UNIVERSITY COLLEGE LONDON

Faculty of Mathematical and Physical Sciences

Department of Physics & Astronomy

DECODING ASTRONOMICAL SPECTRA USING MACHINE LEARNING

Thesis submitted for the Degree of

Doctor of Philosophy

by

Damien de Mijolla

Supervisors:

Prof. Serena Viti

Prof. Ioanna Manolopoulou

Examiners:

Prof. Ofer Lahav

Dr. Payel Das

June 27, 2022

I, Damien de Mijolla, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Spectroscopy is one of the cornerstones of modern astronomy. Using spectra, the light from far-away objects measured on Earth can be related back to the physical and chemical conditions of the astronomical matter from which it is emitted. This makes spectroscopy an essential tool for constraining the physical and chemical conditions of the matter in stars, gas, galaxies and all other types of astronomical objects. However, whilst spectra carry a wealth of astronomical information, their analysis is often complicated by difficulties such as degeneracies between input parameters and gaps in our theoretical knowledge.

In this thesis, we look towards the rapidly growing field of machine learning as a means of better extracting the information content of astronomical spectra. Chapters 2 and 3 of the thesis are dedicated to the study of spectra originating from the interstellar medium. Chapter 2 of this thesis presents a machine learning emulator for the UCLCHEM astrochemical code which when combined with a Bayesian treatment of the radiative-transfer inverse problem enables a rigorous handling of the degeneracies affecting molecular lines (all within short enough computational timescales to be tractable). Chapter 3 extends upon the work of Chapter 2 on modelling molecular lines and investigates the appropriateness of Non-negative Matrix Factorization, a blind source separation algorithm, for the task of unmixing the gas phases which may exist within molecular line-intensity maps.

Chapter 4 and 5 are concerned with the analysis of stellar spectra. In these chapters, we introduce machine learning approaches for extracting the chemical content from stellar spectra which do not rely on manual spectral modelling. This removes the burden of building faithful forward-models of stellar spectroscopy in order to precisely extract the chemistry of stars. The two approaches are also complimentary. Chapter 4 presents a deep-learning approach for distilling the information content within stellar spectra into a

representation where undesirable factors of variation are excluded. Such a representation can then be used to directly find chemically identical stars or for differential abundance analysis. However, the approach requires measurements of the to-be-excluded undesirable factors of variation. The second approach which is presented in Chapter 5 addresses this shortcoming by learning which factors of variation should be excluded using spectra of open clusters. However, because of the low number of known open clusters, whilst the method constructed in Chapter 4 is non-linear and parametrized by a feedforward neural network, the approach presented in Chapter 5 was made linear.

Impact statement

This thesis presents original research on using machine learning approaches to extract the information content from astronomical spectra. The work within lies at the intersection between the fields of astronomy and machine learning, but with contributions primarily targeted towards astronomy audiences.

The work presented in Chapter 2 describes an emulator for accelerating the astrochemical code UCLCHEM. Using the emulator offers multiple factors of magnitude speed-up over running UCLCHEM directly which enables the application of chemical modelling to many previously unfeasible problems. In particular, the emulator allows for incorporating astrochemical information within the radiative-transfer inverse problem needing solving for the interpretation of molecular-line observations of the interstellar medium. We also publicly release on github the source code for the emulator to the wider research community.

The work presented in Chapter 3 describes experiments testing the effectiveness of a blind-source separation algorithm for the task of unmixing gas components in observations of the interstellar medium. This offers another approach for addressing the degeneracies within radiative-transfer modelling of the interstellar medium.

Chapter 4 describes a new approach for measuring the chemical similarity of stars from spectra that reduces the need for accurate stellar modelling. As accurate synthetic spectra are often a bottleneck in the determination of abundances, this new approach could potentially measure chemical similarity much more accurately than is currently possible.

Chapter 5 builds upon the approach developed in Chapter 4 and presents an extension in which open-clusters are used to find other near-chemically identical stars. This work lays the foundations for a new, entirely data-driven, approach of analysing stellar spectra

that has the advantages of being free from the biases introduced by synthetic spectra and applicable to spectra for which accurate synthetic spectra may not exist. On tests applied to existing stellar survey data, we found our developed approach to compare favorably with existing methods.

The work in Chapter 2 has been published in Astronomy & Astrophysics, the work in Chapter 4 has been published in the Astrophysical Journal and the work in Chapter 5 has been submitted to the Astrophysical Journal. In addition, the work presented in this thesis has also been presented at research conferences and at presentations in other universities as well as cited by independent researchers.

Finally, this PhD was undertaken as part of a Centre for Doctoral Training in Data Intensive Science. This has meant that, in addition to the astronomical research presented in this thesis, I have taken courses in data intensive science, and applied my newfound skills in an industrial environment through group projects and a six-month secondment. During my secondment, which took place at Faculty Science Ltd, I contributed to research on machine learning explainability with the placement resulting in two publications, "Human-interpretable model explainability on high-dimensional data" (de Mijolla et al. 2020) which is a preprint available on arxiv for which I am joint first author and "Shapley explainability on the data manifold"(Frye et al. 2021) which was published at the International Conference on Learning Representations (ICLR) and for which I am second author.

Acknowledgements

The past four years, during which I have been working on my PhD, whilst stressful have also been some of the most fun and formative years of my life. And whilst the positive environment that enabled this was the product of far too many individuals to individually name, I would still like to highlight a few as it would feel unfair to them to not do so.

I am immensely grateful to my supervisor Serena Viti for fostering an exceptional positive research environment. Even during the pandemic and after her move to Leiden, she has never felt farther away than a ten minute wait for an email reply, whilst also somehow giving me the space to build my own independent research projects. I am also grateful to all the other role models I have met along the way. I am grateful to Melissa Ness for her patience, guidance and expertise on our collaborations, to Jon Holdship for his support and cheerful attitude and to Ioanna Manolopoulou for her selfless help.

I also owe a debt of gratitude to all the other friends I've made along the way, who have kept me entertained throughout this journey. Especially Davide, Gordon, Greg, James, Johannes, Mala, Marcus, Maria, Paddy and all my other officemates past and present. I would also like to thank the whole administrative team behind the Physics & Astronomy Department and the Center for Doctoral Training in Data Intensive Science.

These thanks would not be complete without acknowledging my parents and sisters for their unwavering support and for providing, throughout my lifetime, the supportive environment which made me who I am today.

Finally, but perhaps most importantly, I would like to thank Maggie, my partner become coworker during the pandemic, who has been next to me at every step of the PhD journey providing me with her unwavering support and enriching every day along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Whilst many of the core ideas underpinning modern machine learning were conceptualized in the second half of the 20$^{\text{th}}$ century, such as the perceptron by Rosenblatt (Rosenblatt 1958) in the 1950s, the modern version of the backpropagation algorithm by Mater's student Linnainmaa (Linnainmaa 1976) in the 1970s, the convolutional neural network by Lecun in 1989 (LeCun et al. 1989) and the long short-term memory network in 1997 Hochreiter & Schmidhuber (1997), few scientific fields have ever seen as rapid progress as machine learning has over the last decade, from AlexNet winning the ImageNet competition in 2012 (Krizhevsky et al. 2012), to AlphaZero unequivocally beating Lee Sedol in the game of GO in 2015 (Silver et al. 2017) - a game long thought to be a bastion for human intelligence, to breakthroughs in protein folding in 2020 (Senior et al. 2020). With such a rapid pace of growth, machine learning has established itself as a new tool in the arsenal of the modern scientist. But as with any other newly released tool, its limits and strengths are not fully understood by practitioners initially. Rapid adoption in the sciences requires a cross-disciplinary research effort for new ideas to permeate.

This thesis presents such a cross-disciplinary research effort. The subject of this thesis is the application of machine learning techniques towards the determination of astronomical chemical abundances from spectroscopy. Two chapters are dedicated to the determination and use of molecular abundances occurring in the interstellar medium. The interstellar medium is diffuse and cold and hosts a rich molecular chemistry. In this context, deriving abundances is difficult as molecules have many formation and destruction

mechanisms which are all highly sensitive to the physical conditions of the astronomical environment. Deploying machine learning also presents its challenges as sub-mm observations - as required to measure molecular abundances - cover non-uniform wavelengths ranges, region sizes and astronomical environments.

Two further chapters are dedicated to the determination of stellar abundances from spectra. In this context, telescopes have gathered extremely large datasets, containing hundreds of thousands of stellar spectra. These spectra cover uniform wavelength ranges and have high signal-to-noise. Additionally, the physics governing the spectra is fairly well understood. The challenge however lies in extracting extremely precise measurements of stellar elemental abundances. Such extremely precise stellar abundances have uses in downstream astronomical applications. For example they can be used in galactic archaeology, the branch of astronomy making use of stars and stellar populations to probe the formation history of galaxies, including our own.

1.1 Spectroscopy

Astronomy is one of the few branches of science where it is not possible to interact with the system being studied. Instead, astronomers must rely on interpreting the light and other forms of radiation emitted by regions of interest.

Materials emit, absorb and scatter light. And so spectra, which are measurements of the amount of light received at different wavelengths are rich in information about the composition of astronomical regions (see Figure 1.1 for an example stellar spectrum). However, interpreting such spectra is a difficult task. Radiative transfer - the physical theory describing the transfer of electromagnetic radiation through matter- is the bridge for relating spectra obtained on Earth back to the physical and chemical conditions of the astronomical regions being observed.

Within this framework, the radiative transfer equation is used to parameterize the change in intensity I_ν of a beam of light at a wavelength ν, along a path ds through a material:

$$\frac{dI_\nu}{ds} = j_\nu - \alpha_\nu I_\nu \tag{1.1}$$

in this equation j_ν is referred to as the emission coefficient and α_ν as the absorption coefficient. Together, the values of j_ν and α_ν capture all emission, absorption and scattering

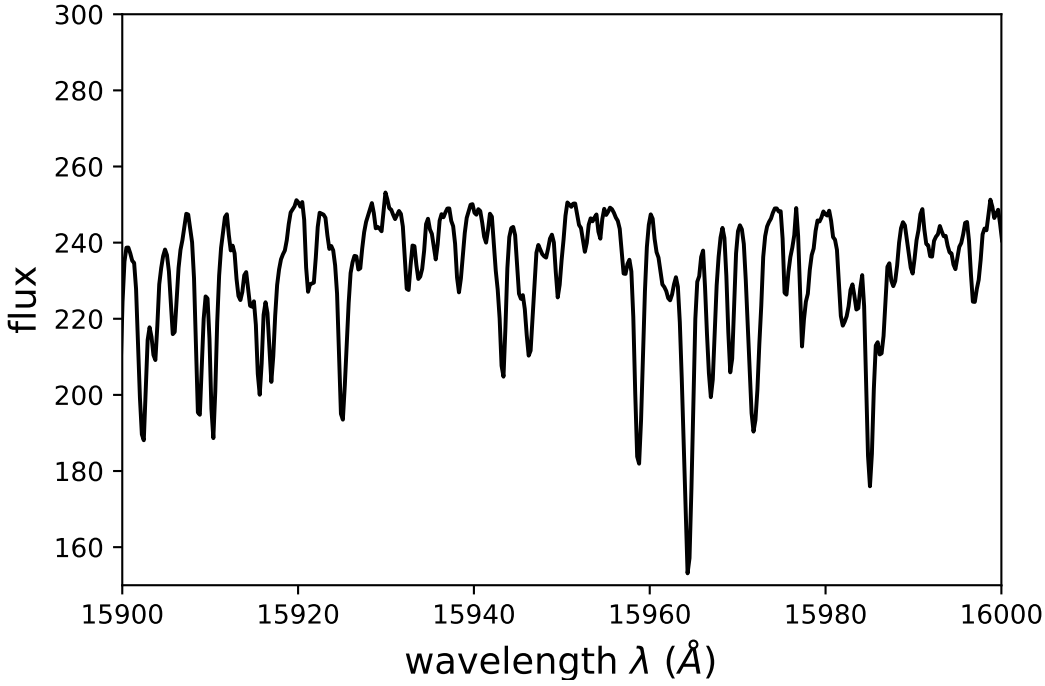

Figure 1.1: *Example of a stellar spectrum. Dips correspond to absorption lines. This spectrum was obtained by the APOGEE survey (Ahumada et al. 2020; Wilson et al. 2019) and, for clarity, we only show a small subset of the wavelength range measured by the APOGEE instrument.*

processes occurring in the beam at the given frequency ν. Their value will depend on the wavelength ν and on the chemical composition of the material (Peraiah 2001).

Defining the emissivity per unit optical depth $d\tau_\nu$ where $d\tau_\nu = \alpha_\nu ds$ and the source function $S_\nu = \frac{j_\nu}{\alpha_\nu}$ allows for integrating over the path of the beam yielding:

$$I_\nu(\tau_\nu) = I_\nu(0)e^{-\tau_\nu} + \int_0^{\tau_\nu} S_\nu(\tau'_\nu)e^{-(\tau_\nu - \tau_{\nu'})}d\tau_{\nu'} \tag{1.2}$$

As atoms and molecules reside in discrete energy levels they emit and absorb photons with specific energies corresponding to the differences between energy levels. Such emission and absorption lead to narrow spectral features known as absorption and emission lines. These spectral lines are unique to every chemical species and so can be used to work out the chemical composition of astronomical regions of interest.

Spectral lines are incorporated into the framework of radiative transfer through the absorption and emission coefficients. With the contribution of a transition from an upper

(u) to lower (l) energy state to the emission and absorption coefficients at the wavelength of the spectral line given by:

$$j_{ul} = n_u A_{ul}, \tag{1.3}$$

and

$$\alpha_{ul} = n_l B_{lu} - n_u B_{ul}, \tag{1.4}$$

where n_u and n_l are the population density of the upper and lower states, A_{ul} is the Einstein coefficient parameterizing the rate of spontaneous decay to the lower energy level and B_{lu} and B_{ul} are the Einstein coefficients for excitation and de-excitation.

The population densities required for calculating the emission and absorption coefficients can be found by assuming statistical equilibrium between the levels:

$$\frac{dn_i}{dt} = \sum_{j>i} \left[n_j A_{ji} + \left(n_j B_{ji} - n_i B_{ij}\right) J_\nu\right] - \sum_{j<i} \left[n_i A_{ji} + \left(n_i B_{ij} - n_j B_{ji}\right) J\right] + \sum_{j\neq i} \left[n_j C_{ji} - n_i C_{ij}\right] = 0, \tag{1.5}$$

where we have introduced coefficients for collisional excitation and de-excitation C_{ji} and C_{ij} and J_ν is the mean intensity per solid-angle for the given frequency:

$$J_\nu = \frac{1}{4\pi} \int I_\nu d\Omega d\nu \tag{1.6}$$

Typically, the mean intensity J_ν and level populations n will vary spatially. Fully solving these set of coupled equations - as required for finding the strength of spectral lines using radiative transfer - is thus time-consuming. Oftentimes shortcuts are taken to estimate the strength of emission and absorption lines. Many codes for analyzing molecular lines, such as RADEX (van der Tak et al. 2007), make simplifying assumptions which allow for decoupling the calculations of the level populations and the mean intensity. Also common, is the even stronger assumption that the gas is in local thermodynamic equilibrium in which case the population levels of states are entirely governed by the temperature of the gas.

An additional complexity in astronomical spectroscopy is that the processes underpinning radiative-transfer are probabilistic and so spectra are affected by sampling noise

(noise from measuring a finite sample of photons). Because of this, it is often not sufficient to just observe an object of interest but it is also necessary to observe it for long enough to measure a statistically significant number of photons at each wavelength of interest. In practice, this signal-to-noise ratio of spectra, will depend on many factors such as the apparent brightness of the object, the required wavelength resolution, the telescope used and the duration of observations.

Spectroscopy is ubiquitous in astronomy but in this thesis we concentrate on two specific applications: The analysis of stellar spectra and the analysis of molecular lines in the interstellar medium.

1.2 Molecular abundances

1.2.1 The Interstellar Medium

When thinking about the constituents of our galaxy, it is easy to forget about the interstellar medium (ISM) - the gas occupying the vast space between stars. Yet, that would be a glaring omission. The interstellar medium is a vital component of our galaxy, the Milky Way, and all other galaxies. It is from the interstellar medium that stars are born and die. It also contains a significant fraction of a galaxy's mass, estimated for the Milky Way to be equivalent to roughly around 15% of its stellar mass (Kalberla & Kerp 2009; Klessen & Glover 2014).

Although we speak of *the* interstellar medium, the interstellar medium is a rich and multifaceted environment and so, when studying it, it may be more appropriate to further sub-categorize it. In some locations, the interstellar medium is cold and dense, for example reaching temperatures of 10K and densities of $10^6 \mathrm{cm}^{-3}$ in prestellar cores, in other locations it is hot and diffuse, as is the case for the diffuse coronal gas in between stars. In Table 1.1, re-transcribed from Williams & Viti (2013), we summarize the main phases of the ISM and their physical conditions.

The interstellar medium is primarily composed of gas (99% of mass) but also contains dust grains (1% of mass (Savage & Mathis 1979)). This gas is itself mostly composed of lighter elements with heavier elements being much less abundant. Within our galaxy at present times the vast majority of the gas is in the form of hydrogen. In terms of gas mass, 70% of the ISM gas in the form of Hydrogen, 28% in the form of Helium and the remaining 2% in the form of Carbon and other heavier elements (Asplund et al. 2009;

Region	Temperature(K)	Density(cm^{-3})
Coronal gas	5×10^5	$< 5 \times 10^{-2}$
HII regions	10^4	> 100
Diffuse gas	70	100–300
Molecular clouds	5–50	10^3–10^5
Prestellar cores	10–30	10^5–10^6
Star-forming regions	100–300	10^7–10^8
Protoplanetary disks	10–500	10^4–10^{10}
Envelopes of evolved stars	2000–3500	10^{10}

Table 1.1: *Regions of the ISM and their physical conditions (re-transcribed from Williams & Viti (2013)).*

.

Klessen & Glover 2014).

The elemental composition of the ISM is however not a static quantity. Heavier elements are constantly being formed within stars and supernova in a process known as stellar nucleosynthesis (Johnson 2019). These heavier elements are then released into the ISM through stellar winds or supernovae where they enrich it with their newly formed heavier elements. As a natural consequence of this enrichment of the ISM, the metallicity - the fractional abundance of heavy elements with respect to solar abundance values- increases over time. Additionally, because of difference in the yield of stellar nucleosynthesis pathways (supernova, neutron star merger...), small local variations in the ISM elemental composition are believed to exist (Lebouteiller & Kunth 2005). Such differences in composition are a cornerstone of strong chemical tagging, an astronomy technique of interest in this thesis.

Dust grains in the ISM are formed from the accretion of carbonaceous and silicate material as well as Polycyclic aromatic hydrocarbons (PAHs)(Draine 2003). Dust grains play an important role as a chemical catalyzer. In a process known as freeze-out, atoms and molecules can stick to the surface of a dust grain and form ice mantles. Many complex chemical species which cannot reliably be formed in the gas phase are formed on icy mantles. This is for example the case for H_2, the most abundant molecule in the ISM (Wakelam et al. 2017). These newly formed molecules can then be released back into the gas phase through desorption from the dust grains (Roberts et al. 2007).

We show in Figure 1.2 a schematic depiction of the cyclic processes occurring in the ISM. Stellar ejecta mixes and enriche the diffuse interstellar medium. This newly enriched material then makes its way into molecular clouds where it forms the next generation of stars and planets. These newly formed stars will then themselves contribute to enriching

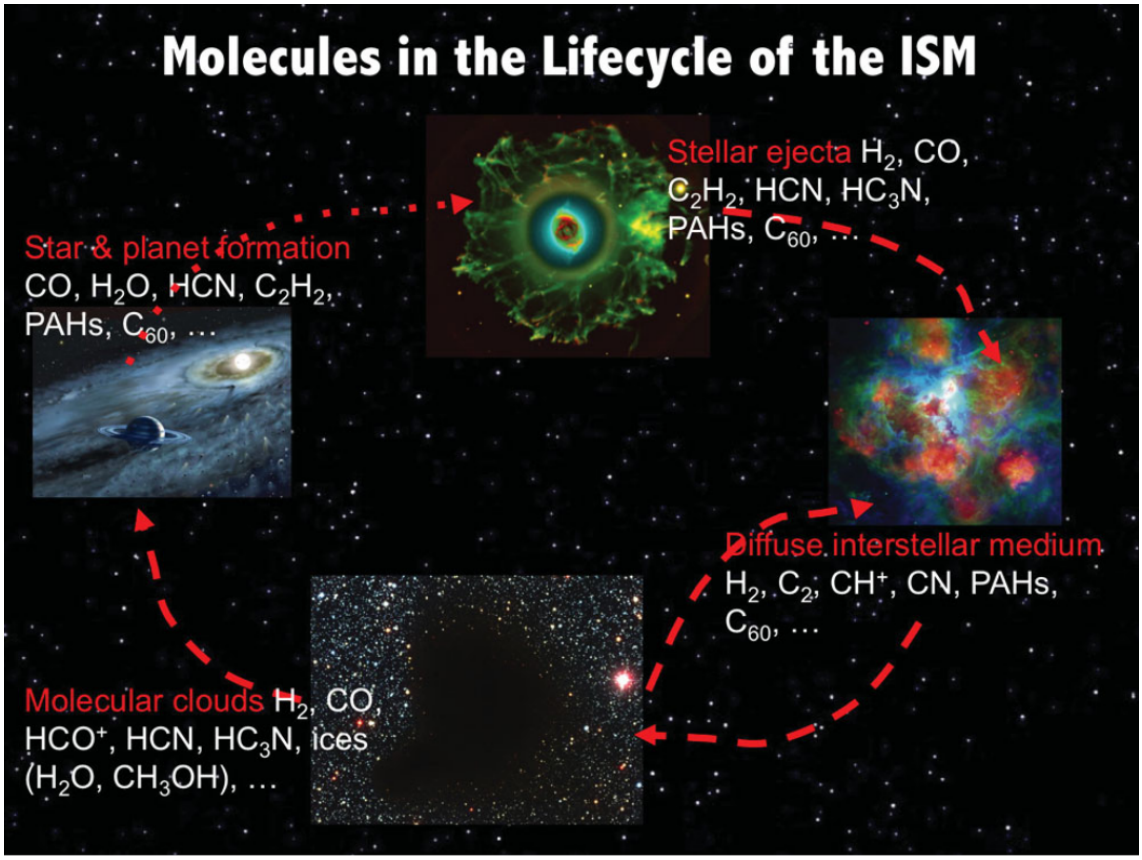

Figure 1.2: *Schematic depiction of the cyclic processes occurring in the ISM (original image from Tielens (2013)).*

the interstellar medium in the future. Every phase in this cyclic process of the ISM is characterized by distinct molecular chemical signatures. The molecules characteristic of each phase are shown in the diagram.

1.2.2 Molecular signature of the ISM

With such a diversity of astronomical environments, the ISM gas can exist in many different forms. In diffuse regions, where photons easily ionize and dissociate the gas, it is primarily in atomic form. In dark interstellar clouds, also sometimes referred to as stellar nurseries, where the high densities shield the center of clouds from photon-ionization and photo-dissociation and only allow the more energetic cosmic-rays to enter, the interstellar medium hosts a complex and rich molecular chemistry. To date, more than 200 distinct molecules have been detected in the ISM (van Dishoeck 2017).

The molecular composition of the ISM is highly-dependent on physical conditions. For example, silicon based molecules, such as SiO, are overabundant in the gas phase

in shocked regions due to sputtering off of ice grains triggered by shocks Holdship et al. (2019); Langer & Glassgold (1990). HCN and other more complex organic molecules can only efficiently be formed in high-density regions as found in bound molecular clouds Kauffmann, Jens et al. (2017). CO is relatively easy to form and so is abundant across a wide range of regions. These varieties of formation conditions enable molecular species to act as tracers of the physical conditions of the interstellar medium Bolatto et al. (2013). This is particularly important as dense ISM regions, where the bulk of molecular species form, are also regions which are opaque to direct observation at visible wavelengths.

But beyond simple heuristics, it is a complex problem to interpret the ISM molecular composition. There exist many formation and destruction pathways for molecular species and the strength of these pathways is dependent on the physical parameters of the environment (temperature, density etc). A thorough understanding of the molecular composition of the ISM requires a thorough understanding of all of these pathways which is still an ongoing research effort.

It is through the analysis of molecular lines and the use of radiative-transfer modelling that the molecular composition of astronomical regions can be determined. Molecular lines occur primarily within the sub-millimeter and far-infrared and so their ground-based measurement require the use of radio telescopes. Examples of such telescopes are the James Clerk Maxwell Telescope (JCMT) (Buckle et al. 2009), a single dish radio telescope, and the Atacama Large Millimeter Array (ALMA) (Wootten & Thompson 2009) which is an array of 66 radio telescopes.

Radiative-transfer codes, such as RADEX (van der Tak et al. 2007), can be used to constrain the column density of molecules from measurements of molecular lines, where the column density is the integral of the number density along the line-of-sight (usually expressed in cm^{-2}). However, even with generous assumptions such as spherical geometry and all gas emitting identically, the radiative-transfer inverse problem is degenerate. This was, for example, found to be the case in Tunnard & Greve (2016) where popular radiative-transfer codes were found to struggle to recover any parameter better than to within half a dex.

1.2.3 Astrochemical Models

To connect molecular abundances with the physical conditions of the ISM they live in requires the usage of astrochemical models. These models are based on the numerical

integration of systems of ordinary differential equations (ODEs) describing the time evolution of the abundances of the chemical species populating the ISM. These system of ODEs are constructed from reaction networks describing the dominant reactions occurring in the ISM, with reaction rates obtained from databases of gas-phase reactions such as UMIST or KIDA (McElroy et al. 2013; Wakelam et al. 2012). However, a comprehensive modelling of the ISM also requires incorporating gas-grain and grain-grain processes such as the freeze-out of species on gas grains (Rawlings et al. 1992), the desorption of species from the grains (Roberts et al. 2007) and grain-surface chemistry (Occhiogrosso, A. et al. 2014).

In this thesis we make use of the UCLCHEM code which is a gas-grain and time-dependent chemical model, that is to say a model including reactions involving dust grains and evolving abundances over time. The code was first developed in Viti & Williams (1999), but was then further improved in Viti et al. (2004) and Holdship et al. (2017). Today, UCLCHEM is open-source and can be found at `https://github.com/uclchem/UCLCHEM`.

In UCLCHEM, chemical evolution of the gas is typically divided into two phases. In the first phase (phase I), supposed to approximate the molecular gas formation processes, the gas starts in a diffuse atomic state and evolves following a collapse to a dense state. In the second phase of the model (phase II), the physical conditions (temperature, density, ionization rate...) are modified so as to approximate specific observable environments.

Unfortunately, there are still gaps in our understanding of astrochemical processes occurring in the ISM. Whilst there are large experimental and computational efforts to constrain astrochemical processes (Collings et al. 2004; Fulvio et al. 2017), because of the difficulty of reproducing ISM conditions on Earth, the long timescales on which the chemistry occurs, and the overwhelming number of possible chemical pathways, it is difficult to make proper headway and the uncertainties associated with chemical processes remain high (Linnartz et al. 2015). Whilst improvements to astrochemical models are primarily driven by laboratory experiments, there have also been early efforts to use data-driven techniques to better constrain reaction rates (Holdship et al. 2018). Unfortunately, these uncertainties on the equations governing astrochemical models translate into non-negligible uncertainties on the outputs and make it so that the predictions from such astrochemical models carry significant uncertainties.

1.3 Stellar abundances for galactic archaeology

1.3.1 Stellar Surveys

The last decade has seen the advent of large-scale stellar surveys obtaining spectroscopy, photometry and parallaxes for millions to billions of stars in the Milky Way. These large sample sizes provide an unprecedented view into the inner workings of our galaxy and the means for astronomers to begin studying the formation history of our galaxy.

Perhaps the most noteworthy of these surveys is the Gaia survey (Gaia Collaboration et al. 2018a; Gilmore et al. 2012; Randich et al. 2013), based around a space-based observatory of the same name launched by the European Space Agency in 2013 aiming to chart a three-dimensional map of the Milky Way, which has measured parallaxes as well as photometry for more than a billion stars from which high-precision stellar distances and low-precision stellar parameters have been derived. The quality of data provided by Gaia is a quantitative leap from that previously available.

The high quality photometric and kinematic information provided by Gaia is being augmented by a number of ground-based spectroscopic surveys. The APOGEE, LAMOST and GALAH surveys (Majewski et al. 2017; Cui et al. 2012; De Silva et al. 2015) have obtained spectra for hundred of thousands of stars and future large-scale missions are on the horizon (de Jong et al. 2016; Kollmeier et al. 2017; Bonifacio et al. 2016; Tamura et al. 2016). The spectra delivered by these missions allow for deriving abundances for many species across many nucleosynthesis channels.

1.3.2 Galactic archaeology

The Milky Way - our Galaxy - is not a static object. Stars are constantly being formed in stellar clusters, which are groups of stars originating from a common molecular cloud. Most often, because of phase-mixing, stellar clusters do not remain gravitationally bounded for very long after birth and the progenitors end up scattered across the Milky Way. However, in some rare cases, stellar clusters will remain gravitationally bounded up to present days, in which case they are known as open clusters.

Galactic archaeology is a branch of astronomy which attempts to understand the history of the Milky Way and other galaxies using the present-day information about stars. As the present-day locations of stars only weakly correlate with birth locations and ages, they provide an incomplete picture into the history of the Milky Way and there is value in

using other complimentary sources of information. Stellar abundances, which are largely unchanged during the course of a star's lifetime, are such a source of information. The approach of using such abundances as tools to study the galaxy's history, first proposed in Freeman & Bland-Hawthorn (2002), is known as chemical tagging. Today, two distinct forms of chemical tagging have emerged.

Weak chemical tagging uses broad structures in the space of abundances to better understand the history of the Milky Way. As stellar abundances reflect the ISM elemental composition at a star's location and time of birth, any structure in the abundance space typically has a historical or kinematic origin. For example, major merger events of the Milky Way lead to chemically distinct populations of stars (Bonaca et al. 2020). The bimodality visible in the [α/Fe] against [Fe/H] plane for disk stars coincides with a kinematic decomposition into a thin and thick disk (Haywood, Misha et al. 2013) with stars in the thick disk typically being older and on more eccentric orbits.

Weak chemical tagging is a well-established research approach and the astronomical literature contains many studies making use of chemical tagging. For example the authors in Bovy et al. (2012), through studying spatial positions of mono-abundance stellar populations assumed to trace same-age stars, were able to provide evidence for an inside-out formation of the Milky Way. In Hayden et al. (2015), the radial dependency of the probability distribution of stellar abundances was found to exhibit patterns suggestive of radial migration.

