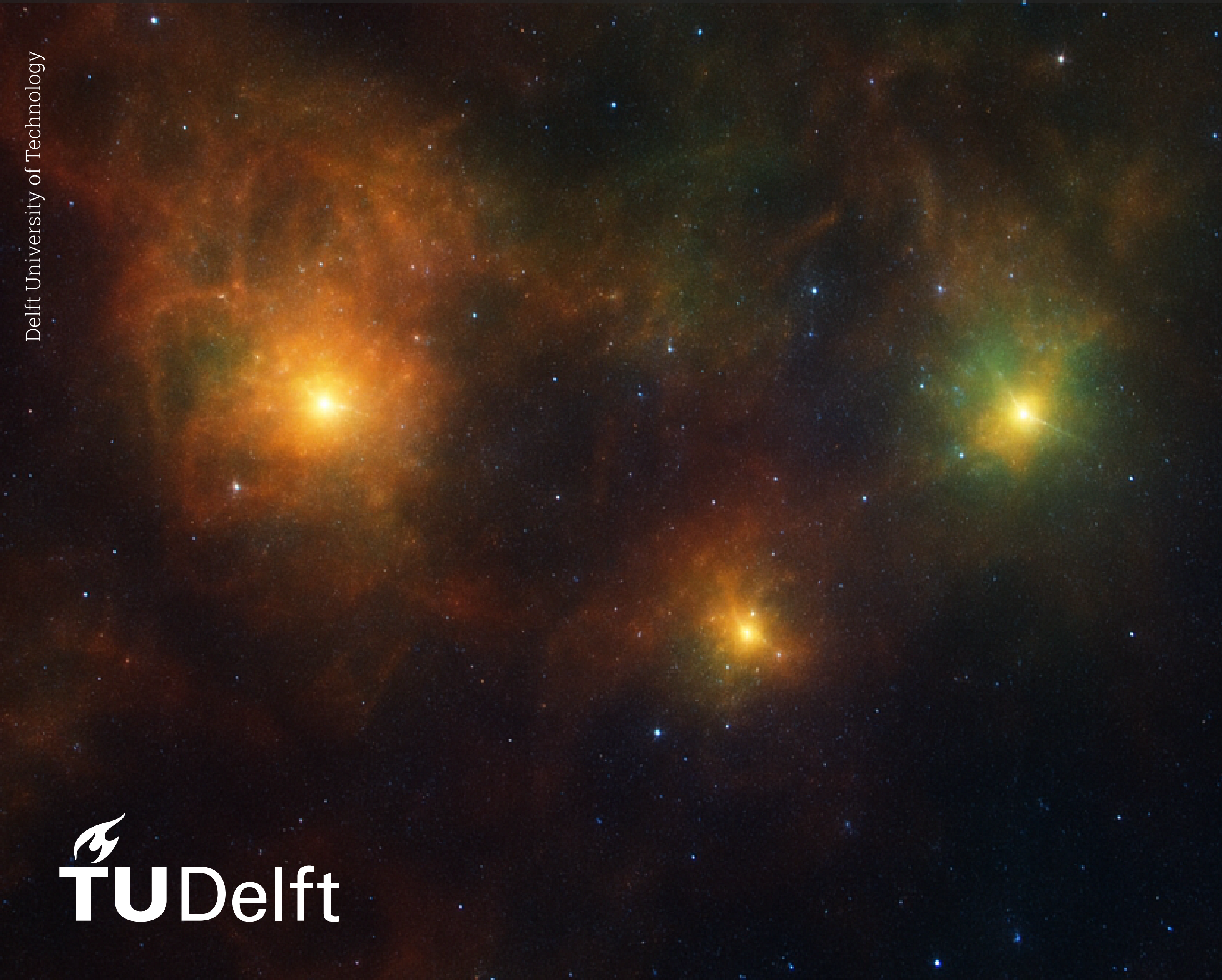
Unsupervised Machine Learning on Astrochemical Spectra

A study on high-mass star-forming regions

AE5822: Thesis Space

Javier Alonso García



Delft University of Technology

TU Delft

Unsupervised Machine Learning on Astrochemical Spectra

A study on high-mass star-forming regions

by

Javier Alonso García

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday November 7th, 2025 at 13:00.

Student number:	5228530	
Project duration:	March 1, 2025 – September 22, 2025	
Thesis committee:	Dr. K. J. Cowan MBA,	TU Delft, chairman
	Dr. S. Cazaux,	TU Delft, responsible supervisor
	Dr. W. van der Wal,	TU Delft, independent examiner
	Dr. A. Sánchez-Monge,	ICE-CSIC, external expert

Cover: Generated using ChatGPT.

Acknowledgements

I would like to thank my supervisors, Álvaro Sánchez-Monge (ICE-CSIC) and Stéphanie Cazaux (TU Delft), for their patience and their guidance throughout these months.

I would also like to thank my friends and family for their support and for motivating me to deliver the best work I could.

Javier Alonso García
Delft, October 2025

Summary

High-mass stars are one of the main drivers that shape the galaxy. Understanding the process through which they form is therefore of the utmost importance. This process, however, is not yet fully understood. Contributing to this field, the ALMAGAL survey has studied over 6000 star-forming regions with a higher resolution than any other survey before. On one hand, this data will help scientists study these obscure regions and draw a clearer picture of the high-mass star-formation process. On the other hand, the sheer volume and complexity of the data produced by this survey is far too great for conventional methods to handle swiftly.

This thesis therefore explores the use of unsupervised machine learning (ML) methods to cluster astrochemical spectra from the ALMAGAL survey. The aim of this thesis is to explore which models are best suited for the task, and to use the resulting clusters to establish a chemical evolutionary sequence for high-mass star-forming regions.

The work is done in three steps: pre-processing, clustering and analysis. In the first step, the data is re-sampled, filtered, and an algorithm is developed to calculate each region's relative velocity to Earth. The velocities obtained through this algorithm are compared to the 3672 signals for which a velocity has been manually calculated, matching in 99.16% of cases with a margin of ±5 km/s. Finally, two dimensionality-reduction techniques are investigated: Principal Component Analysis (PCA) and Uniform Manifold Approximation Projection (UMAP).

In the second step, three clustering algorithms are explored: K-means, Gaussian Mixture Models (GMMs), and Agglomerative Clustering models. The first model successfully captured 8 clusters with a high similarity within the cluster, however, the similarity between clusters was also high, indicating the presence of overlap. Furthermore, this cluster failed to identify small clusters present in the data and did not improve when using dimensionality-reduction methods. On the other hand, this model served to compare the Euclidean and cosine distance metrics, showing how the former differentiates best signals with many peaks, whereas the latter is better at differentiating signals with few peaks. GMMs were then tested on both reduced datasets, as its time scalability makes it impossible to use the complete datasets. This model performed comparable to the K-means model using the UMAP-reduced dataset, greatly reducing the clusters' overlap thanks to its soft labelling assignment. A manual inspection of the clusters was needed to achieve these results, however, reducing this model's reproducibility. Finally, the agglomerative clustering model was tested in different versions and with different distance metrics. The first version of the model made use of neighborhood maps and graph-directed linkages. This model did yield good results when applied to the UMAP-reduced dataset, however, much manual intervention was needed to achieve these results. The second version of the model, on the other hand, was a conventional agglomerative clustering model. This model was able to achieve the best results of all models tested when using the Euclidean distance metric and a Ward linkage, yielding a high similarity within clusters and a low similarity between clusters. Two final cluster arrangements were selected from this model: one with 8 clusters and another one with 19.

The final stage of this thesis focused on analyzing the physical properties of the cores being studied, and comparing the properties of the different clusters. Firstly, the size of each cluster in both arrangements was compared to the average number of spectral lines in the cluster. This revealed an exponentially decaying relationship in both scenarios, proving that the star's evolution accelerates as the region evolves. Next, the number of different clusters present in a clump was compared to the number of cores in each clump. This revealed a correlation of 0.683 between both variables in the 8-cluster arrangement, and 0.806 in the 19-cluster arrangement. Furthermore, both arrangement showed that no regions with numerous cores had few clusters, consistent with the Competitive Accretion (CA) model for high-mass star formation. Finally, the luminosity-to-mass (L/M) ratio, the median number of spectral lines, and the density of each core (both the volumetric and surface densities) were used to establish an evolutionary sequence in both arrangements. These sequences showed a Spearman correlation coefficient of 0.775,

with the correlation being near perfect for the more advance clusters in the sequence.

Overall, the study demonstrated that unsupervised ML can be used to study the chemistry in star forming regions, and that this chemistry correlates to the evolutionary stage of the region. Despite the promising results, some limitations were also identified. Most importantly, the availability of physical data was limited, as many of the variables are still being calculated. Due to this reason this study uses the overall properties of the clumps rather than that of each core for the most part. Given how regions with multiple cores were shown to have cores with different chemistries, it stands to reason to assume that their physical properties would also change. Once these properties have been calculated, a new iteration of this study should be carried out to confirm its results. Additionally, unsupervised neural networks have shown promise in literature but were not tested in this study due to time constraints. Future iterations of this work should explore these models as well to achieve an even better clustering of the data.

In conclusion, this thesis shows that unsupervised ML can be used to study the chemical structure of star-forming regions in large datasets. The results include automated and replicable methodologies, and physical interpretations of the data being studied. Once these results become validated through a manual analysis of the variables, it should be possible to apply these methods to future studies, increasing the speed at which our understanding of the high-mass star-formation process evolves.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ALMA	Atacama Large (sub)Millimeter Array
AU	Astronomical Unit (~1.496E8 km)
BCS	Between-Cluster Similarity
CA	Competitive Accretion
COM	Complex Organic Molecule
CNN	Convolutional Neural Network
D	Deuterium
DR	Dimensionality Reduction
GC	Galactic Center
GHC	Global Hierarchical Collapse
GMC	Giant Molecular Cloud
GMM	Gaussian Mixture Model
HMC	Hot Molecular Core
HMMYSO	High-Mass Young Stellar Objects
HCH_{II}	Hyper-Compact H_{II}
ICA	Independent Component Analysis
IR	Infrared
ISM	Interstellar Medium
k-NN	k-Nearest Neighbors
KDE	Kernel Density Estimation
LTE	Local Thermal Equilibrium
MC	Molecular Cloud
MDC	Massive Dense Core
ML	Machine Learning
NMI	Normalised Mutual Information
pc	Parsec (~3.086E13 km)
PC	Principal Component
PCA	Principal Component Analysis
Sgr	Sagittarius
SNR	Signal-to-Noise Ratio
SOM	Self-Organizing Map
spw	Spectral window
SVM	Support Vector Machine
TC	Turbulent Core
UCH_{II}	Ultra-Compact H_{II}
UMAP	Uniform Manifold Approximation and Projection
UV	Ultraviolet
WCS	Within-Cluster Similarity
WN	Weighted Neighbors
YSO	Young Stellar Objects

Symbols

Symbol	Definition	Unit
c	Speed of light. Constant at $3 \cdot 10^5$	[km/s]
d	Distance	[-]
d_c	Cosine distance	[-]
d_e	Euclidean distance	[-]
f	Frequency	[Hz]
$L_\odot$	Solar luminosity	[W]
L/M	Luminosity-to-mass ratio [$L_\odot/M_\odot$]	
$M_\odot$	Solar mass	[kg]
S	Cosine similarity	[-]
s_s	Silhouette score	[-]
T	Temperature	[K]
v_s	Velocity with respect to Earth of a particular source	[km/s]
x_s	Signal data	[Jy]
x_r	Residual data	[Jy]
μ	Mean	[-]
σ	Standard deviation	[-]

Literature Study

0.1. Star formation

A star's mass greatly influences the reactions that take place in its interior and therefore it is the most used criterion for determining different types of stars. Stars with masses smaller than $0.08M_{\odot}$[1] cannot sustain deuterium burning within their cores. This is therefore the barrier between stars and brown dwarfs ([50, 5]). Stars with higher masses are then sub-divided into two groups, namely, low- and high-mass stars. As seen in Figure 1, low mass stars will expand into a planetary nebula upon their deaths, which will ultimately result in the creation of a white dwarf. High-mass stars (those with masses higher than $8M_{\odot}$), on the other hand, will lead to a supernova explosion upon their deaths, which will create either a neutron star or a black hole ([5]).

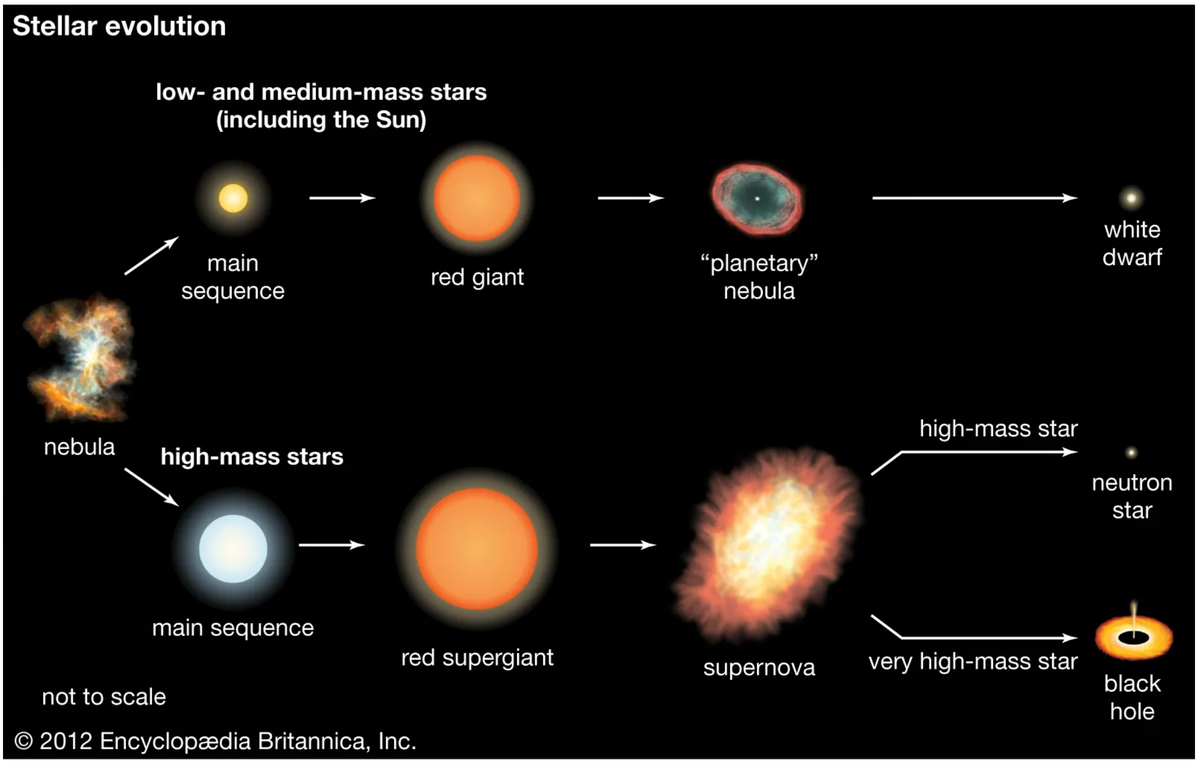


Figure 1: Evolution of low- and high-mass stars depending on the mass of the star.[2]

Understanding the formation of both low- and high-mass stars is of utmost importance as these objects supply most of a galaxy's energy and thus shape it into the way we see it today. For this reason, the star formation process has been a topic of much interest since the 1960s([5]).

In this chapter, a brief overview of the interstellar medium (ISM) will be given, followed by the formation processes of low- and high-mass stars. Within these sections, the observed and theorized processes, as well as the techniques and limitations used to study them will be provided.

[1] 1 solar mass ($M_{\odot}$) = 2×10^{30} kg

[2] Retrieved from `https://www.britannica.com/science/star-astronomy/Star-formation-and-evolution`

0.1.1. The ISM

As its name indicates, the ISM is the space between the stars. It is composed of matter, radiation, and gravitational and magnetic fields. Around 99% of all matter in the ISM is in the gaseous state, and made up of hydrogen (70%), helium (28%), and heavier elements such as oxygen, carbon and nitrogen ([62]).

Matter, however, is not homogeneously distributed throughout the galaxy, but rather clumped together in different structures due to gravity and stellar feedback. This space was originally divided into two phases by Field, Goldsmith, and Habing [25], who modeled the ISM into a cold-dense phase ($T < 300K$) consisting of atomic and molecular clouds and a warm ($T \approx 10^4 K$) inter-cloud phase consisting of neutral and ionized gas. A third group was later on added by McKee and Ostriker [40], which consisted of very hot gas ($T \approx 10^6$) which had been shock-heated by supernovae events. Finally, some authors sub-divide the first class into two groups: small, mainly atomic gas clouds, and larger, denser, mainly molecular gas clouds ([62]). These last group is often referred to as “dense clouds" , “dark clouds" or “molecular clouds" (MCs) in literature. The latter will be used for the remainder of this work.

The largest of these last group of clouds are known as “giant molecular clouds" or GMCs and extend for tens of parsecs[4] ([62]). The densities within MCs can be of up to 10^{12} molecules per cubic meter (m^{-3}), whereas the average density in the ISM is of just 10^6 m^{-3} ([62]). Furthermore, the presence of dust within these clouds makes them optically thick at visible wavelengths as seen in Figure 2. This is because dust will absorb the visible and ultraviolet (UV) radiation. This will lower the destruction rate of molecules, as they will not be photodissociated ([62]). Additionally, the increased density will increase the reaction rate of elements within the cloud, enhancing the production of molecules ([62]). The increase in molecule production, coupled with the decrease in their destruction rate causes MCs to become a molecule-rich environment. These molecules follow turbulent supersonic motions which eventually lead to an inhomogeneous mass distribution within the cloud. This distribution leads to the creation of elongated features known as filaments, within which one may find a clump of mass with a density much higher than that of its surroundings, usually of sizes around 0.1 pc. Finally, within these clumps one may find regions with an even higher density known as “dense cores" of size around 0.01 pc which is where stars are believed to originate ([62, 50]). These cores are subsequently divided into 2 categories, starless and pre-stellar, depending on whether or not they have formed a protostar. Similarly, starless cores can be gravitationally bound or not, and only those bound by gravity are expected to form stars ([50]).

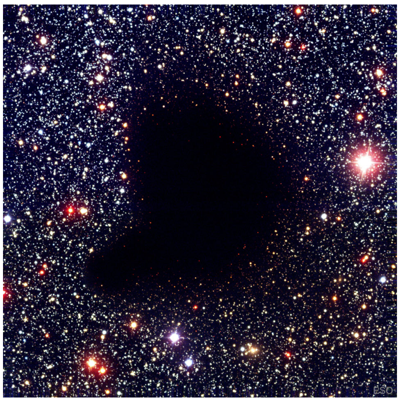


Figure 2: Image of the Dark Molecular Cloud Barnard 68, captured by the FORS team.[3]

0.1.2. Low-mass stars

Low-mass stars are the most studied out of all types of stars due to their higher abundance and proximity to Earth, making the process through which they are formed is well documented. Their formation begins with the collapse of the MC’s dense core and ends with the start of the main sequence, marked by the sustained hydrogen burning at the star’s core ([62, 50]). This main sequence makes reference to the Herzsprung-Russel diagram, which tracks how the mass and luminosity of a star evolves from the point when its luminosity spikes due to sustained fusion reactions and up to the stars death.

Collapsing cores where a star is being formed but that do not yet have a central hydrostatic central object are known as pre-stellar cores. These cores begin their collapse into a central concentration of matter, releasing gravitational energy as radiation (known in literature as “accretion luminosity"). As the core grows denser, however, this radiation finds it harder to escape the core, and so the luminosity decreases with time. The temperature in the outer edge, however, is kept roughly constant as the interior remains convective. This process is known as the **Hayashi track** (represented by the downwards vertical motion in Figure 3) and ends when the central object enters its main accretion phase. ([62])

[3] Retrieved from `https://apod.nasa.gov/apod/ap171008.html`

[4] 1 parsec (pc) = $3 \times 10^{16} m$

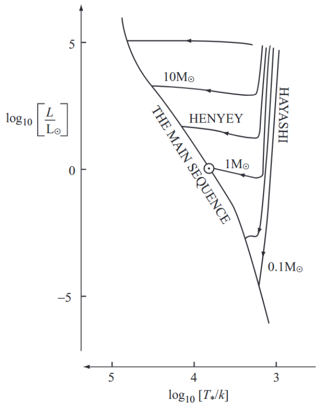
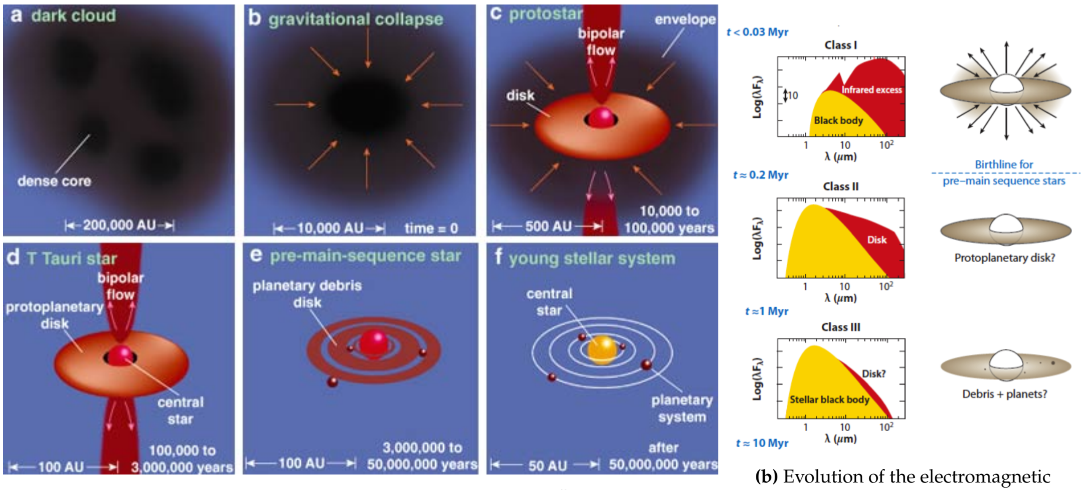


Figure 3: Theoretical pre-main-sequence tracks on the Hertzsprung-Russell diagram. Retrieved from Ward-Thompson and Whitworth [62].

During the main accretion phase, the central object builds most of its mass and forms an accretion disk in a process known as the **Heney track**. This process radiates more accretion luminosity, therefore increasing the object's temperature as represented in Figure 3. Furthermore, said radiation becomes the dominant mode of energy transportation whereas up until till now it had been convection. This causes the temperature to increase, making the envelope slightly more luminous. At the same time, a fraction of the accreted material is ejected at high velocity, in well-collimated flows known as bipolar outflows ([62]). The outflow mechanism is not yet fully understood, however, it is believed to help release the excess angular momentum of the accreted material and studies have successfully correlated their opening angle to the object's evolutionary stage ([62, 50]). The Heney track stops when the star reaches the main sequence, marked by sustained hydrogen burning by the protostar which greatly increases the star's luminosity. These fusion reactions exert radiation pressure onto the remaining envelope, which causes the star to stop contracting and the surrounding gas to disperse ([62]).

The first type of young stellar objects (YSOs) which have not yet reached the main phase are called **Class 0** protostars. They are deeply embedded into a circumstellar envelope more massive then the protostar itself, and by an emission spectrum similar to that of a black body at 15-30K ([62]). They are represented in Figure 4a by step b, where the thick envelope still completely surrounds the protostar. These protostars evolve to **Class I** protostars once the surrounding dust has been heated up to the point where there is non-trivial dust emission, which occurs at a bolometric temperature of about 70K. This type of protostar is represented by step c in Figure 4a, with the additional radiation being labeled “infrared excess" in Figure 4b. Finally, once the surrounding envelope is consumed or cleared due to stellar feedback, **Class II** protostars appear ([50]). These objects, also called T Tauri stars, do not have a surrounding envelope of material, but still present an accretion disk (see step d in Figure 4a). This disk will mass to the protostar and emit its own radiation as shown in Figure 4b. This will continue until it reaches enough mass to sustain fusion reactions in its core, giving way to the star's main phase. This is, however, not its only feature as planets will originate from this circumstellar disk too, sometimes called a protoplanetary disk.



(a) Stages of the pre-main-sequence star formation[5].

(b) Evolution of the electromagnetic spectrum in Class I, II and III stars.[6]

Figure 4: Visual representation of the stages of star formation and the evolution of the detected electromagnetic spectrum at different stages of the star's evolution.

[5] Retrieved and adapted from https://www.americanscientist.org/sites/americanscientist.org/files/2003426104214_307.jpg

[6] Retrieved from http://astronomy.swin.edu.au/~smaddiso/teaching/yso.html

0.1.3. High-mass stars

High-mass stars are much rarer than their low-mass counterparts, however, a star's luminosity is proportional to the cube of its mass, meaning that these star provide most of a galaxy's luminosity [7]. These stars shape the entire galaxy throughout their lifetime via intense radiation fields and at their deaths through supernovae explosions. This stellar feedback is directly linked to the dissolution of star clusters, the destruction of GMCs and the creation of heavy elements which allow life as we know it, as well as being the main driver of galactic-scale outflows ([50]).

Despite their undisputed importance, much is yet to be known about the process through which these stars are formed. This is because high-mass stars are fewer in number, which means that the closest high-mass star-forming regions are further away, limiting the spatial resolution with which they can be observed ([62]). This has been mitigated in recent years with the creation of observatories like ALMA (Atacama Large (sub)Millimeter Array), which should be able to resolve the cores, filaments, clumps and even the accretion disks according to numerical simulations ([50]). The biggest problem, however, is that high-mass stars begin their main sequence before clearing their environment ([2, 50, 62, 44]). Specifically, the Helmholtz time (the time until fusion begins) for a high-mass star is approximately 10^4 years, however, the free-fall timescale of the collapsing GMC where these stars are formed is about 10^5 years for a cloud density of 10^5 m^{-3}, implying that the high-mass stars will be burning hydrogen during 90% of the cloud's collapse ([2]). For this reason, all pre-main-sequence events will be deeply embedded into a GMC which will absorb most of the emitted radiation, limiting our ability to observe them.

Additionally, there are also theoretical limitations that limit our understanding of the high-mass star formation process. Original models attempted to simply scale-up the process documented for low-mass stars, however, the radiation pressure exerted by the protostar would stop the material from collapsing before the star had reached levels similar to those observed throughout the galaxy. This would constitute what is to date known as the "pressure barrier", which made scientists develop two main theories to high-mass star formation.

The first of these theories is known as **Competitive Accretion (CA)**, proposed by Bonnell, Bate, and Zinnecker [7] and refined in subsequent publications. This theory makes use of the fact that high-mass stars always form in groups (also referred to as clusters in literature) to propose a model where several low-mass protostars are initially formed within a GMC. These stars would then favour accretion to those at the center of the cloud due to the combined gravitational potential, which would increase the accretion rate and generate enough pull to overcome the radiation barrier. With time, the central stars would then acquire enough mass to be considered high-mass stars and their increased luminosity would disperse the cloud. A simplified image depicting this process can be seen in Figure 5. An alternative theory to this one was also proposed by the same authors where the low-mass stars would combine and fuse together into a high-mass star. However, this process known as coalescence was identified by the authors as an exceptional case and not the common rule due to the low likelihood it occuring ([2]).

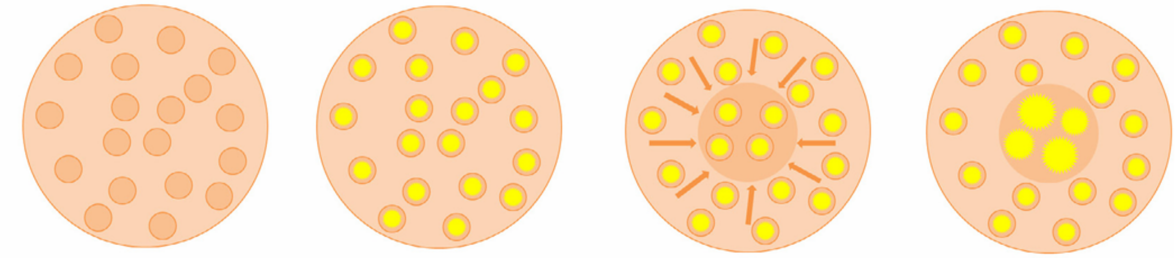


Figure 5: Stages of the high-mass star formation according to competitive accretion. From left to right: (1) denser cores form within the GMC, (2) low-mass protostars appear within those cores, (3) the combined gravitational potential promotes the infall of gas to the central part of the cluster, (4) high-mass stars form from the most massive and centered stars. Retrieved and adapted from Rivilla et al. [49]

.

An additional theory, known as the **Turbulent Core (TC)** theory, was later proposed by McKee and Tan

[7] https://www.mpia.de/en/psf/research/high-mass-star-formation

[41]. In this model, the process followed by high-mass stars would be very similar to that of low-mass stars. In this scenario, a massive near-virial-equilibrium starless core would collapse to form either a massive star or a close binary system with an accretion disk ([41]). This accretion disk would provide an anisotropic accretion flow behavior, allowing the radiation to escape primarily on the polar directions. Numerical simulations have proven that this is one of the possible mechanisms through which the pressure barrier is avoided ([50]).

Both theories have their merits, however, none of them paints a full picture of the process involved in high-mass star formation. On one hand, the TC model implies a rotating motion around high-mass young stellar objects (HMYSOs) which has been detected in several high-resolution ALMA observations, although only Csengeri et al. [18] has been able to demonstrate the existence of a Keplerian accretion disk ([50]). On the other hand, the TC model also assumes the existence of starless and pre-stellar massive dense cores (MDCs) which are prevented from fragmenting due to higher turbulence than in low-mass cores and/or strong magnetic fields ([41]; retrieved from [44], and [5]). These starless cores should be 1-10 times more common than protostellar MDCs, however, only 2 high-mass pre-stellar candidates have been found to date ([44]). Furthermore, these pre-stellar cores have been calculated to have a statistical lifetime close to a single free-fall time, which directly contradicts the TC model's assumptions. Finally, the turbulence level observed in high-mass star-forming regions is similar to that of low-mass star-forming regions, also contradicting the main assumption of the TC model ([5]).

The lack of detected high-mass pre-stellar cores has given raise to a new interpretation: that high-mass pre-stellar cores do not exist. This interpretation follows from the **Global Hierarchical Collapse (GHC)** model proposed by Vázquez-Semadeni et al. [61], which can be interpreted as an expansion of the CA model. In the GHC model, the different sub-structures within MCs are never in equilibrium but rather either contracting or dispersing, and always accreting from their parent structure until stellar feedback becomes significant enough to disperse the clouds ([61, 60]). As seen in Figure 6, low-mass pre-stellar cores would grow by accreting from its surroundings, therefore bypassing the need for high-mass pre-stellar cores ([44]). These low-mass cores would then go on to create a protostar, which would turn into a high-mass protostar as it continued accreting from its envelope. The envelope itself would not run out of mass since it would also be accreting from the filaments that transport the mass from the surrounding cloud ([61]).

Whichever the process through which high-mass young stellar objects (HMYSOs) are formed, the remaining stages of the star-formation process are clearer. First, the central massive object will begin burning hydrogen and turn the dense core into a hot molecular core or HMC ([62]). This process heats up the surrounding dust envelope to temperatures between 100 and 200 K, which releases the frozen molecules in its surface, therefore creating a much more chemically rich environment than its predecessor ([62, 50]). Some of the molecules released in this process are carbon chains such as C^+, HCO^+, C_2H_2, C_3H^+ and $C_6H_7^+$ ([62]).

Once the protostars at the center of the cluster become hot enough, they will start to ionize their environment, leading to the H_{II} region phase. These regions typically have several young massive stars at the center with luminosities of at least $10^4 L_\odot$[8] and surface temperatures surpassing 20,000 K. These stars thus emit a large number of photons which ionize the surrounding hydrogen gas. Due to the rapid increase in temperature and subsequent dissociation of molecular hydrogen into H+ (H_{II} represents singly ionized hydrogen), the total pressure within the cloud expands rapidly, absorbing the dense neutral gas around it. This will occur until the point where enough cool, neutral gas has been absorbed for the H_{II} region to become gravitationally unstable and fragment. These new fragments may then create new stars or even massive star that will then lead to a new H_{II} region ([62]).

H_{II} regions can be sub-divided into 4 categories depending on their ionization expansion. These regions grow from hyper-compact H_{II} regions (HCH_{II}), to ultra-compact H_{II} regions (UCH_{II}), then to compact H_{II} regions and finally into classical/developed H_{II} regions ([44]). These regions have been used several times to attempt to derive an evolutionary sequence based on their chemistry, maser types and luminosity. This is because their physical and chemical properties change as the envelope evolves, allowing researchers to use them to establish a timeline. An example of this timeline parameter is given by the bolometric luminosity ratio, which can be used to qualitatively separate early and late stages of

[8] $1\ L_\odot = 3.84 \times 10^{26}$ W.

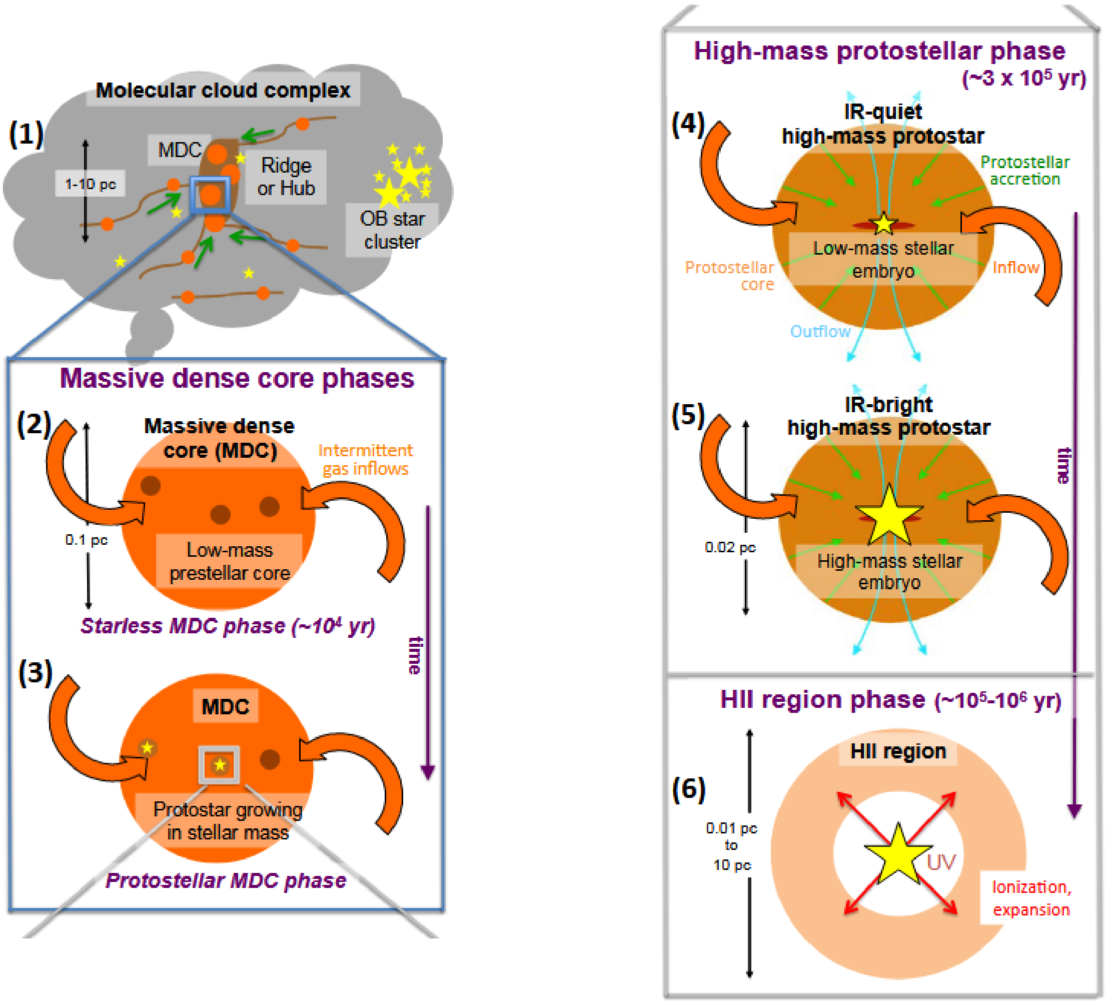


Figure 6: High-mass star formation diagram according to the GHC model. (1) Star-forming massive dense cores (MDCs) form within the filaments and clumps present in the collapsing giant molecular cloud (GMC). (2) Low-mass pre-stellar cores appear within the MDC. (3) As the MDC collapses onto the pre-stellar cores, these form low-mass protostars which continue feeding from their environment. (4) Gravitationally-driven inflows lead to the formation of high-mass IR-quiet protostars. (5) The protostars evolve into IR-bright protostars once the stellar embryo grows to a mass of at least 8 $M_{\odot}$. (6) UV radiation from the high-mass protostar ionizes the surrounding envelope, creating an H_{II} region. Retrieved from Motte, Bontemps, and Louvet [44].

their evolution ([44]). It was originally believed that the formation of the H_{II} region marked the end of the accretion process, however, recent studies show that ionized material may still be accreted onto the central protostar from the accretion disk in the HCH_{II} phase ([5]).

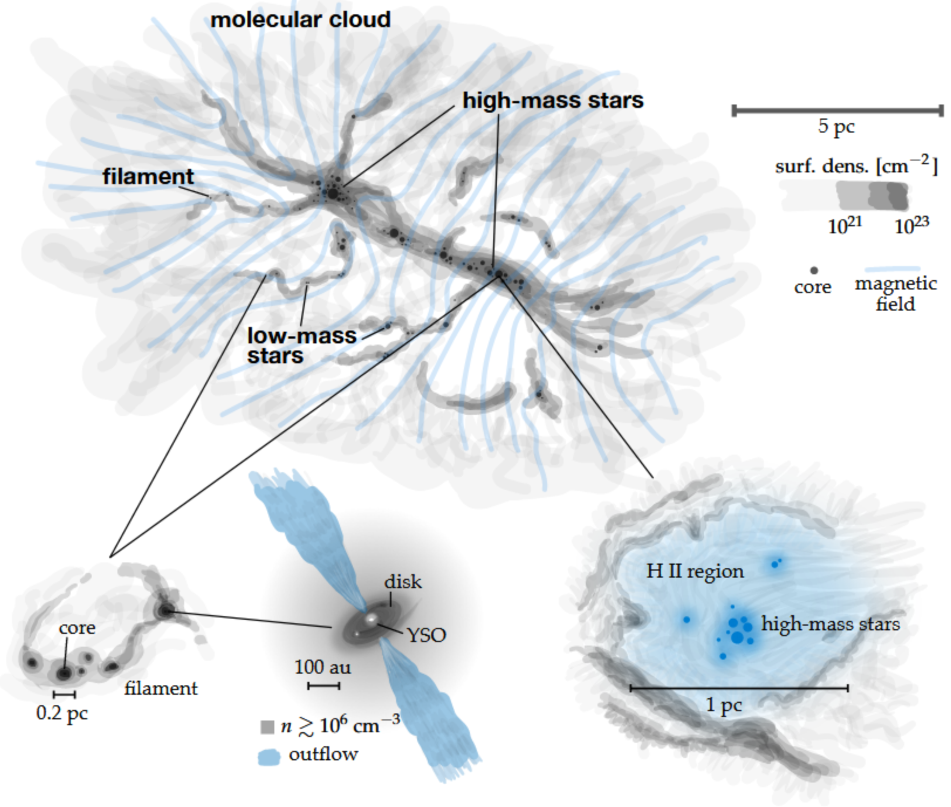


Figure 7: Sketch of a star-formation complex with both low- and high-mass stars. The top part of the image shows a large, magnetized, filamentary cloud with different densities. The bottom part shows sketches of the cores, outflows, accretion disks and H_{II} regions together with their size scales. Figure created by André Oliva, retrieved from Beuther, Kuiper, and Tafalla [5].

Overall, much is yet to be known about the star-formation process of high-mass stars, such as the role and importance of magnetic fields, or the concrete accretion processes through which they form. On the other hand, many advancements have been made in these last years. The presence of rotating motion around HMYSO and the detection of a Keplerian accretion disk around a high-mass protostar by Csengeri et al. [18] support the TC model where high-mass stars form in a scaled-up version of the low-mass star process. On the other hand, the lack of detected high-mass pre-stellar cores points to alternative mechanisms such the GHC model, where high-mass protostars form through the accretion of low-mass protostars from the surrounding core, which is in turn replenished by accreting material from the surrounding GMC. Despite the mechanism used to build them, the high-mass protostars will then heat up their surrounding core into a hot molecular core (HMC) and later ionize it to form an H_{II} region, which will eventually fragment as it grows.

0.2. Molecular Clouds and Dense Cores

Molecular clouds (MCs), and particularly giant molecular clouds (GMCs), are the birthplace of high-mass stars. They are regions in the ISM characterized by the presence of molecular hydrogen and dust, with a higher-than-average density. The dust within these clouds absorbs light in the visible spectrum, giving them the appearance of dark patches in the sky and the name of “dark clouds" (see Figure 2). Studying these clouds can be challenging since only the radio and sub-millimeter wavelengths can escape the clouds, however, their study is of great importance in order to understand the origins of stars. For this reason, this chapter will focus on reviewing the literature pertaining to giant molecular clouds and their dense cores within them.

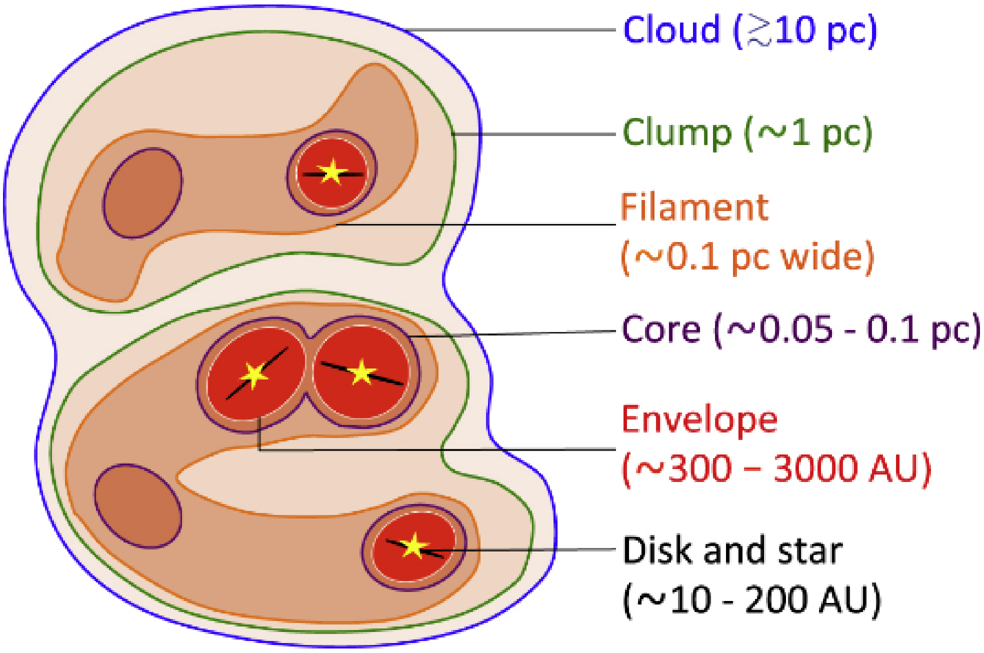


Figure 8: Hierarchical structure of a molecular cloud (MC). Retrieved from Rosen et al. [50].

0.2.1. Giant Molecular Clouds (GMC)

Giant molecular clouds are regions of the ISM with average molecular densities of about 10^{12} molecules per cubic meter and where the presence of dust absorbs UV-radiation, allowing hydrogen to convert to H_2 ([62]). These GMCs are many times found close to one another in regions known as giant molecular complexes, defined as cloud ensembles of size close to 100 pc and masses of $3 \times 10^5 - 3 \times 10^6\ M_\odot$ ([44]). Several authors also argue that parameters such as the density and star-forming capacity should also be used to define these complexes, however, no consensus has been reached.

MCs have a hierarchical structure such as the one shown in Figure 8. The schematic represents a MC responsible for the formation of low-mass stars so the cloud size is not representative of GMCs. The structure and nomenclature employed, however, is still the same and the scale of the smaller sub-structures remains valid for most GMCs. Additionally to these features, when two or more filaments or ridges (a denser, more elongated type of filament) intersect in GMC complexes, they form structures known as hubs. These structures are parsec-scale accumulations of mass prone to the formation of high-mass stars ([44]).

Normal MCs are bound by certain mass and size limitations known as the Jeans mass and Jeans length. The Jeans mass represents the mass a cloud can have before its gravitational potential overcomes the cloud's thermal pressure, causing the cloud to collapse[9]. Similarly, the Jeans length is the critical radius at which the thermal pressure will be in equilibrium with the gravitational potential[1]. These parameters, however, do not hold for GMCs where high-mass stars are formed. These clouds usually have masses higher than the Jeans mass, which means that other mechanisms not contemplated in this theory (turbulence or electromagnetic fields, for example) prevent the cloud from collapsing ([50]). In the TC model, this mechanism was assumed to be strong supersonic turbulence within the cloud, however, numerical simulations show that the turbulence generated within these clouds would not be high enough to support the GMC ([60]).

An alternative mechanism that could prevent the collapse of a GMC could be the presence of strong magnetic fields, however, these are very hard to study. They are typically measured through dust polarization, although only recently have studies managed to do so with sub-parsec resolutions ([44]). Furthermore, some studies have found that the measured magnetic fields are too weak to sustain massive dense cores (MDCs), and therefore the initial setting of the TC model ([44]).

[9] https://en.wikipedia.org/w/index.php?title=Jeans_instability&oldid=1262919161

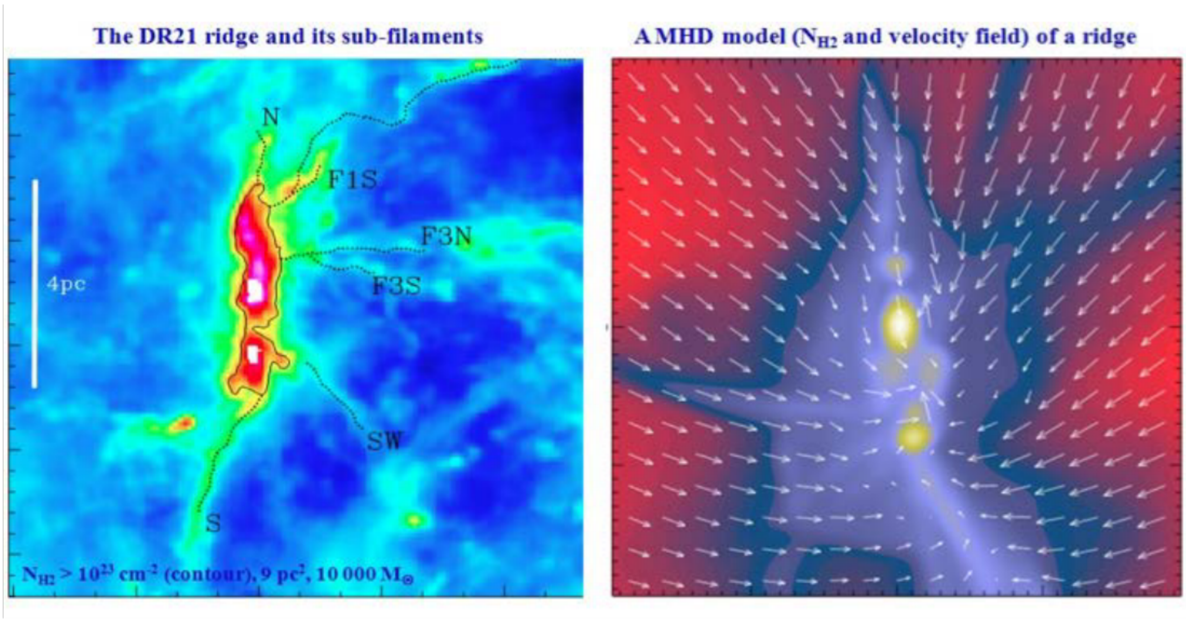


Figure 9: Comparison between the DR21 ridge and its feeding sub-filamentary network (left), and numerical simulations of the collapse of a massive elongated clump (right). Retrieved from Motte, Bontemps, and Louvet [44].

Given all this evidence, Vázquez-Semadeni et al. [61] proposed the GHC model. Here, high-mass dense cores do not exist as the cloud starts contracting as soon as it exceeds the Jeans mass. This contraction, however, is highly non-linear being initially very slow and much more accelerated at later stages ([60]).This allows the MC to continue growing in mass as it accretes from its surroundings until such a time where stellar feedback from the high-mass stars formed in its center dissipate the cloud ([61]). While the cloud collapses, all sub-structures will collapse too at different rates and accreting from their parent structures until stellar feedback dissipates the cloud, keeping the star-formation rate (SFR) low ([61]). This model seems to be supported by numerical simulations of high-density collapsing clumps which predict the existence of ridges and hubs in colliding flow simulations, as seen in Figure 9 ([44]). This explains the most used counterargument against the GHC model, which is the lack of globally-present infall profiles in observed GMCs. This is because, as the matter flows towards the ridges and filaments that then transport the matter to the main hubs, the motion relative to the observer stops being that of a spherically-collapsing cloud which is necessary for the creation of the infalling spectral signatures ([60]).

Overall, both theories have their merits and drawbacks. The TC model's greatest drawback is the low levels of turbulence and magnetic fields detected to date, which would not be able to sustain a high-mass core. On the other hand, the formation of an accretion disk around high-mass protostars as predicted by this model has been observed by Csengeri et al. [18] and is supported by several numerical simulations ([50]). Similarly, the GHC model explains the absence of high-mass cores, however, there is still not enough direct evidence that supports this model.

0.2.2. Dense cores

Cores are found within clumps, which are sub-structures found within MCs with a much higher-than-average density (10^{-5} cm^{-3}), masses between 10 and 100 $M_\odot$ and sizes on the scale of 0.1 pc ([50, 62]). In the case of high-mass star formation, they are usually referred to as massive dense clumps (MDCs). Each MDC can host several cores where either stars or binary systems may form (see step 2 in Figure 6). These terms may be defined differently in literature, so in order to avoid confusion, the substructure capable of forming stars will be referred to as core, whereas its parent structure will be referred to as clump for the remainder of this literature study.

Once cores exceed their Jeans' mass, they will begin to collapse onto a central object unless some additional force stops it from doing so. Until the YSO enters the main phase, the core will be considered a pre-stellar core. These cores can be either cold cores or hot cores (hot corinos in low-mass stars)

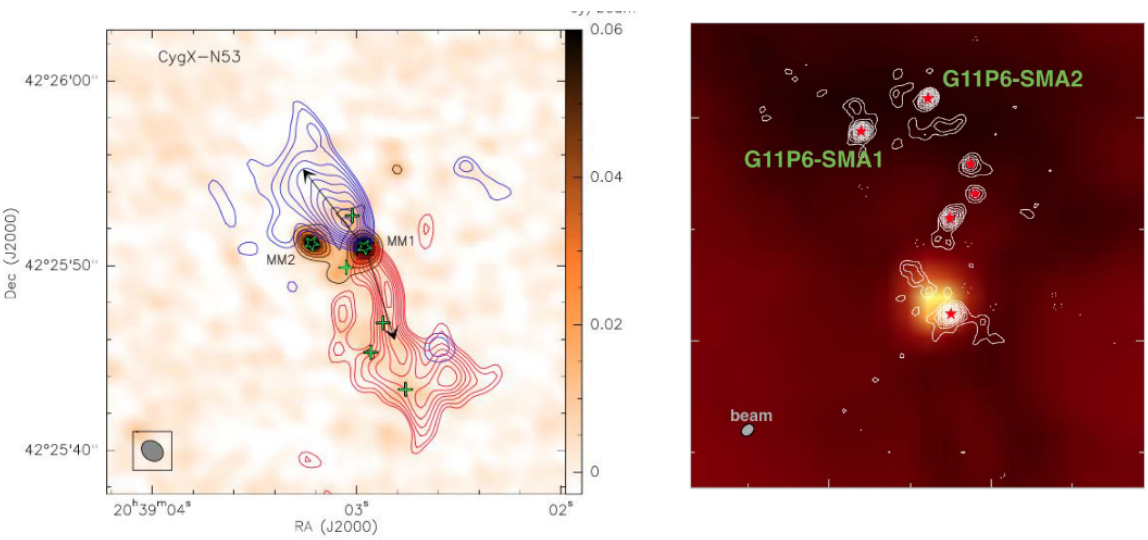


Figure 10: The only known high-mass pre-stellar cores candidates: CygXN53-MM2 and G11P6-SMA1. Retrieved from [44]

depending on their temperature. Finally, these high-mass young stellar objects (HMYSO) will begin to sustain fusion and either finish consuming their envelope or dissipate it through stellar feedback. The increased radiation will ionize its surroundings, creating an H_{II} region. In this chapter, the evolution of the core, its properties and chemistry will be analyzed from the beginning of the collapse, up to the formation of the H_{II} region.

Pre-stellar cores

High-mass pre-stellar cores are defined as gravity-bound high-mass starless cores with the capacity of forming high-mass stars. These cores have proven to be extremely difficult to locate since, despite theoretically being one to ten times more common than protostellar ones, only two have been found to date. These are CygXN53-MM2, found in the Cygnus X star complex; and G11P6-SMA1, found in the IRDC G11.11-0.12 (IR-dark cloud), as seen in Figure 10 ([44]). Originally, it was believed that starless cores had a temperature of around 20 K, however, later studies showed that 15 K was a more appropriate temperature. A new spectrum using the lower temperature was then modeled and similar spectra were searched for, prompting the discovery of more high-mass starless cores ([44]). These cores, however, were deemed unable to form high-mass stars. This was because they had similar sizes to their neighboring protostellar cores, but densities 3-10 times lower, implying that not enough mass was present in these cores to form high-mass stars. ([44]).

On the other hand, studies are now showing how high- and low-mass star formation share the same overall process at different scales. In their review, Beuther, Kuiper, and Tafalla [5] show how the turbulence in high- and low-mass cores is similar and how the star-formation process is comparably insensitive to the density structure within the core. Some theories even claim that high-mass stars do not form from MDCs, but rather from low-mass stars still accreting material from their environment. While it is still unclear which theory can explain the formation of high mass stars, low-mass dense cores can be readily found in space and have been much more studied. For this reason, the remainder of this section will focus on cores that have not yet formed a protostar, independently of their mass and the size of a star they may form.

Dense cores form due to instabilities in their MC due to either turbulence or shocks, which create regions with a greater density than the average cloud. These regions then become gravitationally bound and accrete material from their surroundings until they become gravitationally unstable and begin to collapse. At this point in their evolution, their temperature is believed to be between 10 and 20 K ([44]). Determining the exact temperature is challenging because dust grains present in these MDCs will absorb UV, visible and IR radiation, which means that only far-IR radiation can be observed. This is problematic since Earth's atmosphere will interfere with those readings, so space-based telescopes are required to measure the temperature. These telescopes, however, do not have the necessary angular

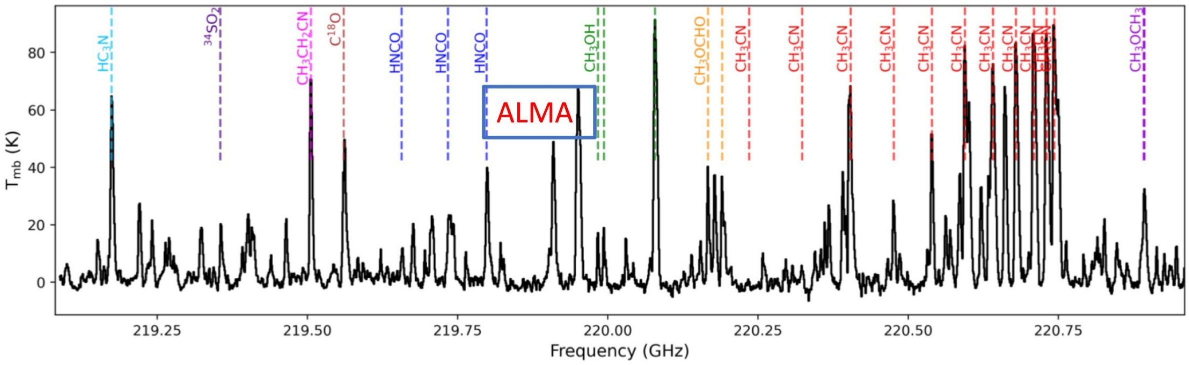


Figure 11: Spectral signature of the hot molecular core G10.43+0.03. The black line represents the registered intensity at every channel and the dashed colored lines represent the different molecules that can be observed in this example. Image created by the ESO/VPHAS+ team, retrieved from Das [19].

resolution to resolve the cores forming within the MDCs (~ 0.01-0.02 pc).

There are therefore two methods of estimating the core's temperature. The first method is simply assuming that, due to the high density in these regions, the local temperature fluctuations will be small and that the temperature of the cores is the same or close to that of the surrounding clumps. These, in turn, can be easily determined by fitting its detected continuum emission to that of a gray body ([62]). The second method attempts to determine the core's temperature independently of the surrounding clump by analyzing the spectral signatures incoming from these regions. Said spectral signatures (many time referred to as spectral lines in literature) show the chemical composition of the gas and dust present in a MC and its cores, as shown in Figure 11. Given the low temperature of cores at this stage, most molecules will be frozen on the surface of dust grains and many less chemical reactions will form, leading to a much poorer chemical richness than shown in Figure 11. Therefore, depending on which molecules can be identified, theoretical models can provide an estimate for the core's temperature based on simulations and ground tests. Additionally, assuming local thermal equilibrium (LTE) the different lines of a single molecule can be used to construct a rotational diagram that will yield the core's temperature ([42]).

Spectral lines can also be used to determine many more properties of the core, making them a primary tool for investigation. For instance, spectral signatures can be used to determine whether or not a starless core has begun collapsing. This is possible because, when a core collapses spherically, half of the material will move towards the observer and the other half will move away. Additionally, the observer will preferentially see the material closest to him, which will be the central warmer material for the half falling towards the observer, and the outer cooler material for the half moving away from him. This will create a blue- and red-shift respectively, which will generate two peaks in the reading instruments. Furthermore, since the inner material is warmer than the outer one, the blue-shifted profile will be more intense than the red-shifted one, creating a blue-skewed asymmetric profile such as the one seen in Figure 12 ([62]). Similar profiles can be created due to other effects such as the superposition of two different clouds on the same line of sight with different temperatures and velocities or an asymmetry in the cloud's velocity field, so this profile is generally considered only an infalling profile if there also is a symmetric profile in a thinner tracer such as N_2H^+ ([5]). If this is the case, however, not only can it be proven that the core is collapsing, but the rate at which it does so can be calculated based on the separation between both peaks ([62]).

Once enough material has accumulated at the center of the core, a hydrostatic object known as a protostar will form, giving way to the protostellar phase of the core.

IR-quiet cores

Once a protostar has formed within the dense core, an accretion disk will form around it independently of the core's or protostar's mass. This has been confirmed by observations such as the one by Csengeri et al. [18] and various numerical simulations ([5]). In order to preserve the material's angular momentum, polar outflows are created perpendicular to the accretion disk which eject part of the material accreted onto the protostar. This makes them easier to locate as their outflows can be observed through markers

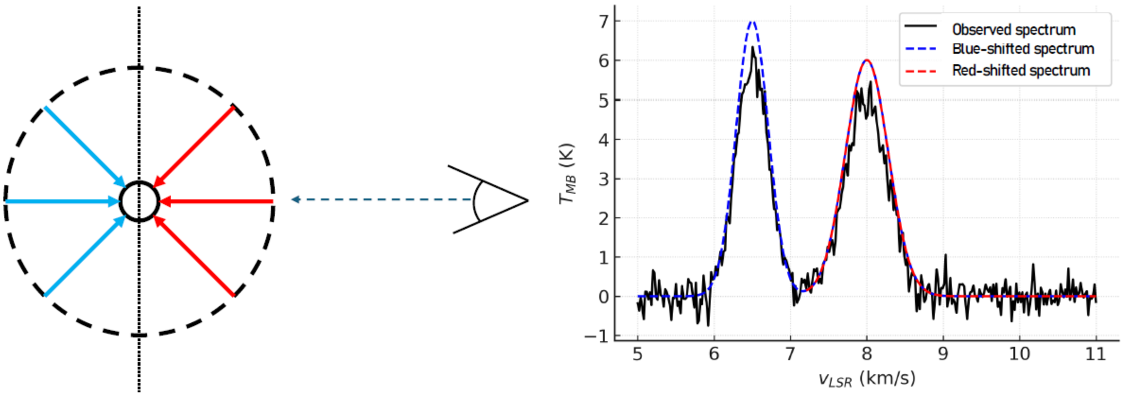


Figure 12: Schematic of a spherically-collapsing molecular cloud (left) and the corresponding blue-skewed asymmetric profile. The spectral lines on the right were retrieved and adapted from Ward-Thompson and Whitworth [62].

such as the spectral signature of ^{12}CO ([50]). Furthermore, the study of these outflows will reveal additional data about the protostar. For instance, the outflow velocity, which can be estimated based on the width of the spectral lines detected, is related to the accretion rate from the accretion disk. Similarly, the collimation angle of the outflows has been linked to the evolutionary stage of the protostar in recent studies ([50]).

