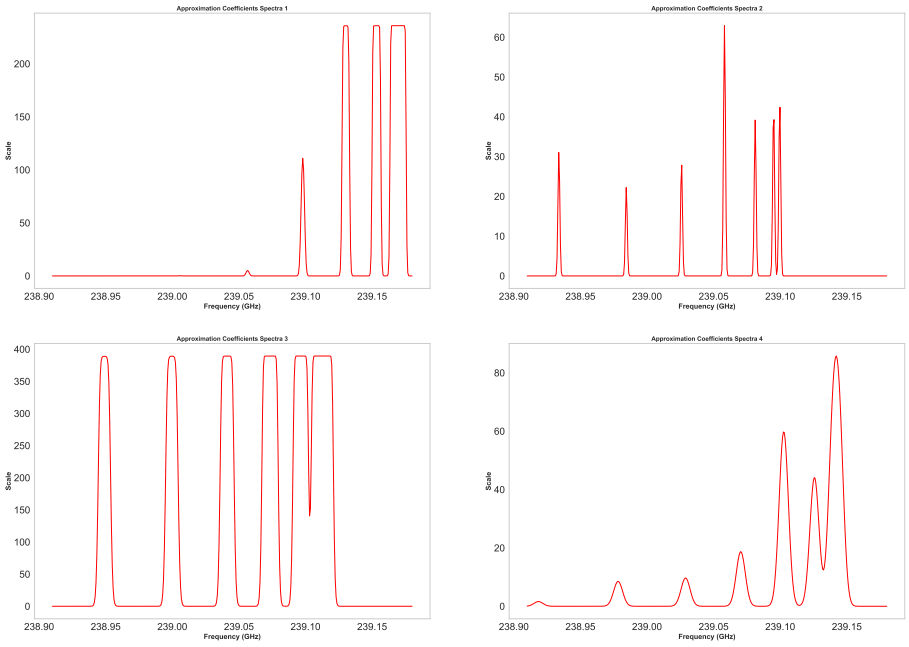


FIGURE 5.2: Several CH_3CN spectra of the approximation representations from the DWT method.

5.2 Observational Data

Figure 5.3 displays the observational data for making predictions, which includes 30 sources. The ALMA telescope provided the CH_3CN observational data in the vibrational ground state (J=13–12) in band 6, and the spectra's frequency range was between 238.91 GHz and 239.18 GHz, with ALMA synthesized beam (θ_{syn}) $\approx$ 0″.8. The data were obtained from the ALMA project 2015.1.01312.5. Details of the observations and their calibration are given in Avision et al (2023, submitted) and also in Asambre Frimpong et al (2023, submitted). This was carried out in order to match the frequency range that our models' training data set employed. Because the observational data is noisier than the synthetic data generated by our LTE code, thresholding will be necessary in order to obtain spectra that are significantly closer to those utilised during training (see algorithm 2).

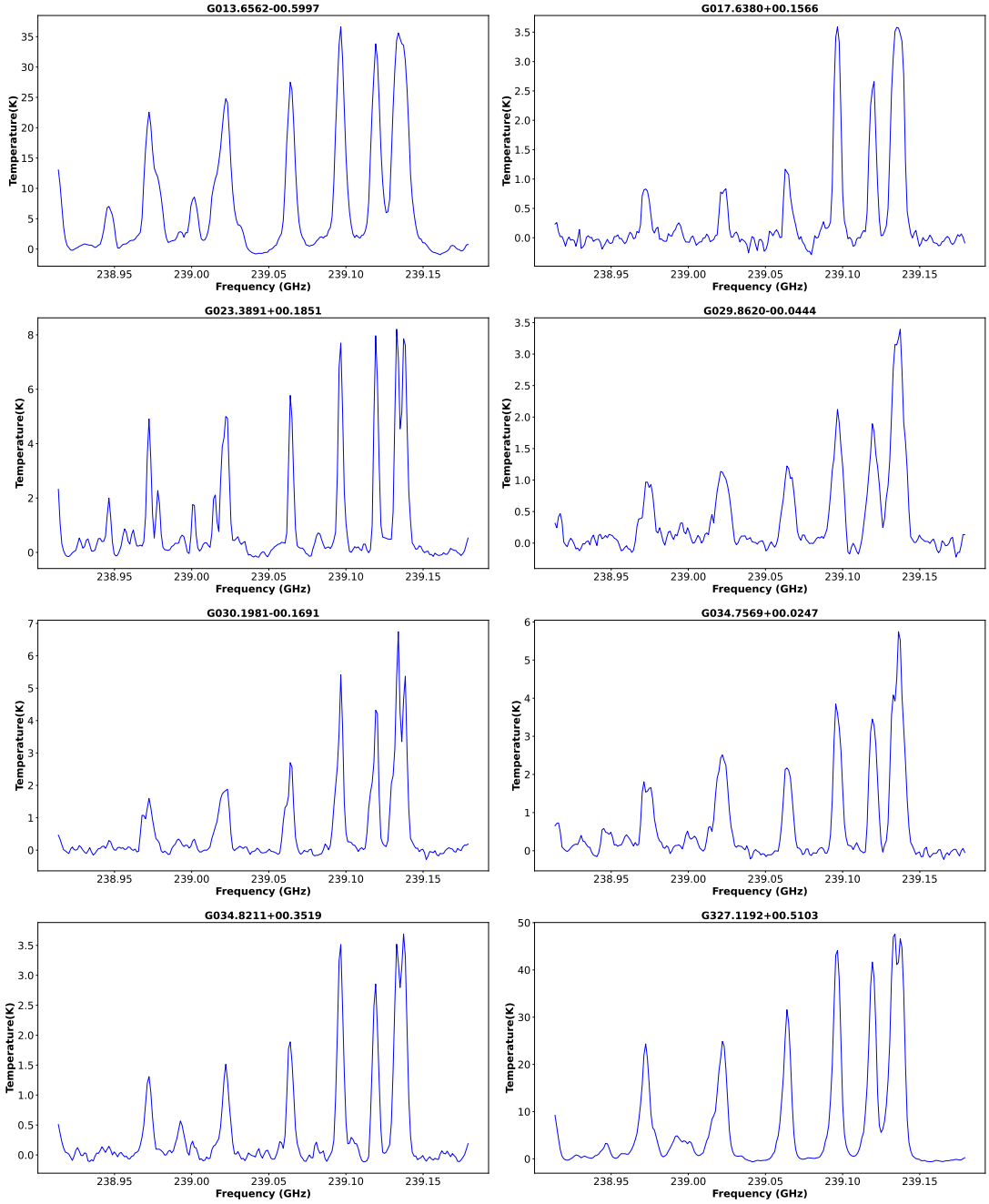


FIGURE 5.3: Observational data of some of the CH_3CN spectra. The name of each source is shown for each panel.

Although there were 236 data points from each source in Figure 5.3, this is fewer than the number of data points required by our machine learning models to make predictions. Using a one-dimensional cubic linear interpolation to match 422 data points, we interpolate our data to have equal amounts of data points. Figure 5.4 illustrates this; when plotted together, the sources in Figure 5.3 match precisely since there are no significant differences between them.

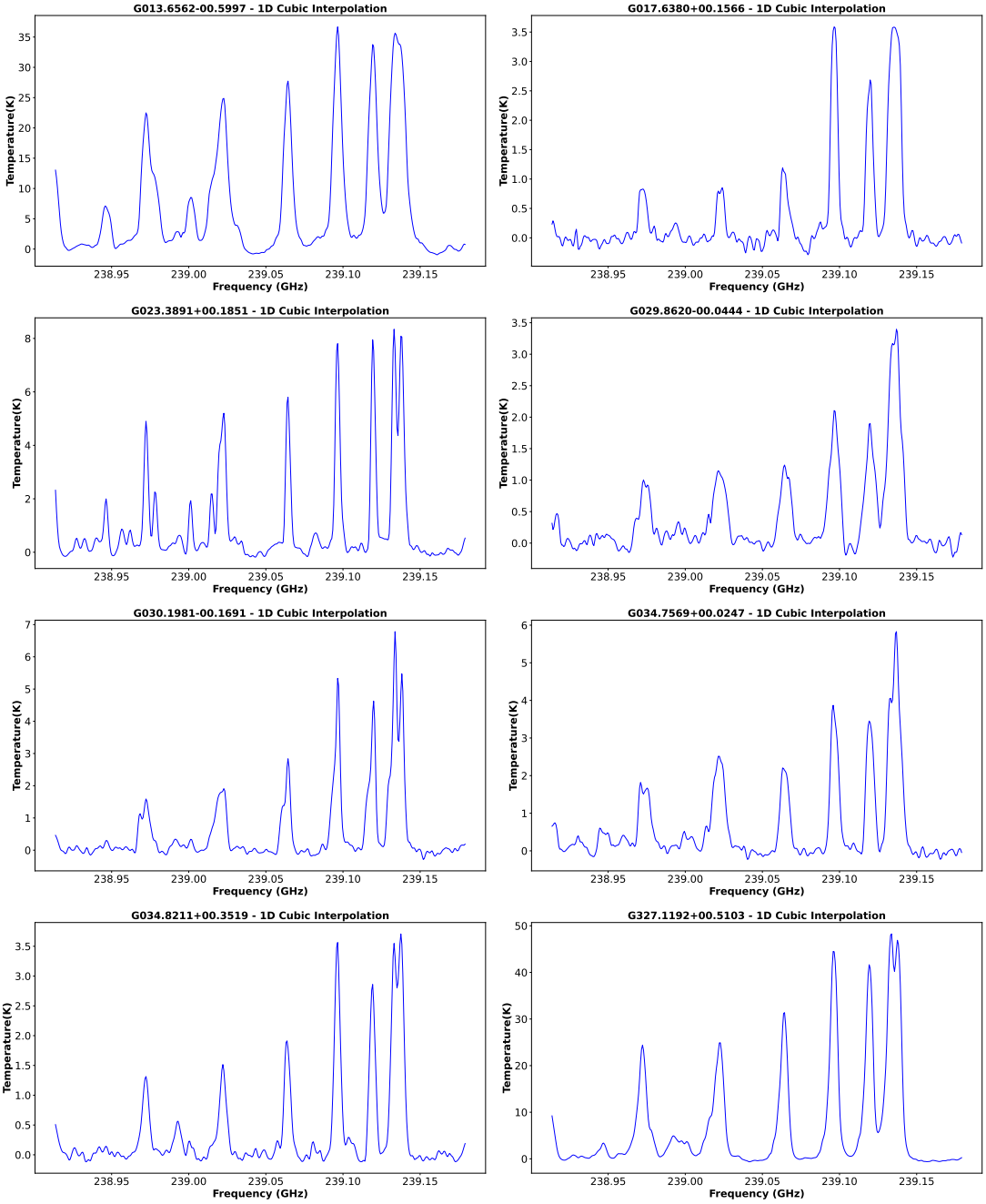


FIGURE 5.4: A representation of some of the observational data of CH_3CN sources using one-dimensional cubic interpolation. The name of the each source is shown for each panel.

We use thresholding to reduce the noise and compress the one-dimensional cubic interpolated data before doing the parameter estimation. Figure 5.5 displays the observational source plots with the approximation coefficients thresholded (see Fig. 4.1 for the denoising step). It is clear that the smoothed signal has improved slightly, albeit it is still not nearly comparable to the synthetic spectra that served as our models' training data.

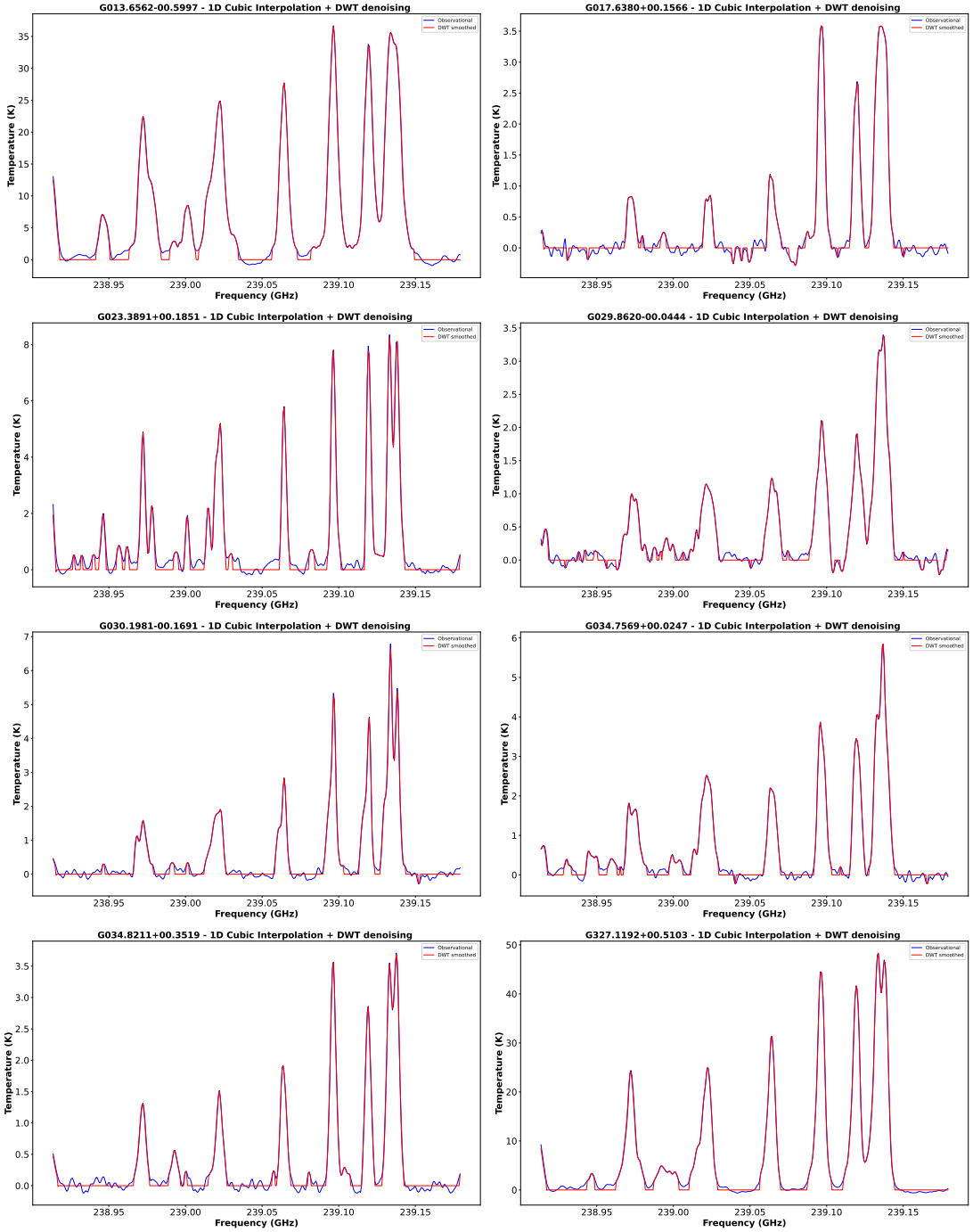


FIGURE 5.5: Plots showing the one-dimensional cubic interpolation of some of the observational data (in blue) from the 30 CH_3CN sources with the approximation coefficients subjected to thresholding (in red). The name of the each source is shown for each panel.