Strong chemical tagging goes one step further than weak chemical tagging. Instead of searching for broad structures in the space of chemical abundances, strong chemical tagging approaches seek - using abundances - to resolve the individual stellar clusters and identify those stars which are stellar siblings (ie those stars born from a common molecular cloud). Strong chemical tagging is an extremely ambitious endeavour and at time of writing it is still a matter of debate whether it will ever be theoretically achievable for "typical clusters"; although some limited success has already been achieved. For example. in Hogg et al. (2016), the author's were able to recover known globular clusters using chemical abundances. In Price-Jones et al. (2020), candidate open clusters were identified from abundances and a subsequent evaluation showed their kinematic patterns to be inline with them being stellar siblings. However, in both of these studies, the abundances of the stars identified are to some degree anomalous from the background population of stars making chemical identification easier.

The success of strong chemical tagging depends on a number of assumptions being satisfied. These are that: i) stellar siblings are born with nearly identical chemical compositions ii) the present-day surface chemical composition of a star - as obtained by spectroscopic measurements - reflect its chemical composition at birth iii) there exists sufficient diversity in chemical composition from cluster-to-cluster for clusters to be distinguishable. In the next few paragraphs we examine the validity of these assumptions.

There is experimental evidence in favour of the veracity of the first assumption - that stars are born with near identical compositions - and the second assumption - that birth abundances can be probed from present-day measurements of surface abundances. Observational studies of open clusters, binary stars and to a lesser conclusive degree globular-clusters, have found those stars presumed to have been born together to exhibit a high-degree of chemical homogeneity although small scatter may exist at the 0.01 to 0.02 dex level and up to <0.05 dex level (Bovy 2016a; Poovelil et al. 2020; Hawkins et al. 2020). This chemical homogeneity is, for most elements, in line with what would be expected if the stars were chemically homogeneous. Although some caveats do exist. A major caveat is that the precision with which abundances are estimated today is insufficient for strong chemical tagging, and so it is unclear whether stellar siblings will still appear chemically homogeneous at the higher precisions required for chemical tagging. Also it is now well understood that stellar evolutionary processes, such as atomic diffusion, mixing and dredge-up (Liu et al. 2019; Casey et al. 2019), can act to change some elements abundances away from their birth abundances. But it is likely that these stellar processes can be accounted for, as they do not strongly affect all the elements and are tied to the evolutionary state of the star. Chemical tagging may also not work for binary stars and other types of stars for which environmental effects, such as accretion of a companion star, may have changed abundances away from their birth values.

The third assumption - that clusters have sufficiently unique chemical signatures to be separated out - is arguably the most contentious of the assumptions. The ease with which clusters can be discovered from chemical abundances depends on the effective number of dimensions required for describing the variability in abundance amongst stars. This is because of the curse of dimensionality, the phenomena describing how data is more sparsely distributed when it occupies higher-dimensional volumes. For strong chemical tagging to be possible would likely require a high-level of variability in abundances between stars. However, recent studies seem to indicate that many species abundances are linked to

each other and so, in fact the variability in chemical abundance between stars occupy a low-dimensional manifold of the abundance space. For example, it was found in Ness et al. (2019) that [Fe/H] and stellar age could predict most other elemental abundances to within or close to measurement precision. If this remains true at higher precisions it may spell the end for strong chemical tagging as separating clusters would require unfeasibly precise abundance measurements. However, it may be that our abundance precisions are insufficient for uncovering the full variability of stellar abundances. A more recent study presented in Ting & Weinberg (2021) provides some evidence for this. In this study, the authors through training a density-based machine learning model on stellar abundances cautiously find evidence for abundances living on a higher dimensional manifold. If this reveals itself to be true, an increase in the abundance precision may uncover new structure in the space of abundances.

1.3.3 Estimating Chemical Abundances

Although spectroscopic surveys measure stellar spectra, it is the physical and chemical parameters of stars that are of use for chemical tagging and other galactic archaeology endeavours. Automatic pipelines are needed for efficiently and accurately extracting with minimal human supervision these stellar parameters from the large quantities of spectra delivered by surveys.

Most pipelines for extracting stellar parameters (temperature, surface gravity, abundances...) are based on synthetic stellar spectra. In this approach, observed spectra as measured by surveys are compared to grids of synthetic spectra generated by model atmospheres combined with radiative-transfer simulations taking stellar parameters as inputs. For any observed spectrum, the stellar parameters of the best-fitting synthetic spectra in the grid are considered to be the true parameters of the star.

Synthetic-spectra based approaches for deriving abundances have drawbacks. As recovering abundances depends on matching synthetic spectra to observations it is crucial that synthetic spectra are faithful to observations. However, generating faithful synthetic spectra is a difficult task. Simulations used to generate stellar spectra make simplifying assumptions. For example they often assume that the stellar atmospheres is one-dimensional, in hydrostatic equilibrium and in local thermodynamic equilibrium. Furthermore, stellar line lists, as required for measuring the opacity at different wavelengths, are incomplete and contain poorly constrained parameters. In addition, even beyond these issues, obser-

vations are affected by further systematics such as telluric lines introduced by the earth's atmosphere (e.g. Holtzman et al. 2015) and telescope imperfections/aberrations. As a result of these mismatches between observed and generated spectra, there exists a gap between the abundance precisions obtained by existing spectroscopic pipelines and those theoretically achievable after accounting for the signal-to-noise in observations (Ting et al. 2019).

Given the importance of abundance precisions for downstream applications such as galactic archaeology, there is value in improving the precisions with which abundances and other stellar parameters are estimated. A popular approach for doing so is differential abundance analysis. In differential abundance analysis studies, a sample of stars with similar physical properties (temperature, surface gravity...) are analysed jointly so as for the sample of stars to share similar systematics in abundance, and reduce the overall impact of systematics on the conclusions of the analysis.

In recent year, machine learning methods have also been used for improving abundance precisions. Methods such as those proposed in Ness et al. (2015); Casey et al. (2016); Leung & Bovy (2018); Ting et al. (2019); O'Briain et al. (2021) leverage data-driven interpolators between stellar spectra and labels to reduce the impact of noise and systematics on derived parameters. Recently, methods for finding chemically similar stars directly from stellar spectra without reliance on synthetic spectra have been developed (Bovy 2016b; Price-Jones & Bovy 2017a; Cheng et al. 2021). This category of methods works by removing the effect of non-chemical parameters on stellar spectra, thus isolating the chemical information within the spectra. Such approaches are not without drawbacks. To be effective, these methods require for all non-chemical factors of variation to be removed from stellar spectra which usually requires some form of simplifying assumptions on the interaction between chemical and non-chemical parameters.

1.4 Machine learning

1.4.1 Introduction

Machine learning is a subfield of computer science concerned with the design of algorithms which learn automatically from data. It has experienced a rapid growth in the last decade spurred by a rise in both data and computational power. The field presents a paradigm shift in the way in which algorithms are designed. The conventional approach of creating

algorithms requires handcrafting all the individual steps in the algorithm. This handcrafting is a laborious task which often requires domain knowledge and quickly becomes prohibitively complex. For example, while it is easy for humans to recognize animals from pictures, it is much more difficult to write down an algorithm for doing so. Machine learning algorithms instead of being handcrafted learn automatically how to solve a task from a dataset containing examples. Because the human is removed from the process, machine learning algorithms are capable of learning arbitrarily complex functions provided enough data and computational power is available. For example, modern natural language processing models have reached sizes of 100s of billions of parameters requiring hundreds of GB for storing model parameters (Brown et al. 2020).

Machine learning algorithms can be used to solve many different types of tasks but a distinction is often made between supervised and unsupervised learning. Supervised learning refers to the task of learning a mapping from a set of inputs to a set of outputs from example input-output pairs. For continuous output variables, supervised learning is also known as regression, for categorical output variables as classification. Unsupervised learning is an umbrella term describing all the uses of machine learning not relying on output labels. Unsupervised learning encompasses dimensionality reduction - the task of compressing inputs into a lower dimensional representation - as accomplished by algorithms such as principal component analysis or autoencoders, as well as clustering, anomaly detection and many other tasks.

Early applications of machine learning, in the field of astronomy, can be dated back to the 1980s and early 1990s (Fluke & Jacobs 2020). Early examples of applications of matrix factorization algorithms include the use of Principal Component Analysis for galaxy classification (Whitmore 1984) and the use of Positive Matrix Factorization for studying molecular lines (Juvela et al. 1996a). Use of artificial neural networks can be traced back to the late 1980s and 1990s, where they were, for example, used for telescope scheduling and classification of galaxies (Storrie-Lombardi et al. 1992; Lahav et al. 1996; Johnston 1989).

Today, machine learning has found many uses in astronomy. In applications where it is laborious to label data by hand, such as is the case for the citizen-science Galaxy Zoo project in which galaxies were classified by volunteers (Willett et al. 2013), supervised learning can reduce the need for human labelling (Walmsley et al. 2019). For the task of redshift determination where spectroscopy allows for determining high-precision redshifts

but is only available for a small sample of galaxies, supervised learning enables the determination of redshifts from the comparatively cheaper photometry (Leistedt & Hogg 2017). In model emulation, supervised learning is used to to bypass the running of a computationally expensive simulation by learning the outputs of the simulation directly from the inputs (Schmit & Pritchard 2017). However, whilst the previously discussed applications of machine learning are examples of supervised learning, unsupervised machine learning algorithms have also been used in astronomy. For example, clustering algorithms, such as DBSCAN, have been used to automatically discover candidate open-clusters from the GAIA dataset (Castro-Ginard, A. et al. 2020) which is too large for manual inspection.

1.4.2 The dangers of overfitting

Machine learning algorithms are capable of learning complicated mappings directly from data with minimal human input. This reliance on data for modelling is one of the core strengths of machine learning but is also a limitation. Machine learning algorithm can only be as good as the data they are trained on. And datasets are never perfect. They always contain only a finite sample of data, often noisy, and so accordingly models can always only give an approximately correct answer, with an accuracy and precision in line with the quantity and quality of data available.

One particularly important aspect of training machine learning model is what is known as overfitting. From only limited sample sizes it is not possible to differentiate between genuine correlations and spurious correlations (correlations due to peculiarities in the dataset which would not appear in a newly sampled dataset). Because of this, when making predictions, machine learning models will not only leverage the genuine correlations in the dataset but also spurious correlations. As such, evaluating an algorithm on the same dataset as used for training will lead to the algorithm's performance being overstated. This overstated performance on the trained dataset is referred to as overfitting.

To obtain unbiased measurement of a model's performance it is thus important to not evaluate the model on the same datasets as was used for its training. A common practice is then to subdivide the available data into three subsets. A training split used for model training, a validation split used for model evaluation and model comparison, and finally a test split used only once at the end of model selection to provide an unbiased estimate of model performance. As it is also possible to overfit to the hyperparameters - the parameters which are kept fixed during training but changed between training runs -

any quoted model performance must be evaluated on a test set separate from the validation set. When data is scarce, a common practice known as cross-validation involves replacing the validation dataset with random subsets of the training dataset. In this approach, model performance is estimated by averaging the performance from many separate runs, each trained on a unique subset of the training dataset and evaluated on the remainder training data.

More complex models will typically exhibit stronger overfitting. John Von Neuman was famously quoted saying "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk". Because complex models are more data-hungry, for small data sizes, simpler models, such as linear regression, often perform better than more complex models. This is related to something known as the bias-variance trade-off of machine learning models (Bishop 2006). Bayesian approaches, discussed in Section 1.4.5, offer one avenue for combating overfitting in machine learning models. They replace pointwise estimates of model parameters with distributions over model parameters.

1.4.3 Feedforward neural networks

Feedforward neural networks are a type of machine learning model composed of sequential layers of learnable parameters parameterizing a mapping from an input representation to an output representation. The building blocks of feedforward neural networks are nodes, called neurons, organized into layers. Every neuron in a layer is connected by edges to every neuron in the succeeding and preceding layer. The layers of neurons serve to store values used in the calculation. The first layer contains the inputs fed to the neural network, the last layer the model predictions and all other layers act as place-holders for intermediate values used in the calculations. In a feedforward network, the outputs associated with a given input are found by successively calculating the values of each layer's neurons, starting from the input layer up to the last layer. For neuron j, the formula for calculating its value a_j is $a_j = \Phi\left(\sum_i w_{ij} a_i + b_j\right)$ where $\sum_i$ refers to a sum over the outputs a_i of all neurons of the previous layer, b_j is a parameter associated with every individual neuron usually referred to as a 'bias', and w_{ij} is a parameter associated with every individual edge and is referred to as the 'weight'. The function $\Phi(x)$, called the activation function, is a (usually simple) non-linear function. It's presence ensures that the neural network is capable of learning non-linear functions.

Feedforward neural networks are very powerful and flexible function approximators.

In fact, as proven by the universal approximation theorems of neural networks, any "well-behaved" function can be approximated by a neural network, provided the neural network is wide or deep enough (Cybenko 1989; Lu et al. 2017).

Training a neural network consists of fine-tuning the neural network parameters (weights and bias) such that it parameterizes a mapping faithful to the relationship between inputs and outputs displayed by a training dataset. Measuring the faithfulness of the neural network is done through the intermediary of a loss function quantifying how well the neural network outputs match with the dataset labels. Training proceeds through finding a set of parameters (locally) minimizing the loss function. A typical example of such a loss function for the task of regression is the mean-squared loss:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{1.7}$$

where Y_i is the label for a datapoint i and $\hat{Y}_i$ is the label predicted by the neural network.

Determining the neural network parameters is done through gradient descent, a procedure in which parameters starting from a random initialization are repeatedly updated by a small amount in the direction locally minimizing the loss function. The direction locally minimizing the loss function is the opposite direction to the gradients of the loss function with respect to parameters. For neural networks, these gradients are calculated using the backpropagation algorithm which is a special case of the chain-rule for partial differentiation tailored to neural networks. As calculating the loss and its gradients over the full dataset is computationally expensive, gradient descent is typically done on small random subsets of the dataset known as batches.

After sufficiently many iterations of the gradient descent algorithm, the neural network parameters will reach a local minima of the loss function. At this point, the loss function (on the training data) will transition from decreasing between iterations to plateauing. When such a local minima is reached, as identified from the plateau in the loss function, the neural network is judged to have converged and training to be completed.

In recent years, designing and training deep neural networks capable of converging quickly and robustly towards low-valued minima of the loss function has become a veritable engineering field in and of itself. The initialization of the neural network weights before training (Glorot & Bengio 2010), preprocessing of inputs, exact details of the gradient

descent procedure(Kingma & Ba 2015), activation functions (Klambauer et al. 2017) and types of layers used (Ioffe & Szegedy 2015) all affect the final network convergence. As a detailed description of all these components and their interplay is far beyond the scope of this introduction and not required for understanding this thesis, we refer readers to further resources (Goodfellow et al. 2016) for more details.

1.4.4 Matrix factorization algorithms

Matrix Factorization algorithms are another class of algorithms that have found uses in machine learning. Matrix factorization algorithms decompose a matrix into a product of matrices. The matrices in the product will customarily be lower dimensional than the matrix being decomposed so as for the decomposition to act as a form of data compression.

As different types of matrix factorization algorithms lead to factorized matrices with different properties, different algorithms will have different purposes. Such matrix factorization algorithms can, amongst other uses, be used for blind source separation - the separation of a signal into its independent components - as well as data exploration and data compression.

Principal Component Analysis

Principal component analysis is arguably the most famous matrix factorization algorithm.

The principal components of a dataset X, of shape $N_D \times N_F$ containing N_D data points and N_F features (i.e. observed quantities associated to each datapoint), are an ordered orthogonal basis of the feature space with special properties. In the principal component basis, basis vectors are ordered by the amount of variance they capture. They have the property that for any k, the hyperplane spanned by the first k-axes of the basis is the k-dimensional hyperplane, which maximally captures the data variance.

Principal Component Analysis (PCA) is an algorithm for dimensionality reduction that consists of performing a change-of-basis to the principal component basis and discarding all but the k-largest principal components. In PCA, the number of principal components k is a hyperparameter controlling the trade-off between the amount of information preserved in the dataset X after compression and the degree of compression. The PCA transformation can be written as a matrix factorization $X = WV$ where W of shape $N_D \times N_k$ denotes the data after transformation to the PCA basis and V of shape $N_k \times N_F$ denotes the PCA basis components.

The principal component basis corresponds to the unit-norm eigenvectors of the covariance matrix of X ordered by eigenvalue magnitude. This can be obtained through diagonalization of the covariance matrix. The principal component basis can also be formulated as the maximum likelihood solution of a probabilistic model which is known as Probabilistic Principal Component Analysis (PPCA) (see Bishop (2006)). This probabilistic formulation is used in this thesis because it can be applied to datasets containing missing values. However, it has several other advantages, such as that it can be used for automatically selecting the dimensionality of the PCA basis, and that it can, in some regimes, be solved more computationally efficiently than traditional PCA (Bishop 2006).

Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) (Lee & Seung 1999) is another matrix-factorization algorithm used in this thesis. In NMF, one seeks to approximate a non-negative matrix X of shape $N_D \times N_F$, as the product of two smaller matrices WV, where W and V are of shape $N_D \times N_C$ and $N_C \times N_F$ and constrained to contain only non-negative entries. The dimensionality of the matrix-factorization bottleneck is controlled by the parameter N_C. N_C is a hyperparameter of NMF and so must be preliminarily set before the running of the algorithm.

Formally, the NMF algorithm proceeds by determining non-negative W and V minimizing a loss function

$$L = \|X - WV\|_F \tag{1.8}$$

where $\|A\|$ is a matrix-norm which is often chosen to be the Frobenius norm $\|A\|^2_{\text{Fro}} = \sum_{i,j} A^2_{ij}$. Minimizing the loss function is a non-convex optimization problem which is typically approached through randomly initializing W and V and then using iterative solvers such as coordinate-descent solvers (Cichocki & Phan 2009).

Matrix factorization is an ill-posed problem. Indeed, if $X = WV$ then it will also be the case that $X = WPP^{-1}V$ for any invertible matrix P (of shape $N_C \times N_C$) and so if W and V are solutions to the matrix factorization problem WP and $P^{-1}V$ will also be solutions, provided they are non-negative matrices. Because of this, there is in general no formal guarantees that the NMF algorithm will be capable of recovering W and V even in the limit of infinite data. However, there exist some special cases where W and V are

perfectly recoverable (Fu et al. 2018).

In NMF, to further constrain the matrix factorization problem towards desirable solutions, it is common to add to the loss function additional regularization terms on W and V. In such cases the loss function minimized by NMF becomes:

$$L = \|X - WV\|_F + \lambda_1 \|W\|_R + \lambda_2 \|V\|_R \tag{1.9}$$

where $\|W\|_R$ and $\|V\|_R$ are regularization terms whose strength is controlled by scaling factors λ_1 and λ_2. Adding regularization terms introduces a trade-off between minimizing the original loss function and the regularization terms which are intended to control the "complexity" of the recovered matrices. Examples of possible regularization are $\|W\|_R = \|W\|_{\mathrm{F}}^2$ (L2 regularization) or $\|W\|_R = \|W\|_1 = \sum_{i,j} \mathrm{abs}\,(W_{ij})$ (L1 regularization). In addition to regularization, the initialization of matrices W and V can impact the convergence of NMF and hence the matrices retrieved by the algorithm.

1.4.5 Bayesian Statistics

Bayesian statistics is a branch of statistics built around the mathematically rigorous handling of uncertainties. In the Bayesian interpretation of statistics, probabilities express a subjective degree of belief about events which is informed by the available knowledge. This interpretation of statistics is in contrast with the Frequentist interpretation of statistics which views probabilities as encoding the frequencies of events in repeatable experiments (Murphy 2013).

Bayes theorem is at the root of Bayesian statistics and describes a principled approach for updating probabilities in light of new data. It can be used to fit the parameters of models describing the underlying mechanisms behind how some observable data are generated. Let θ be some unknown parameters of a generative model of data which we are interested in constraining and let $P(\theta)$ be a distribution, referred to as the prior, describing our current belief in the value of these parameters. Then if we are to observe new data d, Bayes theorem gives us an approach for obtaining $P(\theta \mid d)$, referred to as the posterior, that is to say for updating our beliefs about the value of θ after observation of d.

According to Bayes theoreom:

$$P(\theta \mid d) = \frac{P(d \mid \theta) P(\theta)}{P(d)} \tag{1.10}$$

$P(d \mid \theta)$, when interpreted as a function of θ, is defined as the likelihood; it describes how plausible the obtained data is given a set of parameters. The denominator is referred to as the evidence. For many projects, because it is not a function of θ, the evidence behaves as a normalization constant and can be ignored. However, the evidence can also be used for comparing the suitability of competing models on the basis of how well-supported they are by the data.

In such cases one can write:

$$P(\theta \mid d) \propto P(d \mid \theta)P(\theta) \tag{1.11}$$

Often, for applications of interest to astronomers, the posterior $P(\theta \mid d)$ will not have a closed-form solution. In such cases, it can be approximated using algorithms such as Markov Chain Monte Carlo (MCMC) (Foreman-Mackey et al. 2013) or Nested Sampling (Skilling 2006).

1.5 This Thesis

This thesis is concerned with the development of machine learning techniques for the analysis of spectral lines. Two separate research questions are investigated in this thesis.

In a first part, we try and tackle the many degeneracies affecting the scientific interpretation of molecular line measurements in the ISM. Because of degeneracies in the radiative-transfer problem, observations of molecular lines can often be consistent with many different physical interpretations (Viti et al. 2014). To address such issues, in Chapter 2, we present an emulator for the UCLCHEM astronomical code and illustrate how it can be used, in conjunction with Bayesian techniques, for constraining solutions to the radiative-transfer problem. In Chapter 3, we study another degeneracy in radiative transfer modelling: the possibility of multiple blended gas phases contributing to measured line intensities. For this chapter, we investigate through experiments on synthetic data, the suitability of NMF - a matrix factorization algorithm - for recovering gas components in blended molecular line-intensity maps.

The second topic of this thesis addresses the need, by modern galactic archaeology studies, for high-precision chemical abundance measurements. The work presented in these chapters introduces new approach for analysing spectra which address some of the bottlenecks found in existing approaches. Chapter 4 presents a method for removing the

effect of physical parameters on stellar spectra so as to isolate the chemical information. Chapter 5 expands upon the ideas presented in Chapter 4 and presents a method to directly learn the chemical similarity between stars using the information available in open-clusters, using a custom machine learning approach.

Finally, Chapter 6 presents a summary of the overall thesis and discusses future possible investigations.

This page was intentionally left blank

Chapter 2

Incorporating Astrochemistry into Molecular Line Modelling via Emulation

The work presented in this Chapter is based on the paper de Mijolla et al. (2019), in collaboration with Serena Viti, Jonathan Holdship, Ioanna Manolopoulou and Jeremy Yates.

2.1 Introduction

Molecules form in the interstellar medium provided it is dense enough for collisions to bring the chemical reactants together, and cool enough to suppress the complete dissociation of chemical products. Observations of the interstellar medium inside and outside of our galaxy have revealed a rich and diverse chemistry (Shematovich 2012), spanning a wide range of physical environments. As stars form through the collapse of over-densities inside this optically opaque dense interstellar medium (Young & Scoville 1991), the study of molecular gas can provide a window into the star-formation process and shed light onto the life cycle of galaxies.

Typically, the chemistry of the cold and dense interstellar medium is probed through measurements of molecular lines. As already mentioned in Chapter 1, in order to interpret these, radiative transfer models are used to relate the line strengths back to the physical

conditions of the interstellar medium they trace. RADEX (see van der Tak et al. 2007)) is a popular non-local thermodynamic equilibrium (non-LTE) radiative transfer model that models, for a given set of physical conditions, the expected strength of molecular lines. By matching the observed molecular lines to predictions from grids of such models, it is possible to constrain the density, temperature, and column density of the molecular gas (e.g. Viti 2017; Tunnard & Greve 2016; Salak et al. 2018).

However, as molecular line intensities are dependent on a complex interplay between the physics, chemistry, and radiative transfer of the observed region, their interpretation is often ambiguous. As such, estimating chemical parameters via the use of radiative transfer models is a notoriously degenerate problem. Even under a set of idealized assumptions, when one assumes that the gas can be accurately represented using a single component at a unique temperature and density, there exist wide ranges of parameter values capable of fitting a set of observations (see Tunnard & Greve 2016; Kamenetzky et al. 2018). These degeneracies are amplified when studying external galaxies as the telescope beam sizes often encompass a wide variety of different physical environments that must be disentangled.

Since line intensity predictions obtained from radiative transfer models depend on the column densities of the studied region, it is customary to treat column densities as free parameters of the radiative transfer modelling to be constrained alongside the temperature and density. Unfortunately, this makes the radiative transfer modelling highly degenerate, oftentimes leading to many very different models being able to fit the same observations. There have been attempts to address these degeneracies through the use of astrochemical models.

Astrochemical models, such as UCLCHEM presented in Holdship et al. (2017) and described in Section 1.2.3, are computational codes designed for modelling the chemical composition of gas under well-defined physical conditions. This is done through the numerical integration of a set of differential equations constructed from a network of chemical reactions. Recently, various studies have proposed to use the outputs of chemical models (column densities) as inputs to radiative transfer models (Viti et al. 2014; Harada et al. 2019; Viti 2017). Without including chemistry into the forward model, parameter retrieval can give retrievals that are inconsistent with our current knowledge of chemistry as the forward model has no knowledge of which species should be abundant for a given set of conditions. With inclusion of the chemistry, not only are column densities no longer free

parameters, leading to tighter bounds on the retrieved parameters, but it becomes possible to constrain the parameters driving the chemistry, such as for example the metallicity and cosmic-ray ionization rate. However, the widespread integration of astrochemical models into the radiative transfer process is hindered by their long running times and their complexity. The long running times result in even a relatively small grid of astrochemical models requiring a large amount of computational resources.

In this project, we address both of these issues through the creation of a *publicly available* statistical emulator for the UCLCHEM astrochemical model [1]. The statistical emulator is built using a set of neural networks trained to find a multidimensional fit to a training dataset of chemical simulations. The emulator offers a considerable speed-up in modelling time, being able to estimate molecular abundances in milliseconds, much faster than full chemical models which run in minutes. In addition to the speed-up, the UCLCHEM emulator has been simplified, now being dependent on only six variables: density, temperature, metallicity, visual extinction, cosmic-ray rate, and radiation field rate of the region being modelled. With our emulator, the complexity and computational power required to include chemical models into the radiative transfer process has been considerably reduced. As a by-product, we have also created a RADEX emulator for a select few key species for an even faster inference, although the accuracy of this second emulator is lower.

Statistical emulators, sometimes also referred to as surrogate models, are a type of machine learning model trained to learn a mapping from inputs to outputs of some computational simulations from a limited number of simulation runs. They have been used across many scientific fields (usually as a tool for bypass running costly simulations by instead running the emulator), from seminal applications in geostatistics (Matheron 1963) to applications in chemistry (Nentwich & Engell 2016) and many other fields (Beck & Guillas 2016).

Although emulators have gained some traction in the cosmology community (e.g. Schmit & Pritchard (2018) and Kwan et al. (2015)), they remain uncommon in astrochemistry. The astrochemical community instead usually favours comparison to tables of precomputed models (e.g. Mondal et al. (2019), Meijerink et al. (2007), Maffucci et al. (2018), Bisbas et al. (2019)). In Grassi et al. (2011), a neural network was trained to replace the chemical network calls in N-body simulations. Our work differs from this pre-

[1] https://github.com/drd13/emulchem

vious work in its scope. In Grassi et al. (2011), a neural network was trained to simulate the outputs from running an astrochemical model over a limited number of timesteps. In contrast, in this paper, we use an emulator to predict from initial conditions of the gas the final abundances after full chemical evolution. By simulating the full chemical astrochemical model rather than a timestep, the approach presented in this chapter, restricts the parameter space of the emulator, which allows for more precise predictions and avoids issues of error accumulation from repeated application of the emulator. Additionally, another difference with the work in (Grassi et al. 2011), is that we use a more complicated astrochemical model.

This thesis chapter is structured as follows. In Section 2.2, we give some details on the physical and chemical models being emulated as well as the emulation procedure. In Section 2.3 and 2.4, we give a technical overview of our emulation procedure and quantify, over our selected parameter range, the ability of our emulator to accurately predict molecular abundances and intensities. In Section 2.5, using a set of toy observations, we demonstrate how the emulator can help to lift some of the degeneracies present in radiative transfer modelling. In Section 2.6, we further apply the emulators to ALMA observations of the nearby prototypical Seyfert 2 galaxy NGC1068 (presented in García-Burillo et al. (2014) and Viti et al. (2014)) and touch upon some of the weaknesses and strengths of our emulator.

2.2 Modelling molecular gas

In this section we give a brief overview of the specifics of the chemical and radiative transfer models used for emulation. We explain how we combined these to create a simple forward model capable of reproducing observations of the interstellar medium. We finish the section by covering how we can use a neural network to create an emulator of astronomical models.

2.2.1 Chemical models

UCLCHEM is a time-dependent gas-grain open-source chemical model described in Holdship et al. (2017) and introduced in Section 1.2.3. In UCLCHEM, chemical evolution of the gas is divided into two phases. In the first phase (phase I), supposed to approximate the molecular gas formation processes, the gas starts in a diffuse atomic state and evolves following a freefall collapse. The aim of this first phase of UCLCHEM is to create a set

of gas and grain abundances that are self consistent with the chemical network and can be used as a starting point for the second phase of UCLCHEM. In the second phase of the model (phase II), the physical conditions are modified so as to approximate specific observable environments. For an extragalactic application, this could involve high cosmic-ray rates as would be expected in an AGN-dominated galaxy, or high UV flux as would be expected in a starburst galaxy.

For our models, during phase I the gas started at a density of 100cm^{-3} and was then compressed via freefall to a final density, which was left as a free parameter. This freefall collapse was isothermal, with a gas temperature of 10K. The radius of the region, which was also used to calculate the visual extinction, was allowed to vary. During this phase, gas phase desorption and freeze-out on the dust grains proceeds.

During phase II, the gas was assumed to have reached its final density and the physical parameters were varied so as to model a range of environments. The models were all run for 10^7 years, long enough for the gas to reach chemical equilibrium. The parameters that were allowed to vary were the following.