These cores have been much studied in recent years in an attempt to derive an early evolutionary sequence for high-mass stars, however, much is still unknown. Duarte-Cabral et al. [21] made large contributions to this field by developing a protostellar evolutionary diagram of mass vs. luminosity and outflow momentum vs. mass (see Figure 13). In both cases, the ratios correlated with the ones found for low-mass protostars, suggesting that high-mass stars also form through protostellar accretion but with a higher accretion rate ([44]). It is thus believed that high-mass protostars and their surrounding envelopes will evolve as shown by the cyan lines in Figure 13.

Hot cores

Eventually, the accreting protostar will heat up its surrounding envelope to temperatures about 100 - 200 K, generating luminosities which may even exceed $10^4 L_\odot$ and evaporate the frozen molecules from the surface of the dust grains ([62]). Other than this, however, these cores are similar to IR-quiet cores indensity (10^6 cm^{-3}) and size (smaller than 0.1 pc) ([19, 46]).

The increased temperature, the radiation from the star and the presence of ionized atomic hydrogen will cause some of the newly evaporated molecules such as H_2O or CO_2 to dissociate and recombine into much more complex molecules than before, such as long hydrocarbon chains such as CH_3OCH_3 ([19, 42]). This will lead to a much more chemically rich environment, with many more spectral lines than before as seen in Figure 11. The process of sublimation is not immediate, however, as numerical simulations state that it may take up to 10^4 years for hot cores to go from their initial to their observed conditions ([46]). These molecules will continue to interact as parent molecules (those formed in the cold-phases of stellar formation such as H_2O) dissociate and recombine into child molecules (those formed during the hot-core phase) until the envelope reaches chemical equilibrium. This equilibrium is never reached, however, as theoretical models predict that it would take 10^6 years for this to occur but the entire star-formation process takes between 10^5-10^6 years since the beginning of the core's collapse ([42]). On the other hand, this dynamic environment allows scientists to study the hot-core's evolution by tracing the abundance of child molecules and comparing them to that of their parent molecules. More on this topic will be explained in the next section.

0.3. Chemistry in star-forming regions

The molecules present throughout the star-formation process change depending on the stage of the star formation. This is because different ambient conditions allow for different chemical reactions to take place within the region. This allows scientists to investigate the star-formation process by looking at the

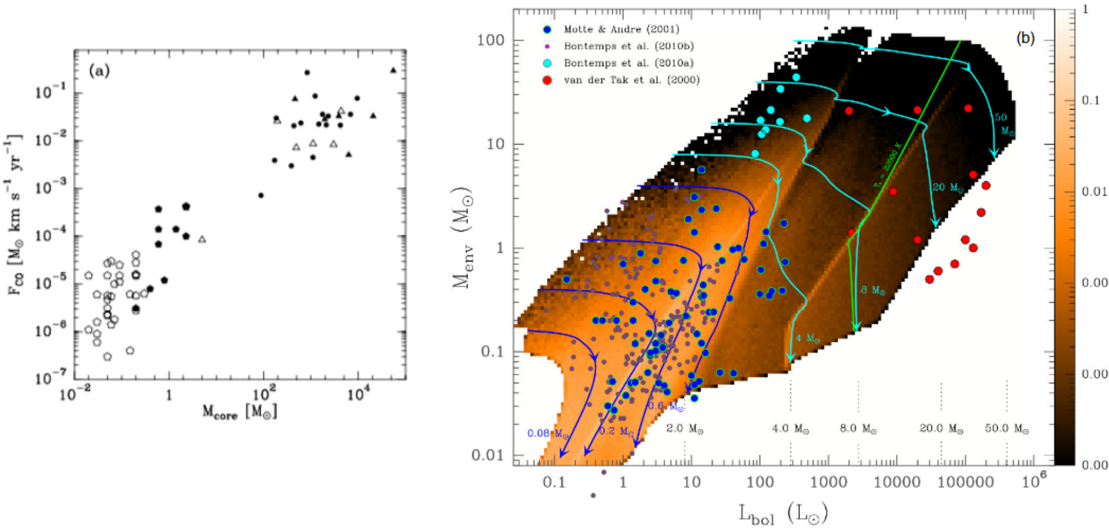


Figure 13: Mass relationship to outflow momentum and luminosity in low and high-mass protostellar cores. On the left, dots and triangles represent high-mass protostellar objects (HMPOs) whereas pentagons represent low-mass protostars. On the right, violet and cyan curves represent the evolutionary tracks of HMPOs derived by Duarte-Cabral et al. [21]. Additionally, the colored area represents the surface density predicted for protostars and the green curve separates high-mass protostars from sources developing an H_{II} region. Retrieved from Motte, Bontemps, and Louvet [44].

spectral signatures of the molecules present in the area of interest.

The evolution of the chemical processes is investigated through the combination of models and observations. There are differences between both, however, most literature agrees on the distinction of three different **processes** through which molecules form ([20, 33]):

1. **Ion-molecule chemistry**: this is a gas-phase process that takes place mostly in the cold-dark clouds before the core's collapse. The molecules formed through this mechanism are limited by high activation barriers and can only form simpler molecules.
2. **Grain-surface chemistry**: this process occurs in the surface of dust grains when molecules accrete onto their surface and react with one another. More complex molecules can be formed through this process which takes place from the beginning of the star-formation process up until the hot-core phase.
3. **Hot core chemistry**: this is also a gas-phase process, however, the high temperatures in the hot cores enable the formation of different molecules than in the first process. As its name indicates, this process takes place at the hot-core stage of star-formation.

In addition to these main types, there is also shock chemistry, which mainly occurs in regions with outflows. The effect of this chemistry is more localized, however, it needs to be taken into account when extracting conclusions about the abundance of the individual molecules.

In this chapter a breakdown of the current understanding of the chemical evolution of star-forming regions will be presented. The outline of the chapter will follow its expected chronological order. Each section will discuss the evolution of the main chemical groups (oxygen (O), carbon (C) and nitrogen (N)) in high-mass star formation regions and compare it to findings in low-mass stars.

0.3.1. Pre-collapse chemistry

Star formation begins in dense clumps within cold, quiescent clouds. The initial state of these clumps, in terms of their chemical abundance, will greatly determine the future state of the core at later stages ([13]). For this reason, the state of these clouds has been studied from observations and back-propagation models of the state of later stages. The combination of both of these methods has lead to the following inventory of O-, C-, and N-bearing species:

- **Oxygen** is found both in the solid and gaseous state in cold clouds. Solid O-bearing species account for 25 ± 5% of all oxygen in the form of silicates, metal oxides and ice water. The first two, in turn, also contain practically all available Si and Fe ([20]). Regarding gas-phase oxygen, models predict that almost all initial oxygen will be in O_2, O or CO form. Observations, however, show that O and CO (and not O_2) make up at least 40% and 20% of all oxygen respectively ([20]).

- **Carbon** can also be found in both solid and gaseous state. The composition of solid carbon has not been completely determined, although proposed materials include graphite, diamonds and polycyclic aromatic hydrocarbons (PAHs) amongst others. These carbonaceous grains are believed to make up at least 60% of interstellar carbon ([20]). The remaining 40% of carbon is present in the gaseous CO previously mentioned.
- **Nitrogen** is mostly found in gaseous phase. Models predict that N and N_2 will dominate gas-phase chemistry. The exact percentages, however, are unknown since both of them are undetectable in dense clouds. Their abundance can thus only be inferred from observations of N_2H^+. These observations seem to indicate that N_2 contains only ~10% of all available nitrogen in quiescent clouds, but that the percentage may increase in some cores such as NGC 2264 IRS1 up to nearly 100% ([20]).

The interaction of these elements with each other will change depending on the density profiles found within the core. These, in turn, will affect which molecules freeze onto the surface of the grain mantle, so understanding the different paths is of utmost importance. The effect of the **density profile** can be readily studied through the H/CO ratio as follows ([20]):

- At lower densities ($< 5 \times 10^3$ cm^{-3}), the abundance of atomic hydrogen is high and most heavy elements accrete onto dust grains as atoms. This leads to the formation of ices formed by the combination of a single non-hydrogen atom and hydrogen, namely H_2O, CH_4 and NH_3 ices.
- At intermediate densities (10^4-10^5 cm^{-3}), the higher abundance of carbon and oxygen, coupled with the decrease in atomic H, will lead to the formation of CO molecules. This molecule will then accrete onto the dust grains, where it may react with hydrogen or oxygen to make H_2CO, CH_3OH or CO_2 amongst others.
- At higher densities ($> 10^5$ cm^{-3}), most oxygen and nitrogen will be in O_2 and N_2 form. Some CO will also form, however, these molecules will not readily interact with hydrogen in the gas phase. They will therefore accrete onto the grains, where CO_2 and H_2O may be formed, the latter occurring through the H_2O_2 route.

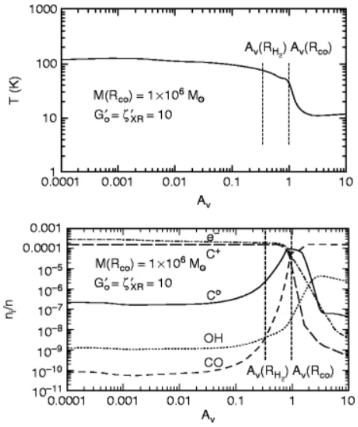


Figure 14: Temperature (top) and photodissociation regions as a function of extinction (bottom). Retrieved from Wolfire and Kaufman [63].

The ices in the first category are called "polar ices" due to the dipole moment of the molecules that form them. These ices may lead to more chemical reactions than the non-polar ones, however, since the density profiles of the cores change with time, it is also possible that both of them will co-exist. In this scenario, the ices would be layered, with the polar ices condensing first and the non-polar ones forming a volatile crust around it ([20]). Within 10^6 years, virtually all species will have condensed onto the ices, with the exception of N_2 and H_2. Molecular ions such as N_2H^+ and HCO^+ will remain in higher concentrations too due to the abundance of H_3^+ caused by the condensation of its main removal (CO, O, H_2O, ...) ([20]).

The effects of the **pressure and temperature profiles**(which, in turn, relate to the extinction profile as shown in Figure 14) can similarly be studied through the formed ices. Particularly, H_2O and CO will condense into ice only below certain temperatures, leading to the creation of "snowlines". These snowlines are the distances from the protostar at which each particular molecule will condense onto the grains, and are commonly used to determine the physical properties of a core. These lines, however, are not fixed in space. This is because accretion is not uniform, which leads to variations in the accretion luminosity emitted by the central protostar, which in turn raise or lower the core's temperature. This effect is enhanced by the formation of accretion disks, which undergo **accretion bursts** every 2-5×10^4 years which result in luminosities 1-2 orders of magnitude higher than the average

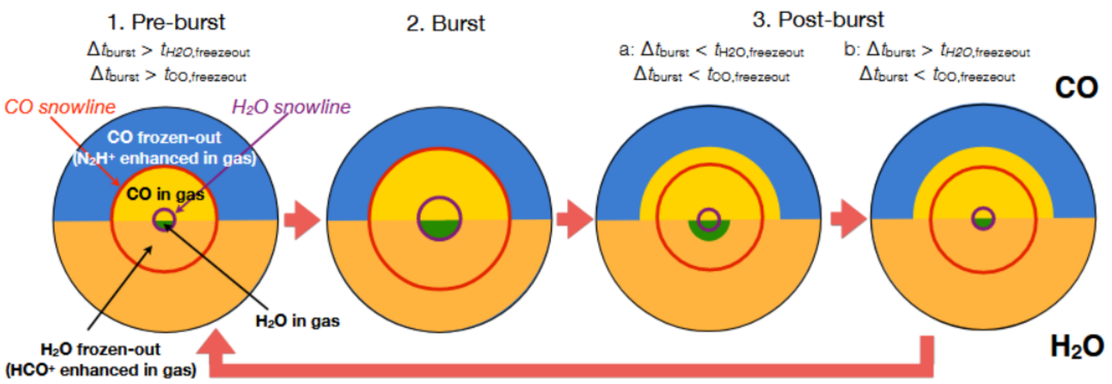

1. Pre-burst
$\Delta t_{burst} > t_{H2O,freezeout}$
$\Delta t_{burst} > t_{CO,freezeout}$
CO snowline
H_2O snowline
CO frozen-out
(N_2H^+ enhanced in gas)
CO in gas
H_2O in gas
H_2O frozen-out
(HCO^+ enhanced in gas)
2. Burst
3. Post-burst
a: $\Delta t_{burst} < t_{H2O,freezeout}$
$\Delta t_{burst} < t_{CO,freezeout}$
b: $\Delta t_{burst} > t_{H2O,freezeout}$
$\Delta t_{burst} < t_{CO,freezeout}$
CO
H_2O

Figure 15: Schematic diagram of the effects of an accretion burst on the CO and H_2O snowlines. Panel 1 shows the core before the burst. Panel 2 the enhanced snowlines created by the burst. Panel 3 shows how the snowlines return to their original position once the burst ends, and how the sublimated species persist, given how their freeze-out times are longer than the bursts duration. Finally, panel 4 sh ows how H_2O freezes faster than CO due to its smaller freeze-out time. Diagram retrieved from Jørgensen, Belloche, and Garrod [33].

([33]). The effect of this burst can be seen in Figure 15, where the snowline is pushed back by the sudden increase in luminosity from panel 1 to panel 2. The duration of the burst, however, is much shorter than the freeze-out time of the different molecules. That is why panel 3 shows the snowlines being pushed back to their original positions, however, the sublimated molecules still persist. Finally, in panel 4 it can be seen how the longer freeze-out time of CO compared to H_2O makes the latter revert back to its original state faster ([33]). This difference in freeze-out times, together with the observed and theoretical snowlines allow scientists to determine how long ago the accretion burst occurred. Furthermore, the abundance of molecules destroyed through gas-phase interactions with the sublimated ones (N_2H^+ for CO and HCO^+ for H_2O) can be used to study the degree to which the other species have been frozen onto the grains ([33]).

0.3.2. Deuteration

Molecules will mainly react with hydrogen (H), as it is the most abundant element of them all. Hydrogen can be present in many forms such as H, H_2 or H_3^+, however, the most interesting of them due to the insights it provides on the early star-formation environment is **deuterium** (D). Deuterium is an isotope of hydrogen that contains a neutron in its core besides the proton. It is present throughout the ISM, however, star-forming regions show ratios of deuterated species orders of magnitude higher, particularly for organic molecules ([12]). Furthermore, a higher D/H ratio is characteristic of cold ($\lesssim$ 10K) regions, due to the reaction below.

$$H_3{}^+ + HD \longleftrightarrow H_2D^+ + H_2 \tag{1}$$

This reaction is important since $H_3{}^+$ and H_2D^+ dissociate through electron recombination, which make them the largest sources of H and D, respectively ([33]). The higher abundance of one over the other will thus lead to a larger or smaller deuteration fraction of neutral species as they combine through grain-surface chemistry (see subsection 0.3.5).

The sensitivity to temperature of this reaction, however, causes it to be heavily influenced by environmental effects. For instance, the deuteration levels of Sagittarius B (from now on referred to as “Sgr B" as is common in literature) and other cores near the galactic center (GC) are much lower than those on more other regions such as Orion KL. This is because the increased stellar feedback present in the GC will raise the minimum cloud temperature, therefore decreasing the D/H ratio ([33]). An alternative explanation could be that less deuterium is available near the GC due to stellar processing, however, the higher deuteration levels of isolated cores investigated by Jensen et al. [32] and measurements of other hot cores around the galactic disk seem to favor the temperature interpretation. Similarly, the deuteration level around low-mass protostars seems to be 5-8% higher than in their high-mass counterparts, which can at times translate to deuterated species making up to 20% of the total species of a given molecule ([33]).

Overall, the D/H ratio achieved in this stage of the star-formation process is inherited by later stages. This means that the D/H ratio at later stages will mostly depend on the D/H ratio achieved in the pre-collapse stage since warm gas exchange reactions play a smaller role in the deuterium enrichment of molecules. The only exception found to date to this rule belongs to the CH_3CHO molecule, where the CHO- group shows a higher D/H ratio than the CH_3- group. The reason for this imbalance is unkonwn. ([33])

0.3.3. Complex Organic Molecules (COMs)

It was originally believed that the same low temperatures that favored a high deuterium fraction, also did not provide enough energy for complex molecules to form. This, however, has been shown not to be the case through the discovery of **complex organic molecules** (COMs) in cold clouds ([33]). The definition varies from paper to paper, but most authors agree on defining COMs as hydrocarbons with at least 6 atoms, often including different functional groups such as alcohol (-OH) or amino (H_2N) groups. They were believed to form through hydrogenation of simpler hydrocarbons such as methane (CH_4) through grain-surface chemistry, however, the detection of large, saturated COMs towards dense, prestellar cores have questioned this assumption ([33]).

The formation of complex hydrocarbons in cold clouds requires either neutral (C) or ionic (C^+) carbon, and can occur through three different mechanisms ([20]):

1. Carbon insertion: $C^+ + CH_4 \longrightarrow C_2 + H_2H_2^+$ or $C + C_2H_2 \longrightarrow C_3H + H$
2. Condensation: $CH_3^+ + CH_4 \longrightarrow C_2H_5^+ + H_2$ or $C_2H + C_2H_2 \longrightarrow C_4H_2 + H$
3. Radiative association: $C^+ + C_n \longrightarrow C_{n+1}^+ + hv$

Before the acceptance of grain-surface reactions, carbon insertion was thought to be the dominant route. However, since this would lead to the loss of a hydrogen atom and because larger ions ($C_nH_m^+$) do not react rapidly with H_2, this mechanism could only produced highly unsaturated hydrocarbons ([20]). Furthermore, provided there was a high enough extinction, this could only happen in short time scales as carbon would get locked into CO and form ices on the surface of grains as the cloud evolved ([58]). The impossibility of these mechanisms to form the observed abundances of COMs in cold clouds has led scientists to believe that either there is an unknown mechanism besides grain-surface chemistry through which these molecules can be formed, or that COMs do originate on grains but are released onto the gas phase through some non-thermal mechanism. The only proposed theory as to a new formation route is the reaction of H and o atoms with the surface of carbonaceous grains, however, this could also just lead to the formation of water ([33]). Alternatively, COMs could form through grain-surface chemistry (see subsection 0.3.5) and be desorbed through one of these mechanisms:

1. Cosmic-ray sputtering of the ices
2. Photo-desorption by UV photons
3. Spontaneous desorption of newly-formed molecules, referred to as “*reactive desorption*".

The first and second theories are supported by the fact that COMs have been detected primarily offset from the core’s center. The latter, however, as well as the third theory, have proven to be too inefficient in laboratory experiments to explain the observed abundances ([33]).

0.3.4. Outflows

The presence of outflows associated with the formation of protostars can have a significant effect on the chemistry of the region. These outflows will drive material formed in the protostar at very high velocities towards the surrounding envelope. This will lead to collisions between the flow and the envelope, giving raise to shock chemistry.

There are two types of shocks that occur due to outflows: C (continuous) shocks and J (jump) shocks. **C shocks** have velocities $\lesssim 40 - 50$ km/s and maximum temperatures in the range 2000-3000 K. This temperature is not high enough for molecular dissociation, but it does allow for reactions with high energy barriers to occur. This makes the formation of H_2O and H_2S occur more rapidly, driving all oxygen and sulfur into those forms in the affected area. This effect, however, is somewhat dampened by the fact that back reactions with atomic hydrogen can destroy these species again, returning them to OH and SH form. This is important since OH is a necessary molecule for the formation of molecules such as SO, SiO and NO. The abundance of these molecules will therefore depend on the H/H_2 ratio in the outflow. **J shocks**, on the other hand, have temperatures so high ($\sim 10^5$ K) that all molecules are dissociated. This effect may even be expanded to the areas around the shock due to photodissociation caused by UV photons produced in the shock. Ahead of the shock, however, a warm zone forms where molecules form again, with a similar chemistry to C shocks once enough H_2 has formed. ([20]).

The presence of outflows is also characterized by strong and broad line profiles for certain enhanced species. These include both refractory and volatile species (e.g. SiO, HCN, CH_3OH, ...) whose abundance is increased by factors of 10^6 and 5-500 respectively ([20]).

Finally, it is also believed that the creation of some masers associated with star-forming regions such as H_2O may be the product of outflow chemistry ([20]).

0.3.5. Grain-surface chemistry

Grain surface chemistry takes place throughout the entire star-formation process up until the formation of the hot-core. As the core starts to collapse, heavier molecules will accrete onto the surface of dust grains due to the low temperatures. These molecules will then be surrounded by mostly H_2O ice, preventing them from being released into the gas-phase. Molecular hydrogen, however, can pass through the ices and react with the embedded molecules (and with the ice itself) to form more complex molecules than the ones available through gas-phase reactions. This can all be observed in Figure 16, where the top part shows the hydrogen being accreted. This hydrogen atom will then be diffused through the ice and react with other embedded molecules. These molecules, in turn, may be desorbed or dissociated by incoming photons or thermally desorbed and released into the gas-phase.

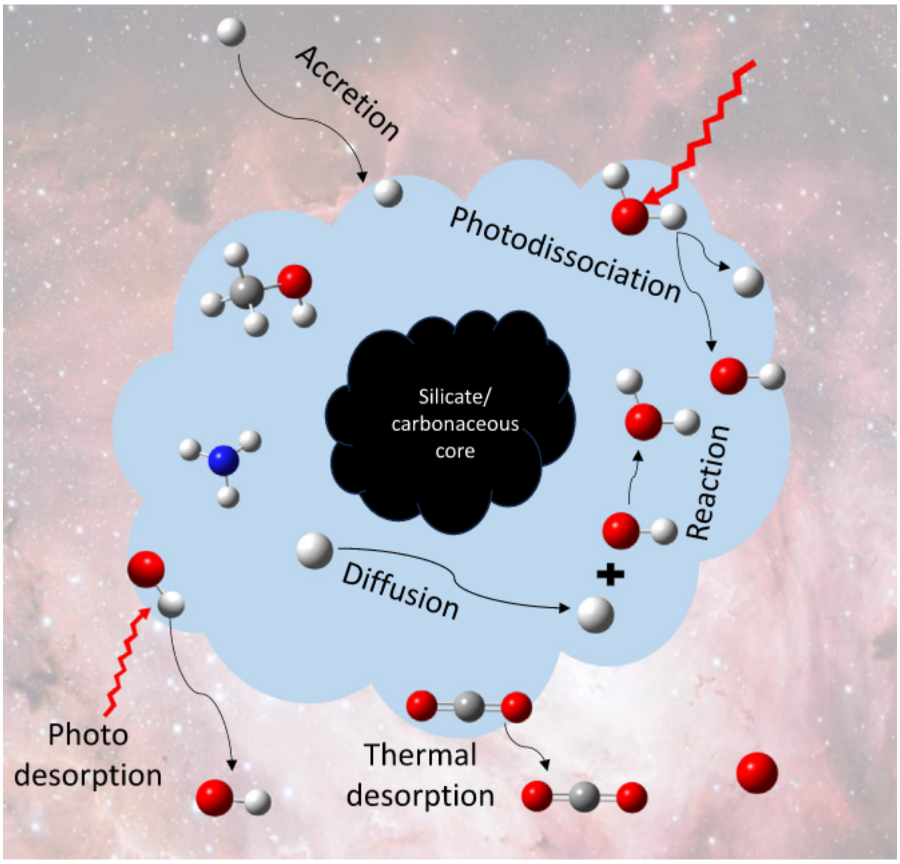


Figure 16: Schematic view of a dust grain surrounded by ice and all the processes that it may undergo. Retrieved from Das [19].

This is the process through which saturated **hydrocarbons** form in star forming regions (or, at least, the only process we have successfully identified). The simplest scenario is the one that leads to the formation of formaldehyde (H_2CO). This molecule will form through the hydrogenation of CO ice, however, this ice only reacts with atomic hydrogen and not its moleculear form, so two separate reactions will be needed ($CO + H \longrightarrow HCO$ and $HCO + H \longrightarrow H_2CO$). Since this is the only reaction path that leads to this molecule, the deuteration levels of formaldehyde, HDCO/H_2CO and D_2CO/H_2CO, will depend on D/H and $(D/H)^2$ respectively ([12]). The activation of CO through this mechanism is so efficient that observational values of CH_3OH/CO in interstellar ices have reached values as high as 1 ([58]).

The scenario is somewhat more complex for the deuteration of **water**. Most of the observed water is not formed through gas-phase reactions since it has a very high energy barrier that needs to be overcome for H_2 and O to react ($T \geq 230$ K) ([20]). Instead, oxygen atoms get accreted onto the surface of grains, where they react with either H or H_2 to form H_2O. The same path takes place for reactions with deuterium, however, the particular process that takes place depends heavily on the temperature of the ice. This can lead to one of three scenarios ([12]):

- At low temperatures ($T_{\text{dust}} \leq 12$ K), oxygen can react with both H and H_2 to form water. The latter, however, is much more abundant, so the deuteration of water will depend on the D/H_2 ratio.
- At intermediate temperatures ($T \sim 15$ K), H_2 has a higher chance of evaporating from the ice's surface. Reactions in short timescales will therefore predominantly involve H. On longer timescales, however, most atomic hydrogen will react with other molecules, decreasing its abundance and making reactions with H_2 the predominant route once again. The deuteration of water will therefore scale with the D/H ratio in short timescales and D/H_2 in longer timescales.
- At higher temperatures ($T_{\text{dust}} \geq 17$ K), H_2 molecules will sublimate from the ice's surface too fast to react with oxygen. This leaves successive hydrogenation of oxygen as the primary mechanism of water formation. Deuterated water will therefore scale with D/H.

These ratios correspond to the HDO/H_2O ratios found in distinct clouds, however, doubly-deuterated water will also form. Statistically, HDO/H_2O should be four times higher than D_2O/HDO, however, direct interferometric measurements showed that the latter is seven times higher. The reason for this is not yet known, however, researchers believe that most deuterated water will form at the earlier stages of star-formation, when deuteration is more efficient. ([33])

Finally, one avenue of research not too deeply explored is the effect of magnetic fields on the evolution of interstellar ices. If true, new chemical routes similar to those of the ion-molecule gas-phase would become possible. One scenario that has good experimental support is the following ([58]):

1. Polar ices or those with traces of alkali metals become charged on the outskirts of the core due to UV and visible photons.
2. These ices get heated up during the collapse to ~ 80 K, point at which methanol segregates out of the water.
3. This methyl transfer allows for sterochemistry to take place, leading to the formation of a weakly-bound dimethyl ether molecule, which can readily sublimate.

0.3.6. Hot core chemistry

Hot cores are regions around high-mass protostars with temperatures higher than 100 K, densities around 10^6 cm^{-3} and with a very rich chemistry. This chemistry is composed of a variety of hydrogenated molecules, as well as an abundance of COMs ([20, 58]). These molecules, just like before being accreted onto dust grains, are once again produced through gas-phase reactions once the ices sublimate. In this instance, however, the higher ambient temperatures allow for more and faster reactions than in previous phases. The hot core region will still be deeply embedded into a dense core, as seen in Figure 17. In the figure, the different snowlines and their corresponding temperatures can be seen in a schematic way. It is also shown how outflows are still present and potentially causing shocks, as well as the most characteristic molecules of each region.

Once the first-generation molecules have been released into the gas phase, most models and observations agree that it will take about 10^4 years for the second-generation molecules to reach a significant level of abundance ([46]). This happens because the parent molecules need to be broken down and recombined into children molecules, which takes time. The breaking of molecules that initiates hot-core chemistry can occur through two mechanisms ([20]):

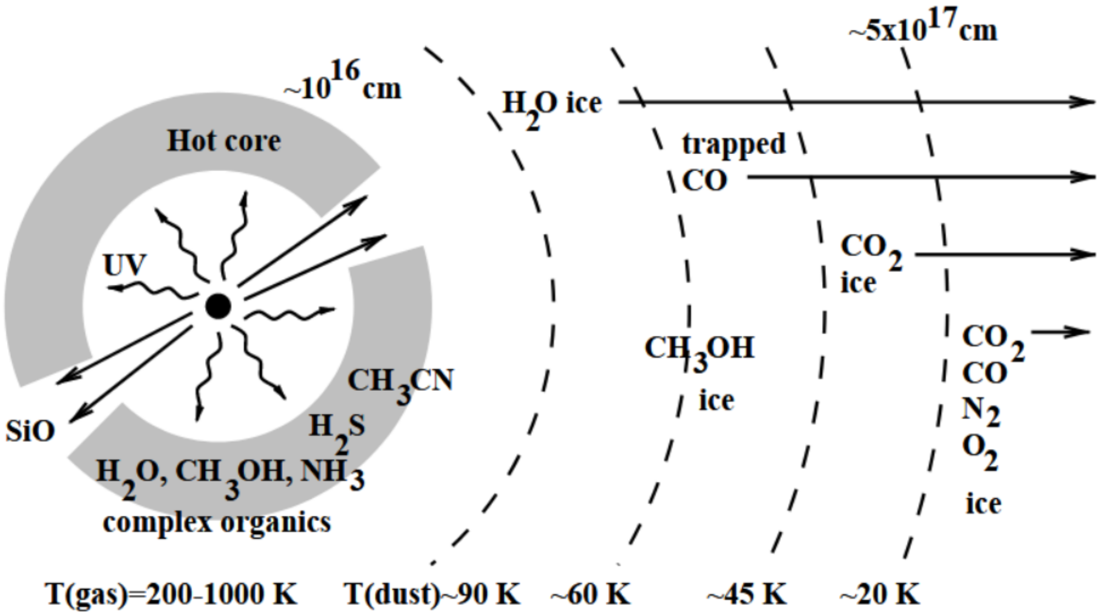


Figure 17: Schematic illustration of the chemical environment around a high-mass protostar surrounded by a hot core. Retrieved from Dishoeck and Blake [20].

1. Reactions with atomic hydrogen ($H_2S + H \longrightarrow SH + H_2$).
2. Reactions with protonated molecules, followed by dissociative recombination ($H_2O + H_3^+ \longrightarrow H_3O^+ + H_2 \xrightarrow{e^-} OH + 2\,H_2$).

Water is the main parent molecule of **O-bearing species**. It is, however, destroyed by HCO^+, which leads to the formation of H_3O^+. Hydroxyl (OH) molecules are then born through dissociative recombination ([46]). Organic molecules then formed, generally through the transfer of an alkyl (-CH_3) group to a neutral molecule ([20]). Once the core reaches a temperature of 230K, however, almost all oxygen will be transformed into H_2O as overcoming the activation barrier of the gas-phase reaction between O and H_2 will become possible ([20]). This is further enhanced by the presence of ionic helium (He^+) in the innermost region, which destroys CO and pushes yet more oxygen into forming water rather than organic molecules ([46]).

N-bearing species, on the other hand, start off either as N_2 (which never accreted onto the dust grains - [20]) or as NH_3 ([58]). Interactions with He^+ will lead to the dissociation of N_2 into N or N^+, both of which will react with hydrogen to form NH_2. Alternatively, that same molecule can be produced through the dissociation of NH_3. Interactions with carbon then lead to the formation of HCN and CN, which are the precursors of N-bearing molecules ([58]). The formation of these molecules is also enhanced above 230 K. This is because O and O_2 are the main removal partners of CN, therefore, their conversion to H_2O stops them from destroying the CN molecules ([20]). The overabundance of water also prevents H_3^+ from producing CH_3^+ from its reaction with CH_3OH, therefore reducing the abundance of HC_3N and CH_3CN. It can be therefore established that nitrogen-rich species are suppressed near the hot core, where the abundance of water prevents their formation ([46]).

In addition to the hot-core's temperature, the initial abundances of parent molecules will also play an important role in its chemical evolution. For instance, models show that at high temperatures (300 K), cores with a high presence of methanol and low ammonia will lead to the formation of O-rich species in 10^4 years, followed by N-rich species after 10^5 years. However, in those same models, N-bearing species will never become abundant at lower temperatures (100 K). On the other hand, when ammonia is injected together with methanol at 300 K, no O-bearing species are formed, while at 100 K both N- and O-bearing species coexist. ([13])

Other species of interest include S-, Si- and P-bearing species. It is believed that these elements get released into the gas-phase in the form of H_2S, SiH_4 and PH_3. These molecules, however, get rapidly destroyed at temperatures of a few hundred kelvin due to their low energy barriers, being returned to their elemental form. This atoms will then react with species such as OH and O_2 to form SO, SO_2, SiO and PO. Once again, the abundance of oxygen will limit the formation of these molecules. ([20])

Similar regions have been found around low-mass protostars, although they are given the name of hot-corinos. Hot-cores and corinos are mostly similar in their chemistry, although some systematic differences have been found. In their review, Jørgensen, Belloche, and Garrod [33] cite how CH_3OH, and HNCO or CH_3CN can be used as proxies for the abundance of O- and N-bearing species in both types of cores, especially in IRAS 16293 and Sgr B2(N2). In this particular case, O-bearing molecules appear to be even more closely related, with -CHO speices showing almost a 1-1 agreement ([33]). The relationship of these molecules to methanol, however, varies between the cores with O-bearing species showing a much higher proportional abundance in IRAS 16293 ([13]). Another difference is the level of deuteration in both cores, with the low-mass core showing as much as twice the level of deuteration of hot cores ([13]).

0.4. Machine Learning in astrophysics

Studies on star-formation processes have radically changed throughout the last years. This is partly because the access to more powerful observational instruments has led to the generation of unprecedented amounts of data for every survey carried out. Just in the last decade, data from sky surveys grew from the gigabytes to terabytes, and the trend is predicted to keep increasing up to several tens or hundreds of petabytes in the next decade ([8]).

This has led scientists to develop more computer-based techniques that rely less on direct human inspection to analyze the produced data. One of the most promising approaches is the adoption of machine-learning (ML) techniques on the classification and analysis of astronomical data. On the one hand, the use of machine learning techniques in other fields has shown an increased processing capacity, however, many of the methods employed lack explainability. That is, the reasoning behind how the results produced come to be is not understandable simply by looking at the model. Additionally, the collected data needs to be pre-processed in a different manner depending on which models it is intended to be used on, making the adoption of these methods not as straight-forward as one could have hoped.

Despite these drawbacks, the potential benefits of applying ML to streamline the data-processing makes it worth investigating. In this chapter, the work currently being done on the use of ML in the field of astrophysics will be presented. This work will be divided into three sections representing the common stages of any data-science project: data pre-processing, ML models and evaluation techniques. The works studied will be focused on (but not limited to) the study of spectroscopic data preferably from stellar sources, but also from similar processes such as Raman spectrography.

0.4.1. Data pre-processing

Data pre-processing involves several fields within the field of data science. The most relevant of these for the applications at hand, however, are three: filtering, scaling and dimensionality reduction. **Filtering** attempts to separate the noise from the readings so the data produced by the source in question is clear. **Scaling**, on the other hand, attempts to enhance certain features of the data so the models to be later used on place a higher emphasis on particular features of the data. Another common use of scaling is normalizing the data so that factors such as the distance to the source do not affect the performance of the ML models. Finally, **dimensionality reduction** attempts to decrease the input space necessary to represent the data. In practical terms, this means that a vector of size N where each point represents the intensity of the emitted radiation at a given wavelength could potentially be represented as a vector of size M where $M < N$. Depending on the particular method used, vectors with thousands of datapoints can be converted into two-dimensional vectors. These are specially useful for visualization purposes.

Filtering techniques

Filtering techniques in the analysis of electromagnetic spectra are very much comparable to filtering techniques used in signal analysis. This is useful since the latter have been much more developed given their applications to the field of communications.

One very basic but still widely used filtering technique is based on the **Fourier transform**. The Fourier transform will assume that any given finite signal is part of an infinitely repeating signal that can be represented by a sum of sines and cosines. The Fourier coefficients will then represent the magnitude of the sine for a given frequency (f). Three types of filters then stem from this information: low-pass, band-pass and high-pass filters. Low-pass filters will disregard frequencies higher than a given cut-off frequency and reconstruct the signal based on the remaining Fourier coefficients ($f < f_{max}$). High-pass filters, on the other hand, will reconstruct the signals based on the coefficients that are smaller than a given frequency ($f > f_{min}$). Finally, band-pass filters will only consider the coefficients bound between a minimum and a maximum frequency ($f_{min} > f > f_{max}$). Additionally, band-stop filters can also be used to suppress the information in a given frequency range ($f < f_{min} \cup f > f_{max}$). The use of each particular filter will depend on the application for which it is intended, however, low-pass filters are usually used for noise-reduction purposes as the background noise for most applications will have a higher frequency than the measured signal. The results of applying this filter can be seen in Figure 18, where the yellow line represents the raw signal and the red line shows the recomposed signal after passing through a low-pass filter. Here it can be seen how the fast oscillations are removed from the data, showing only the underlying trend.

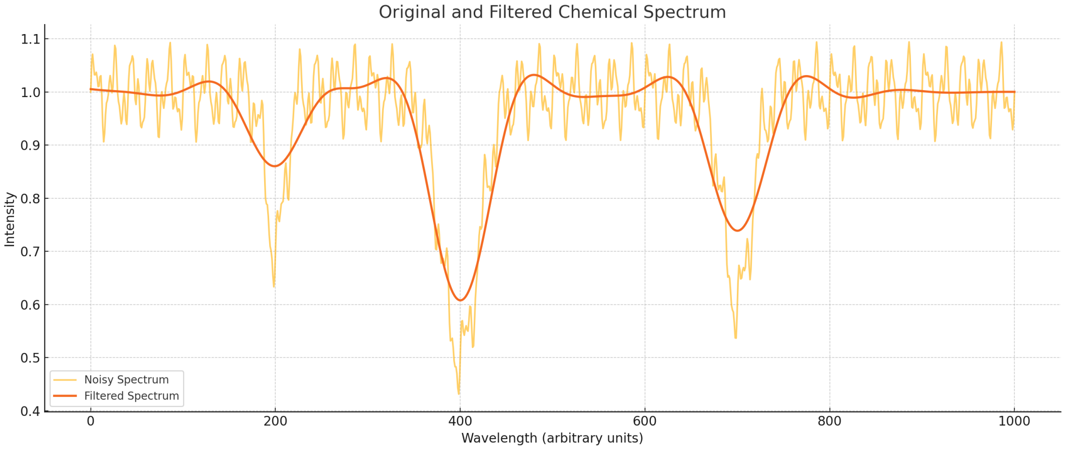


Figure 18: Comparison between a raw and low-pass filtered signal.

One drawback of the Fourier-transform is that it assumes the signal to be stationary in time. This, however, is not the case with many astronomical objects. For this reason, some authors such as Manteiga et al. [37] decide instead to use **wavelet transforms**. These transformations attempt to decompose the signals into different wavelets, which are localized functions dependent on both time and frequency, and whose magnitude represents the coefficient of each wavelet. For a more detailed description on wavelets and wavelet transforms, readers are directed to Lau and Weng [34]. The resulting wavelet will lead to a two-dimensional heatmap representing the intensity of each wavelet at every time and frequency component. At this point there are several techniques that can be used to filter the signal:

- Firstly, one may simply apply the same filters as with the Fourier transform. This time, however, the time-evolution of the frequency components can be used to better discern an appropriate cutoff frequency.
- Alternatively, one may select the highers N wavelet coefficients and use them to re-build the signal.
- Finally, by squaring the magnitude of the coefficients one will obtain the power spectrum of the signal. This spectrum will represent the energy contained within the signal which can be used for feature-extraction purposes. This will be further explained in Figure 0.4.1.

In their paper, Manteiga et al. [37] showed how using Fourier transforms can be useful for signals with high signal-to-noise (S/N) ratio, but that wavelet transforms performed better for S/N < 25. Bromová, Škoda, and Vážný [9] also used wavelets in the classification of spectral emissions, however, their best results were obtained using them for feature-extraction purposes. Overall, the main use of wavelets resides in the analysis of non-stationary signals where no single time-window can resolve the entire frequency content through Fourier transforms. This can be readily observed in Figure 19, where the signal on the left is analyzed first through a short-time Fourier transform with a time resolution of 50 ms and then through a wavelet transform.

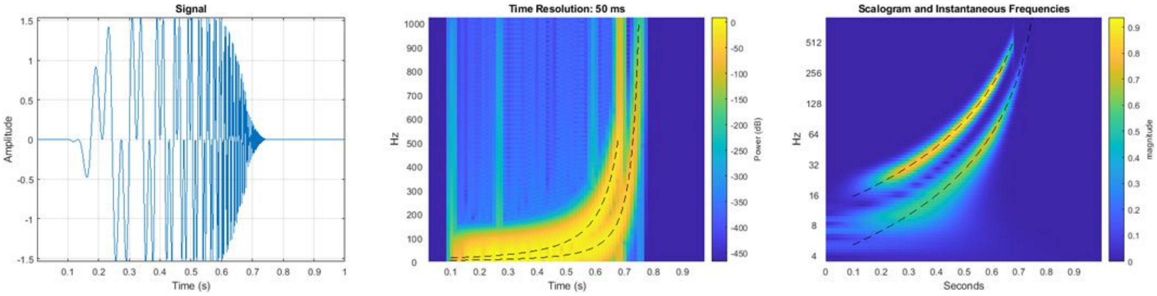


Figure 19: Analysis of the hyperbolic chirp signal (left) through short-time Fourier transforms (middle) and wavelet transforms (right).[10]

Finally, another filtering method often found in literature was the **Savitzky–Golay filter**. This filter attempts to smoothen the underlying data in a signal by successively fitting low-order polynomials to a subset of adjacent points ([53]). This procedure has been successfully applied by several authors to spectroscopic signals in order to make them more robust to noise (e.g. [30, 10]), however, it does require the user to determine both the number of points to consider for the polynomial fitting, as well as the order of the polynomial to be fitted. These are parameters that can heavily influence the outcome of the signal as seen in Figure 20. In the figure the red line considers 51 points for each fitting interval, therefore loosing many of the details of the underlying signal. The green line, on the other hand, considers just 10 points and possibly captures noise-induced perturbances. For this reason, selecting the appropriate amount of points and the degree of the polynomial to consider are not trivial tasks and can therefore be more time-consuming than other filtering methods. Additionally, [10] reported that these pre-processing steps did not significantly improve their classification performance.

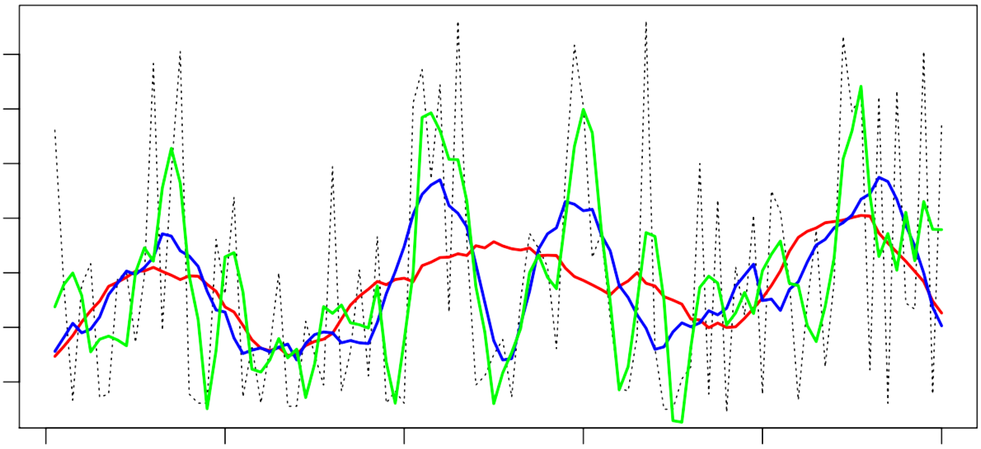


Figure 20: Example of the Savitzky-Golay filter applied to a signal (dotted black line). All filtered lines were made with fourth-degree polynomials but considering 51 points for the red line, 25 for the blue line and 10 for the green line.[11]

Scaling

The most common type of scaling in the ML field is **normalizing**. This process is intended to transform the data from whatever range it is provided in, into a standardized range. For instance, the electromagnetic spetrum detected from stars will reach Earth with different intensities, so scientists may

[10] Retrieved from https://www.mathworks.com/help/wavelet/ug/time-frequency-analysis-and-continuous-wavelet-transform.html

[11] Retrieved from https://jeffreyevans.github.io/spatialEco/reference/sg.smooth.html

want to normalize the spectrum to a [0, 1] range to better compare the different sources. Alternatively, one may want to ensure that all sources as represented with the same energy level, and so the spetral intensities will be divided by the integral of the entire spectrum. These two normalization strategies are called min-max normalization and L2 normalization respectively, and are two of the most common normalization techniques. Reviews by Singh and Singh [55] and Fontoura Costa [26] investigated several transformations for signal analysis, the most important ones being the following:

- **Standardization** - also called z-score normalization, this method is employed to make the data fit a normal distribution with mean (μ) equal to 0 and a standard deviation (σ) of 1. This method makes the data become dimensionless and is appropriate for several models that assume a Gaussian data distribution, or for problems that deal with data in different units.

$$\tilde{x}_i = \frac{x - \mu_X}{\sigma_X} \tag{2}$$

- **Pareto scaling** - it is similar to standardization, although the scaling factor used is the variance instead of the standard deviation. This minimizes the effects of noise on the data and keeps the structure of the data partially intact.

$$\tilde{x}_i = \frac{x - \mu_X}{\sqrt{\sigma_X}} \tag{3}$$

- **Min-max normalization** - this method linearly maps the data such that all sources will have their minimum value at 0 and their maximum value at 1. Alternatively, one may use max-normalization instead, which divides the data by its maximum value, ensuring that all points lie in the [-1, +1] range. Both of these methods are very used since they prevent models from being biased towards features with higher magnitudes.

$$\tilde{x}_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{4}$$

- **Baseline correction** - this method is used to ensure the signal's mean is 0. This is specially important in spectral analysis where the baseline radiation detected by instruments is not constant. In this cases, the baseline (denoted below by $\mu(x)$) needs to be corrected before applying any other transformations. Readers are directed to Eilers and Boelens [22] for details on how the baseline can be calculated.

$$\tilde{x}_i = x_i - \mu(x_i) \tag{5}$$

- **Probability density function transformation** - this method transforms the data such that the minimum value is 0 and the area under the graph is 1, making it akin to a probability density function (PDF).

$$\tilde{x}_i = \frac{x_i - \min(x)}{\sum(x_i - \min(x_i))} \tag{6}$$

- **Average scaling** - the presence of noise may cause methods which scale the signals based on the extreme values more dangerous to use. In this cases, a common approach is to scale the signal based on its mean intensity.

$$\tilde{x}_i = \frac{x_i}{\mu_X} \tag{7}$$

In their review, Singh and Singh [55] compared these and more transformations to a total of 21 datasets in order to find the best ones. They found that no single method outperformed the others across all datasets, however, max-normalization and baseline correction methods performed best for data with high variance. Similarly, Fontoura Costa [26] applied several normalization techniques to both artificially generated signals and measurements of hand-written digits. In their review, they find that standardization provides the best feature separation out of all tested methods.

An alternative use of scaling parameters is to **enhance** certain features. For instance, in the analysis of chemical spectra such as the one seen in Figure 11 one may want to enlarge peaks that separate from the baseline but that are not too large in comparison to the most intense peaks. Some popular methods include the following ([55]):

- **Power transformation** - this method is mostly used for data with an uneven distribution of its noise (heteroscedasticity) and with a standard deviation proportional to the root of its mean. One large limitation of this transform is that it does not work with negative numbers, so the data must be previously shifted.

$$\tilde{x}_i = \sqrt{x_i} - \mu_i(\sqrt{x}) \tag{8}$$

- ***Tanh* normalization** - this method reduces the influence of the values at the extremes, making the data more insensitive to outliers.

$$\tilde{x}_i = \frac{1}{2}\left[\tanh\left(\frac{x_i - \mu_X}{100\,\sigma_X}\right) + 1\right] \tag{9}$$

- **Logistic sigmoid normalization**: similarly to *tanh* normalization, this method also reduces the effects of the data on the extremes. However, it does so using the logistic sigmoid. The main advantage of this is that data at the extremes will be squashed into the (0, 1) range, whereas the data in the middle will be linearly transformed.

$$\tilde{x}_i = \left[1 + \exp\left(-\frac{x_i - \mu_X}{\sigma_X}\right)\right]^{-1} \tag{10}$$

- **Sigmoidal transformation** - this function scales the signal in a similar way to the previous two. It was not in the previously-mentioned reviews, but authors such as Carey et al. [10] and Sevetlidis and Pavlidis [54] use them in their analysis of Raman spectra.

$$\tilde{x}_i = \frac{1 - \cos(\pi x_i)}{2} \tag{11}$$

An alternative to these methods was used by Sevetlidis and Pavlidis [54] who used ML methods to classify Raman spectra. In their work, they first applied a square root transformation ($f(x) = \sqrt{x}$) followed by a max-normalization of the data ($g(x) = f(x)/f_{\max}(x)$) and a sigmoid transformation ($h(x) = [1 - \cos(\pi g(x))]/2$). This was done in an attempt to diminish the effect of peak intensity differences while normalizing the dataset. The results can be seen in Figure 21.

Which method to choose is task-dependent, however, this choice may have a large impact on later processing stages. For example, Sevetlidis and Pavlidis [54] experienced a 13.8% increase in their model's accuracy when applying pre-processing as opposed to classifying the raw spectra.

Dimensionality reduction

The final pre-processing step usually involved in ML projects is dimensionality reduction (DR). This process attempts to reduce the number of points needed to represent the signal, in order to improve computing speed and, on occasions, for visualization purposes. The first of this reasons is especially important in ML since many of the models employed suffer what is known as "the curse of dimensionality". This is the name given to the phenomenon involving several ML models which do not linearly scale with the amount of data and thus, are very much slowed down by increased amounts of data. This occurs for instance with the Gaussian Mixture Model (GMM), which calculates the covariance matrix of all the input features. This means that for an input vector of size N, N^2 parameters will need to be calculated. This is not a problem when hand-picking a few data features, however, if one wanted to use a star's entire electromagnetic spectrum as input, each wavelength channel would be interpreted as a feature, leading to millions of datapoints needed to be calculated.

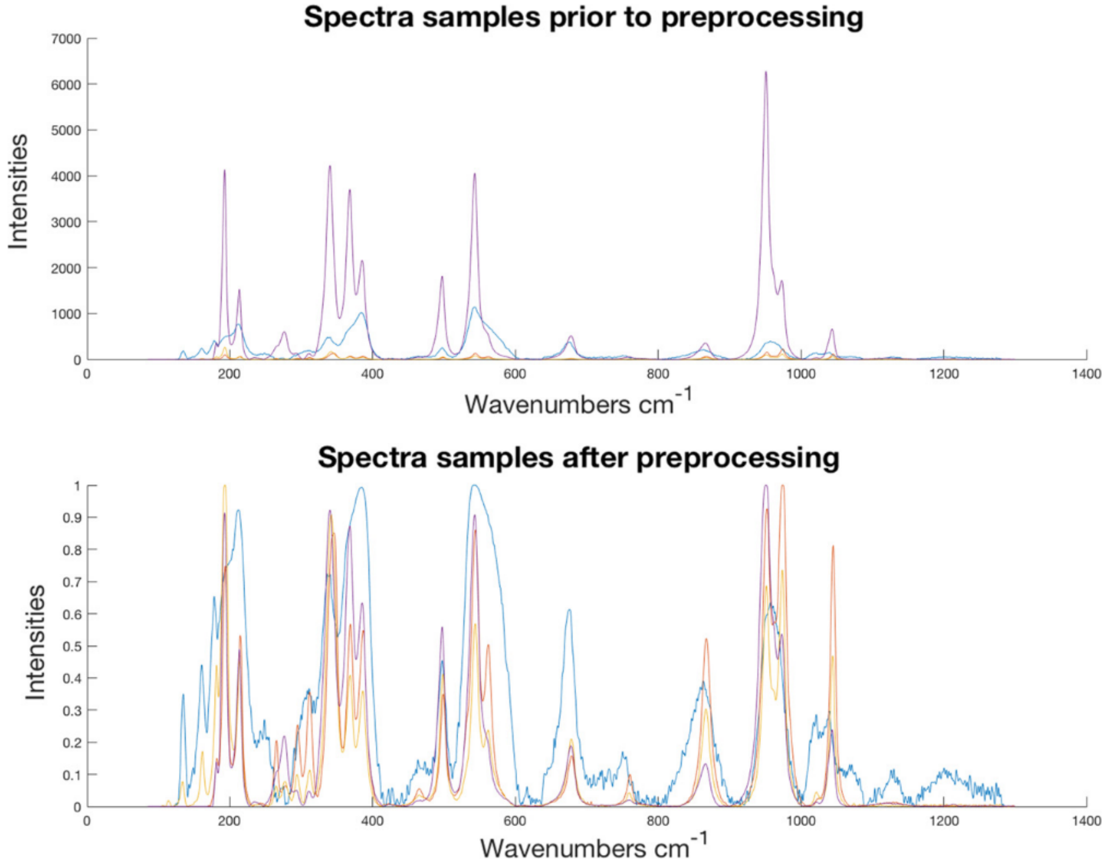


Figure 21: Effect of applying scaling and normalization methods to several Raman spectra. Retrieved from Sevetlidis and Pavlidis [54].

DR methods can be broadly split into two different categories: linear and non-linear methods. The most popular of the linear methods is **Principal Component Analysis (PCA)**. This method calculates a set of patterns (called principal components or PCs) which, when linearly combined, can be used to reconstruct any given signal from the data set. Furthermore, each component will have an associated importance depending on how many underlying signals share that structure. This can lead to the reconstruction of signals to a high degree of accuracy with just a few components as shown by Ishikawa and Gulick [30], whose analysis of Raman spectra showed that just the first 3 components contained 88.1% of the variance in the data (which originally consisted on 765-dimensional vectors). An example of these PCs can be seen in Figure 22, where the images on the left show pre-processed magnetic resonance spectra studied by Yang et al. [67], followed by their top three principal components. There are two sets of PCs in this figure since the paper uses two separate methods to calculate them. However, in all cases it can be seen how the middle peak in the middle, which shared by all signals, is shown in the first component. Similarly, features shared by fewer samples such as the right-most peak showed by the blue lines are only captured by later components.

There are, however, two caveats to the use of PCA. Firstly, the reconstruction will not be perfect unless the number of components is equal to the number of features. This, however, will never happen since identifying as many components as there are features implies that there will be no DR involved. Secondly, the components identified will depend on the number of components being calculated. This can be harder to understand, however, consider the case where the number of components is equal to the number of features. In this scenario, all components would be a unitary point at every channel. This would mean that any possible signal with that number of channels could be reconstructed from these components, however, it stands to reason that a different set of underlying structures needs to be provided when the number of components is smaller than the number of features.

PCA is one of the most widely used DR techniques since it is fast to compute and easy to understand.

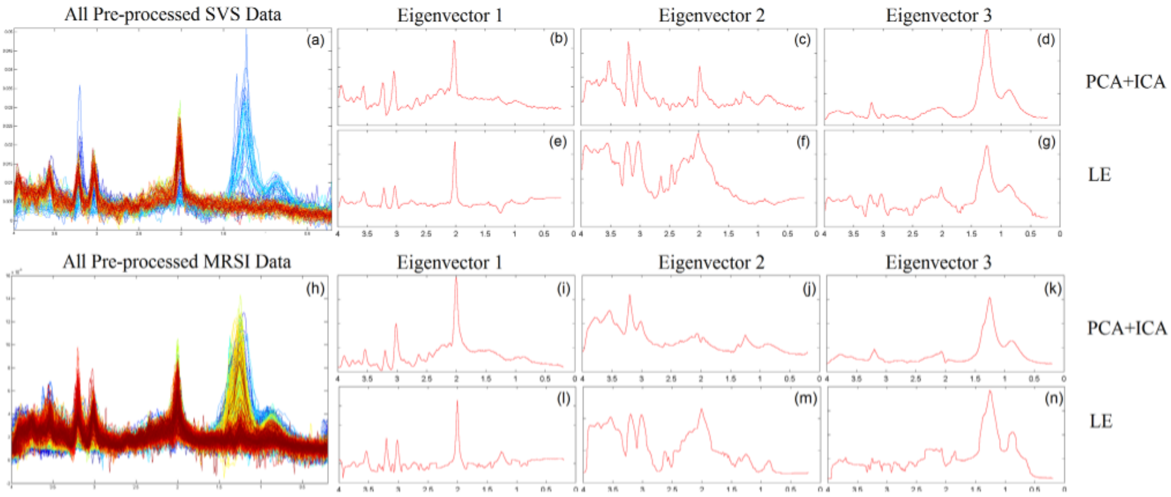


Figure 22: Representation of two sets of magnetic resonance spectra and their respective principal components. Retrieved from Yang et al. [67].

Furthermore, this method allows scientists to greatly reduce the number of datapoints being considered while preserving the accuracy of the model. For instance, one may consider the example of Carey et al. [10]. In their paper, the original data set consisted on several 1715-dimensional Raman spectra for which they managed to obtain a classifier accuracy of ~83% with the entire input space. After applying PCA, however, the accuracies were 82.5% and 81% when considering 100 and 50 PCs, representing 98% and 90% of the data variance respectively. The case of Ishikawa and Gulick [30] is even more extreme, since in their paper they report that the accuracy of their classifiers stagnated when considering more than 5-8 PCs, depending on the pre-processing techniques involved.

PCA, however, has certain limitations that are exploited by other non-linear dimensionality reduction techniques. The main disadvantage of PCA in the analysis of chemical spectra is that it ignores the sequential nature of these signals, treating each channel individually. This potentially results in the loss of information which could be exploited by other algorithms ([10]). For this reason some authors prefer non-linear DR techniques, also known as **manifold learning**. There are several options that explore manifolds in the data to find a lower dimensional representation, however, authors in spectroscopic analysis seem to favor UMAP (Uniform Manifold Approximation and Projection) ([39]). This is because this method preserves both the local and global structure of the signal, as opposed to other manifold learning techniques such as t-SNE which only preserve local information ([38]). One drawback of UMAP, however, is that there are four hyperparameters that need to be chosen: how much of the local structure is preserved, the minimum distance between the reduced signals, the number of output dimensions and the distance metric used for comparing the different signals.

The balance between the local and global information is chosen through the number of neighbors that the algorithm considers when fitting the local manifold. The effect of this parameter can be seen on Figure 23. In this example, choosing only 2 neighbors results in the algorithm preserving only the local information, as evidenced by the unstructured output. Considering 20 neighbors, on the other hand, leads to a balance between the preservation of local and global information as shown by the clusters that start to appear, as well as paths that connect them. On the other hand, choosing a number of neighbors too high will lead to the preservation of mostly the global structure, leading again to a mostly unstructured output. The number of neighbors to be used will depend on the desired application, however, both Xu and McCord [66] and McConville et al. [38] both use 20 neighbors in their analysis.

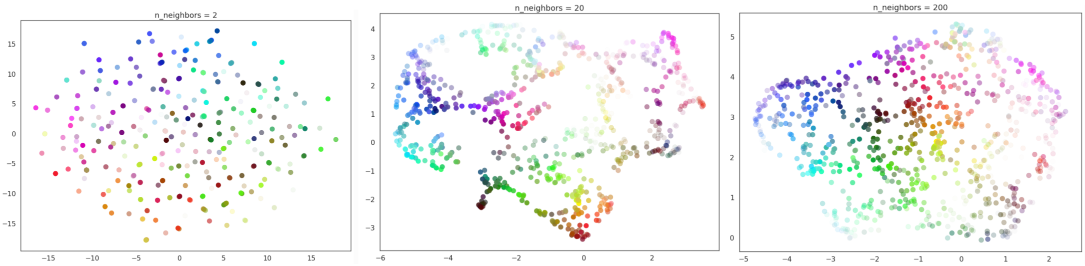


Figure 23: Effect of the number of neighbors considered by UMAP in the clustering of the output signals. From left to right, the number of neighbors considered are 2, 20 and 200.[12]

The minimum distance between output points establishes a compromise between how accurate the dimensionality reduction is and how visible are the results. For computing purposes the distance is set to 0, however, for visualizations such as those Figure 23 this number may be set higher.

The number of output dimensions is usually set to 2 or sometimes 3, so that visualizing the plots becomes possible. This number, however, can be set higher in case this could improve the classification performance of the algorithms.

Finally, the distance metric used will depend on the task at hand. The most widely used distance metric is the euclidean distance ($||\vec{u} - \vec{v}||_2$), however, the **cosine similarity** is many times used in the analysis of chemical spectra (e.g. [10, 31]). This metric, calculated through Equation 12, measures how similar the profiles from two spectroscopic signals are irrespectively of their magnitude. This measure will range from -1 (perfect inverse correlation) to +1 (perfect correlation). This allows for the creation of the "*cosine distance*", defined as 1 - S and therefore ranging from 0 (perfect correlation) to 2 (perfect inverse correlation).

$$S = \frac{\vec{u} \cdot \vec{v}}{||u|| \, ||v||} \tag{12}$$

Finally, there are also deep-learning approaches to dimensionality reduction. The first of these methods is the **autoencoder**, which is a neural network with the structure shown in Figure 24. The aim of this neural network is to compress the data to a lower dimension (encoder) and then reconstruct it into its original form (decoder) in order to ensure that the compressed data is representative of the original input. Once this is ensured, the decoder part can be discarded and the compressed data can be used for classification or visualization purposes. This approach was followed by Xie, Girshick, and Farhadi [65] and McConville et al. [38], who decided to use a number of dimensions of the compressed data equal to the desired number of clusters. They reported that autoencoders perform well in feature representation tasks, however, they do not preserve the distances between the original data in their learned representation.