5.3 Model Performance

In order to evaluate the effectiveness of our model, evaluation measures were first presented in the preceding chapter. All of the models' R^2 scores are displayed in Table 5.1. R^2 provides us with a solid performance overview of our model, hence it was the only

metric utilised to validate the performance of our ML model, as opposed to all the other metrics mentioned. Overall, the tuned XGBoost model performed better than all RF and XGBoost models, achieving an R^2 score of 0.850 on the test set. While the model does not appear to be consistently overfitting or underfitting, the validation set score is slightly lower than the test set score, indicating some degree of variability in performance.

Model	Training set	Validation set	Testing set
Random Forest (RF)	0.964	0.781	0.783
XGBoost	0.956	0.834	0.839
Tuned XGBoost	**0.9898**	**0.846**	**0.850**

TABLE 5.1: The R^2 performance metric of all the 30K synthetic data of CH_3CN across all ML algorithms.

The total metrics for all ML techniques are, however, displayed in table 5.2. Overall, the performance of the tuned XGBoost was superior on all criteria. We had high values for the MSE, which may have been due to the scale of our values, which were in the hundreds. Additionally, the RF model performed poorly for all of the criteria used to evaluate our models, while the XGBoost model had higher prediction metrics than the RF model although lower than the tuned XGBoost model. Due to the compute time while utilising the grid search optimization outlined in Chapter 4, we did not perform hyper-parameter tweaking in the instance of the RF model.

Model	Performance metric			
	MAE	MSE	RMSE	R^2
Random Forest (RF)	10.85	875.89	29.60	0.78
XGBoost	6.98	430.78	20.76	0.84
Tuned XGBoost	**6.28**	**372.51**	**19.30**	**0.85**

TABLE 5.2: Different performance metric of all the 30K synthetic data of CH_3CN across all ML algorithms.

Table 5.3 shows the MAPE for each parameter for all ML models. Looking at the MAPE, the models' predictions of column density and FWHM were the most accurate. In addition, the V_{LSR} MAPE of the RF model was the best predicted physical parameter, and its regression plot together with the FWHM shows a linear relation, supporting that they are well predicted (see Appendix A for the regression plot relationship between the true values and predicted values). Physical parameters like the source size and excitation temperature had errors of between 40% and 46% for the former and between 20% and 41% for all models. However, there is significant cause for concern because the regression graphs for column density, source size, and temperature do not demonstrate a linear link, particularly for the column density when the MAPE across all models is below 2% error. Despite the fact that all of these variables are interdependent, the excitation temperature depends significantly more on the flux and beam size, so for an unresolved image, the

flux is the same for both a large and a small beam. However, understanding that the Stefan-Boltzmann's law which states that the flux is dependent on the source's distance and brightness, can also help to explain why the source's size and excitation temperature had a huge error compared to other physical parameters.

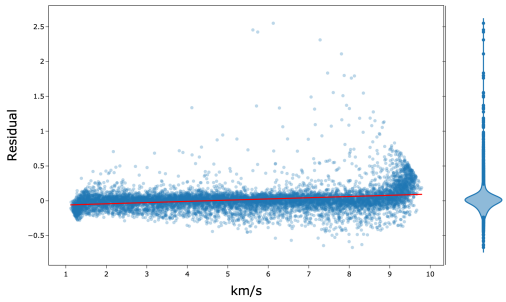
Model	MAPE				
	FWHM	column density	source size	excitation temperature	V_{LSR}
Random Forest (RF)	2.3	2.1	47.0	41.2	**1.4**
XGBoost	1.6	1.9	42.3	25.3	5.7
Tuned XGBoost	**1.2**	**1.8**	**40.6**	**20.7**	2.7

TABLE 5.3: MAPE of all the physical parameters for all the 30K synthetic data of CH_3CN across all ML algorithms.

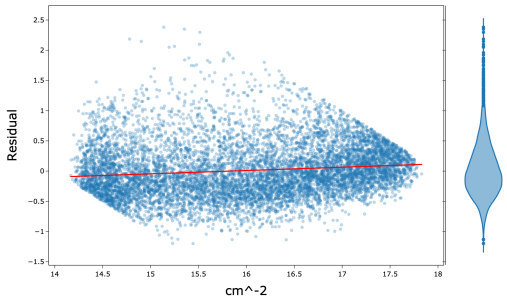
With the exception of the V_{LSR}, where the RF was superior, the tuned XGBoost model exhibited lower MAPE across all of the physical parameters. It is unclear why this is the case, however it could be that some models do better than others at forecasting particular physical parameters.

Residuals

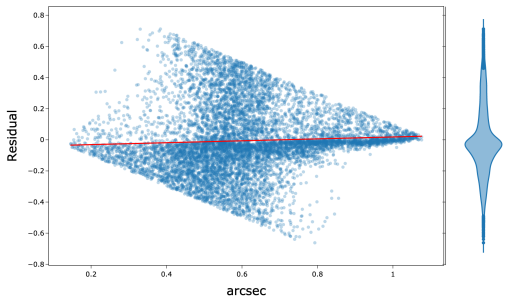
Our ability to evaluate the accuracy of our models is improved because the residuals are based on the predictions provided by our models. A residual plot of all the parameters for the RF model is shown in Figure 5.6. Based on the distributions depicted by the violin plots on each image, as well as how crowded the points are near to the zeroth line, the FWHM and V_{LSR} are the parameters that are most accurately predicted overall. Additionally, the variance is reduced for the FWHM, column density, source size, and V_{LSR} . The excitation temperature, on the other hand, has the biggest variance and is in agreement with the MAPE due to the considerable distance from the centre, making it the worst predicted parameter. According to the data patterns, the only physical parameters with constant residual variance (vertical spread) are the source size, excitation temperature, and V_{LSR}; all other physical parameters have almost constant variance.



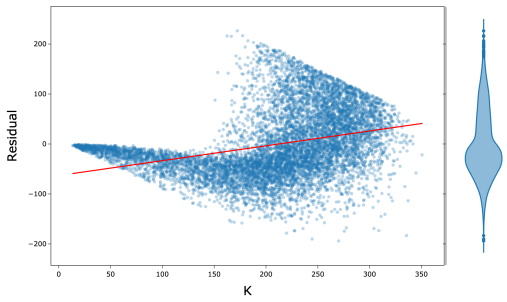
(A) FWHM



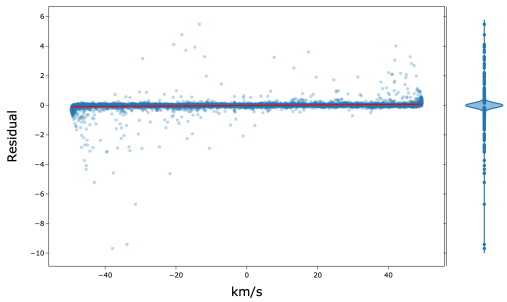
(B) Column Density



(C) Source Size



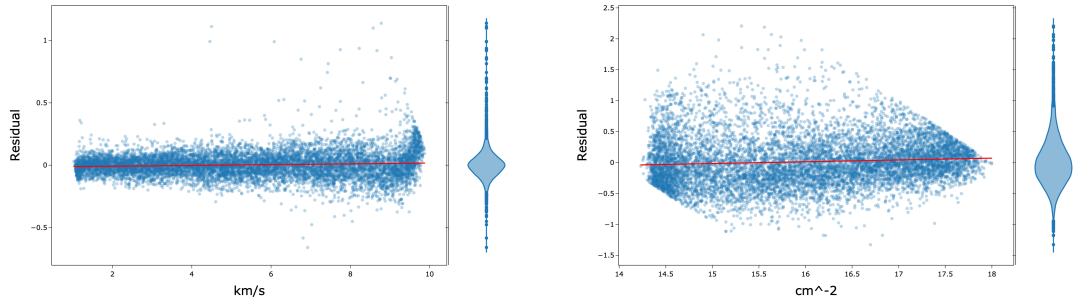
(D) Excitation Temperature



(E) V_{LSR}

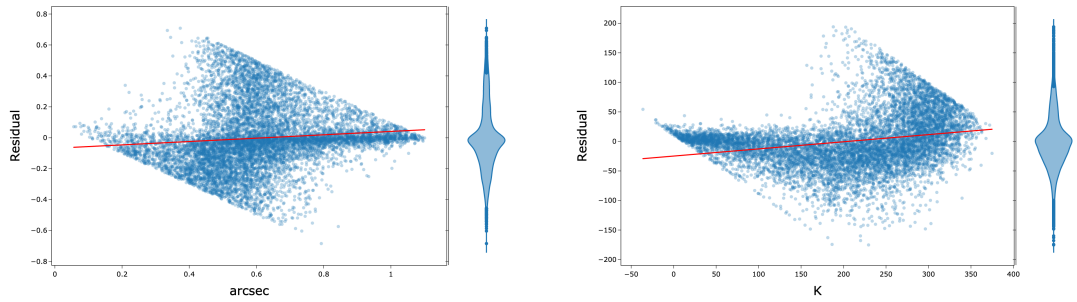
FIGURE 5.6: Each predictor is shown against the residuals individually for the RF model from the training and test datasets. When comparing the distribution of the predictors across the residuals using the violin plots on either side of each plot, it appears that the number of outliers is rising as the residuals grow and shrink away from the central point. The red line represents the data's OLS fit.

Figure 5.7 displays the residuals for the XGBoost model. The variance in the FWHM, column density, source size, and V_{LSR} are all modest, just like in the RF model. However, when the model becomes more accurate at anticipating them, they are substantially lower. The excitation temperature, on the other hand, continues to evade our model because of its higher variance. The violin plots display the residual distributions, where outliers appear as you travel out from the centre, demonstrating that certain data points are among the worst predicted across all physical parameters.



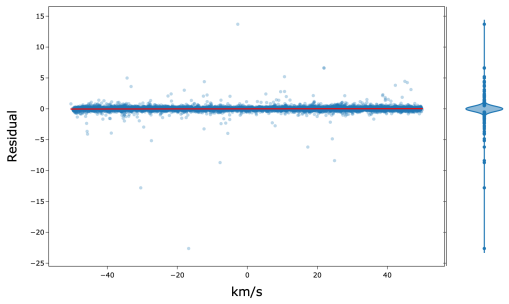
(A) FWHM

(B) Column Density



(C) Source Size

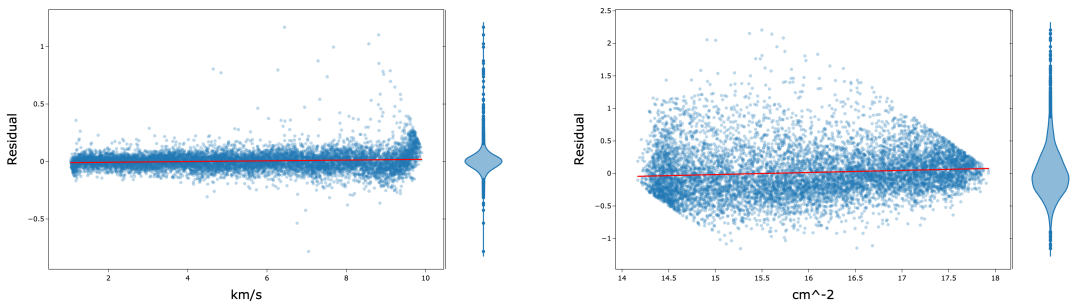
(D) Excitation Temperature



(E) V_{LSR}

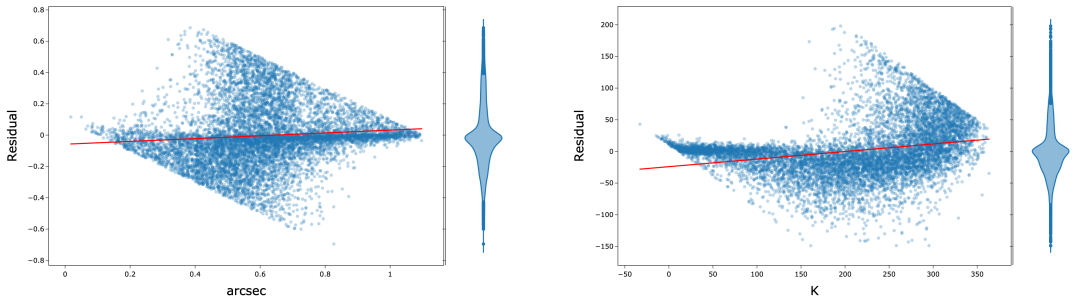
FIGURE 5.7: Each predictor is shown against the residuals individually for the XGBoost model from the training and test datasets. When comparing the distribution of the predictors across the residuals using the violin plots on either side of each plot, it appears that the number of outliers is rising as the residuals grow and shrink away from the central point. The red line represents the data's OLS fit.