- Temperature (T): The temperatures of the phase II models were increased over time up to a value T following the same procedure as in Viti et al. (2004), where the temperature increases with time as a function of the luminosity of an evolving star. Indeed, the temperature dependence on time will differ depending on which objects the chemical model is simulating, but for the purpose of this study we simply adopted the procedure already present in UCLCHEM. The models were then further run at fixed temperatures until a cumulative time of 10^7 years.

- Gas density (n): The phase I models were run following a parametric freefall collapse (Rawlings et al. 1992) until a density n was reached. The density was then kept constant during phase II.

- Metallicity (m_Z): The initial atomic abundances used by the chemical network were constrained to be a fraction of the solar metallicity. We defined metallicity as a multiplicative factor with a metallicity of 1 corresponding to elemental abundances as found in Asplund et al. (2009).

- Cosmic-ray ionization rate (ζ): This parameter represented the cosmic-ray ionization rate used in Phase II. Additionally, as we do not model the X-ray ionization rate,

we use the cosmic-ray ionization rate as its proxy (Xu & Bai 2016). As noted in Viti et al. (2014) this approximation has its limitations in that X-ray heating is more efficient than cosmic-ray heating.

- Ultraviolet-photoionization rate (χ): This parameter represented the UV-photoionization rate used for phase II models. The UV-photoionization is measured in Draine where 1 Draine is equivalent to 1.6×10^{-3} erg/s/cm^2 (Draine 1978; Draine & Bertoldi 1996).
- Visual extinction (A_V): In UCLCHEM the visual extinction of the molecular gas is controlled by the size of the modelled region.

At its core, the UCLCHEM chemical model is centred around a chemical network specified by the user. For this project we used a chemical network based on the UMIST database (McElroy et al. 2013). For each time-step, a set of coupled ordinary differential equations is generated from the chemical network and solved. We refer the reader to the UCLCHEM release paper (Holdship et al. 2017) for a thorough overview of the effect of various parameters on these rate equations.

2.2.2 The radiative transfer model

The non-LTE radiative transfer code RADEX (van der Tak et al. 2007) in conjunction with collision files obtained from the LAMBDA database (Schöier et al. 2005) was used for estimating line intensities. RADEX is a non-LTE radiative transfer model that decouples the non-local radiation field from the local-level population calculation through the escape probability approximation.

For our emulator, all radiative transfer models were run with H_2 as the unique collisional partner, assuming a background temperature of 2.7K and assuming a spherical geometry. As both Krips et al. (2011) and Viti et al. (2014) found that different geometry choices in RADEX gave comparable outputs for the fitting of CO, HCN, and HCO^+ in a nearby galaxy, we restrict ourselves to using a spherical geometry.

2.2.3 The forward model

Using molecular line intensities to constrain the physical conditions of the interstellar medium constitutes an inverse problem. In practice, such an inverse problem can be tackled by comparing synthetic line-intensity predictions obtained using a forward model

with the measured molecular lines. In this thesis chapter we contrast two distinct forward-modelling approaches:

- The forward model can encompass solely the radiative transfer physics (from now on referred to as chemistry-independent). This is the more established methodology for analysing molecular lines (Imanishi et al. (2018), Michiyama et al. (2018)).
- The forward model can alternatively encompass the chemistry of the molecular gas in addition to the radiative transfer (from now on referred to as chemistry-dependent). This approach is less common but has been used in Harada et al. (2019), Viti (2017), and Viti et al. (2014).

In the chemistry-independent forward-modelling approach, the temperature, gas density, line-width, column-densities, and beam-filling factor are the only parameters allowed to vary. The first four parameters correspond to the input parameters used in the RADEX modelling. The beam-filling factor is treated as a multiplicative scaling factor.

In the chemistry-dependent forward model, the parameters allowed to vary are the inputs to the chemical model, the line widths, and the beam-filling factor. In this case the synthetic observations are created by converting the output abundances from the chemical model into column densities and using these, as well as the final chemical model temperature and density, as inputs to the radiative transfer model. The predicted intensities from the radiative transfer model are then transformed into mock observations through multiplication by a beam-filling factor. In order to convert abundances into column densities, the fractional abundances are multiplied by the column densities of hydrogen as measured at 1 mag and by the visual extinction. We used $N(\mathrm{H_2}) = 1.6 \times 10^{21}\mathrm{cm}^{-3}$ for the column density of hydrogen at 1 mag. This conversion procedure is known as the "on-the-spot" approximation (e.g. Dyson & Williams (1997)).

These models can easily be extended to include more than one phase of gas. For example, for a two-phase gas, one can run two single-phase models and add up their intensities (after rescaling by the beam-filling factor). A summary of the newly introduced parameters and their notation is as follows:

- The line width (Δv): This is the line width used as an input in the RADEX radiative transfer calculations. For simplicity, we assume in our forward model that all molecular lines share a common line width.

- The filling factor (f): For each phase, the output intensities from RADEX are rescaled by a beam-filling factor representing the fraction of the beam occupied by the emission.

2.2.4 Artificial neural networks

Artificial neural networks (ANNs) (Rosenblatt 1958; Linnainmaa 1976), first introduced in Section 1.4.3 of this thesis, are a class of algorithms used for learning mappings between an input space and an output space (Goodfellow et al. 2016), and are trained by tuning a set of parameters to match a training dataset composed of input–output pairs. The building blocks of ANNs are nodes, called neurons, connected together by edges. In an ANN, information is passed between nodes through these edges and combined through non-linear functions to obtain a mapping from the input to the output space.

In feed-forward neural networks (Figure 2.1), as used in this project, the neurons are organized into sequential layers. The neurons from each layer are connected by edges to every neuron of the succeeding and preceding layer. Neurons are place-holders in which numbers are stored with the first layer containing the inputs fed to the neural network and the last layer containing the associated predictions from the neural network. All the neurons in other layers are place-holders for intermediate values used in the calculations. The predicted outputs, given some inputs, are found by successively calculating for each layer the values of the neurons, starting from the input layer up to the output layer. For neuron j, the formula for calculating its value a_j is $a_j = \Phi\left(\sum_i w_{ij} a_i + b_j\right)$ where $\sum_i$ refers to a sum over all the neurons of the previous layer, b_j is a parameter associated to every individual neuron usually referred to as 'bias', and w_{ij} is a parameter associated to every individual edge and is referred to as the 'weight of the edge'. The function $\Phi(x)$, often called the activation function, is a non-linear function whose presence makes it possible to approximate non-linear combinations of the inputs. For this specific project a rectified linear unit (ReLU) activation function (Vinod & Hinton 2010) was used : $f(x) = \max(0, \mathrm{x})$

The neural network parameters (bias and weights) are determined using a set of training data. Typically, the training data contain example inputs and their associated outputs. The neural network then attempts to fine-tune the parameters so as to minimize a user-specified loss function, designed to assess how well the neural network predictions match the example outputs.

In this project we utilize neural networks as emulators. This consists of using the

inputs of a model as the inputs to a neural network and training the neural network to reproduce the outputs of the model. By training a neural network to emulate a model, it becomes possible to bypass the model. This is advantageous if the model is computationally intensive to run, as it allows samples to be obtained at a fraction of the original computational cost.

In the application presented here, because there is a strong overlap in the parameter space explored when modelling different galaxy observations, very similar grids of parameters are run even when interpreting radically different observations. This makes the use of an emulator particularly advantageous as the overhead required in training an emulator will quickly be smaller than the accumulated run-time from running redundant models.

We choose to use a feedforward neural network as the basis of our emulator because feedforward neural networks are a class of machine learning model in which inference is computationally cheap. However, there are many other algorithms which could have alternatively be used for similar effect, such as boosted decision trees (Chen & Guestrin 2016). Whilst we initially considered using a Gaussian Process model as the basis of our emulator in order to benefit from the rigorous handling of uncertainty which comes with using Gaussian Processes (Murphy 2013), in the end we decided against because Gaussian processes scale poorly in computational cost with dimensionality and size of dataset.

2.3 UCLCHEM emulator

In this section we discuss the creation and evaluation of our emulator.

2.3.1 The training dataset

A dataset of $N = 120000$ chemical models was generated. The parameter ranges for the training dataset can found in Table 2.1. Since emulators are only capable of interpolating and not extrapolating, these parameter ranges define the usable range of the emulator.

A Latin hypercube sampling scheme was used for generating the samples in the training dataset. Our latin hypercube sampling scheme used a budget of $N = 120000$ samples and was applied in linear space in order to reduce complexity in the sampling procedure. However, for the ultraviolate-photoionization rate (χ) and cosmic-ray ionization rate (ζ) for which sampling spanned 3 orders of magnitude, it may have been more sample-efficient to sample in log-space. Latin hypercube sampling (McKay et al. 1979) is a statistical

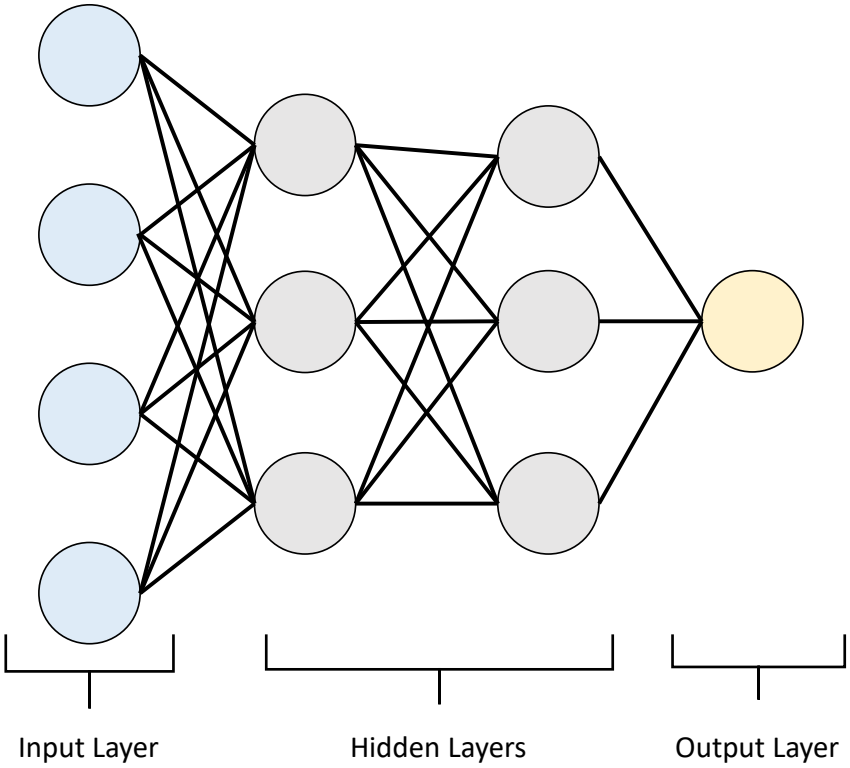

Figure 2.1: *Illustration of a multilayer perceptron neural network.*

Parameter sampling ranges			
Parameter	minimum	maximum	unit
A_V	1	100	mags
n	10^4	10^6	cm^{-3}
ζ	1	10^3	$1.3 \times 10^{-17} s^{-1}$
χ	1	10^3	Draine
T	10	200	K
m_Z	0	2	Z

Table 2.1: *Emulator parameters and their range.*

method for generating near random samples, which are particularly suitable for exploring parameter spaces under a restricted computational budget. It has been used, for example, in Schmit & Pritchard (2018) for the emulation of the epoch of reionization simulations and in Bower et al. (2010) for emulation of semi-analytical galaxy models.

Latin hypercube sampling works by dividing the parameter space into a grid and then placing samples on this grid in such a way that there is only ever one single sample in every dimension (i.e. a single sample per column and row for a two-dimensional parameter space) of the grid. This maximizes the diversity amongst samples but also ensures that no region of parameter space is undersampled (which can happen with fully random sampling).

2.3.2 The algorithm

Feed-forward neural networks were used to predict molecular log-abundances from the UCLCHEM inputs. A separate neural network was trained for each molecular species in our chemical network with each neural network sharing the same five-layer architecture. We chose to model each molecule with a separate smaller neural network rather than using one larger neural network for all molecules, as it decouples the modelling of individual molecules. This can then be more computationally efficient when one needs predictions for only a small number of molecules at once. The final neural network architecture, which was chosen after manual experimentation, was an architecture where the input layer was six neurons wide, with each neuron being assigned to one of the six input parameters. The next three layers were successive hidden layers of width 200, 100, and 50. Finally the last layer represented the predicted output by the neural network. All of the layers used a ReLU activation function (Vinod & Hinton 2010).

For each molecular species, the neural network was trained through the back-propagation algorithm (see Rumelhart et al. (1986)) to minimize the MSE loss over the whole dataset between the chemical simulation outputs and the neural network outputs :

$$\sum_{i=1}^{N}(y_i - \hat{y}(x_i))^2, \tag{2.1}$$

with y_i being the log10-abundance predicted by UCLCHEM for a datapoint x_i, and $\hat{y}(x_i)$ the log10-abundance predicted by the neural network. As abundances cover several orders of magnitude in scale, in order to treat the whole parameter range equally we trained the neural network to minimize the log-abundances. In addition, the input parameters were

scaled to lie within a range from zero to one before being fed to the neural network by using the following transformation:

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)}, \tag{2.2}$$

where X is an input parameter before rescaling. This scaling of input features serves to standardize input features to a common scale, as required by neural networks for good performance.

Training of the neural networks was done in Python using the pytorch framework (Paszke et al. 2017). The neural network parameters were optimized using the stochastic gradient descent optimizer Adam (Kingma & Ba 2015) with an initial learning rate of 0.001 and training occurring over 20 epochs, with batches of 500 in size (see Section 1.4.3 for more detail on neural network training procedure). Hyperparameter selection for choosing this architecture was manually carried-out using hold-out sets of this training dataset.

2.3.3 Error analysis

We quantified the approximation error in our emulation procedure by plotting the distribution of differences between the emulated log-abundances and the UCLCHEM log-abundances for a test dataset, constructed by running 10,000 UCLCHEM models with randomly sampled input parameters (Fig. 2.2). Reassuringly, the emulation error is small for all of our plotted species. For example for CO, 95% of the models agreeing with the emulator predictions to within a multiplicative factor of 1.13 (a log10-abundance difference of 0.05). Whilst some other plotted species like CS had larger uncertainties, all plotted species had 95% of the models agreeing with the emulator predictions to within a multiplicative factor of 1.30. In fact, the worst emulator prediction in this test set on these plotted species was still within a factor of ten of the correct abundances (a log10-abundance difference of 1). Such levels of errors in the emulator are more than acceptable when considering that, for most species, abundances cover a wide range of scales and that astrochemical models have large uncertainties Viti et al. (2014); Holdship et al. (2017). It is also worth mentioning that the figure shows a small but noticeable systematic offset in the CS predictions. This suggests that there was still room for improving the performance of our models. Perhaps this offset could have been removed by training the models for longer.

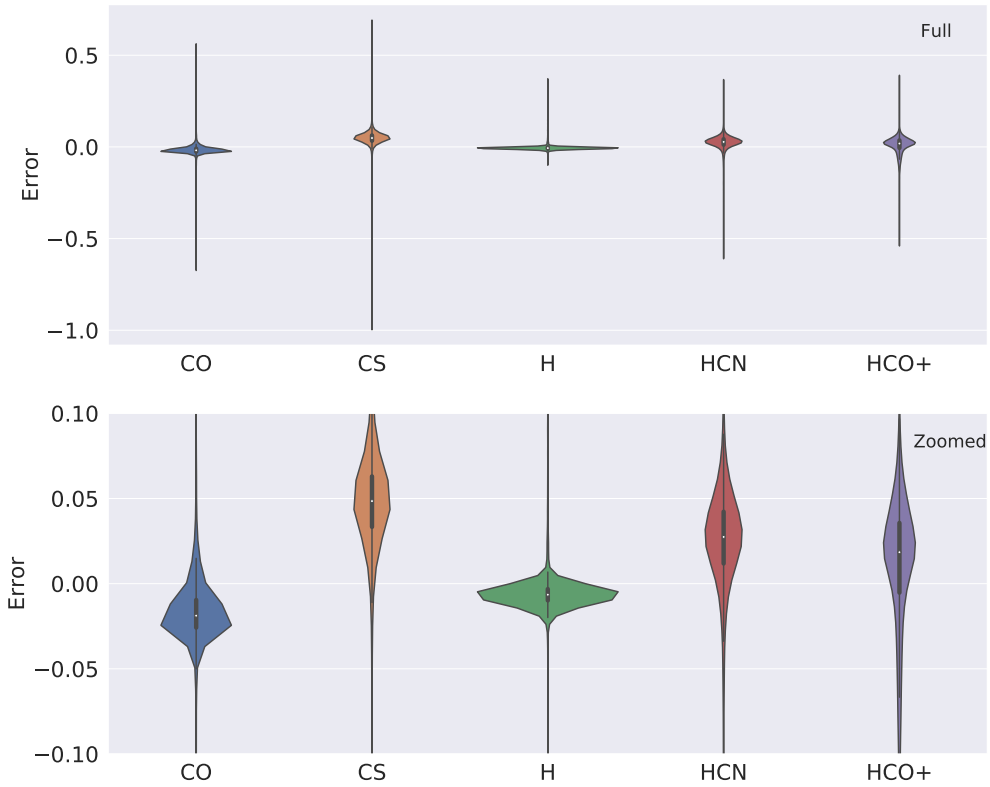

Figure 2.2: *Violin plot of the distribution of the difference between the log10 abundance predictions from the astrochemical models and those from the emulator using a kernel density estimate from the 10,000 simulations in the test dataset for CO, CS, H, HCN, and HCO^+. The bottom plot is a zoomed-in version of the top plot. In the bottom plot, the thick black lines represent the interquartile range and the thin black line the 95% confidence interval.*

.

2.3.4 Effect of the dataset size

We investigated the effect the training dataset size had on the predictive power of our emulator. To do this, we trained our neural networks on differently sized subsets of the training dataset and quantified how the training dataset size affected predictions. For each dataset size, we averaged the MSE model error on the remainder of the training set (i.e. the subset of the training dataset which was not used for training) over multiple runs. The outcome is shown in Figure 2.3, where we can see that further increasing the dataset size

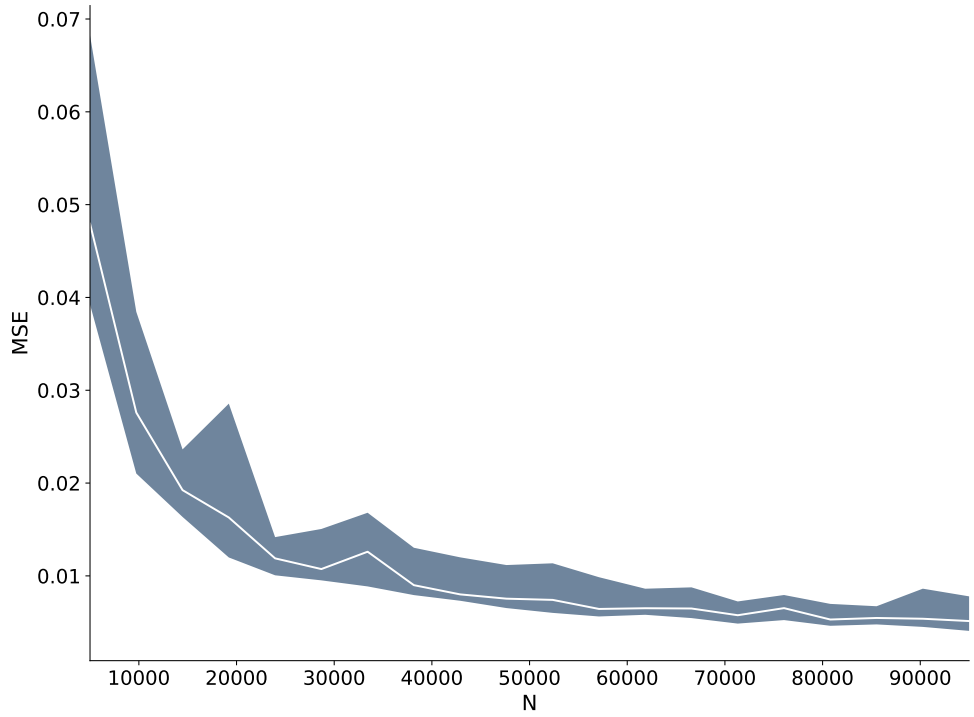

Figure 2.3: *Effect of training set size on emulator prediction. The y-axis shows the mean squared error between the log10 ground truth abundances and neural network prediction evaluated on the remainder of the training dataset which was excluded from training. The x-axis shows the size of the training dataset. The shaded area represents the spread of mean squared error obtained across runs; the* 68.2% *percentiles centered around the mean are shaded.*

beyond the 120,000 samples already used would only offer very marginal improvements to the predictive ability of our emulator. Similar diminishing returns with increasing dataset size have been noted in the literature such as, for example, in Henghes et al. (2021).

2.4 Radiative transfer emulator

In addition to creating an emulator for UCLCHEM, we created a radiative transfer emulator that replicated the results obtained by RADEX for $J < 10$ molecular line transitions. As each molecule required a brand new set of RADEX models to be run, we restricted ourselves to emulating HCO^+, HCN, CO, and CS. Although individual RADEX models are relatively quick to run, exploring a high-dimensional radiative transfer parameter space can require hundreds of thousands of simulations. This makes the use of a RADEX

emulator particularly useful, especially in cases where a slight loss in accuracy is acceptable, such as when used in conjunction with chemical models, which themselves have high associated uncertainties.

2.4.1 Training dataset

The emulator took the temperature, density, line-width, and molecular column densities as inputs. By exploiting the degeneracy between line width and column density for optical depth, we were able to remove the line-width dependency from our training dataset. All the parameter ranges were kept the same as for the UCLCHEM emulator, with the maximal column densities for our RADEX emulator rounded up to be an order of magnitude larger than the maximal column densities in our chemical model dataset. A Latin hypercube sampling scheme was run over the chosen parameter range. We sampled from the log-column density.

The escape-probability formalism that underpins RADEX can break down for some parameter choices, leading to spurious intensities. To mitigate this effect we applied a visually chosen cut-off to the intensities in our training-set; all simulations with intensities higher than the cutoff were excluded from the dataset.

2.4.2 Algorithm

We used the same neural network preprocessing, training, and architecture as that used for the UCLCHEM emulator. We trained the neural network using an L1 loss:

$$\sum_{i=1}^{n} |y_i - \hat{y}(x_i)| \,, \tag{2.3}$$

with y_i referring to the intensity predicted by RADEX, x_i the corresponding RADEX input parameters, and $\hat{y}$ the neural network prediction on a dataset of approximately 100,000 RADEX outputs. Because our method for removing spurious intensities was not perfect, there remained some spurious models. This meant that an L1 loss, which put less emphasis on fitting every single data point perfectly, was more suitable then an L2 loss. To prevent the neural network from predicting nonphysical values, we rounded all negative intensity predictions to zero.

To assess our emulators effectiveness we plotted the difference between RADEX and our emulator alongside the distribution of RADEX intensities for comparison, for a dataset

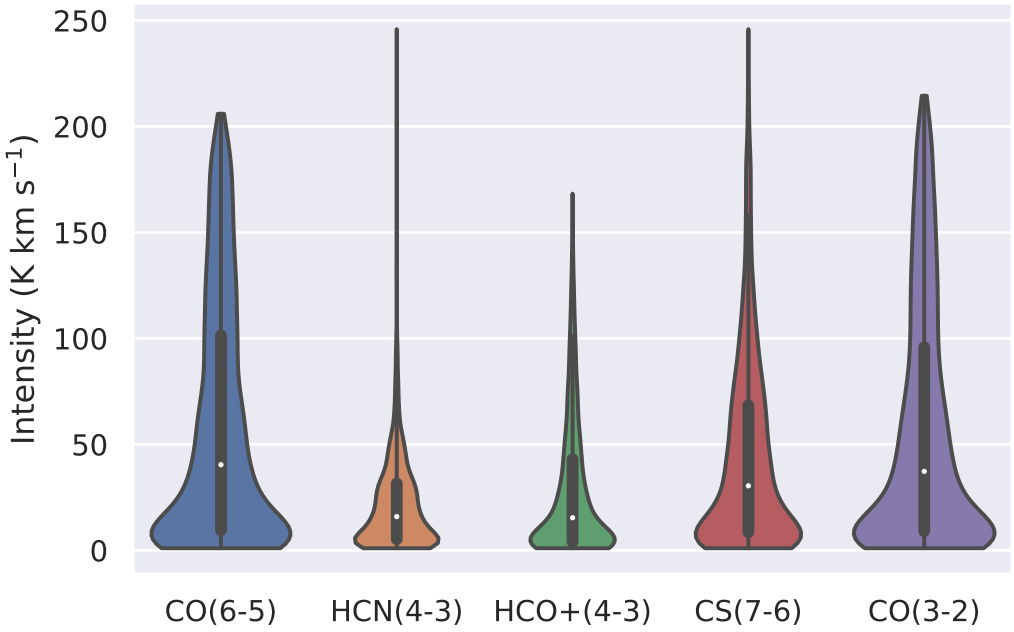

Figure 2.4: *Violin plot of the distribution of RADEX intensities for different molecular lines. The distributions are obtained using a kernel density estimate from the 10,000 simulations in the dataset. The thick black lines represent the interquartile range and the thin black lines the 95% confidence intervals.*

of 10,000 unseen RADEX simulations for CO,CS, HCO^+ and HCN (Figure 2.4 and 2.5). From these figures, we can see that the errors in molecular intensity associated to using the RADEX emulator are comparatively small compared to the errors associated to using the UCLCHEM emulators. In addition, as it is unlikely that all unphysical RADEX simulations have been removed, the tails of our violin plot may be skewed from the unphysical models. From this, we see that the uncertainties associated with the RADEX emulator should be much smaller than those associated to the UCLCHEM emulator. As such, the RADEX emulator should have a minimal impact on our predictions. However, we caution that this situation would change if we were modelling multiple line transitions as, because of the shared column density, the uncertainty from the RADEX emulator would then have a much greater impact.

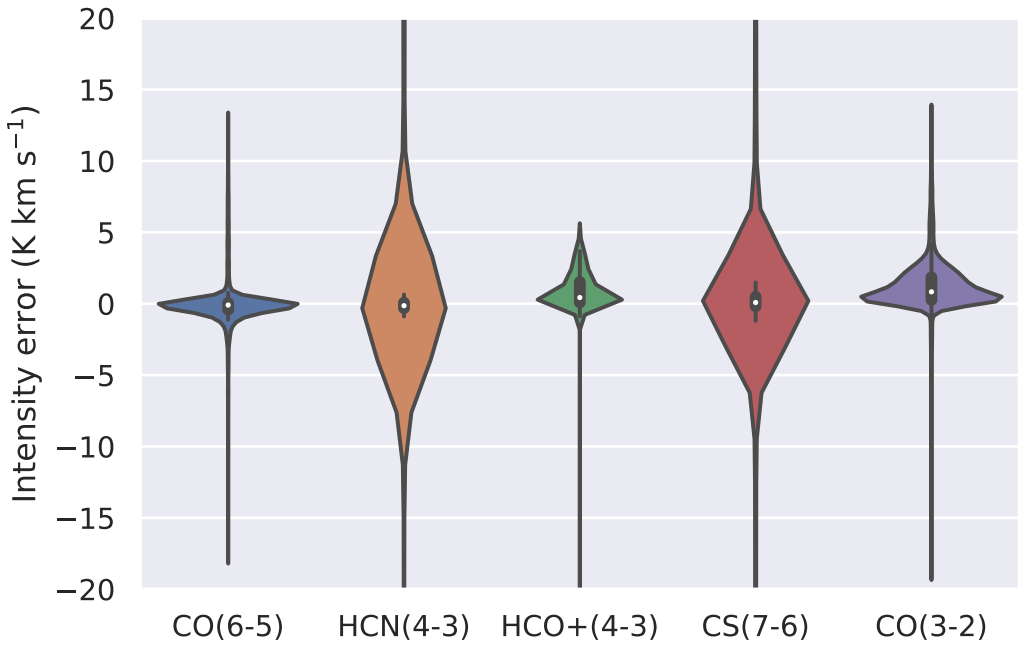

Figure 2.5: *Violin plot of the distribution of the difference between intensity predictions from the emulator and from RADEX for different molecular lines. The distribution is obtained using a kernel density estimate from the 10,000 simulations in the dataset.*

2.5 Bayesian posterior evaluation

In this thesis chapter we advocate the use of emulators as a computationally efficient way of incorporating chemical models into the estimation of model parameters.

To assess the benefits obtained from the inclusion of a chemical model in the forward model, we contrast parameter estimation with and without chemical models. The parameter estimation was performed using Bayesian statistics. The PyMultinest implementation of the Nested Sampling algorithm (see Skilling (2006), Feroz et al. (2009) and Buchner et al. (2014)) was used for sampling from our posterior probability distributions.

In the following sections we begin by giving a brief overview of the Bayesian formalism and Nested Sampling algorithm. We then cover, using increasingly complex models, the advantages and disadvantages of the parameter estimation using our chemical emulator.

We wish to emphasize that in the following sections our objective is to highlight how the emulator may be used for parameter estimation. There is some level of flexibility in how the likelihood may be parametrized, and we do not claim that the parametrization we used is necessarily optimal.

2.5.1 Bayesian formalism

As described in Section 1.4.5, given a model governed by a set of parameter distributions, Bayesian statistics make it possible to mathematically quantify the effect previously unseen data has on further focusing the parameter distributions. In this framework, a probability distribution is associated with each parameter. The prior probability distribution reflects the belief of the user in terms of the parameters before accounting for the new data. This probability distribution will be high in regions of the parameter space likely to coincide with the true parameters and low in other regions. The posterior probability distribution represents the updated probability distribution after accounting for observations; it is mathematically related to the prior distribution through Bayes rule:

$$P(\theta \mid d) = \frac{P(d \mid \theta)P(\theta)}{P(d)}, \tag{2.4}$$

where $P(\theta \mid d)$ is the posterior distribution and $P(\theta)$ is the prior distribution. $P(d \mid \theta)$ (as a function of θ) is defined as the likelihood; it describes how plausible the obtained data is given a set of parameters. The denominator is referred to as the 'evidence'. In this project, because the data are constant, the evidence behaves like a normalization constant and can be ignored.