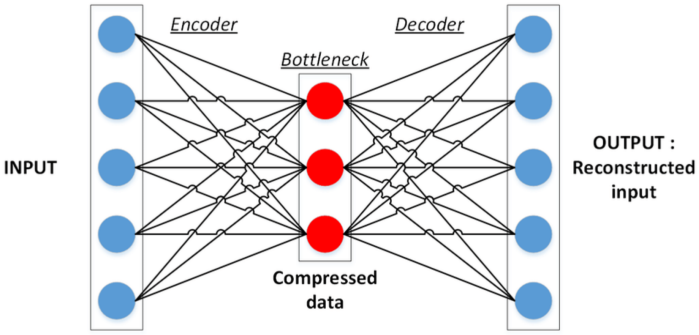


Figure 24: Schematic representation of the structure of an autoencoder.[13]

[12] Retrieved from `https://umap-learn.readthedocs.io/en/latest/parameters.html`

The main drawback of autoencoders and UMAP is that the dimensions learned in their new representation are not explainable. This means that changes in the input data cannot be traced to changes in the compressed data. This issue is addressed by **Self-Organizing Maps** (SOMs). SOMs are represented by a square or hexagonal grid of neurons, where each neuron has the same length as the input vector. The network is then trained on the input data in order to learn a representation of the data that preserves its topological structure. This can be readily seen in Figure 25, where the shapes of the studied galaxies are represented by the SOM and a clear transition between different groups can be observed. This dimensionality reduction technique therefore combines the use of the global structure characteristic of UMAP together with the explainability of PCA.

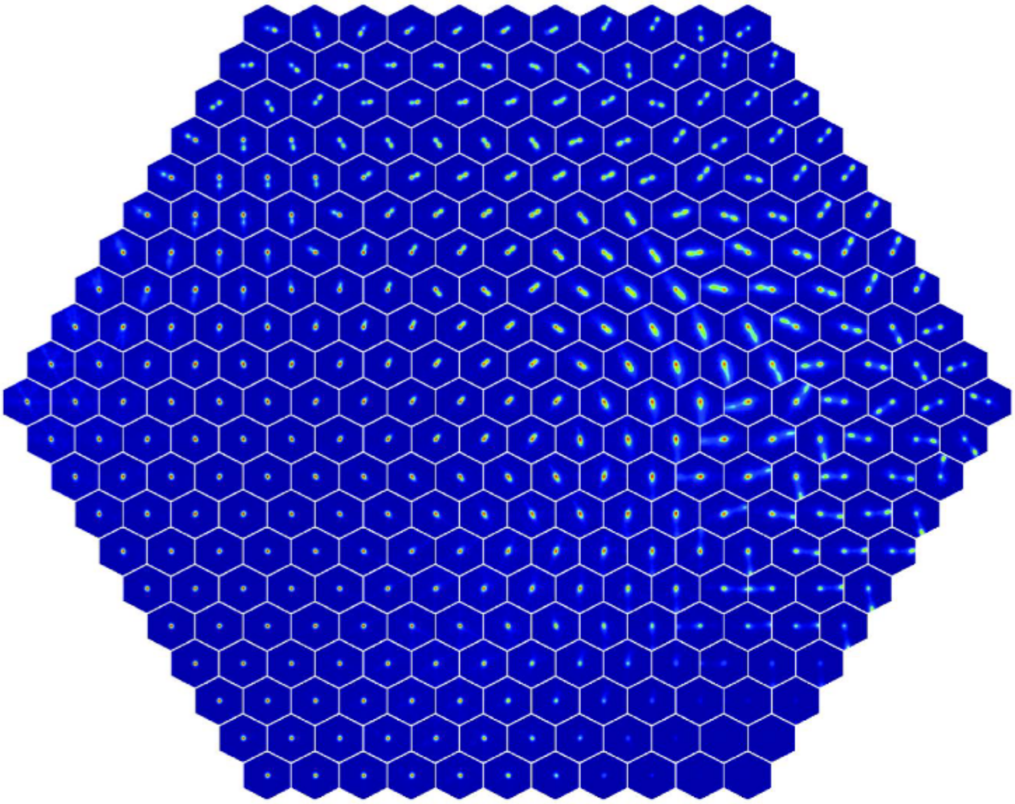


Figure 25: Self-organizing map representing the shapes of observed galaxies. Retrieved from Polsterer, Gieseke, and Igel [48].

Overall, there are several algorithms that can be used for dimensionality reduction; ranging from the linear PCA, to manifold learning techniques such as UMAP and neural network approaches such as autoencoders and SOMs. Each of the methods presents several benefits in terms of explainability and ease of visualization. Furthermore, not only have all methods been applied to different fields in astrophysics, but different studies obtained different results with them. This can be seen by comparing the use of PCA by Carey et al. [10] and Ishikawa and Gulick [30], where the former obtained better classification performance without applying PCA, as opposed to the latter who saw a slight increase in accuracy after reducing the dimensionality. It can therefore only be concluded that no single method will be most appropriate for all tasks. Furthermore, several methods can even be combined as shown by McConville et al. [38], who combined autoencoders and UMAP to obtain a better 2-dimensional representation of the space. Finally, there are many other DR algorithms that have been used in spectroscopy such as Independent Component Analysis (ICA) or Laplacian eigenmaps, both of these studied by Yang et al. [67]. There are even other widely accepted DR techniques such as Isomap ([57]) or t-SNE ([28]), however, the ones presented more deeply in this review are the ones which present the most promising results in the field of astrophysics and spectral analysis.

[13] Retrieved from `https://stackabuse.com/autoencoders-for-image-reconstruction-in-python-and-keras/`

0.4.2. ML models

ML models can be broadly divided into two separate fields: supervised and unsupervised learning. In a **supervised** setting, the data has an associated label to it and the models accuracy is measured based on its capacity to predict this label. In an **unsupervised** setting, however, the data has no label associated with it and the model attempts to group the data into "*clusters*" based on their similarities and differences. For instance, consider a model attempting to differentiate images of cats and dogs. In a supervised setting, the user would know which image corresponded to which animal and therefore the would be able to adapt its predictions based on the feedback provided by these labels. In an unsupervised setting, however, the model would compare the different images and split them into two separate categories depending on the different patterns that it would see repeating most often.

On one hand, supervised learning models tend to be more accurate since their capacity to get feedback for their predictions based on the data labels allows them to fine tune their outputs. On the other hand, unsupervised learning algorithms can be faster to implement since they do not need the user to manually input the labels and have been shown to make connections that humans had not thought of.

Since both of these methods have their benefits, this section will be split into two subsections, respectively supervised and unsupervised learning algorithms.

Supervised learning

Supervised learning has been the most explored option when analyzing chemical spectra of stars. These models can be divided into both ML models such as support vector machines (SVM), tree-based models or nearest neighbor algorithms, and deep learning models, namely neural networks.

Tree-based models are models who build upon basic decision trees. Decision trees are simple models who split the data into different groups based on a set of criteria from the input section. As seen in Figure 26, these trees have a root node at the beginning, leaf nodes at the end (which represent the final classification decision) and several decision nodes in between. Depending on how many sets of decision nodes there are, the depth of the tree will vary. The tree shown in Figure 26, for reference, is a tree of depth 1. Several decision trees can be combined in order to get a more robust output, leading to a random forest classifier.

The main advantage of tree-based methods is that the reasoning behind the classification choice can be readily understood by the user. These models have been used in the analysis of spectroscopic data by several authors such as Ishikawa and Gulick [30], who obtained a classification accuracy of ~85% for Raman spectroscopic signals with just the first 5 PCs. Similarly, Sevetlidis and Pavlidis [54] obtained a classification accuracy of 88.8% with this classifier after applying pre-processing with no DR.

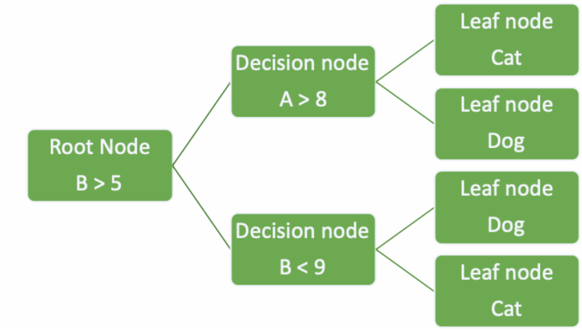


Figure 26: Schematic representation of the structure of a decision tree.[14]

Another very used and simple method for classification tasks is englobed by the family of **linear classifiers**. This classifiers simply attempt to define a linear discriminant within the input space (usually a polynomial whose weights are linearly calculated) in order to distinguish between two classes. Several linear discriminants can be applied together in order to differentiate between more classes. The main drawback of these discriminants, however, is that the data needs to be linearly separable in order for them to work. This can be fixed to a certain degree by adding margins to the data classification, however,

[14] Retrieved from `http://machinelearningintro.uwesterr.de/MlAlgoTrees.html`

the data still needs to be mostly separable for the algorithm to work. If this assumption holds true, however, the most commonly used type of model will be the Support Vector Machine (SVM) ([15]) which will attempt to maximize the distance between the different classes and the criterion that separates them. This is shown in Figure 27, where two classes (red and blue) are being separated by a linear discriminant (the red line) which maximizes the distance between the classes and the classification frontier.

Using this classifier, Bromová, Škoda, and Vážný [9] achieved a classification accuracy of 95.5% when applying it with no DR techniques, and 96% when using wavelet power spectra for feature extraction purposes. Similar results were obtained by Howley et al. [29], who obtained better results when classifying Raman spectroscopic data with a SVM after applying PCA.

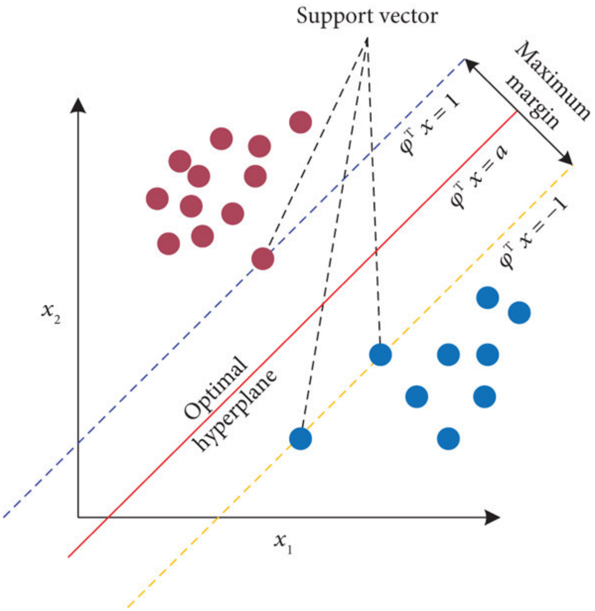


Figure 27: Schematic representation of a Support Vector Machine (SVM) using a linear divider to separate two classes.Retrieved from **qu_predictive_2022**.

Additionally, one may decide to use a **k-nearest neighbors** (k-NN) classifier ([16]) as opposed to the previous ones mentioned. The working of this classifier is also very simple: given a set of labeled training data, a new object will be assigned to the most popular class within its nearest *k* neighbors. An example can be seen in Figure 28, where the training data are the blue and red datapoints and the new object is the green circle. If the number of neighbors (*k*) is chosen to be 3, the new object will be classified as red, whereas if *k* is set to 5 the new object will be assigned to blue. Selecting an appropriate number of neighbors is therefore a non-trivial task. One way to mitigate this effect is to use a weighted neighbors (WN) classifier instead, were the class of the new data point will be also influenced by the distance towards its neighbors. In Figure 28, this could imply that the new data point would in both cases be red since both of their points are close and therefore would contribute more to the final class calculation.

A great example on the use of this classifier in spectroscopic data is outlined by Carey et al. [10]. In their work, they compare the performance of a k-NN, a WN classifier, a decision tree, and a multilayer perceptron (a simple multi-layer neural network) on the analysis of Raman spectroscopic data with different pre-processing steps. They conclude in their paper that the WN classifier outperforms all other algorithms, except on one particular case where it is equivalent to the 1-NN classifier. They also manage to get accuracies of up to 84.8% in the identification of particular species and up to 94.8% when identifying a mineral's class. Additionally, k-NN has also been employed in the field of astrophysics, as done by Pasquet-Itam and Pasquet [47] in their analysis of quasars. Depending on the variable to classify, they report the most accurate number of neighbors to vary between 2 and 6 and show a predictive accuracy of up to 73.72%. They do report that neural networks outperform this algorithm, however, they also found that combining both neural networks and k-NN achieved the highest performance of all with a classification accuracy of 80.43%.

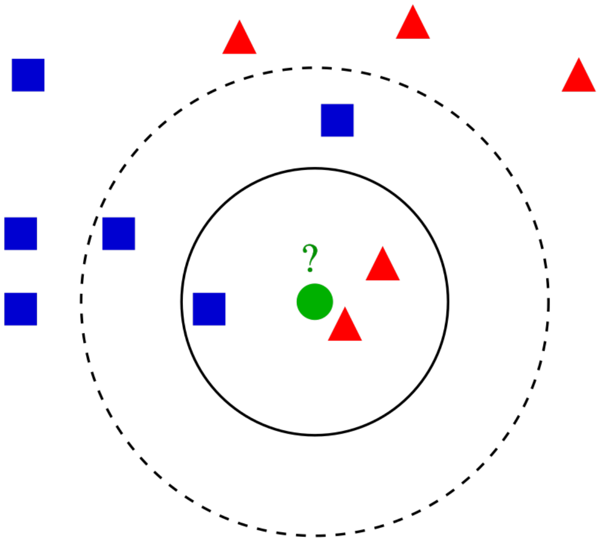


Figure 28: Representation of the k-NN's classification process for two classes. The solid smaller circle represents the first 3 neighbors and the broader dashed one represents 5 neighbors. [15]

Finally, the remaining algorithms to explore fall under the category of **neural networks**. The most simple type of network is the (multi-layer) perceptron which consists on just (several layers of) fully-connected neurons. Applying this architecture directly onto the original spectrum has shown very poor results, as evidenced by the classification accuracy of 35.8% in Carey et al. [10]. On the other hand, applying multi-layer perceptrons to the PCA decomposition of those signals was shown to achieve a classification accuracy of ~92% in Ishikawa and Gulick [30]. Additionally, one of the main advantage of these networks is that they are not only suited for classification purposes, but also for the prediction of properties in the observed target (regression). This was the case for Manteiga et al. [37], who utilized them to calculate T_{eff}, Fe/H, α/H and $\log g$ for several stellar spectra.

The most commonly used type of neural network used for the analysis of chemical spectra (and for any sort of signal analysis in general) are **convolutional neural networks** (CNNs). These networks are based on the presence of kernels (sometimes referred to as filters) which analyze small parts of the signal at each time. For instance, if one had an input vector with 5 channels (a 1×5 vector) and ran it through a CNN with a kernel size of 3 and a stride of 1, the kernel would first perform the dot product of itself and the first 3 values in the vector, then repeat the process for the second to fourth values and finally for the third to fifth values. Typically, and as shown in Figure 29, several convolutional layers will stacked together and connected in the end to one or more fully-connected layers that lead to the final output. This sequential analysis of the signal allows it to better identify patterns and particular waveforms which may appear at different points in the input but can be used to identify a given signal.

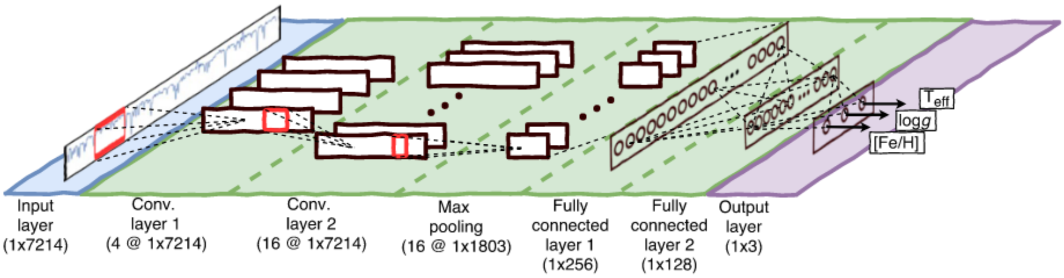

Teff
logg
[Fe/H]
Input layer (1x7214)
Conv. layer 1 (4 @ 1x7214)
Conv. layer 2 (16 @ 1x7214)
Max pooling (16 @ 1x1803)
Fully connected layer 1 (1x256)
Fully connected layer 2 (1x128)
Output layer (1x3)

Figure 29: Structure of a simple 1-dimensional convolutional neural network. Retrieved from Fabbro et al. [23].

These networks have shown outstanding performance in both classification and regression tasks.

[15] Retrieved from `https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm`

Regarding the former, Pasquet-Itam and Pasquet [47] showed a classification performance using simple CNNs of 79.32%, 5.6% higher than with k-NN. Similarly, Liu et al. [35] achieved an accuracy of 86.34% when classifying Raman spectral signals. Finally, Jahoda et al. [31] reported a classification performance of 89.31% by combining the output of 6 different CNNs. Furthermore, regression tasks also show very promising results. This can be seen in the work by Jahoda et al. [31] who obtained a mean absolute error of 0.053 and 0.025 in their calculations of log g and Fe/H respectively.

Overall, neural networks, and in particular CNNs, have been reported to yield the most accurate results. However, their complexity in designing and training, coupled with their intrinsic lack of explainability makes other algorithms with a slightly worse accuracy such as WN, random forests or SVMs compelling options.

Unsupervised learning

The main benefit of unsupervised learning is that it can be readily applied to surveys without the need of manually labeling the data. These models were traditionally only ML models, with the most common ones seen in literature being k-means, gaussian mixture models (GMMs) and hierarchical clustering. Recently, however, the field of deep clustering has come to exist, where neural networks are trained in an unsupervised manner for feature extraction and clustering purposes.

K-means clustering ([36]) is the most basic clustering algorithm but still used very often due to its ease of implementation and good results. The algorithm works as follows for k desired clusters, illustrated by Figure 30 for 3 clusters:

a) Initialize k centroids at random locations in the input space.
b) Assign each data point to its closest centroid.
c) Recalculate the centroid location based on the data points assigned to it.
d) Repeat steps b) and c) until convergence.

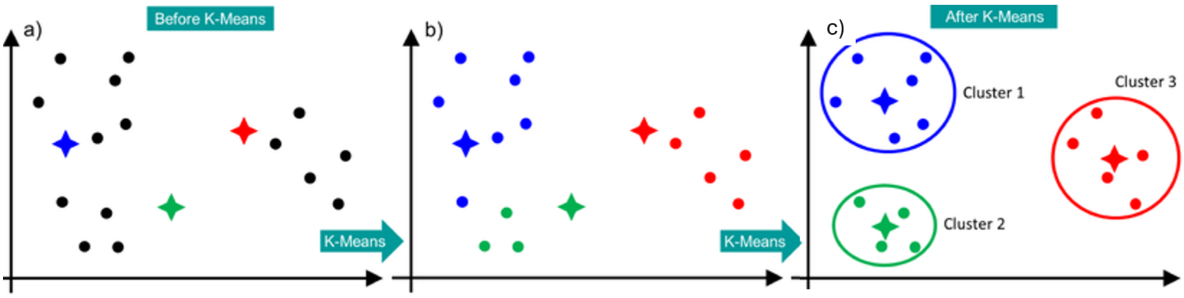


Figure 30: Evolution of the k-means classifier from initialization until convergence. Retrieved and adapted from Aguiar Nascimento et al. [1].

This algorithm has been used in many instances in the analysis of spectral signatures. For instance, Zhang [70] used it together with PCA to cluster 60 spectral signatures from star-forming regions. In their work, however, the number of clusters was not appropriately selected due to the limited amount of sources (typical ML projects have thousands of sources) and the imbalance between the cluster populations. A more sophisticated approach also in the field of astrophysics was used by Woods, Sainz Dalda, and De Pontieu [64], who looked at the clustering performance of their algorithm between 2 and ~400 clusters. The final number of clusters in their work, however, was still decided on the post-processing stage. This was because a larger number of clusters than their plots indicated would allow them to better explain the physical phenomena behind the observed processes. Overall, the only hyperparameter to optimize in the k-means algorithm is the number of clusters, which makes it a very easy to implement algorithm for many applications.

The main drawback of the k-means algorithm is that it will attempt to form circular clusters irrespectively of the data distribution. Therefore, if the data follows an alternative distribution, the k-means algorithm will be unable to capture the correct clusters. For this reason, mixture model algorithms were invented, the most known of which is the **Gaussian mixture model** (GMM) ([6]). This algorithm takes as a hyperparameter also the number of clusters and attempts to fit as many Gaussian distributions to the data. This can be seen in Figure 31 for an example with 3 classes, where the contour lines moving away from each cluster's center represent one standard deviation in a 2-dimensional Gaussian distribution. What is most interesting about these models, however, is that objects are not assigned to a given class, but rather have a probability of belonging to all of them. This is of special interest in points which are caught between two distributions such as the ones in the overlapping blue and green areas in Figure 31. This soft assignment (as opposed to the hard assignment, where a point has to belong to just one class), allows the user to better detect outliers and points which may have been misclassified.

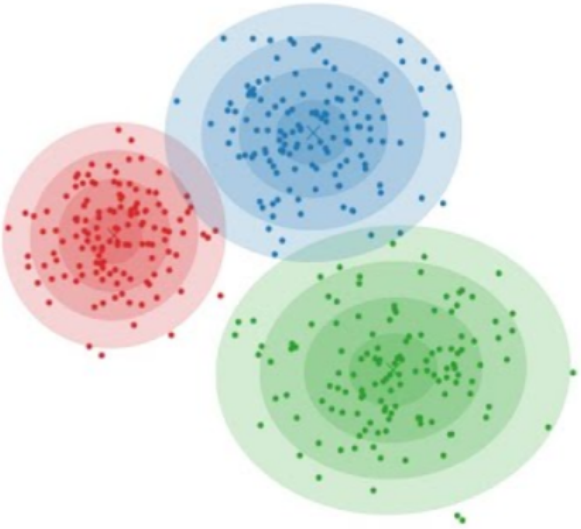


Figure 31: Example of a Gaussian mixture model clusterization with 3 classes.[16]

GMMs, as opposed to k-means, have not been widely adopted in the field of astrophysics. This is mostly because they scale very poorly with the dimensionality of the data, meaning that some DR technique is needed for them to be applicable. In Xu and McCord [66], UMAP was chosen as the DR algorithm in their analysis of gene expressions, as well as other commonly used test data sets such as MNIST. This was also the case for McConville et al. [38], who combined an autoencoder and UMAP before applying a GMM to the clusterization. Other DR algorithms can also work, however, UMAP seems to be preferable due to its ability to accurately map the input data to a 2-dimensional space where the GMM is easy to visualize and interpret.

In addition to these two methods, there is plenty of work being done using **hierarchical clustering** methods ([56]). The main advantage of these algorithms is that they do not require the user to know beforehand the number of clusters, but rather allows them to find it. The most common method found in literature is agglomerative clustering, where all data points start as their own cluster and clusters successively merge based on the distance between them. Other advantages of this algorithm are that it can accurately identify clusters with no particular shape or underlying distribution, and that it can be very easily adapted by the user. This, in turn, also means that it is more complex to implement and that there are more hyperparameters to optimize as opposed to k-means of GMMs.

A very nice review on the use of hierarchical clustering algorithms is given by Yu and Hou [69], although the referenced works mostly center around the identifying which stars belong to which star cluster rather than in the analysis of chemical spectra.

Finally, the remaining clustering methods belong to the **deep clustering** category. The first type of networks used for this purpose are networks which do not need apriori knowledge of the data labels such as autoencoders. This could be done with a structure such as the one reported in Xie, Girshick, and Farhadi [65], where the number of compressed dimensions equals the number of desired clusters and so applying a softmax[17] layer to this representation could lead to a soft label assignment. Another example could be that of SOMs where each node in the grid would correspond to a different cluster (see

[16] Retrieved from `https://blog.csdn.net/qq_44214428/article/details/140001995`

[17] Softmax is an activation layer added to some neural networks that makes its output's sum equal to 1, therefore acting as a probability density function.

Figure 25 for a visual example). More often than not, however, deep clustering methods involve the combination of a neural network and a classical clustering algorithm. For instance, both McConville et al. [38] and Xie, Girshick, and Farhadi [65] use classical clustering algorithms in the compressed dimension learned by the autoencoder in order to find clusters.

Many types of networks could be employed for this purpose, however, as CNNs performed best in the supervised setting for the analysis of chemical spectra, it is reasonable to assume that the same will hold for the unsupervised case. One very interesting example of an unsupervised clustering using CNNs is shown by Yang, Parikh, and Batra [68] who jointly utilizes CNNs and hierarchical clustering to cluster images. In order to validate their work, they use their architecture on an already labeled dataset, reaching an accuracy of 78.55% on the CIFAR-10[18] data set. They even compared their results to those obtained with a supervised CNN on a face-verification dataset for different sample sizes. In this experiment they achieved accuracies at most 0.9% lower than the supervised network, and even achieved a higher accuracy when training with 30,000 samples. Another example is the one reported by Caron et al. [11], who trained a CNN together with a k-means algorithm on the recognition of images. The downside of this combination is that the authors had to optimize the number of clusters used by the k-means algorithm, however, the implementation was simpler than the one laid out by Yang, Parikh, and Batra [68], which allowed them more time for testing. One of this testing stages also involved the comparison with a supervised CNN which had an accuracy difference of 14.4% in this case.

Overall, unsupervised learning algorithms are not as explored in the field of chemical spectral analysis of stars. Some classical algorithms have been explored in this context, however, the performance of supervised CNNs and analogue works on unsupervised deep clustering indicate that there is unexplored potential in this field.

0.4.3. Model evaluation

After the raw data has been pre-processed and fed to a ML model for its training, it is time to evaluate the model's performance. The metric used for this will depend on the model's task, which will primarily be split into classification (supervised) or clustering (unsupervised). Regression can also be a possible task of these models, however, most of the literature on the use of ML in astrophysics deals with the previous two. Therefore, those will be the focus of this section.

Classification evaluation

Evaluating a classifier is many times reported through the model's accuracy. This, however, does not paint the entire picture and can be confusing when there are more than 2 classes. In order to clear the latter case, a **confusion matrix** is usually built, such as the one shown in Figure 32. Despite its name, the underlying concept is very simple: the rows represent the true labels of the data whereas the columns represent the predicted labels. Ideally, only the main diagonal should be filled, however, mistakes can be made. In the figure, for example, one can observe that 6 objects whose true label was B were predicted to be C, which is the highest error rate in the example. This tool is therefore very helpful in analyzing which classes can the model better predict and which classes does the model confuse more often.

Similarly, some authors (e.g. [47]) also report their model's *recall* (or precision) together with its accuracy. A model's recall, shown in Equation 13, is a measure of how many false negatives there are, whereas the accuracy, shown in Equation 14, measures how many correctly classified objects there are. The distinction is often used to show that the model will not only be good at identifying positives, but also that it makes few mistakes.

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{13}$$

$$\text{precision} = \frac{\text{true positive} + \text{false negative}}{\text{number of samples}} \tag{14}$$

Combining both of these scores leads to the *F1 score* (see Equation 15), which is also many times used in supervised learning (e.g. [9]).

[18] https://www.cs.toronto.edu/~kriz/cifar.html

[19] Retrieved from https://www.vrogue.co/post/what-is-confusion-matrix-in-machine-learning

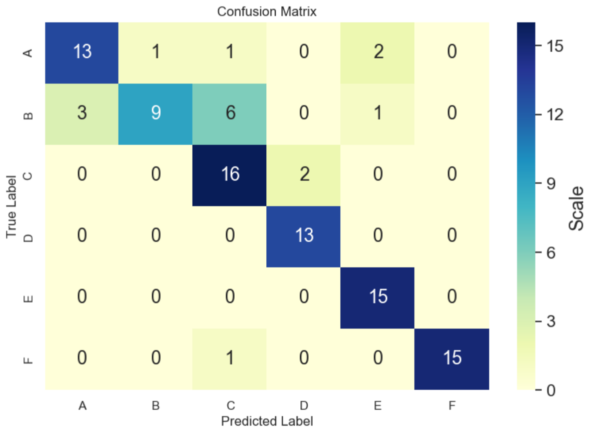


Figure 32: Example confusion matrix for a classifier with 6 different labels.[19]

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{15}$$

Clustering evaluation

When performing clustering tasks, the metric to be used will depend on whether or not the true label of the data is known, independently of whether or not it was used during the model's training. When the true label is known, the most popular metric is the **Normalized Mutual Information** (NMI). This metric measures how much correlation there is between the true and the predicted labels, ranging from 0 (no correlation) to 1 (all labels have been assigned to distinct clusters with no misclassified objects). Examples on the use of this metric can be found in McConville et al. [38] and Xie, Girshick, and Farhadi [65] among others.

On the other hand, the **silhouette score** ([51]) is the most popular metric when the true labels are not known. This metric, calculated as shown in Equation 16 (where C_I represents a given cluster, $|C_I|$ the number of samples in the cluster and $d(i, j)$ the distance between two samples), measures both the distance between clusters and the distances within clusters. The values of this metric range from -1 to +1, with negative numbers indicating misclassification of points, and numbers close to 1 indicating a good separation between clusters. Its popularity in the analysis of chemical spectra stems from the fact that different distance metrics can be used to calculate this parameter. This was the case in Carey et al. [10], who used the cosine distance to calculate the silhouette score of their clustering algorithm.

$$s = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \tag{16}$$

1 Introduction

High-mass stars greatly influence the way our galaxy evolves. They provide most of the galaxy's luminosity and greatly shape their surroundings both through their lifetime and at their extinction. Despite their importance, however, very little is known about their formation, and so much emphasis has been put into this research in the last decades. Nevertheless, it has not been until recently that advanced computer models and new instruments such as the Atacama Large (sub)Millimeter Array (ALMA) have allowed scientists to gather more data about these obscure regions. These new instruments, however, are also generating an unprecedented amount of data for which the current research tools are not ready.

This is the case of the ALMAGAL research consortium who, using ALMA, have observed 1017 high-mass clumps, totalling almost 6000 cores with a detectable emission [24]. For all cores, frequencies between 217 GHz and 221 GHz have been observed with a resolution of ~0.5 MHz [24]. This dataset has the potential to greatly advance our knowledge on high-mass star-formation processes, but manually reviewing all emissions could take years.

One possible solution is the introduction of machine learning (ML) models to assist with the study of the data. This new field of research is starting to gain popularity in similar fields, such as the study of Raman spectra ([10, 31, 35, 54]); however, not much research has been carried out on astrochemical spectra.

The aim of this research is therefore to study how can unsupervised ML techniques be used to study the emission patterns of high-mass star-forming regions. This will be done in two steps. Firstly, different models will be trained on the emissions from the ALMAGAL dataset and compared based on performance and complexity. The output from the best performing model will then be studied through the physical properties of the cores with similar spectra.

1.1. About the Data

The data gathered by the ALMAGAL consortium will be used for this research. This data consists of 1013 target fields observed, spanning evolutionary stages that range from IRDCs to HII regions [14]. Each field was observed through three different telescope arrays imaging the continuum emission at 219 GHz, achieving spatial resolutions ranging from ~1 000 AU to ~ 0.1 pc [52]. This high resolution allowed for the differentiation of up to ~6000 dense cores within those regions [43]. In addition to these spectra, the continuum emission of the area immediately surrounding each core was also recorded [52]. This signal, referred to as residual signal, was also provided for this research. For this MSc thesis, only 4396 sources were made available. These spectra, furthermore, are split into two different observational windows called spectral window 0 (spw0, [217-219 GHz]) and spectral window 1 (spw1, [219-221 GHz]) [52]. Note that the data is not continuous as there is a ~60 MHz gap between both spectral windows. More information on the survey and its processing can be found in [43, 52].

In addition to the chemical spectra, the following physical properties were provided [14]:

- Number of fragments

- Heliocentric distance
- Clump luminosity
- Clump mass
- Clump temperature
- Core temperature
- Core mass
- Core diameter
- Core volumetric density
- Core surface density

Some of the core's properties have only been estimated rather than calculated. For instance, the temperature of the cores was calculated based on the luminosity-to-mass (L/M) ratio of the clump [14]. This implies that all cores within a clump were assigned the same temperature, rather than calculating a core temperature based on chemical transitions in their spectra. This assumption is further propagated to the core's density and mass, which were calculated using the observed diameter and the assumed temperature [14].

1.2. Research Questions

Given the available data and the research gaps found in literature, the main research question is:

> **RQ:** To what extent can unsupervised ML be used to establish an evolutionary sequence of the star-formation process based on the spectral line content of the dense cores?

There are two main aspects to this research question: unsupervised ML (methodology) and the evolution of star-forming regions (results). It is therefore convenient to develop two groups of sub-questions. Firstly, the questions regarding the methodology are:

- **RQ-1.1:** How does pre-processing affect the clustering performance of the different models?
- **RQ-1.2:** Which models are best suited for the clustering of astrochemical spectra?

Each model's pre-processing steps and their output will be studied through these questions. Similarly, the evolutionary sequence of the star-formation process will be guided by the following research questions:

- **RQ-2.1:** Is there a correlation between the number of identified peaks in the spectral line content and the evolutionary stage of a core?
- **RQ-2.2:** Which molecular groups can be used to identify a dense core's evolutionary stage?
- **RQ-2.3:** How does the spectral line content of the cores correlate to that of the cloud they are embedded in?

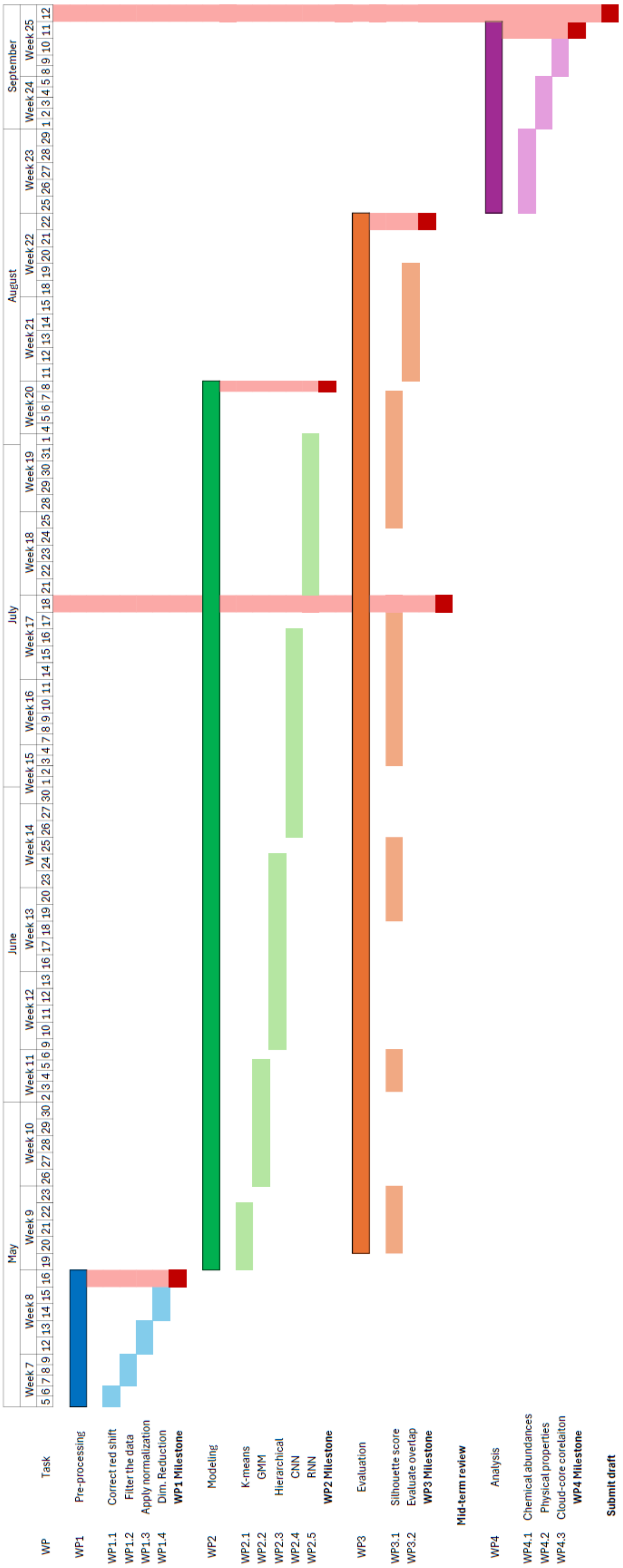
Chemical abundances within the spectra grouped by the models and the physical properties of the cores will be studied to answer these questions. Additionally, the similarities of cores within the same cloud will be studied to understand the entire region's evolution.

1.3. Structure and Planning

Following this introduction, the thesis will be structured as follows:

- Chapter 2 details all the pre-processing steps carried out during the investigation, including filtering of signals and red-shift correction.
- Chapter 3 details the different models tested and their outputs.
- Chapter 4 analyzes the physical properties of the clusters formed in the previous chapter and establishes an evolutionary sequence between the different cores.
- Chapter 5 explains the limitations of the study and guides future research on the topic.

The original planning for this thesis can be found in the next page. This planning was followed up until WP2.4, point at which the analysis of the clusters was started. This analysis, however, took longer than expected, leaving WP2.4 and WP2.5 for future work.


May
June
July
August
September
Week 7
Week 8
Week 9
Week 10
Week 11
Week 12
Week 13
Week 14
Week 15
Week 16
Week 17
Week 18
Week 19
Week 20
Week 21
Week 22
Week 23
Week 24
Week 25
5 6 7 8 9 12 13 14 15 16 19 20 21 22 23 26 27 28 29 30 2 3 4 5 6 9 10 11 12 13 16 17 18 19 20 23 24 25 26 27 30 1 2 3 4 7 8 9 10 11 14 15 16 17 18 21 22 23 24 25 28 29 30 31 1 4 5 6 7 8 11 12 13 14 15 18 19 20 21 22 25 26 27 28 29 1 2 3 4 5 8 9 10 11 12
WP
Task
WP1
Pre-processing
WP1.1
Correct red shift
WP1.2
Filter the data
WP1.3
Apply normalization
WP1.4
Dim. Reduction
WP1 Milestone
WP2
Modeling
WP2.1
K-means
WP2.2
GMM
WP2.3
Hierarchical
WP2.4
CNN
WP2.5
RNN
WP2 Milestone
WP3
Evaluation
WP3.1
Silhouette score
WP3.2
Evaluate overlap
WP3 Milestone
Mid-term review
WP4
Analysis
WP4.1
Chemical abundances
WP4.2
Physical properties
WP4.3
Cloud-core corelaiton
WP4 Milestone
Submit draft

2

Pre-processing: from raw data to model inputs

This chapter reviews all the pre-processing steps carried out prior to the training of the models. The data considered here is limited to the electromagnetic spectra from the 4396 sources observed in the ALMAGAL consortium [43]. The pre-processing steps described in this chapter therefore include data sampling, filtering, red-shift correction of the raw electromagnetic spectra. Different dimensionality reduction methods are also discussed, as they are necessary for some of the models used in this work.

2.1. Data re-sampling

The original data is split into two observational windows, referred to as spw0 and spw1 in this work. These windows cover the frequency ranges of 216.9-218.9 GHz and 219.0-221.0 GHz respectively. With a channel size of 0.4883 MHz, each window contains approximately 3800 channels. These windows, however, are not common to all observations, and neither are the frequencies around which they are centered. The models used in later chapters require all signals to have the same number of channels and for these channels to be aligned. The data was therefore re-sampled to ensure these requirements were met. To prevent the loss of data during the process, the number of channels per spectral window was increased to 10 000, with a channel width of 0.201 MHz. The resulting spectral windows were determined by the highest and lowest frequency across all signals. Channels for which a signal did not have information were assigned a value of zero. The resulting windows ranged as follows:

- spw0: 216 919.67 MHz - 218 929.75 MHz
- spw1: 218 991.72 MHz - 221 003.40 MHz

2.2. Data filtering

The radiation detected by the interferometer was contaminated by other radiation sources in the same section of the sky. This meant that not all the radiation associated with a source was actually produced by the source. In order to differentiate the signal from the surrounding noise, the radiation detected around the source (referred to as residual signal) was also recorded. An example can be seen in Figure 2.1, where the signal and residual from the 12th core of region 103421 are shown.

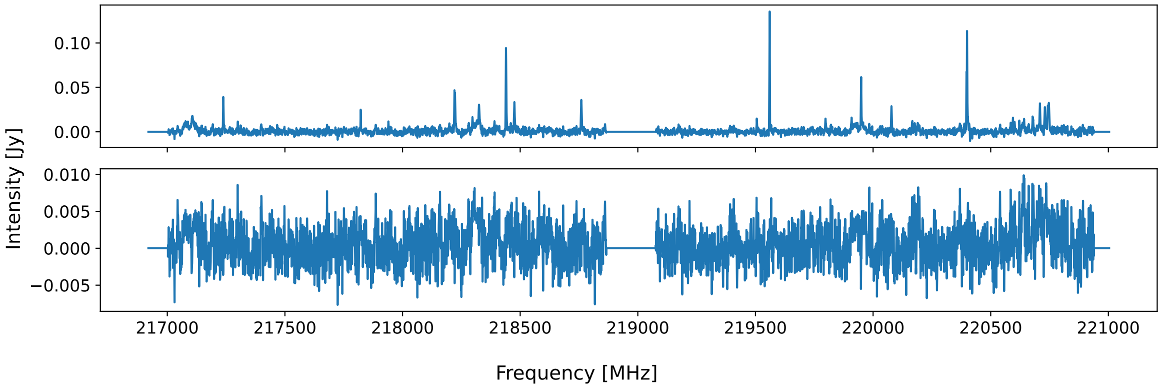


Figure 2.1: Detected signal intensity from the 12th core of the region 103421 (top) and its residual (bottom).

When studying the star-forming regions it is important to differentiate between the signal and its residual. The classical approach to analyse these signals is therefore to establish a threshold based on the standard deviation of the residual. Once the threshold is determined, all channels with an intensity lower than the threshold are suppressed, as seen in Figure 2.2. The blue line in the figure represents the original raw signal from the 1st core of region 554367, whereas the dashed red line represents the threshold of significance (in this case, three residual standard deviations) and the orange line represents the filtered signal.

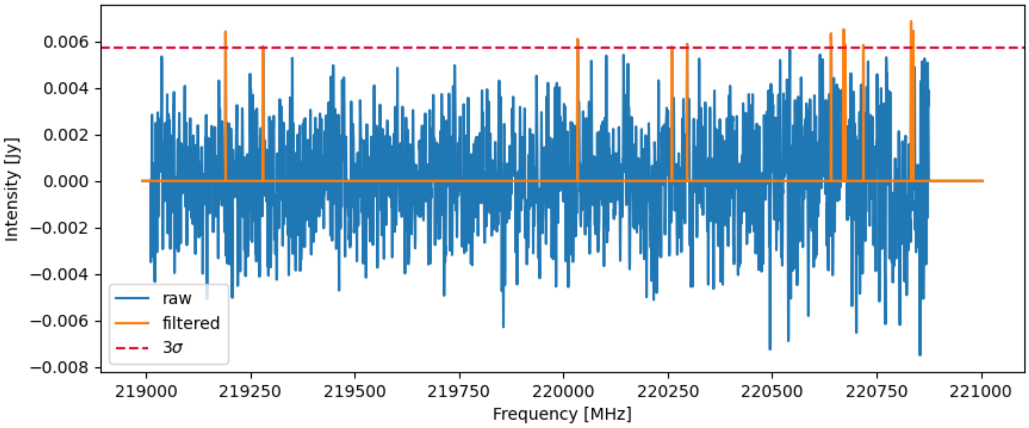


Figure 2.2: Raw (blue) and threshold-filtered (orange) signals from the 1st core of region 554367. The dashed red line represents the threshold of significance, set at three times the standard deviation of the residual. Only spw1 is shown.

This methodology ensures that most of the residual noise is filtered-out, however, it has two main drawbacks. Firstly, determining the threshold value is not trivial, specially for signals with a low signal-to-noise ratio (SNR). This is the case of the signal seen in Figure 2.2, where small changes to the threshold value will change the number of lines detected. Setting the value too high will imply the suppression of relevant data, and a value too low will result in residual noise being considered part of the signal. Secondly, this filtering method will always suppress absorption patterns which may hold relevant information about the signal.

A band-pass filter could potentially circumvent these limitations. This filter works by converting the signal and residual to the Fourier domain and suppressing the frequencies where the residual shows activity. However, this was not possible since, as seen in Figure 2.3, both signals share the same frequency bands.

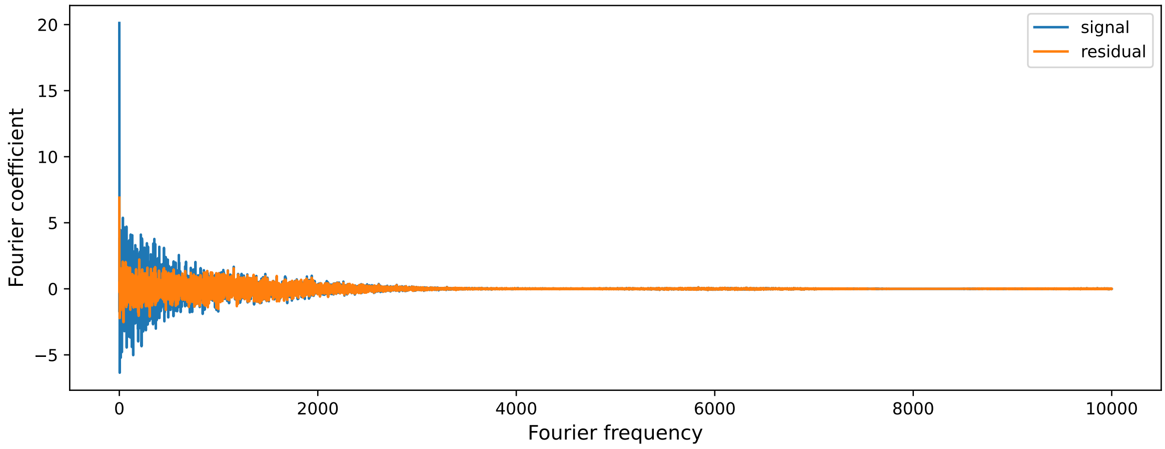


Figure 2.3: Absolute values of the Fourier coefficients of the signal (blue) and residual (orange) from the 12th core of region 103421.

This, however, prompted the idea of subtracting the residual from the signal itself. Doing so for the same core observed in Figure 2.3 leads to the results in Figure 2.4. In this figure, the blue signal represents the normalised raw signal and the orange signal the normalised result from subtracting the residual. Note that the signals were normalised by their maximum values for visualization purposes. For this particular core, it can be seen how the most significant peaks are all present in the filtered signal, as well as some of the smaller ones. The benefit of this new approach is that there is no longer a need to manually set a threshold, reducing the amount of effort needed to filter the signal. On the other hand, some very small peaks are still present near the baseline, which are most likely artifacts created by the filtering mechanism itself.

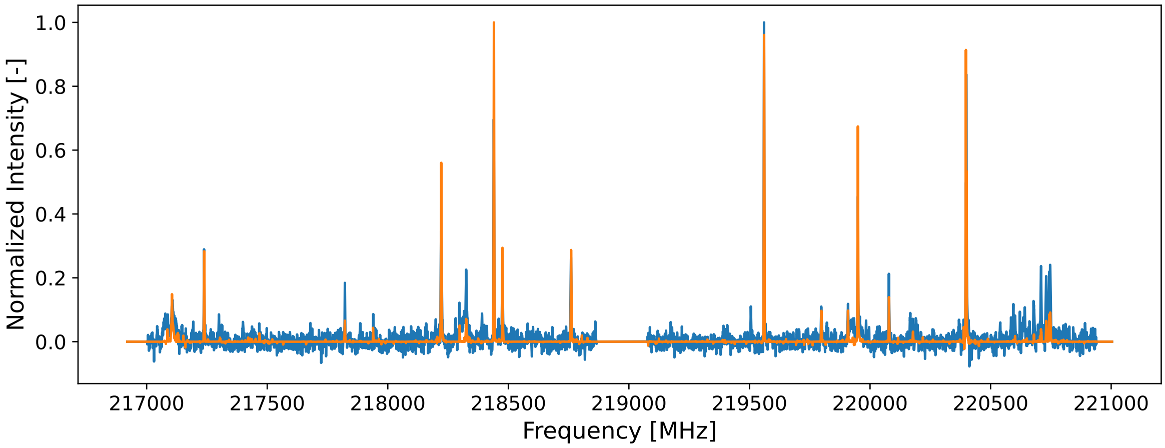


Figure 2.4: Superposition of the raw (blue) and subtraction-filtered (orange) signal intensities from the 12th core from region 103421.

This filter, although potentially more promising than the threshold-based filter, still has problems when dealing with signals with very low SNR. This can be seen in the top part of Figure 2.5, which corresponds to the first core from source 554367. The signal's SNR was calculated using Equation 2.1, where x_s represents the signal's data and x_r the residual's data. This particular signal has a SNR of 0.017 and, as seen in the figure, the subtraction-filter yields a single significant (positive) peak and many artifacts. An alternative method that did not depend on the residual could therefore outperform these for signals with low SNR. The Savitzky-Golay filter is an alternative that meets the criteria, and which was used in Carey et al. [10] and Ishikawa and Gulick [30]. This filter performs a polynomial regression to successive windows of observation within the signal, and so it does not rely on the availability of a residual signal [53]. The result can be seen in the bottom half of Figure 2.5, where the blue signal represents the raw signal, and the overlaid red signal represents the filtered signal using the same hyperparameters as Ishikawa and Gulick [30] (3rd degree polynomial and a 25-point window). The figure shows how an underlying pattern is unveiled within the noisy figure, however, it is unclear how much of the noise remains in the filtered signal.

$$SNR = 10 \log \left(\overline{x_s^2} / \overline{x_r^2} \right) \tag{2.1}$$

Originally, the Savitzky-Golay filter was used together with the raw, subtraction- and threshold-filtered signals to correct the red-shift. Using these four datasets yielded four different sets of velocities. When comparing these, it became apparent that the Savitzky-Golay filter did not yield good results. This was because, out of the 1172 occasions where three velocities matched and one did not, the Savitzky-Golay velocity was the odd one out 1145 (97.7%) times. This filter was therefore not considered further, and not included in the final methodology of section 2.3. The only filters considered for future work were therefore the *subtraction filter* and the *threshold filter*.

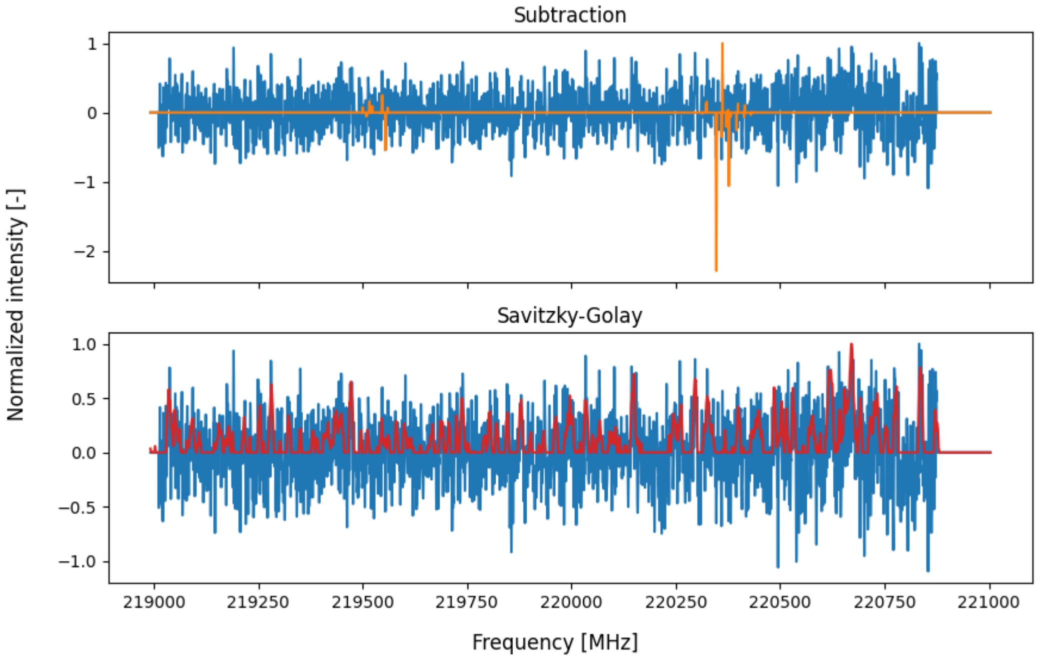


Figure 2.5: Comparison between the subtraction filter and the Savitzky-Golay filter for the 1st core of region 554367.

2.3. Red-shift correction

Due to the relative velocity between Earth and the star-forming regions, the perceived radiation detected from those points shifts due to the Doppler effect. This shift can be calculated using Equation 2.2 if the velocity of the source relative to Earth is known.

$$f = f_0 \left(1 - \frac{v_s}{c} \right) \tag{2.2}$$

Where:

- f is the observed frequency.
- f_0 is the original emission frequency.
- v_s is the velocity of the source with respect to Earth.

- c is the speed of light. Taken to be constant at $3 \cdot 10^5$ km/s.

v_s is unknown a priori. The standard approach is therefore to first identify a peak or a series of peaks from a known molecule. Thanks to laboratory experiments, the frequencies at which those peaks should appear are known, so the source's velocity can be found by solving for v_s in Equation 2.2. This approach is solid, however, it has a series of drawbacks. Firstly, as seen in Figure 2.5, not all signals have clearly identifiable peaks that one may use to calculate the velocity. Secondly, it can become a very tedious work when dealing with a large number of signals. This last point is of particular importance for this investigation as, despite the data having been taken in 2019 [59], only 3672 out of the 4396 cores have had their velocities calculated.

For these reasons, a new algorithm for calculating the velocity of the sources with respect to Earth was developed. The first step consists on checking the mean spectrum of the unshifted data, which is shown in blue in Figure 2.6. Based on the locations of highest activity and general knowledge on the relative abundance of molecules in dense cores, a series of peaks most likely to appear across the dataset can be identified. Five peaks were selected in this dataset, shown in Figure 2.6 as vertical lines. The frequencies of these peaks were retrieved from the Cologne Database for Molecular Spectroscopy [45], which reported the following values:

- H_2CO ($3_{0,3} - 2_{0,2}$): 218222.192 MHz
- H_2CO ($3_{2,2} - 2_{2,1}$): 218475.632 MHz
- $C^{18}O$ (J = 2 − 1): 219560.358 MHz
- SO ($6_5 - 5_4$): 219949.442 MHz
- ^{13}CO (J = 2 − 1): 220398.684 MHz

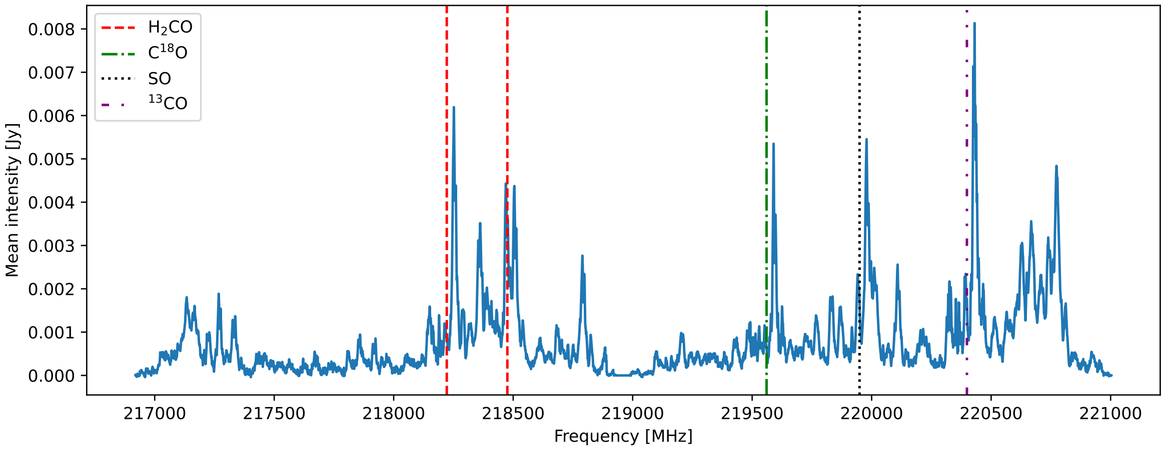


Figure 2.6: Mean unshifted spectrum from all cores in the dataset. The vertical lines show the molecules most likely to appear in most signals given this spectrum and the physical properties of these regions.

Once these peaks have been identified, a reference signal can be created using Gaussian distributions around those frequencies. Each signal is then shifted using every velocity from -250 km/s to +250 km/s at intervals of 1 km/s and compared to the reference signal. This comparison was done using the cosine similarity (see Equation 2.3) as done in similar studies [10, 31].

$$S = \frac{\vec{u} \cdot \vec{v}}{|u||v|} \tag{2.3}$$

Once the velocity with the highest similarity has been identified, a finer search can be done. This new search tested velocities on a range of ±1 km/s around the previously-found velocity at intervals of 0.01 km/s. The overall algorithm can be seen in Algorithm 1.

Running this algorithm can take some time, so a recursive bisection algorithm for the fine search phase was also tested. This algorithm is faster, however, it runs the risk of getting stuck in a local maximum.

Therefore, the observation window (in this case ±1 km/s) must be small enough for there to only be one maximum within. The new algorithm, seen in Algorithm 2, is slightly faster and provides a much more precise velocity than Algorithm 1. Despite this, the time performance improvement is marginal, which makes this algorithm less suited for this dataset when coupled with the risk of getting stuck at a local maximum.

The use of Algorithm 1 automates the correction, and so it is suitable for handling large datasets. Signals with a low SNR, however, are still hard to correct. This is because the peaks created by the background noise are comparable in magnitude to those of the signal itself, leading the algorithm to identify a velocity that aligns the noise peaks with the reference signal instead of the actual signal peaks. To address this issue, Algorithm 1 was also applied to the threshold-filtered and subtraction-filtered signals. Using a majority-voting scheme, this allowed the algorithm to identify the velocities of sources for which the raw signal had too much noise. An example of this is the 1st core of region 100309, shown in Figure 2.7. As seen in the figure, the velocity identified through the raw signal is -187.58 km/s, however, both filtered signals yield a velocity of 112.30 km/s.

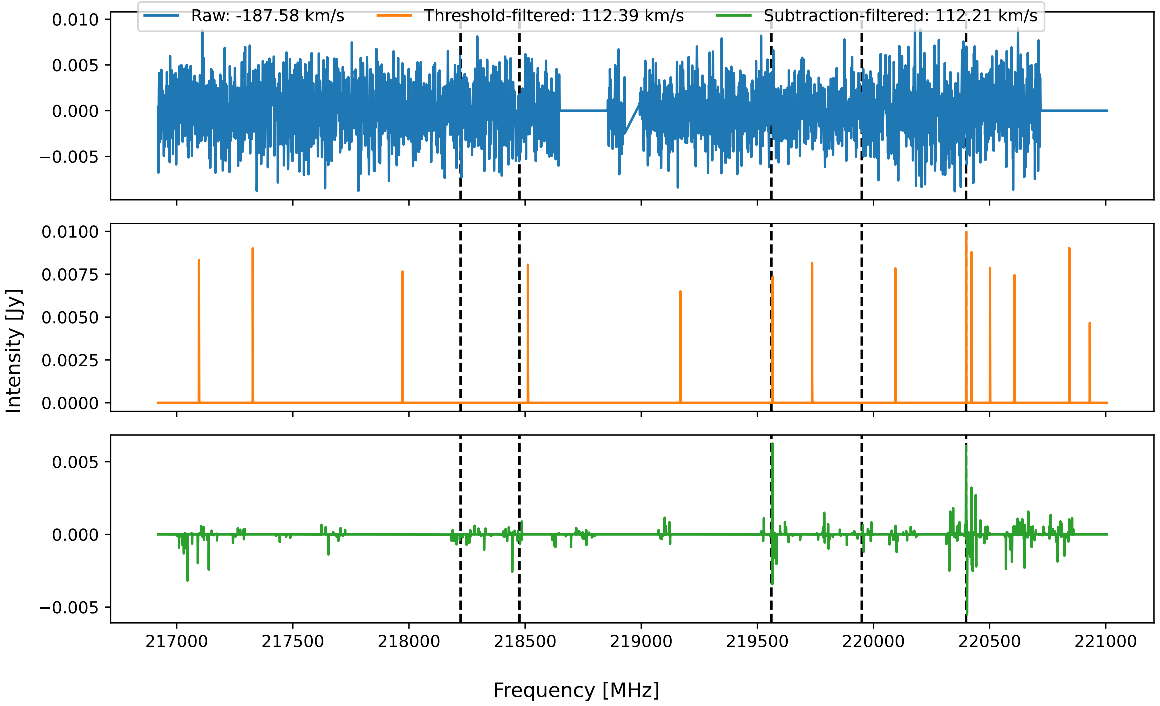


Figure 2.7: Raw (top, blue), threshold-filtered (middle, orange) and subtraction-filtered (bottom, green) signals from the 1st core of region 100309.

Using this scheme, the velocities identified using the three signal types mostly coincided. However, some signals yielded a velocity with a magnitude greater than 200 km/s. It is unlikely to encounter an object with such a velocity relative to Earth, so these signals were manually inspected. Doing so revealed that many such signals had strong absorption features near the ^{13}CO and C^{18}O peaks, which decreased the overall cosine similarity between the signal and the reference. This was the case of the signal shown in Figure 2.8. In light of this effect, all negative values were suppressed before using Algorithm 1, resulting in a better alignment, as seen in the bottom half of Figure 2.8.