In Figure 5.8, the tuned XGBoost residuals are displayed. The residuals of the physical parameters follow the same pattern as the previous two models, and the tuned XGBoost is the best model overall. Because there are so many points along the zeroth line, the FWHM and V_{LSR} are once more accurately predicted. Although the MAPE is now lower for the tweaked XGBoost model, the variance is still significant with the excitation temperature, and we now see more scatter points along the central point. Additionally, as the excitation temperature rises, the model tends to perform poorly; this is demonstrated by the previous two models as well.



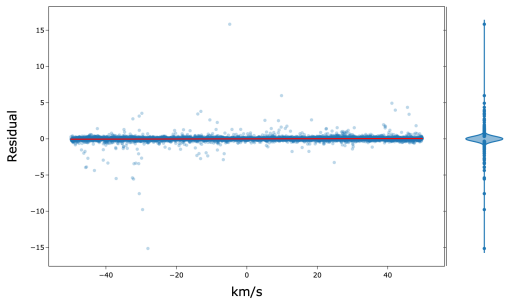
(A) FWHM

(B) Column Density



(C) Source Size

(D) Excitation Temperature



(E) V_{LSR}

FIGURE 5.8: Each predictor is shown against the residuals individually for the tuned XGBoost model from the training and test datasets. When comparing the distribution of the predictors across the residuals using the violin plots on either side of each plot, it appears that the number of outliers is rising as the residuals grow and shrink away from the central point.The red line represents the data's OLS fit.

5.4 Model Errors and Reconstruction of Synthetic Spectra

Although all of the machine learning models produced highly accurate parameter predictions, we observed a few instances where our forecasts diverged noticeably from the actual results generated by the simulations. The predictions of our physical parameters have been used to evaluate our machine learning models. Using the LTE code script and software package CASSIS database, we will attempt to rebuild the spectra by modelling them using the forecasted physical parameters and assessing how closely they match the original spectra. This could clear up any confusion regarding the possibility that various

spectra with the same spectra profiles could have different CH_3CN characteristics. Additionally, it may demonstrate that some of the physical parameters are difficult to forecast by our model since they are interdependent and so affect the spectra profiles. There is coupling between the input parameters which affects the line intensity.

A portion of the true values utilised to create some of our synthetic spectra and those used in the training set for our ML model are displayed in Table 5.4.

spectra index	FWHM (km s^{-1})	column density ($\times 10^{16}$ cm^{-2})	source size ($''$)	excitation temperature (K)	V_{LSR} (km s^{-1})
2308	8.8	57.39	0.19	59.6	30.91
22404	6.5	0.27	0.69	65.3	-6.93
23397	7.5	0.03	0.11	128.2	4.99
25058	8.3	0.01	0.31	184.9	30.99
2664	4.5	0.32	0.95	250.3	8.94
8511	5.6	0.04	0.62	265.4	49.59

TABLE 5.4: Examples of a few true value physical parameters from synthetic CH_3CN spectra produced by the LTE script that were utilised to train our ML models.

Random Forest

Table 5.5 displays the physical parameters that were predicted for random forest. The close V_{LSR} and FWHM prediction values were closer to the actual values when compared to the true parameter values in Table 5.4. Considering the MAPE of all the physical parameters, this was covered in the section before. We compare the same spectra with the original spectra profile while plotting the same spectra created from the predicted values to put this in context.

spectra index	FWHM (km s^{-1})	column density ($\times 10^{16}$ cm^{-2})	source size ($''$)	excitation temperature (K)	V_{LSR} (km s^{-1})
2308	9.1	30.38	0.22	151.9	30.91
22404	6.4	0.64	0.62	83.2	-6.96
23397	7.5	0.02	0.28	263.4	4.97
25058	8.2	0.04	0.28	268.6	30.94
2664	4.3	0.92	0.51	228.7	8.95
8511	5.6	0.06	0.55	272.8	49.50

TABLE 5.5: Examples of the CH_3CN predicted physical parameters - from the RF model.

To compare the spectra, Figure 5.9 shows the synthetic and reconstructed spectra from the predicted physical parameters using our model. Looking at the plots for the predicted physical parameters given in Table 5.5 they are not close to the original ones although the velocity positions are the same.

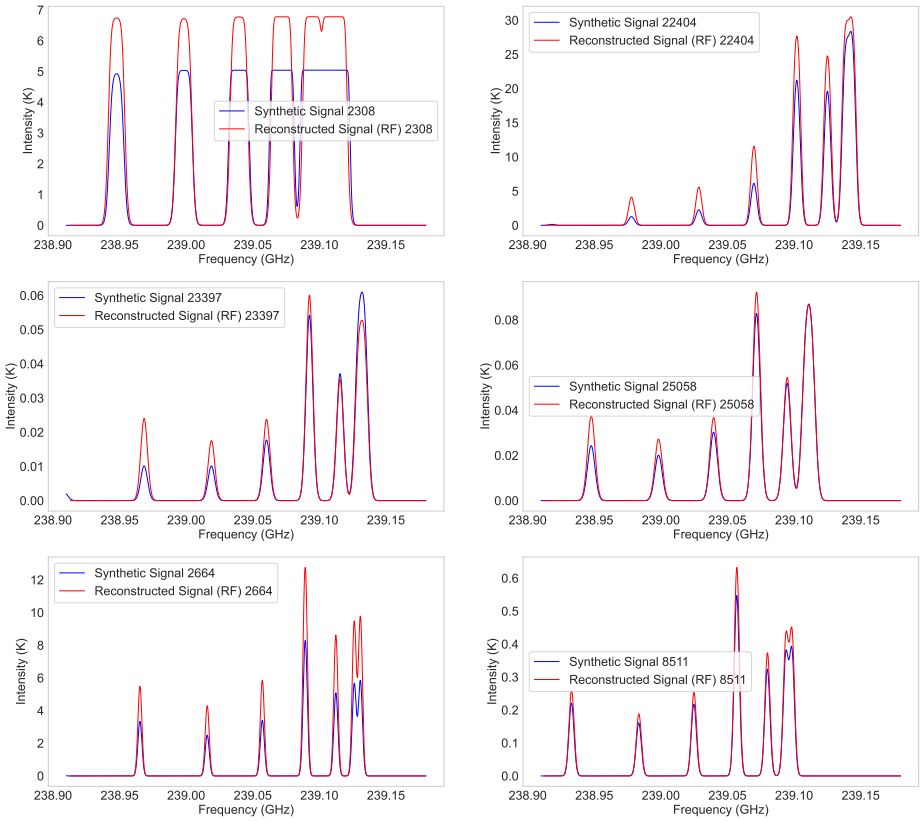


FIGURE 5.9: Spectra from the simulated model (in blue) and reconstructed spectra (in red) from the predicted physical parameters used in the RF model.

According to Figure 5.9, for each of the shown spectra, the intensity of the simulated data and that of the reconstructed data from the forecasted RF model agree. In contrast to the simulated data, the intensity of the reconstructed lines from the RF model is higher for some line emissions at lower frequencies as compared to higher frequencies. Furthermore, both the simulated data and the reconstructed data agree on the velocity positions. Overall, these reconstructions fit quite well. The fact that the intensities are close even though the parameters might seem to be relatively poorly determined compared to the input values results from the coupling between the parameters in setting the line intensity. This coupling may ultimately limit the accuracy which can be achieved in determining the physical parameters, and this contributes to the overall uncertainty in the derived parameters.

XGBoost

In comparison to the RF model, the XGBoost model had a superior assessment of the physical parameters. Table 5.6 shows some of the predicted physical parameters as compared to the initial input physical parameters used in the test set of our model.

spectra index	FWHM (km s^{-1})	column density ($\times 10^{16}$ cm^{-2})	source size ($''$)	excitation temperature (K)	V_{LSR} (km s^{-1})
2308	8.8	74.44	0.12	124.5	30.38
22404	6.5	0.51	0.62	59.3	-6.69
23397	7.6	0.02	0.20	250.8	3.94
25058	8.2	0.03	0.27	217.8	30.75
2664	4.3	0.94	0.54	268.1	9.12
8511	5.6	0.05	0.52	269.9	49.48

TABLE 5.6: Examples of the CH_3CN predicted physical parameters - from the XGBoost model.

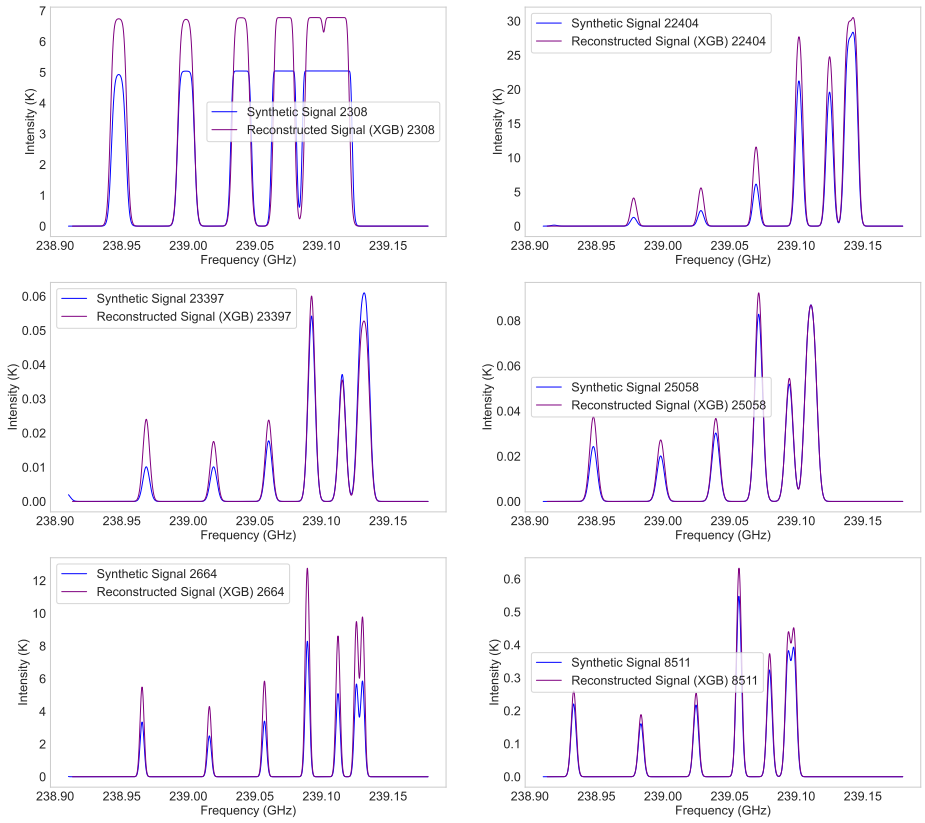


FIGURE 5.10: Spectra from the simulated model (in blue) and reconstructed spectra (in red) from the predicted physical parameters used in the XGBoost model.

Additionally, there is consistency between the original and reconstructed data's intensities. Although our XGBoost model overestimates the intensity at lower frequencies, similar to the RF model, the results are not very significant for the spectra displayed. Additionally, just like in the RF model, the line positions of the reconstructed data from our model completely match the simulated data. Overall, there are different types of spectra for the spectra displayed in Figure 5.10, including those from optically thin (*peaked tops*) and optically thick (*flattened tops*) sources. Our reconstructed data appears to fit the optically thin sources the best. Like with the RF model, overall the XGBoost reconstructions fits are reasonable.

Tuned XGBoost

Overall the tuned XGBoost had a better model than all then other model in terms of the physical parameters predictions. This is well explained in the previous section. Table 5.7 shows some of the predicted spectra physical parameters and comparing them to the physical parameters used in the test set as shown in Table 5.4.

spectra index	FWHM (km s^{-1})	column density ($\times 10^{16}$ cm^{-2})	source size (″)	excitation temperature (K)	V_{LSR} (km s^{-1})
2308	8.8	51.57	0.15	116.1	30.55
22404	6.5	0.68	0.68	58.5	-6.88
23397	7.7	0.02	0.21	208.6	4.46
25058	8.3	0.03	0.28	238.8	30.87
2664	4.3	0.96	0.53	242.0	8.98
8511	5.6	0.06	0.53	269.1	49.53

TABLE 5.7: Examples of the predicted CH_3CN physical parameters from the tuned XGBoost model.

Similar to the previous two models, Table 5.7 displays the intensity of the data we were able to recreate using the tuned XGBoost model physical parameters. All of the intensities are consistent with the simulated data, so, are the positions of the line emissions. All signals exhibit an overestimation of lower frequency strengths. For some signals, the reconstructions are, however, fairly close to the simulated data, demonstrating how superior the adjusted model is to the competing models. One noteworthy aspect is that our model undervalues the intensities for optically thick sources, which may be a result of the model's miscalculation of the excitation temperature for optically thick sources. Overall, the reconstruction fits are all good like the previous two models.