2.5.2 Application

In this section we describe how we applied the Bayesian statistical formalism towards constraining physical parameters from observations of molecular lines. We consider the case where we have observed N molecular line transitions obtaining observations

$$X = \begin{pmatrix} x_1 \\ x_2 \\ .. \\ x_N \end{pmatrix}, \tag{2.5}$$

with x_i being the intensity of the i^{th} molecular species. Here, we also consider that we wish to estimate $P(\theta \mid X)$ with θ a vector describing the molecular gas parameters of a forward model f. Using Bayes Rule we can express this in terms of the likelihood and

the prior distribution. For our purposes, we consider that our observations correspond to the intensities predicted from our forward model with an independent additive Gaussian noise. This leads to a log-likelihood of the form

$$\ln(L(\theta)) = \mathrm{A} - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - f_i(\theta))^2}{\sigma_i^2}, \tag{2.6}$$

with A being a normalization constant, σ_i the uncertainty associated with the ith species, and $f(\theta)$ the vector of intensities predicted by the forward model f for model parameters θ. The forward model can either be chemistry-independent or chemistry-dependent (see Section 2.2.3).

We set uniform or log-uniform priors on all the parameters with ranges matching the emulator range. The choice of uniform prior meant that all values within the prior bounds were expected to be equally likely, while the choice of log-uniform priors meant that all scales within the bounds were expected to be equally likely. Uniform priors were used for the temperature, visual extinction, line width, filling factor, and metallicity. Log-uniform priors were used for the cosmic-ray ionization rate, UV-photoionization rate, number density, and scaled column densities. Unless otherwise stated, we used parameter priors as found in Table 2.2.

2.5.3 Posterior evaluation

It can be prohibitively expensive to evaluate a high-dimensional posterior distribution because of the rapid scaling with dimensionality of the volume of the parameter space needing exploring. This motivates the use of efficient parameter-space exploration techniques which prioritize resource allocation towards exploring high-probability regions of parameter space. Because of the multimodality found in some of our posterior distributions due to the high-level of degeneracies between parameters of astrochemical models, we used the pymultinest Python module (Buchner et al. 2014) to evaluate our posterior distributions; this is a python wrapper for the multinest package (Feroz et al. 2009) which is itself an implementation of the nested sampling algorithm (Skilling 2006) redwhich is a particularly effective algorithm when the posterior is multimodal.We used the corner module (Foreman-Mackey 2016) to visualize the marginalized posterior parameter distributions.

Nested sampling is an algorithm which can be used to approximate the posterior

Parameter	prior type	range
A_V (mags)	uniform	1-100
n (cm^{-3})	log-uniform	10^4-10^6
ζ ($1.3 \times 10^{-17} s^{-1}$)	log-uniform	10^0-10^3
χ (Draine)	log-uniform	10^0-10^3
T (K)	uniform	10-200
m_Z (Z)	uniform	0.2-2
f (-)	uniform	0-1
Δv (km s^{-1})	uniform	1-100
N(CO)/Δv (cm^{-2}/(km s^{-1}))	log-uniform	10^{13}-10^{19}
N(CS)/Δv (cm^{-2}/(km s^{-1}))	log-uniform	10^{10}-10^{18}
N(HCN)/Δv (cm^{-2}/(km s^{-1}))	log-uniform	10^9-10^{17}
N(HCO^+)/Δv (cm^{-2}/(km s^{-1}))	log-uniform	10^8-10^{15}

Table 2.2: *Default prior distributions on model parameters.*

distibution. It was originally developed, in Feroz et al. (2009), as a tool for estimating the evidence $P(d)$. The algorithm breaks down the task of estimating the posterior distribution into that of sampling from constrained subset of the prior bounded by some hard bound on the likelihood value. The algorithm progresses iteratively, where at each iteration samples are drawn from shrinking subset of the prior distribution. As a complete explanation of the inner workings of the algorithm are beyond the scope of this thesis, we refer the reader to Skilling (2006) for an overview of the algorithm.

2.5.4 One-phase model

Generation

So as to contrast and evaluate the two forward-modelling approaches, we generated a set of synthetic observations using the nonemulated UCLCHEM and RADEX. The parameters used for generating the observations can be found in Table 2.3 and the resultant intensities can be found in Table 2.4. In this case we assumed a known beam-filling factor of $f = 1$.

		Model
T (K)		150
n (cm^{-3})		5×10^5
m_Z (Z)		0.9
A_V	(mags)	40
χ (Draine)		10
ζ ($1.3 \times 10^{-17} s^{-1}$)		100
Δv (km s^{-1})		50

Table 2.3: *Parameters used for creating the single-phase model.*

Beam-Adjusted Intensities		
Transition	Intensity (K km s^{-1})	Emulated Intensity (K km s^{-1})
CO(3-2)	6760.0	6869.1
CO(6-5)	6905.0	6844.5
HCN(4-3)	905.5	958.3
HCO^+(4-3)	16.1	19.9
CS(7-6)	361.05	443.7

Table 2.4: *Intensities (K km s^{-1}) of the single-phase model.*

Posterior estimation

We attempted to retrieve the parameters using the chemistry dependent forward model and the chemistry independent forward model. In both cases, we assumed observational uncertainties of 20% and prior distributions as defined in Table 2.2. Marginalized one- and two-dimensional posterior probability distributions for the temperature, density, and line width, as obtained using the corner module, are shown in Figure 2.6 for the chemistry-independent forward model and in Figure 2.7 for the chemistry-dependent forward model. For the chemistry-independent forward model, the posterior distribution was evaluated using the emulated models, while for the chemistry-dependent forward model, the posterior distribution was evaluated using both the emulated (black) and nonemulated model (blue).

We see that the posterior distributions obtained using the emulated and nonemulated forward models (Figure 2.7) are in excellent agreement with each other. This further supports our findings from previous sections that the emulator can reproduce the RADEX and UCLCHEM model predictions with high fidelity.

From these figures, it is apparent that the chemistry-dependent and the chemistry-independent parameter estimations give very different predictions. While the chemistry-independent parameter estimation struggles to constrain the temperature and density of the molecular gas, the chemistry-dependent estimation is able to return tight and accurate

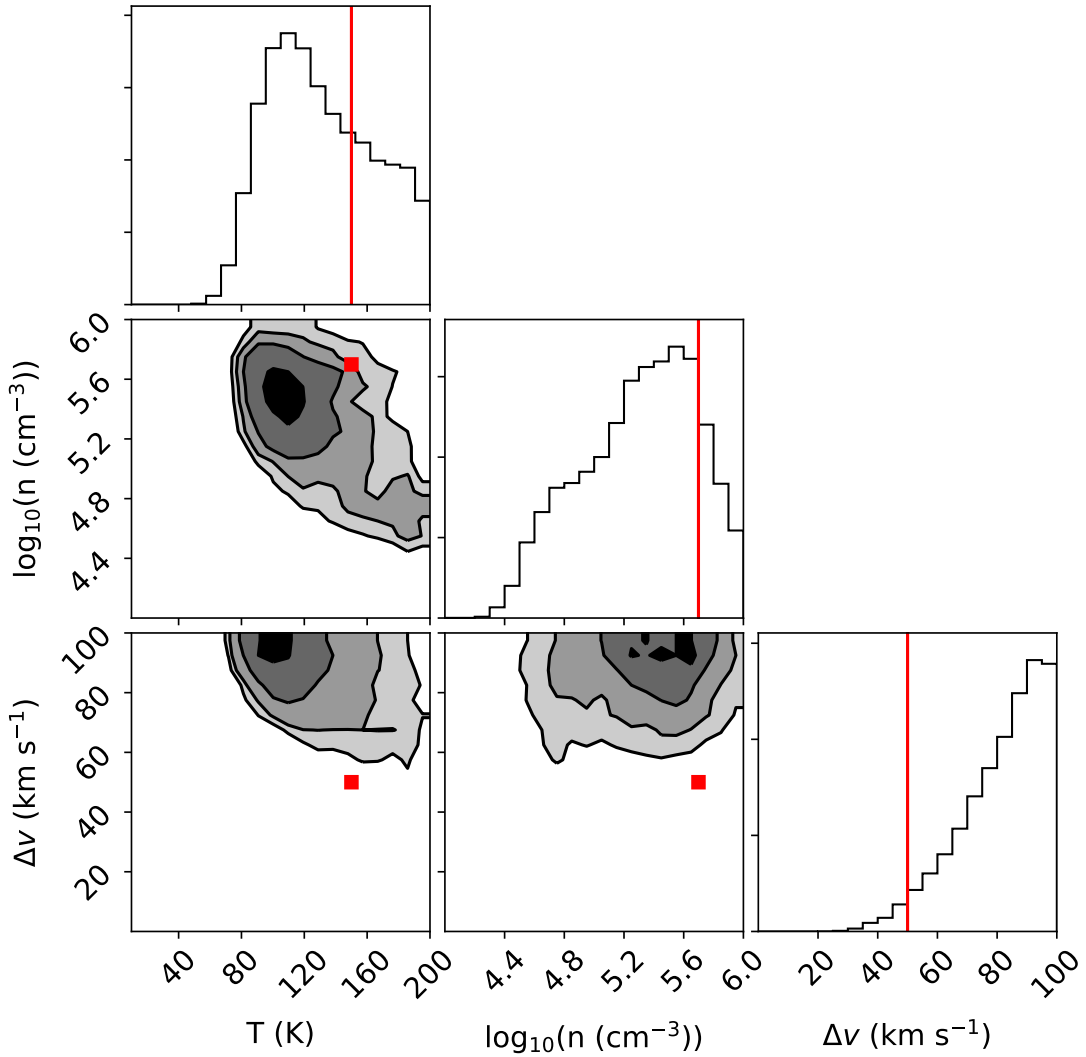

Figure 2.6: *Marginalised posterior distributions obtained when using a single-phase "chemistry-independent" forward model. The true parameters, plotted in red, can be found in Table 2.3.*

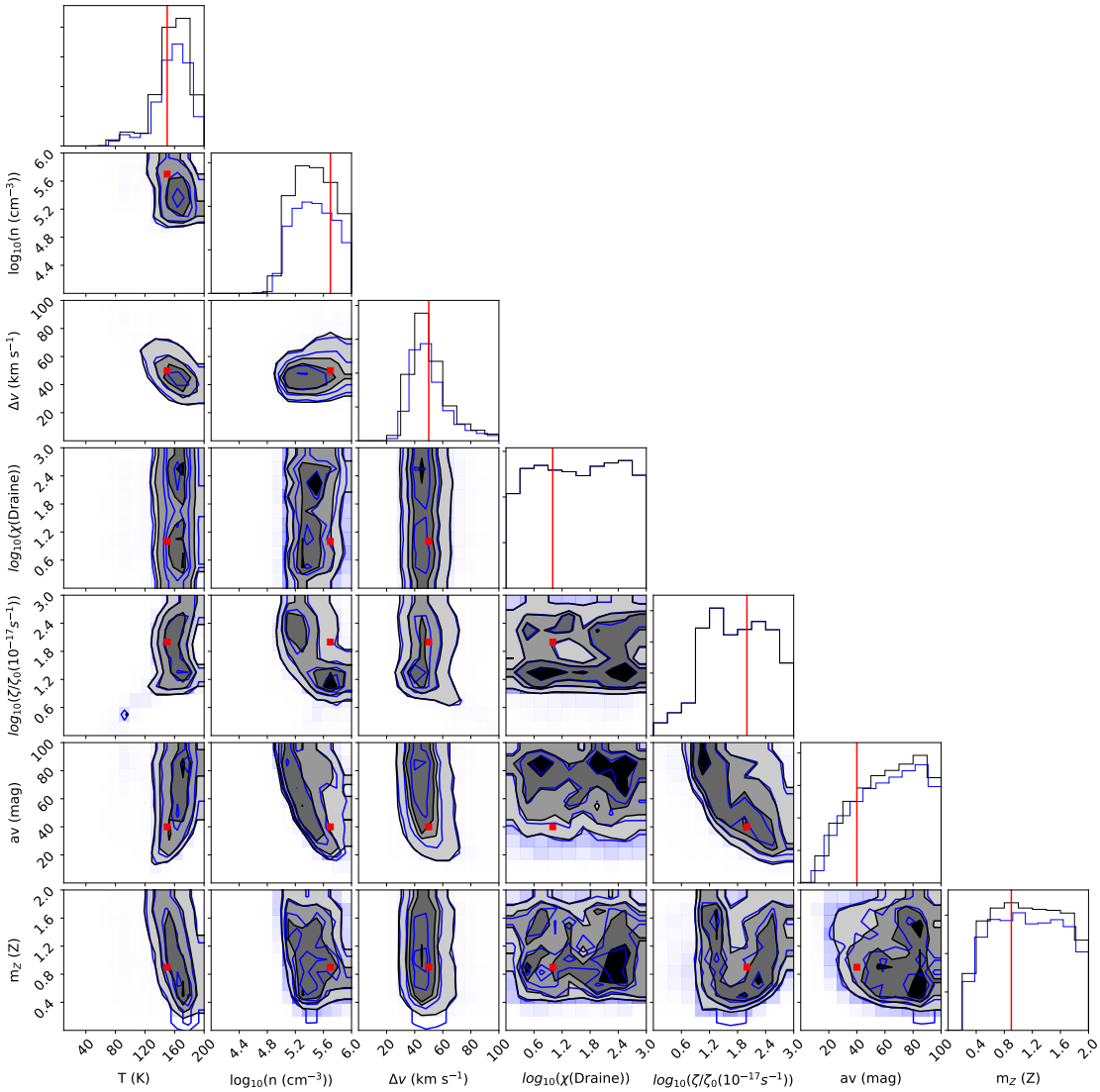

Figure 2.7: *Marginalized posterior distributions obtained when using a single-phase chemistry dependent forward model. The posterior distributions obtained using the emulators are plotted in black while those obtained using the nonemulated models are plotted in blue. The true parameters, plotted in red, can be found in Table 2.3.*

	Model I	Model II
T (K)	50	150
n (cm^{-3})	2×10^4	5×10^5
m_Z (Z)	0.9	0.9
A_V (mags)	3	40
χ (Draine)	10	10
ζ ($1.3 \times 10^{-17} s^{-1}$)	10	100
f (-)	0.7	0.3
Δv (km s^{-1})	50	50

Table 2.5: *Parameters used for creating the two-phase model.*

Beam-Adjusted Intensities		
Transition	Intensity (K km s^{-1})	Emulated Intensity (K km s^{-1})
CO(3-2)	2865.2	2930.0
CO(6-5)	2233.4	2225.7
HCN(4-3)	271.9	287.5
HCO^+(4-3)	4.9	6.9
CS(7-6)	108.5	124.0

Table 2.6: *Intensities (K km s^{-1}) of the two-phase model.*

confidence bounds on these parameters. However, it can also be seen, from the relatively poorly constrained posteriors on some UCLCHEM parameters (m_Z, ζ, χ, A_V), that the chemistry-dependent model struggles to constrain some of the UCLCHEM parameters. We believe this to be due to the high-level of degeneracies between UCLCHEM parameters which makes it difficult to simultaneously constrain all parameters.

2.5.5 Two-phase model

Generation

To further test our parameter retrieval process, we modeled a new molecular gas phase by generating an additional set of intensities, approximating a lower temperature molecular phase, using the nonemulated UCLCHEM and RADEX. We then modeled a beam containing two molecular gas phases by adding the two phases after scaling them by beam-filling factors. The parameters used for generating the two phases and the intensities of the two phases can be found in Tables 2.5 and 2.6. These parameters roughly correspond to a beam filled with one hot and dense phase and one cool and diffuse phase.

Posterior estimation

Much like in the previous section, we next attempted to fit our two-phase observations using chemistry-dependent and chemistry-independent forward models. In both cases, we used the emulated two-phase forward models with prior distributions identical to those used in the single-phase forward models (Table 2.2) on both phases and assumed uncertainties of 20% on the observations (Viti et al. 2014). As we set both prior distributions to cover the full emulated range, the forward model could hypothetically fit the observations with two hot-gas phases or two cold-gas phases. This is in contrast with what has sometimes been done, such as in Tunnard et al. (2015), where the authors constrained the two phases to nonoverlapping parameter ranges, thus artificially forcing the gas to exist in two very distinct phases. In our study, we do not enforce such constraints on the parameter range and thus allow the gas to occupy similar gas phases if the data supports such a scenario. In this analysis we have forced the two phases of gas to share the same metallicity and line width.

The marginalized posterior distributions obtained are plotted in Figs. 2.8 and 2.9. Once again, we see that both approaches give very different constraints on the posterior probability distributions. Although both methods arguably struggle to retrieve the correct temperature and density, the chemistry-dependent forward-model predictions occupy a more narrow range of the parameter space.

In addition, the chemistry-dependent forward model also partially recovers the two-phase bimodality. Even though this is most apparent in the plot of filling factor against density, it is also visible in the temperature-versus-density plot. The above results collectively indicate that the chemistry-dependent estimation is capable, at least partially, of picking out the two distinct phases of gas, even when the chemistry-independent forward model struggles to constrain any of the parameters. On the other hand, we see that our parameter estimation underestimates the line width.

Interestingly, the posterior distributions obtained from two-phase models appear smoother than those of one phase models. This is likely due to a smoothing effect on the posterior from fitting two phases simultaneously.

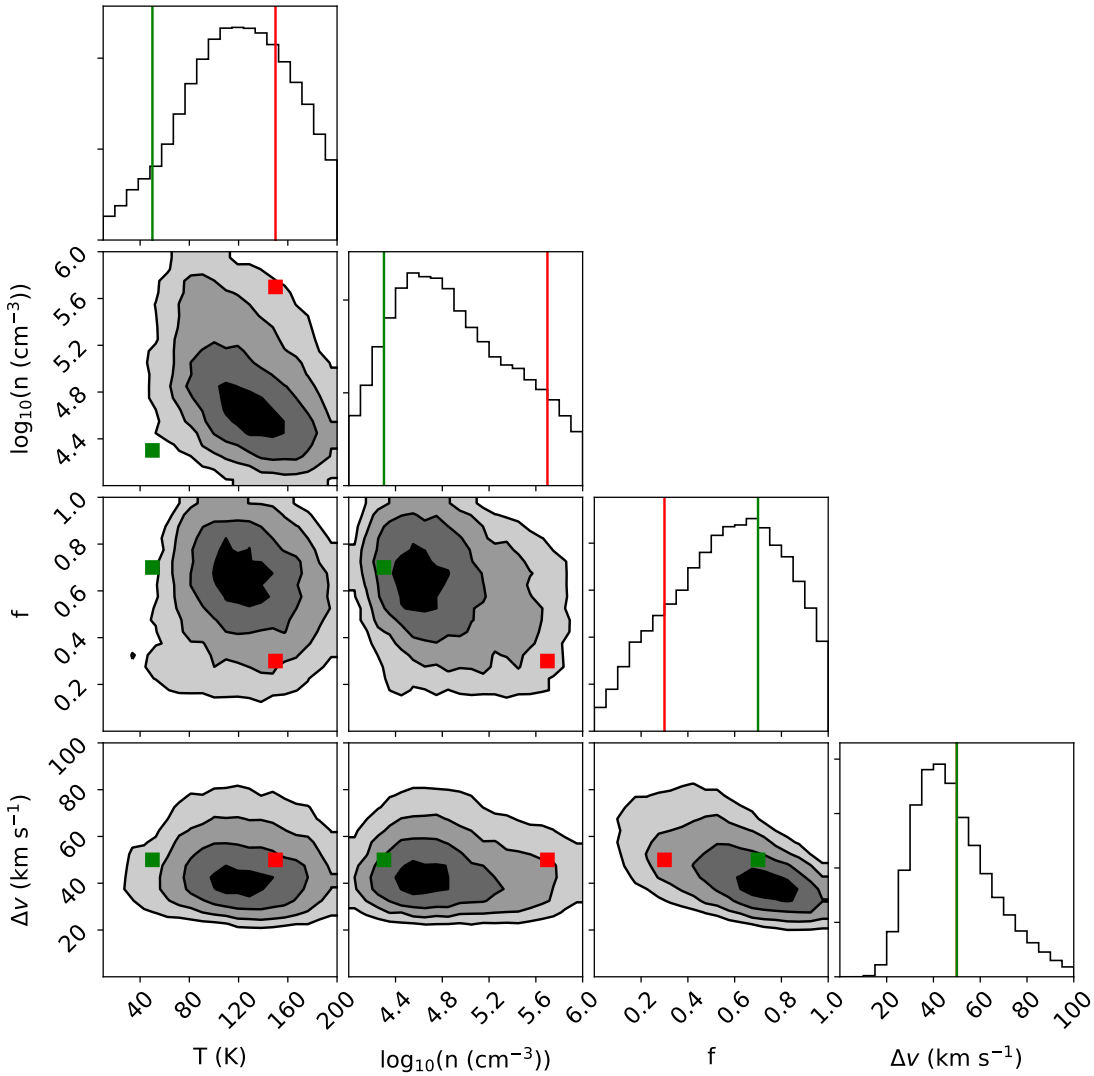

Figure 2.8: *Marginalized posterior distributions obtained when using a two-phase chemistry-independent forward model. The true parameters, plotted in green and red, can be found in Table 2.5.*

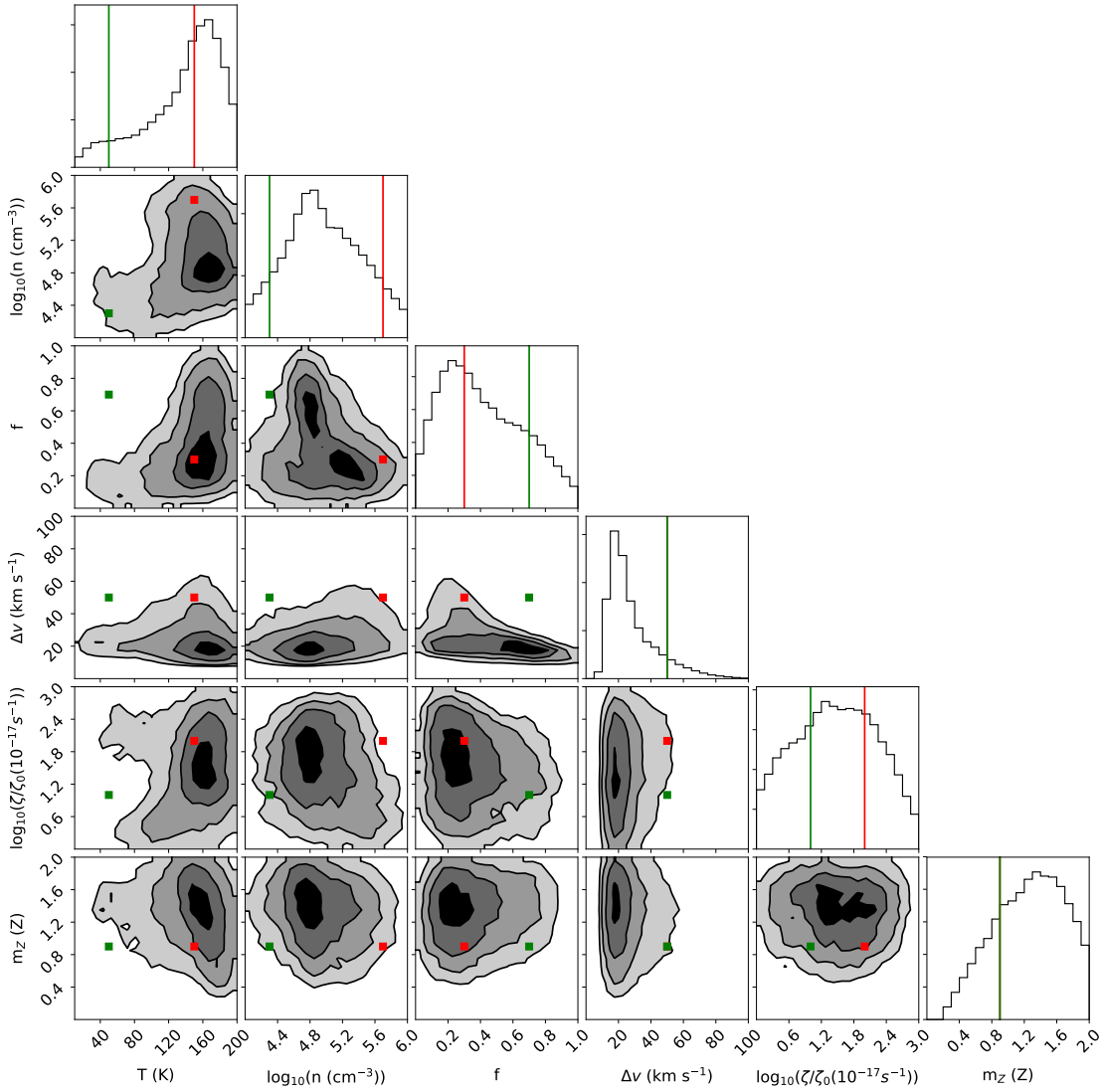

Figure 2.9: *Marginalized posterior distributions obtained when using a two-phase chemistry-dependent forward model. The true parameters, plotted in green and red, can be found in Table 2.5.*

2.6 Application to real line ratios

In the previous section, we used synthetic observations to assess the benefits brought by incorporating chemistry into radiative transfer forward models. However, although synthetic observations are useful in that they offer a controllable and well-understood test bed, there are aspects of working with real regions that cannot be easily understood with synthetic observations. Below is an inexhaustive list of these complications.

- Even though in recent years there has been tremendous progress towards understanding the chemistry in the interstellar medium (Williams & Viti 2013), there are still significant uncertainties associated with the reactions therein. Because of these, the chemistry in our forward model may not accurately match that occurring in real regions.
- Our emulator is only usable or valid for the parameter range under which it was trained (Table 2.1).
- The molecular abundances predicted by our emulator are those reached after the chemical models have been run long enough for an equilibrium to be reached. As such, any transient chemical variation will not be captured by our emulator.
- For molecular observations, and particularly for observations with large beam sizes, the approximation that the gas can be represented using a small number of components is likely to break down.

For the reasons highlighted above, we thought it important to showcase the performance of the emulator on real observations. To do this, we used ALMA observations of NGC1068, a prototypical nearby Seyfert barred galaxy as presented in Viti et al. (2014) and Viti (2017). This galaxy is believed to host a rich chemical diversity and as such it is expected to be very difficult to disentangle the chemistry occurring within. We focused on the spectral lines measured in the East Knot, a region of the molecular ring showing strong emission, and used the degraded resolution measurements as presented in the original paper (Viti et al. 2014); the measured intensities have been retranscribed and can be found in Table 2.8. Although we have every reason to believe that the molecular gas spans a wide range of physical conditions, to avoid using an excessively complex forward model, we fit the region using a single-phase molecular component.

After running an exploratory single-phase chemistry-dependent parameter estimation, we found that our models were consistently unable to reproduce the HCO^+ intensities. To further investigate this issue, we evaluated the HCO^+(4-3)/CO(3-2) line-intensity ratio for all of the UCLCHEM models used by our emulator. We found that all of the chemical models in our training dataset resulted in line-intensity ratios smaller than the ratio observed with ALMA. This suggests that our models, over the parameter range studied, does not produce enough HCO+ to match the observations.

This was in itself not surprising, as chemical models have been known to struggle with producing enough HCO^+ to match observations (e.g. Godard et al. (2010) and Viti et al. (2014)). In Papadopoulos (2007), it was argued that because of the sensitivity of the HCO^+ column density to the ionization degree of the molecular gas, HCO^+ can be an unreliable tracer of hot dense gas. Furthermore, if the HCO^+ is a transient species or if it is created in low-visual-extinction and high-density clumps not covered by our emulation, then it is likely that our models would not capture it.

In light of our inability to reproduce HCO^+ observations, we reran a parameter estimation identical in all but the fact that HCO^+ was excluded from the fitting. The posterior plots obtained without HCO^+ can be found in Figure 2.10. From this we can see that the Bayesian parameter estimation, for fitting a single phase, favors a component with moderately high temperature ($T \sim 120K$) but very high density ($n \sim 10^6 cm^{-3}$).

It is informative to quantify the goodness of fits of some of the models from the posterior distributions. We show, for a small representative sample of well-fit models, the model parameters in Table 2.7 and the corresponding intensities in Table 2.8. From these tables, it becomes clear that the intensities predicted by the emulated and nonemulated forward model are in excellent agreement. These tables also highlight the strong degeneracies which exist in the forward modeling process.

Of note is that the HCN intensity recovered by the Bayesian parameter estimation, although at the correct order of magnitude, was consistently lower than the observed HCN intensity. Almost none of the best-fitting models predicted an HCN intensity greater than the observed intensity. This could be interpreted as the models struggling to create a high enough HCN intensity which could be indicative that the molecular phase, not captured by our models, which is responsible for the high HCO^+ intensity may also be at least partially responsible for a portion of the HCN intensity.