In the end, the velocity of 4322/4396 (98.32%) signals was identified through this methodology. To test the accuracy of these values, the velocity of the 3672 cores with a manually-identified velocity (provided by the ALMAGAL consortium) was compared to the one calculated using this procedure. The results can be seen in Figure 2.9, which shows how both methods identified the same velocity for most signals. Specifically, out of 3672 signals, only 8 differ by more than 15 km/s and only 31 by more than 5 km/s.

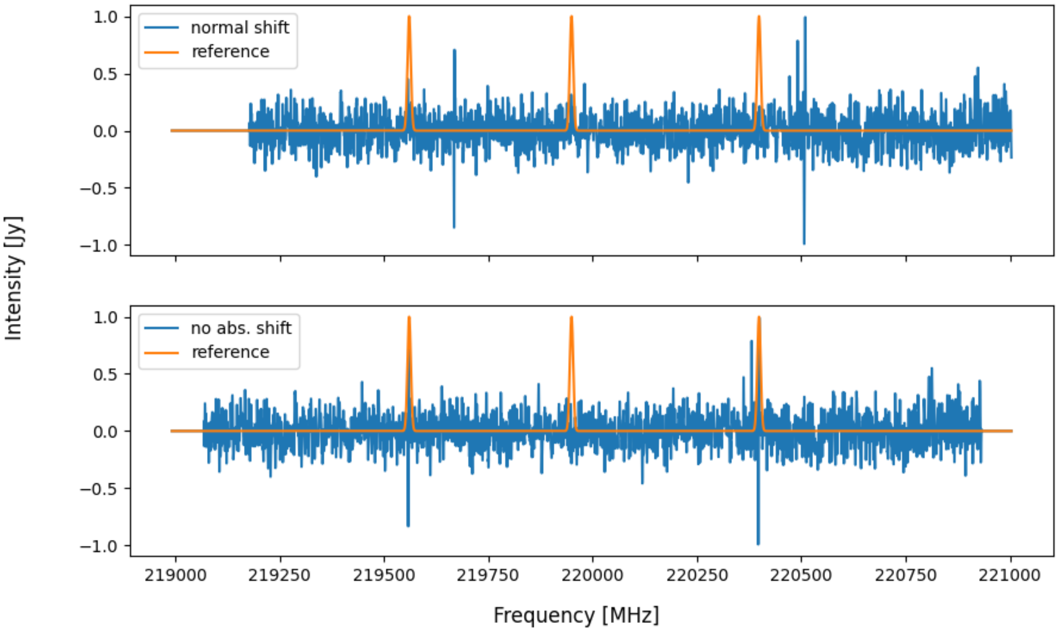


Figure 2.8: Shifted signals with (top) and without (bottom) considering the absorption features near the ^{13}CO and C^{18}O peaks. The orange signal in both figures represents the reference signal used.

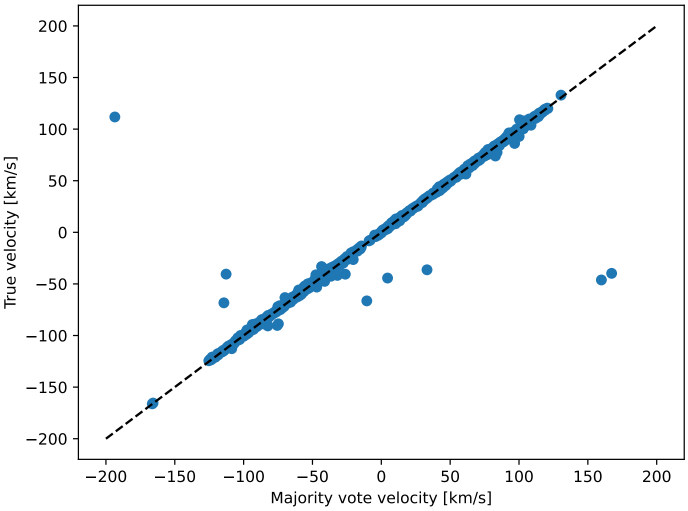


Figure 2.9: Visual comparison between the velocities identified through the procedure described in this section (Majority vote velocity) and the manually-identified velocities (True velocity).

Finally, the signals with a difference greater than 15 km/s where manually inspected, revealing that, in almost all instances, the manual velocity showed a better alignment to the reference peaks. The only exception to this rule would be the 7th core from source 644237, which can be seen in Figure 2.10. In the figure, it is clear how the signal shifted according to the procedure described here shows a better alignment with both H_2CO lines, the SO and the ^{13}CO line. Overall, this trend shows that the manually identified velocities are marginally more reliable, however, it also serves as validation for Algorithm 1.

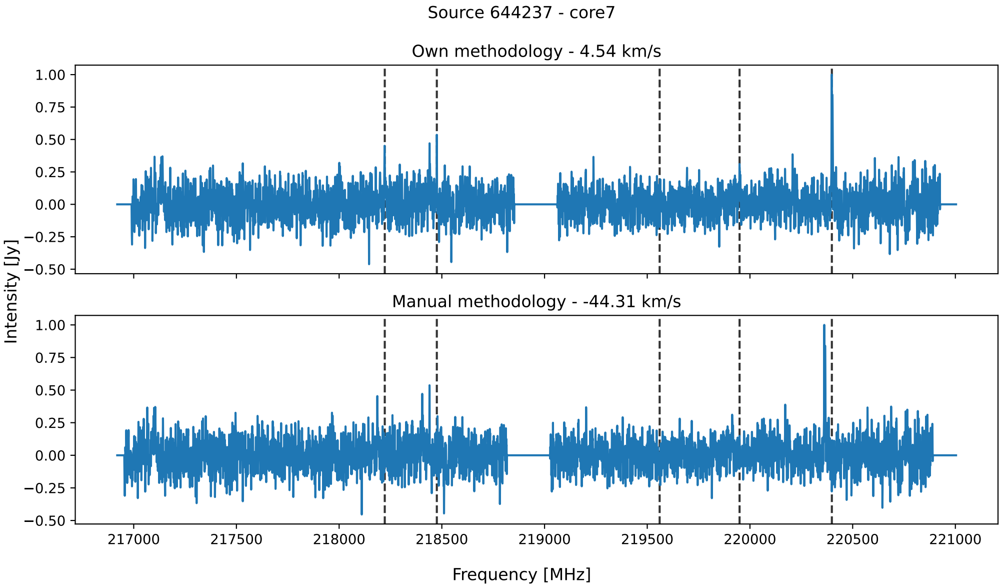


Figure 2.10: Comparison of the alignment of the signal from the 1st core from region 644237 to the reference lines when shifted according to this section's procedure and a manual alignment.

Despite this observed trend, however, the automatic velocity assignment showed a greater alignment than the manual velocities. This can be observed in Figure 2.11, which shows the mean spectrum of all shifted signals using both velocities. As seen in the figure, the spectrum that uses the automatic velocities reaches a higher maximum height and has fewer peaks around the main peaks, both of which indicate a better alignment.

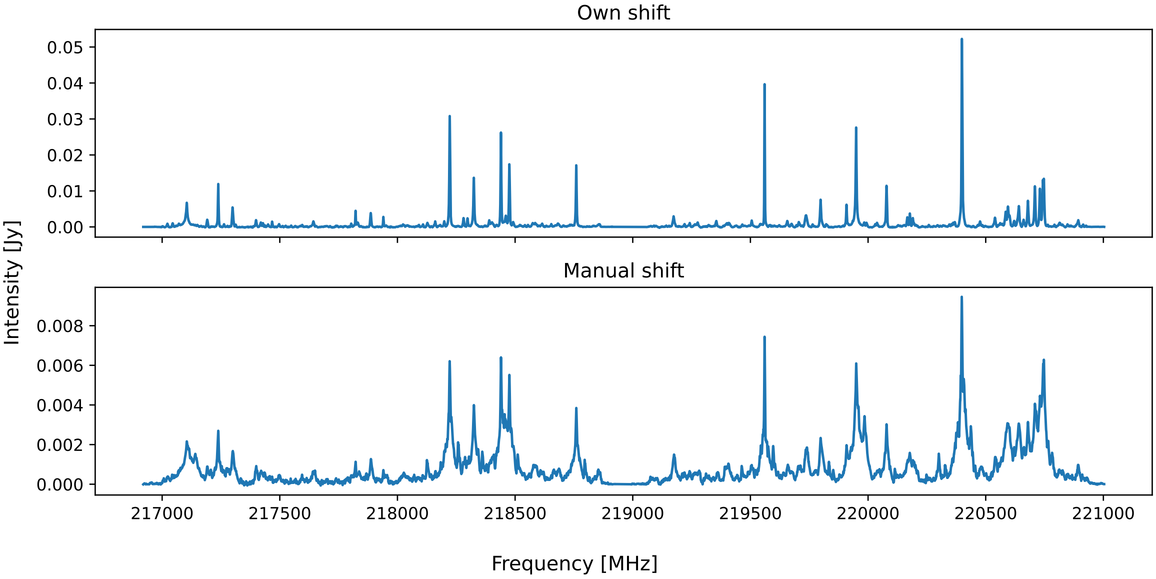


Figure 2.11: Mean spectra for all signals with an assigned velocity, calculated using the velocities from the method described in this section (top) and the manually-identified ones (bottom).

One explanation for the better alignment of the automatically-calculated velocities could be the use of a re-sampled dataset. The manually-calculated velocities use the original dataset for their calculations, which can lead to minor changes in velocity. This is made even more likely when taking into account that 99.16% of all velocities coincided within less than 5 km/s between both method. Despite the reason, given the spectra in Figure 2.11, the automatically-calculated velocities were used in the work that followed. Furthermore, the high percentage of signals with a common velocity across the methods serves as validation for Algorithm 1, ensuring its reliability for signals without a manually-identified velocity.

2.4. Dimensionality reduction

Dimensionality reduction methods are techniques through which the same information can be expressed in a lower dimensional space. For this particular case, the objective was to reduce the length of each vector expressing the detected signal. There are two groups of methods that serve this purpose: linear and non-linear methods.

The most commonly used linear method is Principal Component Analysis (PCA) [10, 30, 31, 67, 70]. The method works by identifying underlying structures in the data (called principal components or PCs) common to as many signals as possible and expressing each signals as a linear combination of these components. The first components represent the underlying structures most common across all signals, and so they will be significantly insensitive to random noise. This can be observed in Figure 2.12, where the first five components contain a similar amount of variance, independently of whether the dataset is filtered. Once a certain number of components is reached, however, each additional component will be more governed by random noise and so their individual explained variance ratio will decrease. This is important since 193 components are needed to explain 95% of the variance in the raw dataset, whereas only 28 and 15 are needed when using the threshold- and subtraction-filtered respectively.

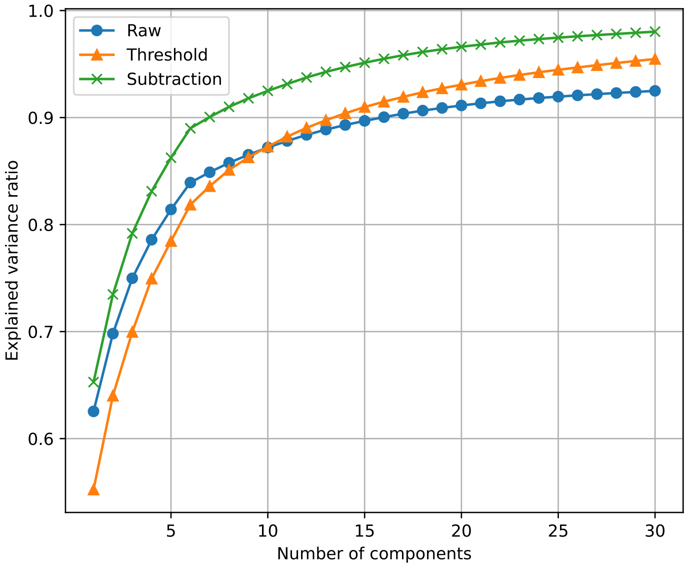


Figure 2.12: Cumulative variance per principal component.

Non-linear methods, on the other hand, are not as standardised as linear ones. Several options such as Laplacian Eigenmaps [4] (used by Yang et al. [67] to analyse MRI spectra), t-SNE or Isomap. These last two in particular are more standard depending on the application, since the former focuses on local distances within different signals (distances around a particular frequency range) whereas the latter focuses on the global distance between entire signals. Both of these have benefits, leading to both McConville et al. [38], and Xu and McCord [66] utilising UMAP [39], which combines both local and global distance measures. Using 5 neighbors when transforming the data yields Figure 2.13. The top of the figure shows the actual UMAP projection for the raw and filtered datasets. Since there is much overlap, the density distributions are plotted in the bottom panels. This overlap in the projection shows how there are no clearly-defined groups. The density distributions, however, do show that the data points cluster around particular areas. This implies that several signal archetypes exist, where each archetype represents a particular spectral pattern. These archetypes could be linked to different evolutionary stages of high-mass star-forming regions, such as hot and cold cores. Additionally, the data points between them could be transition stages. Assuming this to be true, it could imply that there are stages where the cloud's evolution slows down and transition stages where the cloud evolves faster. This, however, is still too early to be determined at this stage.

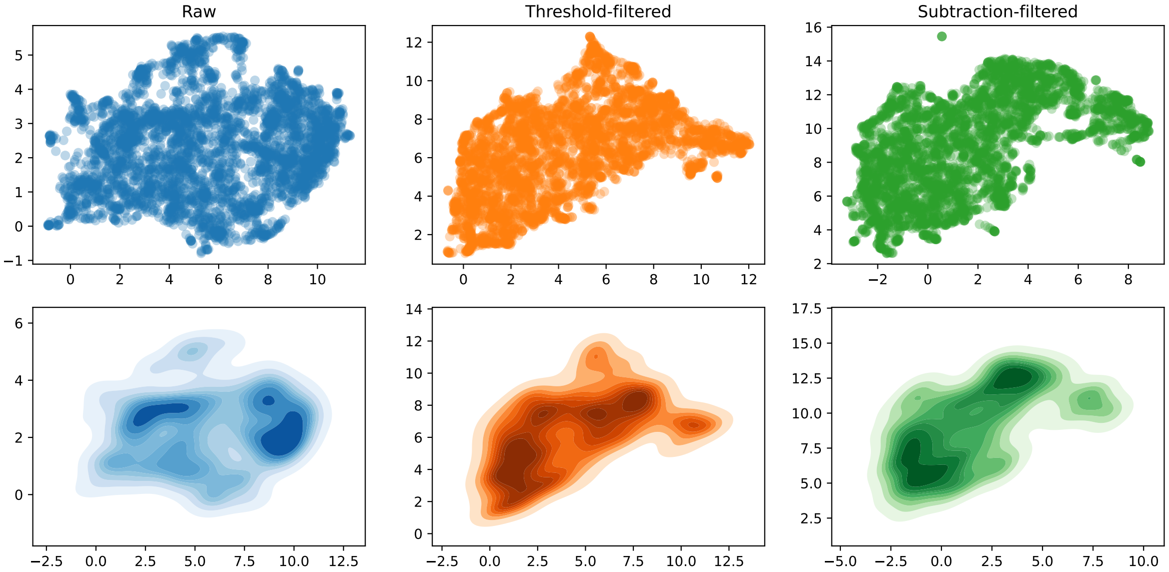


Figure 2.13: UMAP projection for the filtered and un-filtered datasets with the number of neighbors set to 5.

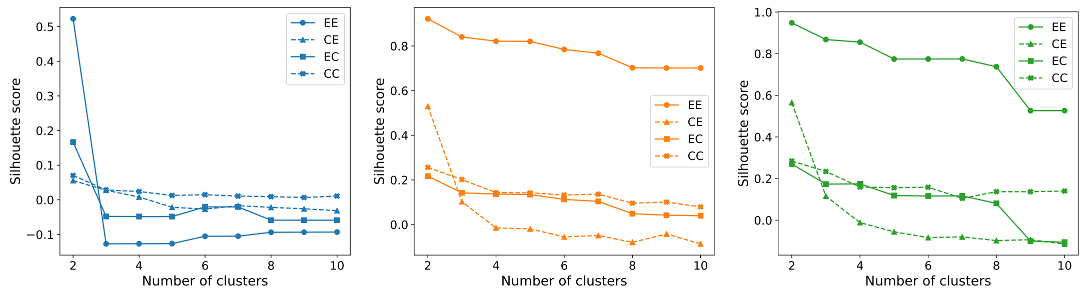
3 Clustering

This chapter describes the different unsupervised ML models tested on the different datasets and their performance. This performance will be limited to the quality of the clusters formed and will not take into account the physical properties of the signals that form the clusters. Therefore, only how similar the signals are to other members of their cluster and how different they are from the rest will matter, and not other properties such as the temperature of the cores. Readers unfamiliar with distance and evaluation metrics used in unsupervised ML can find a detailed description of the ones used in this study in Appendix A.

3.1. K-mean clustering

K-means clustering is the simplest means of clustering. It has been widely used in literature for this reason, often used as a baseline or with the objective of figuring out how the dataset responds to several pre-processing methods [1, 9, 64, 67, 70]. This was also the case during this research, as the first thing that was tested was the most appropriate distance metric. The two methods evaluated were the Euclidean distance, which evaluates the difference in intensity in every channel equally; and the cosine distance, which places more importance on the more intense peaks. Cosine distance is the more promising metric between the two described in Appendix A according to Carey et al. [10]; however, filtering mechanisms could neglect the effect of random noise on the signals. This, coupled with the fact that the Euclidean distance is the standard built-in distance metric for most clustering algorithms, makes it worth it to test which one works best. The performance of both distance metrics was first evaluated by comparing the silhouette score of models using each metric, as seen in Figure 3.1. This evaluation metric measures how similar signals are to their own cluster compared to other clusters, with a score closer to +1 indicating well-defined clusters and a score close to -1 indicating misclassification (see Appendix A for more details). The silhouette score also requires a distance metric, so four lines are present in each sub-figure: one for each combination of modeling and evaluation distance metrics.

For all datasets, the silhouette score almost continually decreases when adding additional clusters up to a point where the metric stabilizes. Filtering greatly improves the asymptotic score of the clusterings. This effect is greatly enhanced when computing the silhouette score using the Euclidean distance metric. This effect can be observed by comparing the scores of the model trained using the Euclidean distance, represented by solid lines in Figure 3.1. Both lines (EE and EC) show the exact same outline, but the one computed using the Euclidean distance is significantly higher, proving that the silhouette score is biased towards this distance metric. Despite this bias, the model using the Euclidean distance metric obtained a higher silhouette score than the one using the cosine distance when computing the silhouette score using the same distance metric. This is contrary to what was found in literature, therefore, the differences between both clusterings were investigated. There are no maxima in the silhouette score of the cosine-distance model beyond 2 clusters, however, an increase in the metric is seen when reaching 8 clusters and computing the score using the cosine distance. Since this number of clusters is also the one that precedes a large drop in score in the Euclidean distance model, this number of clusters was selected for further investigation. The centroids of both models were then plotted, as seen in



(a) Raw signal **(b)** Threshold-filtered signal **(c)** Subtraction-filtered signal

Figure 3.1: Silhouette score of the k-means classifier using 2-9 clusters. Each subplot uses a different filtering algorithm and the lines correspond to the distance metric used in the classification and in the evaluation (EE - Euclidean in both, EC - Euclidean classification and cosine evaluation, CE - cosine classification and Euclidean evaluation and CC - cosine classification and cosine evaluation).

Figure 3.2 and Figure 3.3. Inspecting these figures revealed some differences between the behaviour of the models, despite the centroids being fairly similar overall. Firstly, the cosine distance classifier grouped together all signals with many spectral lines into a single cluster (see the second cluster of Figure 3.3). These signals were split into two different clusters when using the Euclidean distance metric, namely the second and fifth clusters of Figure 3.2. Furthermore, the total number of signals is lower in the Euclidean classifier, with both clusters adding up to only 183 clusters. The remaining signals were most likely clustered in the third cluster of Figure 3.2, where the features at the end of spw1 are present but secondary. This indicates that the Euclidean distance places an emphasis on both the number of

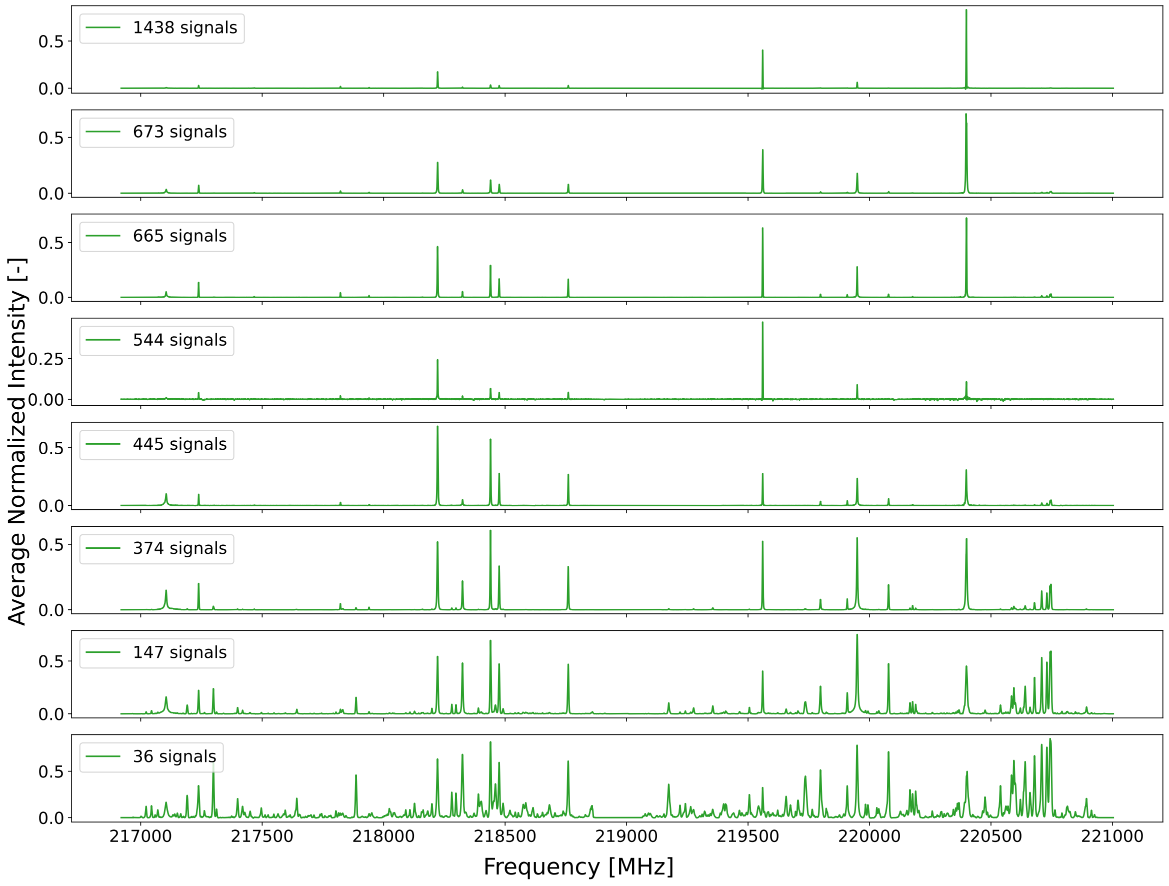


Figure 3.2: Centroids from the Euclidean distance classifier.

peaks and their intensity, whereas the cosine distance classifier disregards the intensity of these peaks and clusters signals based on the number of features. Additionally, the cosine distance seems to also place more emphasis on absorption features, as they are visible in several of the centroids of Figure 3.3 but not on those of Figure 3.2.

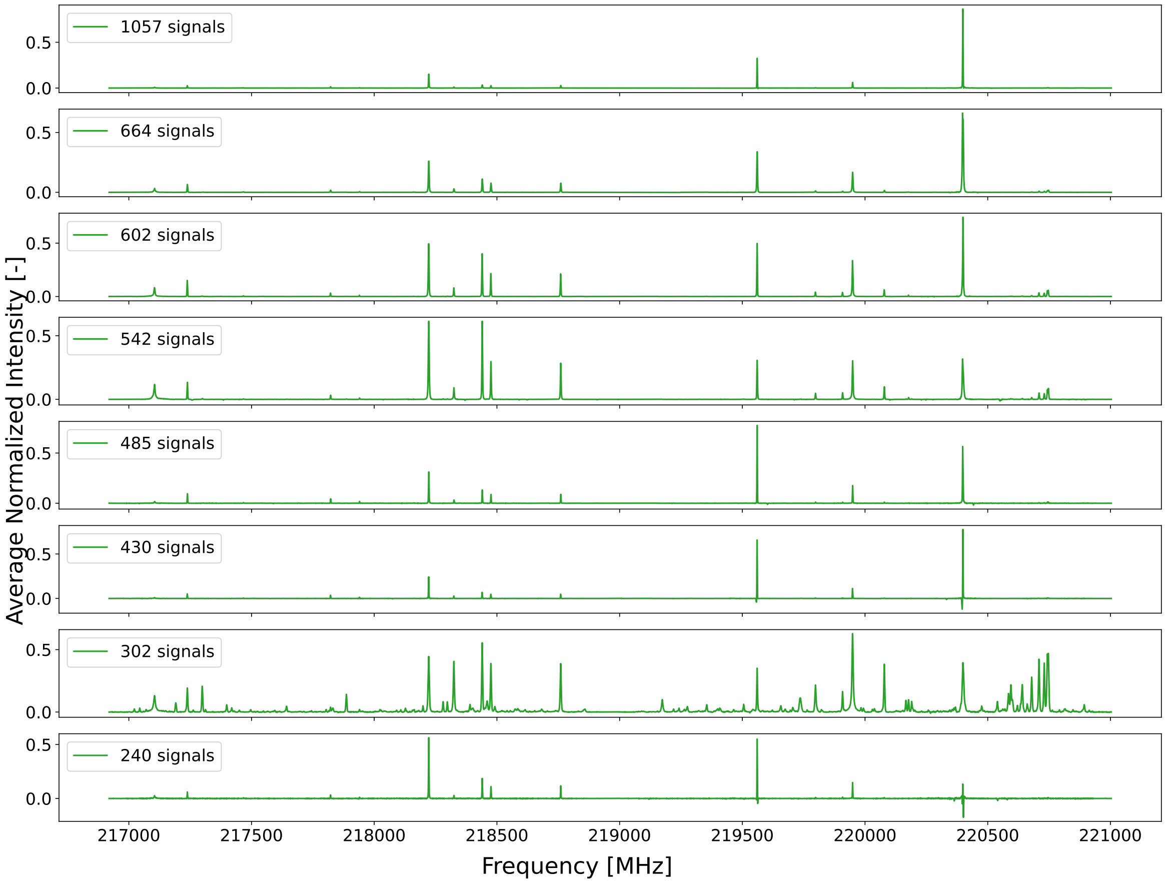
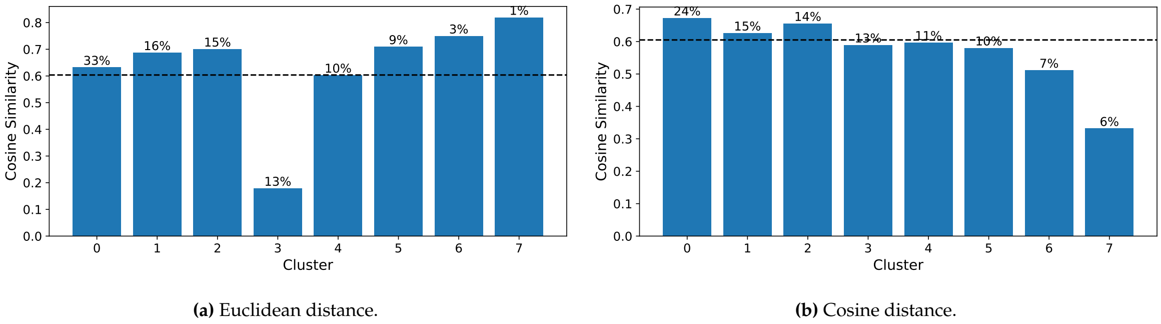


Figure 3.3: Centroids from the cosine distance classifier.

Next, the within-cluster similarity (WCS) of both clusterings was calculated. This metric is computed by averaging the cosine similarity of the signals within each cluster, as proposed by Carey et al. [10] (see Appendix A for more details). The results can be seen in Figure 3.4, where the percentage on top of each bar represents the percentage of the dataset contained in the cluster, and where the dashed horizontal line represents the size-averaged WCS. Both arrangements have a WCS slightly higher than 0.6, however, this mean is skewed in the Euclidean-distance classifier by cluster 3, which contains 13% of the data and has a WCS of just 0.179.



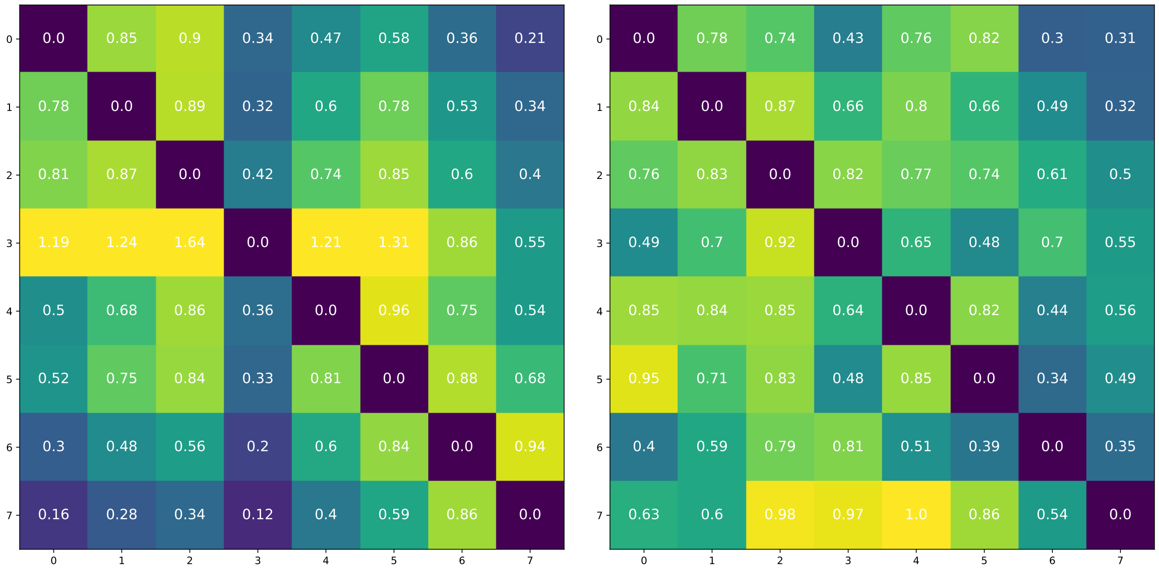
(a) Euclidean distance.

(b) Cosine distance.

Figure 3.4: Within clusters similarity of both label assignments (Euclidean- and cosine-distance) using the max-normalized subtraction-filtered signals.

Finally, the between-cluster similarity (BCS) was calculated. This metric is computed by averaging the cosine similarity between the signals of one cluster and all the others. This values can then be divided by the WCS of one cluster, leading to a non-symmetric matrix where each cell represents the similarity of the signals from a given row to those of a given column, relative to the similarity of the signals that make up the row's cluster. Ideally, this BCS/WCS ratio should be as low as possible, indicating that the similarity of the signals within a cluster is much higher than the similarity of those signals to those of the other clusters. The results can be seen in Figure 3.5. As expected, the low WCS of cluster 3 of the Euclidean-distance classifier leads to very high BCS values in that row. The values in that row, however, are much lower, indicating that the similarity between cluster 3 and the other clusters is much lower than the WCS of the other clusters. In addition to this, it can also be observed that the best differentiated clusters in Figure 3.5a are the last two clusters, who only have a BCS higher than 0.6 between them. Beyond these two observations, however, nor many more patterns can be observe as the values in the lower triangle of the first 6 rows are very similar to those in the upper triangle. This is far from being the case in Figure 3.5b, where the values in the lower triangle are consistently higher than those in the upper one. This, however, was also to be expected, as the WCS almost continually decreases in Figure 3.4b. Overall, the mean BCS/WCS ratio of each model was calculated as shown in Equation 3.1, which yielded a value of 0.3749 for the Euclidean-distance model and 0.4044 for the cosine-distance model, including all clusters in both cases. This indicates that, on average, the Euclidean distance classifier will form clusters slightly better defined than the cosine-distance one.

$$\mu = \text{mean}(\text{BCS}/\text{WCS} \cdot \text{BCS}/\text{WCS}^T) \tag{3.1}$$



(a) Euclidean distance.

(b) Cosine distance.

Figure 3.5: BCS/WCS ratio of both label assignments (Euclidean- and cosine-distance) using the max-normalized subtraction-filtered signals.

Determining the most appropriate distance metric is not straight-forward. The silhouette score is biased towards the Euclidean distance metric and the WCS of both arrangements was almost identical. Inspecting the clusters formed using the max-normalized subtraction-filtered dataset revealed that the Euclidean-distance classifier performs best when clustering signals with many peaks. The cosine-distance classifier, on the other hand, is better at differentiating the relative magnitude of the most intense peaks while still separating the signals with many lines from those with few. Finally, analyzing the BCS/WCS ratio of both models revealed that many of them have values higher than 0.6, which is the size-averaged WCS of both models. This indicates that there is likely some overlap between the clusters, however, not cluster in either arrangement had a BCS higher than 1 both in its row and its column, indicating that, in all instances, at least one of the clusters was more similar with itself than with the other.

3.1.1. Masking major peaks

Observing the clusters formed in the previous section, it is clear that most clusters are dominated by the same peaks. An attempt was therefore made at masking these peaks to see if the clustering algorithm could identify more subtle features in the signals. The suppressed areas covered a 50 channel window around each of the peaks listed in section 2.3, as shown in grey in Figure 3.6.

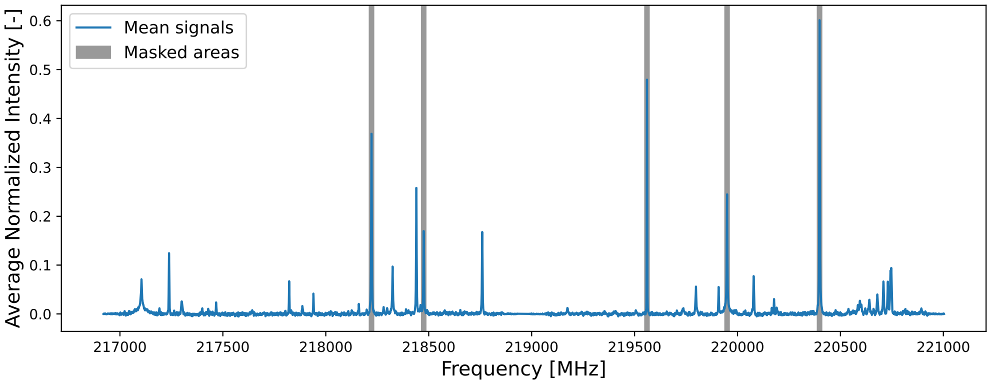
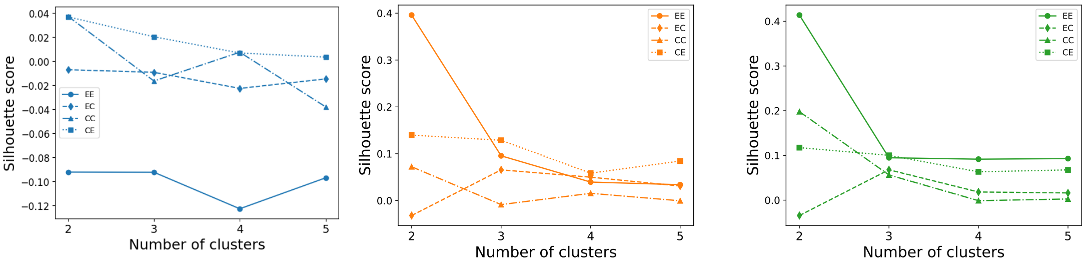


Figure 3.6: Mean signal and masked windows.

The results, seen in Figure 3.7, show that the clustering performance is greatly decreased overall. This effect is further enhanced with the raw signals, whose maximum silhouette score is 10 times smaller than that of the filtered signals.



(a) Raw signal **(b)** Threshold-filtered signal **(c)** Subtraction-filtered signal

Figure 3.7: Silhouette score of the k-means classifier using 2-5 clusters. Each subplot uses a different filtering algorithm and the lines correspond to the distance metric used in the classification and in the evaluation (EE - Euclidean in both, EC - Euclidean classification and cosine evaluation, CE - cosine classification and Euclidean evaluation and CC - cosine classification and cosine evaluation).

The centroids when the number of clusters is set to 2 were investigated using both distance metrics on the subtraction-filtered dataset. When using a Euclidean distance metric (see Figure 3.8), the signals are split into a small group with an active spw1, consisting mostly of a series of peaks at the end of the signal that can be matched to CH_3CN, and a larger group which is mostly active only in spw0. Similarly, using a cosine distance yields a similar result with a larger group of signals where spw1 is active. These signals have a worse alignment, however, as indicated by the reduced maximum intensity.

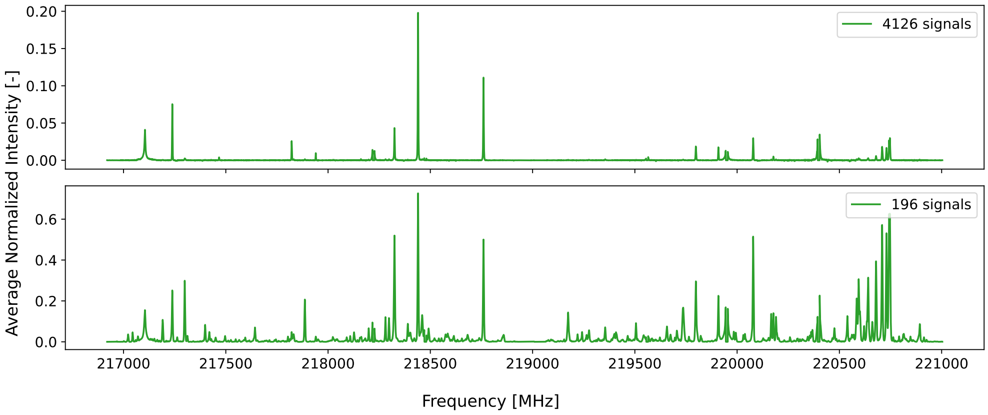


Figure 3.8: Masked centroids using a Euclidean distance and normalized subtraction-filtered signals.

Once again, it was proven that the Euclidean distance will identify smaller groups with more features, whereas the cosine distance will identify more balanced groups where the differences lie in the distribution of the main peaks rather than on the number of features. Beyond this, however, it can only be stated that this algorithm was not suited to do a deeper clustering of the masked signals.

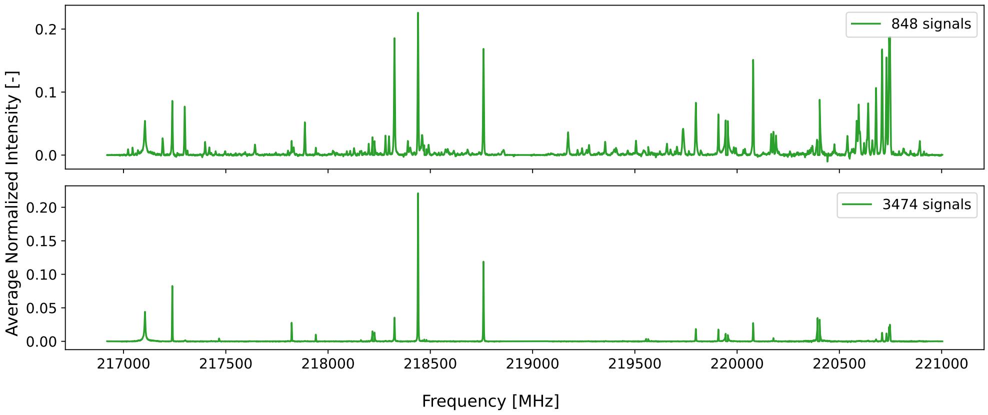
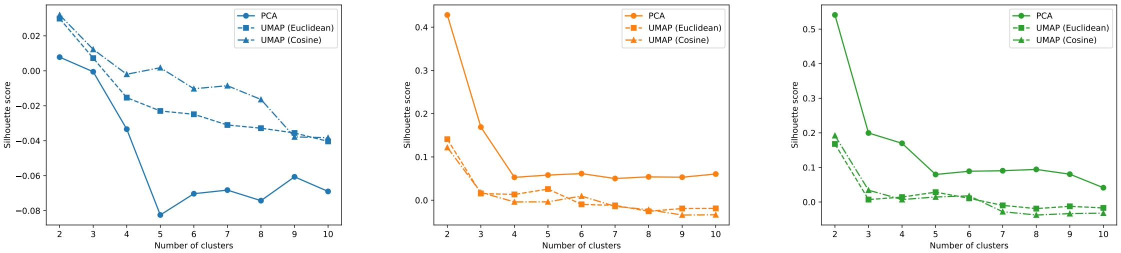


Figure 3.9: Masked centroids using a cosine distance and normalized subtraction-filtered signals.

3.1.2. Dimensionality reduction

The training of the models can be made much faster if the dimensions of the dataset are reduced. Using PCA and UMAP (the latter using both the Euclidean and cosine distances) to train the models and the underlying signals to calculate the silhouette score resulted in Figure 3.10. As seen in the figure, using PCA on the filtered datasets is the most promising clustering scheme, however, the clustering is limited to two clusters and the score is still worse than when using the complete signals. Note also that the UMAP projection used for this section used the cosine-distance metric to calculate the neighboring signals.



(a) Raw signals.

(b) Threshold-filtered signals.

(c) Subtraction-filtered signals.

Figure 3.10: Silhouette score, calculated using the Euclidean distance metric, of the of the k-means classifier using PCA (30 dimensions) and UMAP on differently-filtered datasets.

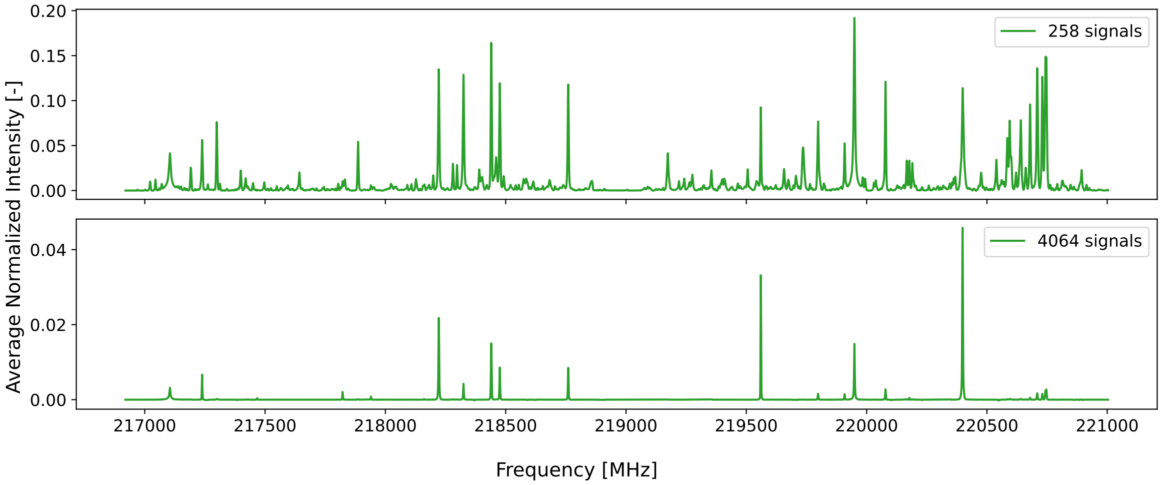


Figure 3.11: Cluster centroids obtained when training a k-means classifier with 2 clusters on PCA data with 30 dimensions.

3.1.3. Conclusion

The k-means classifier managed to identify 8 clusters with a size-averaged WCS of 0.6, and where in no instance the similarity between two clusters was greater than the similarity within both clusters. This was only possible thanks to the use of filtering mechanisms, as the performance of the classifier when using the raw dataset was much poorer. The centroids of the clusters formed using the subtraction-filtered dataset were then analysed. Many of the centroids resembled one another which, added to the numerous instances where the BCS/WCS ratio was greater than 0.6, suggests that some overlap exists between the different clusters.

Additionally, this clustering algorithm made it possible to understand the behavior of the different distance metrics. The Euclidean distance metric made it possible to discern signals with many peaks from those with a moderate amount and those with few significant features. As seen in Figure 3.2, the most common type of signal was one with virtually no other features beyond H_2CN, ^{13}CO, $C^{18}O$ and SO. These signals dominate the first 5 clusters although differences in the relative intensity of each peak can be discerned. Specifically, these signals tend to have a more intense spw1 than spw0. Following this signal, the second most common signal type is one with the same features as the previous, however, this signal's spw0 and spw1 were of comparable magnitudes. Finally, the final two groups exhibit a much greater amount of spectral lines, with the one having the most amount of features consisting of just 36 signals. Regarding the quality of the clusters, the WCS of 77% of the signals is greater than the average one. This is due to the extremely low WCS of cluster 3 which contains 13% of all signals and has a WCS just under 0.18. Without this cluster, the average WCS would have been 0.663, although only 87% of the dataset would have been clustered.

The same WCS was achieved using a cosine-distance metric, however, all clusters had a comparable similarity, with only 55% of the signals having a higher-than-average value. As seen in Figure 3.3, the centroids found using this metric are very similar with a few exceptions. Firstly, all signals with a high number of features were clustered into the same cluster whereas with the Euclidean distance metric they had been split into two. Secondly, absorption features are present in two of the centroids of Figure 3.3, which was not the case for the Euclidean-distance model. This highlights how the cosine-distance metric places more emphasis on the presence of peaks and absorption features, rather than on their relative magnitudes.

These observations make it possible to state that the Euclidean distance metric is better for differentiating groups with a large amount of features, whereas the cosine distance is best suited for analysing the overall composition of the signal. Overall, however, the Euclidean-distance classifier yielded better results. This is because, despite having a similar WCS, the BCS/WCS ratio of the Euclidean-distance classifier was 0.3749, compared to the 0.4044 of the cosine-distance one.

An attempt was then made to cluster the signals while suppressing the main peaks that dominate the dataset. This proved to be futile, however, as only two different clusters were identified. When using the Euclidean distance metric, these clusters had 4126 and 196 signals respectively, with the latter group having many spectral lines and the former just a few and mostly in spw0. Using the cosine distance, on the other hand, yielded more balanced clusters (848 and 3474 signals), although the composition of the centroids was still the same.

Finally, this classifier was tested using reduced datasets. These datasets were obtained applying PCA (using 30 dimensions) and UMAP to the raw and filtered datasets. Figure 3.10 shows that the PCA-reduced dataset achieved a higher silhouette score than both UMAP datasets. However, only 2 clusters could be obtained using this dataset. Furthermore, the sizes of the clusters were completely imbalanced (4275 and 47 signals in each), discouraging any further attempts at clustering the reduced dataset with this classifier.

Table 3.1: Summary of the k-means classifier's research.

Aspect	Finding	Comment
Distance metric	Ambiguous	Euclidean distance was found best for signals with many features, whereas the cosine distance performed better at differentiating the relative intensity of the main peaks.
Filtering method	Subtraction filter	This filter yielded the best silhouette scores and WCS. The threshold filter is comparable, and both are far better than the raw signal.
Masking major peaks	Not useful	Masking the main peaks greatly decreased the silhouette score and yielded just two clusters, providing no additional insights.
Dimensionality reduction	Not useful	PCA with 30 dimensions yielded a slightly better silhouette score than UMAP, however, only 2 clusters could be obtained.
Final clustering	8 clusters	The model found 8 differentiable clusters when using either distance metric, in both cases with an average WCS of 0.6. The Euclidean-classifier, however, could increase this similarity by discarding its fourth-largest cluster. This would result in 87% of the data being clustered with a WCS of 0.663.

3.2. Gaussian Mixture Models

Gaussian Mixture Models (GMMs) work by calculating the covariance of all channels across all signals to identify a given number of Gaussian distributions present in the data. The main issue with this model is that such a covariance matrix is impossible to calculate when the input signals have 20 000 channels. This is because doing so would imply calculating the covariance for 400 million points, which would take a normal computer far too long to compute for it to be acceptable. It is therefore necessary to use dimensionality reduction methods before training the model.

Similarly to the k-means classifier, the number of clusters is a hyperparameter that needs to be determined. Unlike k-means, however, this model uses a soft-labelling assignment. This means that the model assigns to each signal a probability of belonging to each cluster, rather than assigning a single label. This simplifies detecting signals transitioning between two groups, as they will have a lower maximum probability of belonging to any one label.

3.2.1. PCA

This classifier was first tested with the PCA-reduced dataset. Given the higher performance of the subtraction-filtered signals in section 3.1, only that dataset was used. Once more, the first step consisted on finding the most optimal number of clusters. This was also done through the silhouette score (calculated using the Euclidean distance metric), shown in Figure 3.12.

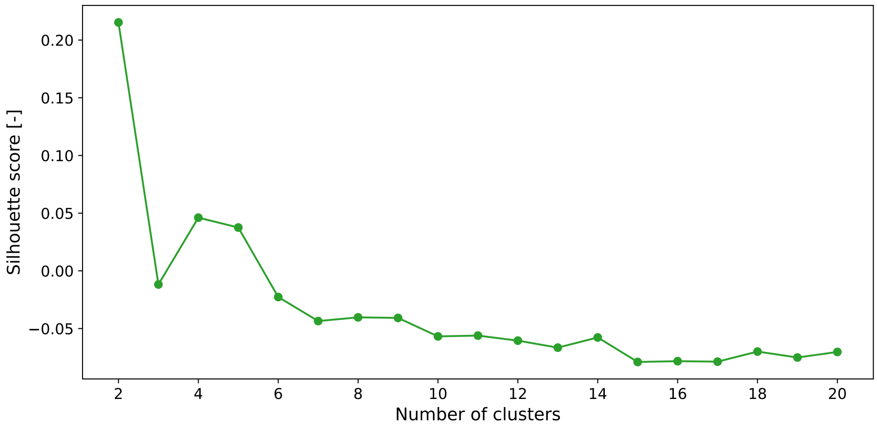


Figure 3.12: Silhouette score of the GMM clustering using the PCA-reduced dataset with 30 dimensions.

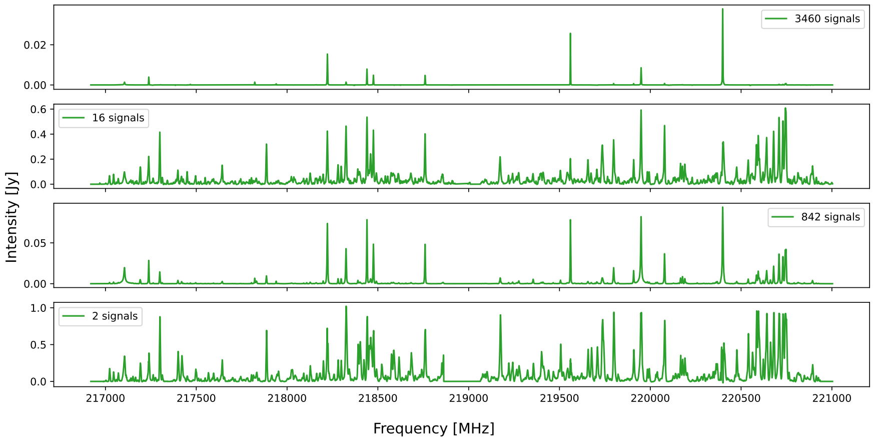


Figure 3.13: Cluster centroids obtained through sklearn's `GaussianMixture` classifier when using the subtraction-filtered dataset.

The maximum value was once more obtained at just 2 clusters, however, there is another local maximum at 4 clusters. The corresponding centroids using the subtraction-filtered signals can be seen in Figure 3.13. With two clusters consisting of just 18 signals in total, it can be stated that the GMM split the data into essentially 2 clusters. This is a far worse performance than that of the k-means classifier, so no more work was done with the PCA-reduced dataset.

3.2.2. UMAP

There are two possible UMAP projections available, depending on the distance metric chosen. These projections, as seen in Figure 3.14, differ in shape; however, they both concentrate the data around a number of areas. Furthermore, both projections also show groups of data far apart from the main clump, which is a strong indicator of a smaller cluster being present. Note that both projections were obtained using the subtraction-filtered dataset due to its higher performance with the previous classifier.

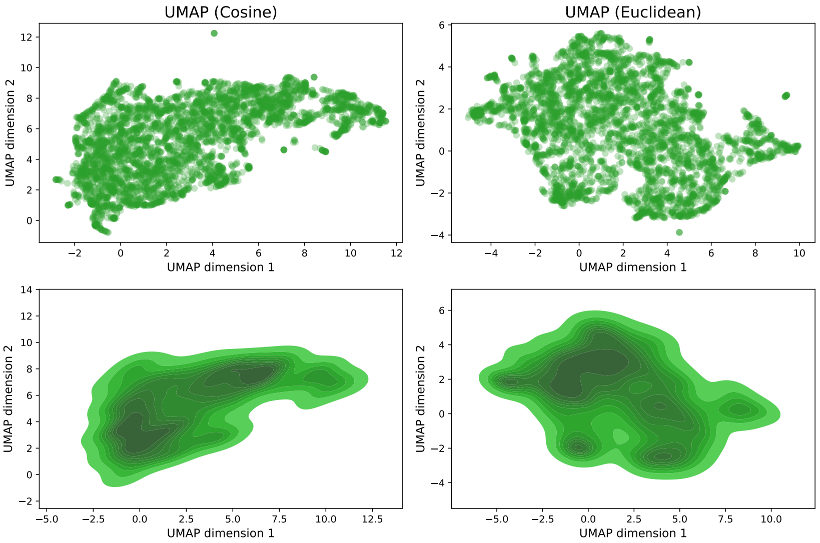


Figure 3.14: UMAP projections using the cosine distance metric (left) and the Euclidean distance metric (right).

Small cluster detection

Firstly the signals far apart from the main clump were investigated. These smaller clusters have been marked in Figure 3.15 and Figure 3.16. Five small clusters were identified in both cases, however, the only overlap between both groups is the one characterised by the broad-peaked signals (last cluster in Figure 3.15 and second in Figure 3.16). This cluster is not perfectly aligned, however, the 12 signals of the Euclidean projection form part of the 16 signals making the cosine-distance cluster, so those were selected as an independent cluster.

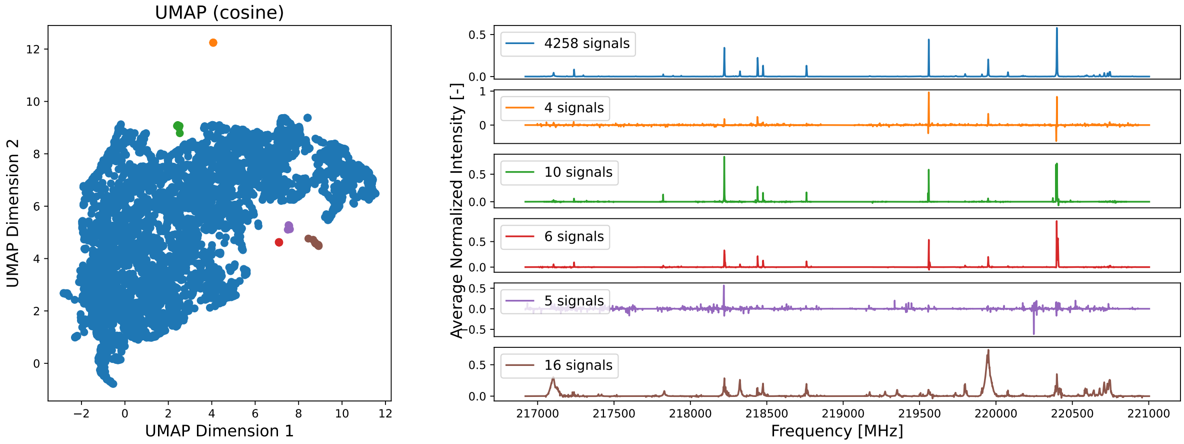


Figure 3.15: Manually identified small clusters in the cosine-distance UMAP projection (left) and their centroids (right).

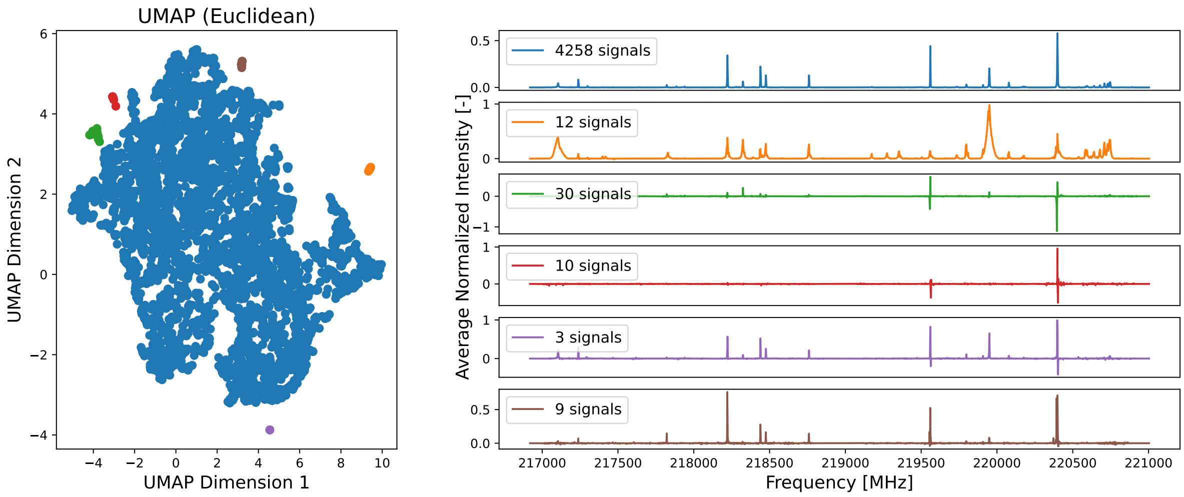


Figure 3.16: Manually identified small clusters in the Euclidean-distance UMAP projection (left) and their centroids (right).

One other small cluster shows a partial overlap between both projections: the third cluster (green) from the cosine projection and the last cluster (brown) of the Euclidean projection. Eight signals overlap between these two clusters, making up the second and last outlier cluster.

All other signals were used to train two different GMM models, one for each projection. The silhouette scores of the resulting clusters were then calculated using the subtraction-filtered signals and the labels from the GMM, resulting in Figure 3.17. In the figure, the silhouette score was calculated using both all the signals and using only the signals which had a probability higher than 90% of belonging to their assigned cluster. All lines share a local maximum at 5 clusters, however, the highest point for the Euclidean distance projection after the one at the start is at 7 clusters.

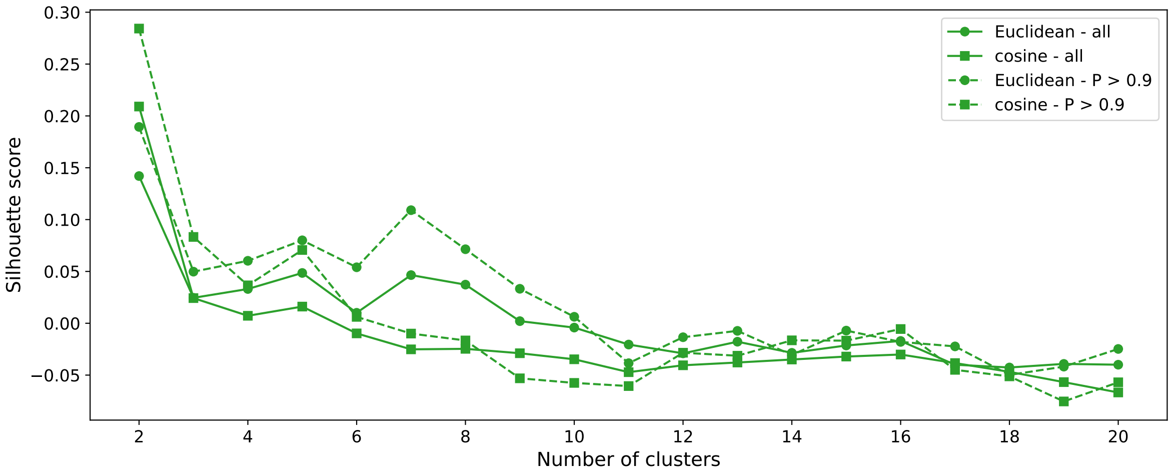
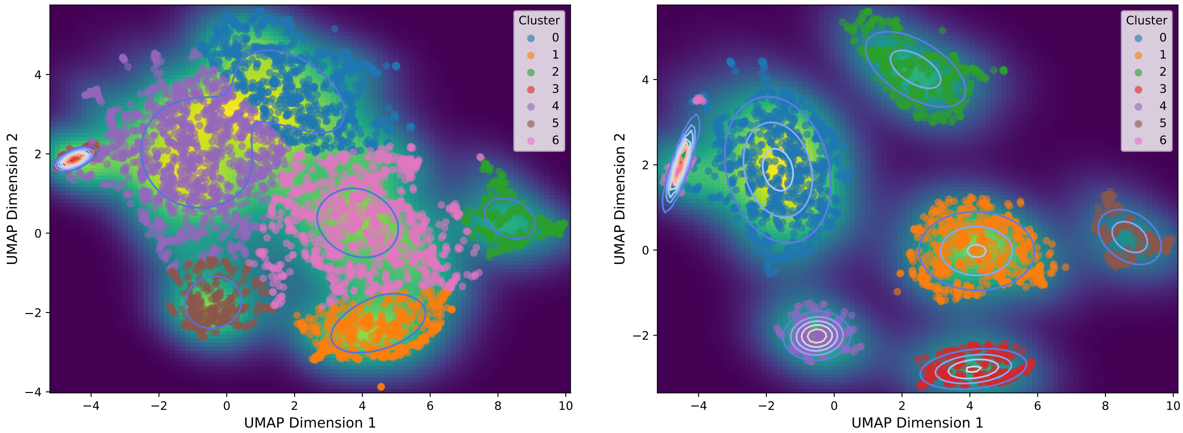


Figure 3.17: Silhouette score of the GMM classifier using the UMAP-reduced dataset with 2-20 clusters. Round markers correspond to models trained using the Euclidean-distance UMAP projection, whereas square markers correspond to models trained using the cosine-distance UMAP projection. The solid line represents the silhouette score for all signals and the dashed line represents the silhouette score obtained using only the signals with a probability higher than 90% of belonging to their cluster.