In order to compare all of our models fairly, Figure 5.12 displays a plot of the simulated data together with all of the data that was previously shown in Figures 5.9, 5.10 and 5.11. The spectra patterns of all the reconstructed data match those of the simulated data exactly, indicating that our LTE model code is deterministic. The figures show that our best model (*tuned XGBoost*) reconstruction data performs poorly in fitting the synthetic data, overestimating the intensities for the majority of the spectra while underestimating several. Even though all of the models exhibit the same pattern of line emissions, there

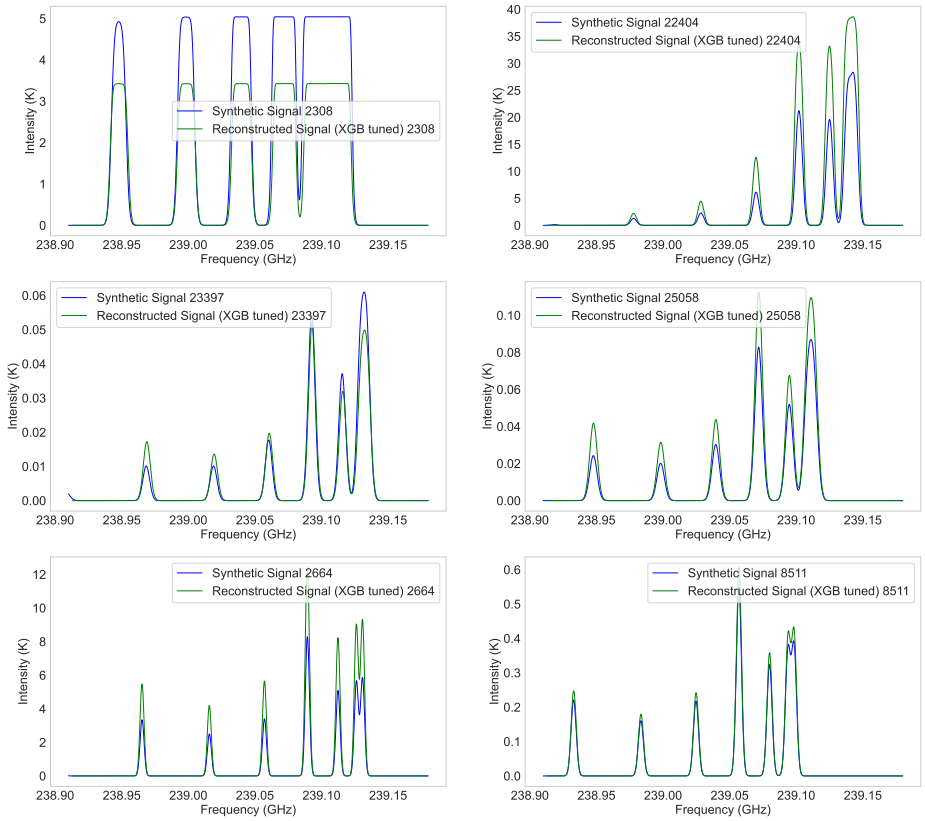


FIGURE 5.11: Spectra from the simulated model (in blue) and reconstructed spectra (in green) from the predicted physical parameters used in the tuned XGBoost model.

are a few occasions (right-middle and bottom left panel) where the reconstruction for all of the models was substandard at lower frequencies (high energy), with overestimations of the line emission intensities.

5.5 Reconstruction of Observational Data Using ML models

The ML models were used to forecast the physical parameters, such as column density, source size, excitation temperature, FWHM, and V_{LSR}, from the CH_3CN observational data sources. In order to rebuild the spectra and compare them to the observational data, all ML models were applied to the data and made predictions about the physical parameters. Due to the noise in the observation data, we used thresholding and used the physical parameters that were predicted from the thresholded values to reconstruct the spectra (explained in Chapter 3).

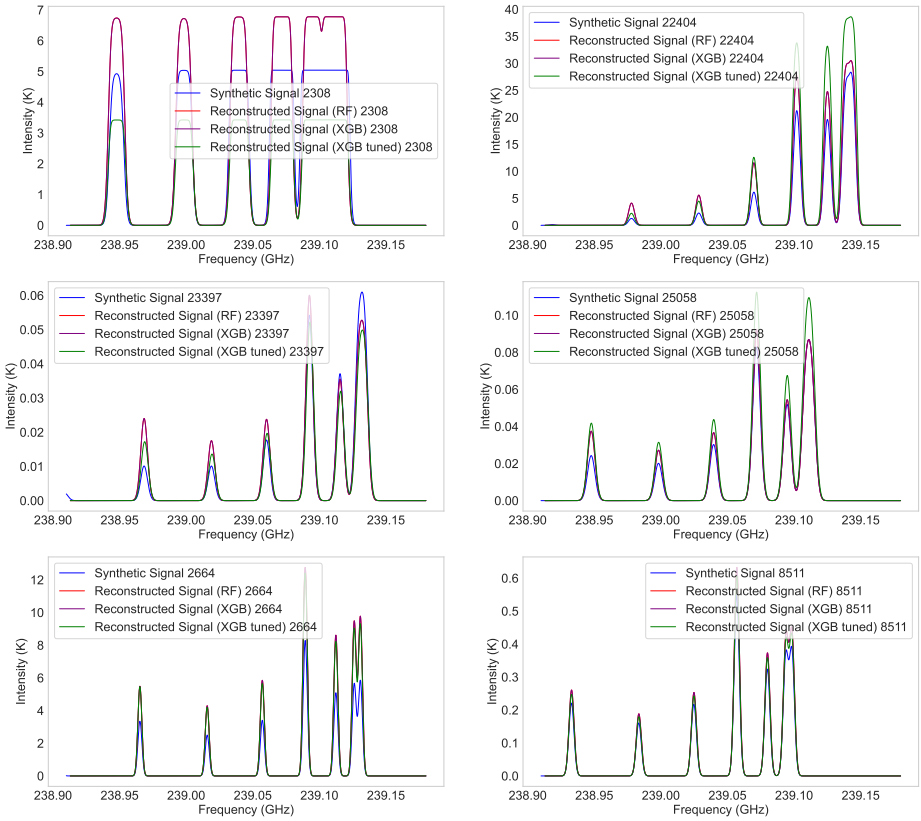


FIGURE 5.12: Spectra from the simulated model (in blue) and reconstructed spectra from the predicted physical parameters used in the RF model (in red), XGBoost model (in purple) and tuned XGBoost model (in green).

As seen in figure 5.13, the physical parameters predicted by our RF model and the reconstructed spectra line emission patterns do not even come close to matching the observational data. One explanation could be that our RF model was trained on data that does not perfectly match our observational data. Additionally, it is clear that there is an error in the line emission positions between the observational spectra of all the sources and the reconstructed spectra. All of the reconstructed spectra have an intensity which is above much of the observational spectra. The next step would have been to analyse the system using model spectra with added noise, but time was limited, so I decided to experiment with the observed data instead.

Although the XGBoost model had a higher R^2 score, the line emission profiles produced by the XGBoost approach (Figure 5.14) do not reflect the line emissions produced by the RF model (Figure 5.13). The line positions for line profiles, however, appear to be out of position, and the majority of them do not match the observational data very well. Overall, the reconstruction for the observational data is not very good, and the

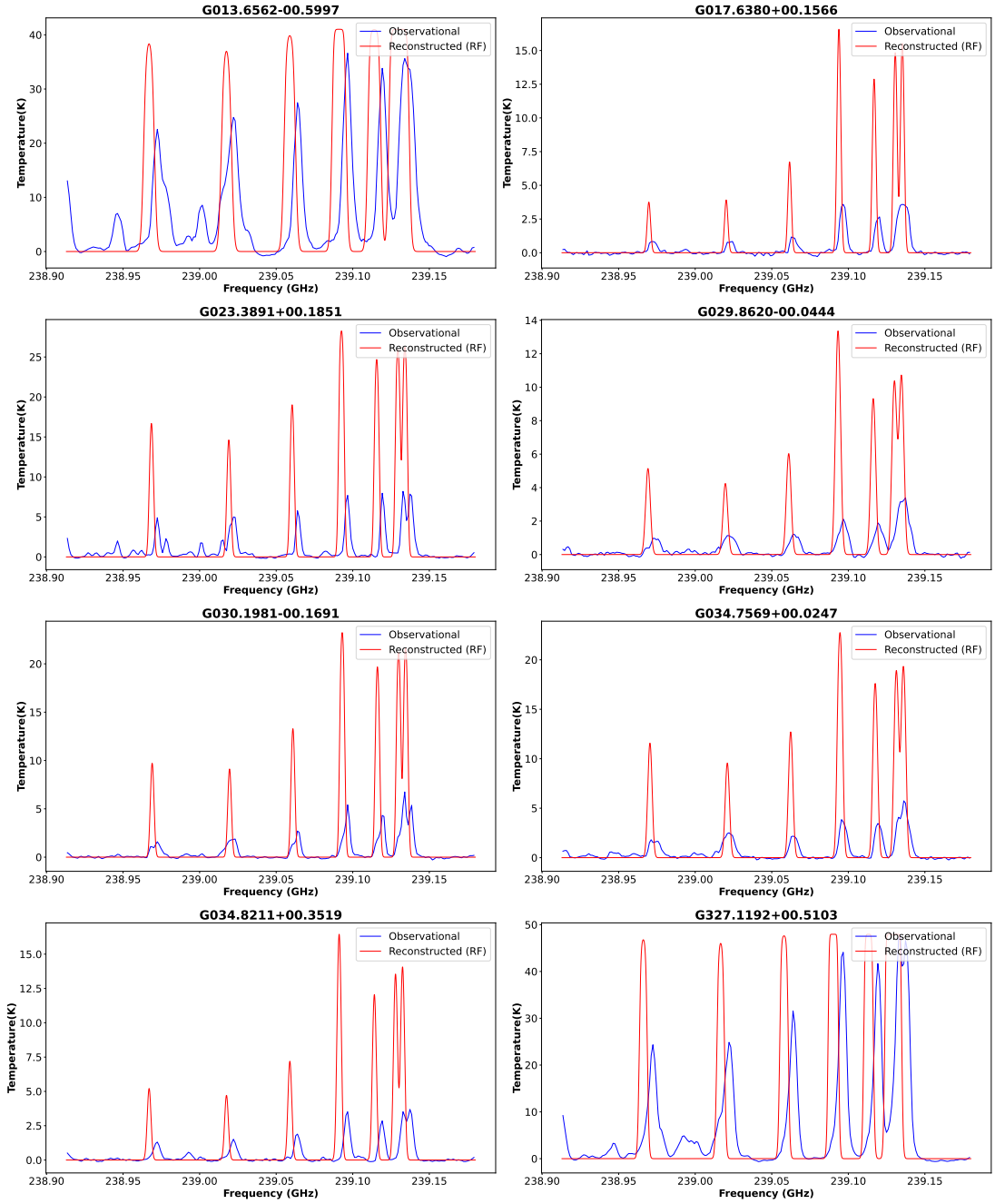


FIGURE 5.13: Spectra of some of the observational data (in blue) and the reconstructed spectra (in red) from the predicted physical parameters using the RF model.

wavelet family utilised to decompose the signals may be another factor. The Daubechies 1 wavelet, however, performed better at dissecting the signals from our training data because they closely matched the synthetic signals and had a considerably higher accuracy when trained using the ML methods used.

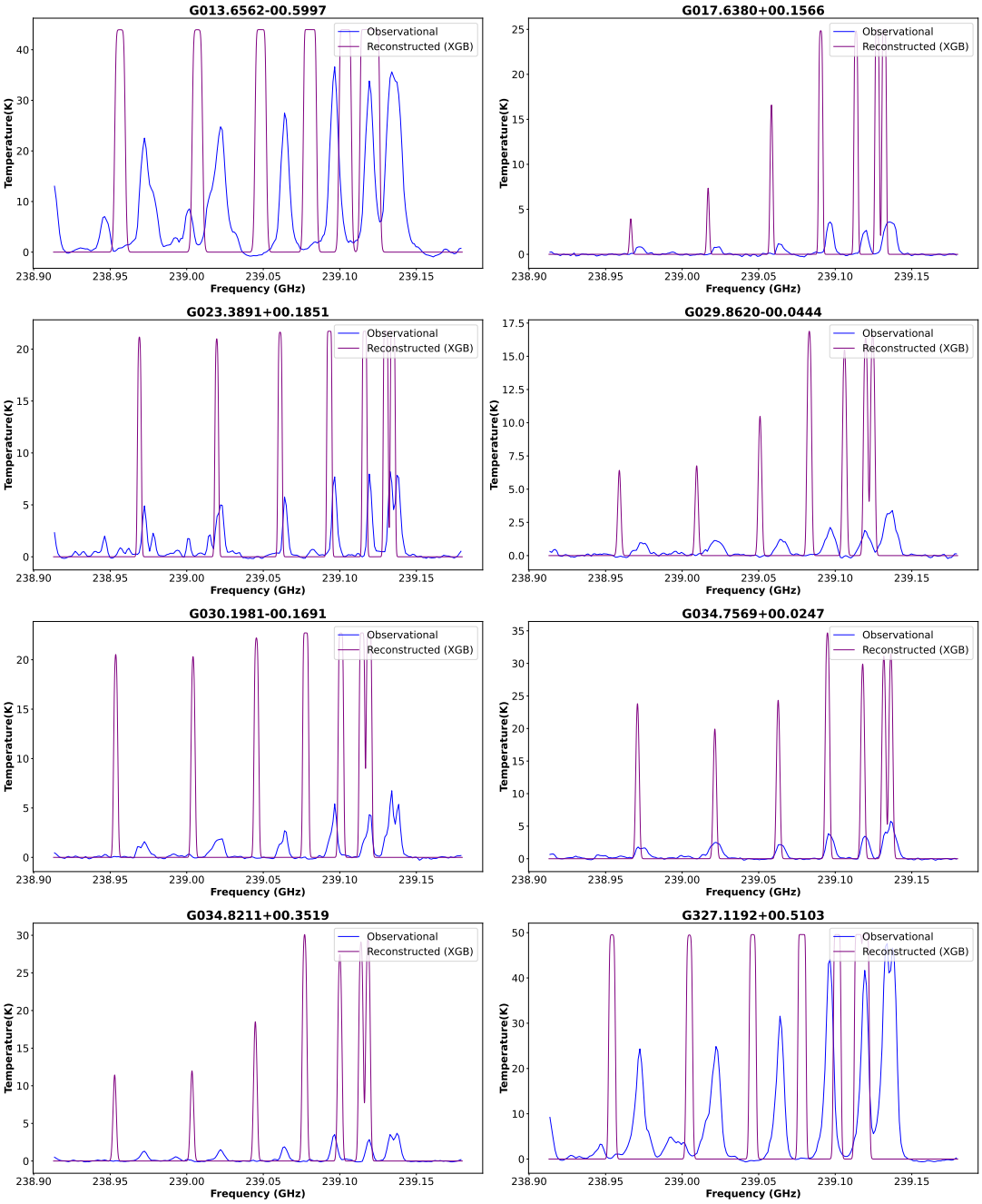


FIGURE 5.14: Spectra of some of the observational data (in blue) and the reconstructed spectra (in purple) from the predicted physical parameters using the XGBoost model.