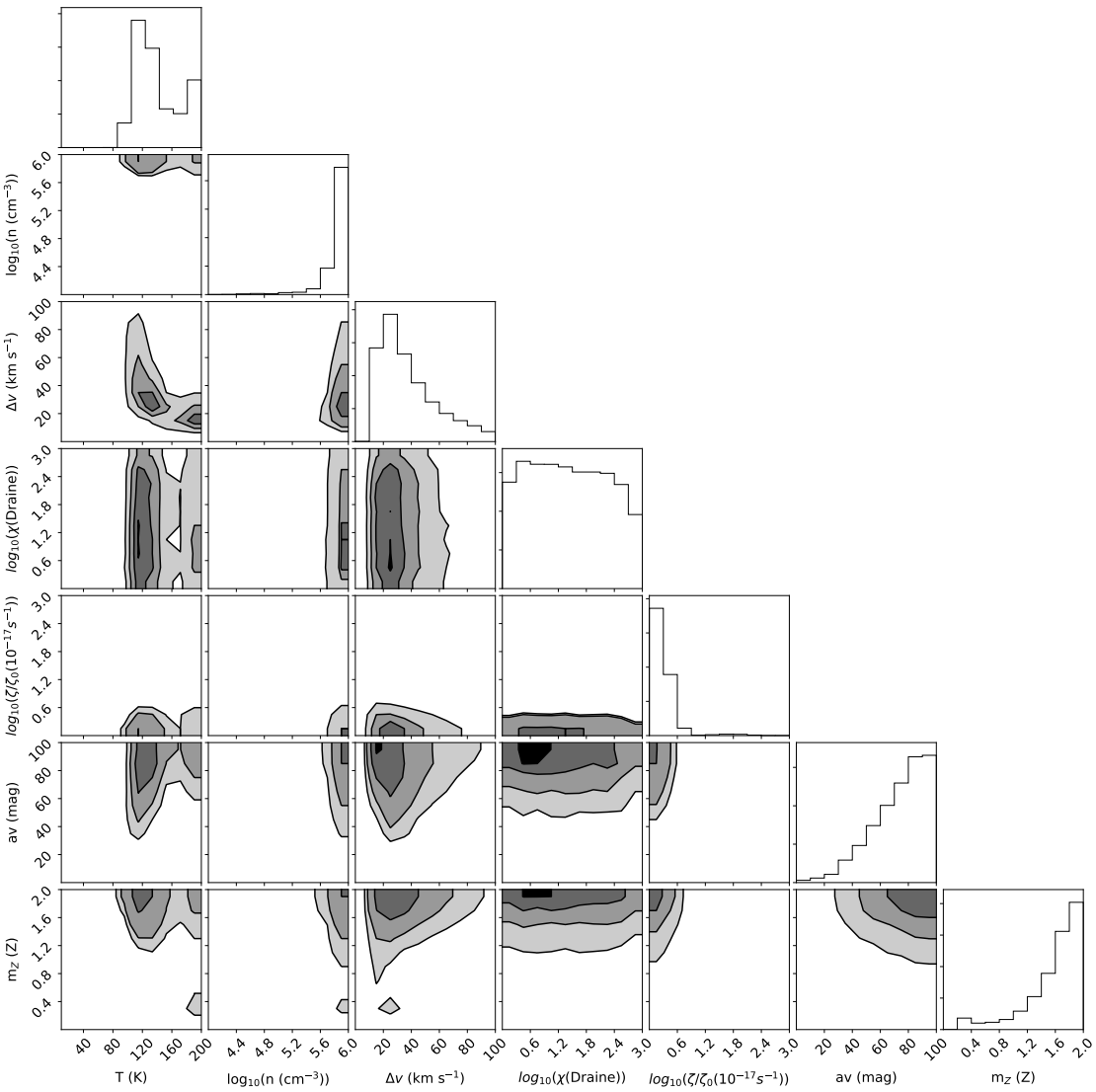

Figure 2.10: *Marginalized posterior distributions obtained when using a single-phase chemistry-dependent forward model on the ALMA observations excluding HCO^+.*

model	χ (Draine)	ζ (1.3×10^{-17})	n (cm^{-3})	A_V (mags)	T (K)	m_Z (Z)	Δv (km s^{-1})	f (-)
(1)	237.90	1.24	966810.74	89.10	199.24	0.26	18.81	0.75
(2)	90.46	1.02	979300.34	86.86	108.46	1.30	78.15	0.31
(3)	31.28	2.95	907548.96	71.29	102.80	1.90	99.09	0.27
(4)	13.94	1.39	857614.52	47.84	96.72	1.85	58.51	0.44

Table 2.7: *Example input model parameters. For the associated intensities see Table 2.8.*

model	CO(3-2)	CO(6-5)	HCN(4-3)	HCO^+(4-3)	CS(7-6)
(1) emul	2397.37	2656.13	532.01	8.38	8.89
(1) direct	2566.15	2705.82	485.44	5.95	9.07
(2) emul	2611.54	2347.36	470.32	1.73	7.86
(2) direct	2577.70	2377.35	554.30	0.94	3.96
(3) emul	2710.83	2426.85	451.47	1.55	10.30
(3) direct	2688.81	2465.41	593.14	1.13	11.54
(4) emul	2491.62	2215.54	454.70	1.47	9.97
(4) direct	2422.29	2209.38	656.22	0.72	5.33
observed	**2346.28**	**2712.70**	**639.47**	**251.18**	**8.26**

Table 2.8: *Intensities (in K km s^{-1}) obtained for the models as defined in Table 2.7. The emul columns correspond to the intensities obtained using the emulated UCLCHEM and emulated RADEX. The direct columns correspond to the intensities obtained using the true UCLCHEM and true RADEX. The last column contains the measured NGC1068 intensities for comparison.*

2.7 Conclusions

In this thesis chapter we propose an alternative method for interpreting the physical conditions of the interstellar medium from the analysis of molecular line intensities. This is typically approached by running many forward models and finding the input parameters to the non-LTE radiative transfer forward model whose predictions match well with observations.

By feeding the outputs of a chemical model into the RADEX radiative transfer model, as was expanded upon in Sect. 2.2.3, it is possible to re-parametrize the radiative transfer forward model. In this formalism, the forward-model dependency on column densities is replaced with a dependency on physical parameters. This offers the potential to lift parameter degeneracies. However, the subsequent forward model is not computationally practical as the required chemical models are computationally expensive to run at scale.

We present and evaluate an emulator, created using neural networks, capable of predicting molecular abundances at a fraction of the run-time of the UCLCHEM astrochemical model, as well as an emulator approximating the outputs of the RADEX radiative transfer codes. We show that by using our emulators it becomes computationally tractable to run the chemistry-dependent forward models accurately at scale.

By applying our emulator to mock observations as well as real observations we show that incorporating chemistry into the parameter retrieval cannot only lead to tighter constraints on the retrieved physical parameters but also constrain parameters introduced by the chemical models such as the cosmic-ray ionization rate and metallicity. We also show that our emulator-based approach is able to distinguish between two distinct phases of molecular gas where a traditional radiative transfer approach fails and that emulator errors are small enough to not affect overall conclusions.

Finally, by applying our formalism to real observations of the galaxy NGC1068, we show that our emulator can effectively be applied to obtain information from real observations. This comes with the caveat that the emulator may struggle to reproduce certain molecular lines such as the HCO^+ molecular lines. However, we argue that this inability to reproduce HCO^+ may be indicative of molecular gas with extreme physical conditions not within our emulation range.

We would like to conclude this thesis chapter by emphasizing that we have had to make choices in defining the likelihood for our experiments, but that these choices may not be optimal. For example, the likelihood could be designed to not only put emphasis on having the line intensities at the correct scale, as was done in our experiments, but also to put emphasis on preserving the relative strength of lines tracing the same species. Finally, the likelihood could also be designed to put stronger constraints on species for which the chemical modeling has been benchmarked and well understood. Furthermore, in most applications it is probably sensible to further constrain or fix some of the forward model parameters, such as the metallicity.

Chapter 3

Non-negative matrix factorization for unmixing molecular components

The work presented in this Chapter constitutes unpublished work led by myself with inputs from Serena Viti and Jonathan Holdship.

3.1 Introduction

As discussed in thesis Chapter 2, understanding the chemical and physical processes regulating the ISM is necessary for a thorough understanding of many astronomical processes, such as star formation (Benedettini et al. 2013) and galaxy quenching (Colombo et al. 2020). As molecular lines are the primary means through which the conditions of the interstellar medium can be probed, their interpretation is of vital importance for accomplishing these goals.

Radiative-transfer modelling is the framework through which radiation emitted by the interstellar medium, typically taking the form of molecular lines, can be connected back to the physical conditions of the gas it is emitted from. Through computational radiative-transfer modelling codes, such as RADEX (van der Tak et al. 2007), the strength of molecular transitions can be used to constrain the chemistry, temperature, density and

other astrophysical conditions of the emitting interstellar medium. This, in turn, then enables a better understanding of the local environment in which the ISM is embedded.

Unfortunately, as demonstrated by experiments in the previous thesis chapter, the radiative-transfer modelling problem needing solving when analysing molecular lines is typically degenerate. Because of the large number of free parameters in radiative-transfer models and their similar influence on intensities, usually multiple set of input parameters are capable of reproducing any given set of molecular line intensities. In Tunnard & Greve (2016), it was found that radiative transfer modeling was typically incapable of recovering parameters to better than half a dex. Although such degeneracies can be partially addressed through using rigorous statistical methodologies (Tunnard et al. 2015; James et al. 2021) or through incorporating astrochemical knowledge (Viti 2017; de Mijolla et al. 2019), the interpretation of molecular lines still remains difficult.

Notably, the analysis of emission arising from blended components is particularly challenging. In sub-mm astronomy, as used to probe molecular line transitions, because of spatial telescope resolution limits, often the flux measured at any spatial location does not originate from a single component but rather from multiple non-resolved components each with unique physical conditions. This is for example the case when observing far-away galaxies as the beam size is far too small to resolve the individual molecular clouds. When this is the case, interpreting the physical conditions of the ISM becomes especially difficult as it is unclear what fraction of the measured emission originates from each component or even how many components are responsible for the observed emission. Moreover, as observers increasingly turn their attention towards the analysis of external galaxies, where multiple blended components are expected, mitigating these degeneracies is likely to become an ever more pressing concern.

To address these degeneracies associated with the the radiative-transfer problem, there is value in approaching the task of disentangling gas components within observations from a more data-driven perspective. In this thesis chapter, we investigate the use of data-driven approaches for separating out the flux emitted by molecular lines into the individual components within observations. More precisely, we investigate the use of NMF (Lee & Seung 1999) which is a matrix-factorization algorithm used to decompose a non-negative matrix into a product of non-negative matrices. In NMF, components are estimated globally from all pixels available. As there will typically be far fewer components than pixels, NMF can then exploit the redundancy across pixels to constrain the possible shapes

of the components.

Our work is not the first attempt at using data-driven approaches to understand the ISM. For example, there has been a long line of research into using Principal Component Analysis to interpret observations of the ISM (Lo et al. 2009; Gratier, Pierre et al. 2017; Neufeld et al. 2007). Perhaps more similar to our work, there have also been efforts into using NMF to interpret interstellar spectra. For example, NMF has been used to analyze interstellar bands (Berné, O. et al. 2007; Foschino, S. et al. 2019). Efforts at using NMF in astronomy date as far back as 1997, where positive matrix factorization, a precursor to NMF, was used to interpret molecular observations (Juvela et al. 1996a).

Specifically, in this study we attempt to answer the question of whether NMF can be used to recover gas components from a set of blended line intensity maps of transitions tracing different physical conditions. Here, unlike previous approaches such as those presented in Boulais, A. et al. (2021); Juvela et al. (1996a), our interest is less about identifying comoving parcels of gas within observations and more about recovering components tracing a common physical environment even if they may not have originated from the same spatial location. Whereas previous approaches have operated on the spectral datacube where frequency information is available and only made use of a small number of line-transitions, here we operate on the derived line-intensity maps for a large number of lines and in doing so ignore the spectral shape of line-transitions. This makes our factorization ignore the radial velocity of spectral lines and only use the molecular composition when constructing components. Our hope is then for NMF components to group together all emission arising from gas under similar physical conditions leading to NMF components capturing the archetypical signatures of physical environments in observations.

3.2 Data model

We begin by detailing a model for the blending of components which relates line intensities and spatial locations of components in a given region to the intensities measured in line-intensity maps.

Let us represent a set of molecular line intensity-maps, observed over a region of $N_x \times N_y$ pixels, by a non-negative matrix X of shape $N_M \times N_p$ whose entries contain the intensities for N_M molecular line transitions measured at the $N_p = N_x \times N_y$ pixels. We assume that multiple components contribute additively to the intensities such that the

intensity at every pixel is the additive sum of the intensities of the separate components. We represent the strength of emission of the components in the various line-transitions by a (non-negative) matrix W_{true} of shape $N_M \times N_D$ whose N_D columns each represent the intensities emitted in the N_M molecular transitions for a component. Similarly, we encode the spatial contributions of each of the N_D components to the N_p pixels into a (non-negative) matrix V_{true} of shape $N_D \times N_p$. Under this model in which component intensities are assumed to be additive, the intensity of a given line at a given pixel will simply be the sum of the emission in that line for all components multiplied by their contribution to that same pixel (a number between 0-1). That is to say the line-intensity maps can be written as $X = W_{true}V_{true} + N$ where N is an additional noise term chosen to be Gaussian.

So far, we have neglected the impact of convolution by a telescope beam factor on observations. Provided that all line intensities are downsampled to a common spatial resolution, convolution can be modeled through a matrix multiplication by a convolution matrix K of shape $N_p \times N_p$. The data-generating process then becomes $X = W_{true}V_{true}K + N$ which can be written using the associativity property of matrix multiplication as $X = W_{true}V_{conv} + N$ where $V_{conv} = V_{true}K$. Hence, including convolution by a beam-filling factor does not change the original model beyond the replacement of V_{true} with V_{conv}.

A number of conditions must be met for our model of line-intensity maps to be accurate. Molecular transitions must not be optically thick so as for intensities to be (near) additive. Additionally, as explained in the previous paragraph, transitions must be preliminarily downsampled to a common spatial resolution for the convolution to be parameterizable by a matrix K.

3.3 Non-negative matrix factorization

In this project, we consider the problem of recovering components W_{true} and their convolved spatial locations V_{conv} from intensity maps X containing blended components, under the assumption that $X = W_{true}V_{conv} + N$. Although many algorithms could be relevant for this task, here we focus on investigating the non-negative matrix factorization algorithm (NMF) (Juvela et al. 1996b). We briefly reintroduce NMF going beyond the explanation provided in Section 1.4.4 of the thesis introduction. Given a matrix X, the NMF algorithm aims to find non-negative matrices W and V such that $X \approx WV$, where W and V are

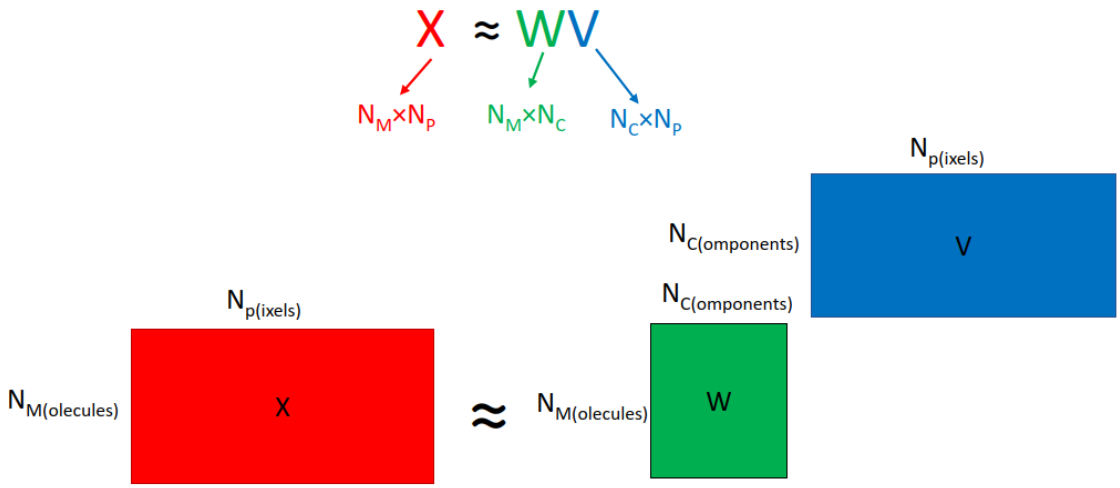

Figure 3.1: *Schematic depiction of NMF algorithm.*

matrices of shape $N_M \times N_C$ and $N_C \times N_p$ and N_C is the number of components used (see Figure 3.1 for schematic depiction). In the context of our project, N_C is analogous to the number of components assumed to exist within observations and is a user-defined parameter for the algorithm. NMF falls within the realm of blind-source separation algorithms which is to say that it is an algorithm making minimal assumptions on the contents of matrices W and V.

Formally, the NMF algorithm proceeds by determining non-negative W and V minimizing a loss function

$$L = \|X - WV\| \tag{3.1}$$

where $\|A\|$ is a matrix-norm which is often chosen to be the Frobenius norm $\|A\|_{\mathrm{F}}^2 = \sum_{i,j} A_{ij}^2$. Minimizing the loss function is a non-convex optimization problem which is typically approached using iterative solvers. For this paper, we use the scikit-learn implementation of NMF making use of a coordinate-descent solver (Cichocki & Phan 2009) with W and V initialized as random matrices.

Matrix factorization is an ill-posed problem. Indeed, if $X = WV$ then it would also be the case that $X = WPP^{-1}V$ for any invertible matrix P (of shape $N_C \times N_C$) and so if W and V are solutions to the matrix factorization problem WP and $P^{-1}V$ will also be solutions, provided they are non-negative matrices. Because of this, there is no formal guarantees that the NMF algorithm applied to line-intensity maps X will converge towards W_{true} and V_{conv}, even in the limit of infinite data.

In NMF, to further constrain the matrix factorization problem towards desirable solutions, it is common to add to the loss function additional regularization terms on W and V. In such cases, the loss function minimized by NMF becomes:

$$L = \|X - WV\|_F + \lambda_1 \|W\|_R + \lambda_2 \|V\|_R \tag{3.2}$$

where $\|W\|_R$ and $\|V\|_R$ are regularization terms whose strength is controlled by scaling factors λ_1 and λ_2. Adding regularization terms introduces a trade-off between minimizing the original loss function and the regularization terms which are intended to control the "complexity" of the recovered matrices. Here we consider regularization on W and V of the form $\|A\|_R = \|A\|_{\mathrm{F}}^2$ (L2 regularization).

In addition to regularization, the initialization of matrices W and V can impact the convergence of NMF and hence the matrices retrieved by the algorithm. In this project, we initialize the algorithm with non-negative random matrices as defined in the scikit-learn NMF documentation.

It is also worth mentioning that the NMF algorithm, like many other matrix factorization algorithms, is most effective when there are significantly fewer components than molecular lines that is to say when $N_C << N_M$. This is because NMF relies on finding a set of components capturing the most information about observations X which requires grouping together correlated molecular lines. In the limit where the number of components matches the number of molecular lines observed ($N_C = N_M$), the NMF loss is trivially minimized by setting W to the identity matrix (i.e. one component reconstruct each line) and V to be equal to X. Such a solution perfectly reconstructs observations but brings no useful information about the correlations amongst molecular lines.

3.4 Synthetic data generation

In this thesis chapter, we investigate how helpful NMF is for interpreting molecular line-intensity maps by testing its practical effectiveness at retrieving components closely matching ground-truth components from synthetic line-intensity maps. We work on synthetic data in order to have knowledge of the ground-truth components. Our aim when constructing these synthetic observations is to approximately reproduce components and observations found within real-world datasets. Our aim is not to build extremely accurate simulations. As such, we have striven for simplicity over complexity in our data-generating

process. We now describe how these maps are constructed.

To construct synthetic line-intensity maps, we first specify the spatial location and emission profile of components as well as the beam-filling factor and noise in the form of matrices W_{true}, V_{true}, K and N. We then assemble line-intensity maps from these using our model of component blending presented in Section 3.2: $X = W_{true}V_{true}K + N$.

The matrix W_{true} contains the emission of each component for the molecular lines considered in X. For our experiments, components are initialized at a user-defined temperature and density. The intensity of molecular lines are then determined using radiative transfer modelling from these user-defined temperature and density and from column densities which are derived from astrochemical models. The UCLCHEM astrochemical code is used to determine the column densities of species (see Section 1.2.3). The UCLCHEM code takes as inputs a gas temperature, density and initial elemental abundances as well as other physical parameters and outputs ISM molecular abundances. In all our experiments, when running UCLCHEM, we run models for 2×10^6 years. When running these models we assume standard Milky values for the cosmic-ray ionization rate and UV-photoionization rate of respectively $1.3 \times 10^{-17}\mathrm{s}^{-1}$ and 1 Draine. We convert abundance to molecular column densities by finding the factor that would convert the maximum component's CO abundance to a CO column density of $10^{19}\mathrm{cm}^{-2}$. We then scale the other components by the same factor so that their column density ratios are equal to their abundance ratios. Since emission scales with column density and we choose the noise level, the absolute value of the column density is completely arbitrary and has no effect on the experiments. On the other hand, the differences in each species' column density between components does affect the results. We therefore use the abundance ratios in the experiments and explore the effect of larger column density differences in Section 3.5.4. Finally, the RADEX radiative transfer code (see Section 1.1) then uses these molecular column densities and the temperature to predict molecular line intensities for lines in W_{true}. These intensities form the entries of W_{true}.

The matrix V_{true} encodes the spatial strength of each component at the p pixels of line-intensity maps X. For each component, we model the spatial contribution at every pixel by a 2d Gaussian whose dimensions correspond to the x and y coordinates of the pixel. That is to say, the contribution at a pixel v_{xy} is $v_{xy} = A \times N(\mu, \Sigma)$ evaluated at the x,y coordinates of the pixel center. In this data-generating process, the mean μ of a component controls its location, the covariance Σ its extent and the amplitude A its

relative strength. For all components, the amplitude A is set such that $v_{xy} = 1$ at the peak location of the component. The entries of V_{true} are then the contribution v_{xy} of the N_p pixels for the N_D components.

The telescope beam-response is approximated by a Gaussian kernel K. After creating noiseless line-intensity maps according to $W_{true}V_{true}K$, in our final step gaussian noise is added to line-intensity maps such that every map has a signal-to-noise-ratio of 100 with noise uniformly added to all pixels.

3.5 Experiments

We now present our experiments on applying NMF to synthetic observations. All experiments are carried out using a common blueprint for the characteristics/properties of the line-intensity map. All line-intensity maps are constructed on a 50×50 dimensional grid (i.e. 50×50 pixels), for the transitions observed in Viti et al. (2014), to which Gaussian noise is added such that maps have a mean per-pixel SNR of 100 and to which a convolution with a Gaussian kernel of width 1/10 of the overall width of maps has been applied. Although this blueprint is only one amongst many possible parametrizations, we found our study's conclusions to be robust to changes to the blueprint.

In the following experiments, before application of NMF, we always apply a series of preprocessing steps to the line intensity maps. These preprocessing steps would likely need to be reproduced when working on real observations. As a first step, in order to satisfy the non-negativity requirement of NMF, all negative-entries in the line-intensity maps are zeroed-out. As a second step, so as for the NMF loss to equally weight all line-intensity maps, we rescale line-intensity maps prior to running NMF through multiplication by a scaling factor such that, after multiplication, all line-intensity maps have an equal average flux of unity. After the NMF algorithm has finished running, this scaling factor is erased from the NMF components through dividing entries in W by the scaling factor associated to their transition.

In addition, because the convergence of the NMF algorithm was found to be sensitive to initial conditions and regularization (because of the under-constrained nature of the factorization), we study the performance of NMF in the aggregate by running ensembles of runs in which hyperparameters (regularization and initialization) are varied.

We then study the convergence of the overall population of NMF runs. To do so, we

have created a goodness-of-fit metric condensing all information about individual runs into a single number quantifying the level of agreement between recovered and true components. This metric is defined as the mean-absolute error (MAE) between the true components emission profiles W_{true} and those retrieved by the NMF run (W). However, since components returned by NMF can only be correct up to a scaling factor and a permutation of components, for our MAE metric calculation, we preliminarily rescale emission profiles to have an average flux (averaged over all transitions) of unity and quote the MAE value for the one permutation amongst all possible permutations of the components in W minimizing the MAE. Under this definition, the line intensities (W) of components retrieved by a run with a MAE of 0.1 will on average differ from the ground-truth intensities (W_{true}) by 10% of the average line-intensity of their closest matching ground-truth component.

3.5.1 Protostellar environment

Here, we present insights derived from applying NMF to the first of two sets of mock line-intensity maps X_{proto}, which models emission from blended protostellar environments. The maps in X_{proto} are generated from three components, each meant to approximate a gas component found in star forming regions. Generated components are i) An extended "low density" component (T=20 K and n=10^4cm^{-3}), ii) a "medium density" component (T=10 K and n=10^5cm^{-3}), and iii) a "dense" warmer component (T=100 K and n=10^7cm^{-3}). These three idealized components may, for example, be representing the surrounding ISM, the envelope, and the more central compact region of a protostellar core.

The synthetic-line intensity maps (X_{proto}) as well as emission profiles and convolved contributions of emitting components are shown in Figure 3.2. From this Figure, we see some moderate spatial overlap between components as well as some components emitting more strongly than others. Both of these factors make it difficult, at least visually, to recover components from the line-intensity maps motivating the use of specialized retrieval algorithms.

We show in Figure 3.3 outputs obtained from running an ensemble of NMF runs on X_{proto}. Runs in the ensemble use three NMF components, the same number as used in the modelling, which means that the algorithm should have enough capacity to fit all components. Figure 3.3 a (on the left) characterizes the goodness of fit of NMF across this ensemble of runs. Each panel of the figure represents a binned density plot of runs in which x-axis coordinates quantify the amplitude of the regularization term, y-axis coordinate the

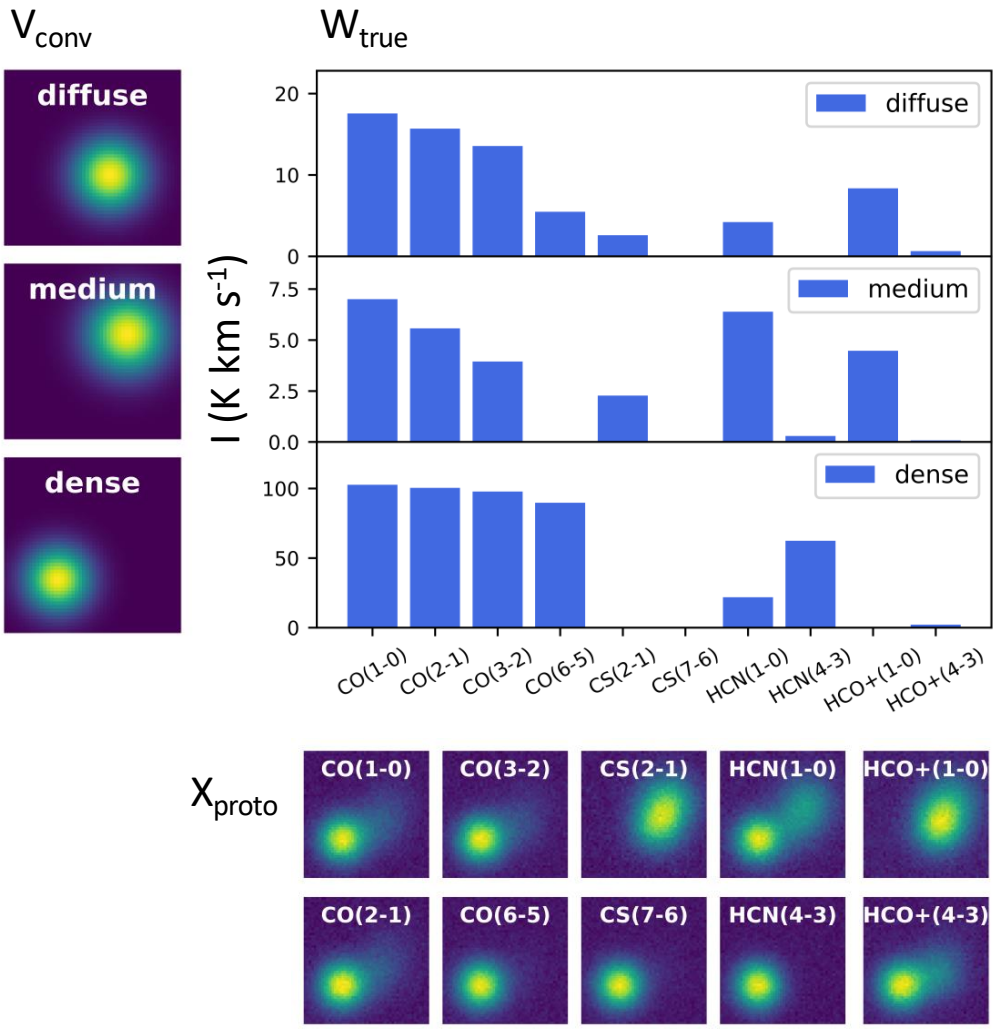

Figure 3.2: *Synthetic line-intensity maps of molecular clouds (X_{proto} : bottom) alongside the emission profile (W_{true} : top-right) and convolved (unitless) spatial contribution (V_{conv} : top-left) of each components contributing to the maps.*

goodness of fit of the retrieved components as defined by the MAE metric introduced in Section 3.5. Each panel considers a different type of regularization: i) only on W (top panel), ii) only on V (central panel), or iii) on both W and V (bottom panel). Figure 3.3 b and c focus-in on two runs within the ensemble and compare the relative line strengths (W) and spatial contributions (V) of these runs to those of the ground-truth components. The locations of these runs within the density plots are shown by markers #1 and #2. These two runs are selected for further study from the ensemble of NMF runs because they represent two extremes (in terms of agreement with W_{true}) amongst all runs.

We find an excellent agreement between the components from run #1, one of the better fitting runs in the ensemble, and the true components. Although there are minor

mismatches between true and recovered components, such as most of the HCN(1-0) originating from the diffuse (top-panel) true component being mistakenly attributed to the middle-panel retrieved component, the recovered components are by and large in excellent agreement with the true gas components. However other runs in the ensemble differ more strongly from the true components. Run #2 is an extreme example of such a run with one of the lowest y-values across the ensemble. This run exhibits a much higher level of disagreement with the true components, with in particular its top-panel component not matching particularly well with any of the true components. The disagreement observed stems from the run learning a different decomposition into components than that used to generate the data. In this decomposition, the top panel component captures emission from both the top and bottom panel of the true components. Such a decomposition, while different from the data-generating decomposition, was found to match the lower MAE runs (e.g. run#1) in terms of its quality of fit to the line-intensity maps. The decomposition found in run#2 is thus an equally valid decomposition that cannot be ruled-out from studying the quality of the match between X_{proto} and the factorized approximation WV.

By comparing the distribution of y-coordinates of runs in the models ensemble to the y-values of runs #1 and #2, we can understand how effective and reliable NMF is at recovering components closely matching with the underlying component. We observe that the vast majority of runs, irrespective of regularization, have y-values comparable to #1 indicating that most runs retrieve components matching well with the underlying components. While poor-match runs comparable to #2 do exist, these runs are relatively rare. This seems to suggest that, at least for the considered dataset, NMF is an effective albeit imperfect algorithm for recovering the underlying components.

We can also study the impact of hyperparameters on NMF results. Regularization appears to have a beneficial effect on the agreement between recovered and ground-truth components with even small amounts of regularization on W being sufficient to steer the factorization away from the higher MAE solutions. Whilst regularization on V does not have as strong of an impact it also improves the match between true and recovered gas components as can be seen from how the best performance is obtained with high-levels of regularization on both W and V. For now we do not discuss how the choice of number of NMF components affects results (but such a discussion can be found in Section 3.5.3). However, it is worth acknowledging that we expect performance to degrade when using a sub-optimal number of components.