Euclidean-distance UMAP projection

As mentioned previously, Figure 2.13 shows that the Euclidean-distance UMAP projection points cluster around 7 different areas, which perfectly matches the number of clusters identified through the silhouette score. The resulting classification can be seen in Figure 3.18, where Figure 3.18a shows the classification using the entire dataset and Figure 3.18b shows the classification using only the signals with a probability greater than 90% of belonging to their assigned cluster.



(a) Classification using the entire dataset.

(b) Classification using only the signals with a high probability.

Figure 3.18: GMM classification of the Euclidean-distance UMAP projection using 7 clusters.

The centroids corresponding to the high-probability clusters shown in Figure 3.18b can be seen in Figure 3.19. These signals represent only 59.14% of the dataset, however, the missing signals are most likely transitioning between stages. This means that these high-probability signals most likely represent the different stages best. The remaining signals will be analysed later.

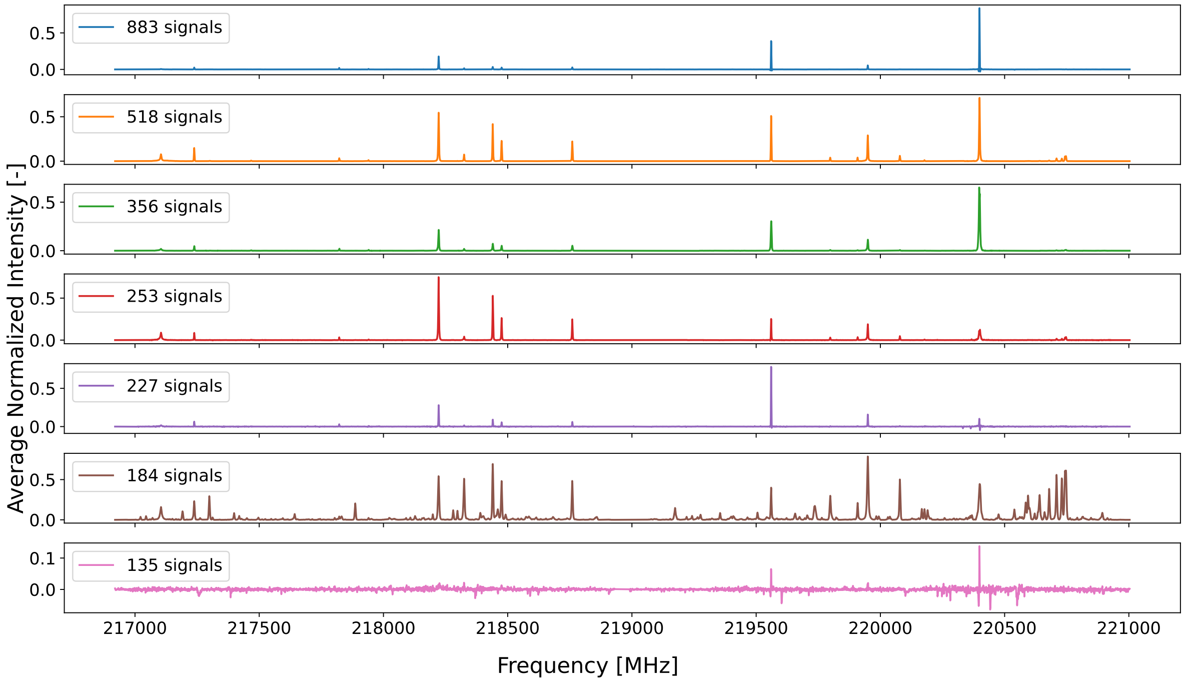
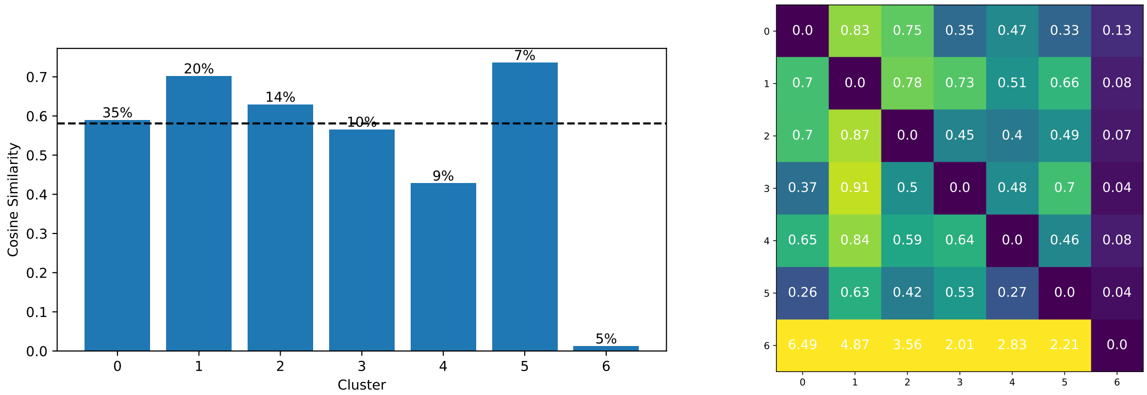


Figure 3.19: Centroids of the high-probability clusters shown in Figure 3.18b.

At first glance it would seem that the clusters are differentiable as all centroids are different from one another. The only exception to this may be the first and third clusters, where the only difference is that the latter has a slightly more intense spw0 and a slightly more significant SO peak. Beyond these observations, the quality of the clusters was assessed through the WCS and BCS/WCS ratio, which can be seen in Figure 3.20. Overall, the average WCS is slightly lower than the one obtained for the k-means classifier, although this is due to the almost null similarity of the last cluster. Excluding cluster 5, the average WCS increases up to 0.613, which is comparable to the quality of the clusters found through k-means. The BCS/WCS ratio, on the other hand, is significantly lower, with an average value of 0.2939 after discarding cluster 5. It is import to keep in mind, however, that these signals represent just 59.14% of all signals, whereas the k-means model utilized all signals.

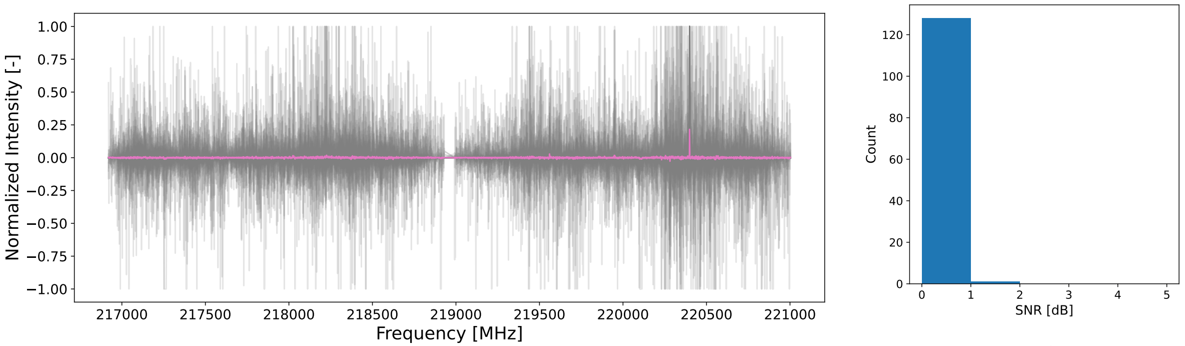


(a) Within-cluster similarity

(b) BCS/WCS ratio.

Figure 3.20: WCS and BCS/WCS of the clusters resulting from applying a GMM with 7 clusters to the Euclidean-distance UMAP projection of the subtraction-filtered dataset.

The first point to address was the almost null similarity between the signals in the last cluster (pink). Figure 3.19 shows that the signals that make it up share only the ^{13}CO signal. However, the centroid also displays several artifacts around the baseline and the maximum intensity of the centroid is just above 0.1, which indicates a very poor alignment. For this reason, Figure 3.21a shows all signals that make up the cluster in grey, and the mean signal in yellow. As seen in the figure, the signals greatly differ between them, both in the features present and in which features are the dominant ones. Furthermore, there is an abundant amount of absorption signals. These features are most likely residual noise that remained after filtering, which was enhanced when normalizing. This was confirmed through Figure 3.21b, which shows the distribution of SNR values in the signals from the sixth cluster. Given this information, this cluster was considered a cluster of noisy signals and was excluded from further analysis.

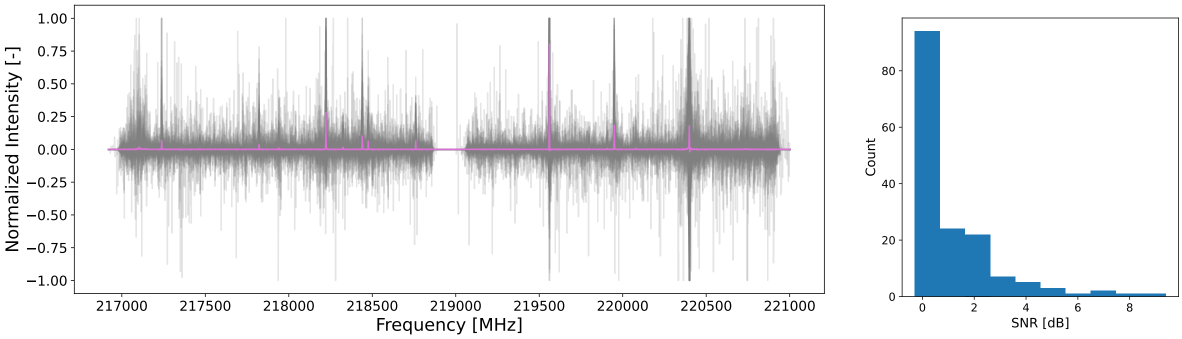


(b) SNR distribution of the signals that make up the sixth cluster.

(a) Overlaid signals of cluster 5, shown in grey; together with the mean cluster signal, shown in pink.

Figure 3.21: Signals and SNR distribution of cluster 5.

The only other cluster with a WCS lower than 0.5 was cluster 4 (purple). The same analysis was therefore carried, leading to Figure 3.22. As seen in the figure, there is also significant disparity between the signals that make up this cluster, although the features present are more consistent than in the previous case. Similarly, the SNR distribution, shown in Figure 3.22b, also shows that the signals in this cluster also have a low SNR, although not as low as the previous cluster's signals. Overall, most signals in this cluster show $C^{18}O$ and H_2CO features, which appear in the mean signal; however, a significant amount of signals also show ^{13}CO features. This last feature is also accompanied by a strong absorption signal near the same location. Given this information and the fact that the ^{13}CO line is almost not present in the mean signal, it is possible that a slight misalignment in the red-shift correction was present. On the other hand, it is also possible that the amount of signals that show this feature is not high enough compared to the amount of signals that show the $C^{18}O$ line. Due to time constraints, these hypotheses could not be tested further.

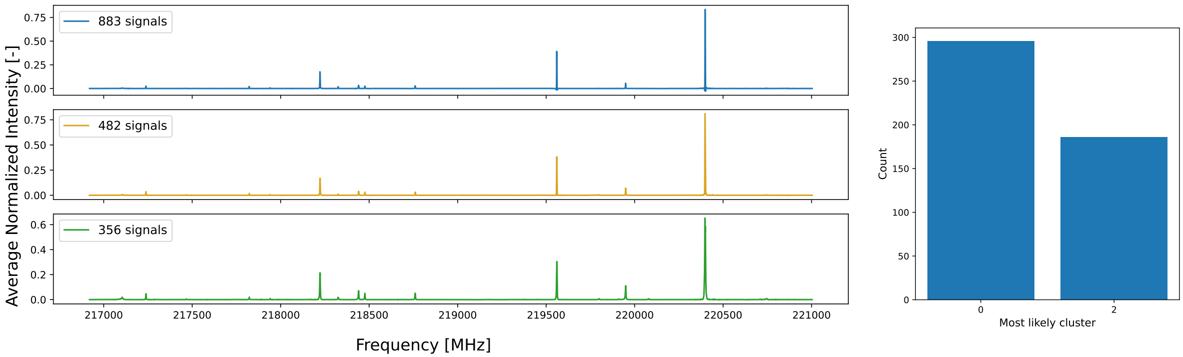


(a) Overlaid signals of cluster 4, shown in grey; together with the mean cluster signal, shown in light purple.

(b) SNR distribution of the signals that make up the thrid cluster.

Figure 3.22: Signals and SNR distribution of cluster 4.

The final step of the analysis was to determine the shape of the signals that lied between two clusters. To do this, the signals whose probability of belonging to either of 2 clusters was greater than 0.9 were split into groups depending on which two clusters they lied between. The first group investigated was the one made up of the signals lying between clusters 0 (blue) and 2 (green). A total of 566 signals make up this group, with a mean spectrum shown in Figure 3.23a. Both clusters are very similar between them so not many differences can be observed in the transitioning signals, however, it can be observed that the SO peak's intensity in the transitioning signals has an intermediate magnitude between the ones found in both cluster centroids.



(a) Mean signal of clusters 0 (top) and 2 (bottom), together with the mean signal of the signals lying between both clusters (middle).

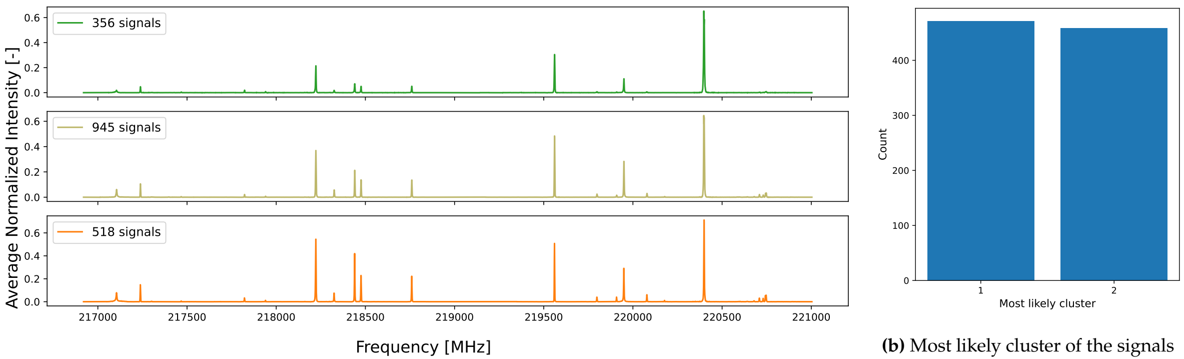
(b) Most likely cluster of the signals whose probability of belonging to clusters 0 or 2 add up to more than 90%.

Figure 3.23: Mean signal comparison of the signals lying between clusters 0 and 2, together with the distribution of the most likely cluster.

The transition is more visible in the signals lying between clusters 1 (orange) and 2 (green). As seen in Figure 3.24a, te signals transitioning between both groups have a stronger spw0 than cluster 2's but weaker than cluster 1's. The same is applicable to the intensity of the SO peak relative to that of the CO peaks and the smaller features in between the three.

Similarly, these transitions continued to be true in the subsequent analyses. It can be most easily seen in Figure 3.26, where the heavy contrast between the low number of features of cluster 1 and the large number of features of cluster 5 results in a signal with many spectral lines but where the intensity of the CH_3CN peaks is still lower than that of the CO peaks. The only transition that was not studied was the one between clusters 0 and 6 due to the low WCS of the latter.

Overall, this classifier managed to yield 6 clusters with a comparable WCS to that of the k-means algorithm (0.613) and a significantly lower BCS/WCS (0.2939). Only 59.14% of all signals were clustered,

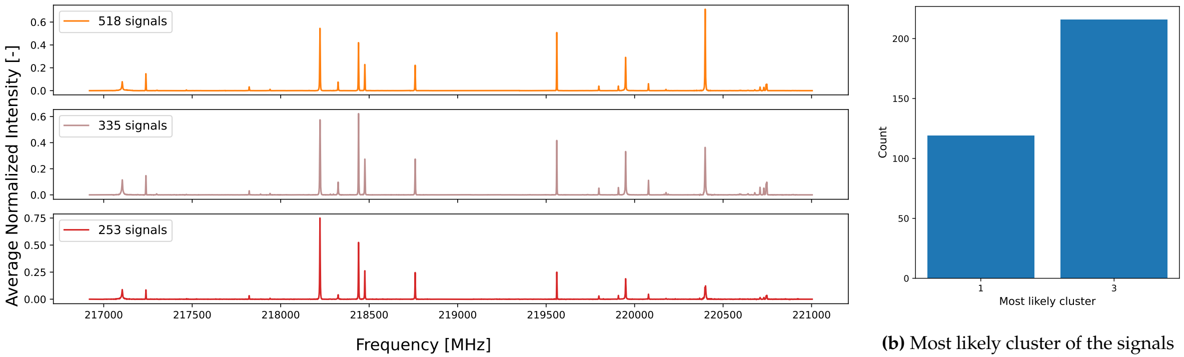


(a) Mean signal of clusters 2 (top) and 1 (bottom), together with the mean signal of the signals lying between both clusters (middle).

(b) Most likely cluster of the signals whose probability of belonging to clusters 1 and 2 add up to more than 90%.

Figure 3.24: Mean signal comparison of the signals lying between clusters 0 and 3, together with the distribution of the most likely cluster.

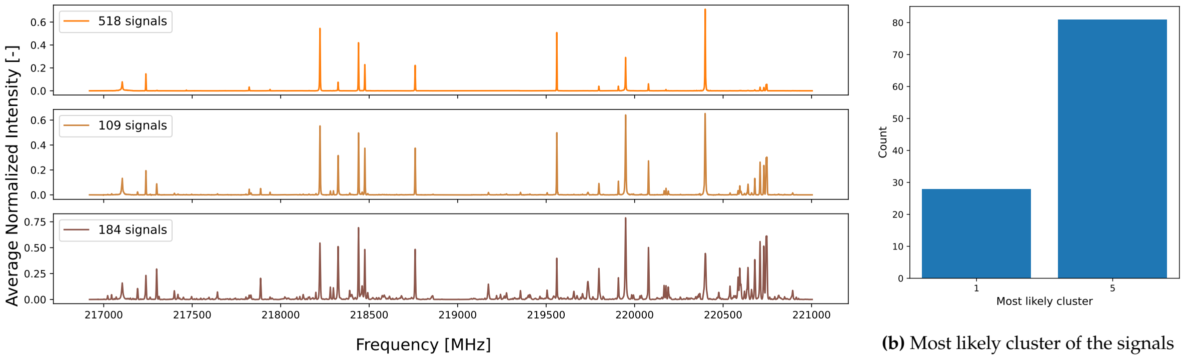
however, the remaining signals were proved to be a mixture of two other clusters, supporting the theory that those signals are in a transition state between two of the defined clusters.



(a) Mean signal of clusters 1 (top) and 3 (bottom), together with the mean signal of the signals lying between both clusters (middle).

(b) Most likely cluster of the signals whose probability of belonging to clusters 1 and 3 add up to more than 90%.

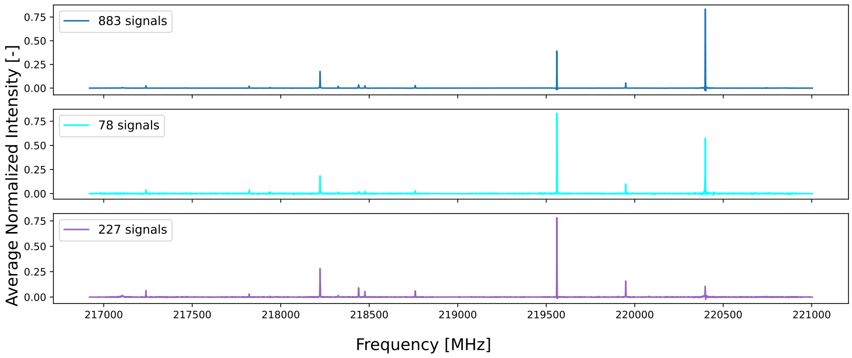
Figure 3.25: Mean signal comparison of the signals lying between clusters 1 and 3, together with the distribution of the most likely cluster.



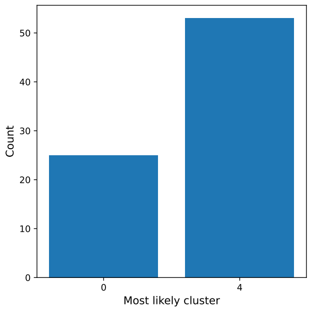
(a) Mean signal of clusters 1 (top) and 5 (bottom), together with the mean signal of the signals lying between both clusters (middle).

(b) Most likely cluster of the signals whose probability of belonging to clusters 1 and 5 add up to more than 90%.

Figure 3.26: Mean signal comparison of the signals lying between clusters 1 and 5, together with the distribution of the most likely cluster.



(a) Mean signal of clusters 0 (top) and 4 (bottom), together with the mean signal of the signals lying between both clusters (middle).



(b) Most likely cluster of the signals whose probability of belonging to clusters 0 and 4 add up to more than 90%.

Figure 3.27: Mean signal comparison of the signals lying between clusters 0 and 4, together with the distribution of the most likely cluster.

The final (high-confidence) cluster arrangement would therefore consist of the 7 clusters from Figure 3.19 and the two small clusters previously identified. The sencond of these latter clusters, however, shares a very high resemblance to several of the larger clusters, causing the BCS/WCS ratio to increase from 0.2696 to 0.3029. Therefore, only the small cluster consisting of the broad-peaked signals was selected, leading to the centroids seen in Figure 3.28. The quality of this final arrangement can be seen in Figure 3.29, with an average WCS of 0.6138 and an average BCS/WCS ratio of 0.2696.

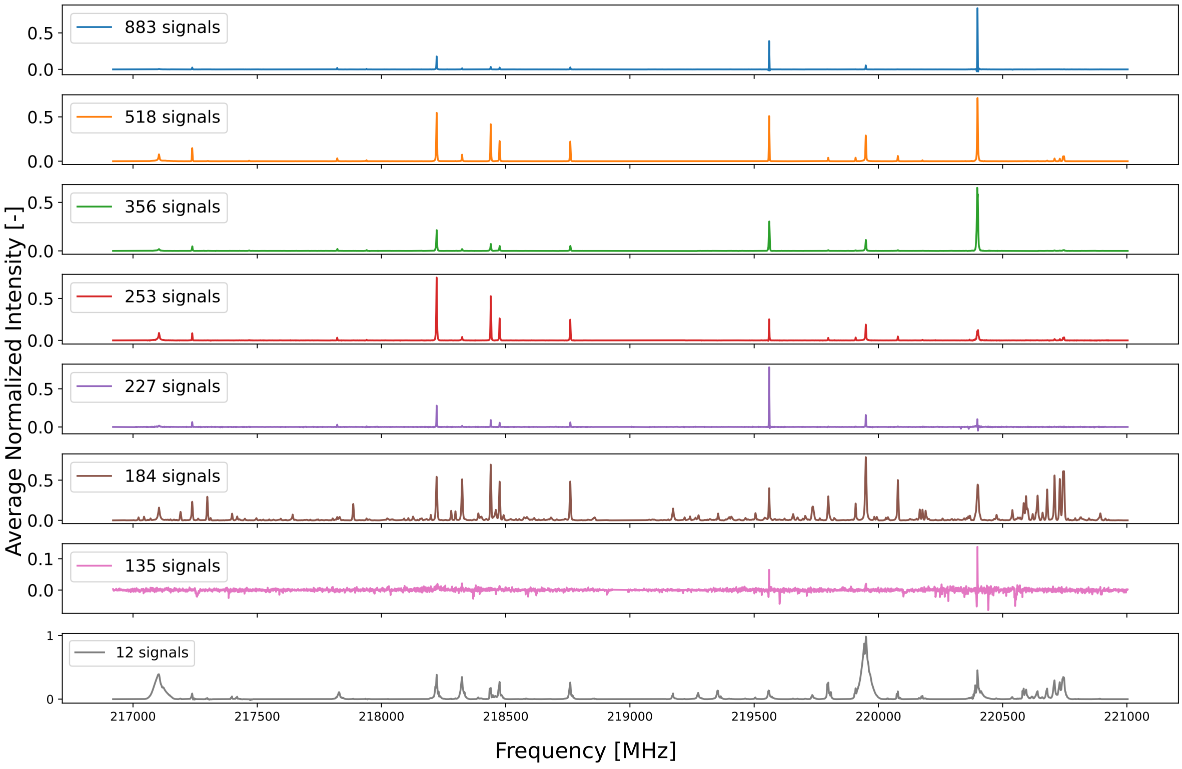
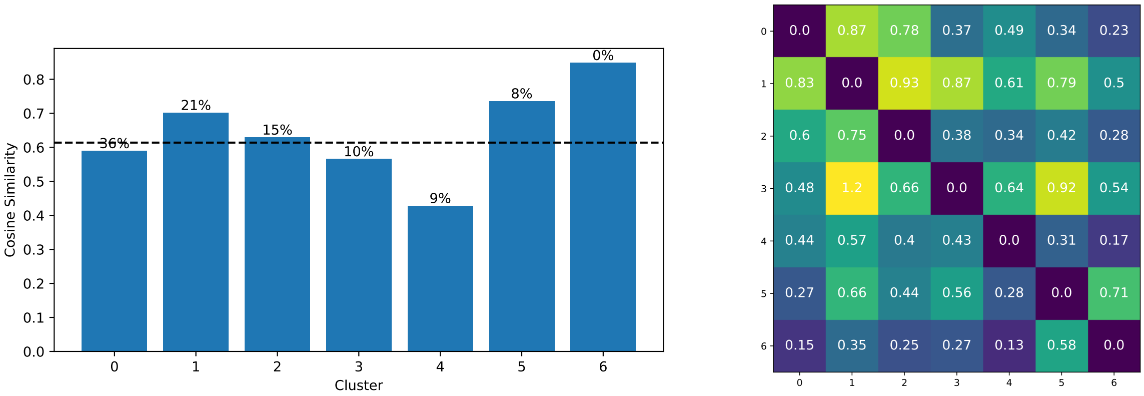


Figure 3.28: Centroids of the final cluster arrangement using the Euclidean-distance UMAP projection of the subtraction-filtered dataset.



	0	1	2	3	4	5	6
0	0.0	0.87	0.78	0.37	0.49	0.34	0.23
1	0.83	0.0	0.93	0.87	0.61	0.79	0.5
2	0.6	0.75	0.0	0.38	0.34	0.42	0.28
3	0.48	1.2	0.66	0.0	0.64	0.92	0.54
4	0.44	0.57	0.4	0.43	0.0	0.31	0.17
5	0.27	0.66	0.44	0.56	0.28	0.0	0.71
6	0.15	0.35	0.25	0.27	0.13	0.58	0.0

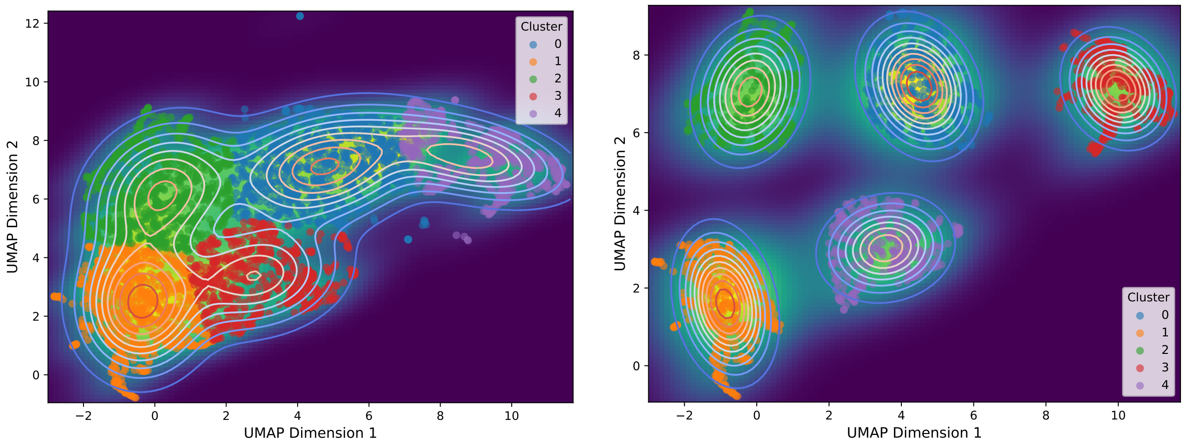
(a) Within-cluster similarity

(b) BCS/WCS ratio.

Figure 3.29: WCS and BCS/WCS of the final cluster arrangement using the Euclidean-distance UMAP projection of the subtraction-filtered dataset.

Cosine-distance UMAP projection

Following the analysis of the Euclidean-distance UMAP projection, the same steps were followed for the cosine-distance UMAP projection. In this instance, the silhouette score showed a local maximum at 5 clusters, so that arrangement was investigated. The arrangement and corresponding cluster centroids can be seen in Figure 3.30 and Figure 3.31 respectively.



(a) Classification using the entire dataset.

(b) Classification using only the signals with a high probability.

Figure 3.30: GMM classification of the cosine-distance UMAP projection using 5 clusters.

These clusters show clear differences between them, especially cluster 3 (red), which shows far more peaks than the rest. The only two clusters that show high similarity are clusters 1 (orange) and 2 (green), similarly to the case with clusters 0 and 2 in the Euclidean-distance projection (see Figure 3.19). In fact, all clusters are very similar to another one from the Euclidean projection, leaving just 2 undetected clusters: cluster 3, which a stronger spw0; and cluster 6, which was characterised by an abundance of signals with a low SNR. Other arrangements with a larger number of clusters were investigated, however, those clusters remained unidentified.

Following the same procedure as with the Euclidean-projection GMM, the WCS and BCS/WCS ratio were calculated. The results, seen in Figure 3.32, show how the size-averaged WCS is just 0.5607 and the average BCS/WCS ratio is 0.3363. These results are far worse than those obtained with the Euclidean projection, with the latter having a WCS 9.33% higher and a BCS/WCS ratio 12.61% lower. Furthermore, these values refer to just 52.79% of the dataset, 6.35% less than with the Euclidean-distance projection.

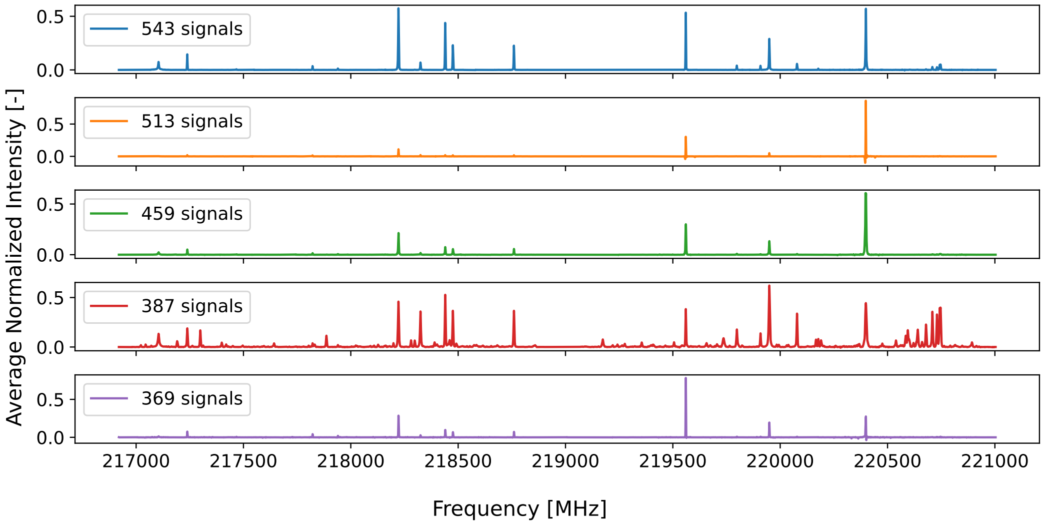
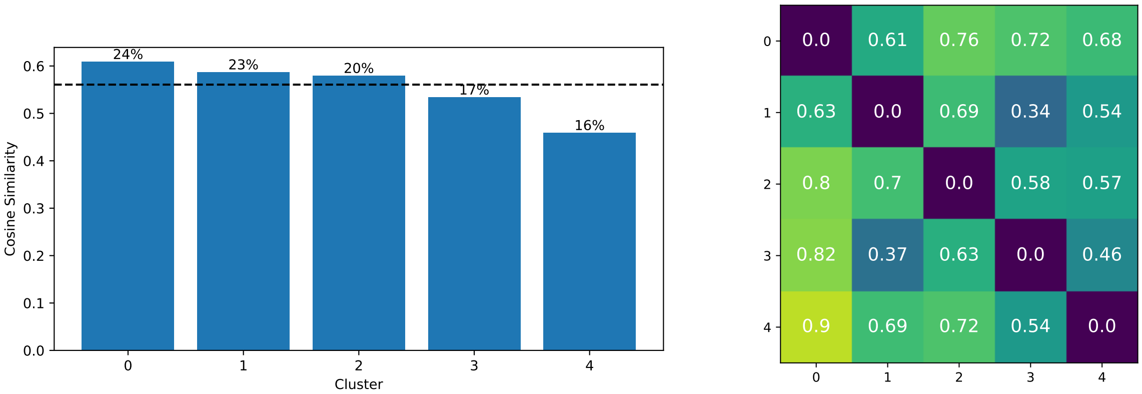


Figure 3.31: Centroids of the high-probability clusters shown in Figure 3.30b.



(a) Within-cluster similarity

(b) BCS/WCS ratio.

Figure 3.32: WCS and BCS/WCS of the clusters resulting from applying a GMM with 5 clusters to the cosine-distance UMAP projection of the subtraction-filtered dataset.

Seeing how all evaluation metrics show that this clustering is much worse than the one obtained with the Euclidean projection, no further research was made into this model.

3.2.3. Conclusion

When using Gaussian Mixture Models (GMM) to cluster the data, it is imperative to use a dimensionality reduction method. In this study, PCA and UMAP were used to this end, with the latter greatly outperforming the former. Using the PCA reduction, the GMM only managed to distinguish 4 clusters, with two of them adding up to just 18 signals. Using UMAP, on the other hand, two different results were obtained depending on whether the reduction method used the Euclidean- or cosine-distance metric. Comparing both projections, two small clusters were identified due to their distance from the main clump of points. The rest of the signals were then used to train the GMM, with the Euclidean-distance projection's silhouette score showing a local maximum at 7 clusters, and the cosine-distance projection's at 5 clusters. The former proved to be a much better clustering, with a 9.33% higher WCS and a 12.61% lowe BCS/WCS ratio, all the while clustering 6.35% more signals. The identified clusters were then analysed together with the smaller ones, which caused one of the small ones to be discarded due to its high similarity to the larger clusters. The final arrangement therefore consisted of a total of 8 clusters, with one of them consisting entirely of signals with a SNR lower than 1 dB. Not counting this last cluster, this arrangement has an average WCS of 0.6138 and an average BCS/WCS ratio of 0.2696.

Compared to the KMeans classifier, the GMM identified the same number of clusters, although it only clustered 56.29% of the dataset with a high degree of confidence. The remaining signals, however, were analysed and proven to be in a transition stage between two different clusters, supporting the hypothesis that these clusters represent different evolutionary stages. Furthermore, even though the WCS of this classifier is lower than the maximum one achieved through the k-means classifier (which achieved a WCS of 0.663 on 87% of the dataset), its average BCS/WCS ratio is 28.09% lower (0.2696 vs 0.3749). This proves that these clusters have less overlap than those found through the k-means classifier, making them better defined.

This model's performance is therefore comparable to the one obtained through k-means, however, the process followed to reach these cluster assignments included manually identifying and discarding clusters. This greatly decreases the reproducibility of the procedure which is one of the main objectives of this study.

Table 3.2: Summary of the GMM classifier's research.

Aspect	Finding	Comment
PCA	Unsuccessful	Using the PCA-reduced dataset yielded only 2 clusters with more than 20 signals.
Euclidean UMAP	Successful	8 clusters were detected, including a small cluster with 12 signals and a cluster made up entirely of signals with a SNR lower than 1 dB. This arrangement had an average WCS of 0.6138 and an average BCS/WCS ratio of 0.2696. On the other hand, only 56.29% of the signals were confidently clustered.
Cosine UMAP	Unsuccessful	Only 5 clusters were detected, with much worse WCS and BCS/WCS ratio than those obtained with the Euclidean-distance projection.
Reproducibility	Low	The small clusters were manually identified. Furthermore, one of them was manually discarded as it greatly increased the BCS/WCS ratio.

3.3. Agglomerative Clustering

Agglomerative clustering models work by initially considering each signal as its own cluster, and then merging the closest clusters until a stopping criterion is met. This stopping criterion can be a set number of clusters, a distance threshold or a more complex rule. Finally, the distance between clusters can also be defined in different ways, with common methods being single linkage (minimum distance), complete linkage (maximum distance), average linkage (average distance), and Ward's method (minimizing variance). Variants of this base model also exist, such as first model tested in this section, which considers only a certain number of neighbors.

This section will show the results for three different models. The first one will be a neighborhood-based agglomerative clustering model based on Zhang et al. [71], first using the subtraction-filtered dataset and then both UMAP-reduced datasets. The next two models are basic agglomerative clustering models using the Euclidean- and cosine-distance metrics on the subtraction-filtered dataset.

3.3.1. Neighborhood-based agglomerative clustering

The idea behind this model is that it will stop merging signals once each signal's first k neighbors belong to the same cluster. This would provide an easier-to-optimize stopping criterion, as the number of final clusters would be determined by the number of neighbors set. However, this approached turned out unsuccessful as only one large cluster was identified independently of the hyperparameters used. Such an example can be seen in Figure 3.33, which shows the final clusters obtained when using a cosine distance metric and setting the number of neighbors to 25.

The largest cluster was therefore studied. Backtracking all the merging steps, it was possible to identify the moments where two large clusters were being merged. Using a cosine distance metric and 25

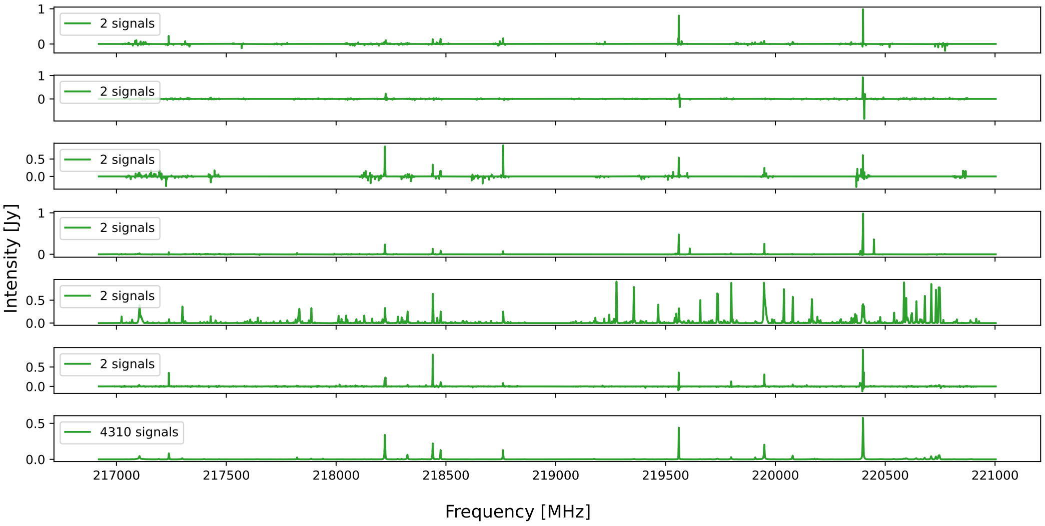


Figure 3.33: Mean signal of the final clusters obtained from a neighborhood-based agglomerative clustering model with the number of neighbors set to 25 and using a cosine distance metric.

neighbors, this last occurred at step 593/612. Figure 3.34 shows how both centroids being merged and Figure 3.35 the locations of their signals in both UMAP projections.

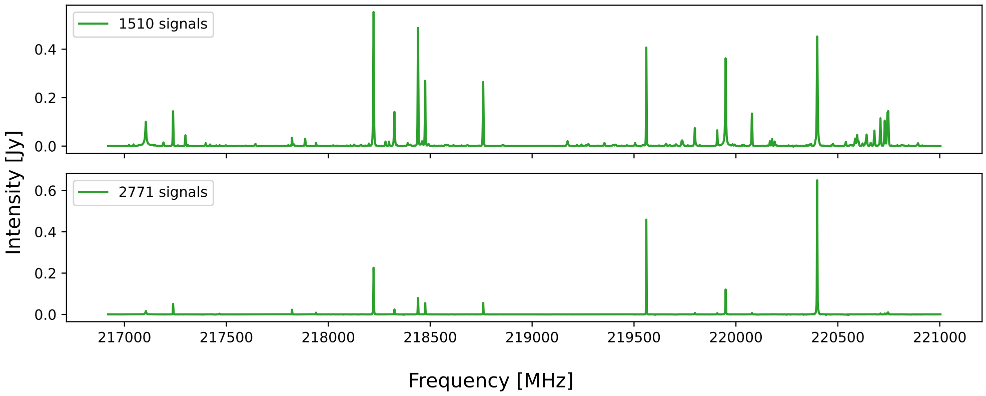
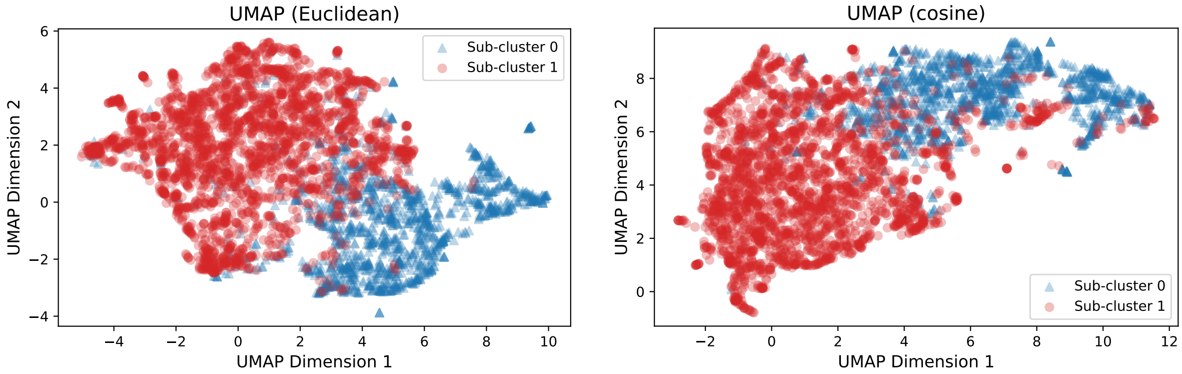


Figure 3.34: Mean signals of the clusters being merged in step 593 of the model.



(a) Euclidean-distance projection depicting the location of the signals being merged in step 593 of the neighborhood-based model.

(b) Cosine-distance projection depicting the location of the signals being merged in step 593 of the neighborhood-based model.

Figure 3.35: Location in both UMAP projections of the signals that make up the clusters being merged in step 593 of the neighborhood-based model.

The centroids of both clusters are widely different, proving that the model did not find an appropriate stopping point. Furthermore, even though the members of each cluster are grouped around a given section of the UMAP projection, there are outliers present in both clusters. For instance, there are two red points in Figure 3.35a with an x-value greater than 7. Those signals can be seen in purple in Figure 3.36, together with the mean signal they belong to (shown in red at the top of the figure) and the surrounding signals of the other cluster (shown in blue at the bottom). The signals shown in the middle show a much higher similarity to the surrounding cluster than to its own, indicating that they have been misclassified. This, combined with the success obtained when using the GMM model on the UMAP projection, motivated the study of applying the neighborhood-based agglomerative model directly to the UMAP projection.

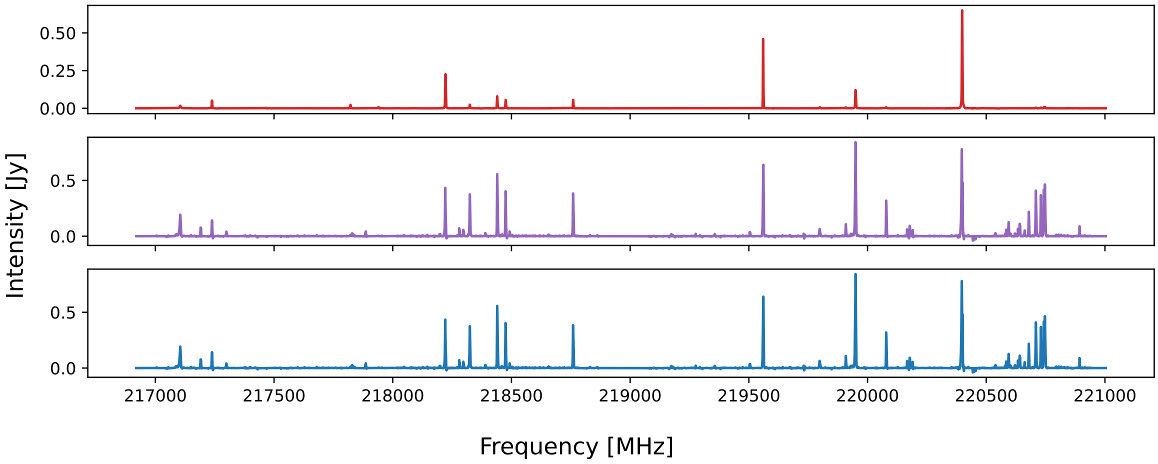


Figure 3.36: Top: mean signal of the red cluster shown in Figure 3.35. Middle: average signal of the points belonging to the red cluster shown in Figure 3.35a that have an x-value greater than 7 in the Euclidean projection. Bottom: Mean signal of the signals from the blue cluster shown in Figure 3.35a with an x-value greater than 7.

UMAP projection

Given the different nature of both projections, the first parameter checked was the number of final clusters as a function of the number of neighbors. Figure 3.37 shows that the number of final clusters differs at most by two, except at $k = 10$ neighbors, where the difference is of three clusters.

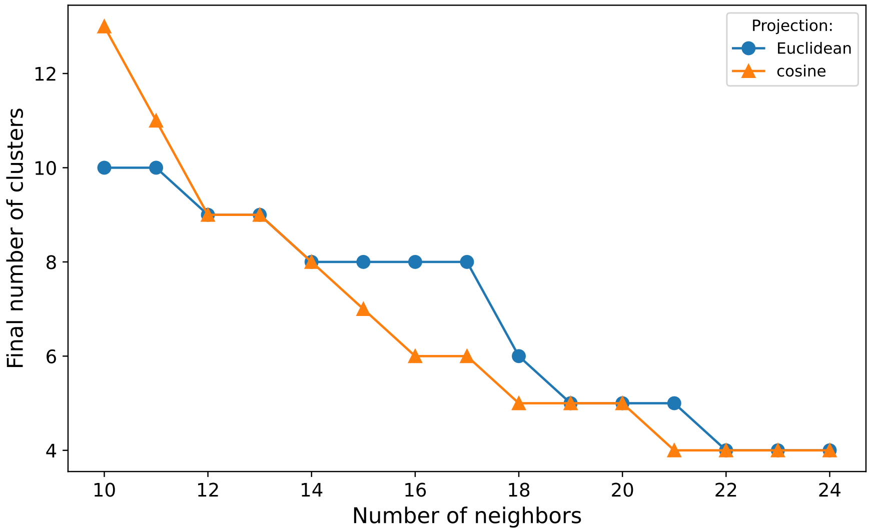


Figure 3.37: Number of final clusters as a function of the number of neighbors when applying the agglomerative model to the UMAP projections directly.

The first setting tested was the one where $k = 12$ (with $k_0 = 2$) as that was the first value for which both projections identified the same number of signals. Figure 3.38 and Figure 3.39 show the final centroids obtained using the Euclidean and cosine UMAP projections respectively. In both cases, eight small clusters and one large cluster were identified. Furthermore, both models successfully identified similar clusters to those found at the beginning of subsection 3.2.2 within the small clusters. Some of these clusters, however, are very similar between them or to the large cluster. For instance, clusters 2 and 3 from Figure 3.38 only differ from the main cluster in small details around the ^{13}CO line. The next step was therefore to sub-cluster the large cluster in both projections.

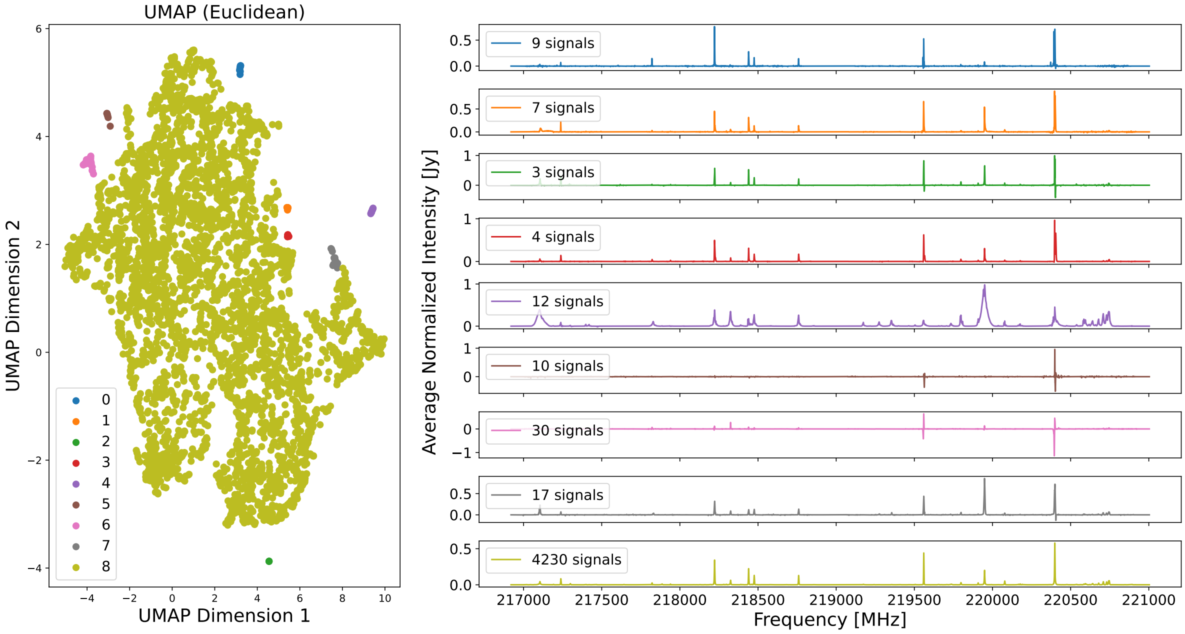


Figure 3.38: Cluster location in the Euclidean-distance UMAP projections and their centroids using the neighborhood-based agglomerative clustering model on the UMAP projection.

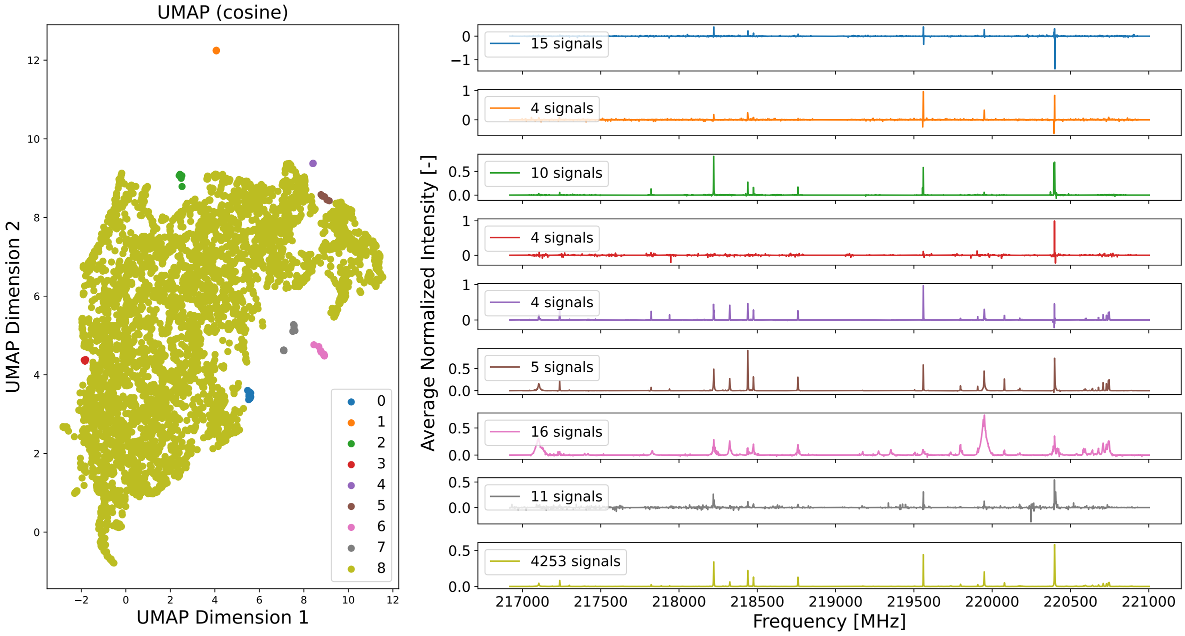


Figure 3.39: Cluster location in the cosine-distance UMAP projections and their centroids using the neighborhood-based agglomerative clustering model on the UMAP projection.

The first attempt to sub-cluster the large cluster in both projections consisted on applying the same algorithm to the points making up the cluster. This, however, yielded one large cluster and several small ones once more. Instead, the merging process was inspected. This was done by inspecting the size of the largest cluster of each model throughout the last 20 steps, as seen in Figure 3.40.

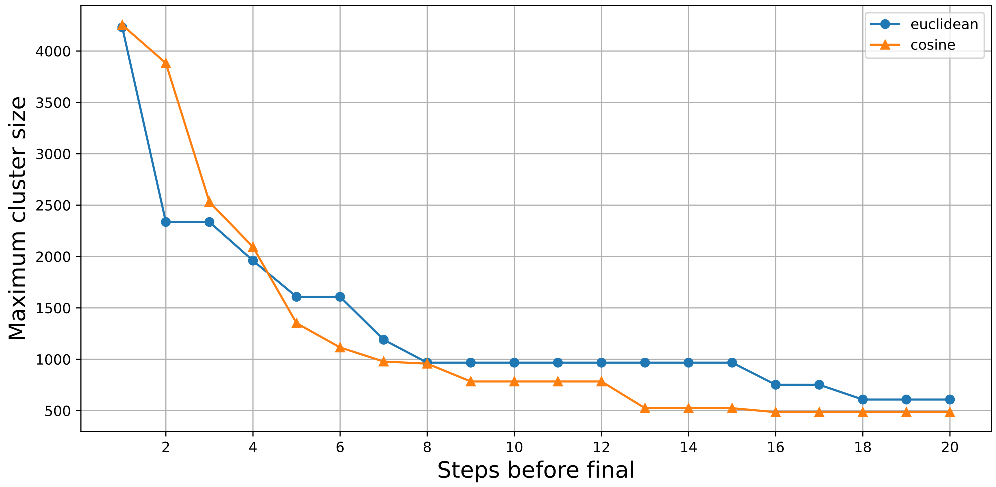


Figure 3.40: Size of the largest cluster in the last 20 steps of each model.

Both model's largest cluster had approximately 1000 signals at the eighth step before the last. Furthermore, the small clusters remaining at the end were formed in the first step of the algorithm, indicating that both models subdivided their largest cluster into 8 sub-clusters at that step. Given these reasons, that point was manually inspected, leading to Figure 3.41 and Figure 3.42.

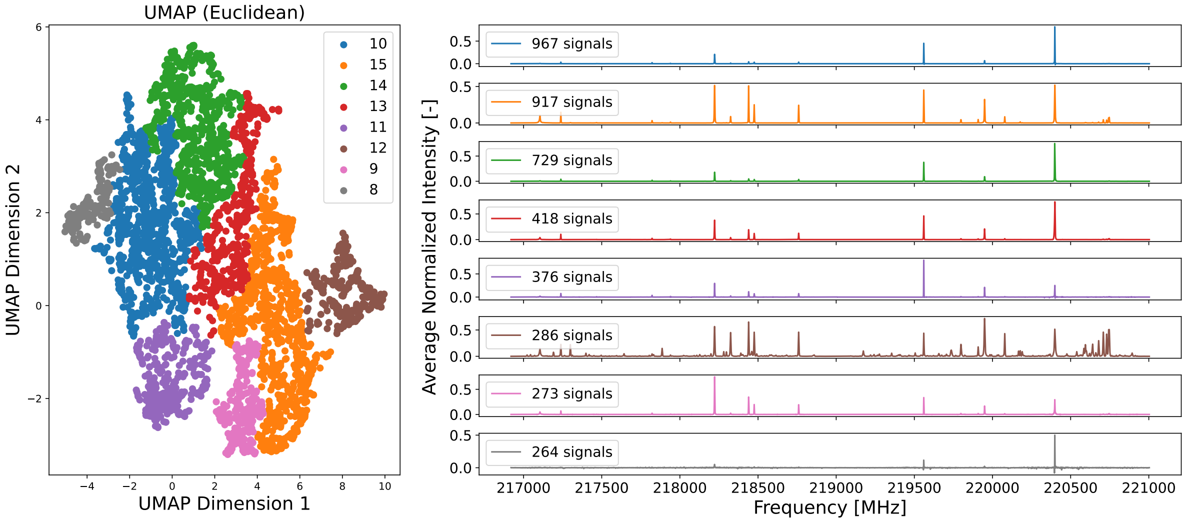


Figure 3.41: Projection and centroids of the sub-clustering achieved at the 8th merging step before the last, using the neighborhood-based agglomerative model on the Euclidean-distance UMAP projection of the subtraction-filtered dataset.

In both cases, several of the cluster centers are very similar to one another. For instance, clusters 0, 2 and 3 in Figure 3.41 differ only slightly in the magnitude of their spw0 peaks. Before deciding whether to merge them, the WCS of the sub-clusters was investigated. The results can be seen in Figure 3.43, with a size-weighted average WCS of 0.592 for the Euclidean projection and 0.589 for the cosine one. Note, however, that the Euclidean projection's WCS is heavily penalized by sub-cluster 0, which has a similarity of 0.166 and is made up of the low SNR signals identified through the GMM. Discarding this sub-cluster due to these reasons leads to a new WCS of 0.621.

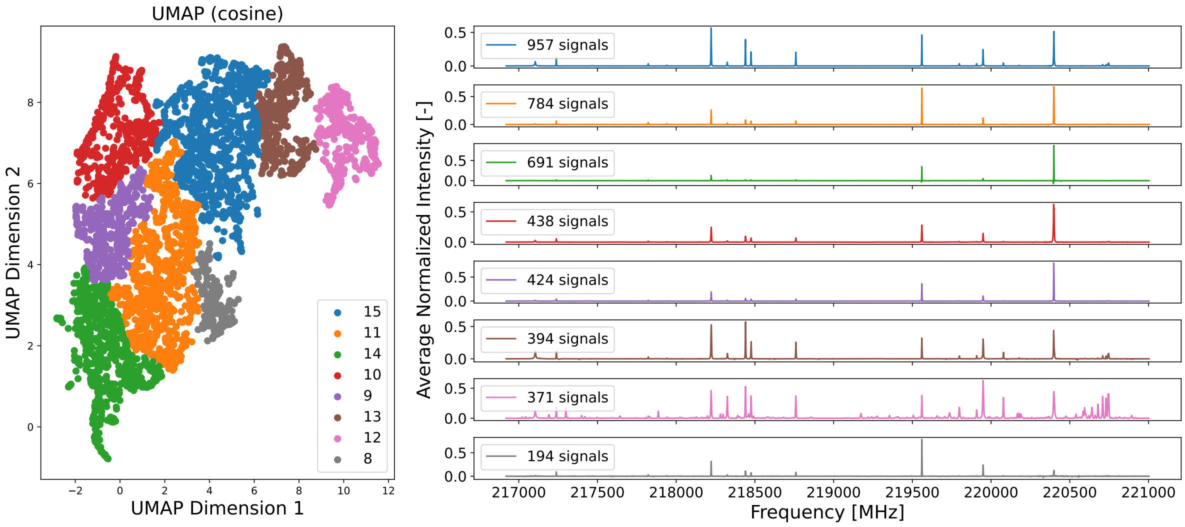
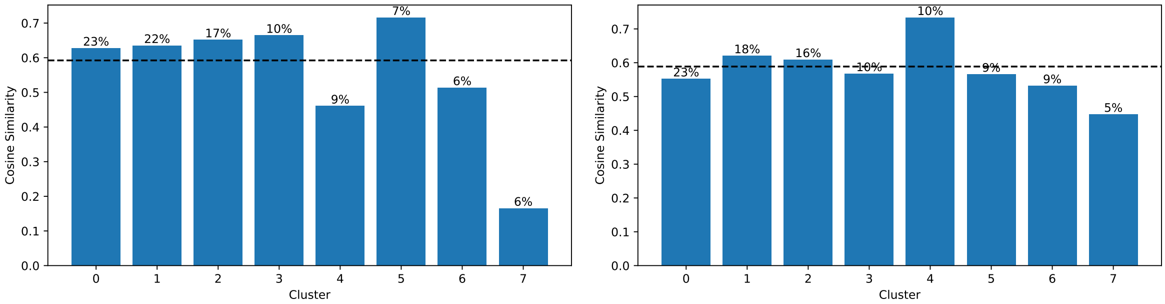


Figure 3.42: Projection and centroids of the sub-clustering achieved at the 8th merging step before the last, using the neighborhood-based agglomerative model on the cosine-distance UMAP projection of the subtraction-filtered dataset.