Additionally, while having a higher R^2 score, the best model's emission line profiles only replicate the XGBoost model's line emissions, not those from the RF model. As far as the positions of the line emissions are concerned, not much has changed from the

XGBoost model. Our models can determine whether the source was optically thick or thin since all of the line emission profiles across all frequency ranges somewhat reflect the line emission profiles from the observational data.

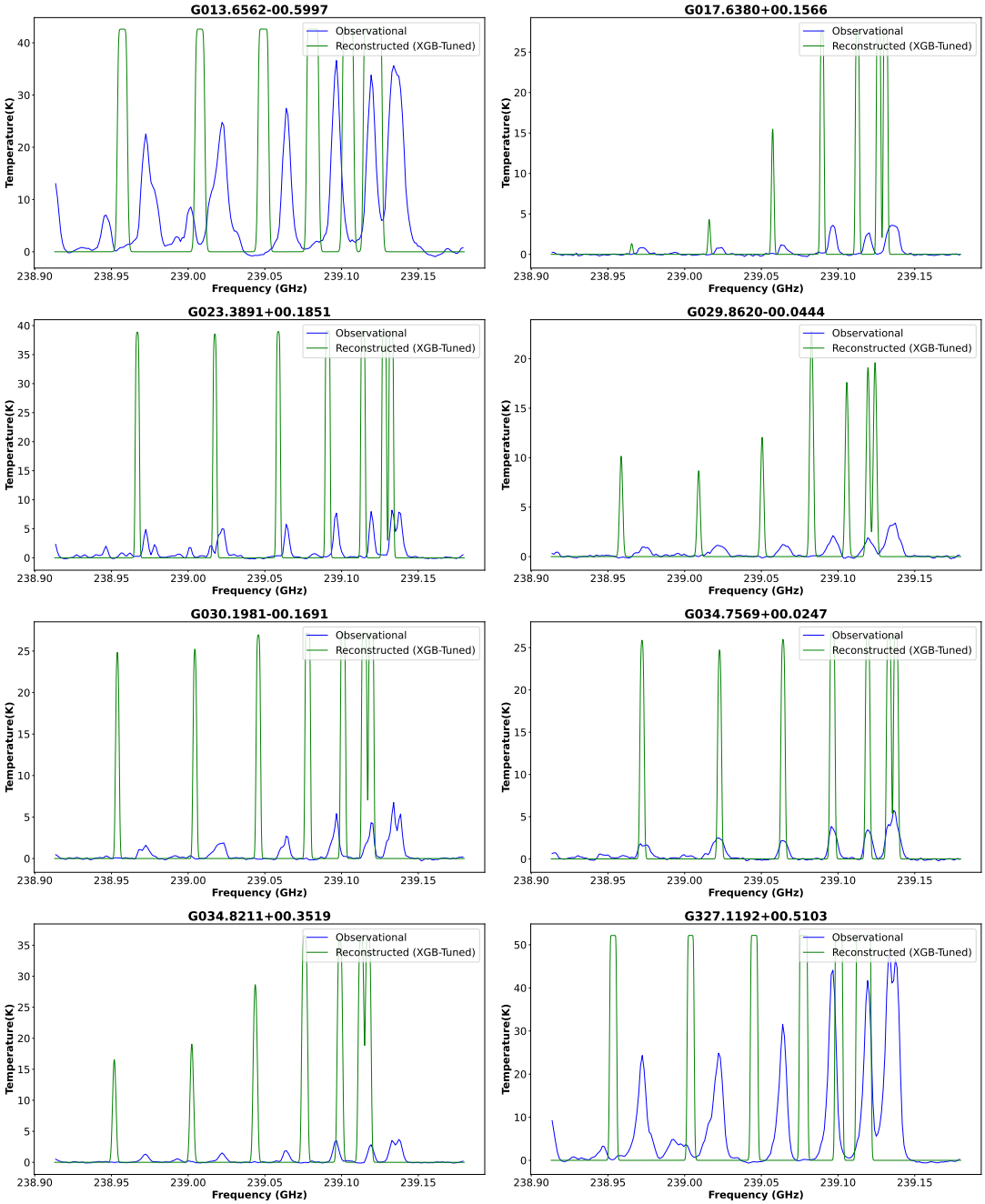


FIGURE 5.15: Spectra of some of the observational data (in blue) and the reconstructed spectra (in green) from the predicted physical parameters using the tuned XGBoost model.

Figure 5.16 displays every reconstructed ML model at once. For the XGBoost models, there are notable differences in intensity between the observational and reconstructed spectra. Overall, the RF model's reconstructed spectra were significantly better and closely matched the observational data in terms of the locations and intensities of the

emission lines. Furthermore, the temperature used in the LTE approximation determines the level populations that affect transitions' strength (Pols et al., 2018) which is also affected by the column density, which could explain why there is a discrepancy in our line intensity from our ML models and that of the observational data. This is because the temperature of the gas influences the spectral characteristics of a molecule's line emission.

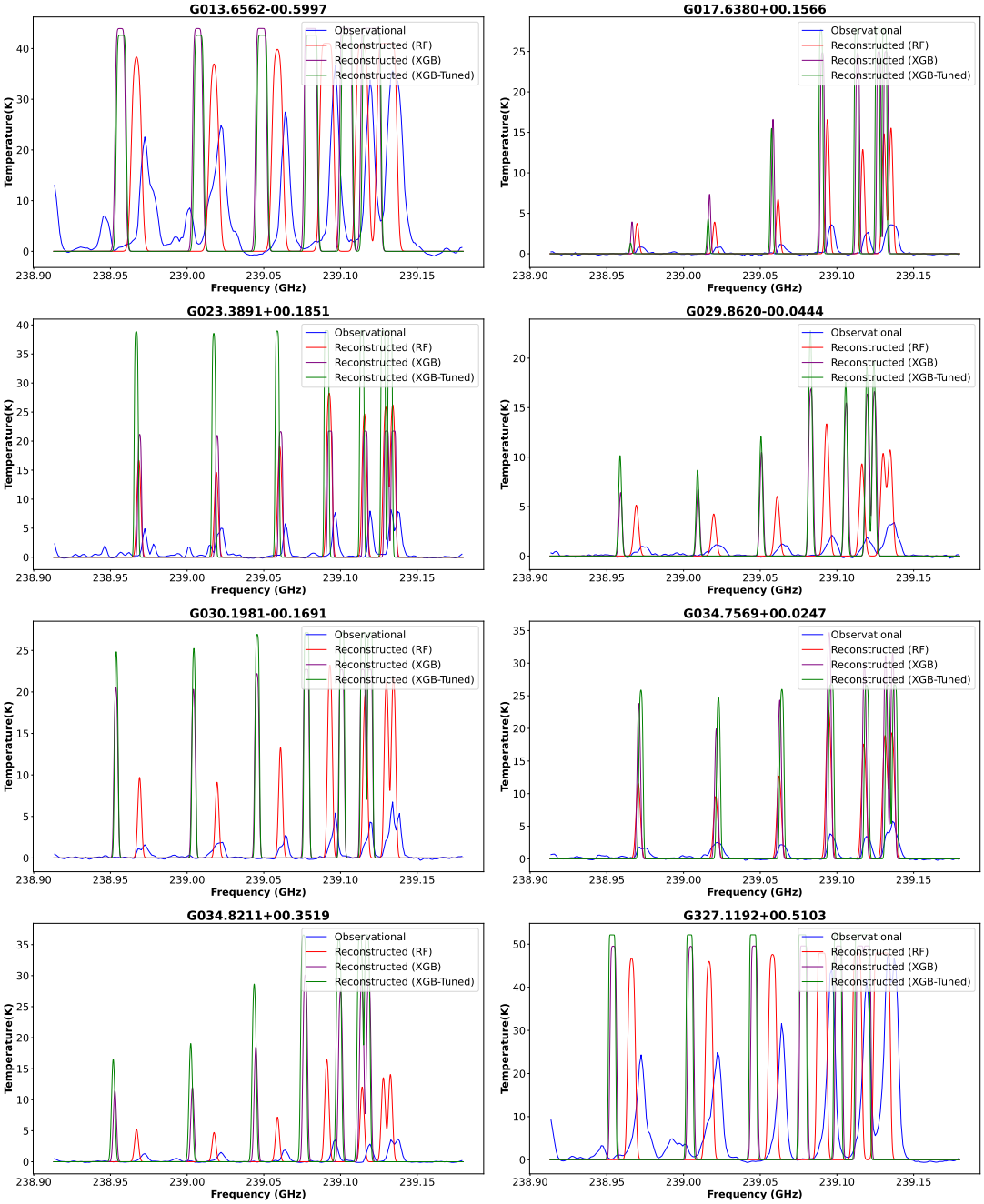


FIGURE 5.16: Spectra of some of the observational data (in blue) and the reconstructed spectra from the predicted physical parameters of all the ML models.

The physical parameters predicted by our ML models are listed in Appendix C using our LTE Python code script from the CASSIS team. Even though the reconstruction does

not perfectly match the observed data over the frequency range, the models' estimates of the physical parameters differ greatly. According to Rosero et al. (2013), if the assumption of the optically thin circumstances is erroneous, the resulting temperatures of our spectra are overstated and the column densities are underestimated, which has an impact on the line emission profiles generated. This could be one of the reasons why the line emission profile reconstruction of our observational data using projected physical parameters, notably for the XGBoost and tuned XGBoost models, only reproduces optically thick line profiles.

Furthermore, a further factor contributing to the poor performance of our ML models when applied to the observational data could be the mismatch in the intensity range of the training set and that from the observational data. This suggests that the training data may not have fully captured the range of variability present in the observational data or unusual cases not present in the training set. As a proportion of the intensity values, Figure 5.17 displays the distribution of the synthetic data intensity utilised in the training set for our ML models. The figure shows a close distribution in the intensity for both the training and observational data with the maximum intensity for the observational being 65.6 K and 62.2 K for the training set.

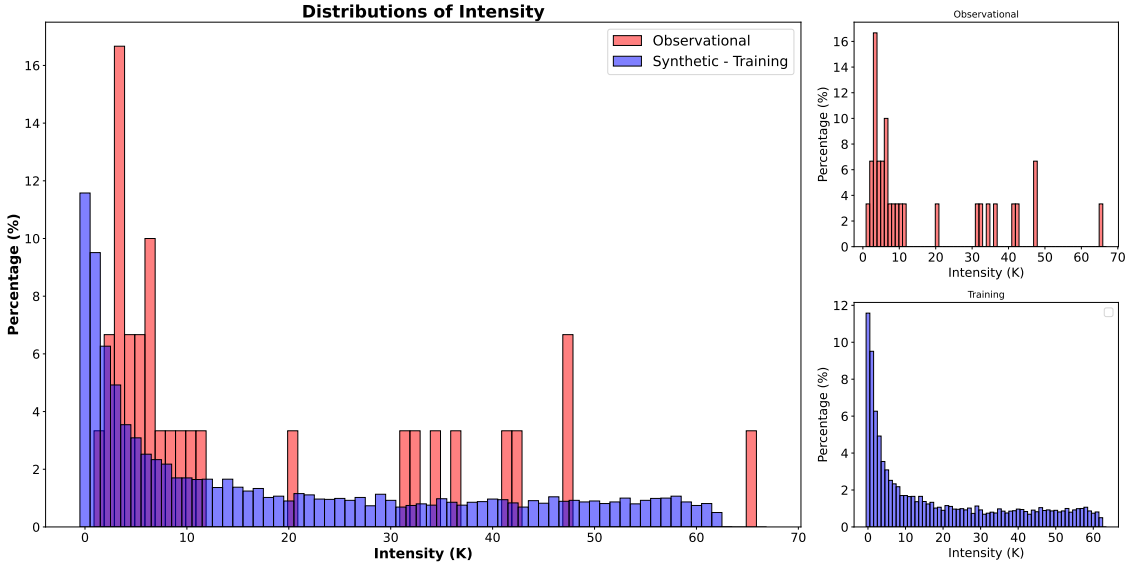


FIGURE 5.17: Distribution of the intensity range from the training set derived from our synthetic data (in blue) and the intensity range from the observational data (in red).

Nevertheless, figure 5.17 demonstrates that the intensity for the observational data in the 0.01–1 K range is roughly 58%, compared to 72% in the training set. Although the observational data shape only comprises 7080 points across all sources, compared to the 151 730 743 points used in the training set, the conclusions should still hold true if we interpolate the observational data. Thus, it is evident that even if the data was smoothed to match the data used in the training set, which was produced by our LTE code script, our ML models do poorly when applied to noisy observational data.

5.6 Summary

In order to attempt to deduce the physical parameters from CH_3CN line emissions of the observational data, we developed ML models in this work about using machine learning to star formation. The results and discoveries were reported in this chapter. The ML models that were developed do not work effectively when tested on synthetic data. When these models were applied to observational data, they faced difficulties in reconstructing line profiles, despite generating physical parameter predictions that were relatively similar to the original ones.

Chapter 6

Conclusion

To summarise the results of this research, we used the LTE code script to generate spectral line data for CH_3CN, a complex organic chemical that is present in star-forming areas, and using the CASSIS database as a reference. In order to supply our machine learning models with smaller data dimensions, the synthetic data had to be processed. The discrete wavelet transform (DWT) was used to break down the synthetic signals. The Daubechies 1 wavelet family typically performs well in decomposing the signal at the decomposition level 6 while still holding a large component of the original signal information compared to other wavelet family types.