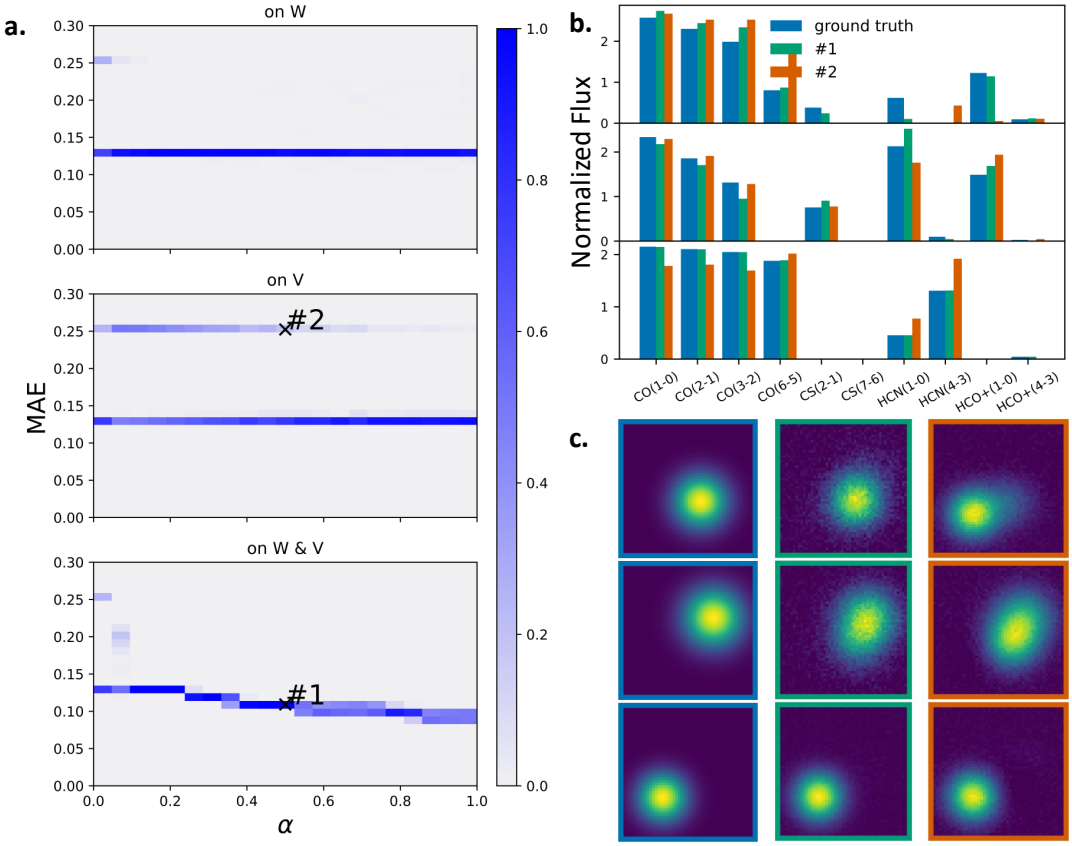

Figure 3.3: ***a)*** *Binned density plots characterizing the performance of an ensemble of three-component NMF runs on X_{proto} in which x-axis coordinates quantify the amplitude of the regularization term and y-axis coordinates the goodness of fit of the retrieved components as defined by the MAE metric introduced in Section 3.5. Each panel considers a different type of regularization: i) only on W (top panel), ii) only on V (central panel), or iii) on both W and V (bottom panel).* ***b)*** *Emission profiles of the true components alongside those of two runs in the ensemble whose location in the scatter-plot are represented by markers #1 & #2.* ***c)*** *Convolved spatial contributions of the same two runs.*

3.5.2 High-z galaxy (three-component fits)

In this next section, we study the performance of the NMF algorithm on a synthetic dataset approximating line-intensity maps from a high-redshift galaxy X_{gal}. Unlike in the previous experiment where components had only moderate spatial overlap, in this scenario all components peak in the same location - the center of the galaxy - but have different spatial extents. This high-overlap between components complicates the retrieval process as there are fewer pixels which can clearly be attributed back to a unique component. This dataset thus constitutes a more challenging test for NMF but an important one as the NMF algorithm will only be practical if it can reliably produce useful retrievals across a range of different scenarios.

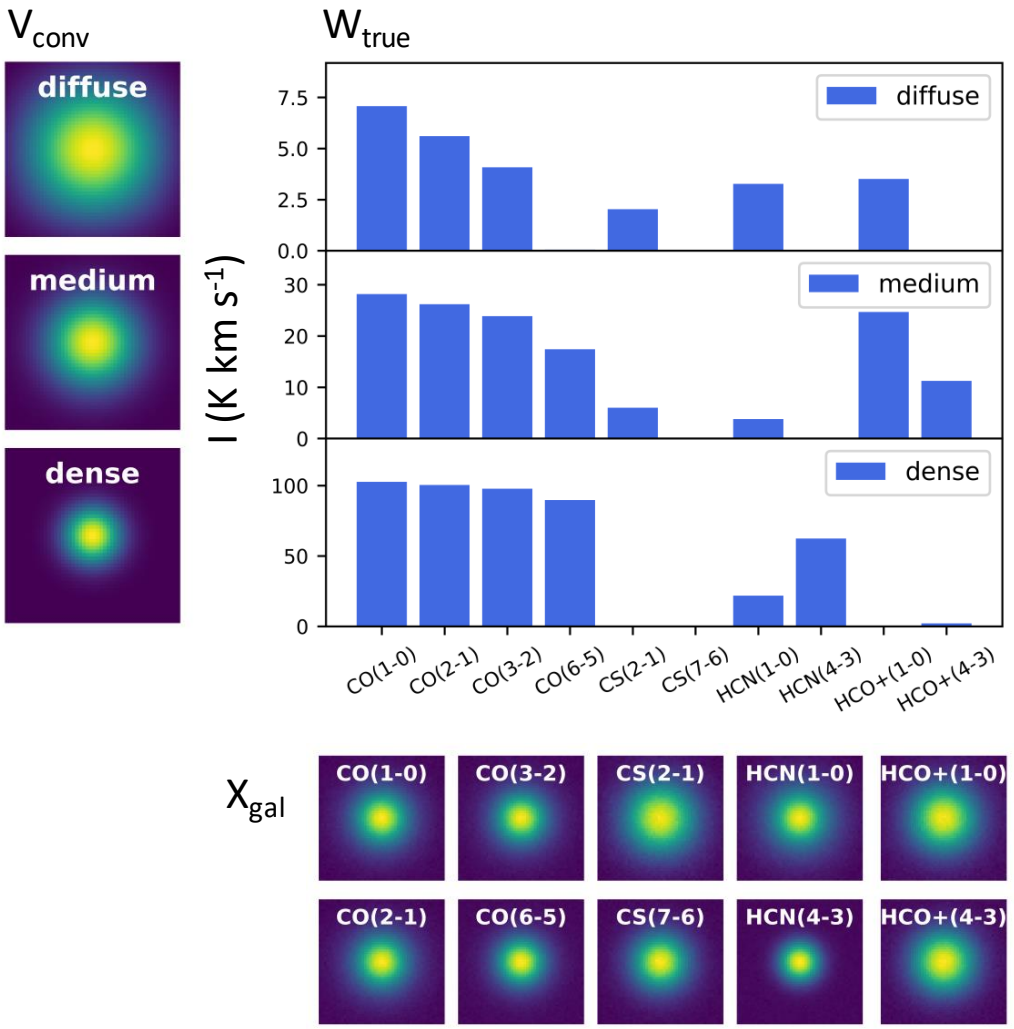

Figure 3.4: *Synthetic line-intensity maps of high-redshift galaxy (X_{gal} : bottom) alongside the emission profile (W_{true} : top-right) and convolved (unitless) spatial contribution (V_{conv} : top-left) of each components contributing to the maps.*

The components used in this run are a less dense component (T=10 K and n=10^4cm^{-3}), a medium density moderately extended and warm component (T=30 K and n=10^5cm^{-3}), and a dense compact hotter central component (T=50 K and n=10^7cm^{-3}). We note that, for a spatially unresolved high redshift galaxy, our three gas components are definitely more arbitrary in nature than previous examples. Nevertheless, we may identify the first component with the average cold molecular gas, typical of a giant Molecular Cloud, the second component with dense molecular gas heated by either nearby star formation or an AGN, and the third component with compact very dense gas found in the proximity of the nucleus of a galaxy.

Once again, we validate the algorithm through plotting i) line-intensity maps and

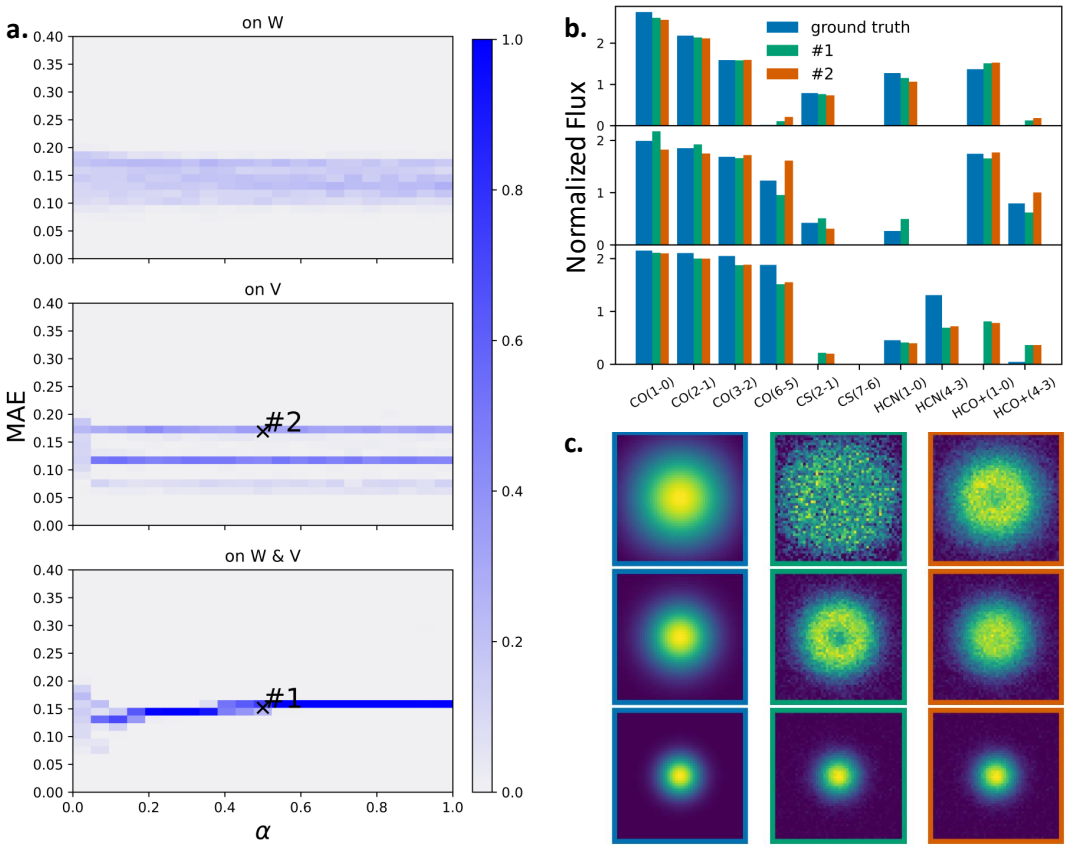

Figure 3.5: ***a)*** *Binned density plots characterizing the performance of an ensemble of three-component NMF runs on X_{gal} in which x-axis coordinates quantify the amplitude of the regularization term and y-axis coordinates the goodness of fit of the retrieved components as defined by the MAE metric introduced in Section 3.5. Each panel considers a different type of regularization: i) only on W (top panel), ii) only on V (central panel), or iii) on both W and V (bottom panel).* ***b)*** *Emission profiles of the true components alongside those of two runs in the ensemble whose location in the scatter-plot are represented by markers #1 & #2.* ***c)*** *Convolved spatial contributions of the same two runs.*

associated components, ii) density plots showing the MAEs for an ensemble of three-component runs, and iii) the retrieved components for two runs whose locations in the scatter plot are shown by markers #1 and #2. These plots are shown in Figure 3.4 and 3.5. NMF runs use three NMF components to fit the line-intensity maps which is a number equal to the number of components within observations and as many as was used in the previous experiment.

We can see from these figures that there is a relatively good level of agreement between the true components and those recovered in runs #1 and #2, with runs having at most a MAE of 0.2 (i.e. at most 20% of the total flux being misplaced). This suggests that for this high-overlap scenario, NMF can approximately recover the underlying components.

However, the mismatch between true and recovered components appears larger here (at least in terms of contributions) than in the previous low-overlap scenario (Section 3.5.1). In particular, the algorithm seems to struggle with decomposing the fully overlapping central component favouring a decomposition into spatially non-overlapping components despite how the underlying components do in fact have a high overlap. This can be seen by how, in both runs, emission arising from more extended components in the central portion of the observations is misattributed to the less extended retrieved component (bottom panel). This in turn then leads to a mismatch between the true and retrieved emission profile for this less extended components as well as the creation of donut-like holes in the spatial profiles of the other more extended components (at the location of misattributed emission).

Overall, these results highlight the presence of degeneracies when doing retrievals of observations containing highly overlapping components. When such high-spatial overlap components exist, the NMF algorithm struggles to recover the proper overlap between components. Furthermore, no choice of regularization appears capable of lifting these degeneracies. Nevertheless, recovered components still match fairly well with observations.

3.5.3 High-z galaxy (two-component fits)

In real-world applications, the number of components used in the NMF retrieval will need to be estimated from observations. However, because of subtle differences in physical conditions and chemical make-up, different spatial locations may emit slightly differently leading to the existence of a high number of components. In such cases, to maintain the regime of fewer NMF components than observed line-transitions, NMF retrievals must be run with fewer NMF components than required to fully capture observations. The aim is then not to retrieve the true components but instead to retrieve useful components. We examine here how NMF behaves in this regime.

To do this, we run NMF on our high-z line intensity maps but using only two components instead of the three components actually required for reproducing the line-intensity maps. This provides a test for the behaviour of the NMF algorithm when data is more complex than the available number of components. Figure 3.6 shows the output of such two-component runs. Here, we do not show density plots since our goodness of fit metric is not applicable when the number of retrieved and ground truth gas components differ.

Comparing these runs to their three-component counterparts (Figure 3.5) reveals that

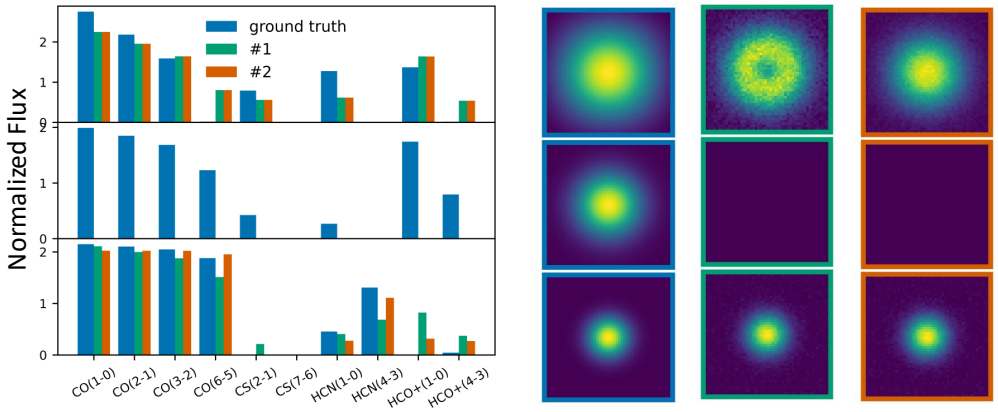

Figure 3.6: *Emission profiles and (unitless) convolved spatial contributions of the true components alongside those of two random runs in an ensemble of two-component runs fitted to* X_{gal}.

the two-component runs seem to only differ from the three-component runs in their treatment of the top and middle-panel components. Because the two-component runs are incapable of fully capturing the emission from all three components, they groups together the top and middle-panel components and instead return a component whose emission profile is approximately a mixture of the emission from these two components. On the other hand, the lowest-panel component of the two-component runs, characterizing the densest component, is almost identical to the analogous component in the three-component runs. In fact, similarly to the three-component runs, it (mistakenly) captures emission from other gas-phases within the central region leading to a similar mismatch between retrieved and true dense component and a similar (incorrect) donut shape spatial emission in other components.

Conceptually, it appears that because the two-component NMF runs did not have enough capacity to reproduce all three components, they selectively grouped the two most similar components - the top and middle panel components - into a single "archetypical component" capturing the average emission across both phases whilst dedicating the remaining component to the densest component because it emitted more uniquely. Such behaviour suggests that, when lacking enough components to fit observations, the NMF algorithm will fall-back on grouping together the most similar components into archetypical components. This means that the algorithm could allow for synthesizing highly-complex observations into a manageable number of components for human interpretation.

3.5.4 Discussion

We have run a series of experiments aiming at recovering the underlying components within observations of molecular line-intensity maps. These experiments were run on a set of synthetic datasets designed to replicate conditions expected from real observations while also capturing the full diversities of environments where the algorithm could be applicable. Because NMF ignores the spatial locations of pixels and would perform equally well if pixels were scrambled, we did not put a strong emphasises on reproducing the geometrical structure of real molecular maps and instead focused on reproducing the expected overlap between components. To capture the diversity of environments we built two synthetic line-intensity maps on which to test our observations: one with a moderate overlap between components and a second dataset with a high overlap. We validated the effectiveness of NMF on these synthetic environments through studying the algorithm's ability to recover the components used in the data-generating process. Because the convergence of the NMF algorithm was found to be sensitive to hyperparameters, we took careful care to characterize the performance across a range of hyperparameters rather than for a single set of hyperparameters.

The results of our investigations were as follows. We found that the NMF algorithm was a valuable tool for separating out components within line-intensity maps as demonstrated by its ability to often recover components resembling those used in the data-generating process. We found that when the NMF algorithm lacked the capacity to reproduce all components, it would group together the most similarly emitting components into one archetypical component. However, in general the matrix-factorization problem solved by NMF did not have a unique solution and so its convergence to a set of components closely resembling the underlying components was not guaranteed. Instead, initialization and regularization played a crucial role towards influencing the type of solution the algorithm converged towards with sometimes very different solutions depending on the hyperparameters. This sensitivity of NMF to model hyperparameters is obviously not ideal as it means that the algorithm can sometimes recover components that match poorly with observations. In general we also found that the NMF algorithm struggled with recovering the proper spatial overlap of overlapping components.

Our experiments provide a proof-of-concept towards the applicability of NMF for analyzing line-intensity maps. However, some differences do still exist between real data

and the data within our experiments. One such difference is that our mock line-intensity maps shared a common beam size. When working with real data, it will be crucial to down-sample observations so as to guarantee that parcels of emitting gas have the same spatial extent across all line transition maps. Another difference is that because of effects such as optical depth and differences in local physical conditions, it is unlikely that real data will be perfectly characterized as having arisen from a small number of components. This means that the idealized conditions under which our experiments were generated will partially break down. However, as demonstrated by our experiment in which we used a smaller number of NMF components than the number of components, even when the data-generating process is not perfectly satisfied by observations NMF still retains the ability to uncover meaningful components in the form of archetypal components.

A weakness of this work is that our experiments only considered scenarios in which observations were created from components at similar column densities and thus in which all components contributed more or less equally to the line-intensity maps. When components have dissimilar column densities then it becomes much more challenging for NMF to recover the fainter components. This is because as the lower column density components emit noticeably more weakly, correctly fitting them does not contribute much to the minimization of the loss function. For synthetic observations containing low-column density components, in order to recover the weakest components it is crucial that the signal-to-noise within observations be high enough for these weakest components to not be drowned by noise.

In astronomical observations, this assumption of the observed emission being produced by N distinct components is typically untrue. Instead, a spectrum of gas conditions will contribute to the emission; the majority of which may be well approximated by N distinct components. Astronomers use this assumption to fit data and the NMF procedure described in this chapter presents a method of retrieving these components without bias from the astronomer. In light of this, if one NMF component's contribution to the overall emission is negligible, it makes sense to fit one fewer component to the data rather than attempting to force the algorithm to retrieve a negligible NMF component which may not reflect any real gas component. Additionally, on real observations it may be especially hard to retrieve weak components as the NMF algorithm will prioritize obtaining a good fit to the high-emission regions before attempting to fit the much weaker low column density regions.

We also wish to emphasize that our work provides only one, amongst potentially many, approaches for incorporating matrix factorization constraints into the component retrieval process. A potential extension of this work would be to combine NMF with radiative-transfer codes to enforce on retrieved components both the matrix factorization constraints of NMF and the physical realism constraints of radiative-transfer codes. This could perhaps be done through a two-stage approach in which Bayesian Non-negative matrix factorization (Schmidt et al. 2009) - a probabilistic extension of NMF - is first used to find a probability distributions functions over components plausibly matching observations and RADEX is secondly used to weight this distribution (which could for example be done through rejection sampling) based on a likelihood of such phases being produced according to RADEX. Application of this procedure would result in a posterior capturing both RADEX and factorization constraints. Alternatively, geometrical constraints could be incorporated into the NMF factorization, as was for example done in Melchior et al. (2018).

Regardless of the exact approach taken, the large degeneracies associated with radiative transfer currently make it extremely difficult to constrain the properties of the interstellar medium especially at high-redshift. Combining the information available across all spatial locations during inference, as is done in NMF and other matrix factorization approaches, is, we believe, a promising approach to address such degeneracies. However, whilst NMF is useful for understanding the relationship between lines within observations, on its own it is not sufficient for fully constraining components.

This page was intentionally left blank

Chapter 4

Disentangled Representation Learning for Chemical Tagging

The work presented in this Chapter is based on the paper de Mijolla et al. (2021), in collaboration with Melissa Ness, Serena Viti and Adam Wheeler.

4.1 Introduction

Galactic archaeology, the sub-field of astronomy interested in reconstructing the Galaxy's history, has recently experienced substantial growth. This growth has been spurred on by stellar surveys such as RAVE, APOGEE, GALAH, LAMOST, Gaia and Gaia-ESO (Steinmetz et al. 2006; Majewski et al. 2017; De Silva et al. 2015; Cui et al. 2012; Gaia Collaboration et al. 2018a; Gilmore et al. 2012; Randich et al. 2013). These surveys have obtained spectra, and in the case of Gaia astrometry and photometry as well, for hundreds of thousands to millions of stars across the Galaxy. These data have enabled measurement of stellar abundances, distances, proper motions and ages across the Galaxy. Future missions are also on the horizon which will provide large, high-quality stellar samples probing the different Milky Way components including the faintest stars (de Jong et al. 2016; Kollmeier et al. 2017; Bonifacio et al. 2016; Tamura et al. 2016).

Chemical element abundances derived from stellar spectra are core to archaeological pursuits. While there are evolutionary and environmental factors which can impact the

surface abundance of a star (e.g. Liu et al. 2019; Casey et al. 2019), abundances link stars to individual molecular clouds, which give their stellar brood similar chemical fingerprints (Feng & Krumholz 2014; Bovy 2016b; Krumholz et al. 2019; Ness et al. 2018; Liu et al. 2019). However, the chemical space of stars in the Milky Way's disk seems fairly low-dimensional. Recent research suggests that stars born at the same radius and time are chemically similar or even identical within measurement precision (Ness et al. 2019; Weinberg et al. 2019; Ting et al. 2012; Price-Jones & Bovy 2019) which would indicate that, at current precisions, only two dimensions (radius and age) are required for modelling abundances in the Milky Way disk. Such results are further corroborated by studies of the chemical similarity of field disk stars which find that, at solar metallicity, around 1 percent of field stars in the APOGEE survey are as chemically similar as stars that are known to be from the same individual birth cluster (Ness et al. 2018). This doppelganger rate alone renders chemical tagging of stars to their individual birth sites, using ≈ 20 abundances alone, rather difficult.

Chemical element abundances derived from stellar spectra are core to archaeological pursuits. While there are evolutionary and environmental factors which can impact the surface abundance of a star (e.g. Liu et al. 2019; Casey et al. 2019), abundances link stars to individual molecular clouds, which give their stellar brood similar chemical fingerprints (Feng & Krumholz 2014; Bovy 2016b; Krumholz et al. 2019; Ness et al. 2018; Liu et al. 2019). However, the chemical space of stars in the Milky Way's disk seems fairly low-dimensional, with stars born at the same radius and time being chemically similar or even identical within measurement precision (Ness et al. 2019; Weinberg et al. 2019; Ting et al. 2012; Price-Jones & Bovy 2019). Indeed, at solar metallicity, the APOGEE survey shows that 1 percent of field stars are as chemically similar as stars that are known to be from the same individual birth cluster (Ness et al. 2018). This doppelganger rate alone renders chemical tagging of stars to their individual birth sites, using ≈ 20 abundances alone, rather difficult. Nevertheless, identifying chemically identical or near-identical stars, has high utility in reconstructing the galaxy's formation. For example, in estimating the number of star-forming clusters in the galactic disk (e.g. Kamdar et al. (2019) and Ting et al. (2016)) or for understanding how stars have moved over time (e.g. Beane et al. (2018); Coronado et al. (2020); Price-Jones et al. (2020) and Frankel et al. (2018)). Furthermore, detailed abundances allow connecting stars to their birth radii as well as their time of formation (Ness et al. 2019; Bedell et al. 2018; Feuillet et al. 2019; Casali et al. 2020)

Typically, efforts to identify chemically identical stars have involved estimating surface abundances by comparing observations to synthetic spectra, and then running a clustering algorithm (Price-Jones & Bovy 2019; Hogg et al. 2016). This procedure is hampered by its reliance on imperfect stellar models to obtain the abundance labels that describe the spectra. Typically employed 1D non-LTE stellar simulations do not fully capture the complexity of stellar photospheres. Often only a fraction of the spectrum (the locations of a subset of cleanly identified features) is utilized. There may also be systematic abundance offsets in the derived abundance labels due to signal-to-noise dependencies of their derivation, or unmodelled instrumental imprints on the spectra, meaning that abundance estimates are subject to artefacts (e.g. Holtzman et al. 2015). Data-driven approaches have provided higher precision abundances for stars across surveys (e.g. Ness et al. 2015; Ho et al. 2017; Casey et al. 2017; Wheeler et al. 2020). However, these approaches still at their core rely on stellar models to provide stellar parameter and abundance labels for the training data.

In this thesis chapter, we demonstrate the feasibility of identifying chemically identical stars without explicit use of measured abundances. We apply a neural network with a supervised disentanglement loss term to a synthetic APOGEE-like dataset of spectra. The model learns a representation of spectra that traces abundances independently from the non-chemical factors of variation. That is, it controls for changes in the spectra caused by, for example, effective temperature, $T_{\rm eff}$, and surface gravity, $\log g$. This isolates the chemical variation expressed in the spectra. Stars with identical chemical compositions but differing $T_{\rm eff}$ and $\log g$ are mapped to nearly identical representations.

Unlike approaches based on explicit abundance estimates, this model naturally exploits the full available wavelength range including blended lines to estimate, effectively, chemical composition. Additionally, it does not depend on stellar models and so does not suffer from associated systematics. Furthermore, we find that the learned low-dimensional representation of synthetic spectra can be transformed linearly into abundances with high precision.

Our method relies on the assumption that there does not exist any correlation, nor statistical dependencies, between physical and chemical factors of variation. Although such an assumption has been used in the past (Jofré et al. 2019; Valenti & Fischer 2005), stellar processes, such as atomic diffusion and dredge-up, contribute to modifying surface abundances away from their birth values (Dotter et al. 2017). Because our model

learns a representation of stellar spectra in which all variation dependent on non-chemical parameters is removed, assuming evolutionary changes in abundances correlate with the non-chemical factors we parametrize, the effect of these processes should also be removed from the representation, meaning that it will reflect birth, rather than present-day, abundances.

Studies of open cluster populations have demonstrated that stars can change in their element abundances by 0.1-0.3 dex across the main sequence to the giant branch (Souto et al. 2019; Bertelli Motta et al. 2018). This is also in line with theoretical expectations and a consequence of physical processes like atomic diffusion (Dotter et al. 2017). It is therefore a relevant and important distinction that we interrogate the spectra of a star for its birth abundance composition as opposed to its present day composition.

This thesis chapter is structured such that technical work on supervised disentanglement, of potential interest outside the astronomy community, is presented separately from its astrophysical application. After introducing related work in Section 4.2, we present our method in Section 4.3. Finally, in Section 4.4, we adapt this method to the astronomical task of chemical tagging and show experimental results on an APOGEE-like dataset, demonstrating the recovery of chemically identical stars in the presence of noise and comparing to a baseline method (Price-Jones & Bovy 2019). We finish by discussing in Section 4.6 some important aspects of our method which are not fully explored by our experiments on synthetic data.

4.2 Related Work

This thesis chapter makes use of neural networks. We refer the reader to Section 1.4.3 of this thesis for an overview on neural networks.

4.2.1 Disentangled representation learning

There is a growing body of literature on using neural networks for learning to encode data into interpretable representations. Unsupervised disentanglement methods, such as beta-VAE (Higgins et al. 2017) and InfoGAN (Chen et al. 2016), attempt to find representations in which distinct informative factors of variation (for example lighting conditions and object orientation in the context of images) are encoded in separate dimensions (Bengio et al. 2013). However recent results suggest that finding such disentangled representations

in a fully unsupervised setting is fundamentally ill-posed without additional assumptions or priors being set (Locatello et al. 2019).

Supervised disentanglement methods (Schmidhuber 1991; Ganin et al. 2016; Lample et al. 2017) specify labels for factors of variations that should be excluded from the learnt representation. These methods aim to find a representation of inputs in which the specified factors of variation are selectively removed from the representation. Ideally, a perfectly disentangled representation would be statistically independent from the specified factors of variation. However, there will often be a trade-off between disentanglement and reconstruction (Lezama 2019).

Supervised disentangled learning has primarily been implemented through an adversarial training scheme, in which an autoencoder—a neural network with a low-dimensional bottleneck that is trained at reconstructing inputs—learns to encode its input in such a way that a second network is unable to predict the to-be-disentangled labels from the encoded representation (Lample et al. 2017; Edwards & Storkey 2016; Hadad et al. 2018). It has also been proposed to obtain a disentangled representation by enforcing that an autoencoder learn a representation in which the latent (i.e. the representation at the autoencoder bottleneck) and the labels are statistically independent. This has been done within the variational autoencoder framework in Louizos et al. (2016), but also with adversarial autoencoders in Polykovskiy et al. (2018). Another existing avenue for obtaining supervised disentanglement can be found through a cyclic training scheme, that encourages the latent to remain unchanged after re-encoding outputs, obtained after modifying the factors of variation. This approach has been demonstrated in the context of variational autoencoders in Chen et al. (2019) and in Jha et al. (2018).

Supervised disentanglement could be a very useful technique in the field of astronomy and we hope that this thesis chapter will be beneficial for showcasing its potential. For example, supervised disentanglement could be used in astronomical calibration to remove the effects of individual fibers or weather conditions on spectra. This could be done by learning a representation that is, for example, statistically independent from the fiber number for a multi-object spectrograph. Such an approach would be complimentary to the work proposed in this chapter, as our proposed method requires precisely calibrated spectra and additional augmentation to handle systematic artifacts.