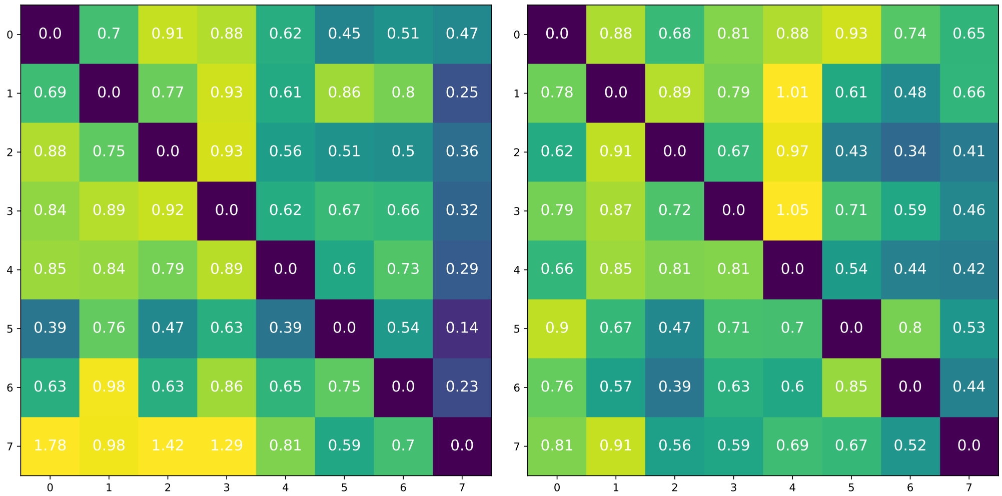


(a) WCS of the sub-clusters shown in Figure 3.41.

(b) WCS of the sub-clusters shown in Figure 3.42.

Figure 3.43: Within cluster similarity (WCS) of the sub-clusters identified for both projections.

Similarly to other models, the BCS/WCS ratio of both arrangements was computed. This resulted in the matrices seen in Figure 3.44, with average values of 0.418 (Euclidean) and 0.435 (cosine).



(a) Euclidean-distance sub-clusters.

(b) Cosine-distance subclusters.

Figure 3.44: BCS/WCS ratio between the sub-clusters identified in the Euclidean- and cosine-distance UMAP projections.

Since no pair of clusters had a BCS/WCS greater than one in both cases, no clusters were merged. This meant that the clusters found were also the final clusters, resulting in the centroids seen in Figure 3.45 (Euclidean UMAP) and Figure 3.46 (cosine UMAP). The WCS of the large clusters is comparable to that of the previous models, however, the BCS/WCS ratio is much 55.17% higher than the one found through the GMM. This decrease in performance, together with the low reproducibility of this approach, motivated the study of conventional agglomerative clustering models.

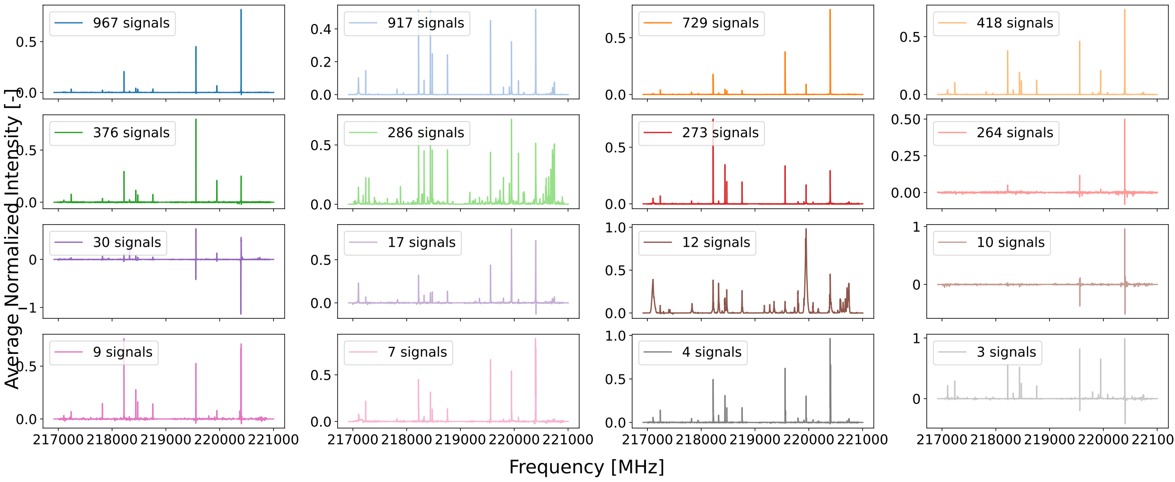


Figure 3.45: Final clusters identified with the neighborhood-based agglomerative model, using the Euclidean UMAP projection.

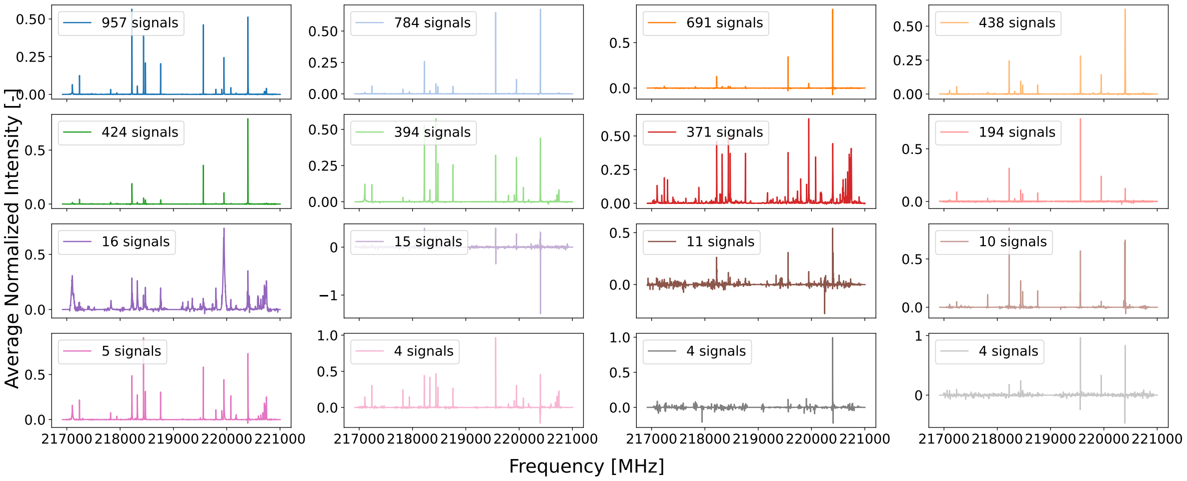


Figure 3.46: Final clusters identified with the neighborhood-based agglomerative model, using the cosine UMAP projection.

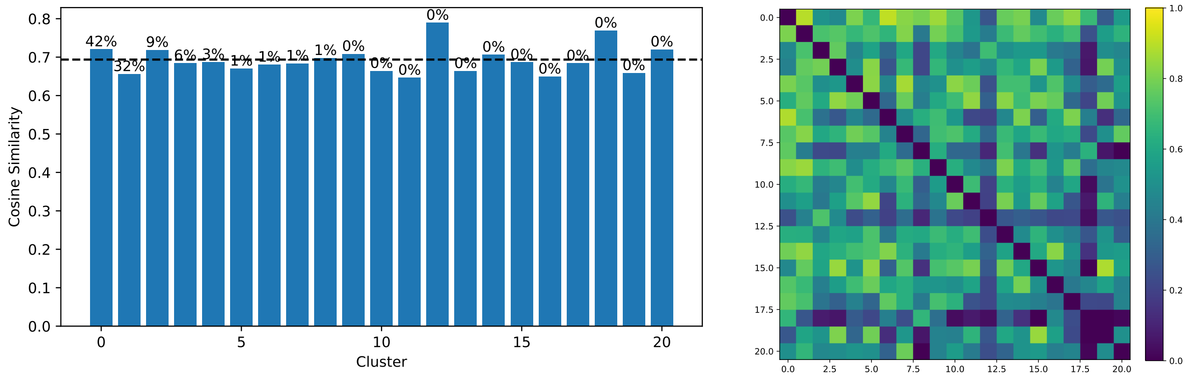
3.3.2. Conventional agglomerative clustering

Cosine-distance agglomerative clustering

The main advantage of conventional agglomerative clustering is that the stopping point can be manually decided based on the distances between clusters. This means that the user can decide to stop the merging process once the distance between two clusters is larger than a certain threshold. Since the previous models achieved a WCS of ~ 0.6, that threshold was first set at 0.3 (equivalent to a similarity of 0.7). The next parameter that needed to be decided was the linkage type. A conservative approach was to use a complete linkage at first, and if the clusters turned out to be too similar to one another, then an average linkage would be tested.

A total of 1076 clusters were identified using this configuration directly on the subtraction-filtered dataset. From these clusters, only 97 had more than 10 signals, and only 37 more than 20. This clustering only achieved a silhouette score of 0.077 when using the cosine distance metric, most likely due to heavy overlap between the different clusters. An average linkage was therefore tested, which yielded 819 different clusters, with 34 of them having more than 10 signals and 23 more than 20. The silhouette score with this configuration, however, was also low, with a value of 0.068 using the cosine distance metric.

The threshold was then lowered to 0.4, which yielded a clustering assignment where, using an average linkage, 21 clusters had more than 10 signals. These clusters, additionally, contained 81.74% of all signals (3521/4322). With an average WCS of 0.6936 and an average WCS/BCS ratio of 0.2944, as seen in Figure 3.47, this arrangement seemed very promising. However, most of the clustered signals (75%) belong to just 2 clusters.



(a) WCS ratio.

(b) BCS/WCS ratio.

Figure 3.47: WCS and BCS/WCS ratio of the clusters with more than 10 signals obtained through a conventional agglomerative clustering algorithm using the cosine distance metric, with an average linkage and a distance threshold of 0.4.

Figure 3.48 shows the first 20 centroids obtained with this model. Based on the results of the other models, the two largest clusters could be broken down, therefore yielding an unsatisfactory result.

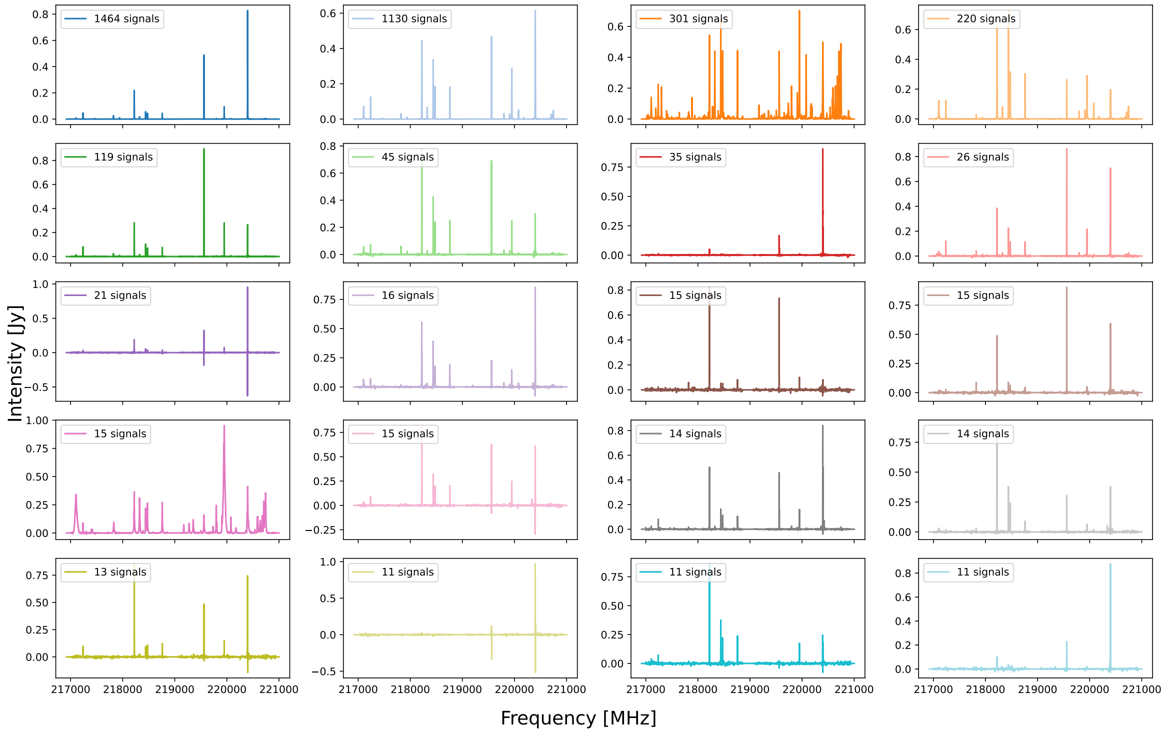


Figure 3.48: Centroids of the 20 clusters with more than 10 signals obtained through a conventional agglomerative clustering algorithm using the cosine distance metric, with an average linkage and a distance threshold of 0.4.

Euclidean-distance agglomerative clustering

Next, the same model was tested but using the Euclidean distance metric. Since two signals with the same outline but different intensities will have a non-zero distance using this metric (see Appendix A), this model was trained using the max-normalized subtraction-filtered dataset. Furthermore, this distance metric allows for the use of the "Ward" linkage, which minimizes the variability within each cluster. This distance is not limited like the cosine distance, so setting a distance threshold is not a simple task. Furthermore, setting such a limit would require a manual inspection of the dataset, which would reduce the reproducibility of the results, as this value would vary between datasets. The silhouette score was therefore evaluated at different stages in the merging process, resulting in Figure 3.49. As seen in the figure, the points of interest are at 10 and 23 clusters, since those are the settings with more than 5 clusters where the silhouette score is locally maximized.

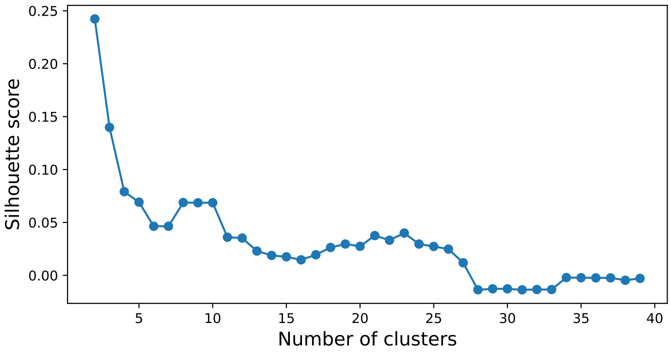


Figure 3.49: Silhouette score of the agglomerative clustering model using the Euclidean distance.

Figure 3.50 shows the centroids of the clusters identified by the model with 10 clusters. As seen in the figure, there are two clusters formed by just one signal due to their strong absorption features. These individual clusters explain the plateau seen in the silhouette score in Figure 3.49 between 8 and 10 clusters, as the next two merging steps would be to add those signals to other clusters.

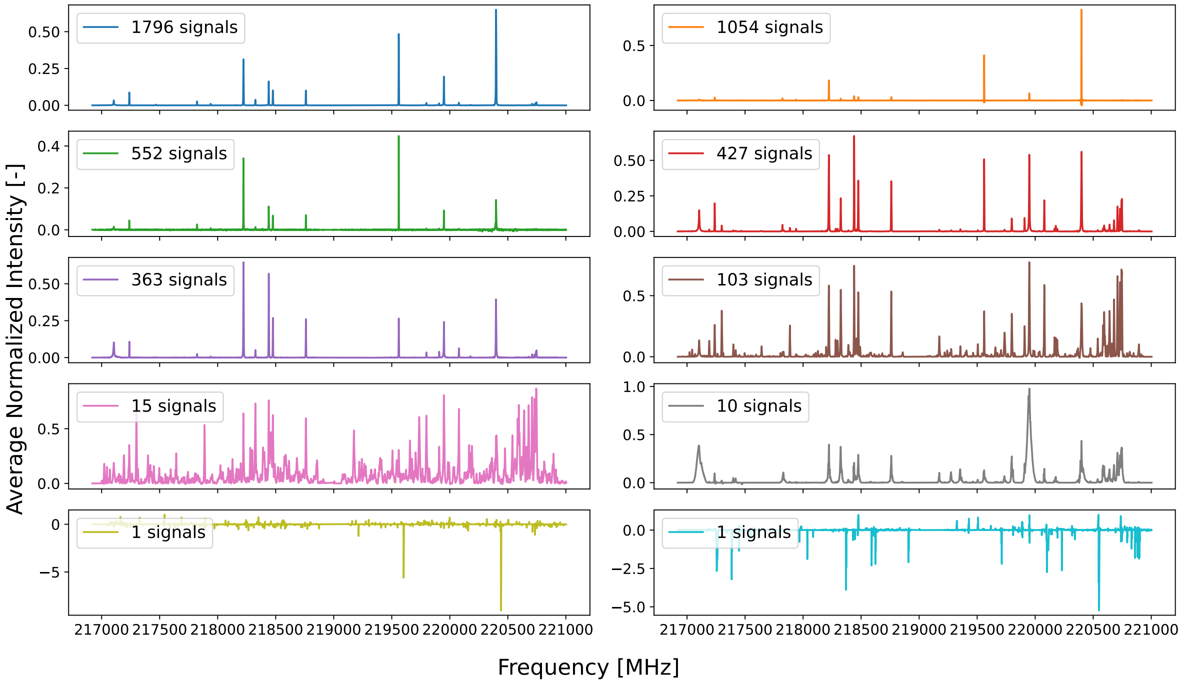
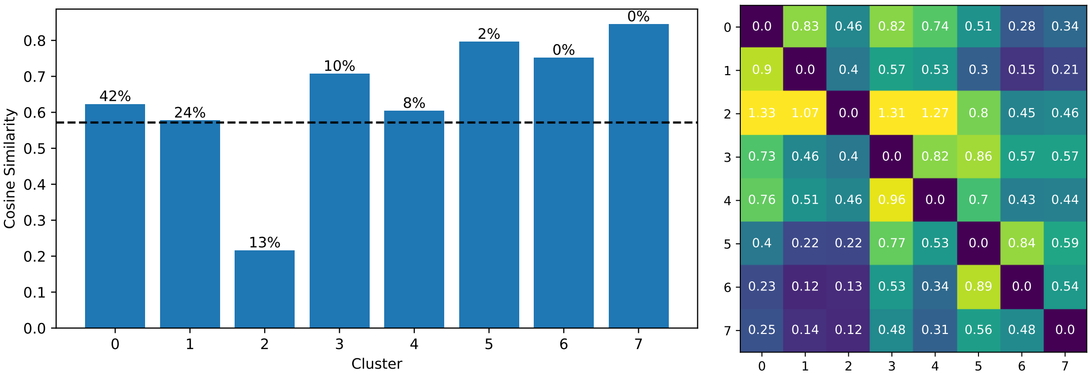


Figure 3.50: Centroids from the Euclidean-distance agglomerative clustering when setting the number of clusters to 10.

The clusters are clearly distinct between themselves, and the small clusters identified share very particular characteristics, making this a very promising cluster assignment. In order to verify this, the WCS and BCS of these clusters was calculated, as shown in Figure 3.51.



(a) WCS.

(b) BCS/WCS ratio.

Figure 3.51: WCS and BCS of the clusters shown in Figure 3.50. The last two clusters were excluded from the analysis due to them having only one signal each.

The weighted average WCS shown in Figure 3.51a is 0.572, due mainly to the low similarity of cluster 2. This cluster contains 13% of the signals but has a similarity of just 0.216, most likely due to the presence of signals with a low SNR. This can be seen in Figure 3.52, which shows the density distribution of the SNR of the signals that form each cluster (excluding the last two clusters).

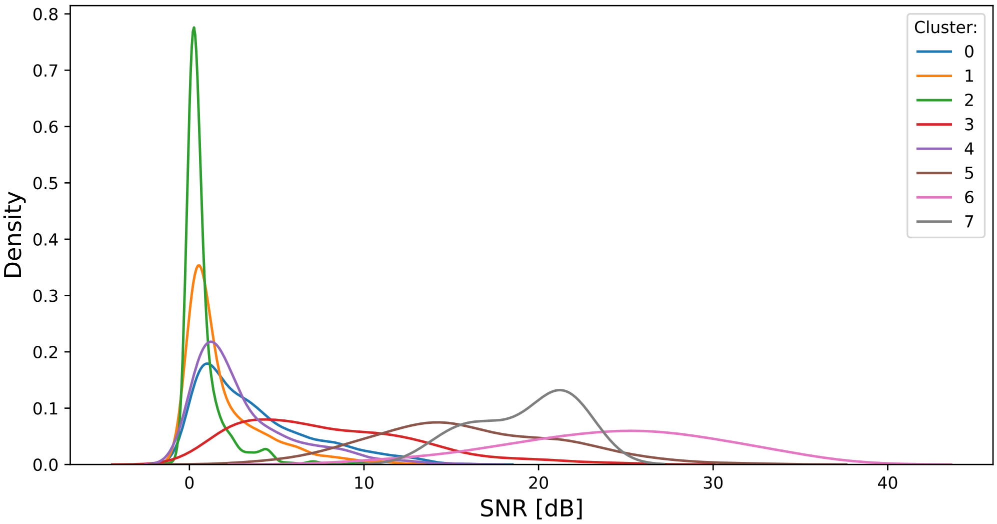


Figure 3.52: Signal-to-Noise Ratio (SNR) density distribution of the signals that form the first 8 clusters of Figure 3.50.

Based on these findings, a deeper study of cluster 2 was made. This analysis consisted on comparing the superposition of all signals forming the cluster to the superposition of only those signals with a SNR greater than 1 dB. The result is shown in Figure 3.53, where it can be clearly seen that the signals with a higher SNR are the ones that determine the outline of the average signal shown in green. It should be noted, however, that the signals with a SNR higher than 1 dB are only 114 out of the 552 signals that make up the original cluster.

Disregarding those signals from cluster 2 with a SNR lower than 1 dB, its WCS increases from 0.216 to 0.4173. This can be seen in Figure 3.54a, where cluster 2 has been re-labelled as cluster 4 due to its decrease in size. Since the new cluster only contains 3% of the total amount of signals, the new

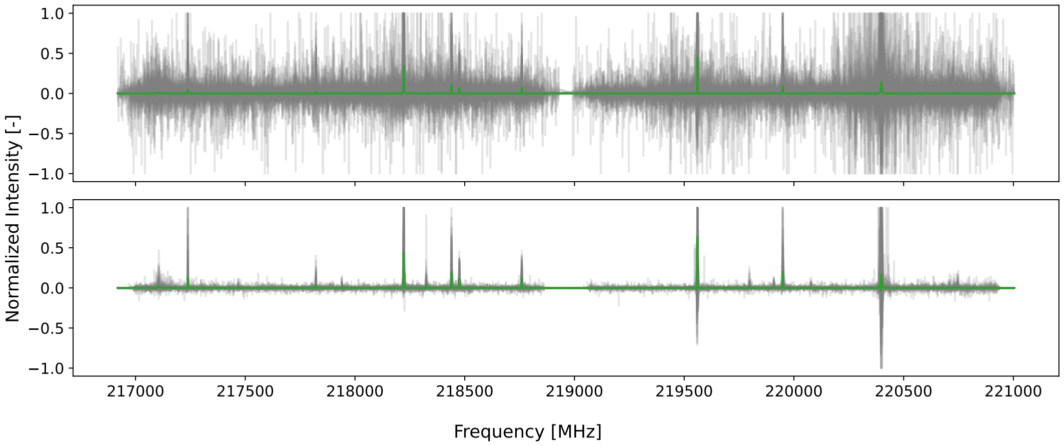
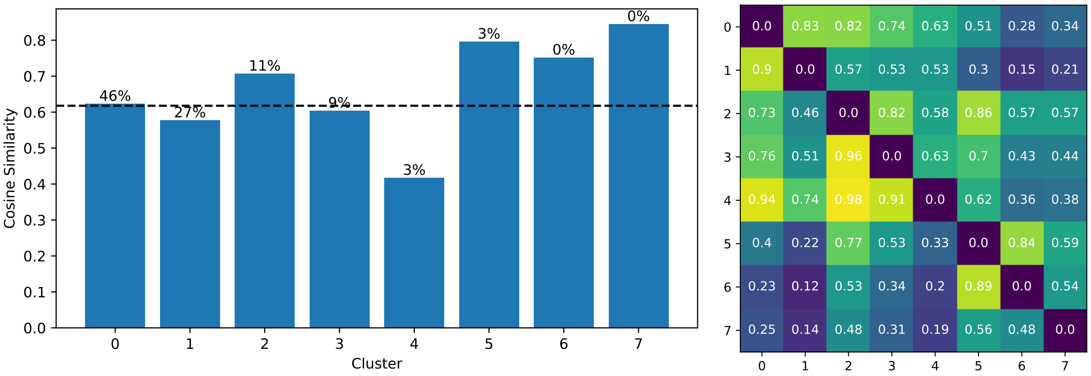


Figure 3.53: Comparison between all signals making up cluster 2 from Figure 3.50 (top) and only those signals with a SNR higher or equal to 1 (bottom).

weighted-average WCS is 0.618. The mean BCS/WCS ratio, on the other hand, has increased slightly, from 0.292 to 0.294.



	0	1	2	3	4	5	6	7
0	0.0	0.83	0.82	0.74	0.63	0.51	0.28	0.34
1	0.9	0.0	0.57	0.53	0.53	0.3	0.15	0.21
2	0.73	0.46	0.0	0.82	0.58	0.86	0.57	0.57
3	0.76	0.51	0.96	0.0	0.63	0.7	0.43	0.44
4	0.94	0.74	0.98	0.91	0.0	0.62	0.36	0.38
5	0.4	0.22	0.77	0.53	0.33	0.0	0.84	0.59
6	0.23	0.12	0.53	0.34	0.2	0.89	0.0	0.54
7	0.25	0.14	0.48	0.31	0.19	0.56	0.48	0.0

(a) WCS.

(b) BCS/WCS ratio.

Figure 3.54: WCS and BCS/WCS ratio of the clusters shown in Figure 3.50. The last two clusters were excluded from the analysis due to them having only one signal each, as well as the signals from cluster 2 (re-labelled as cluster 4 in this figure) with a SNR lower than 1 dB.

The WCS of cluster 4 is still quite low, however, the bottom panel of Figure 3.53 shows that the differences within the signals are now due to a difference in the magnitude of their peaks, rather than due to noise. This cluster was therefore accepted as-is, resulting in the centroids shown in Figure 3.55, which represent 89.82% of all signals. These signals were therefore discarded, leading to Figure 3.55.

The next clustering investigated was the one resulting from using the same model but stopping at 23 clusters. The centroids of these clusters can be seen in Figure 3.56. The figure shows how no clusters with more than 1000 signals had been formed at this stage, having instead 12 clusters with more than 100 signals. These clusters are made up of signals with few peaks, except for the last cluster which shows no pattern. Additionally, clusters 0 and 1 (the two largest clusters) show a great amount of similarity.

The small clusters, on the other hand, are repeated in three instances with the ones from Figure 3.50. These are the broad-peak cluster with 10 signals and the two clusters with just one signal, shown in Figure 3.56 in the last and third-to-last positions. The remaining small clusters all show many spectral lines, although with varying intensities between the lines. The only exception is the fifth small cluster (sixth row, middle column), which is characterised by having two strong absorption features next to both CO lines. The remaining small clusters all show many spectral lines, although with varying intensities between the lines. The only exception to this is the fifth small cluster (sixth row, middle column), which is characterised by having two strong absorption features next to both CO lines.

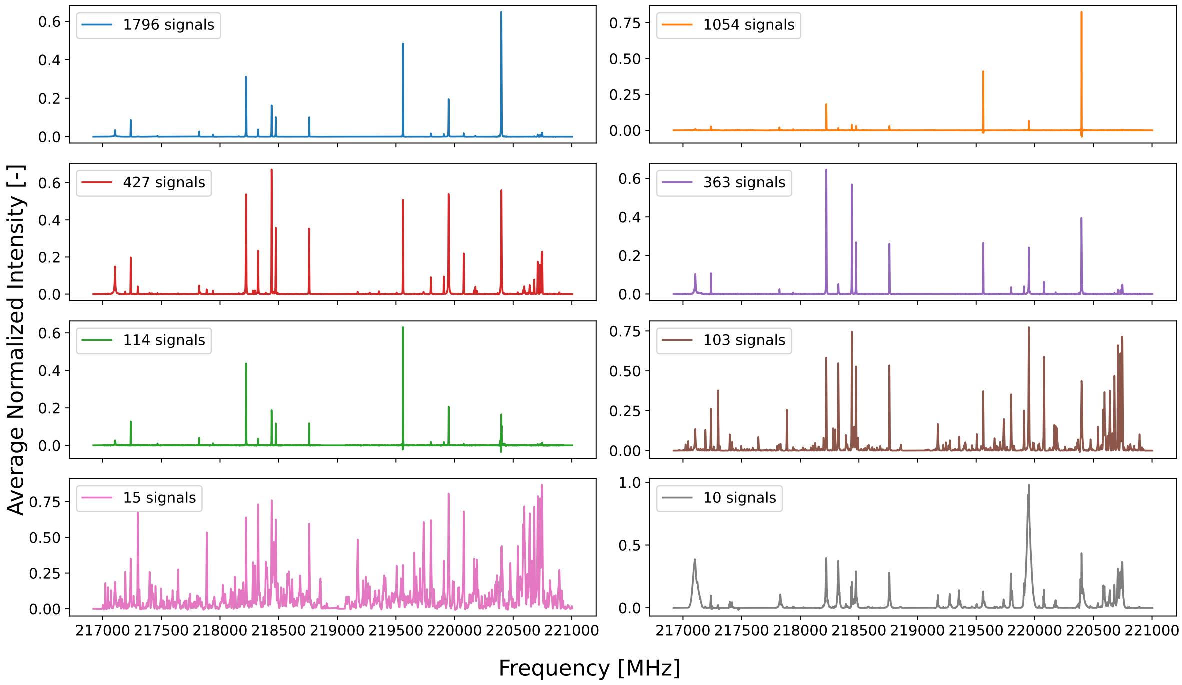


Figure 3.55: Centroids of the final cluster assignment from the Euclidean-distance agglomerative clustering model when set to 10 clusters.

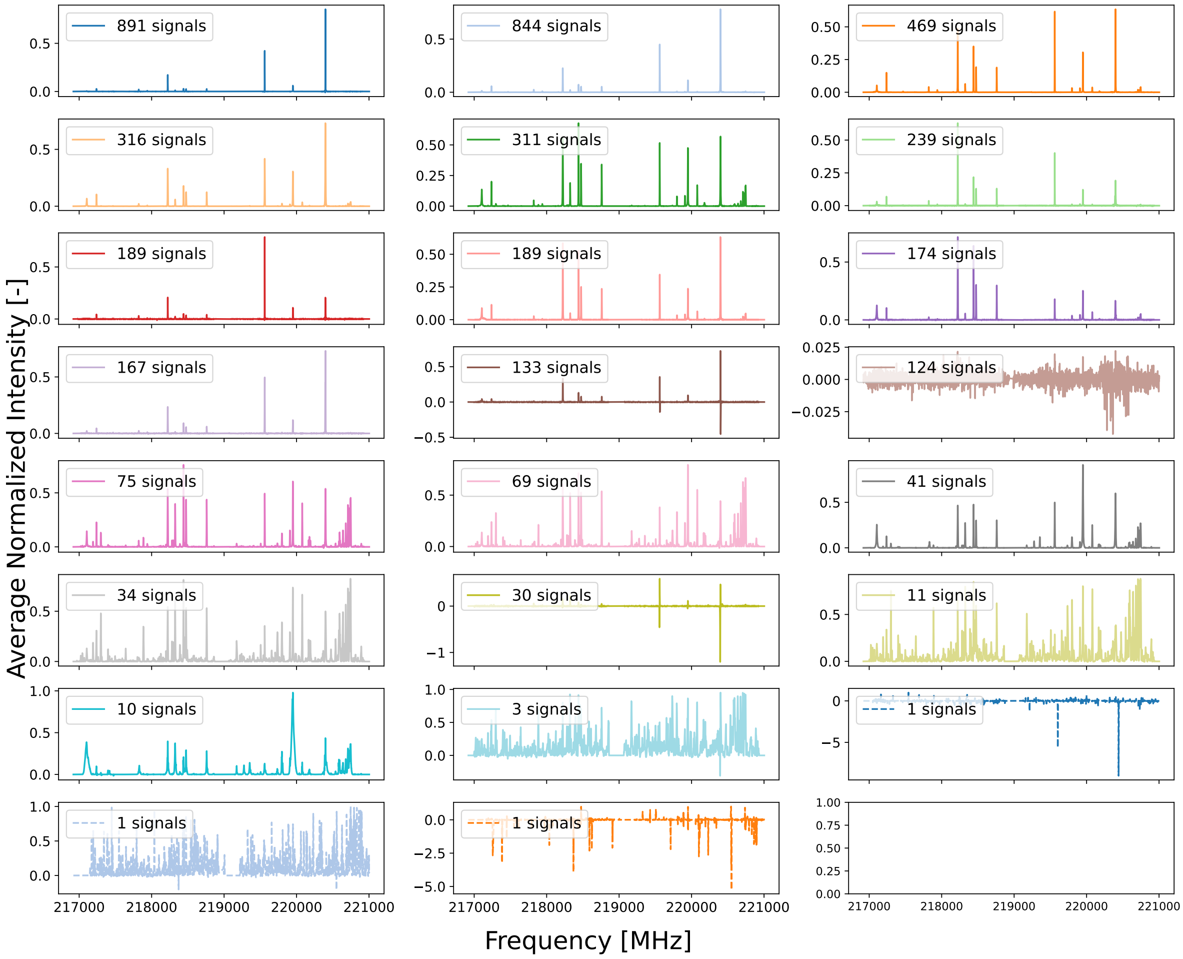
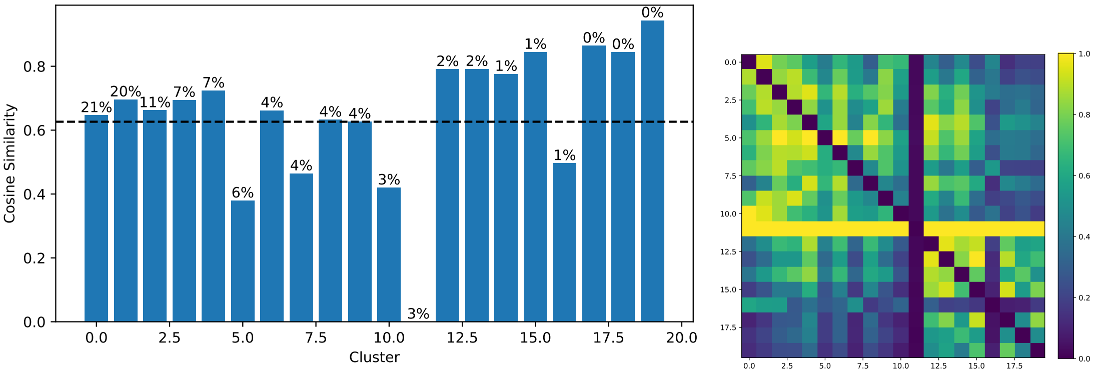


Figure 3.56: Centroids from the Euclidean-distance agglomerative clustering when setting the number of clusters to 23.

Analysing this cluster arrangement, a WCS of 0.626 is obtained after excluding the clusters with just one signal. This can be seen in Figure 3.57a, where it can be seen that the WCS of cluster 11 (fourth row, third column) is 0. The next two clusters with a significantly lower-than-average WCS are clusters 5 and 10, warranting a deeper analysis of these clusters. Looking at the BCS/WCS ratio, on the other hand, it can be seen how the large clusters share more overlap between them than with the smaller ones and vice-versa. The only exceptions to this norm are clusters 11 and 16, the former due to its null WCS skewing the ratio and the latter due to it not sharing similarity with any clusters.



(a) WCS.

(b) BCS/WCS ratio.

Figure 3.57: WCS and BCS/WCS ratio of the clusters shown in Figure 3.56. The clusters with only one signal have been excluded.

The maximum SNR found in the signals that make up cluster 11 (light brown) is 0.63 dB. This caused Figure 3.58 (which shows the distribution of the SNR of the signals that make up each cluster) to be unreadable, so it was excluded from the figure. Besides cluster 11, the next two clusters that show a greater concentration of signals with a low SNR are clusters 5 (light green) and 10 (dark brown), respectively. These two clusters are also the two clusters with the lowest SNR, other than cluster 11. The composition of these two clusters was therefore further investigated.

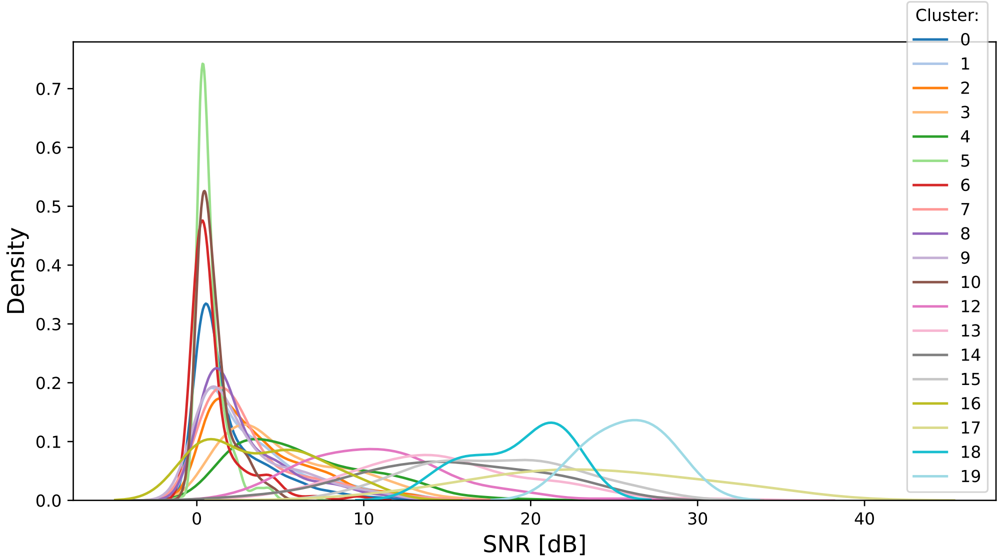


Figure 3.58: Signal-to-Noise Ratio (SNR) density distribution of the signals that form the clusters of Figure 3.56 with more than 1 signal.

The first cluster investigated was cluster 5 (light green). The top panel of Figure 3.59 shows all the signals that make up this cluster and its centroid, whereas the bottom panel shows only the signals with a SNR greater than 1 dB. This exclusion discarded 176 out of the 239 signals, however, the signal's outline is much more consistent. A very similar case can be seen in Figure 3.60, which shows the composition of cluster 10. In this case, 81 out of the 133 signals that make up the cluster where discarded.

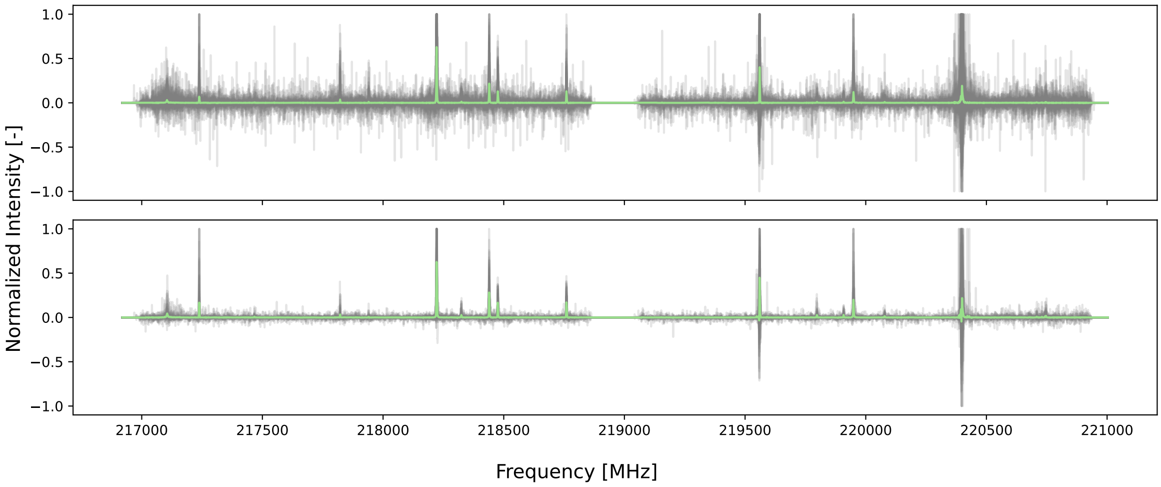


Figure 3.59: Comparison between all signals making up cluster 5 from Figure 3.56 (top) and only those signals with a SNR greater than 1 dB.

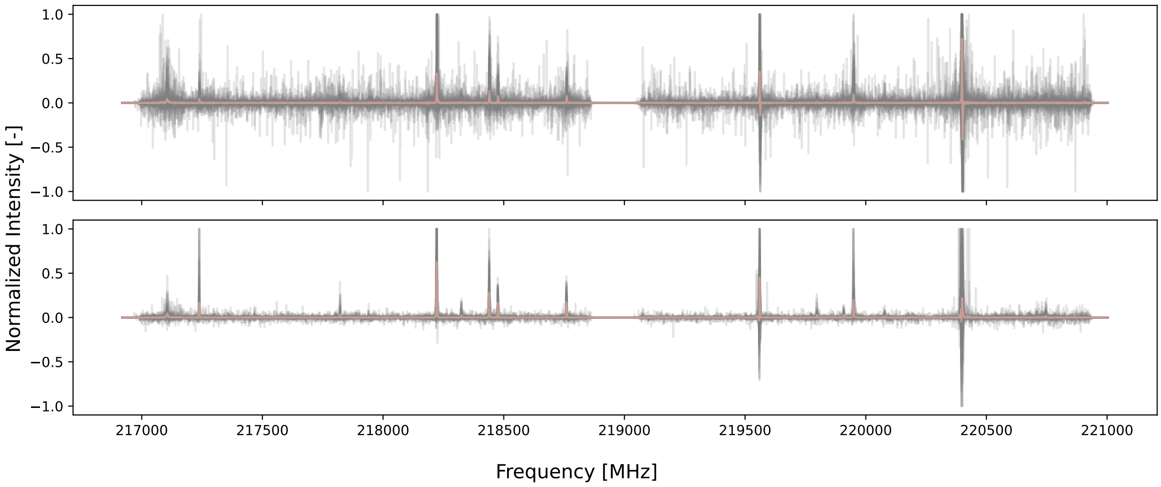


Figure 3.60: Comparison between all signals making up cluster 10 from Figure 3.56 (top) and only those signals with a SNR greater than 1 dB.

The new cluster centroids after discarding the conflictive signals can be seen in Figure 3.61. These centroids were made with 3938 out of the 4322 signals (91.15%). On the one hand, this is higher than when using the 10 clusters model. Furthermore, the weighted-average WCS (seen in Figure 3.62a) increased to 0.663, which is 0.045 points higher than in the previous model. On the other hand, the mean BCS/WCS ratio increased to 0.318 (from 0.298). The approach could have been made more reproducible by directly discarding all signals with a SNR lower than 1 dB. This would have resulted in a cluster arrangement with 18 clusters with a weighted-average WCS of 0.717. However, only 2819 signals (65.22%) would have been clustered. This was not a desirable outcome so, to preserve the reproducibility of the process, signals with a SNR lower than 1 dB were only discarded from clusters with a WCS lower than 0.4.

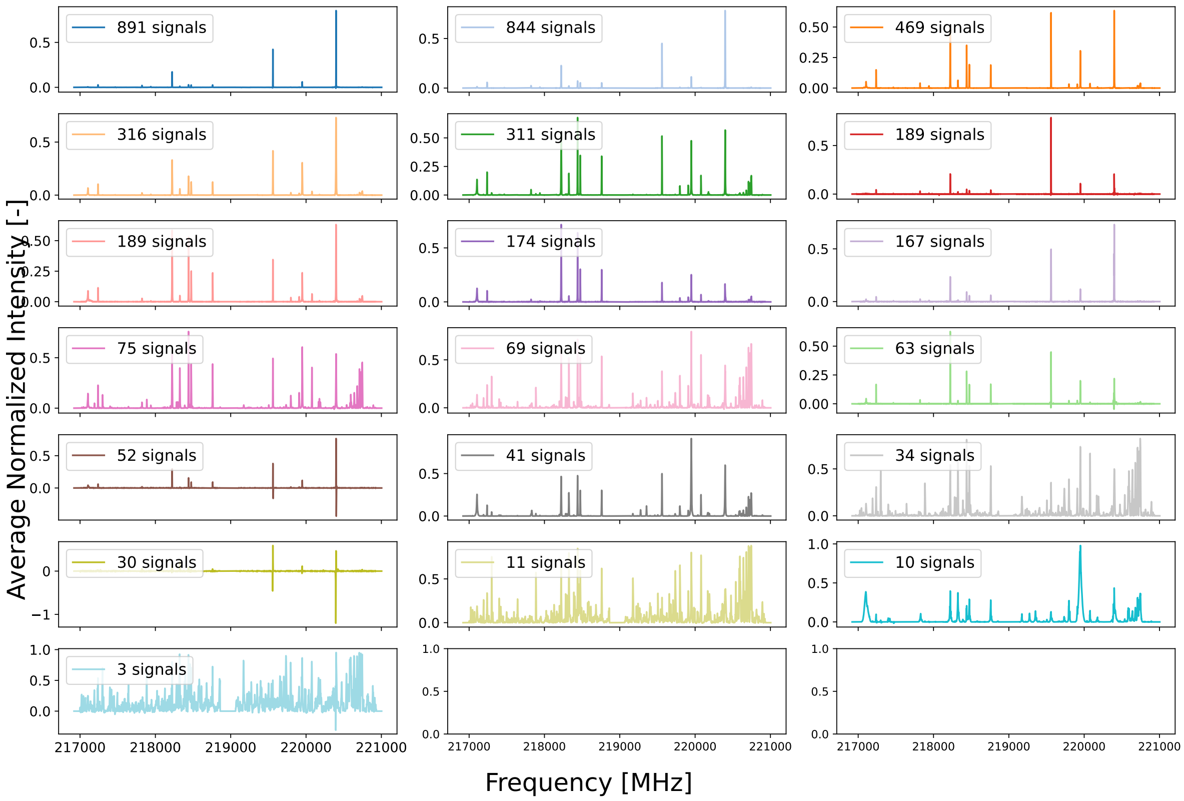
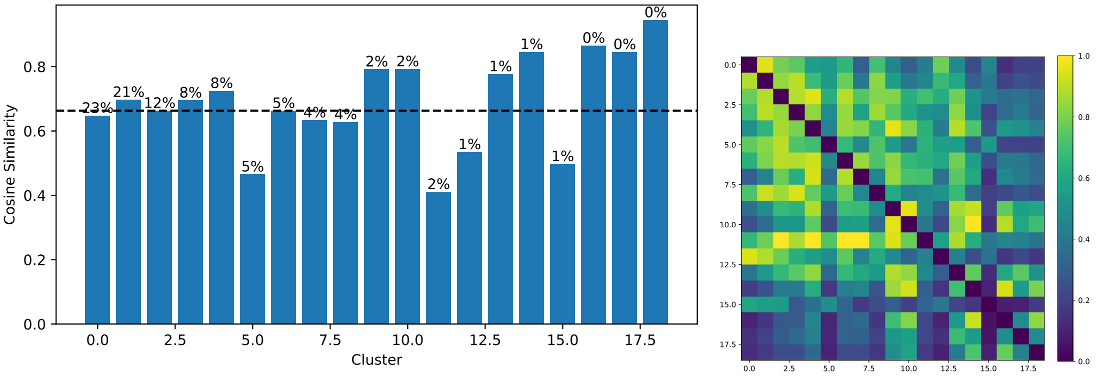


Figure 3.61: Centroids of the final cluster assignment from the Euclidean-distance agglomerative clustering model when set to 23 clusters.



(a) WCS.

(b) BCS/WCS ratio.

Figure 3.62: WCS and BCS/WCS ratio of the clusters shown in Figure 3.61.

3.3.3. Conclusion

This section showed the results for two different agglomerative clustering models. The first one was a neighborhood-based agglomerative clustering model, based on the work of Zhang et al. [71]. The second model, on the other hand, was a conventional agglomerative clustering model tested using both the cosine and Euclidean distance metrics.

The first model was originally tested on the subtraction-filtered dataset, however, the results were far from ideal. One large cluster and several small clusters were found irrespectively of the number of neighbors considered. Furthermore, the clusters showed several outliers throughout the merging process. The model was therefore trained on the UMAP projections of the subtraction-filtered signals. The results were much better, with a weighted-average WCS of 0.621 and 0.589 respectively. However, the BCS/WCS ratio of these clusters was higher than in previous models, with values of 0.418 in the Euclidean-distance projeciton and 0.435 in the cosine-distance one. Furthermore, the merging process was not fully reproducible, as it required a manual inspection of the merging steps.

The conventional model was then tested using the cosine distance metric, setting a distance threshold of 0.4 and using an average linkage. This resulted in 20 clusters with more than 10 signals, containing 81.24% of all signals. The weighted-average WCS of these clusters was 0.6936, however, most of the clustered signals (75%) belonged to just two clusters. This made the clustering assignment unsatisfactory, as the previous models had shown that there are significant differences between the signals that made up those two clusters.

Finally, the same model was tested using the Euclidean distance metric and the Ward linkage. Since this distance is not bounded like the cosine distance, a silhouette score analysis was performed to determine the optimal number of clusters. Two cluster arrangements were tested: one with 10 clusters and another with 23 clusters. The first arrangement contained two clusters with just one signal each, which were discarded as outliers. The weighted-average WCS of the remaining clusters was 0.618 after discarding the signals with a SNR lower than 1 dB from the clusters with a WCS lower than 0.4. This implied discarding 10.18% of the dataset, leaving the remaining 89.82% of the signals arranged into clusters with a BCS/WCS ratio of just 0.298. The 23-cluster arrangement followed this same procedure, which led to 91.15% of the dataset being clustered into 19 clusters with a weighted-average WCS of 0.663 and a mean BCS/WCS ratio of 0.318.

A higher WCS (0.717) could have been achieved by directly discarding all signals with a SNR lower than 1 dB. This, however, would have resulted in only 65.22% of the dataset being clustered.

Overall, this model yielded the best results out of all the tested models. The k-means classifier achieved the same WCS (0.663), however, its mean WCS/BCS ratio was 0.3749, which is 17.89% higher than the highest one obtained using agglomerative clustering. On the other hand, the GMM applied to the Euclidean-distance UMAP projection yielded a WCS 7.42% lower than the one obtained in the 19-cluster arrangement, although with a BCS/WCS ratio was significantly lower. However, the 8-cluster arrangement has comparable a WCS and BCS/WCS ratio, all the while clustering an additional 33.53% of the dataset. This makes these last two arrangements the best and most balanced ones so far, making them the best candidates for the analysis carried on in Chapter 4.

Table 3.3: Summary of the agglomerative clustering models' research.

Aspect	Finding	Comment
Neighborhood-based	Unsuccessful	The signals in this dataset are not distinct enough for the number of neighbors to provide an appropriate stopping point.
Cosine-distance	Unsuccessful	20 clusters with more than 10 signals were found, which contained 81.24% of all data. Two of these clusters, however, contained 75% of all clustered signals. This indicated that the clusters could have been broken down into smaller components, leading to an overall unsuccessful clustering arrangement.
Euclidean distance	Successful	Two final arrangements were detected: one with 8 clusters and one with 19 clusters. The first arrangement has a WCS of 0.618, containing 89.82% of the dataset and with a BCS/WCS ratio of 0.298. The second one, on the other hand, has a WCS of 0.663, a mean BCS/WCS ratio of 0.318 and contains 91.15% of the dataset.
Reproducibility	Complete	The last clustering arrangement is completely reproducible, as no manual selection of clusters was made.

4
Analysis

This chapter will make use of the final two cluster arrangements found in Chapter 3 and the physical properties of the cores to answer the second set of research questions. The chapter begins by studying each cluster's chemical composition and the relationship between the number of lines in a cluster and that cluster's size. Then, the presence of HII regions throughout the clusters is studied, together with differences in the chemical composition of those signals. Then, the relationship between the number of cores in each region and the different clusters found within is studied. Finally, section 4.3 studies the distribution of the different physical properties throughout each cluster and proposes an evolutionary sequence.

4.1. Cluster's chemical composition

4.1.1. 8-cluster arrangement

To analyse the composition of these clusters, two metrics will have to be taken into account. Firstly, the mean signal of each cluster (also referred to as cluster centroid) will show which peaks are most common across all signals. Secondly, the chanel-by-chanel product of all signals will show which peaks are common to all signals in the cluster. Both of these metrics can be seen in Figure 4.2 for all clusters. The figure shows on the mean (top) and product (bottom) of the signals in each cluster, revealing that the largest cluster has no peaks common to all signals within them, and that all others have just one common peak. The only exceptions are the last two clusters, which have 4 and 2 peaks respectively.

Moving on to the mean signals, it is clear that there is an inverse correlation between the number of spectral lines and the size of the cluster. This can be better visualized in Figure 4.1, where the number of spectral lines was determined by the amount of features with a magnitude greater than 1% of the magnitude of the ^{13}CO line. The figure also shows an exponentially decaying curve fit with an R^2 value of 0.776.

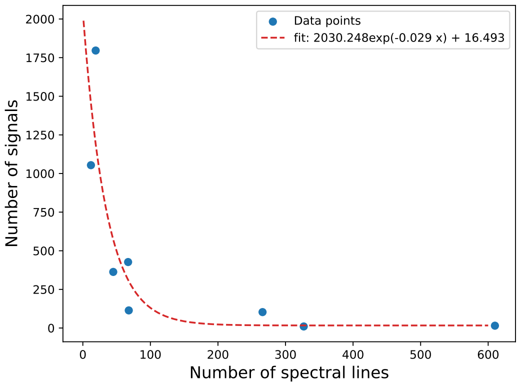
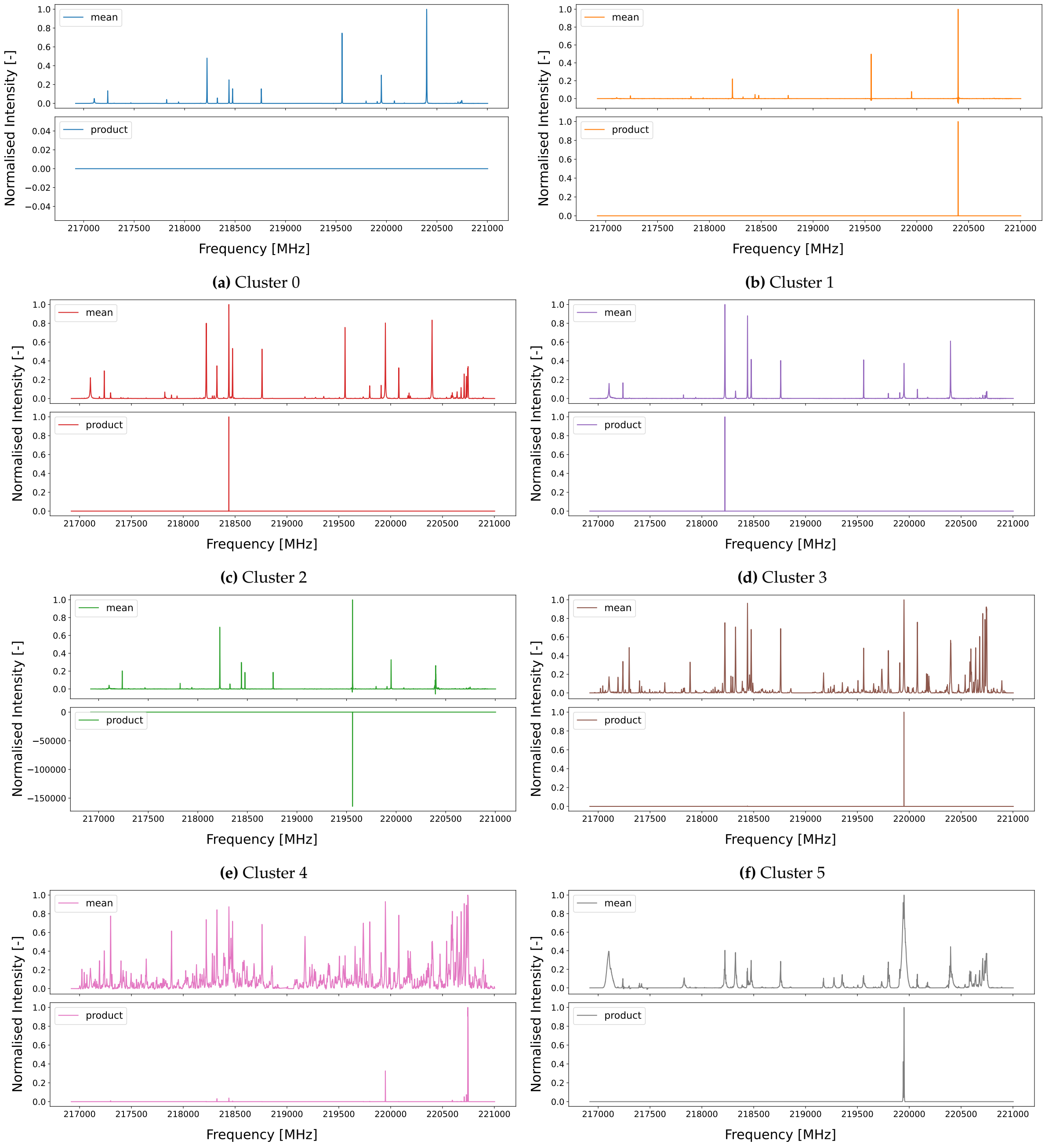


Figure 4.1: Relationship between the number of spectral lines in a cluster's centroid and the size of the cluster.

4.1.2. 19-cluster arrangement

The same analysis was then performed for the 19-cluster arrangement. Once more, as seen in Figure 4.3, the product of all signals in a cluster yielded a null value for the largest clusters. Additionally, smaller clusters with few peaks such as cluster 6 or cluster 8 only have the ^{13}CO peak as common to all signals. This, however, is not the case for the small clusters with many peaks, such as cluster 13 or cluster 17, whose common peak is the SO line.



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

(d) Cluster 3

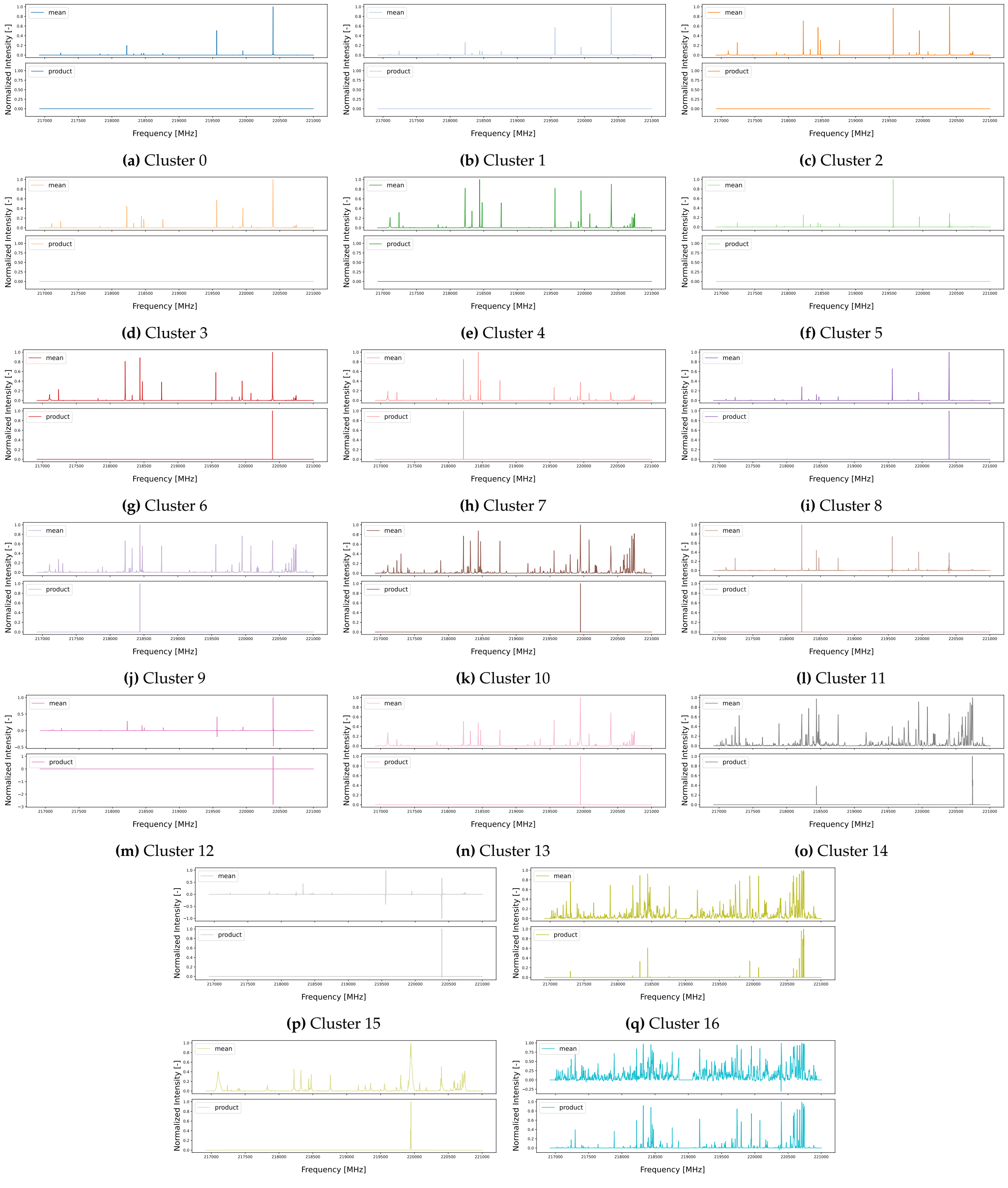
(e) Cluster 4

(f) Cluster 5

(g) Cluster 6

(h) Cluster 7

Figure 4.2: Normalized mean and product of the signals from each cluster from the 8-cluster arrangement.