Furthermore, wavelet coefficients (approximation coefficients) were used as features to predict the physical parameters of each signal. Among the other ML models in the ML implementation, the tuned xgboost model had the highest R^2 score of 0.850. The R^2 values for the RF and xgboost models were 0.783 and 0.839, respectively. Furthermore, because running the Grid-Search Optimization (GSO) algorithm on a larger dataset is computationally expensive, the R^2 score could have been higher if all of the data had been used rather than a small sample. A summary of the evaluation metrics is shown in Table 6.1.

Evaluation Metric	Tuned Xgboost Model	Xgboost Model	Random Forest Model
R^2	**0.85**	0.84	0.78
MAE	**6.28**	6.98	10.85
MSE	**327.51**	430.78	875.89
RMSE	**19.30**	20.76	29.60

TABLE 6.1: A summary of all the model evaluation metrics. Overall, the tuned xgboost model outperforms the xgboost and random forest models.

Although the tuned xgboost model outperformed the other models, the MAPE of the RF model had a better V_{LSR} across all models, while the MAPE of the remaining physical parameters, FWHM, column density, excitation temperature, and source size, were all lower with the tuned xgboost.

To confirm our results of the predicted parameters from our ML models, we used the LTE code script to generate new spectra using the physical parameters from our ML

models. Because our LTE code script is deterministic rather than stochastic, the generated spectra from the predicted physical parameters were all in agreement with the synthetic spectra we trained our models on. In terms of velocity positions, all of the predicted machine learning models' line emission profiles matched the synthetic data, though there were more spectra where the intensities of the emission lines increased. This could be because the MAPE of the physical parameters was not low sufficient because the parameters are mutually dependent, or it could be that variety of physical parameters yield similar line profiles.

The reconstruction of the predicted parameters from our ML models works quite well (see figure 5.12 in chapter 5), however, we need to apply the models to the observational data to assess their performance. On the other, the synthetic data used in the training of our ML models do not reflect real-world data which often has noise. To combat this, we used a thresholding technique to get the data line emissions which are close to those used in the training set. Consequently, the RF model which had a lower R^2 score among all the ML models performed quite well in the prediction of the physical parameter which was better when reconstructed and matched the observational data in terms of the line positions although not satisfactory. While the performance of the other two models was high when using the synthetic data, the reconstruction of the predicted physical parameters was not close to the observational data's line intensity. Furthermore, the lack of similarity between the data used to train the ML models and the observational data may be the justification that all of the models performed poorly on the observational data.

The main motivation was to use ML techniques to analyse star formation and early evolution using spectral data from the ALMA telescope as well as data compression methods. Figure 4.1 depicts the layout of the ML project, while Figure 3.1 depicts the data compression technique. Overall, the ML methods used were properly assessed in terms of prediction accuracy and physical parameter errors using the regression evaluation metrics outlined in section 4.1.5. The machine learning models perform well on synthetic data but poorly on observational data. All of the code is available on GitHub, which can be accessed here: `https://github.com/jpandeinge/wavelet_decomposition`.

The weakness of this work could be attributed to the type of wavelet family used, as well as the LTE code script used to generate data that closely mirrors observational data of CH_3CN spectra obtained with the ALMA telescope. Furthermore, the computational constraints in the case where we needed to optimise the algorithms employed in this research using the GSO algorithm. When compared to the xgboost, the random forest regressor performed better on observational data, and tuning the model may improve its accuracy.

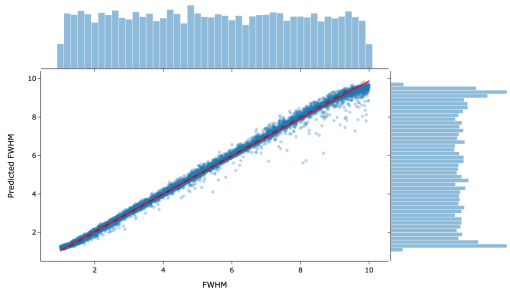
This work satisfied the research's motivation. For future work, increasing the dataset size may not significantly improve the performance of the ML models. Instead, it may be more effective to increase the sampling of the FWHM and velocity to achieve better accuracy. However, the implementation of ML models is still a much faster method for generating physical parameters and spectra than conventional fitting and error estimation, which can take up to an hour per spectrum. It takes about an hour and a half to

generate 40,000 synthetic data points on a 16-GB machine, and the entire cycle, including model validation, can take up to a day for preprocessing and implementing the ML models. When applied to observational data, the ML models can generate the physical parameters and spectra in less than a minute. Overall, future work should consider increasing the sampling of the FWHM and velocity to improve model accuracy while still leveraging the speed advantages of the ML approach.

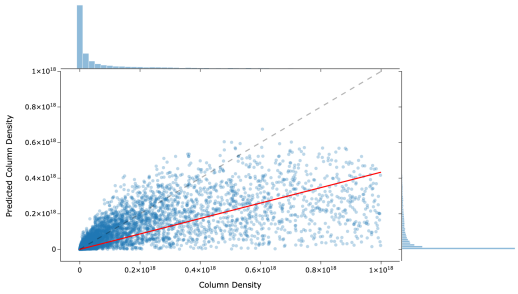
Utilising different wavelet family types, as well as various data compression techniques, could be beneficial. Finally, building a neural network (NN) may be worthwhile. Although it should be noted that NN has weaknesses when it comes to using tabular data when particularly in comparison to the tree-based method used in this research.

Appendix A

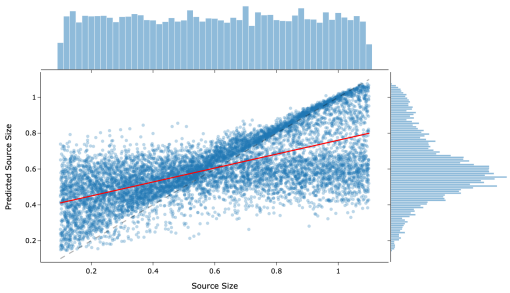
Appendix A



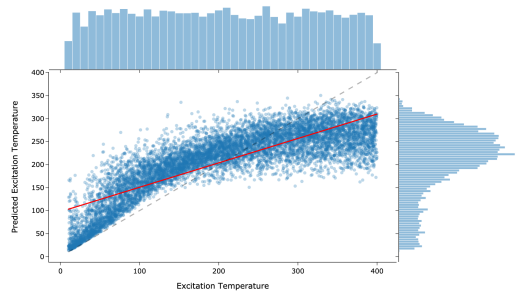
(A) FWHM



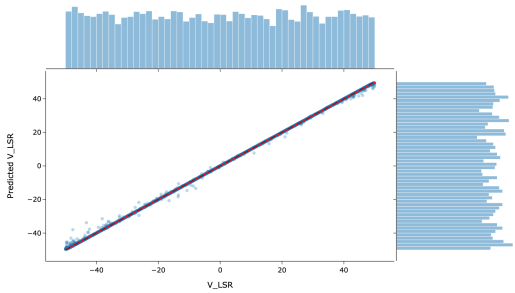
(B) Column Density



(C) Source Size

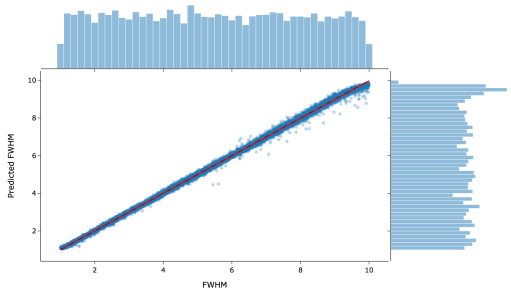


(D) Excitation Temperature

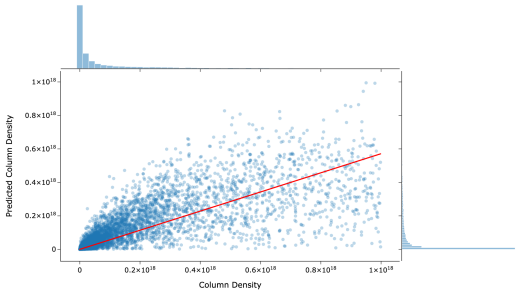


(E) V_{LSR}

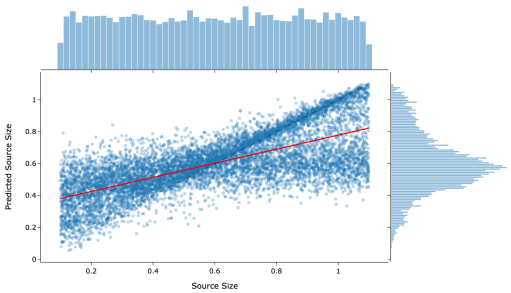
FIGURE A.1: Regression graphs that contrast the parameters predicted by the random forest model with their actual values. The grey line (dashed line), when contrasted to the red line (the ordinary linear square (OLS) fit), shows how well our model fits the data. Most of the scatter dots in a strong model will be located close to the diagonal dashed line. The histogram contrasts the true value distribution with the projected value distribution.



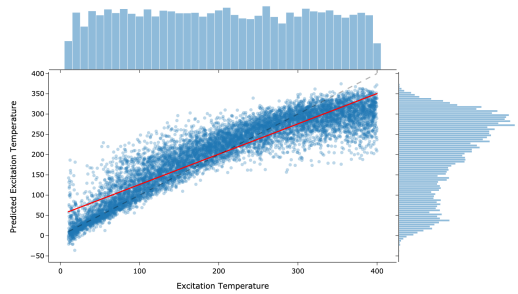
(A) FWHM



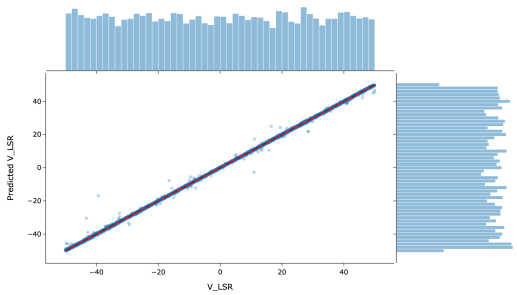
(B) Column Density



(C) Source Size

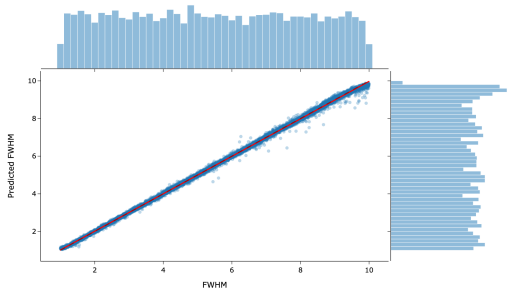


(D) Excitation Temperature

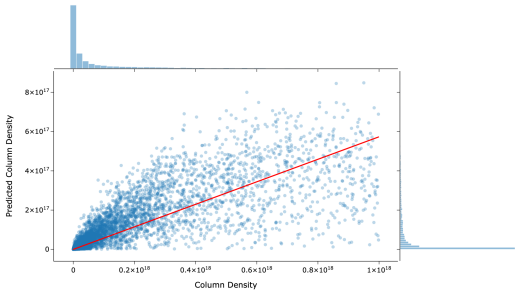


(E) V_{LSR}

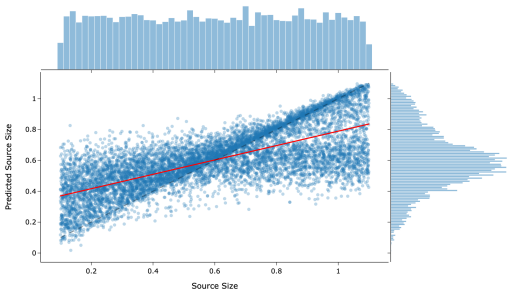
FIGURE A.2: Regression graphs that contrast the parameters predicted by the xgboost model with their actual values. The grey line (dashed line), when contrasted to the red line (the ordinary linear square (OLS) fit), shows how well our model fits the data. Most of the scatter dots in a strong model will be located close to the diagonal dashed line. The histogram contrasts the true value distribution with the projected value distribution.



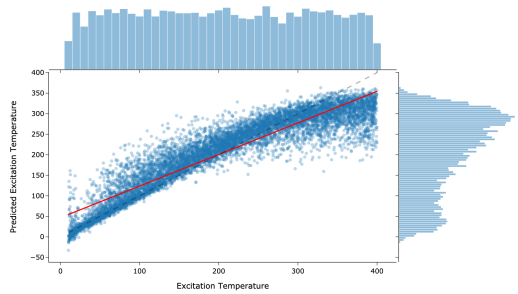
(A) FWHM



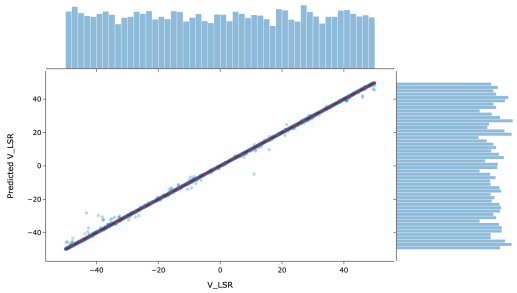
(B) Column Density



(C) Source Size



(D) Excitation Temperature



(E) V_{LSR}

FIGURE A.3: Regression graphs that contrast the parameters predicted by the tuned xgboost model with their actual values. The grey line (dashed line), when contrasted to the red line (the ordinary linear square (OLS) fit), shows how well our model fits the data. Most of the scatter dots in a strong model will be located close to the diagonal dashed line. The histogram contrasts the true value distribution with the projected value distribution.

Appendix B

Appendix B

Spectra of all of the observational data (in blue) and the reconstructed spectra from the predicted physical parameters of all the ML models.