4.2.2 Data-driven chemical tagging

Chemical tagging, as introduced in Section 1.3 of this thesis, describes the reconstruction of individual cluster groups via abundance information (Freeman & Bland-Hawthorn 2002; Ting et al. 2016; Casey et al. 2019). The concept has extended to the identification of chemically anomalous stars of particular formation origins (Hogg et al. 2016; Schiavon et al. 2017), the association of and differentiation between stellar groups and populations using abundances (Simpson et al. 2019; Hawkins & Wyse 2018; Martell et al. 2016) and grouping stars by chemical similarity (e.g. Price-Jones & Bovy 2019). Recent work indicates there is limited feasibility of chemically tagging stars back to their individual cluster origins using the ≈ 20 individual abundance measurements alone from resolution R=22,500 spectra (e.g. Ness et al. 2018). Most approaches use abundance labels that describe the spectra and new approaches have improved the precision of these labels (Ness et al. 2015; Leung & Bovy 2018; Ting et al. 2019). Novel approaches to chemical tagging include those presented in Blanco-Cuaresma & Fraix-Burnet (2018) and Jofré et al. (2017) which use techniques from the field of phylogeny and Price-Jones & Bovy (2019) who identify chemically identical stars without explicit use of abundances. The concurrent work presented in O'Briain et al. (2021) uses a machine learning algorithm loosely similar to ours for improving stellar abundance estimation.

The method proposed in Price-Jones & Bovy (2019), which itself expands upon earlier work presented in Price-Jones & Bovy (2017a), bears some clear similarity to our work in that it uses a data-driven model applied directly to spectra to learn a representation in which undesirable parameters are removed. Additionally, some level of similarity can be found in the fact that they use Principal Component Analysis for dimensionality reduction, which can be viewed as a linear counterpart to our use of an autoencoder. In this approach, the authors fit a polynomial model of the non-chemical parameters to every single wavelength bin. The residuals of this fit are then considered to only contain chemical information. To identify chemically-similar groups, the authors then run a clustering algorithm on a compressed representation of the residuals obtained after principal component analysis. However, as discussed in their paper, this method comes with some limitations. A polynomial fit may not be an optimally flexible functional form, particularly across a breadth of stellar evolutionary states (see for example Ting et al. 2019). As such, it is unlikely to perfectly remove physical parameters of variation from the residuals. Fur-

thermore, by fitting non-chemical parameters in isolation, any joint dependencies between chemical and non-chemical factors of variation on spectral line strengths are ignored.

4.3 Methods

We present, here, an overview of our method. Subsection 4.3.1, introduces our underlying assumptions on the data-generating process, the problem we are trying to answer, and the broad-strokes of our method. In Subsection 4.3.2, we dive deeper and present a neural network architecture for solving our introduced problem. In Subsection 4.3.3, we focus on two different methods of enforcing a disentangled representation - a key component in our method.

4.3.1 Problem statement

We consider a setup in which a dataset $X = \{x_1, ...x_n\}$ is observed. We assume the dataset to be generated deterministically from latent variables through a mapping unknown to us. Despite not knowing this mapping, we assume that a subset of the latent variables can be accurately estimated. As such we can subdivide latent variables into a vector of known variables u and a vector of unknown variables v. For our method to work, we further assume that u and v are (marginally) statistically independent (i.e. $p(v|u) = p(v)$). This corresponds to the notion that u and v cannot be used to predict each other.

The core technical contribution of this chapter is to present a general method for quantifying the similarity of observations x as measured in terms of unknown variables v. In particular, this allows for identifying observations x sharing identical or near identical vector v without knowledge of the mapping from latent to observed variables.

Our method learns a mapping, parameterized by a neural network, from observations x to a vector z acting as a proxy for unknown variables v. More precisely, we learn a mapping such that observations sharing a common parameterization for v, in turn, share a near identical representation for z.

We achieve this through finding a representation z which is statistically independent from the known and provided parameters, u, but when combined with these known parameters, capable of perfectly reconstructing observations x. This ensures that our latent variables contains all the information contained within the unknown variables v but not any additional superfluous information.

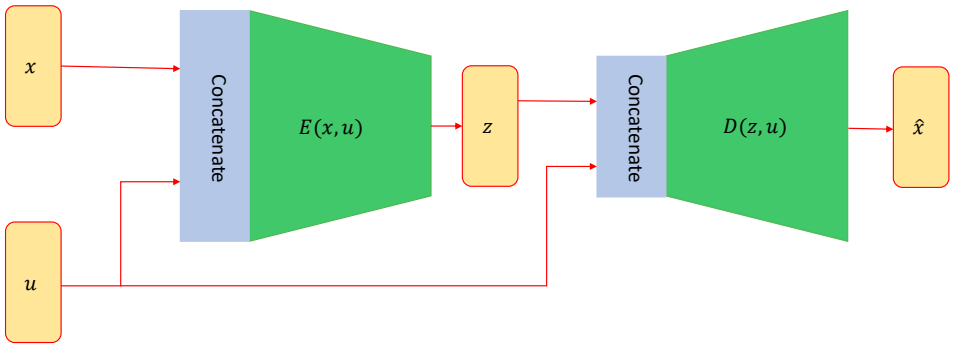

Figure 4.1: *Diagram of the conditional autoencoder architecture. We denote the reconstructed observation as $\hat{x}$. For chemical tagging, x corresponds to stellar spectra and u to phyiscal factors of variation.*

How does this assumed setup relate back to astronomical chemical tagging? For chemical tagging, we have access to stellar spectra of stars, x, from which we seek to identify stars sharing an identical chemical composition, v. Although we are capable of estimating physical parameters, u, fairly accurately, shortcomings in spectral synthesis make it difficult to relate spectra back to their chemical composition.

4.3.2 Approach

We rely on a conditional autoencoder, an autoencoder neural network architecture in which additional inputs are passed to the network beyond those the network is trained to reconstruct, to learn the mapping to the lower dimensional representation z. Our autoencoder (represented in Figure 4.1) is composed of two seperate neural networks. A conditional encoder taking as inputs observations, x, concatenated with known parameters, u, and returning a latent representation, z, (for the remainder of the chapter we adopt machine learning terminology and refer to z as latents) and a conditional decoder taking z and u as input and trained to output reconstructed observations x.

This autoencoder is trained to minimize the following loss function:

$$L_{AE} = L_{rec} + \lambda L_{dis} \tag{4.1}$$

where L_{rec} is a reconstruction loss. In our experiments we used the mean squared loss [1]

$$L_{rec} = E_{(x,u)\sim p(x,u)}[\|D(E(x,u),u) - x\|_2^2] \tag{4.2}$$

L_{dis} is a disentanglement loss acting to ensure that the latent, z, is maximally disentangled from the known and provided parameters, u. λ is a term controlling the trade-off between reconstruction and disentanglement whose value is determined using a hyperparameter search. The disentanglement loss serves to incentivize the network towards learning a latent representation statistically independent from the observed parameters u such that its value be minimal when z and u are independent. We present two formulations of L_{dis} in Section 4.3.3.

During training, the autoencoder is iteratively shown datapoints, grouped into batches (i.e. subsets of the dataset). The autoencoder's loss, as described above, is evaluated on each batch and the derivatives of this loss with respect to the neural network parameters are used to update the parameters in the direction minimizing the loss function. After training, the neural network will have converged to parameterizing a mapping which (locally) minimizes the loss function. Although not a global minima, the learned mapping, in part because of the stochastic nature of the training process, will typically be a good minimizer of the loss function.

Our neural network, in minimizing the loss function described by Eq. 4.1, simultaneously minimize reconstruction and disentanglement terms with a trade-off controlled by λ. Minimizing the disentanglement loss term corresponds to learning a latent representation statistically independent from factors of variation parameterized by u. This is achieved by removing all related information from the latent. The reconstruction term will be minimized when z and u are sufficient for reconstructing observations x. Combined, these two loss terms will be minimized when all the information required for modelling observations x not included within u is contained within the latent z. While it may not always be possible to minimize both loss terms together, we know that it is possible to do so for data generated as described in Section 4.3.1. Indeed, a global minimum of the loss function would be reached for a neural network which encoded observations x into v and decoded back to x.

[1] $\|x\|_2 := \sqrt{x_1^2 + ... + x_n^2}$

In addition to isolating unknown factors of variation, v, we have found that, at least for the problems we have considered, supervised disentanglement maps observations with shared parameter values v, to nearly identical latents, z. We attribute this to our set of assumptions (see Section 4.3.1). This property makes some intuitive sense when we take a moment to consider how our autoencoder might map observations, x, generated from a common shared vector of unknown parameter values, v, but each with different values of the observed parameters, u. If the mapping does not project all of these observations to a common latent value, then the latent value, z, will be informative about the parameter value u (as some u are then more or less likely based on the observed z). Therefore, z and u will no longer be statistically independent.

In practice, our neural network will only approximately minimize our loss function and so will not perfectly map observations sharing common parameter values, v, to the same latent z. Observations sharing common parameter values will thus appear as overdensities in the latent space. These over-densities can then be identified, for example by running a clustering algorithm such as K-means (Lloyd 1982), or by finding those observations particularly close according to some distance metric. Alternatively, we can instead identify such observations in the data-space if we use the decoder to convert all latents with a common set of parameters, u_i.

4.3.3 Implementation of supervised disentanglement

We present two alternative methods, FaderDis and Factordis, for learning a disentanglement loss L_{dis} encouraging statistical independence. FaderDis is an adaptation of the Fader disentanglement architecture presented in Lample et al. (2017) modified for our purposes. FactorDis is, to our knowledge, a novel architecture for supervised disentanglement. We present here the architectures investigated.

Factor Disentanglement (FactorDis)

The FactorDis method enforces independence by training a critic network to differentiate between samples from the joint distribution $p(x, u, z)$ and samples in which the statistical dependency between z and u has been forcibly removed. Analogously to generative adversarial networks (Goodfellow et al. 2014), the conditional autoencoder is adversarially trained to generate samples that hinder the critic network's ability to do its job.

The joint distribution $p(x, u, z)$ can be expressed using Bayes rule as:

$$p(x, u, z) = p(x|u, z)p(u, z) \tag{4.3}$$

This can be rewritten as

$$q(x, u, z) = p(x|u, z)p(u)p(z) \tag{4.4}$$

if and only if u is statistically independent from z. If the joint distribution $p(u, z)$ is not factorizable, the distributions $q(x, u, z)$ and $p(x, u, z)$ will be different. It follows from this, that u and z are statistically independent, if and only if samples (x, u, z) drawn according to $p(x, u, z)$ are indistinguishable from those sampled from $q(x, u, z)$.

How can we generate samples from these two distributions? If we consider our autoencoder to be an idealistic autoencoder capable of perfectly reconstructing its inputs, then the encoder and decoder can be viewed as respectively approximately parameterizing $p(z|u, x)$ and $p(x|u, z)$, which are both deterministic functions. We can thus draw samples from $p(x, u, z) = p(z|u, x)p(u, x)$ by first randomly sampling from the dataset to obtain (u, x), and then using the encoder to obtain the associated z.

We can similarly draw samples from $q(x, u, z) = p(x|u, z)p(u)p(z)$ through reusing our samples drawn from $p(x, u, z)$. By scrambling (u, z) pairs within a batch, we can effectively remove any joint information between u and z (Belghazi et al. 2018) which results in samples drawn from the marginal distribution $(u, z) \sim p(u)p(z)$. We can then use the decoder which approximates $p(x|u, z)$ to obtain approximate samples $q(x, u, z)$ drawn from $p(x|u, z)p(u)p(z)$.

As stated above, enforcing statistical independence is the same as finding a latent representation, z, such that samples drawn according to these two procedures are indistinguishable. This objective bears strong similarity to the training objective of generative adversarial networks which attempt to train a generator such that generated samples are indistinguishable from samples drawn from a dataset. As such, we can take inspiration from existing generative adversarial network architectures to solve our disentanglement objective.

In generative adversarial networks (Goodfellow et al. 2014), a critic network is trained to distinguish between samples drawn from a dataset, and samples created by a generator network fed samples from a well-understood probability distribution. The generator and

critic network are jointly optimized in a minimax game. That is, the critic attempts to maximally distinguish between the two data streams and the generator attempts to minimize the critic network's ability at doing so. The global optimum of this two player game occurs when both the generator and critic network can no longer improve - when the two data streams are identical.

For our disentanglement neural network architecture, we take heavy inspiration from Wasserstein generative adversarial network (Arjovsky et al. 2017). We use an architecture parallel to that of generative adversarial networks. Similarly to generative adversarial neural network, a critic neural network is trained to distinguish between two streams of data and a second network is trained to improve the similarity between the two streams of data using the feedback from the critic network. However, instead of differentiating between real and fake samples, we differentiate between samples from $p(x, u, z)$ and from $q(x, u, z)$ generated using the autoencoder (AE). This leads to optimizing the following minimax objective:

$$\begin{aligned} &\mathrm{min_{AE}max_{C\in\mathbb{D}}\mathbb{E}_{(x,u,z)\sim p(x,u,z)}[C(x,u,z)]} \\ &-\mathbb{E}_{(x,u,z)\sim q(x,u,z)}[C(x,u,z)] \end{aligned} \tag{4.5}$$

where $\mathbb{D}$ is the space of 1-lipschitz continuous functions and $C(x, u, z)$ refers to a critic network that takes as inputs a vector in which observations x, latents z and parameters u are concatenated and attempts to differentiate between the different type of samples generated by our autoencoder.

The critic network attempts to maximize Eq. 4.5. In order to constrain the critic network to learn a lipschitz continuous function, we add a gradient penalty term, weighted by a constant λ, to the loss as was introduced in Gulrajani et al. (2017). This leads to a critic loss function

$$\begin{aligned} L_{critic} &= \mathbb{E}_{(x,u,z)\sim q(x,u,z)}[C(x,u,z)] - \mathbb{E}_{(x,u,z)\sim p(x,u,z)}[C(x,u,z)] \\ &+ \lambda\mathbb{E}_{(x,u,z)\sim r(x,u,z)}[(\|\nabla_{x,u,z}C(x,u,z)\|_2 - 1)^2] \end{aligned} \tag{4.6}$$

where $r(x, u, z)$ is implicitly defined as sampling uniformly along straight lines between pairs of points sampled from the distributions $p(x, u, z)$ and $q(x, u, z)$. Further information about this sampling procedure can be found in Gulrajani et al. (2017).

Our autoencoder, which plays the role of a generator network, is trained to minimize Eq. 4.5 while simultaneously minimizing the reconstruction loss function:

$$L_{AE} = L_{rec} + \lambda_2 \mathbb{E}_{(x,u,z)\sim p(x,u,z)}[C(x,u,z)] \\ - \lambda_2 \mathbb{E}_{(x,u,z)\sim q(x,u,z)}[C(x,u,z)] \tag{4.7}$$

.

This loss function combines the reconstruction loss that is traditionally used for optimizing autoencoders with a Wasserstein loss. In addition, unlike for generative adversarial networks, as both data streams are passed through the generator, they are both used for optimizing the generator. Training involves jointly minimizing the critic and autoencoder losses. The two different types of losses are weighted by a factor λ_2. Experimentally, we found that it was crucial to correctly set the factor λ_2, such that neither the reconstruction term nor the disentanglement term in the loss dominated over the other.

Fader Disentanglement (FaderDis)

The FaderDis method of disentanglement follows the setup presented in Lample et al. (2017) in which an autoencoder is adverserially trained to learn a latent representation from which an auxiliary network is incapable of predicting u. Since the method is designed to operate on discrete variables, we discretize the parameter of space u into n equal sized bins.

In this method, an auxiliary network, A, accepts latents, z, as inputs and outputs a vector of size equal to the number of discretized bin. It is trained using a cross-entropy loss to predict the probability of the corresponding u vector falling in each of the n bins. The autoencoder AE is then trained alongside this auxilary network. The autoencoder attempts to minimize the auxilary networks loss weighted by a factor λ_1 while also maximizing its own reconstruction loss. The autoencoder loss takes the form

$$L_{AE} = L_{rec} - \lambda_1(E_{(x,u)\sim p(x,u)}[-u_n log(A(E(x,u)))]) \tag{4.8}$$

where u_n is a vector containing a single non-zero entry. The non-zero entry in u_n, which has a value of 1, corresponds to the bin (out of n bins) in which the discretized u falls.

The global optimum of this two player minimax game will occur when the autoencoder

has learnt to reconstruct observations using a latent z which does not contain any helpful information for the auxilary network. Since the auxilary network attempts to learn $p(u \mid z)$, this will occur when $p(u \mid z) = p(u)$ or equivalently when u and z are statistically independent.

4.4 Application to Stellar Spectra

We now present the application of our disentanglement framework to stellar spectra. In this context, we aim to learn a representation of stellar spectra that disentangles factors of variation of interest (chemical abundances) from observed parameters, u. Two cases for observed parameters u are considered: i) the case where the metallicity is an observed parameter ($u = [T_{\rm eff}, \log g, [\mathrm{Fe/H}]]$) and ii) the case where it is not ($u = [T_{\rm eff}, \log g]$). As a reminder, we have constructed our architecture such that after training, without any explicit knowledge of abundance labels, v, our neural network will find a mapping from observations, x and parameters, u, to latents, z, such that stars sharing a common abundance are mapped to nearly identical latents. It is worth noting that $T_{\rm eff}, \log g, [\mathrm{Fe/H}]]$ are not technically observed parameters as they must be derived from the spectra using models. We only refer to them as observed parameters as per the narrow definition in our framework (see Section 4.3.1), that is to say we refer to them as observed parameters because we assume that they are preliminarily estimated before the application of our algorithm. Here, we assume perfect knowledge of the true values of these parameters but in real settings they would need to be estimated (for example by using synthetic spectra).

We demonstrate our method using a synthetic dataset described in Section 4.4.1. The dataset is designed to mimic the spectral variability found within the APOGEE red-giant sample. This allows us to carry out a proof of concept for our method in an ideal and controlled environment, in which independence between chemical and physical parameters is guaranteed and for which we were certain to have accounted for all factors of variation. This is an important first step in demonstrating the viability and performance of our method.

We quantify the performance of our generative model with a chemical abundance twin recovery test, comparing to simpler models, that also remove factors of variation $T_{\rm eff}$, $\log g$ and [Fe/H]. We do this for a number of signal to noise qualities. We note that the performance of our method, in practice, will be sensitive to any calibration or instrumental

artifacts that are poorly modelled or not included as observed parameters. We also expect that the dimensionality of real data may be far lower than that of our synthetic library. This is because we do not restrict our realized abundances to the correlations observed in real stars. We therefore make only a comparative analysis of different modeling choices in recovery of abundance twins, rather than make a quantitative prediction of performance for real survey data.

4.4.1 Simulated dataset

For the creation of our spectra, we relied on the APOGEE package introduced in Bovy (2016b), which wraps the Turbospectrum spectral synthesis code (Plez 2012) using ATLAS9 atmospheres (Mészáros et al. 2012). We created identically distributed training and test datasets, both containing 25,000 pairs of chemical abundance twin spectra, sharing identical surface chemical abundances but differing stellar parameters. We generated our spectra assuming solar isotopes ratios.

When creating our spectra, the non-chemical parameters varied were the effective temperature $T_{\rm eff}$ and surface gravity, $\log g$. For each spectrum, $T_{\rm eff}$ and $\log g$ were generated by sampling from uniform distributions as specified in Table 4.1. These parameter ranges were designed to replicate those of red-giant type stars which are the favoured type of stars for chemical tagging (Hogg et al. 2016; Price-Jones & Bovy 2017a) because they are relatively well-mixed and so have, for most elements, surface abundances closely related to birth abundances. Chemical abundances were generated by independently sampling log-metallicity ([Fe/H]), and log-element abundance enhancements ([X/Fe]), assuming Gaussian distributed values. Our Gaussian 1-σ standard deviations were chosen to roughly reflect those observed by the APOGEE survey. Exact values can be found in Table 4.2. These were determined from the 1-σ element abundance dispersion in APOGEE's DR14 for each element, for red giant stars. By fitting separate one-dimensional Gaussians to the element abundance enhancements, we ignore any further correlations that may exist between elemental abundances. In doing this, we will overestimate spectral variability (i.e. dimensionality), which could lead to our chemical tagging predictions being overly optimistic, for the set of stars we consider in our tests. The absolute performance that we later report in recovery of abundance twins, is subject to the number of stars we are evaluating, as well as their density in chemical element abundance space and the dimensionality of the spectra itself. Therefore, it is only the comparative performance between

Parameter ranges

Parameter	Min	Max
T_{eff}[K]	4000	5000
log g[dex]	1.5	3.0

Table 4.1: *Table containing the ranges used for uniformly sampling the non-chemical parameters of variation.*

[X/Fe]	Mean [dex]	Standard Deviation [dex]
[Fe/H]	-0.13	0.24
[N/Fe]	0.28	0.11
[O/Fe]	0.03	0.08
[Na/Fe]	-0.05	0.38
[Mg/Fe]	0.06	0.08
[Al/Fe]	0.07	0.09
[Si/Fe]	0.05	0.07
[S/Fe]	0.05	0.07
[K/Fe]	0.04	0.07
[Ca/Fe]	0.02	0.04
[Ti/Fe]	-0.01	0.06
[V/Fe]	-0.01	0.11
[Mn/Fe]	-0.04	0.07
[Ni/Fe]	0.02	0.04
[P/Fe]	-0.04	0.18
[Cr/Fe]	-0.01	0.06
[Co/Fe]	0.	0.15
[Rb/Fe]	-0.03	0.29

Table 4.2: *Mean and standard deviation used when sampling the chemical factors of variation. Enhancements and metallicity are assumed to be Gaussian distributed in dex.*

the approaches we show that is relevant.

4.4.2 Implementation details

For both FaderDis and FactorDis, we performed a manual hyperparameter search on the training dataset to select the best-performing model. Results for selected models are then shown on the test dataset. We chose to set the latent dimensionality, z, to have dimension 20, slightly exceeding the number of varied abundances in order for the autoencoder to have enough capacity to fully reconstruct inputs. A more comprehensive description of our neural network architectures can be found in Appendix A.1. Our code is available on GitHub at `https://github.com/drd13/tagging-package`.

We also evaluated the performance of the model developed in (Price-Jones & Bovy 2017a, 2019; described in Section 4.2.2), which we refer to as *PolyDis* from now on. We

use PolyDis with a forth-order polynomial, as was found to work best on the training dataset.

To better simulate real data, we also evaluate our methods on a test dataset with added Gaussian noise. For a given signal-to-noise ratio (SNR), we add to every bin of the continuum-normalized spectrum zero-mean Gaussian noise with standard deviation $\sigma = \frac{1}{\text{SNR}}$. For FactorDis, results on the noisy test dataset are obtained by training models on data in which noise of order 1 percent (SNR=100) is added to every observation during training. For FaderDis (and PolyDis), noise of order 1 percent was added to the training data and kept constant for every epoch of training. It was found that adding this type of noise to the training dataset led to worsened spectral reconstruction but improved isolation of chemical factors of variation. The worsened reconstruction can easily be attributed to overfitting to the noise. We do not have a clear explanation for why it led to improved isolation of chemical factors of variation.

4.5 Results

In this section we present a series of experiments comparing and contrasting the capacities of the different models.

4.5.1 Resolving power of latent representation to distinguish chemically identical stars

If non-chemical factors of variation have been perfectly removed from the latent z, stars sharing a common chemical abundance should in turn also share a common latent vector. As such, any difference in latent representation between chemically identical stars can be attributed to imperfections in the learned representation. Here, we use this to compare and contrast how well our considered methods isolate chemical factors of variation.

Figure 4.2 shows histograms of euclidean distances, $\|z_i - z_j\|_2$, calculated on the latents, z, for both chemically identical pairs of stars (blue) and randomly sampled (non-chemically identical) pairs of stars (orange). We show this for our three different disentanglement methods at several SNR. Distances are evaluated on the 20 dimensional latent for both the FaderDis and FactorDis methods, and the 50 dimensional PCA components for the PolyDis approach. We found that 50 principle components explained 99.95% of the variance in the (noiseless) data. In the interest of making the comparison with PolyDis

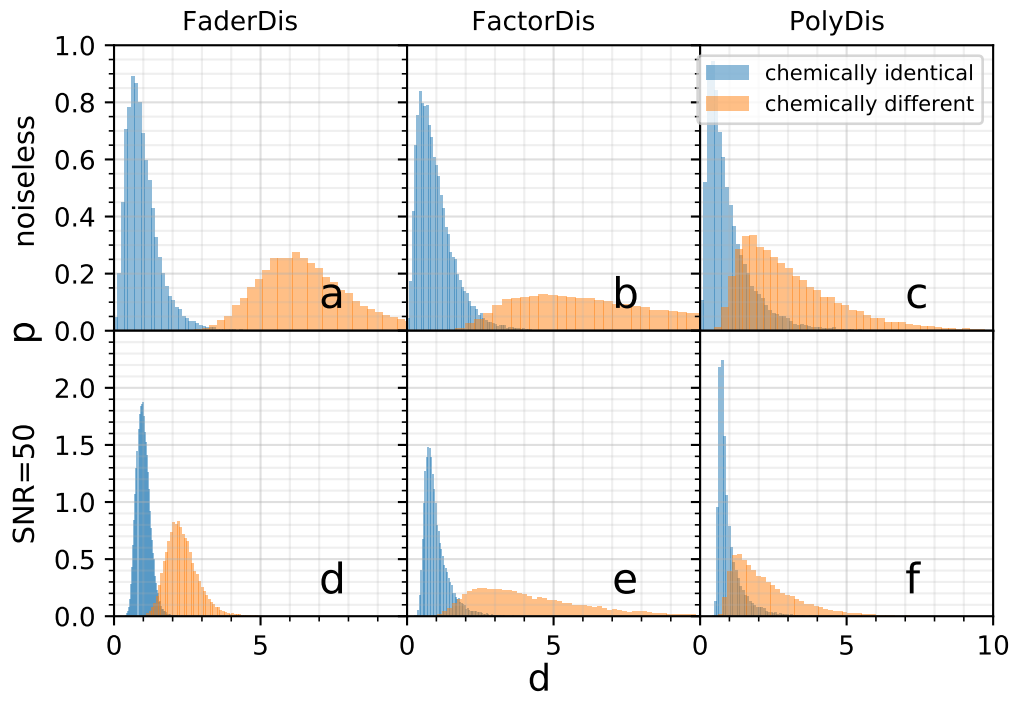

Figure 4.2: *Distribution of scaled euclidian distances, d, for a sample of chemically identical pairs of stars (blue) and fully randomly sampled pairs of stars (orange). For each model, a scaling is applied to the latents such that the mean distance of chemically identical stars is 1. Each model includes $T_{\rm eff}$, $\log g$ and [Fe/H], as the parameters to disentangle from the chemical factors of variation. The top row is evaluated using the noiseless test dataset, the bottom with noise of order SNR=50 added. The first column is evaluated using the FaderDis method, the second using the FactorDis method and the final row using the PolyDis method (after PCA with 50 components).*

fair, we include [Fe/H] as a disentangled parameter during the training of the FaderDis and FactorDis methods.

Reassuringly, we find that for all considered methods, chemically identical stars share more similar latents than random pairs of stars. However not all methods are equally good at this task, with PolyDis underperforming compared to FaderDis and FactorDis. Indeed we find that, unlike the other considered methods, the PolyDis method has a non-negligible overlap between the distributions of chemically identical pairs of stars and of random pairs of stars. This means that for chemical tagging purposes, there will be a larger fraction of random stars in the dataset falsely appearing more chemically similar than genuinely chemically identical stars.

4.5.2 Quantifying Chemical Tagging Performance

In this section, we evaluate the quality of our learnt representations directly on the task of chemical tagging. Since our dataset was designed such that every star has a unique chemical abundance twin, we can evaluate chemical tagging methods based on their capability at recovering these introduced chemical abundance twins. We once again use the euclidean distance in latent space $\mathrm{d} = \|z_i - z_j\|_2$ as our measure of chemical similarity between stars.

We show the results of our analysis in Figure 4.3, where we have plotted the distribution of 'false' chemical abundance twins recovered with each method - considered as stars appearing more similar than the genuine chemical abundance twin. We term these our 'doppelganger' stars. In our plot, the y-axis corresponds to the percentage of stars in the test dataset with fewer false twins than the corresponding value on the x-axis. For example when evaluating the FactorDis model that was trained to remove [Fe/H] from a dataset without noise (as shown in panel d), we found that around 85% of stars in the dataset had fewer than 10 out of the 49998 other stars in the dataset being mistakenly measured as more chemically similar than their genuine chemical abundance twin. Similarly, the y-intercept represents the percentage of stars for which none of the 49998 other stars in the dataset are more similar than the genuine chemical twin.

The Figure suggests that precision disentanglement and removal of non-chemical factors of variation from stellar spectra is valuable for chemical tagging pursuits. The FaderDis method identifies significantly more pairs of chemically identical stars than the baseline PolyDis method. For example the FaderDis method, applied on a noiseless dataset with [Fe/H] removed from the representation (panel c), identifies around 97% of pairs of chemical abundance twins compared to only 50% for the baseline PolyDis method (panel e). For spectra with $SNR = 100$, this number goes down to about 88%. As the neural network performance is sensitive to hyperparameters, architecture and loss function, any improvement in these areas could further improve results. For example, the FaderDis method was found to perform significantly worse when noise was not added as described in the implementation details.