(a) Cluster 0

(b) Cluster 1

(c) Cluster 2

(d) Cluster 3

(e) Cluster 4

(f) Cluster 5

(g) Cluster 6

(h) Cluster 7

(i) Cluster 8

(j) Cluster 9

(k) Cluster 10

(l) Cluster 11

(m) Cluster 12

(n) Cluster 13

(o) Cluster 14

(p) Cluster 15

(q) Cluster 16

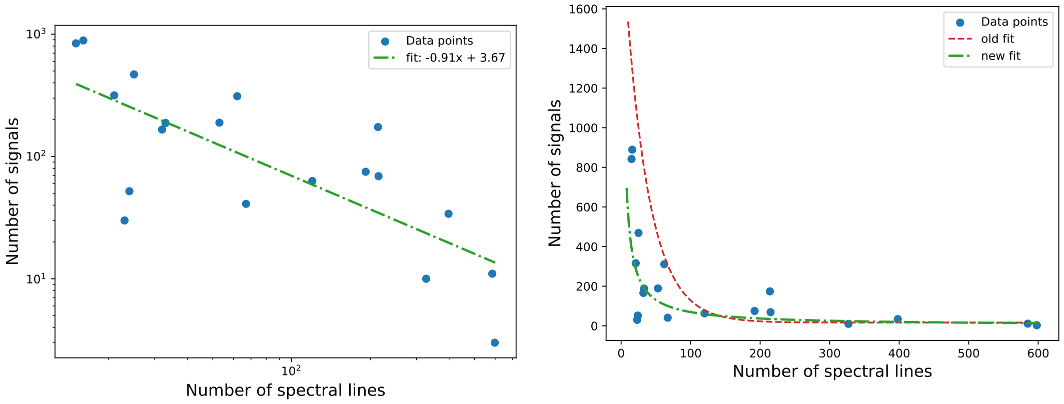
(r) Cluster 17

(s) Cluster 18

Figure 4.3: Normalized mean and product of the signals from each cluster from the 19-cluster arrangement.

On the other hand, analyzing the mean signal of each cluster revealed slight differences. With these labels, a larger proportion of the total clusters had less than 100 lines, compared to the previous assignment. Furthermore, the total amount of signals per cluster decreased since the largest cluster in this arrangement is less than half the size of the previous largest cluster. These two factors combined caused the old fit to overestimate the number of signals for clusters whose centroid had less than 150 lines, as seen in Figure 4.4b. This new fit, which follows Equation 4.1, has an R^2 value of 0.570 as a consequence of the additional data-points, which reduce the quality of the fit.

$$\log(\text{n\_signals}) = -0.912 \log(\text{n\_lines}) + 3.666 \tag{4.1}$$



(a) Relationship between the number of spectral lines in a cluster's centroid and the size of the cluster in logarithmic space.

(b) Comparison between the old and new relationship between the number of spectral lines in a cluster's centroid and the size of the cluster in the nominal space.

Figure 4.4: Number of signals in a cluster as a function of the number of spectral lines in the cluster's mean using the 19-cluster assignment.

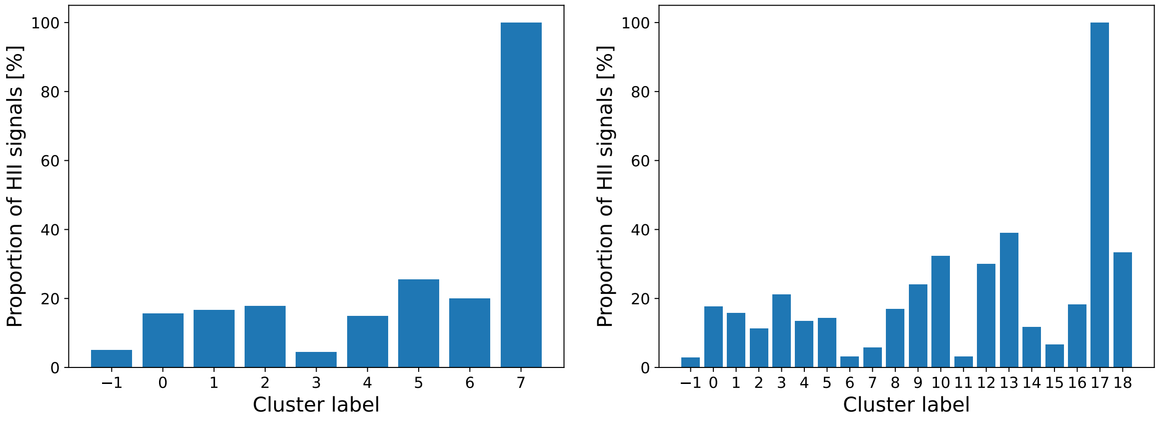
The size of the sample and the number of regions where the signals originate is statistically significant. It is therefore reasonable to assume that there should be no bias towards a given evolutionary stage within the data. Both cluster assignments, however, show that signals with few spectral lines are much more abundant than signals with many features. This suggests that the transition between evolutionary stages accelerates as the core collapses. This is in agreement with the timelines proposed in literature, where the cold-core phase of the star-forming region is significantly longer than the hot-core phase [42, 46].

4.2. Cloud-core correlation

4.2.1. HII region

Some of the collected signals originate at an HII region or near one. Specifically, 623 out of the 4298 signals were marked as possibly contaminated by an HII region. Note that this does not imply that those signals originate within an HII region, as an HII region being present in the section of the sky analysed will lead to all signals in that section being marked as “potentially contaminated". Figure 4.5 shows how these signals are distributed along all clusters, including the unclassified signals. In both label assignments, the cluster formed by the signals with the broad peaks (cluster 7 in the 8-cluster arrangement, and cluster 17 in the 19-cluster arrangement) was entirely formed by HII-contaminated signals. Beyond this, however, there is no clear trend between the centroids with a higher concentration of HII-contaminated signals and those without it.

The clusters were then inspected to find differences between the HII-contaminated signals and the rest. This investigation, however, proved to be mostly unsuccessful. This was due to the clusters having been formed based on their chemical similarity, which led to very few differences being found. In the 8-cluster arrangement, the only clusters to show differences were clusters 2 and 3, shown in Figures 4.6 and 4.7. The figures show the centroids of the cluster with all its signals (top panel), with only the signals without HII contamination (middle panel), and with only the signals potentially contaminated by HII



(a) 8-cluster arrangement.

(b) 19-cluster arrangement

Figure 4.5: Proportion of signals in each cluster that come from an HII region in each cluster arrangement.

regions (bottom panel). In all cases, the centroids were normalized by the intensity of the ^{13}CO signal. No clear patterns can be discerned among the different clusters. This is because the HII-contaminated signals in cluster 2 feature a weaker spw0 compared to their non-contaminated counterpart, whereas the opposite is true in cluster 3. The only common feature between both is an increase in the relative intensity of the SO peak, although this difference is marginal in cluster 3. Both of these patterns can also be seen in the 19-cluster arrangement, specially of clusters 4, 7 and 9, shown in Figures 4.8-4.10.

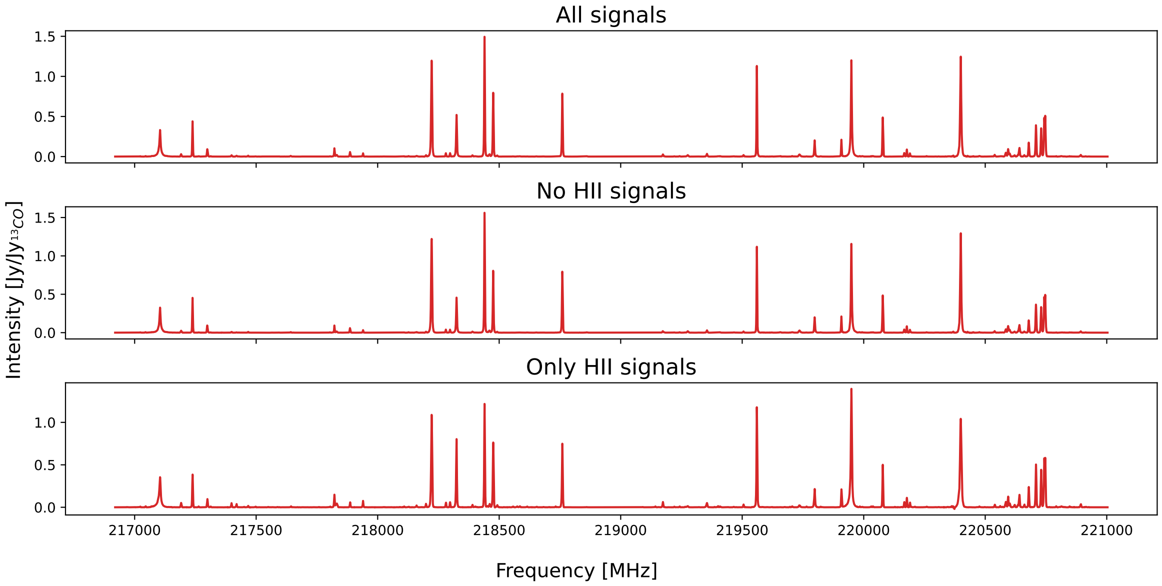


Figure 4.6: Centroid comparison from cluster 2 of the 8-cluster arrangement when considering all signals (top), only signals without HII contamination (middle), and only signals with potential HII contamination (bottom).

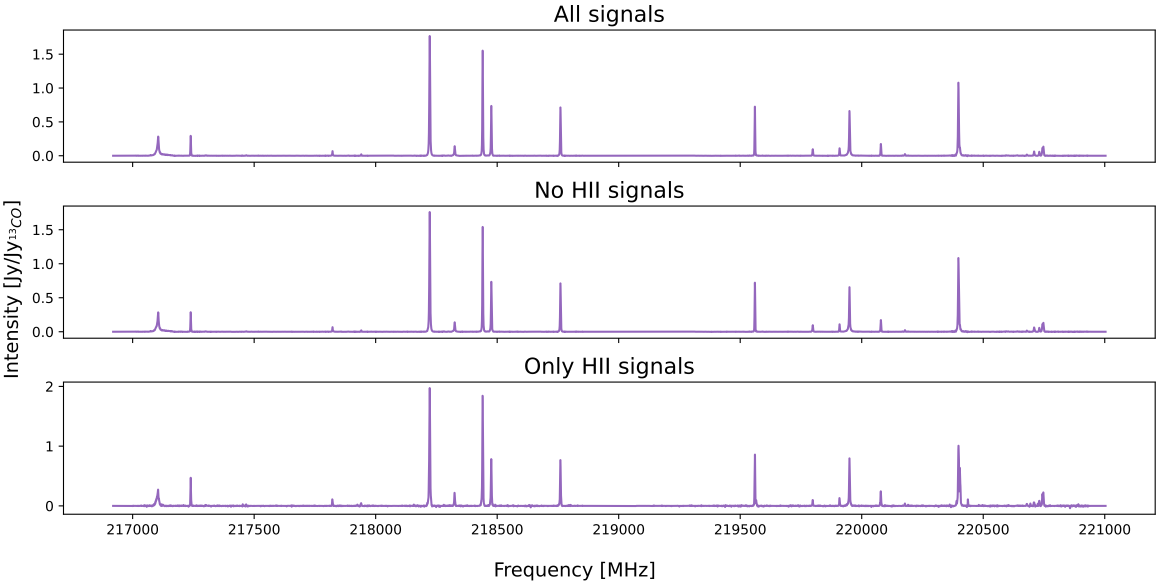


Figure 4.7: Centroid comparison from cluster 3 of the 8-cluster arrangement when considering all signals (top), only signals without HII contamination (middle), and only signals with potential HII contamination (bottom).

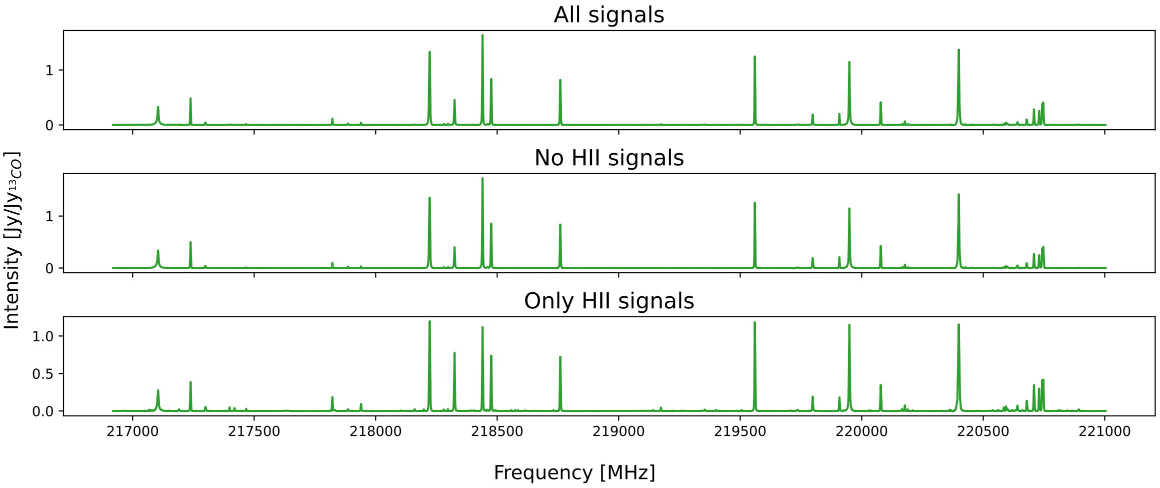


Figure 4.8: Centroid comparison of cluster 4 from the 19-cluster arrangement when considering all signals (top), only signals without HII contamination (middle), and only signals with potential HII contamination (bottom).

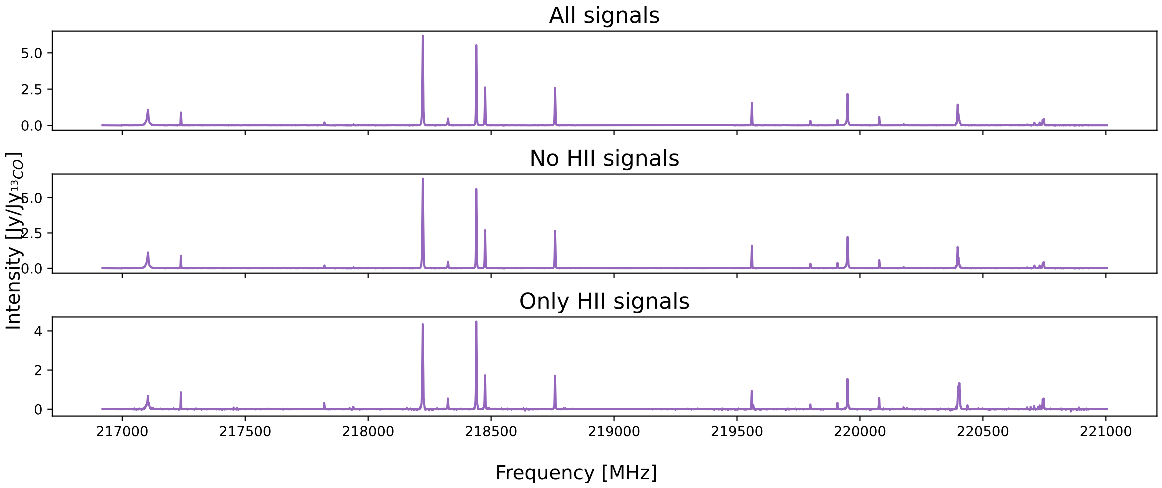


Figure 4.9: Centroid comparison of cluster 7 from the 19-cluster arrangement when considering all signals (top), only signals without HII contamination (middle), and only signals with potential HII contamination (bottom).

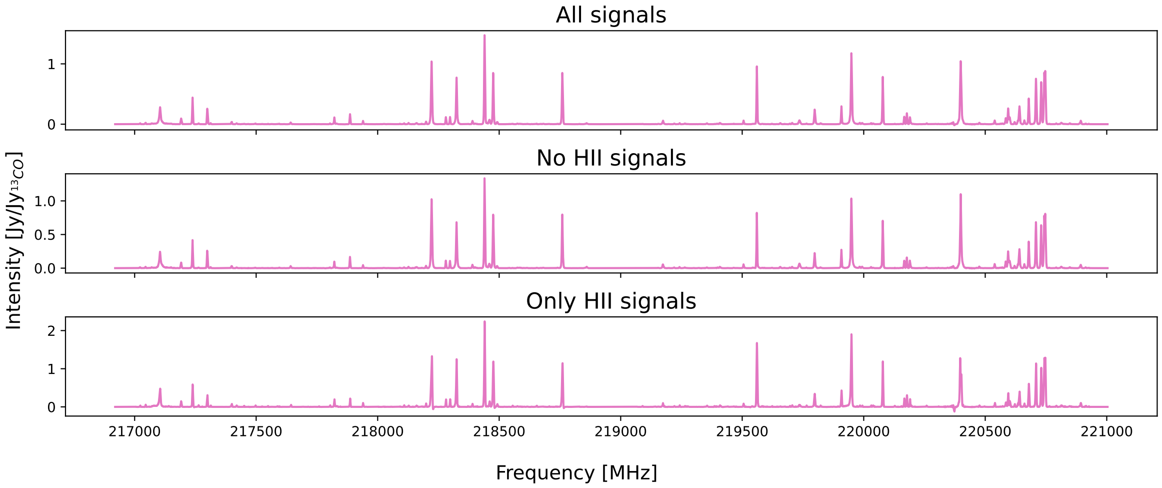


Figure 4.10: Centroid comparison of cluster 9 from the 19-cluster arrangement when considering all signals (top), only signals without HII contamination (middle), and only signals with potential HII contamination (bottom).

4.2.2. Clusters per region

The next parameter verified was the correlation between the number of cores and the number of clusters present in each region. According to the competitive accretion (CA) model, the cores at the center of the cloud will develop faster than those at the edges [7]. Assuming that each cluster represents a different evolutionary stage, the number of clusters present in each region should be positively correlated to the number of cores. This proved to be true in the regions surveyed by ALMAGAL, as both variables have a Pearson correlation coefficient of 0.683 when using the 8-cluster arrangement. This relationship can be observed in Figure 4.11, where the trend line clearly shows the positive correlation between both variables. Furthermore, the kernel density estimation (KDE) shown in the background highlights how most regions have few cores and follow a steeper correlation than the trend line. Additionally, the figure features an empty space in the lower right corner, indicating that no region with a significant amount of cores has less than three clusters present.

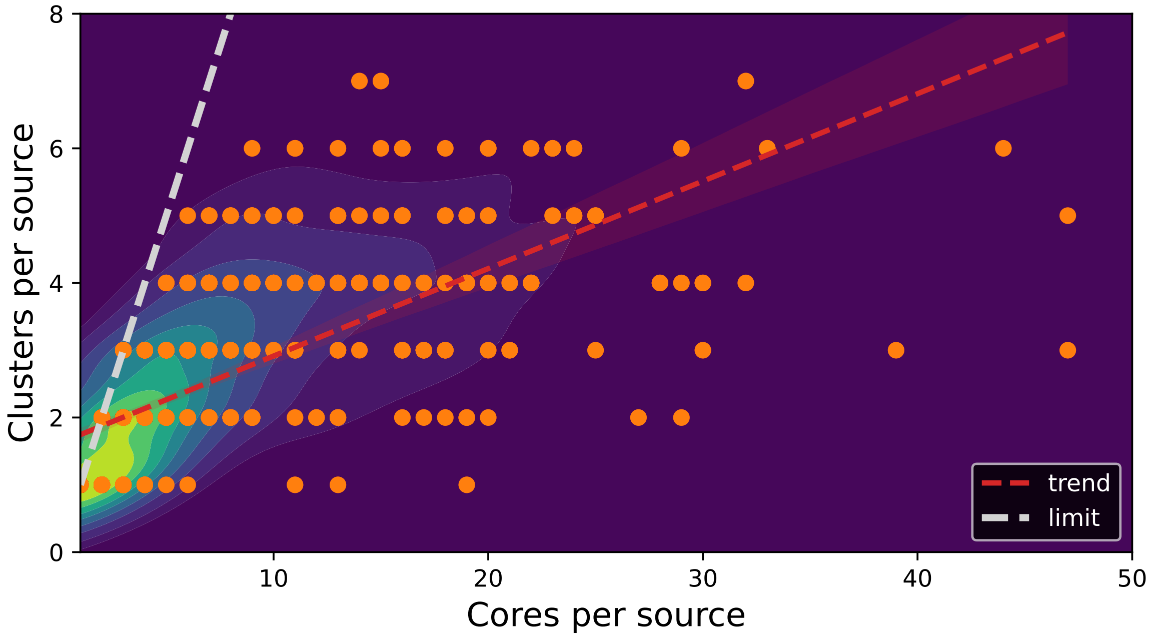


Figure 4.11: Relationship between the number of cores and the number of clusters identified in a region using the 8-cluster assignment.

The numerical correlation is even greater when considering 19 different clusters, where the Pearson correlation increases up to 0.806. This was to be expected, nevertheless, as increasing the possible amount of clusters without changing the possible amount of cores could only lead to an increase in the correlation coefficient. Despite this numerical bias, a similarly empty space can be seen at the bottom of Figure 4.12. This time, however, another empty space can be observed at the top of the figure, showing that there are no regions with more than 12 different clusters present. The reason for this cannot be known with the current data, although it could be possible that some evolutionary stages depend on the initial conditions of the cloud before the collapse. This would, therefore, make it impossible for a single cloud to form all 19 evolutionary stages.

Despite these correlations being in agreement with the CA model, obtaining such distributions could also be possible according to the Turbulent Core (TC) model. For this to be the case, however, the different stars forming in each region would have had to start forming at different times. Furthermore, the time between the start of each region's collapse would have to be comparable to the time taken for the region to evolve to the next evolutionary stage. This way, a region with multiple cores would display multiple evolutionary stages without there being any interaction between the different cores.

Further insights could be gained by investigating which clusters appear in the same regions. This information could then be matched to the evolutionary sequence proposed in section 4.3 to determine

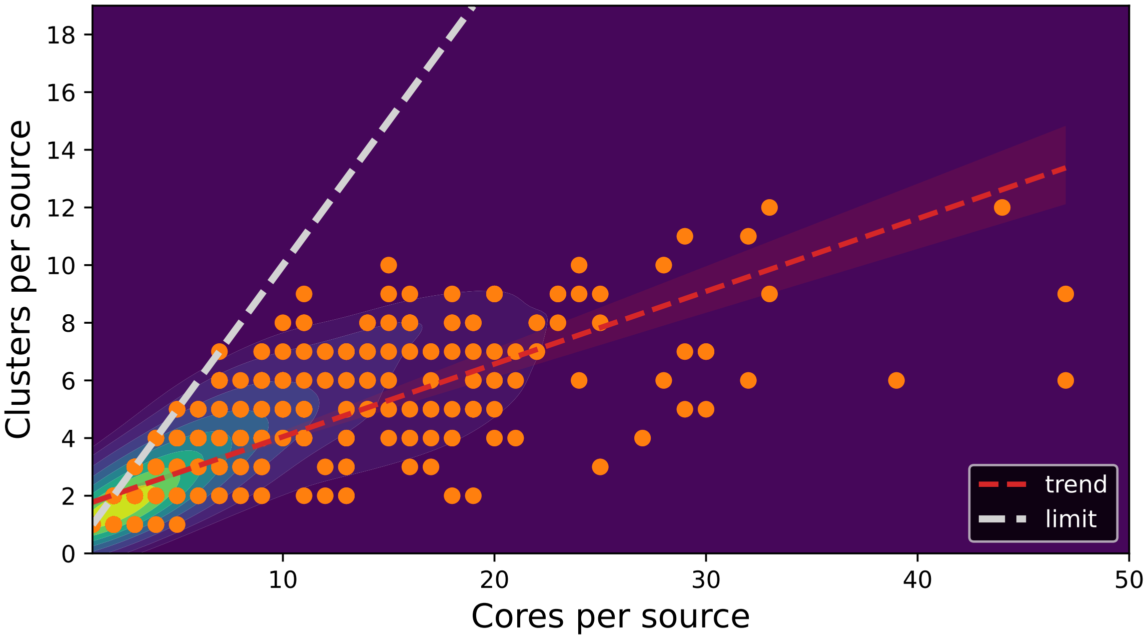


Figure 4.12: Relationship between the number of cores and the number of clusters identified in a region using the 19-cluster assignment.

which model is the data most consistent with. This analysis, however, is left for future work.

4.3. Physical properties of the clusters

In this section, the physical properties of the clusters found through the two different label assignments will be discussed. The objective was to associate each cluster to a different evolutionary stage by comparing properties such as the temperature of the cores assigned to each cluster. These properties, however, are still being studied, which posed limitations on the quality of the data available for this task. Table 4.1 details the different properties used in this study and explains each variable's limitations.

Table 4.1: Description of the physical properties used in the study and their limitations [14, 43].

Tag	Name	Description
Lclump/Mclump	Luminosity-to-mass ratio of the clump	Luminosity-to-mass (L/M) ratio of the region where the source is present. Expressed in $\mathrm{L}_{\odot}/M_{\odot}.Allsourcesinaregionsharethesamevalue.$
Tclump	Temperature of the clump	Average temperature of the region where the source is located. Value retrieved from Herschel observation or estimated based on the luminosity of the region. Expressed in K.
Tcore	Temperature of the core	Temperature of core where the source originates. Value estimated (with low accuracy) based on the luminosity of the region. Expressed in K.
Mclump	Mass of the clump	Mass of the region where the source is present. Expressed in $\mathrm{M}_{\odot}$.
Surfd_nd	Surface density of the dust	Particle density per unit area along the observation line. Expressed in g/cm^{-2}.
n(H2)	Volumetric density	Number of particles per cm^{-3} in the core. Value estimated based on core mass and integrated flux. Expressed in g/cm^{-3}.

Correlation to the number of lines per signal

These physical properties were first matched to the chemical richness of a signal. This was done by correlating the number of peaks in each signal to each physical property, resulting in Table 4.2. The table shows how all correlations to the number of lines are weak, with the volumetric density being the highest with a value of 0.452. This demonstrates that there is not a strong linear correlation between the variables and the chemical richness of a signal. There are, however, two factors affecting this outcome. Firstly, most of the variables were calculated for the entire clump, despite different evolutionary states coexisting within a single clump, as shown in subsection 4.2.2. Secondly, a low Pearson coefficient value does not imply that the variables are not correlated, but rather that the variables are not "linearly" correlated. This can be better seen in Figure 4.13, where each signal's number of lines and physical properties are plotted. In the figure it can be observed how, despite the correlation coefficient of 0.155, signals with a high number of lines always have a high luminosity-to-mass ratio.

Table 4.2: Pearson correlation coefficient of all physical properties and the number of lines in each signal.

	Nlines	Lclump/Mclump	Tclump	Tcore	Mclump	Surfd_nd	n(H2)
Nlines	1.000	0.155	0.234	0.202	0.060	0.183	0.452
Lclump/Mclump	0.155	1.000	0.706	0.823	-0.070	0.038	0.006
Tclump	0.234	0.706	1.000	0.880	-0.015	0.216	0.020
Tcore	0.202	0.823	0.880	1.000	-0.033	0.143	0.003
Mclump	0.060	-0.070	-0.015	-0.033	1.000	0.321	0.020
Surfd_nd	0.183	0.038	0.216	0.143	0.321	1.000	-0.008
n(H2)	0.452	0.006	0.020	0.003	0.020	-0.008	1.000

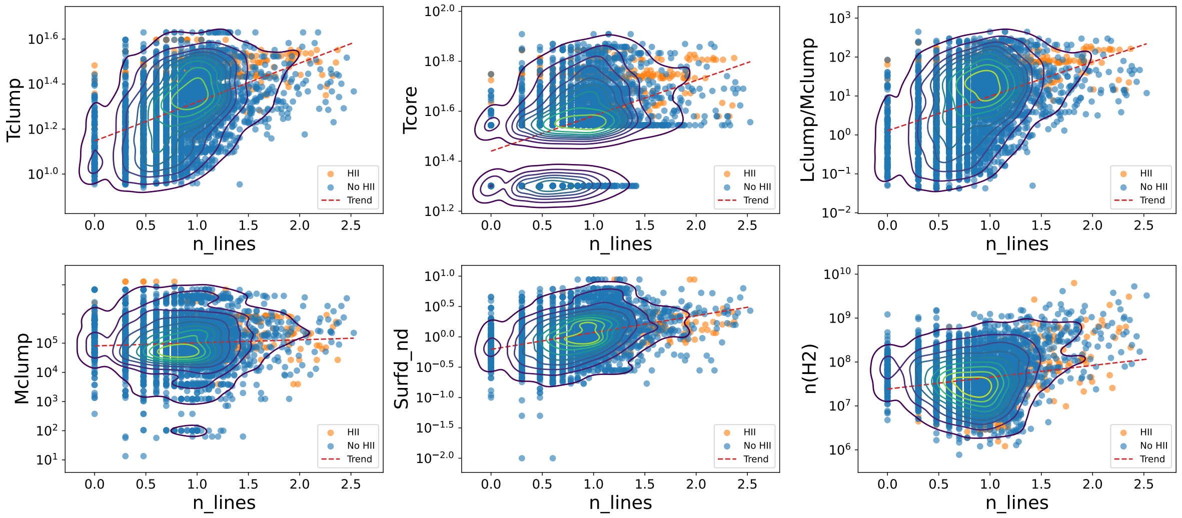


Figure 4.13: Correlation of the number of lines to the different physical properties of the cores.

Variable distributions within each cluster

Having already established that the observed variables are related to the number of lines, the next step was to analyse the distribution of the different variables within the clusters. This was first done on the 8-cluster arrangement, as the lower number of clusters would facilitate extracting conclusions based on the spectral line content. The numerical results for these distributions can be found in Table 4.3, where the 14th, 50th and 86th percentiles of each distribution are reported. These same values are shown in the vertical lines in Figure 4.14, where violin plots are used to show the entire distribution. Note that each violin plot in the figure is divided into two, with the left side showing the distribution of the signals without HII contamination, and the right side showing the distribution of the signals potentially contaminated by HII regions.

Table 4.3: Distribution of the physical properties of the clusters in the 8-cluster arrangement

(a) Clump temperature

#	Tclump (No HII)			Tclump (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	10.40	13.20	20.71	18.74	24.55	30.40
0	13.00	21.00	28.18	22.38	27.60	32.40
1	11.50	17.50	25.10	21.40	27.90	31.40
2	16.30	22.20	29.90	24.10	30.60	33.50
3	12.50	17.00	22.61	19.63	27.60	31.95
4	15.68	20.00	28.09	19.22	30.20	31.40
5	18.65	24.40	31.60	25.00	33.05	34.50
6	20.30	25.20	30.20	31.50	32.00	34.59
7	-	-	-	27.20	27.20	27.20

(b) Core temperature

#	Tcore (No HII)			Tcore (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	20.00	20.00	36.76	35.00	48.50	56.00
0	20.00	36.00	50.00	38.00	47.00	59.00
1	20.00	35.00	46.00	36.00	48.00	56.00
2	35.00	36.00	52.00	40.00	54.00	58.00
3	20.00	35.00	40.00	35.30	40.50	54.00
4	35.00	39.00	48.00	38.48	41.00	47.04
5	35.00	40.50	60.00	41.50	57.00	64.50
6	35.54	45.50	51.76	45.36	54.00	61.92
7	-	-	-	55.00	55.00	55.00

(c) Luminosity-to-mass ratio

#	Lclump/Mclump (No HII)			Lclump/Mclump (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	0.15	0.88	12.26	9.17	42.69	83.28
0	0.74	10.82	52.10	14.67	38.33	106.91
1	0.34	4.18	34.59	11.44	41.38	83.28
2	2.12	10.82	60.06	17.70	71.42	102.18
3	0.47	3.43	17.93	8.50	19.86	71.42
4	1.43	15.71	41.35	15.67	20.69	38.10
5	4.23	19.65	114.47	21.73	92.59	158.29
6	11.24	33.91	61.59	36.00	71.42	137.09
7	-	-	-	75.20	75.20	75.20

(d) Number of lines

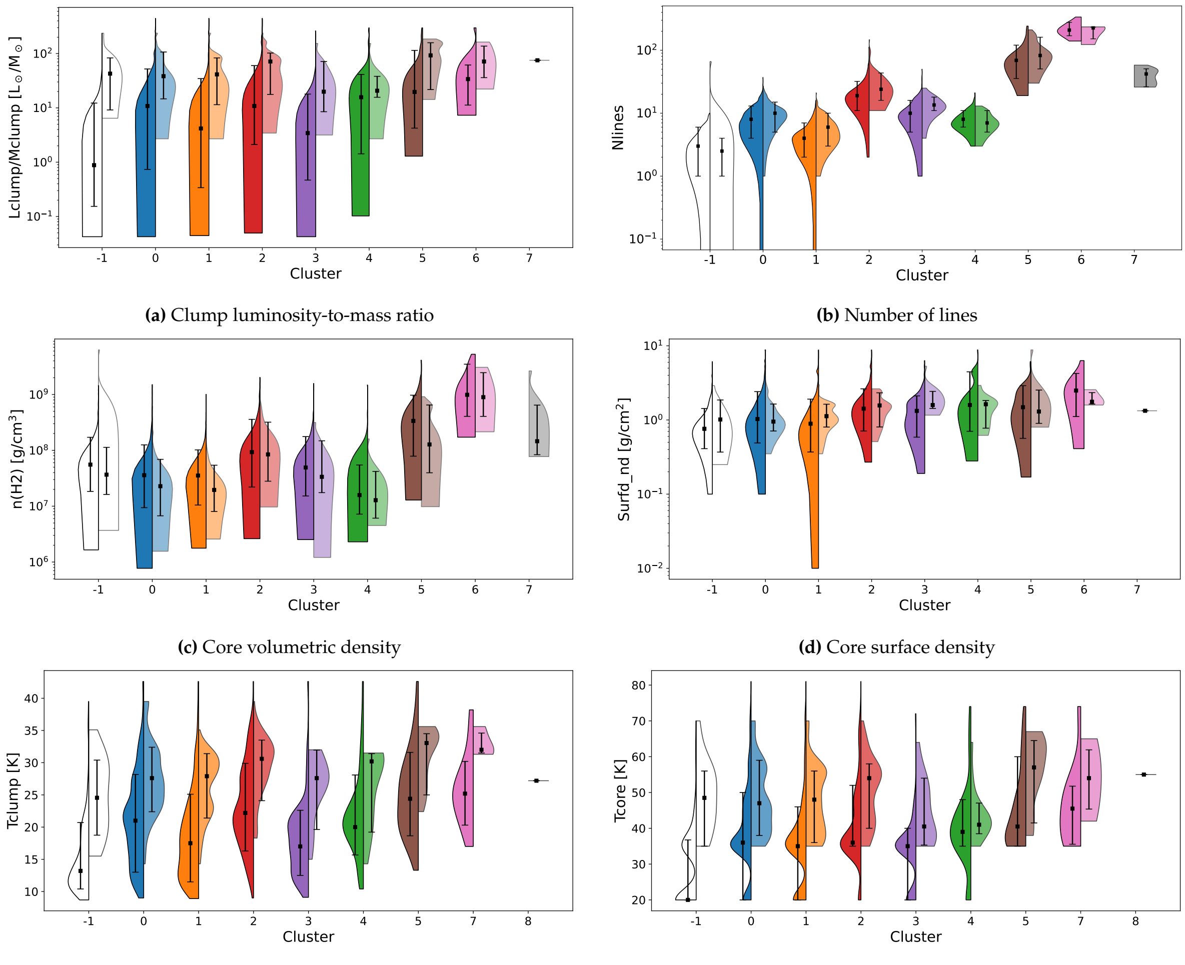
Cluster	Nlines (No HII)			Nlines (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	1.00	3.00	6.00	1.00	2.50	4.00
0	4.00	8.00	13.00	5.00	10.00	15.00
1	2.00	4.00	7.00	3.00	6.00	10.00
2	11.00	19.00	32.00	16.00	24.00	43.50
3	5.00	10.00	16.00	11.00	13.50	18.00
4	6.00	8.00	11.00	5.00	7.00	11.00
5	35.50	69.00	120.50	50.50	82.50	161.00
6	169.80	209.00	277.74	151.84	226.00	232.48
7	-	-	-	26.26	42.00	50.70

(e) Surface density

Cluster	Surfd$_nd(NoHII)$			Surfd$_nd(OnlyHII)$		
	P14	P50	P86	P14	P50	P86
-1	0.41	0.76	1.43	0.37	1.02	1.86
0	0.49	1.03	2.41	0.71	0.95	1.64
1	0.37	0.89	1.91	0.80	1.12	1.63
2	0.71	1.42	2.63	0.81	1.57	2.33
3	0.59	1.33	2.11	1.43	1.59	2.43
4	0.70	1.59	4.46	0.77	1.63	1.83
5	0.57	1.48	2.90	0.90	1.30	2.53
6	1.11	2.49	4.24	1.64	1.76	2.34
7	-	-	-	1.33	1.33	1.33

(f) Volumetric density

Cluster	n(H2) (No HII)			n(H2) (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	18.37	55.50	171.28	16.22	36.70	113.18
0	9.40	35.90	125.44	6.73	22.80	69.05
1	10.48	35.40	102.00	8.04	19.55	54.17
2	22.00	93.00	359.00	28.00	84.60	319.50
3	15.25	49.30	176.40	17.47	33.55	148.30
4	7.19	15.80	54.73	6.09	12.80	41.97
5	78.55	336.50	971.50	39.65	127.50	649.00
6	406.14	987.50	3490.00	405.68	896.00	2461.28
7	-	-	-	83.49	145.70	645.58



(a) Clump luminosity-to-mass ratio

(b) Number of lines

(c) Core volumetric density

(d) Core surface density

(e) Clump temperature

(f) Core temperature

Figure 4.14: Distribution of the physical properties of the clumps/cores along the different clusters identified in the 8-cluster arrangement. In all sub-figures, the left violin plot represents the data without HII region contamination, whereas the right one contains only the data with HII contamination for each cluster.

Several patterns that could lead to the establishment of an evolutionary sequence can be found analyzing these variables. The strongest indicator of a core's evolutionary stage, the luminosity-to-mass (L/M) ratio, was the first to be analysed. The first thing one may notice is that signals potentially contaminated by HII regions have a significantly higher L/M ratio in all clusters. This is specially the case for the unclassified signals (cluster -1) where the difference between the means is of 41.81 $L_{\odot}/M_{\odot}$. A possible explanation for this phenomenon is that the unclassified HII-contaminated signals do not actually belong to an HII region but are simply near one. Since the L/M ratio uses clump parameters, the ratio for these signals would be the same as that of the signals from the HII regions, while still having the chemistry of a cold core. This could also be the case for clusters 5-7, which would explain the high abundance of molecules in these clusters. The rest of the clusters, on the other hand, could either be following this pattern, or simply be late-stage regions where molecular destruction has already taken place. In this last scenario, the chemical signature of the region could resemble that of an earlier-stage region where complex molecules have not yet formed.

The molecular richness of a core is captured in the dataset by the number of lines in each signal. Figure 4.14b shows how this variable is distributed along the different clusters, showing clear differences between the clusters. This variable is not a perfect metric to show molecular richness as random noise or artifacts created by the filtering algorithms could have accidentally been counted as spectral lines. This, however, does not seem to be the case as the unclassified signals, which have the lowest SNR within the dataset, also have the lowest median value. The number of lines can therefore be used to establish an evolutionary sequence, as regions evolving from cold- to hot-cores are expected to have an increasing number of molecular lines [44].

The next property that better establishes the evolutionary sequence of a region is its temperature. High-mass star-forming regions evolve from cold cores, to hot cores, to HII regions, each of them with their own particular chemistry [44, 62]. Given the different amount of clusters that may be found within one region, as shown in Figure 4.11, the most relevant temperature would be the core temperature. This variable, however, was directly obtained from the clump's L/M ratio rather than calculated based on the chemical composition of each core [14]. In fact, there were three possible temperature assignments, as detailed in Coletta et al. [14]: "(i) 20 K to cores belonging to less evolved targets (clump L/M $\leq$ 1 $L_\odot/M_\odot$); (ii) 35 K to core pertaining to intermediate sources (1 < L/M $\leq$ 10 $L_\odot/M_\odot$); iii) a temperature following the relation [see Equation 4.2] (with L/M expressed in $L_\odot/M_\odot$) to cores in more evolved targets (L/M > 10 $L_\odot/M_\odot$)".

$$T_{\text{core}} = 21.1 \cdot (L/M)^{0.22} \tag{4.2}$$

This implies that no information can be extracted from the core's temperature that was not already present in the L/M ratio.

The final parameter of interest is the density of the cores. This value is partly dependent on temperature, which may have induced some errors [14]. However, it is also based on the integrated flux and the size of the core, making it a more robust measurement than core temperature [14]. The expected density distribution would show low values at early stages, increasing as the core collapses, and then decreasing once more as the HII region expands. Regions near an HII region, however, may have an increased density due to the external pressure exerted by the HII region. Density can then be measured in one of two ways: surface density (Figure 4.14d) or volumetric density (Figure 4.14c). The only difference between both distributions is that the former is proportional to R^{-2} whereas the latter is proportional to R^{-3} [14]. When looking at the data, however, the median volumetric density is consistently higher in non-contaminated signals, whereas the surface density is higher for the HII-contaminated signals in 3/7 clusters and for the non-classified signals. It is worth noting that the difference in both groups' means is very small, however, the 14th percentile for all distributions other than that of the unclassifed signals is lower when considering the non-contaminated signals. This indicates that these signals dominate the lower-end of each group, which can only be explained by the radii of the contaminated cores being smaller than that of the non-contaminated ones. This would be consistent with the signals originating in early HII regions where molecular destruction occurred faster than material outflows. According to this theory, the core would have the chemistry of a younger region and the density of a more evolved one. Having said this, the next step was to compare the distributions of the non-contaminated signals.

Cluster-based evolutionary sequence

Having analysed all relevant variables, the final remaining step is to propose an evolutionary order within the clusters. From oldest to youngest, and considering only up until the hot-core phase (no HII regions), the proposed order is:

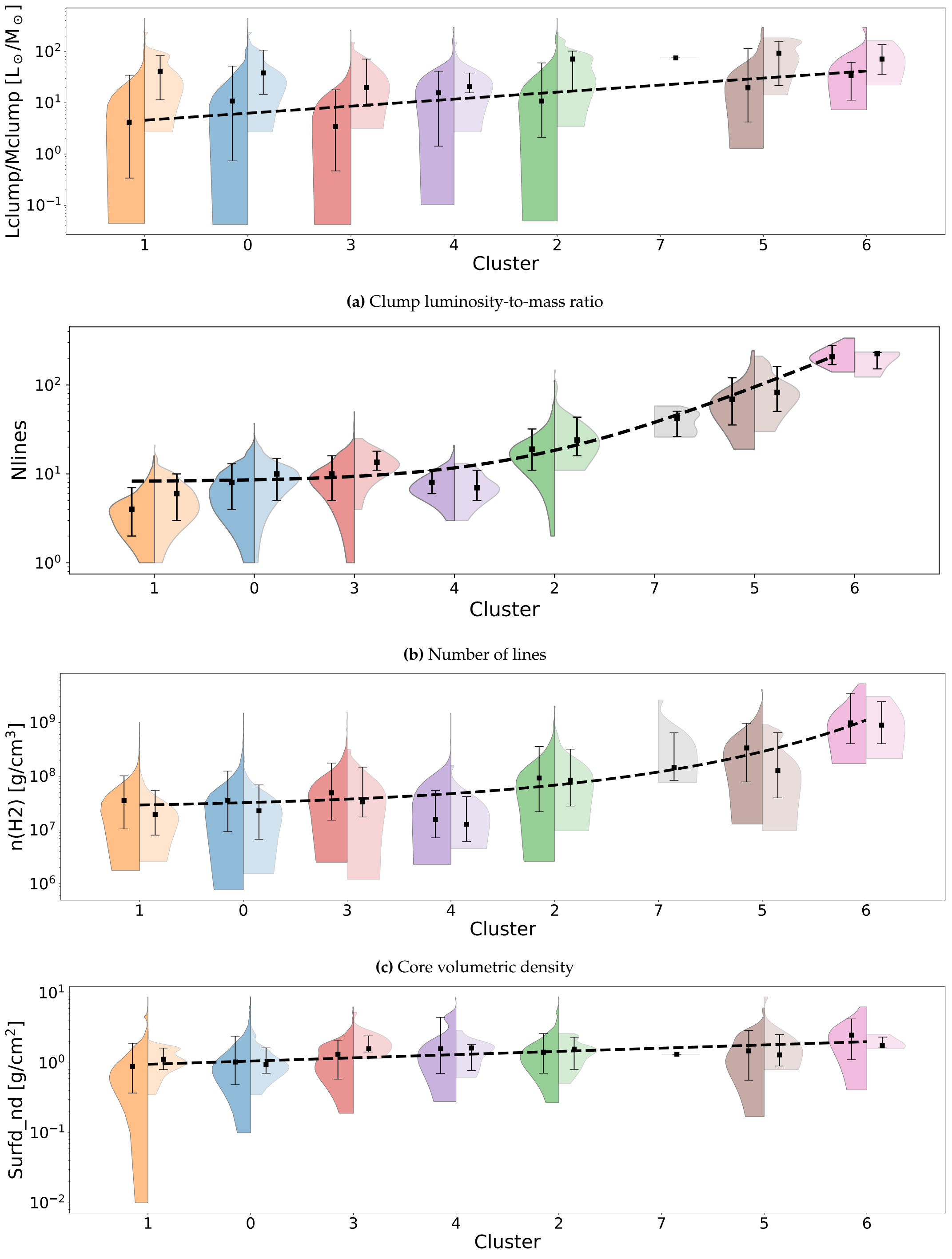
1. **Cluster 6** This cluster has the highest non-HII median value in all measured properties and the most amount of spectral lines. Considering all values, this cluster's core temperature and L/M ratio is only surpassed by cluster 7's. This last cluster, however, is fully contaminated by HII regions, which most likely implies that the assigned L/M ratio is higher than the real one.
2. **Cluster 5** This cluster is second only to cluster 6 in all parameters except surface density when considering the non-contaminated values. When considering all values, this cluster has lower values than cluster 7 in all parameters except the number of lines. However, the possible HII contamination and the outflow signatures present in this last cluster indicate that it is probably on an earlier evolutionary stage than cluster 5.

3. **Cluster 7** Values such as the L/M ratio are a bad indicator of this cluster's evolutionary stage due to the possible HII-region contamination. This cluster, however, places 3rd in the number of lines. Additionally, this cluster also ranks 3rd in volumetric density. This metric is not entirely independent of the HII region's luminosity, making it a more reliable parameter than the L/M ratio. Finally, the wide SO and SiO peaks in spw1 and spw0 are clear indicators of outflow emission, indicating that a protostellar object is present in these cores [20].
4. **Cluster 2** This cluster's number of lines is significantly lower than that of the previous three clusters. Despite this, this distribution's 14th percentile is still higher than the median number of lines of all remaining clusters, thus indicating a higher chemical richness. Additionally, this cluster also ranks fourth in volumetric density and third in surface density (third and fourth respectively when not considering HII-contaminated signals). Finally, cluster 4 has a higher median L/M ratio, however, it has lower 14th and 86th percentiles.
5. **Cluster 4** This cluster has the highest median L/M ratio and surface density out of all remaining ones. Its median number of lines is smaller than that of cluster 3, however, it does have a higher 14th percentile. On the other hand, this cluster has the lowest volumetric density out of all remaining clusters.
6. **Cluster 3** Out of the remaining 3 clusters, this one has the lowest median L/M ratio, but it has the highest median number of lines and density (both volumetric and surface). All three clusters all have a median core temperature of 35 K, indicating that most signals in these clusters have an L/M ratio lower than 10 $L_{\odot}/M_{\odot}$. It is therefore plausible that all three clusters are in the early stages of formation, and that cluster 3's collapse has progressed the most, leading to the observed higher densities.
7. **Cluster 0** The remaining two clusters are broadly similar, however, cluster 0 ranks (slightly) higher than cluster 1 in all categories.
8. **Cluster 1**

The evolution of the physical properties throughout the sequence can be seen in Figure 4.15. All variables increase exponentially throughout the sequence (all plots in Figure 4.15 use a logarithmic scale), with the steepest increases being those of the number of lines and volumetric density. The least indicative parameter is the surface density, as it is the one that changes the least between the first and last stage.

The unclassified signals (labelled as cluster -1) did not share any common chemical structure in their spectra, however, they do share some physical properties. Firstly, they have the lowest L/M ratio out of all distributions without HII regions. Their low SNR makes them also have the lowest median number of spectral lines. Additionally, they also have the lowest surface density out of all groups of signals. This would all be consistent with a dense core that still has not acquired enough mass to break thermal equilibrium, placing them as the youngest regions in the dataset. However, these signals rank fourth in volumetric density, indicating that their average radius is much smaller than that of most other cores. This could potentially be explained by the GHC model, where cores accrete material from their surroundings throughout their collapse [**vazquez-semadeni_gravitational_2019**]. It could therefore be plausible that these unclassified signals are at the very first stages of star-formation, where the core is accreting material faster than it collapses, leading to the next stages having a larger radius.

Having already established an evolutionary sequence for the 8-cluster arrangement, another one was done for the 19-cluster arrangement. Since the temperature distributions were not used in the previous analysis, they were not plotted this time. All other distributions can be found in Figure 3.58, with the numeric data in Tables 4.4-4.7. Most relationships between the contaminated and non HII-contaminated distributions are maintained. A few notable exceptions include the L/M ratio of clusters 13 and 18, the first of which, despite having almost 40% of its signals contaminated by HII regions, has a median L/M ratio 41% higher in its non-contaminated signals. Similarly, the volumetric density of clusters 7, 9 and 15 are all higher for the HII-contaminated signals. No additional conclusions could be extracted from the data itself, so an evolutionary order was directly proposed.



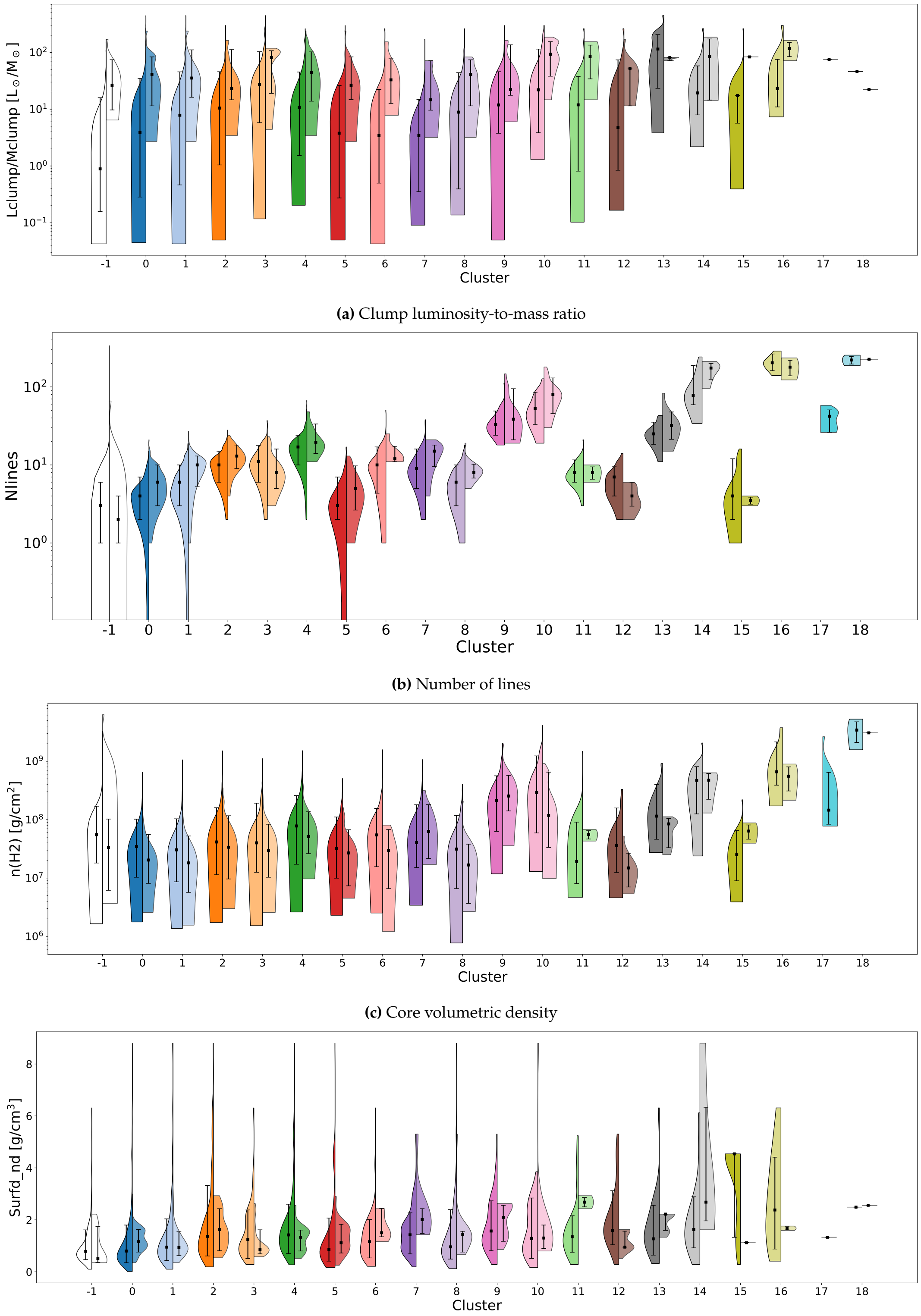
(a) Clump luminosity-to-mass ratio

(b) Number of lines

(c) Core volumetric density

(d) Core surface density

Figure 4.15: Distribution of the physical properties of the clumps/cores throughout the proposed evolutionary sequence. In all sub-figures, the left violin plot represents the data without HII region contamination, whereas the right one contains only the data with HII contamination for each cluster.



(a) Clump luminosity-to-mass ratio

(b) Number of lines

(c) Core volumetric density

(d) Core surface density

Figure 4.16: Distribution of the physical properties of the clumps/cores along the different clusters identified in the 19-cluster arrangement. In all sub-figures, the left violin plot represents the data without HII region contamination, whereas the right one contains only the data with HII contamination for each cluster.

Table 4.4: L/M ratio distribution for the different clusters in the 19-cluster arrangmenet.

Cluster	Lclump/Mclump (No HII)			Lclump/Mclump (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	0.16	0.89	15.92	9.69	26.35	74.27
0	0.28	3.91	34.78	11.44	41.05	83.28
1	0.46	7.81	45.46	16.20	35.34	110.85
2	1.04	10.46	45.69	14.67	23.01	112.15
3	5.78	27.39	102.86	18.97	81.16	106.91
4	1.53	10.82	45.15	13.83	44.65	102.18
5	0.27	3.77	26.08	14.79	26.40	83.28
6	0.50	3.44	22.26	12.61	32.88	77.33
7	0.35	3.43	14.86	9.59	14.67	71.42
8	0.39	8.87	43.93	11.44	40.72	74.02
9	3.76	11.88	45.69	17.54	22.23	135.83
10	3.85	21.91	114.47	38.05	92.59	154.11
11	0.81	11.93	37.73	34.17	84.31	134.46
12	0.84	4.75	73.76	11.44	51.64	52.88
13	23.26	114.47	205.95	71.42	81.16	81.16
14	7.93	19.31	57.87	14.44	84.31	172.22
15	5.66	17.46	17.46	83.28	83.28	83.28
16	10.95	23.16	75.36	84.19	117.02	149.86
17	-	-	-	75.20	75.20	75.20
18	46.23	46.23	46.23	22.23	22.23	22.23

Table 4.5: Caption

Cluster	Nlines (No HII)			Nlines (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	1.00	3.00	6.00	1.00	2.00	4.00
0	2.00	4.00	7.00	3.00	6.00	10.00
1	3.00	6.00	10.00	5.34	10.00	13.00
2	6.00	10.00	15.00	9.00	13.00	18.00
3	6.00	11.00	17.70	5.00	8.00	16.00
4	10.00	17.00	24.00	14.00	19.50	33.52
5	2.00	3.00	7.00	2.64	5.00	9.72
6	4.34	10.00	17.00	11.00	12.00	17.30
7	5.00	9.00	16.00	9.52	15.00	18.00
8	3.00	6.00	10.00	6.78	8.00	10.22
9	24.00	33.00	49.16	21.00	38.50	94.82
10	33.00	53.00	85.40	45.52	80.00	130.08
11	6.00	8.00	11.60	6.56	8.00	9.44
12	4.00	7.00	9.48	2.96	4.00	6.00
13	18.36	25.00	35.28	21.30	32.00	47.70
14	59.12	78.00	188.14	125.40	174.50	199.24
15	2.00	4.00	12.00	3.14	3.50	3.86
16	161.40	204.00	264.32	138.68	179.00	219.32
17	-	-	-	26.26	42.00	50.70
18	196.52	221.00	245.48	226.00	226.00	226.00

Table 4.6: Caption

Cluster	Surfd_nd (No HII)			Surfd_nd (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	0.47	0.79	1.61	0.36	0.51	1.74
0	0.35	0.80	1.80	0.76	1.16	1.63
1	0.43	0.95	2.04	0.62	0.94	1.54
2	0.61	1.37	3.32	0.81	1.63	2.43
3	0.51	1.25	2.38	0.69	0.86	1.62
4	0.70	1.42	2.60	0.80	1.33	1.61
5	0.41	0.86	2.07	0.72	1.12	1.83
6	0.53	1.16	2.01	1.35	1.50	2.44
7	0.69	1.42	2.27	1.48	2.01	2.43
8	0.49	0.96	2.39	0.76	1.44	1.56
9	0.81	1.57	2.73	1.17	2.10	2.56
10	0.52	1.29	2.84	0.90	1.30	1.80
11	0.75	1.35	2.16	2.50	2.68	2.86
12	1.04	1.59	3.12	0.95	0.95	1.56
13	0.64	1.27	2.56	1.59	2.22	2.22
14	0.92	1.63	2.89	1.96	2.68	6.33
15	1.33	4.54	4.54	1.12	1.12	1.12
16	0.88	2.38	4.41	1.61	1.68	1.74
17	-	-	-	1.33	1.33	1.33
18	2.49	2.49	2.49	2.56	2.56	2.56

Table 4.7: Caption

Cluster	n(H2) (No HII)			n(H2) (Only HII)		
	P14	P50	P86	P14	P50	P86
-1	18.10	55.20	170.02	6.16	33.50	102.48
0	10.31	34.60	102.00	8.10	20.35	55.71
1	8.63	30.40	103.44	5.68	18.15	52.46
2	11.41	41.50	159.90	9.65	33.70	116.80
3	12.56	40.10	191.70	10.31	29.30	83.50
4	17.20	78.00	257.96	26.24	51.55	136.52
5	9.94	32.20	110.98	7.37	26.90	67.10
6	15.63	54.75	155.28	6.60	29.70	67.79
7	15.08	40.45	179.80	21.66	62.95	180.60
8	6.61	31.35	118.46	3.69	16.75	38.07
9	63.01	212.00	563.68	141.00	254.00	572.28
10	59.17	292.00	1237.00	33.40	118.50	650.96
11	8.01	19.20	90.86	46.54	55.65	64.76
12	12.42	36.00	158.36	7.02	14.90	26.63
13	46.07	115.00	401.56	33.14	84.60	104.84
14	125.38	470.50	808.34	223.24	473.00	616.62
15	8.98	25.10	64.86	46.29	63.75	81.21
16	388.92	659.00	2143.60	310.34	555.50	800.66
17	-	-	-	83.49	145.70	645.58
18	2095.20	3420.00	4744.80	3070.00	3070.00	3070.00

1. **Cluster 18** This cluster ranks first in number of lines and volumetric density, and second in L/M ratio and surface density. The first cluster in both of these categories, however, is different, making this the most likely candidate for the eldest cluster.
2. **Cluster 16** This cluster ranks just before cluster 18 in all categories when considering all signals. Discarding contaminated signals, cluster 3's L/M ratio becomes larger than cluster 16's. Despite this, the great amount of lines in cluster 16 is more consistent with hot cores near an HII region rather than with cores within an HII region.
3. **Cluster 14** Once more, this cluster ranks just below cluster 16 in all categories except L/M ratio. Its median is surpassed by clusters 3 and 10 in this category, but not its 14th quartile. This implies that this cluster is much less disperse than the other two, giving a higher confidence to its L/M ratio.
4. **Cluster 10** This cluster has the highest median L/M ratio and number of lines out of all remaining clusters. Additionally, it ranks second in number of lines, only surpassed by cluster 9.
5. **Cluster 13** This cluster has the overall largest L/M ratio and ranks second in number of lines and volumetric density out of all remaining clusters. Furthermore, the difference in L/M ratio is much more significant than the difference in number of lines and density, therefore meriting the 5th position.
6. **Cluster 17** This cluster is equivalent to cluster 7 in the 8-cluster arrangement, which placed 3rd in its ranking. Chemically speaking it is very similar to cluster 13, with the main difference being the width of its peaks. This parameter, however, depends on the outflow's alignment with respect to Earth, making it a poor element for comparing both clusters. The decision to place this cluster below cluster 13 was therefore made based on the fact that the luminosity of the latter is significantly greater.
7. **Cluster 9** Out of all remaining clusters, this one has the highest number of lines and the largest volumetric density. It is also second in surface density when considering all signals and third when discarding the possibly-contaminated ones. However, it does rank 9th overall when looking at the median L/M ratio, falling behind clusters 3, 11, 15 and the already mentioned ones. This is why, despite its chemistry resembling that of a hot core, this cluster was placed at this position.
8. **Cluster 3** All remaining clusters have a chemistry that resembles that of a cold core. Out of these clusters, cluster 3 has the highest L/M ratio and the second-largest number of lines. This cluster's density is fairly low, however, the differences between the remaining clusters in those categories are minor.
9. **Cluster 4** This cluster places third when comparing its median L/M ratio to that of the remaining clusters; however, its 86th percentile is the largest. Additionally, this cluster ranks first in number of lines and volumetric density.
10. **Cluster 2** This cluster has the highest number of lines, the second-highest volumetric density and the third-highest L/M ratio out of the remaining clusters. Additionally, the 86th percentile in this cluster's L/M ratio is, once more, higher than that of the clusters with a higher median.
11. **Cluster 11** Despite this cluster having the lowest median volumetric density, it has the second-highest L/M ratio and the third-highest number of lines.
12. **Cluster 12** This cluster has the third-highest number of lines and volumetric density, and its 86th percentile L/M ratio is the highest out of all clusters. Most importantly, however, no other cluster has a good position both in L/M ratio and number of lines.