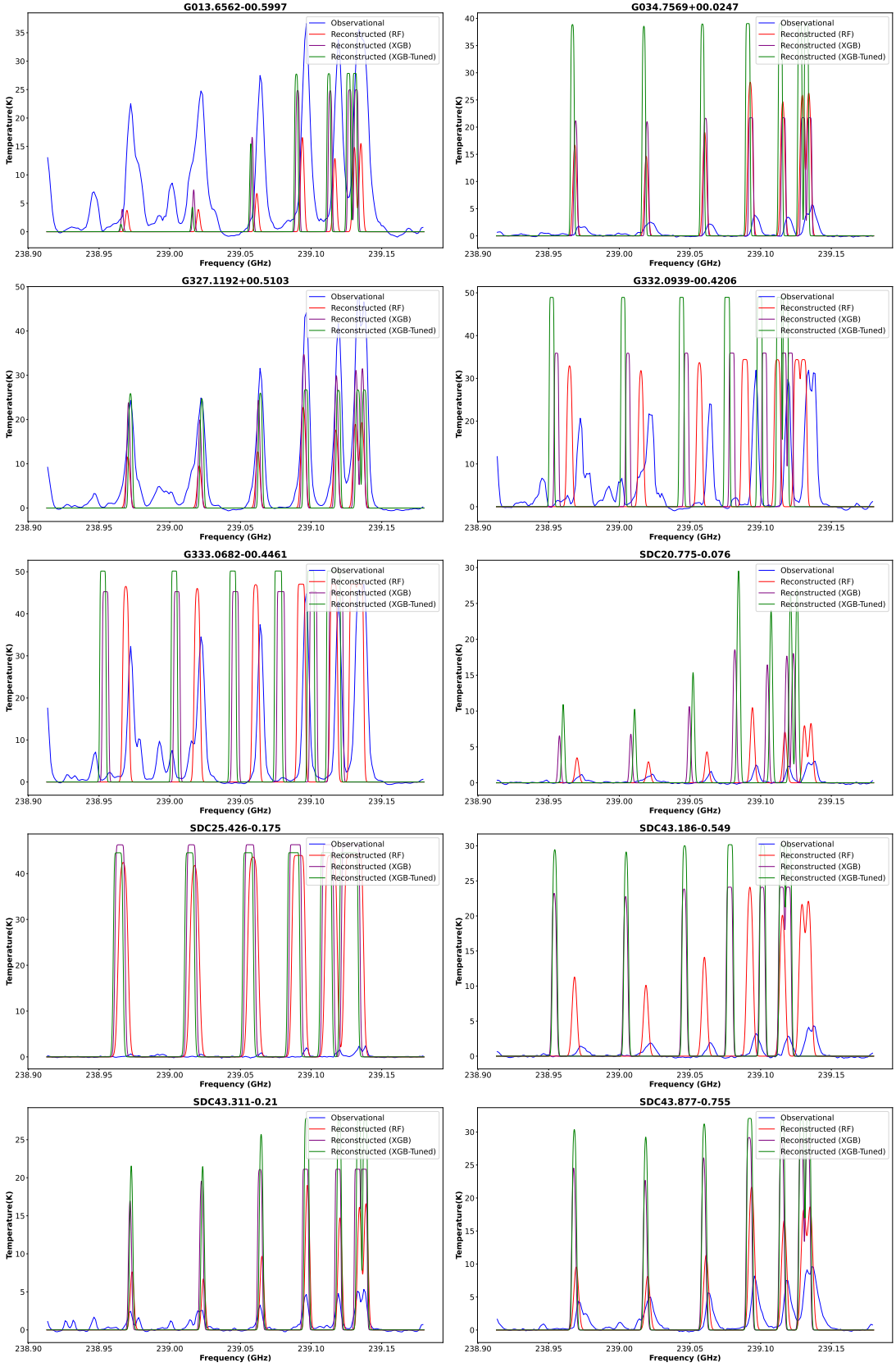


FIGURE B.1: Spectra of some of the observational data (in blue) and the reconstructed spectra from the predicted physical parameters of all the ML models.

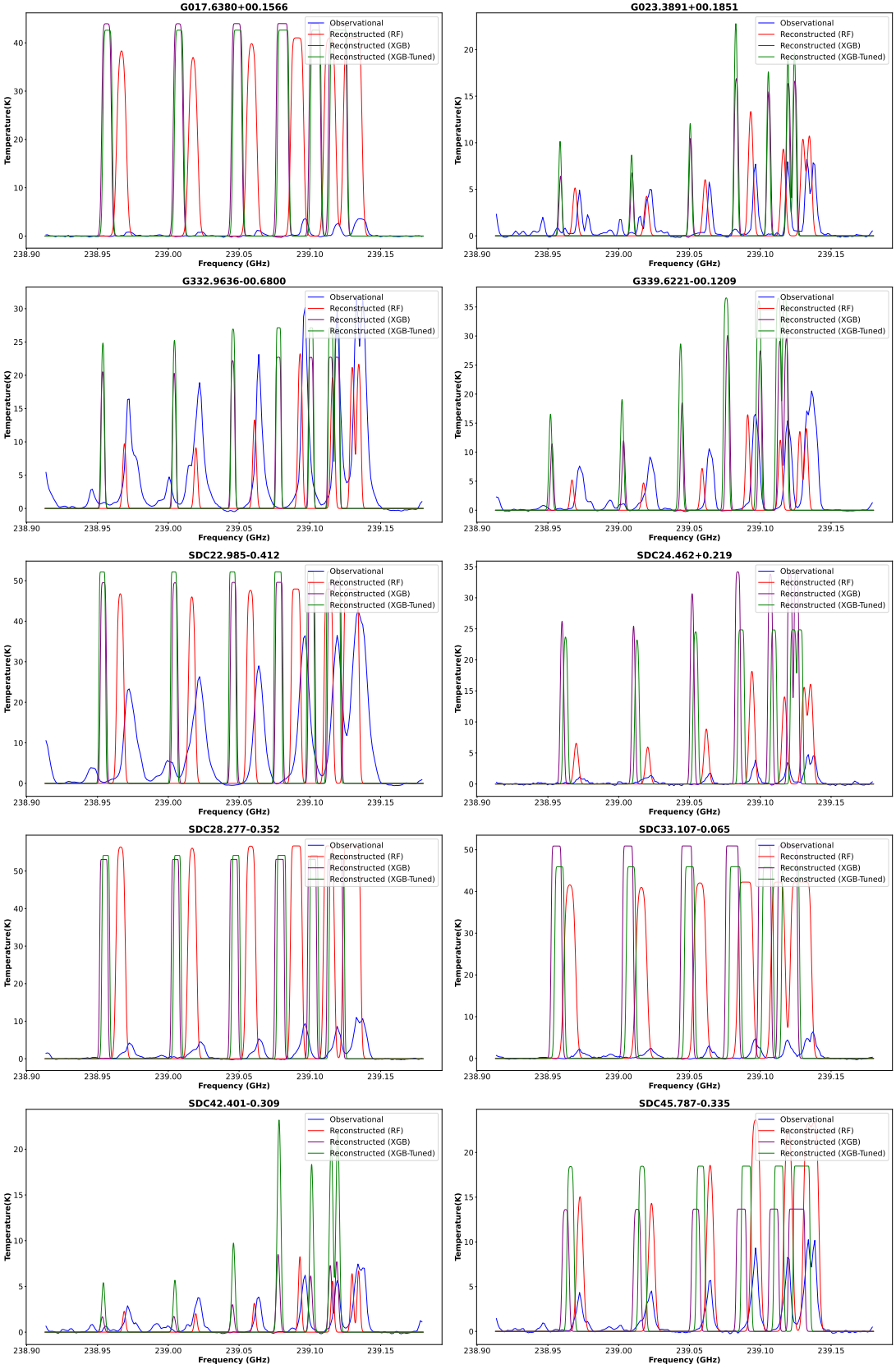


FIGURE B.2: Spectra of some of the observational data (in blue) and the reconstructed spectra from the predicted physical parameters of all the ML models

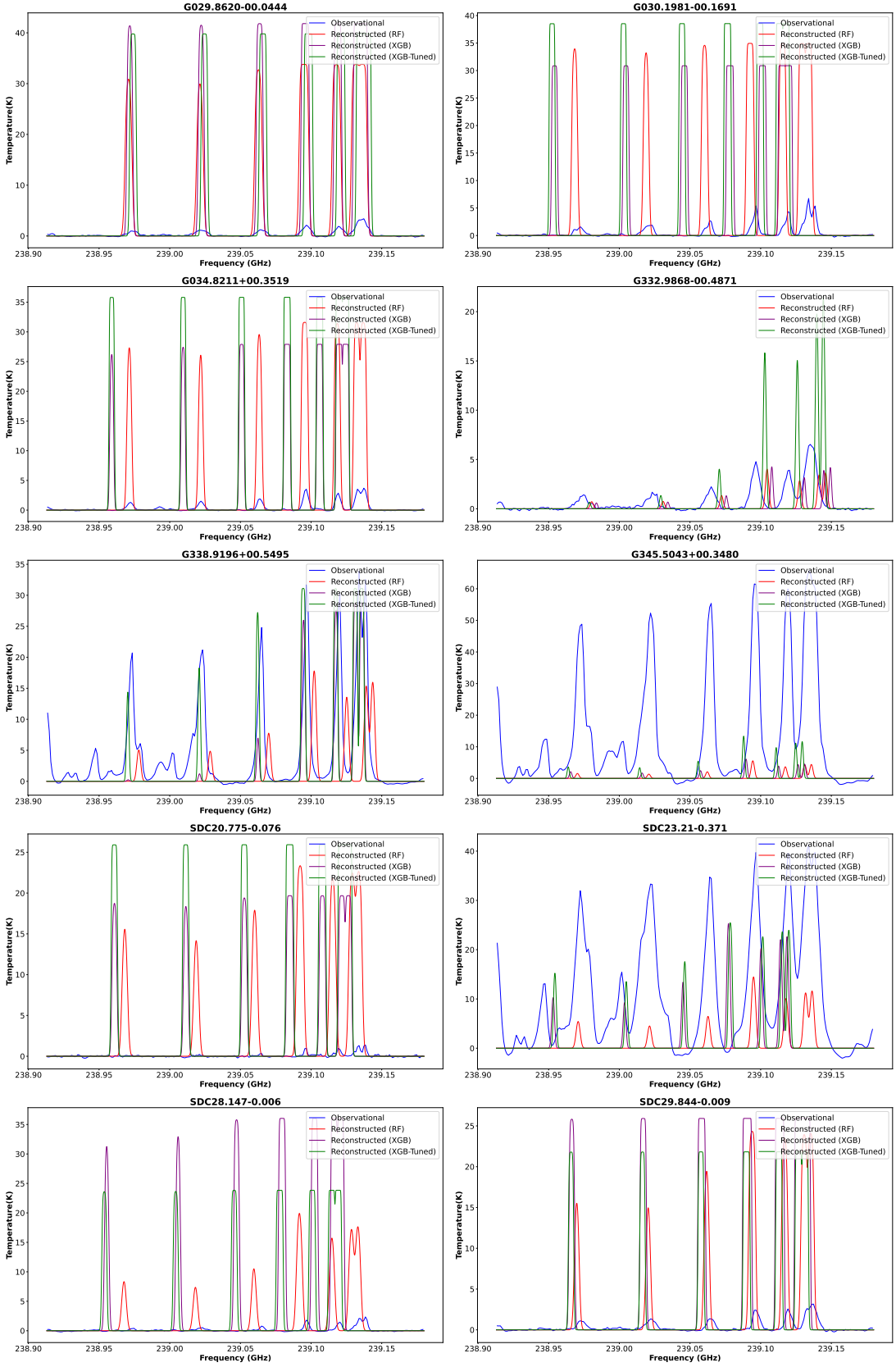


FIGURE B.3: Spectra of some of the observational data (in blue) and the reconstructed spectra from the predicted physical parameters of all the ML models.

Appendix C

Appendix C

source	FWHM (km s^{-1})	column density ($\times 10^{16}$ cm^{-2})	source size ($''$)	excitation temperature (K)	V_{LSR} (km s^{-1})
G013.6562-00.5997	6.1	12.00	0.69	175.2	6.81
G017.6380+00.1566	3.2	0.27	0.53	119.4	3.34
G023.3891+00.1851	3.4	2.07	0.54	187.1	4.85
G029.8620-00.0444	4.4	0.75	0.50	195.0	4.04
G030.1981-00.1691	3.3	0.90	0.50	146.9	4.24
G034.7569+00.0247	3.9	1.71	0.53	211.4	2.44
G034.8211+00.3519	3.5	0.43	0.54	156.2	6.75
G327.1192+00.5103	4.2	10.90	0.80	169.2	7.89
G332.0939-00.4206	3.4	8.36	0.60	189.9	9.57
G332.9636-00.6800	4.5	7.27	0.60	159.7	1.92
G332.9868-00.4871	4.4	0.82	0.49	156.7	3.02
G333.0682-00.4461	3.6	11.40	0.79	172.4	4.36
G338.9196+00.5495	3.5	9.26	0.61	178.0	5.00
G339.6221-00.1209	3.4	4.59	0.56	164.6	1.26
G345.5043+00.3480	4.9	19.50	0.97	178.2	7.58
SDC20.775-0.076_1	3.3	0.05	0.52	120.5	-10.10
SDC20.775-0.076_3	3.8	0.34	0.55	182.8	2.91
SDC22.985-0.412_1	5.9	1.28	0.74	156.8	6.71
SDC23.21-0.371_1	16.7	2.00	0.71	175.3	8.70
SDC24.462+0.219_2	3.8	0.43	0.54	137.5	-7.17
SDC25.426-0.175_6	2.9	0.14	0.58	160.2	4.42
SDC28.147-0.006_1	3.3	0.11	0.57	167.4	2.57
SDC28.277-0.352_1	4.4	3.28	0.46	173.2	5.28
SDC29.844-0.009_4	4.1	0.64	0.54	190.7	1.92
SDC33.107-0.065_2	4.4	1.19	0.49	168.9	5.96
SDC42.401-0.309_2	4.6	1.61	0.51	166.2	5.13
SDC43.186-0.549_2	3.6	0.86	0.49	169.0	-1.17
SDC43.311-0.21_1	14.6	1.39	0.53	186.7	3.63
SDC43.877-0.755_1	5.4	3.27	0.46	148.3	-0.33
SDC45.787-0.335_1	3.9	2.40	0.47	142.4	3.13