Beyond this point, the decision for ranking one cluster before another becomes increasingly hard. For instance, clusters 6 and 7 have the lowest median L/M ratio, but the highest number of lines and volumetric density. The most likely candidates would be either cluster 15 or cluster 1. The former has the highest median L/M ratio and surface density, however, it places third-to-last in number of lines and last in volumetric density. Cluster 1, on the other hand, places third in both median L/M ratio and number of lines, but second-to-last in volumetric density and third-to-last in surface density. The final contender for the 13th place would be cluster 8, which has the second-highest L/M ratio and the fourth-largest number of lines. Additionally, all of these three clusters have very similar centroids, especially taking into account that the absorption features from cluster 15 are probably caused by a saturation of the measuring instrument. Knowing this, all three clusters were tied in the 13th position.

The remaining 4 clusters (0, 5, 6 and 7) were also tied at the last place, given how the ones with the lowest L/M ratio were also the ones with the highest number of lines and density.

Overall, the final evolution of the physical properties of the sequence can be seen in Figure 4.17. The same trends as in the 8-cluster arrangement can be observed, with the number of lines and volumetric density showing the steepest increase. The surface density, once more, is the least indicative parameter.

Sequence overlap

The additional clusters from the 19-cluster arrangement are sub-groups of the clusters in the 8-cluster arrangement. Ideally, each sub-group would concentrate a continuous stratum of the physical properties of the original cluster. This means that if one cluster is broken down into two, one of them should concentrate the upper half while the other one concentrates on the lower one. Each cluster would therefore have distinct values for each variable but the same overall position in the ranking.

This theory was put to the test by comparing both rankings, although some modifications needed to be done due to the different number of elements in each list. The first step was to back-trace which clusters form which, shown in Table 4.8.

Table 4.8: Mapping between the 8- and 19-cluster arrangement

Cluster 8 label	→	Cluster 19 label
0	→	1, 2, 3, 8
1	→	0, 12, 15
2	→	4, 9, 13
3	→	6, 7
4	→	5, 11
5	→	10, 14
6	→	16, 18
7	→	17

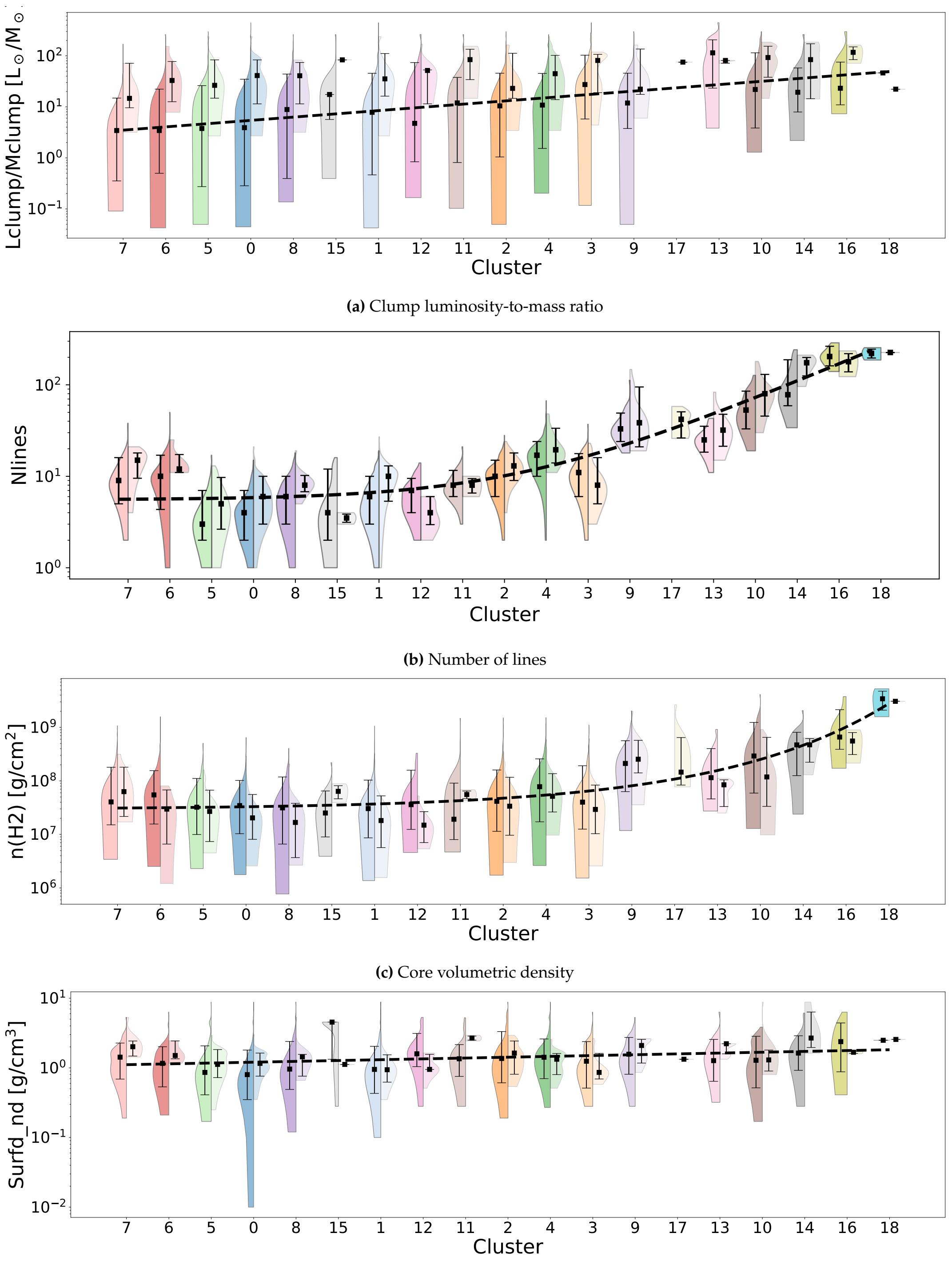
Once this was done, the 8-cluster ranking needed to be expanded to a list with 19 elements. This was done by repeating each cluster a number of times equal to the number of sub-clusters it has. For instance, cluster 0 in the 8-cluster arrangement is formed by clusters 1, 2, 3 and 8 in the 19-cluster arrangement. Therefore, cluster 0 was repeated 4 times in the new list. Finally, the 19-cluster ranking needed to replace each cluster by its corresponding cluster in the 8-cluster arrangement. For example, cluster 18 in the 19-cluster arrangement is part of cluster 6 in the 8-cluster arrangement. Therefore, cluster 18 was replaced by cluster 6 in the new list. The resulting rankings can be found in Table 4.9.

Table 4.9: Reformatted rankings for both cluster arrangements, from youngest (left) to eldest (right).

8C ranking	1	1	1	0	0	0	0	3	3	4	4	2	2	2	7	5	5	6	6
19C ranking	3	3	4	1	0	1	0	1	4	0	2	0	2	7	2	5	5	6	6

These rankings were then compared using Spearman's rank correlation coefficient, which yielded a correlation value of 0.775, with a p-value of 9.63e-05. Having ties in a list reduces the maximum possible correlation value, making 0.775 a very high correlation and proof that the rankings are very similar.

Manually inspecting the rankings, it can also be stated that the older stages of the star-formation process will be in accordance more often than those at earlier stages. There are two factors that lead to this conclusion. Firstly, 8 out of the last 9 elements in the ranking are in agreement between in both instances. Secondly, the last 7 elements of the second ranking were tied in a last position due to the inability of the physical properties as currently calculated from discerning one cluster from the other. Additionally, this conclusion is also in agreement with the one drawn about the Euclidean-distance metric in Chapter 3, where it was established that this distance metric is better at differentiating clusters with many lines than at distinguishing clusters with few features.



(a) Clump luminosity-to-mass ratio

(b) Number of lines

(c) Core volumetric density

(d) Core surface density

Figure 4.17: Distribution of the physical properties of the clumps/cores throughout the proposed evolutionary sequence. In all sub-figures, the left violin plot represents the data without HII region contamination, whereas the right one contains only the data with HII contamination for each cluster.

4.4. Conclusion

This chapter focused on analysing the physical properties of the star-forming regions and their relationship to the clusters found on Chapter 3. Firstly, the relationship between a cluster's size and average number of spectral lines was analysed, revealing an exponentially decaying relationship between both variables in both cluster arrangements. This suggests that the evolution of the star-forming regions accelerates as the core collapses, which is in agreement with what can be found in literature [42, 46].

The chemical composition of each cluster was also investigated, revealing that even small clusters share few to no spectral lines among all signals. This could be proof of a wider chemical variety than previously thought to exist, however, it is more likely that small errors such as a slightly inadequate red-shift correction could have caused this effect.

The next step of the analysis revolved around HII regions. The available data did not specify whether a signal originated at an HII region. Instead, if an HII region was detected within the scanned section of the sky, all signals in that region were marked as "possibly contaminated". It is therefore possible that clusters marked as contaminated did not indeed arise from within an HII region, but rather from a region near one. This is thought to be the case for the 10-signal cluster repeated in both arrangements (cluster 7 or 17 in the 8- or 19-cluster arrangements), where the wide SO and SiO peaks are strong outflow indicators. In spite of this limitation, the proportion of the possibly contaminated signals in each cluster was investigated. In both arrangements, the unclassified signals show very low contamination (~ 5%). The clustered signals, on the other hand, show much more dispersion in the 19-cluster arrangement, where several clusters have over 30% of their signals contaminated while others have less than 10%. The 8-cluster arrangement, on the other hand, is more consistent, with an average contamination level of ~ 18% throughout almost all clusters.

Following the HII regions, the relationship between the number of signals and the number of clusters in a core was investigated. A correlation of 0.683 was found between both variables using the 8-cluster arrangement, scaling up to 0.806 when using the 19-cluster arrangement. Additionally, two empty areas were found in both groups: one at the bottom right of both plots, and another one at the top of the 19-cluster arrangement plot. The first space, common to both arrangements, proved that it is very unlikely for a source to have more than 13 signals and less than 2 clusters present. The second space, on the other hand, showed that no region has more than 12 different clusters within it, despite there being 19 possible clusters. The first finding is consistent with the CA model, where several evolutionary stages are expected to be found within a single collapsing cloud [7, 61]. The second finding, on the other hand, suggests that some evolutionary stages depend on the cloud's initial conditions, making it impossible for a single cloud to form all evolutionary stages.

Finally, the physical properties of the clusters and their relationship to the cluster arrangements was investigated. First, the correlation between each available property and the number of lines (a proxy variable for the chemical richness of a core) was calculated. Numerically, only the volumetric density shows some correlation (0.452); however, the figures in Figure 4.13 illustrate that the variables are correlated, although not linearly. Each variable's distribution within each cluster was then used to establish an evolutionary sequence between the different clusters. The sequences obtained using the 8- and 19-cluster arrangements were then compared, showing a Spearman correlation coefficient of 0.775. This correlation is even higher in the later parts of the sequence, where 8 out of the first 9 elements match. This proves that regions at later evolutionary stages are better defined and can be clustered in more depth by an agglomerative clustering model using a Euclidean distance metric.

Table 4.10: Summary of the analysis of the cluster arrangements and their physical properties.

Aspect	Finding	Comment
Size and abundance	Inverse relationship	Signals with numerous spectral lines are exponentially less common than signals with few lines. This suggests that the evolution is slower at the initial stages of the star-formation process, in agreement with literature.
Chemical composition	Inconsistent	Large clusters in both arrangements have no lines common to all signals in the cluster. Furthermore, small clusters have very few lines (one in most cases) common across all signals. No valid conclusions can therefore be extracted.
HII regions	Inconclusive	The presence of HII regions in the dataset is not known for certain, which limits the capacity for analysis. The centroids of potentially contaminated regions showed little to no difference to their counterparts within the same cluster. Additionally, their distribution throughout the clusters is approximately homogeneous in the 8-cluster arrangement, but much more different in the 19-cluster arrangement.
Physical properties	Non-linear correlation	All variables showed a low Pearson correlation coefficient with respect to the number of lines. Looking at the distributions, however, it can be seen that the variables are related, although not linearly. For instance, high values of L/M ratio are only be found in signals with many lines.
Evolutionary sequence	Successful	An evolutionary sequence was proposed for both arrangements based on the distributions of the physical properties. The sequences were then compared, showing a high correlation, especially in the later stages of the sequence. This last point implies that the cores become more homogeneous as they evolve, showing less chemical variability. Finally, the physical properties throughout the sequence match what would be expected according to literature, giving validity to the process.

5 Future Work

The work presented in this thesis revolved around three major aspects: pre-processing of astrochemical spectra, unsupervised clustering of the processed spectra, and the establishment of an evolutionary sequence based on those clusters. Due to time constraints, not all aspects could be explored in all categories, so this chapter will detail the work that can still be done to complete this investigation.

5.1. Pre-processing

The first aspect to explore in future work should be the addition of new filtering mechanisms. The most used filtering in this work was the subtraction-filtering algorithm, which resulted from subtracting a region's signal and its residual. This signal, however, was inaccurate on several occasions, creating artificial peaks that could have confused the models in later stages. A solution for this could have been the exploration of a combined subtraction and threshold-based filtering, or other algorithms not explored in this study.

Finally, the only remaining aspect that needs completion is the validation of the red-shift correction algorithm. Currently, that algorithm correctly predicts the velocity of 99.16% of signals with a margin of 5 km/s. This statistic, however, is limited to the signals for which a velocity has been manually calculated, which are also the signals with the highest SNR. It is therefore possible that the accuracy will decrease when accounting for the additional 650 signals used in this study that did not have a manually identified velocity.

5.2. Clustering

Only 3 unsupervised clustering models were tested in this study. Future work could therefore focus on testing new algorithms that appear promising in literature. Most notably, unsupervised neural networks that make use of the models developed in this study should be tested. Such algorithms have shown promise in literature, such as in the works by Xie, Girshick, and Farhadi [65] or Xu and McCord [66].

Furthermore, the models tested in this study could be enhanced through the addition of new distance metrics. Only the Euclidean and cosine distance metrics were used in this work, and it was proven that each metric performed better at different tasks. A new distance metric that combined both of them could therefore provide better-defined clusters. Alternatively, other distance metrics could also be tested in order to find the most appropriate one.

Finally, it could also be interesting to train unsupervised ML models on the physical properties of each signal and check how those clusters correlate to the ones found by clustering the chemical spectra.

5.3. Analysis

This part of the study was severely limited by the lack of data regarding the dense cores. Most properties such as the temperature of the cores or the L/M ratio were calculated for the entire clump rather than for each individual core. Given how signals from the same clump proved to be widely different between them, it could be expected that the physical properties may differ too. Those properties are currently being calculated within the ALMAGAL consortium, and a new iteration of the analysis should be done once they become available.

One particular aspect that could be interesting to explore is the ranking of the different clusters present within each cloud. Currently, it was found that each clump has an increasing amount of clusters as the number of cores increases. The placement of these clusters within the evolutionary sequence, however, was not analyzed. It could therefore be of interest to investigate whether the different clusters are or not consecutive in the sequence.

Additionally, the established evolutionary sequence does not consider the HII regions. Once it becomes clear which signals belong to an HII region, it could be interesting to extend the analysis into this regime and study how HII regions evolve.

6 Conclusion

The objective of this work was to investigate the extent to which unsupervised ML models can be used to establish an evolutionary sequence of the high-mass star-forming regions, based on the spectral line content of dense cores. To do this, the work was divided into three main parts: pre-processing, clustering and analysis. The conclusions drawn from each of these sections will be presented in this chapter.

The pre-processing of the data started by re-sampling the spectra to a common set of channels. This data was then filtered according to three separate methods, giving raise to four different datasets. These methods were the following:

1. Threshold filter: The standard deviation of the residual signal is calculated. A threshold of significance is then set based on this value and all channels with an intensity lower than the threshold are suppressed. The threshold in this study was chosen to be 3σ.
2. Subtraction filter: The residual signal is subtracted from the core's signal, leaving (in theory) only the relevant information. The filter generally works well, however, it does create some artificial peaks around the baseline in signals with a low SNR.
3. Savitzky-Golay filter: A polynomial regression is performed on successive windows of the spectra, smoothing the signal and removing random noise. This filter does not depend on the availability of a residual signal.

Out of these three, the Savitzky-Golay filter was quickly discarded as it yielded much worse results than the other two. The remaining three datasets were then used to calculate each region's relative velocity to Earth. This was done by comparing each shifted signal's cosine similarity to a reference spectrum for a set of velocities. The velocity with the highest cosine similarity was then compared across datasets and a majority-voting scheme was implemented. This algorithm found a velocity for 4322/4396 (98.32%) signals. The algorithm was then validated using the manually calculated velocities provided by the ALMAGAL consortium. These velocities were still being calculated as of the writing of this thesis, so only 3672 of them were available. From these velocities, 99.16% matched the ones found using Algorithm 1 with a margin of ±5 km/s, therefore validating the algorithm for the remaining signals. Due to unclear reasons, the alignment of the signals using Algorithm 1 was greater than when using the manual velocities, so those were used in this study.

As a final step in the pre-processing, the data was subjected to two different dimensionality-reduction techniques: PCA and UMAP. Regarding PCA, it was first proven that fewer components were needed to retain the variance of the filtered signals than in the raw signals. For comparison, 193 components are needed to retain 95% of the raw dataset's variation, whereas 28 are needed when using the threshold-filtered dataset and 15 when using the subtraction-filtered one. The effects of filtering were also apparent using UMAP, as the projection changed greatly when using the filtered datasets. Furthermore, different projections were obtained when using different distance metrics, with the Euclidean distance yielding a greater separation between points.

Once all pre-processing had been done, the different clustering algorithms were tested. Three algorithms were studied in this work: k-means, GMM and agglomerative models. The k-means classifier was the first to be tested. Two different distance metrics were used, Euclidean and cosine distance, and the most appropriate number of clusters was determined using the silhouette score (also calculated using both distance metrics). The silhouette score proved to be biased towards the Euclidean distance metric; however, it was still useful to determine the most appropriate number of clusters. The filtered datasets showed a significant improvement in the silhouette score, with the subtraction-filtered dataset yielding the best results. Its silhouette score showed a great decrease after 4 and 8 clusters, with the latter being the most significant. This last cluster arrangement was therefore tested using both distance metrics. Both models yielded similar clusters, with the Euclidean distance model placing more emphasis on signals with many spectral lines and the cosine distance model being better at differentiating signals with few lines. The quality of the clusters was then evaluated using the weighted-average WCS and BCS. Both models had a comparable WCS of around 0.6, but the Euclidean distance model had a lower BCS/WCS ratio (0.3749 vs 0.4044).

In an attempt to avoid the clusters being dominated by spectral lines that could originate at the clouds rather than the cores, the peaks corresponding to H_2CO, ^{13}CO, $c^{18}O$ and SO were suppressed. The clustering with this new dataset, however, only managed to identify two clusters: those with many spectral lines and those with very few. Furthermore, the cluster with very few lines was much larger than the one with many, proving an imbalance within the dataset.

Finally, the k-means algorithm was tested on the PCA-reduced dataset, as done in literature [70]. Only two clusters were detected this way, however, and of much poorer quality than with the entire dataset.

The next model evaluated was the Gaussian Mixture Model (GMM). This cluster required the use of dimensionality-reduction techniques, as the computing time scales with the square of the number of channels. Using PCA, a total of 4 clusters were discovered; however, the quality of these clusters was worse than the one obtained using the k-means classifier. Using UMAP, on the other hand, a much more promising cluster arrangement was found. First, the signals placed at a significant distance from the main group were analyzed, revealing two small clusters significantly separated from the main clump. The silhouette score was then used to determine the most appropriate number of clusters, revealing that 7 clusters should be used in the Euclidean-distance UMAP projection. Inspecting these clusters, it was found that one of the clusters was made up of a group of signals with very low SNR and no particular pattern in their spectral lines. This cluster was therefore discarded, as well as one of the small ones found due to their signals being very similar to those of the remaining clusters.The WCS of the final arrangement was therefore 0.613, with a BCS/WCS ratio of just 0.2939. These numbers, however, relate to just 59.14% of the dataset. The remaining signals were confirmed to be in a transition state between different clusters, supporting the theory that these centroids represent different evolutionary stages of star formation. The same procedure was then applied to the cosine-distance projection, however, only 5 clusters were identified, with a WCS of 0.5607 and a BCS/WCS ratio of 0.3363.

The final clustering method tested in this thesis was agglomerative clustering. Following Zhang et al. [71], a neighborhood-based model was first tested. Using this model directly on the signals yielded very poor results, however, several small clusters were directly found when using it on the UMAP projections, including the one manually identified in the GMM section. This cluster arrangement yielded a very large cluster and 8 very small ones. The large cluster was then broken down by back-tracing the merging steps taken to form it. A total of 8 sub-clusters with an average WCS of 0.59 was found through this method for both UMAP projections, leading to a total clustering formed by 16 different clusters. The BCS/WCS ratio, however, was 0.418 for the Euclidean-distance projection and 0.435 for the cosine-distance projection, which is a worse performance than that of k-means and GMM. Finally, a conventional agglomerative clustering model was tested. When using a cosine distance metric, a total of 21 clusters were found; however, these clusters were also not well-defined as 75% of the data was concentrated in just 2 clusters. In the end, the final clustering arrangement used came from using a Euclidean distance metric in a conventional agglomerative clustering model. This model showed two significant peaks, one at 10 clusters and another one at 23. In both cases, some clusters contained just one signal (which were discarded), leading to a total of 8 and 19 clusters respectively. The clusters with a WCS lower than 0.4 were pruned such that only signals with a SNR larger than 1 dB remained, leading to the clustering of 89.92% and 91.15% of all signals. This final arrangement had a WCS of 0.618 for the 8-cluster arrangement and 0.657 for the 19-cluster arrangement, with BCS/WCS ratios of 0.298 and 0.318 respectively.

Using the last two arrangements, the physical properties of the clusters were inspected. First, the size of the clusters was compared to the amount of spectral lines in their signals. This yielded an exponentially decreasing relationship, indicating that the evolution of a star forming region is slower in the early stages than in the more advanced ones. Next, the number of clusters in each region was compared to the number of cores in each region. The results indicate that most clouds will have cores with several evolutionary stages present, consistent with the CA model. Additionally, no region had more than 12 different clusters present, potentially indicating that some evolutionary stages depend on the initial conditions of the cloud. Finally, the physical properties of the cores were investigated. The only variable that seemed correlated to the number of lines in a signal (used as a proxy for the chemical richness in the core) was the volumetric density of each core. A closer inspection, however, showed how signals with a high number of lines will always have high L/M ratios and temperatures. Finally, an evolutionary sequence was established based on these properties, using both clustering arrangements. The order of the sequence was compared, reaching a Spearman correlation coefficient of 0.775, proving that breaking down the clusters based on their chemistry has little effect on the cluster's placement in the sequence. This effect is enhanced for hot cores, where both lists' correlation was almost perfect.

After all this investigation, the research questions posed in the introduction can now be answered:

> **RQ:** To what extent can unsupervised ML be used to establish an evolutionary sequence of the star-formation process, based on the spectral line content of the dense cores?

Machine learning techniques can confidently be used to cluster signals based on their chemical spectra. Many of the signals will be very close to one another, so simple models such as K-means will yield heavily-overlapping clusters and fail to identify smaller groups. More powerful models coupled with strong pre-processing techniques, on the other hand, can be used to achieve a clustering whose physical properties match what would be expected from literature. The extent to which the evolutionary sequence can be established, however, depends heavily on the quality of the data about the cores' physical properties.

> **RQ-1.1:** How does pre-processing affect the clustering performance of the different models?

The performance of the models is heavily affected by the pre-processing phase. An incorrect alignment when correcting for red-shift will lead to models comparing the incorrect channels and therefore not being able to identify similar signals. However, the magnitude of the effect will depend on the particular model used. For instance, CNNs have been known to not be affected by shifted spectra [31], whereas the GMM only performed satisfactorily when using the UMAP projected data. Additionally, other pre-processing steps such as normalizing the signals are required for most models to work properly. This is the case for any model using the Euclidean distance metric, as signals with the same outline but at different overall intensity will be marked as different by the models.

> **RQ-1.2:** Which models are best suited for the clustering of astrochemical spectra?

Out of the tested models, the agglomerative clustering model performed best. Simple models such as K-means will yield overlapping clusters due to the smooth transitions between one group and another. Other intermediate models, such as the GMM, can be useful only after the correct pre-processing has been applied. Finally, more complex models (especially NNs) where not tested but may prove to outperform agglomerative clustering models.

> **RQ-2.1:** Is there a correlation between the number of identified peaks in the spectral line content and the evolutionary stage of a core?

Yes. This is in agreement with the current knowledge about the star formation process. This investigation, however, proved with a statistically significant dataset that signals with many spectral lines will always have certain physical properties such as high L/M ratios.

> **RQ-2.2:** Which molecular groups can be used to identify a dense core's evolutionary stage?

Based on the sequence established in Chapter 4, signals with advanced evolutionary stages will show intense CH_3CN lines at the end of spw1. Furthermore, the relative intensity of the lines in spw0 will be comparable to those of spw1, whereas signals with a heavy imbalance between the two are generally found earlier in the sequence. Additionally, the SO line increases relative to the $C^{18}O$ line as the region collapses and approaches the hot-core phase.

> **RQ-2.3:** How does the spectral line content of the cores correlate to that of the cloud they are embedded in?

The spectral line content of the cloud was not specifically analyzed in this study, as the data was not made available. However, the majority of the signals display only a few peaks, and their corresponding molecules can be found in the surrounding envelope. As these regions evolve, however, the relative intensity between those lines becomes more homogeneous. This possibly implies that the evolution of a region into a hot core standardises the chemistry to some degree. This effect, however, must be local, as it was proven that a single cloud will contain cores in different evolutionary stages. Additionally, no single cloud had more than 12 different clusters, despite some regions having up to 49 cores. Together, these two factors could indicate that some evolutionary stages are locked depending on the cloud's initial conditions. It is therefore not possible to confidently answer this question with the current knowledge.

References

[1] Rutinaldo Aguiar Nascimento et al. "A new hybrid optimization approach using PSO, Nelder-Mead Simplex and Kmeans clustering algorithms for 1D Full Waveform Inversion". en. In: *PLOS ONE* 17.12 (Dec. 2022). Ed. by Seyedali Mirjalili, e0277900. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0277900. URL: https://dx.plos.org/10.1371/journal.pone.0277900 (visited on 04/14/2025).

[2] Veronica Amber Allen. *Organic chemistry around young high-mass stars: observational and theoretical*. eng. OCLC: 1056491077. Groningen: University of Groningen, 2018. ISBN: 978-94-034-1004-3.

[3] Olatz Arbelaitz et al. "An extensive comparative study of cluster validity indices". en. In: *Pattern Recognition* 46.1 (Jan. 2013), pp. 243–256. ISSN: 00313203. DOI: 10.1016/j.patcog.2012.07.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S003132031200338X (visited on 11/15/2024).

[4] Mikhail Belkin and Partha Niyogi. "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation". en. In: *Neural Computation* 15.6 (June 2003), pp. 1373–1396. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976603321780317. URL: https://direct.mit.edu/neco/article/15/6/1373-1396/6730 (visited on 11/11/2024).

[5] H. Beuther, R. Kuiper, and M. Tafalla. *Star formation from low to high mass: A comparative view*. Version Number: 1. 2025. DOI: 10.48550/ARXIV.2501.16866. URL: https://arxiv.org/abs/2501.16866 (visited on 03/10/2025).

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. eng. Softcover reprint of the original 1st edition 2006 (corrected at 8th printing 2009). Information science and statistics. New York, NY: Springer New York, 2016. ISBN: 978-1-4939-3843-8.

[7] I. A. Bonnell, M. R. Bate, and H. Zinnecker. "On the formation of massive stars". en. In: *Monthly Notices of the Royal Astronomical Society* 298.1 (July 1998), pp. 93–102. ISSN: 0035-8711, 1365-2966. DOI: 10.1046/j.1365-8711.1998.01590.x. URL: https://academic.oup.com/mnras/article/298/1/93/976453 (visited on 03/15/2025).

[8] Kirk D. Borne. *Astroinformatics: A 21st Century Approach to Astronomy*. Version Number: 1. 2009. DOI: 10.48550/ARXIV.0909.3892. URL: https://arxiv.org/abs/0909.3892 (visited on 04/07/2025).

[9] Pavla Bromová, Petr Škoda, and Jaroslav Vážný. "Classification of Spectra of Emission Line Stars Using Machine Learning Techniques". en. In: *International Journal of Automation and Computing* 11.3 (June 2014), pp. 265–273. ISSN: 1476-8186, 1751-8520. DOI: 10.1007/s11633-014-0789-2. URL: http://link.springer.com/10.1007/s11633-014-0789-2 (visited on 04/07/2025).

[10] C. Carey et al. "Machine learning tools for mineral recognition and classification from Raman spectroscopy". en. In: *Journal of Raman Spectroscopy* 46.10 (Oct. 2015), pp. 894–903. ISSN: 0377-0486, 1097-4555. DOI: 10.1002/jrs.4757. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/jrs.4757 (visited on 11/06/2024).

[11] Mathilde Caron et al. *Deep Clustering for Unsupervised Learning of Visual Features*. Version Number: 2. 2018. DOI: 10.48550/ARXIV.1807.05520. URL: https://arxiv.org/abs/1807.05520 (visited on 11/08/2024).

[12] S. Cazaux, P. Caselli, and M. Spaans. "INTERSTELLAR ICES AS WITNESSES OF STAR FORMATION: SELECTIVE DEUTERATION OF WATER AND ORGANIC MOLECULES UNVEILED". In: *The Astrophysical Journal* 741.2 (Nov. 2011), p. L34. ISSN: 2041-8205, 2041-8213. DOI: 10.1088/2041-8205/741/2/L34. URL: https://iopscience.iop.org/article/10.1088/2041-8205/741/2/L34 (visited on 03/27/2025).

[13] S. Cazaux et al. "The Hot Core around the Low-Mass Protostar IRAS 16293-2422: Scoundrels Rule!" en. In: *The Astrophysical Journal* 593.1 (Aug. 2003), pp. L51–L55. ISSN: 0004-637X, 1538-4357. DOI: 10.1086/378038. URL: https://iopscience.iop.org/article/10.1086/378038 (visited on 03/31/2025).

[14] A. Coletta et al. "ALMAGAL: III. Compact source catalog: Fragmentation statistics and physical evolution of the core population". In: *Astronomy & Astrophysics* 696 (Apr. 2025), A151. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/202452706. URL: https://www.aanda.org/10.1051/0004-6361/202452706 (visited on 09/08/2025).

[15] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". en. In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00994018. URL: http://link.springer.com/10.1007/BF00994018 (visited on 04/13/2025).

[16] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (Jan. 1967), pp. 21–27. ISSN: 0018-9448, 1557-9654. DOI: 10.1109/TIT.1967.1053964. URL: http://ieeexplore.ieee.org/document/1053964/ (visited on 04/13/2025).

[17] Simon Crase and Suresh N. Thennadil. "An analysis framework for clustering algorithm selection with applications to spectroscopy". en. In: *PLOS ONE* 17.3 (Mar. 2022). Ed. by Usman Qamar, e0266369. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0266369. URL: https://dx.plos.org/10.1371/journal.pone.0266369 (visited on 11/15/2024).

[18] T. Csengeri et al. "Chemical differentiation in the inner envelope of a young high-mass protostar associated with Class II methanol maser emission". en. In: *Proceedings of the International Astronomical Union* 13.S336 (Sept. 2017), pp. 331–333. ISSN: 1743-9213, 1743-9221. DOI: 10.1017/S1743921318000364. URL: https://www.cambridge.org/core/product/identifier/S1743921318000364/type/journal_article (visited on 03/15/2025).

[19] Ankan Das. "Astrochemistry: The study of chemical processes in space". en. In: *Life Sciences in Space Research* 43 (Nov. 2024), pp. 43–53. ISSN: 22145524. DOI: 10.1016/j.lssr.2024.10.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S2214552424000944 (visited on 03/20/2025).

[20] Ewine F. van Dishoeck and Geoffrey A. Blake. "CHEMICAL EVOLUTION OF STAR-FORMING REGIONS". en. In: *Annual Review of Astronomy and Astrophysics* 36.1 (Sept. 1998), pp. 317–368. ISSN: 0066-4146, 1545-4282. DOI: 10.1146/annurev.astro.36.1.317. URL: https://www.annualreviews.org/doi/10.1146/annurev.astro.36.1.317 (visited on 03/25/2025).

[21] A. Duarte-Cabral et al. "CO outflows from high-mass Class 0 protostars in Cygnus-X". In: *Astronomy & Astrophysics* 558 (Oct. 2013), A125. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201321393. URL: http://www.aanda.org/10.1051/0004-6361/201321393 (visited on 03/17/2025).

[22] Paul Eilers and Hans Boelens. "Baseline Correction with Asymmetric Least Squares Smoothing". In: (Nov. 2005).

[23] S Fabbro et al. "An application of deep learning in the analysis of stellar spectra". en. In: *Monthly Notices of the Royal Astronomical Society* 475.3 (Apr. 2018), pp. 2978–2993. ISSN: 0035-8711, 1365-2966. DOI: 10.1093/mnras/stx3298. URL: https://academic.oup.com/mnras/article/475/3/2978/4775133 (visited on 04/07/2025).

[24] Fabrizio De Angelis and Chiara Mininni. *ALMAGAL*. 2025. URL: https://www.almagal.org/ (visited on 02/28/2025).

[25] G. B. Field, D. W. Goldsmith, and H. J. Habing. "Cosmic-Ray Heating of the Interstellar Gas". en. In: *The Astrophysical Journal* 155 (Mar. 1969), p. L149. ISSN: 0004-637X, 1538-4357. DOI: 10.1086/180324. URL: http://adsabs.harvard.edu/doi/10.1086/180324 (visited on 03/14/2025).

[26] Luciano da Fontoura Costa. "Data Normalization in Signal and Pattern Analysis and Recognition: A Modeling Approach". In: hal-03688208 (2022). URL: https://hal.science/hal-03688208v2.

[27] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. "On Clustering Validation Techniques". In: *Journal of Intelligent Information Systems* 17.2/3 (2001), pp. 107–145. ISSN: 09259902. DOI: 10.1023/A:1012801612483. URL: http://link.springer.com/10.1023/A:1012801612483 (visited on 11/12/2024).

[28] Geoffrey Hinton and Sam Roweis. "Stochastic neighbor embedding". In: *Proceedings of the 16th International Conference on Neural Information Processing Systems*. NIPS'02. MIT Press, 2002, pp. 857–864. DOI: `https://dl.acm.org/doi/10.5555/2968618.2968725`.

[29] Tom Howley et al. "The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data". en. In: *Knowledge-Based Systems* 19.5 (Sept. 2006), pp. 363–370. ISSN: 09507051. DOI: `10.1016/j.knosys.2005.11.014`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S0950705106000141` (visited on 04/13/2025).

[30] Sascha T. Ishikawa and Virginia C. Gulick. "An automated mineral classifier using Raman spectra". en. In: *Computers & Geosciences* 54 (Apr. 2013), pp. 259–268. ISSN: 00983004. DOI: `10.1016/j.cageo.2013.01.011`. URL: `https://linkinghub.elsevier.com/retrieve/pii/S0098300413000253` (visited on 11/08/2024).

[31] Pavel Jahoda et al. "Machine Learning for recognition of minerals from multispectral data". In: (2020). Publisher: arXiv Version Number: 1. DOI: `10.48550/ARXIV.2005.14324`. URL: `https://arxiv.org/abs/2005.14324` (visited on 11/06/2024).

[32] S. S. Jensen et al. "Modeling chemistry during star formation: water deuteration in dynamic star-forming regions". In: *Astronomy & Astrophysics*. Interstellar and circumstellar matter 649.A&A (Nov. 2021), p. 21. ISSN: 0004-6361. DOI: `https://doi.org/10.1051/0004-6361/202040196`. URL: `https://www.aanda.org/articles/aa/full_html/2021/05/aa40196-20/aa40196-20.html` (visited on 02/04/2025).

[33] Jes K. Jørgensen, Arnaud Belloche, and Robin T. Garrod. "Astrochemistry During the Formation of Stars". en. In: *Annual Review of Astronomy and Astrophysics* 58.1 (Aug. 2020), pp. 727–778. ISSN: 0066-4146, 1545-4282. DOI: `10.1146/annurev-astro-032620-021927`. URL: `https://www.annualreviews.org/doi/10.1146/annurev-astro-032620-021927` (visited on 03/25/2025).

[34] K-M. Lau and Hengyi Weng. "Climate Signal Detection Using Wavelet Transform: How to Make a Time Series Sing". en. In: *Bulletin of the American Meteorological Society* 76.12 (Dec. 1995), pp. 2391–2402. ISSN: 0003-0007, 1520-0477. DOI: `10.1175/1520-0477(1995)076<2391:CSDUWT>2.0.CO;2`. URL: `http://journals.ametsoc.org/doi/10.1175/1520-0477(1995)076%3C2391:CSDUWT%3E2.0.CO;2` (visited on 04/11/2025).

[35] Jinchao Liu et al. "Deep convolutional neural networks for Raman spectrum recognition: a unified solution". en. In: *The Analyst* 142.21 (2017), pp. 4067–4074. ISSN: 0003-2654, 1364-5528. DOI: `10.1039/C7AN01371J`. URL: `https://xlink.rsc.org/?DOI=C7AN01371J` (visited on 04/13/2025).

[36] J.B. MacQueen. "Some Methods for Classification and Analysis of Multivariate Observations". In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. Unversity of California Press, 1967, pp. 281–297.

[37] M. Manteiga et al. "ANNs and Wavelets: A Strategy for *Gaia* RVS Low S/N Stellar Spectra Parameterization". en. In: *Publications of the Astronomical Society of the Pacific* 122.891 (May 2010), pp. 608–617. ISSN: 0004-6280, 1538-3873. DOI: `10.1086/653039`. URL: `http://iopscience.iop.org/article/10.1086/653039` (visited on 04/07/2025).

[38] Ryan McConville et al. *N2D: (Not Too) Deep Clustering via Clustering the Local Manifold of an Autoencoded Embedding*. Version Number: 6. 2019. DOI: `10.48550/ARXIV.1908.05968`. URL: `https://arxiv.org/abs/1908.05968` (visited on 11/08/2024).

[39] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426 [stat]. Sept. 2020. URL: `http://arxiv.org/abs/1802.03426` (visited on 11/14/2024).

[40] C. F. McKee and J. P. Ostriker. "A theory of the interstellar medium - Three components regulated by supernova explosions in an inhomogeneous substrate". en. In: *The Astrophysical Journal* 218 (Nov. 1977), p. 148. ISSN: 0004-637X, 1538-4357. DOI: `10.1086/155667`. URL: `http://adsabs.harvard.edu/doi/10.1086/155667` (visited on 03/14/2025).

[41] Christopher F. McKee and Jonathan C. Tan. "The Formation of Massive Stars from Turbulent Cores". In: (2002). Publisher: arXiv Version Number: 2. DOI: `10.48550/ARXIV.ASTRO-PH/0206037`. URL: `https://arxiv.org/abs/astro-ph/0206037` (visited on 03/15/2025).

[42] T. J. Millar and David A. Williams, eds. *Dust and chemistry in astronomy*. eng. Boca Raton, Florida ; London ; New York: CRC Press, 1993. ISBN: 978-1-351-45446-9.

[43] S. Molinari et al. "ALMAGAL: I. The ALMA evolutionary study of high-mass protocluster formation in the Galaxy: Presentation of the survey and early results". In: *Astronomy & Astrophysics* 696 (Apr. 2025), A149. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/202452702. URL: https://www.aanda.org/10.1051/0004-6361/202452702 (visited on 09/08/2025).

[44] Frederique Motte, Sylvain Bontemps, and Fabien Louvet. "High-Mass Star and Massive Cluster Formation in the Milky Way". In: (2017). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1706.00118. URL: https://arxiv.org/abs/1706.00118 (visited on 03/10/2025).

[45] Holger S.P. Müller et al. "The Cologne Database for Molecular Spectroscopy, CDMS: a useful tool for astronomers and spectroscopists". en. In: *Journal of Molecular Structure* 742.1-3 (May 2005), pp. 215–227. ISSN: 00222860. DOI: 10.1016/j.molstruc.2005.01.027. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022286005000888 (visited on 03/20/2025).

[46] H. Nomura and T. J. Millar. "The physical and chemical structure of hot molecular cores". In: *Astronomy & Astrophysics* 414.2 (Feb. 2004), pp. 409–423. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361:20031646. URL: http://www.aanda.org/10.1051/0004-6361:20031646 (visited on 03/21/2025).

[47] J. Pasquet-Itam and J. Pasquet. "Deep learning approach for classifying, detecting andpredicting photometric redshifts of quasars in the Sloan DigitalSky Survey stripe 82". In: *Astronomy & Astrophysics* 611 (Mar. 2018), A97. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201731106. URL: https://www.aanda.org/10.1051/0004-6361/201731106 (visited on 04/07/2025).

[48] K. L. Polsterer, F. Gieseke, and C. Igel. "Automatic Galaxy Classification via Machine Learning Techniques: Parallelized Rotation/Flipping INvariant Kohonen Maps (PINK)". In: 495 (Sept. 2015). Conference Name: Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV) ADS Bibcode: 2015ASPC..495...81P, p. 81. URL: https://ui.adsabs.harvard.edu/abs/2015ASPC..495...81P (visited on 03/02/2025).

[49] V. M. Rivilla et al. "The role of low-mass star clusters in massive star formation. The Orion case". In: *Astronomy & Astrophysics* 554 (June 2013), A48. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201117487. URL: http://www.aanda.org/10.1051/0004-6361/201117487 (visited on 03/15/2025).

[50] Anna L. Rosen et al. "Zooming in on Individual Star Formation: Low- and High-Mass Stars". en. In: *Space Science Reviews* 216.4 (June 2020), p. 62. ISSN: 0038-6308, 1572-9672. DOI: 10.1007/s11214-020-00688-5. URL: https://link.springer.com/10.1007/s11214-020-00688-5 (visited on 03/10/2025).

[51] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". en. In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. ISSN: 03770427. DOI: 10.1016/0377-0427(87)90125-7. URL: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257 (visited on 04/14/2025).

[52] Á. Sánchez-Monge et al. *ALMAGAL II. The ALMA evolutionary study of high-mass protocluster formation in the Galaxy. ALMA data processing and pipeline*. Version Number: 1. 2025. DOI: 10.48550/ARXIV.2503.05559. URL: https://arxiv.org/abs/2503.05559 (visited on 09/15/2025).

[53] Abraham. Savitzky and M. J. E. Golay. "Smoothing and Differentiation of Data by Simplified Least Squares Procedures." en. In: *Analytical Chemistry* 36.8 (July 1964), pp. 1627–1639. ISSN: 0003-2700, 1520-6882. DOI: 10.1021/ac60214a047. URL: https://pubs.acs.org/doi/abs/10.1021/ac60214a047 (visited on 04/11/2025).

[54] Vasileios Sevetlidis and George Pavlidis. "Effective Raman spectra identification with tree-based methods". en. In: *Journal of Cultural Heritage* 37 (May 2019), pp. 121–128. ISSN: 12962074. DOI: 10.1016/j.culher.2018.10.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S1296207418303595 (visited on 11/06/2024).

[55] Dalwinder Singh and Birmohan Singh. "Investigating the impact of data normalization on classification performance". en. In: *Applied Soft Computing* 97 (Dec. 2020), p. 105524. ISSN: 15684946. DOI: 10.1016/j.asoc.2019.105524. URL: https://linkinghub.elsevier.com/retrieve/pii/S1568494619302947 (visited on 11/06/2024).

[56] T. Sørensen. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. Biologiske skrifter. Munksgaard, 1948. URL: https://books.google.es/books?id=rpS8GAAACAAJ.

[57] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction". en. In: *Science* 290.5500 (Dec. 2000), pp. 2319–2323. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.290.5500.2319. URL: https://www.science.org/doi/10.1126/science.290.5500.2319 (visited on 11/11/2024).

[58] A. G. G. M. Tielens. "The molecular universe". en. In: *Reviews of Modern Physics* 85.3 (July 2013), pp. 1021–1081. ISSN: 0034-6861, 1539-0756. DOI: 10.1103/RevModPhys.85.1021. URL: https://link.aps.org/doi/10.1103/RevModPhys.85.1021 (visited on 03/25/2025).

[59] Universität zu Köln. *ALMAGAL*. 2019. URL: https://astro.uni-koeln.de/schilke/research/almagal (visited on 02/28/2025).

[60] Enrique Vázquez-Semadeni et al. "Global hierarchical collapse in molecular clouds. Towards a comprehensive scenario". en. In: *Monthly Notices of the Royal Astronomical Society* 490.3 (Dec. 2019), pp. 3061–3097. ISSN: 0035-8711, 1365-2966. DOI: 10.1093/mnras/stz2736. URL: https://academic.oup.com/mnras/article/490/3/3061/5580657 (visited on 03/18/2025).

[61] Enrique Vázquez-Semadeni et al. "HIGH- AND LOW-MASS STAR-FORMING REGIONS FROM HIERARCHICAL GRAVITATIONAL FRAGMENTATION. HIGH LOCAL STAR FORMATION RATES WITH LOW GLOBAL EFFICIENCIES". In: *The Astrophysical Journal* 707.2 (Dec. 2009), pp. 1023–1033. ISSN: 0004-637X, 1538-4357. DOI: 10.1088/0004-637X/707/2/1023. URL: https://iopscience.iop.org/article/10.1088/0004-637X/707/2/1023 (visited on 03/18/2025).

[62] Derek Ward-Thompson and Anthony P. Whitworth. *An introduction to star formation*. eng. 1. paperback ed. Cambridge: Cambridge University Press, 2015. ISBN: 978-0-521-63030-6 978-1-107-48352-1.

[63] Mark G. Wolfire and Michael J. Kaufman. "Photodissociation Regions". en. In: *Encyclopedia of Astrobiology*. Ed. by Muriel Gargaud et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1236–1243. ISBN: 978-3-642-11271-3 978-3-642-11274-4. DOI: 10.1007/978-3-642-11274-4_1197. URL: http://link.springer.com/10.1007/978-3-642-11274-4_1197 (visited on 04/15/2025).

[64] Magnus M. Woods, Alberto Sainz Dalda, and Bart De Pontieu. "Unsupervised Machine Learning for the Identification of Preflare Spectroscopic Signatures". In: *The Astrophysical Journal* 922.2 (Dec. 2021), p. 137. ISSN: 0004-637X, 1538-4357. DOI: 10.3847/1538-4357/ac2667. URL: https://iopscience.iop.org/article/10.3847/1538-4357/ac2667 (visited on 11/08/2024).

[65] Junyuan Xie, Ross Girshick, and Ali Farhadi. *Unsupervised Deep Embedding for Clustering Analysis*. Version Number: 2. 2015. DOI: 10.48550/ARXIV.1511.06335. URL: https://arxiv.org/abs/1511.06335 (visited on 11/12/2024).

[66] Yang Xu and Rachel Patton McCord. "CoSTA: unsupervised convolutional neural network learning for spatial transcriptomics analysis". en. In: *BMC Bioinformatics* 22.1 (Dec. 2021), p. 397. ISSN: 1471-2105. DOI: 10.1186/s12859-021-04314-1. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04314-1 (visited on 11/07/2024).

[67] Guang Yang et al. "Manifold Learning in MR spectroscopy using nonlinear dimensionality reduction and unsupervised clustering". en. In: *Magnetic Resonance in Medicine* 74.3 (Sept. 2015), pp. 868–878. ISSN: 0740-3194, 1522-2594. DOI: 10.1002/mrm.25447. URL: https://onlinelibrary.wiley.com/doi/10.1002/mrm.25447 (visited on 11/08/2024).

[68] Jianwei Yang, Devi Parikh, and Dhruv Batra. *Joint Unsupervised Learning of Deep Representations and Image Clusters*. Version Number: 3. 2016. DOI: 10.48550/ARXIV.1604.03628. URL: https://arxiv.org/abs/1604.03628 (visited on 11/12/2024).

[69] Heng Yu and Xiaolan Hou. "Hierarchical clustering in astronomy". en. In: *Astronomy and Computing* 41 (Oct. 2022), p. 100662. ISSN: 22131337. DOI: 10.1016/j.ascom.2022.100662. URL: https://linkinghub.elsevier.com/retrieve/pii/S2213133722000762 (visited on 03/19/2025).

[70] Tianwei Zhang. "Characterization of the star-forming region G327.3-0.6". PhD thesis. Cologne: University of Cologne, 2020.

[71] Wei Zhang et al. *Graph Degree Linkage: Agglomerative Clustering on a Directed Graph*. Version Number: 1. 2012. DOI: `10.48550/ARXIV.1208.5092`. URL: `https://arxiv.org/abs/1208.5092` (visited on 11/12/2024).

A

Metrics

A.1. Distance metrics

A distance metric is a way of evaluating how different two samples are. The most common metric is the **Euclidean distance**, which is calculated using Equation A.1. This metric is the standard metric used in most models, including `sklearn`'s `KMeans` classifier class. This metric, however, places the same importance on each channel independently of the intensity of the signal. This may not be desirable as the distance between channels that show great intensity will be just as relevant as that of channels with just random noise.

$$d_e = ||\vec{u} - \vec{v}||_2 \tag{A.1}$$

An alternative to the Euclidean distance is the **cosine distance** metric. This metric has been used in literature for the analysis of chemical spectra with a higher performance than the Euclidean distance [10, 31]. This distance, calculated using Equation A.2, is bounded between 0 and +2, with a score of 0 indicating that the outline between two signals is the same, even if their scale is different. This metric places more importance on channels with a higher intensity than those with a lower one, therefore reducing the effects of random noise on the metric.

$$d_c = 1 - \frac{\vec{u} \cdot \vec{v}}{|u|\,|v|} \tag{A.2}$$

To illustrate the difference between both, the distance between various sample signals will be calculated. The first signal selected is the mean spectrum of all observed spectra in the dataset. The second signal, on the other hand, will be the same spectra but scaled by a factor of 2. Finally, the third signal will consist of the first signal with a random noise component of amplitude equal to 0.2 Jy. These signals can be seen in Figure A.1, and their distances with each metric can be found in Table A.1.

Table A.1: Distances between the three sample signals in Figure A.1 using both the Euclidean and Cosine distance metrics.

(a) Euclidean distance

d_e	Original	Scaled	Noisy
Original	0.00	6.819	8.175
Scaled	6.819	0.00	10.667
Noisy	8.17	10.667	0.00

(b) Cosine distance

d_e	Original	Scaled	Noisy
Original	0.00	0.00	0.361
Scaled	0.00	0.00	0.361
Noisy	0.361	0.361	0.00

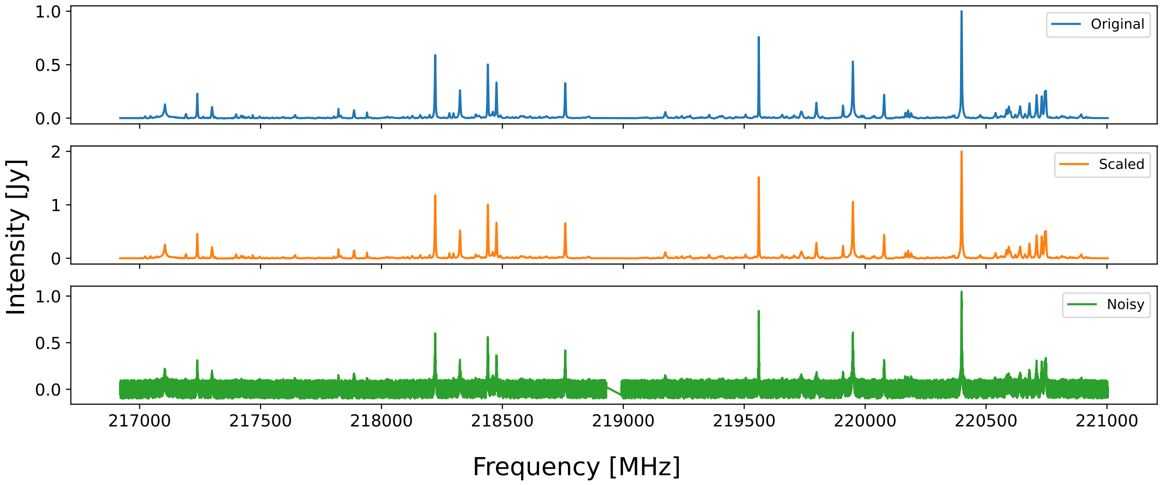


Figure A.1: Sample signals used to compare distance metrics.

As seen in Table A.1, the Euclidean distance between the original and scaled signals is equal to the original signal's magnitude, even though their shape is exactly the same. This is also the case for the noisy signal, which has a Euclidean distance to the original signal almost 50% greater than that of the scaled signal. This is not the case with the cosine distance, however, as the original and scaled signals have a null distance and the distance to the noisy signal is 0.436 out of a maximum value of 2.

A.2. Evaluation metrics

These metrics measure the quality of the clusters. Since the true labels of the signals are unknown, it is impossible to evaluate the models based on their accuracy or precision. However, it is still possible to determine the quality of a cluster based on two factors: how similar are the signals of one cluster to one another and how dissimilar are those same signals to the signals in other clusters. There are a number of metrics that can be used for this purpose, and readers are encouraged to read the reviews from Arbelaitz et al. [3] and Halkidi, Batistakis, and Vazirgiannis [27] on clustering validation techniques. Based on these reviews, and on the works by Bromová, Škoda, and Vážný [9] and Crase and Thennadil [17] on the clustering of spectroscopic data, the first evaluation metric used was the **silhouette score**. This metric, calculated using Equation A.3, measures both the compactness of the clusters (how small are the distances between members of a single cluster) and how well separated are the clusters.

$$s_s = \frac{b-a}{max(a,b)} \tag{A.3}$$

Where:

- a is the average distance between a signal and the other signals in its cluster.
- b is the average distance between a signal and all signals in different clusters.

The score is bounded between -1 and +1 and can be interpreted as follows:

- $s_s \sim -1$: the signals have been misclassified.
- $s_s \sim 0$: the clustering is no better than a random assignation.
- $s_s \approx +1$: the clusters are compact (low distance between members of the cluster) and well separated (long distance between clusters).

The performance of this metric can be evaluated using a 2D example. Using sklearn's `make_blobs` function, three separate datasets were created. The first one consists of 300 signals grouped into 3 well-defined clusters. The second one, on the other hand, consists of 500 signals grouped into 5 clusters where two of them overlap. Finally, the third one consists of 500 signals with just one center but with 5 randomly assigned labels. All datasets and their silhouette score can be seen in Figure A.2.

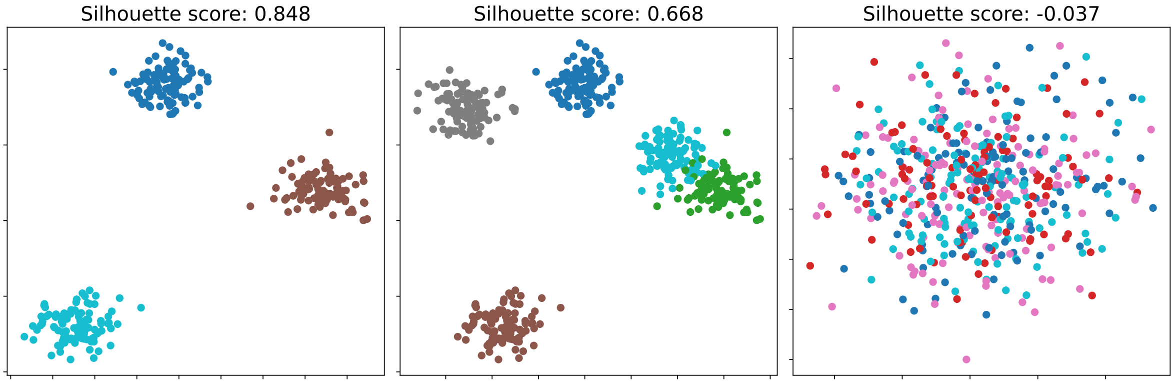


Figure A.2: Sample clusterings and their silhouette scores.

The second evaluation metric, inspired by Carey et al. [10], is the **within-cluster similarity** (WCS). The WCS measures the average cosine similarity of the signals within a cluster. Complementary to this, I also designed the **between-cluster similarity** (BCS), which measures the average cosine similarity between signals of different clusters. Ideally, a cluster should have a high WCS and low BCS.

B

Red-shift correction algorithms

Algorithm 1 Red-shift correction

```
Require: reference, $v_{\text{min}}$, $v_{\text{max}}$, $dv_0$, $dv_1$
  $v_{\text{best}} \leftarrow 0$
  $v \leftarrow v_{\text{min}} - dv_0$
  $s \leftarrow 0$
  $s_{\text{max}} \leftarrow s$
  while $v \leq v_{\text{max}}$ do                              ▷ Coarse search
      $v \leftarrow v + dv_0$
      $s \leftarrow \text{sim}(\text{reference}, \text{shift}(\text{signal}, v))$
      if $s > s_{\text{max}}$ then
          $v_{\text{best}} \leftarrow v$
          $s_{\text{max}} \leftarrow s$
      end if
  end while
  $v \leftarrow v_{\text{best}} - dv_0$
  $v_{\text{max},2} \leftarrow v + 2 \cdot dv_0$
  while $v < v_{\text{max},2}$ do                               ▷ Fine search
      $v \leftarrow v + dv_1$
      $s_{\text{new}} \leftarrow \text{sim}(\text{reference}, \text{shift}(\text{signal}, v))$
      if $s_{\text{new}} > s_{\text{max}}$ then
          $v_{\text{best}} \leftarrow v$
          $s_{\text{max}} \leftarrow s_{\text{new}}$
      end if
  end while
Output: $v_{\text{best}}$
```

Algorithm 2 Fast red-shift correction

```
Require: reference, $v_{\min}$, $v_{\max}$, $dv$, $tol$
  $v_{\text{best}} \leftarrow 0$
  $v \leftarrow v_{\min} - dv$
  $s \leftarrow 0$
  $s_{\max} \leftarrow s$
  while $v \leq v_{\max}$ do                                  ▹ Coarse search
      $v \leftarrow v + dv$
      $s \leftarrow \text{sim}(\text{reference}, \text{shift}(\text{signal}, v))$
      if $s > s_{\max}$ then
          $v_{\text{best}} \leftarrow v$
          $s_{\max} \leftarrow s$
      end if
  end while
  $v \leftarrow v_{\text{best}} - dv$
  $dv \leftarrow dv/5$
  while abs($dv$) > $tol$ do                                  ▹ Fine search
      $v \leftarrow v + dv$
      $s_{\text{new}} \leftarrow \text{sim}(\text{reference}, \text{shift}(\text{signal}, v))$
      if $s_{\text{new}} < s$ then
          $dv \leftarrow -dv/2$
      end if
      $s \leftarrow s_{\text{new}}$
  end while
Output: $v$
```

C

Data and code availability

The code used to generate the results presented in this thesis is available at `https://github.com/javialonso05/MSc-Thesis`