TABLE C.1: Physical parameters of CH_3CN observational spectra predicted using using the RF model.

source	FWHM (km s^{-1})	column density ($\times 10^{16}$ cm^{-2})	source size (″)	excitation temperature (K)	V_{LSR} (km s^{-1})
G013.6562-00.5997	4.5	2.72	0.74	143.0	20.04
G017.6380+00.1566	2.1	0.63	0.47	59.4	7.48
G023.3891+00.1851	2.1	4.78	0.43	147.8	4.17
G029.8620-00.0444	2.8	0.78	0.38	115.6	16.86
G030.1981-00.1691	2.5	3.64	0.44	137.5	23.57
G034.7569+00.0247	2.7	2.98	0.64	241.9	1.98
G034.8211+00.3519	2.8	0.79	0.57	117.1	24.44
G327.1192+00.5103	3.0	17.70	0.83	182.3	22.85
G332.0939-00.4206	2.1	18.90	0.62	197.2	21.17
G332.9636-00.6800	3.2	9.53	0.70	115.8	0.82
G332.9868-00.4871	2.8	2.46	0.60	143.5	15.57
G333.0682-00.4461	2.7	32.40	0.76	135.5	22.42
G338.9196+00.5495	2.6	24.40	0.51	120.4	22.86
G339.6221-00.1209	2.5	6.04	0.51	79.5	16.87
G345.5043+00.3480	3.2	90.00	0.90	201.5	22.49
SDC20.775-0.076_1	2.6	0.03	0.57	104.2	-14.29
SDC20.775-0.076_3	2.5	0.61	0.42	119.9	18.53
SDC22.985-0.412_1	4.3	42.70	0.78	138.7	9.50
SDC23.21-0.371_1	4.2	84.20	0.86	104.0	20.49
SDC24.462+0.219_2	2.5	0.42	0.52	35.6	2.32
SDC25.426-0.175_6	3.3	0.12	0.54	123.0	23.75
SDC28.147-0.006_1	2.6	0.15	0.57	205.6	8.56
SDC28.277-0.352_1	3.1	6.30	0.41	160.0	14.44
SDC29.844-0.009_4	2.8	0.70	0.57	160.4	24.15
SDC33.107-0.065_2	3.0	4.27	0.62	97.3	21.33
SDC42.401-0.309_2	3.0	6.68	0.46	165.2	23.18
SDC43.186-0.549_2	2.4	4.00	0.42	72.5	0.59
SDC43.311-0.21_1	3.0	4.22	0.53	190.4	5.89
SDC43.877-0.755_1	3.7	15.30	0.32	97.3	12.42
SDC45.787-0.335_1	3.0	29.20	0.49	59.7	7.56

TABLE C.2: Physical parameters of CH_3CN observational spectra predicted using using the xgboost model.

source	FWHM (km s^{-1})	column density ($\times 10^{16}$ cm^{-2})	source size ($''$)	excitation temperature (K)	V_{LSR} (km s^{-1})
G013.6562-00.5997	4.1	34.40	0.72	152.1	19.15
G017.6380+00.1566	2.0	0.67	0.51	42.4	8.67
G023.3891+00.1851	2.0	9.63	0.66	207.2	6.96
G029.8620-00.0444	2.7	0.84	0.54	185.4	17.29
G030.1981-00.1691	2.2	3.56	0.50	114.4	23.29
G034.7569+00.0247	2.4	9.11	0.50	243.7	0.22
G034.8211+00.3519	2.9	1.24	0.63	96.5	25.79
G327.1192+00.5103	2.7	43.10	0.88	167.1	23.84
G332.0939-00.4206	2.1	36.40	0.82	221.9	25.39
G332.9636-00.6800	3.0	20.50	0.67	110.6	-2.03
G332.9868-00.4871	2.9	5.86	0.47	159.1	12.50
G333.0682-00.4461	2.4	49.50	0.84	129.5	24.70
G338.9196+00.5495	2.4	42.10	0.66	140.8	25.03
G339.6221-00.1209	2.3	24.90	0.62	97.8	16.86
G345.5043+00.3480	3.1	65.50	0.92	185.8	20.75
SDC20.775-0.076_1	2.7	0.08	0.65	59.9	-8.07
SDC20.775-0.076_3	2.6	0.54	0.64	145.7	15.19
SDC22.985-0.412_1	4.1	44.10	0.75	115.1	11.14
SDC23.21-0.371_1	4.0	54.30	0.77	156.6	17.78
SDC24.462+0.219_2	2.3	1.19	0.55	81.0	2.62
SDC25.426-0.175_6	3.2	0.31	0.64	118.0	22.88
SDC28.147-0.006_1	2.2	0.19	0.52	142.8	10.68
SDC28.277-0.352_1	2.8	22.70	0.49	129.7	14.54
SDC29.844-0.009_4	2.6	1.61	0.50	180.7	22.57
SDC33.107-0.065_2	2.8	8.10	0.46	120.6	23.45
SDC42.401-0.309_2	2.8	6.73	0.54	156.2	22.67
SDC43.186-0.549_2	2.5	2.30	0.51	129.8	-0.47
SDC43.311-0.21_1	2.8	6.33	0.57	187.1	5.31
SDC43.877-0.755_1	3.6	17.20	0.39	91.8	8.02
SDC45.787-0.335_1	2.6	12.60	0.43	117.4	8.22

TABLE C.3: Physical parameters of CH_3CN observational spectra predicted using using the tuned xgboost model.

Bibliography

Andron I., Gratier P., Majumdar L., Vidal T. H. G., Coutens A., Loison J.-C., Wakelam V., 2018, Monthly Notices of the Royal Astronomical Society, 481, 5651

Barrientos A., Solar M., 2019, in Molinaro M., Shortridge K., Pasian F., eds, Astronomical Society of the Pacific Conference Series Vol. 521, Astronomical Data Analysis Software and Systems XXVI. p. 189

Bates S., Hastie T., Tibshirani R., 2021, Cross-validation: what does it estimate and how well does it do it?, doi:10.48550/ARXIV.2104.00673, https://arxiv.org/abs/2104.00673

Bell T., Cernicharo J., Viti S., Marcelino N., Palau A., Esplugues G., Tercero B., 2014, Astronomy & Astrophysics, 564

Bergstra J., Bengio Y., 2012, J. Mach. Learn. Res., 13, 281

Berrar D., 2018, Cross-Validation, doi:10.1016/B978-0-12-809633-8.20349-X.

Beuther H., 2011.

Bonfand, M. Belloche, A. Garrod, R. T. Menten, K. M. Willis, E. Stéphan, G. Müller, H. S. P. 2019, A&A, 628, A27

Breiman L., 2001, Machine Learning, 45, 5

Brown J. M., 1971, Molecular Physics, 20, 817

Calcutt H., et al., 2018, Astronomy & Astrophysics, 616, A90

Cárcamo M., Scaife A. M. M., Alexander E. L., Leahy J. P., 2022, arXiv e-prints, p. arXiv:2205.01413

Carraro G., 2021, Astrophysics of the Interstellar Medium, doi:10.1007/978-3-030-75293-4.

Cesaroni R., Galli D., Neri R., Walmsley C. M., 2014, Astronomy and Astrophysics, 566, A73

Cesaroni R., et al., 2017, Astronomy and Astrophysics, 602, A59

Chen T., Guestrin C., 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, doi:10.1145/2939672.2939785, https://doi.org/10.1145%2F2939672.2939785

Chevance M., Krumholz M. R., McLeod A. F., Ostriker E. C., Rosolowsky E. W., Sternberg A., 2022, The Life and Times of Giant Molecular Clouds (arXiv:2203.09570)

Chicco D., Warrens M., Jurman G., 2021, PeerJ Computer Science, 7, e623

Colombo D., et al., 2014, Astrophysical Journal, 784, 3

Daubechies I., 1992, Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics (https://epubs.siam.org/doi/pdf/10.1137/1.9781611970104), doi:10.1137/1.9781611970104, https://epubs.siam.org/doi/abs/10.1137/1.9781611970104

Dayal P., 2019, Proceedings of the International Astronomical Union, 15, 43

Dobbs C. L., et al., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning T., eds, Protostars and Planets VI. p. 3 (arXiv:1312.3223), doi:10.2458/azu_uapress_9780816531240-ch001

Donoho D. L., Johnstone I. M., 1994, Biometrika, 81, 425

Dopita M. A., Stromlo M., 1988.

Efron B., Tibshirani R. J., 1994, An Introduction to the Bootstrap. CRC press

Frasca, A. Miroshnichenko, A. S. Rossi, C. Friedjung, M. Marilli, E. Muratorio, G. Busà, I. 2016, A&A, 585, A60

Henning T., Feldt M., Linz H., Antolin E. P., Stecklum B., 1990.

Henning T., Feldt M., Linz H., Antolin E., Stecklum B., 2006

Herbst E., van Dishoeck E. F., 2009, Annual Review of Astron and Astrophys, 47, 427

Heyer M., Pillai T., Ossenkopf-Okada V., Bolatto A., Goldsmith P. F., Johnstone D., Leisawitz D., Roman-Duval J., 2019, Bulletin of the AAS, 51, 26

Hougen J. T., 1962, The Journal of Chemical Physics, 37, 1433

Hung T., Liu S.-Y., Su Y.-N., He J. H., Lee H.-T., Takahashi S., Chen H.-R., 2019, The Astrophysical Journal, 872, 61

Ilee J. D., et al., 2021, Astrophysical Journal, Supplement, 257, 9

Jiménez-Serra I., et al., 2016, Astrophysical Journal, Letters, 830, L6

Krumholz M. R., 2015, arXiv e-prints, p. arXiv:1511.03457

Lada C., Lada E., 2003, Annual Review of Astronomy and Astrophysics, 41

Lery T., Combet C., Murphy G., 2005. pp 140–144, doi:10.1063/1.2077178

Li T., Li Q., Zhu S., Ogihara M., 2002, SIGKDD Explorations, 4, 49

Maoz D., 2016, Astrophysics in a nutshell; 2nd ed.. Princeton Univ. Press, Princeton, NJ

Martín S., Martín-Pintado J., Blanco-Sánchez C., Rivilla V. M., Rodríguez-Franco A., Rico-Villas F., 2019, Astronomy and Astrophysics, 631, A159

Matthews H. E., Sears T. J., 1983, Astrophysical Journal, Letters, 267, L53

Meng F., et al., 2019, Astronomy and Astrophysics, 630, A73

Meng F., et al., 2022, arXiv e-prints, p. arXiv:2208.07796

Menten K. M., Wyrowski F., Belloche A., Güsten R., Dedes L., Müller H. S. P., 2010, Astronomy & Astrophysics, 525, A77

Müller, Holger S. P. Drouin, Brian J. Pearson, John C. Ordu, Matthias H. Wehres, Nadine Lewen, Frank 2016, A&A, 586, A17

Öberg K. I., Guzmán V. V., Furuya K., Qi C., Aikawa Y., Andrews S. M., Loomis R., Wilner D. J., 2015, Nature, 520, 198

Opitz D., Maclin R., 1999, Journal of Artificial Intelligence Research, 11, 169

Pacifici C., et al., 2016, The Astrophysical Journal, 832

Pols S., Schwörer A., Schilke P., Schmiedeke A., Sánchez-Monge Á ., Möller T., 2018, Astronomy &amp$$ Astrophysics, 614, A123

Remijan A., Sutton E. C., Snyder L. E., Friedel D. N., Liu S.-Y., Pei C.-C., 2004, The Astrophysical Journal, 606, 917

Rieder S., Dobbs C., Bending T., Liow K. Y., Wurster J., 2021

Rosero V., Hofner P., Kurtz S., Bieging J., Araya E. D., 2013, The Astrophysical Journal Supplement Series, 207, 12

Rowe A. C. H., Abbott P. C., 1995, Computers in Physics, 9, 635

Sava H., Fleury M., Downton A., Clark A., 1997. pp 171 – 173 vol.1, doi:10.1049/cp:19970877

Schapire R. E., 2003, The Boosting Approach to Machine Learning: An Overview. Springer New York, New York, NY, pp 149–171, doi:10.1007/978-0-387-21579-2_9, https://doi.org/10.1007/978-0-387-21579-2_9

Solomon P. M., Jefferts K. B., Penzias A. A., Wilson R. W., 1971, Astrophysical Journal, Letters, 168, L107

Tayman J., Swanson D., 1999, Population Research and Policy Review, 18, 299

Tomáš V., 2018, PhD thesis, Vysoká škola báňská - Technická univerzita Ostrava, http://hdl.handle.net/10084/133114

Topp R., Gómez G., 2004, Statistics in Medicine, 23, 3377

Trypsteen M. F. M., Walker R., 2017, Analysis of the Spectra. Cambridge University Press, p. 76–84, doi:10.1017/9781316694435.010

Vabalas A., Gowen E., Poliakoff E., Casson A. J., 2019, PLOS ONE, 14, 1

Vastel C., Bottinelli S., Caux E., Glorian J. M., Boiziot M., 2015, in SF2A-2015: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics. pp 313–316

Wang J.-J., Chen W.-P., Miller M., Qin S.-L., Wu Y.-F., 2004, The Astrophysical Journal, 614, L105

Watson C., Churchwell E., Pankonin V., Bieging J. H., 2002, The Astrophysical Journal, 577, 260

Wooden D., Charnley S., Ehrenfreund P., 2004

Woodward P. R., 1978, Annual Review of Astronomy and Astrophysics, 16, 555

Wright s., 1921, Journal of agricultural research., 20

Yu T., Zhu H., 2020, Hyper-Parameter Optimization: A Review of Algorithms and Applications, doi:10.48550/ARXIV.2003.05689, https://arxiv.org/abs/2003.05689

Zinnecker H., Yorke H. W., 2007, Annual Review of Astronomy and Astrophysics, 45, 481

Zuo Y., 2022, Least sum of squares of trimmed residuals regression, doi:10.48550/ARXIV.2202.10329, https://arxiv.org/abs/2202.10329