Note that as we randomly generate our stars from a high dimensional distribution, there is a possibility for random pairs of stars to be chemically similar by chance. However we expect a chance of no more than 10^{-12} of doppelganger pairs given the high dimensionality

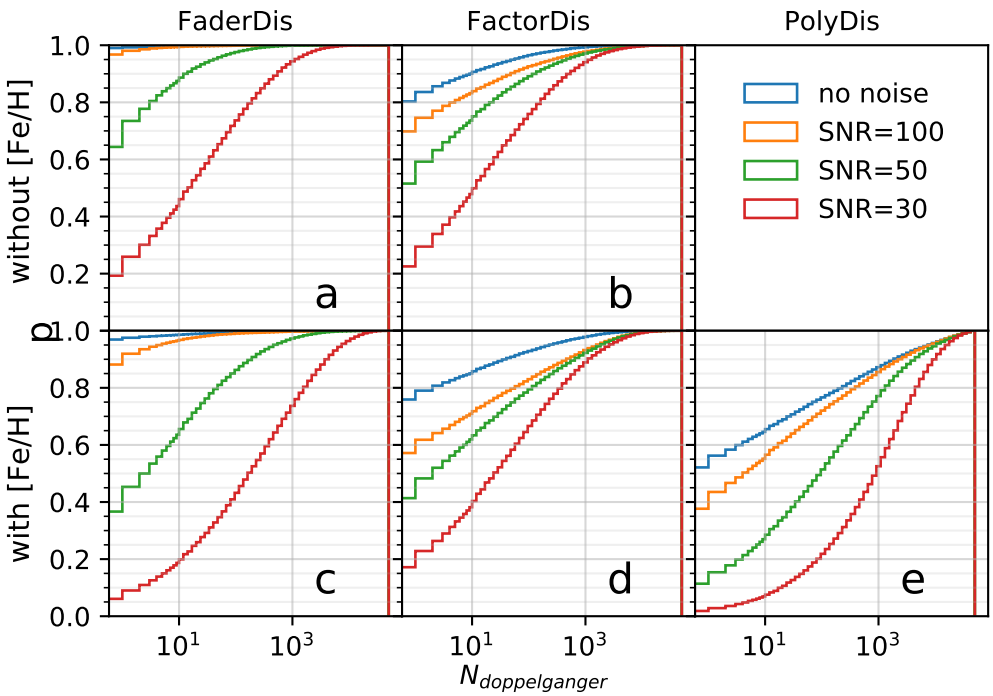

Figure 4.3: *In each panel, we plot the percentage of stars in the test dataset with fewer false twins than x, where x is the x-axis value, denoted as $N_{doppelganger}$ for datasets with varying levels of signal to noise (SNR). In the top row, we show results conditioned on $T_{\rm eff}$ and $\log g$. In the bottom row, we show results conditioned on $T_{\rm eff}$, $\log g$ and [Fe/H]. We plot results obtained for FaderDis in the first column, with FactorDis in the second column and with PolyDis in the third column. It is worth reemphasizing that $N_{doppelganger}$ is highly dependent on the size of the dataset and as such this figure is only intended to be comparative and not as an absolute reference.*

of the artificially generated dataset. Even if these exist however, our Figure is comparative only, to demonstrate how the three different methods work to recover the designated chemical abundance twin stars. As we are generating data from a fixed range of 20 independent abundance labels, recovery of chemical abundance twin stars will become harder under various conditions. This includes as the size of our test set grows within its current abundance ranges, and if correlations between the abundances were included in their prescription. We highlight, however, that this is a comparative test, to demonstrate the relative performance of the three methods, and the performance as a function of signal to noise. The absolute performance would vary in the physical abundance distribution plane of real data.

4.5.3 Interpretability of latent representation

In this section we investigate whether the latent representations that organically emerges from our neural networks are interpretable. As our encoder and decoder are non-linear functions we might pessimistically expect our latent representations to be non-interpretable. We show that this is not the case and instead that, at least on our synthetic dataset, the learned representations align well with the measured abundances.

We approach this question through learning a linear transformation converting from latents to abundances. We represent our dataset of abundances and latents as matrices V and Z of shape $n_{species} \times n_{data}$ and $n_z \times n_{data}$ where $n_{species}$ is the number of chemical species in the spectra, n_z is the dimensionality of the latent space and n_{data} is the number of observations in the dataset. We seek a transformation matrix A converting latents into abundances as faithfully as possible. We can find such a matrix by solving $\mathrm{argmin_A}||\mathrm{AZ} - \mathrm{V}||^2$ which has known solution $A = VZ^+$ (Petersen & Pedersen 2008) with Z^+ the Moore-Penrose inverse of Z. We solve this matrix using all stars stars in the noiseless training data.

In Figure 4.4 we have plotted chemical compositions as estimated from the linearly transformed latents against true chemical compositions. These are shown for 2000 stars in the noiseless test dataset. We see a remarkable agreement between the estimated and true abundances. For almost all species, the linear transformation is nearly as good at estimating chemical compositions as a neural network trained on the latents (denoted "non-linear") on the same stars. Although Na is not as well fit as other species, it is known to be particularly difficult to estimate (in the APOGEE DR14 pipeline is estimated from two weak and possibly blended lines (Ness et al. 2019; Jönsson et al. 2018)). This shows that our method has naturally learned to decompose spectra into a representation nearly equivalent to chemical abundances. Although these results were obtained on a synthetic dataset they are particularly encouraging. Measuring abundance variation quantitatively, without reliance on synthetic spectra would allow for fully circumventing the uncertainties propagated from inaccuracies in spectral modelling.

4.5.4 Spectral Reconstruction

Our neural network encoder allows for converting spectra into a representation in which predefined non-chemical factors of variation are removed. By subsequently applying the

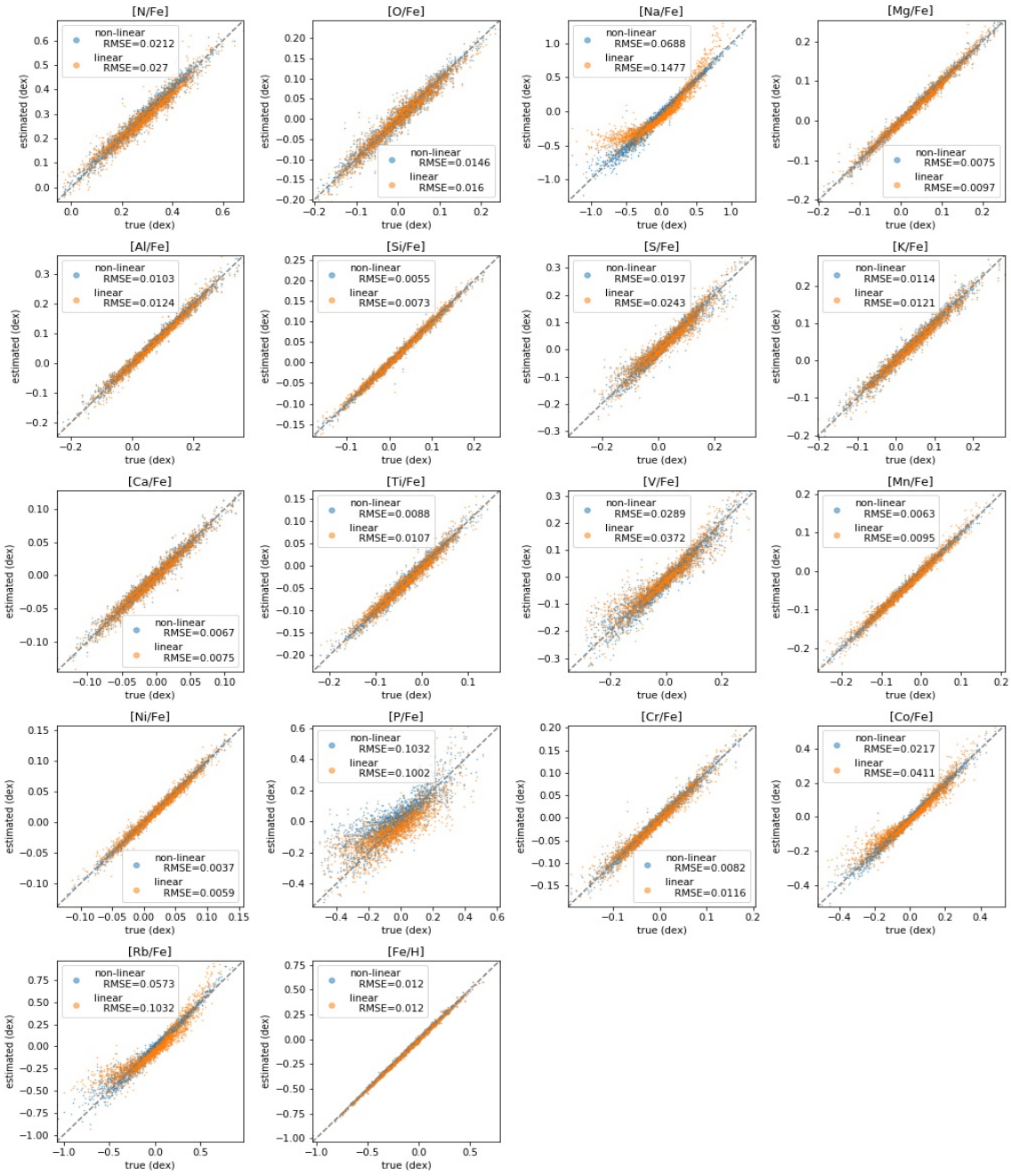

Figure 4.4: *Scatter plot showing estimated against true chemical enhancements and metallicities for synthetic stars in our test dataset. In the legend, linear refers to abundances estimated by multiplying the latent with matrix A, and non-linear to abundances estimated from the latent using a neural network. This figure was obtained using the latent from a FaderDis model trained at disentangling [*$\mathrm{T_{eff}}$*,* $\log \mathrm{g}$*]. For each chemical element, we have also estimated the root-mean-square error (RMSE), the standard deviation of the residuals between predicted and true enhancements/metallicity.*

decoder to this representation, we can generate modified spectra recast to new non-chemical parameters.

In Figure 4.5, we leverage this to visually demonstrate, for the FactorDis approach, how well our learned representation isolates the chemical information in the spectra of a pair of metal-rich stars (top plot, stars x_1 and x_2) and a pair of more metal-poor star (bottom plot, stars x_3 and x_4). These test spectra have been generated as described in Section 4.4.1, with each pair sharing identical chemical compositions but differing physical parameters.

For each sub-figure, in the top panel we plot the original pair of stellar spectra. In the middle panel, we plot how these same chemical abundance twins appear after x_1 is transformed to the physical parameter of x_2, and in the bottom panel we plot the residuals between the twins after the transformation. From these Figures, we see that although the initial spectra are very different, the transformed spectra are near identical. This is because the encoder isolated the chemical information and the decoder generated the recast spectra (for star x_1), at the new provided physical parameters (of the star x_2).

In Figure 4.6, we show the residuals between a star and its transformed twin for FactorDis, FaderDis and PolyDis. For this comparison, we include the three factors of non-chemical variation, $T_{\rm eff}$, $\log g$ and [Fe/H], in the disentanglement network training. Alongside the Figure, we also report the mean absolute residual across the full spectral region considered as well as the standard deviation (per pixel) of the residuals, σ_R for each approach. For the PolyDis method, a star is recast to its chemical abundance twin star's stellar parameters, by replacing its residuals from the polynomial fit with those of its twin star (the fit is meant to isolate the chemical information into the residuals).

In Table 4.3, we report the average mean absolute residual $\langle \mathrm{R} \rangle$ and average mean squared error $\langle \mathrm{MSE} \rangle$, obtained by averaging over random pairs of chemically identical stars in the dataset, transformed to each others physical parameters. The $\langle \mathrm{MSE} \rangle$ metric more severely penalises large deviations in the reconstructed compared to original spectra. Several interesting trends appear in the data.

We observe a difference in performance between methods, depending on whether the residuals or the squared residuals are used for evaluation. Most notably, the FactorDis method outperforms the PolyDis in terms of squared residuals but not raw residuals. As squared values are more sensible to outliers, this seems suggestive that the PolyDis method has comparative better overall reconstruction but struggles with representing some

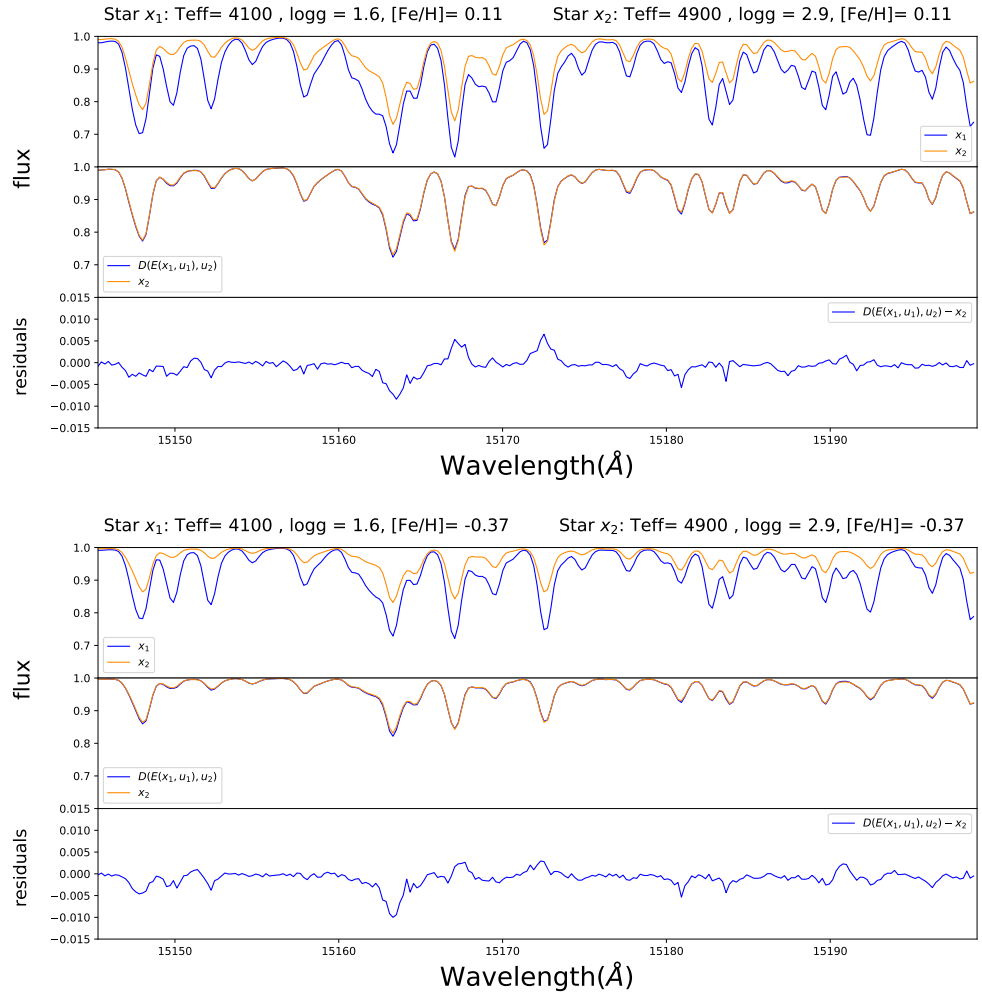

Figure 4.5: *For each subfigure, in the top panel, we show the spectra of two stellar chemical abundance twins (with differing $T_{\rm eff}$ and $\log g$), x_1 and x_2. In the middle panel, the spectra of the second chemical abundance twin, x_2, is shown with a spectra reconstructed by the decoder ($D(E(x_1, u_1), u_2)$) using the other star's latent z_1 but the same physical parameters u_2. In the bottom panel, the corresponding residuals. The stellar parameters are shown above each subfigure (For conciseness The [X/Fe] vector is not shown). We can see that the spectra of chemical abundance twins are nearly indistinguishable after transforming them to a common physical parameter ($T_{\rm eff}$ and $\log g$) parameterization.*

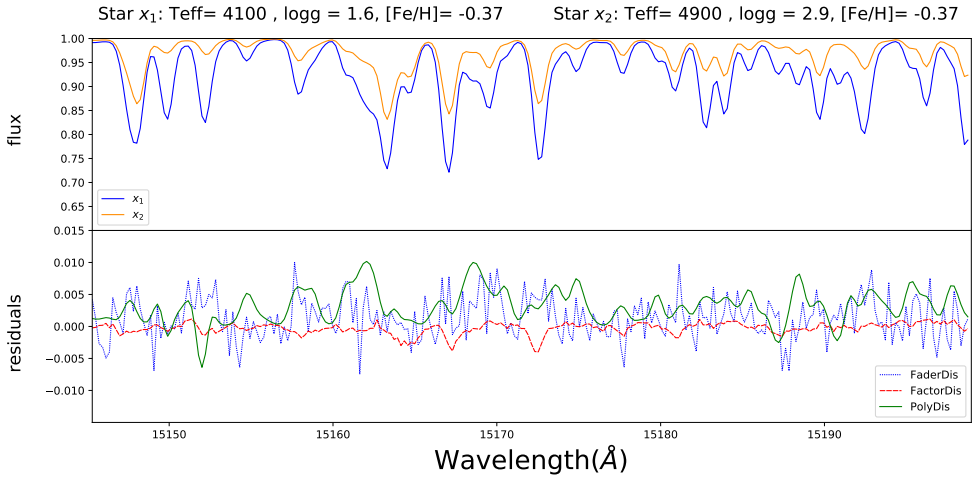

Figure 4.6: *This Figure compares the reconstruction capacities of the three disentanglement methods for the metal-rich star shown in Figure 2. In the top panel, the spectra of two chemical abundance twins, x_1 and x_2, for the first 256 wavelength bins. In the bottom panel, the residuals between the second twin, x_2, alongside the spectra of the first twin x_1, recast by the decoder ($D(E(x_1, u_1), u_2)$) to the physical parameters u_2 for the three disentanglement methods considered. The mean residuals and associated standard deviation (per pixel across the full spectral range) are* $\mathrm{R} = 0.0029$ *and* $\sigma_R = 0.0021$ *for FaderDis,* $\mathrm{R} = 0.0011$ *and* $\sigma_R = 0.0009$ *for factorDis and* $\mathrm{R} = 0.0034$ *and* $\sigma_R = 0.0023$ *for polyDis.*

Method	⟨R⟩	⟨MSE⟩
FactorDis	0.0021	1.26×10^{-5}
FaderDis	0.0030	1.68×10^{-5}
PolyDis	0.0018	1.50×10^{-5}

Table 4.3: *Average MSE between two chemically identical stars transformed to each others physical parameters for the different methods. The quoted number assumes a dataset of stars distributed following the procedure as described in Section 4.4.1.*

portions of the dataset.

The relative mean values reported in Table 3 are also largely indicative of the distribution in parameters of our library of test spectra that we have generated. Simpler modeling approaches likely perform very well when both the training and test data is smaller in overall variability and for pairs of stars that have nearer $T_{\rm eff}$ and $\log g$ parameters. In this case, the spectral variability due to the parameter and abundance labels is nearer to a linear or low order polynomial form (e.g. Ness et al. 2015; Casey et al. 2016). In the regime that the stars considered cover a wide range in $T_{\rm eff}$ or $\log g$ and pairs of stars have much larger differences in these parameters, the move to more complex models, (or as an alternative, local linear models that build non-parametric models using nearest neighbours (e.g. Wheeler et al. 2021)), may have higher return. These differences can

also be understood in terms of the differences between methods at reconstructing spectra. In the PolyDis method, spectra are recast to new physical parameters, by adding residuals to the polynomial fit. This transformation is not parametric in the traditional sense, and so, if two stars being compared are similar to begin with, will give a very small reconstruction loss. For the case of pairs of identical spectra, this would give a perfect reconstruction, even if the residuals do not capture the chemical information. On the other hand, the FactorDis and FaderDis method involves decoding from a lower dimensional representation, and so, even for identical stars, have non-zero residuals. The difference thus boils down to FactorDis and FaderDis being, by design, built for capturing chemical information, but not always (for our exercise) at reconstructing the stellar spectra while the PolyDis method can "cheat" at reconstructing stars. The FaderDis method does not perform particularly well at this task. We believe this to be linked to its training procedure which involves reconstructing noisy rather than clean data. To demonstrate that our

Method	$\langle$R$\rangle$	$\langle$MSE$\rangle$
FactorDis	0.0027	2.13×10^{-5}
FaderDis	0.0033	2.14×10^{-5}
PolyDis	0.0029	3.33×10^{-5}

Table 4.4: *Average reconstruction between two chemically identical stars transformed to each others physical parameters for the different methods on a* ***restricted*** *dataset composed of stellar chemical abundance twin pairs with at least 500K of temperature difference.*

.

method performs better on chemically disimilar stars, we have recalculated our metrics on a dataset restricted to stars with high temperature differences (see Table 4.4). On this partial dataset, the FactorDis method performs better across the board, and the PolyDis method performs worse than both FaderDis and FactorDis, in terms of $\langle$MSE$\rangle$.

4.6 Discussion

Typically chemical abundances are measured using synthetic stellar spectra. Such approaches rely on imperfect stellar models and do not take advantage of the full spectral range. To remedy these shortcomings, there is value in working as closely to the observed data space as possible and developing approaches that circumvent limitations in our current knowledge of stellar physics and incomplete models. Towards this goal, we have developed a neural network architecture that allows for isolating factors of variation of interest.

We compared two deep learning approaches and a simpler polynomial model approximation for the task of removing $T_{\rm eff}$, $\log g$, and also [Fe/H] from model mock stellar spectra – leaving behind the intrinsic variation caused by chemical abundances. All three approaches perform well at generating a disentangled representation of spectra. A reader might note that the mean residuals of the fits, for all three methods we investigate, are lower than a typical Poisson noise level in an observed spectra in large surveys (often SNR $\approx>$ 50-100 per wavelength). The value of this level of precision, and importance of maximising the precision, comes when working with large stellar ensembles. Galactic archaeology typically demands large numbers of stars where the sampling precision of the population increases as $(\mathrm{N})^{1/2}$. The sampling error of the population itself at each wavelength, becomes smaller than the reconstruction precision, for hundred thousand star surveys.

We demonstrated the benefits of using a disentangled neural network, using a chemical abundance twin star recovery, from our 50,000 star test set. This is related to the pursuit of chemical tagging, which is an extremely challenging aspiration with galactic archaeology—to find stars born together using their identical abundances. Even if chemical tagging is prohibited by field contamination, there is still tremendous promise in reconstructing galactic history by modeling the distribution of the most chemically similar stars as a function of orbital properties or abundance density (e.g. Kamdar et al. 2019; Coronado et al. 2020; Price-Jones et al. 2020). When applying our FaderDis approach to a synthetic dataset of APOGEE-like stars, we were able to identify the most chemically identical stars from the ensemble, even with moderate signal to noise. At an SNR of 100 we were able to identify around 87% of pairs of stars. This compared to around 60% of stars for our second neural network approach, FactorDis, and around 40% for our implementation of a polynomial representation to subtract stellar parameters, PolyDis, at this SNR. It was found to be crucial, for getting good results, that high-resolution spectra be available as seen from the drop in performance from 87% to less than 40% when the SNR was changed from 100 to 50.

Our results obtained using synthetic, model spectra are particularly promising. However, these may not translate directly to real survey data, for a number of reasons. As our dataset was handcrafted, we were able to ensure it perfectly satisfied the stringent requirements of our method. That is to say, we ensured perfect knowledge of all non-chemical factors of variation, and expressed these using a determinisitic parametrization,

that is statistically independent from chemical factors of variation. We examine here if these assumptions are accurate for actual stellar surveys, and if not, how we might be able to modify our method to accommodate these discrepancies.

4.6.1 Assumptions about stellar spectra

Our method involves removing all non-chemical factors of variation. If our neural network is conditioned on an incomplete set of non-chemical factors of variation u, our latent will be contaminated by these when isolating chemical factors of variation. For actual observations, these nuisance parameters may arise from imperfect calibration such as from telluric lines or persistance in the detector or any other of a number of systematics. In principle, we may be able to somewhat counteract this phenomenon by restricting the dimensionality of our latent. This would force the latent to only encode the most important factors of variation. However, as some abundances only have a minuscule impact on the overall recorded spectral flux, we require very good knowledge of our factors of variation. An alternative approach for accounting for these systematics would be to add a disentanglement term targeting them. However, this would require additional infrastructure not built into this first demonstration of the approach.

In our proof-of-concept experiments, we first modelled stars using only the effective temperature $T_{\rm eff}$ and surface gravity $\log g$ as non-chemical factors of variation. These two parameters should explain most of the non-chemical variance in the spectral data. Indeed, many data-driven models have been capable of accurately reconstructing spectra using these parameters, plus overall metallicity [Fe/H], (i.e. Ness et al. 2015; Ting et al. 2019; Leung & Bovy 2018), as these are responsible for the majority of spectral variability. However, other parameters, that are independent of the chemical composition, may also impact the observed spectra, and so may need to be included in our conditioning parameters u. Stellar mass, or age, for example, while correlated with effective temperature and surface gravity (Price-Jones & Bovy 2017a), contains additional independent predictive power for generating the spectra (Ness et al. 2016). If this is indeed the case, it may be beneficial to include an independent estimate of stellar mass. This could be achieved by using a training dataset of stars from astroseismology surveys, with mass estimates. Stellar rotation may also affect stellar spectra. These variations may, at least in part, be captured by the micro and macro turbulence parameters. Distance related variations in the spectra must also be preliminarily removed. This requires that spectra be extinction

and radial-velocity corrected before application of the algorithm.

Beyond assuming knowledge of non-chemical factors of variation, we have also so far assumed the ability to perfectly estimate these. In realistic scenarios, this may not be easy. However, similarly to other data-driven methods such as Ness et al. (2016), our method requires precise but not necessarily accurate parameter values. For example our neural network method should still be effective if a change of variable is applied to any of the conditioning variables. Furthermore, we do not account for the correlations between elements when we generate our test data, which will reduce the dimensionality of the spectra and effective sparsity of the data space.

Finally, even if we are unable to fully remove non-chemical factors of variation from spectra, our neural network architecture may still be useful for traditional chemical abundance estimation. Indeed, we may be able to reduce systematic uncertainties in traditional abundance estimation methods by recasting stars to a common temperature and surface gravity (as shown in Section 4.5.4) before comparing to synthetic stellar spectra. Similarly to differential analysis, this would serve to restrict the number of factors of variation changing in stellar spectra.

4.6.2 Assumptions relating to statistical independence

Our approach assumes that abundances are statistically independent from other factors of variation. In our experiments, the synthetic spectral dataset was purposefully constructed to satisfy this assumption. However, this assumption is most likely reasonable for real observations. Indeed, there is evidence that most stellar abundances should, at least to first order, be independent from temperatures and surface gravity (Jofré et al. 2019). In fact, trends between abundances and physical parameters have, in the past, been attributed to systematic uncertainties and sometimes even been corrected for (Valenti & Fischer 2005; Adibekyan et al. 2012). However, overall metallicity does, at some level, affect stellar evolution (e.g. see Gaia Collaboration et al. 2018b) and so ultimately this assumption breaks down to some degree and there is some level of statistical dependency between metallicity and physical parameters in spectra. Including overall metallicity in the disentangled parameters, as done in our experiments, mitigates this issue. Furthermore, spectral synthesis approaches, including at low resolution, typically derive a basic set of $T_{\rm eff}$, $\log g$ and [M/H] (or [Fe/H]) parameters (with errors), and so all parameters, including metallicity, are readily available to use in the disentanglement architecture we

have built. An extension of the approach, would then be to identify chemically identical stars by finding those stars not only sharing a common latent representation but also a common metallicity.

Ultimately, it is the birth abundances of stars rather than surface abundances that are useful for chemical tagging. Whilst stellar processes like dredge-up and diffusion (Masseron & Gilmore 2015; Martig et al. 2016) modify surface abundances away from their birth values across evolutionary state, encouraging statistical independence, as is done by our technique, will act to remove these processes and recover a chemical representation closer to the birth abudances, which is ultimately preferable for using abundances for chemical tagging pursuits.

4.6.3 Beyond synthetic spectra

There are a few challenges associated with applying our method to real observations. Spectral bins in real observations are sometimes flagged as untrustworthy, for example due to cosmic-rays or persistence in the detector (see Jönsson et al. (2020)). These are flagged for each APOGEE spectra and individual pixels are correspondingly masked. Since our neural network methodology requires all spectral bins as inputs, such untrustworthy or missing data need to be imputed somehow, so as to not impact the downstream learned representation. Another more practical challenge with applying the method to real observations is that the method requires hyperparameter tuning and training the algorithm takes a day to run on specialized hardware (GPUs). This makes iterative deployment slow.

4.7 Conclusion

Organising stars by their chemical similarity and investigating the distribution of their other properties (e.g. orbits, density) is a promising avenue for unravelling galactic evolution (i.e. Coronado et al. 2020; Kamdar et al. 2019; Ting et al. 2016). Ranking stars by chemical similarity requires precise chemical information for a large numbers of stars. Chemical similarity is typically determined using measured element abundances. However, these measurements are subject to inaccuracies and systematics inherited from the incomplete and approximate stellar models currently in use. As an alternative to deriving abundances from spectra, using the variability of the spectra itself, becomes possible and

advantageous in the regime of large stellar surveys. Data-driven deep learning methods - applied directly to the spectra itself - find natural applications here.

In this thesis chapter, we have introduced a new deep-learning method for extracting chemical information from spectra that relies on isolating chemical factors of variation from non-chemical factors of variation, through training a neural network using a disentanglement loss. This method removes the need for accurate and precise modelling of the chemical abundance factors of variation in stellar spectra. Instead, it relies only on the parameterization of the other primary sources of variability, namely stellar parameters, including $T_{\rm eff}$ and $\log g$. This requires conditioning a neural network on these factors, in our case, $T_{\rm eff}$ and $\log g$, (and also [Fe/H] for modeling only the variation from abundance enhancements, [X/Fe]).

We validated the approach on a synthetic dataset of spectra and found it capable of accurately distinguishing chemically identical pairs of stars from a field distribution. We were able to identify more than 85% of pairs of chemical abundance twins from a dataset of 50000 synthetic spectra, generated at the resolution of APOGEE and SNR of 100, with 20 independently drawn chemical abundances ([X/Fe]). Furthermore, on these experiments, the method compared favourably to our baselines.

This thesis chapter and associated experiments act as a proof-of-concept in a controlled environment where the data generating process is perfectly understood. This sets out the groundwork for applying such representation learning methods to real observations. The natural next steps of this line of work will be translating the success found on synthetic spectra to real APOGEE observations. Because of the imprint of systematics on real stellar spectra and other possible departures from our assumptions, this will likely require further modifications and/or fine-tuning of the approach.

Beyond our astronomical contributions, we hope that our proposed methodology will find uses in other fields. Our application of supervised disentanglement for identifying observations sharing a common parameterization is a novel method that could be adapted to other tasks. In particular, our chemical tagging experiments and associated datasets could be useful in comparing different supervised disentanglement architectures, something that has so far been lacking in the machine learning community. We believe that our task of evaluating how well a supervised disentanglement neural network maps chemically identical stars to an identical latent is particularly useful for assessing supervised disentanglement. Additionally, our proposed FactorDis supervised disentanglement archi-

tecture has shown good performance at chemical tagging-like pursuits and disentanglement which suggests that it may be a competitive alternative to FaderDis types of architectures.

Chapter 5

Measuring chemical likeness of stars with RSCA

The work presented in this Chapter is based on the preprint de Mijolla & Ness (2021) (submitted), in collaboration with Melissa Ness.

5.1 Introduction

As was discussed in the previous chapter of this thesis, the field of Galactic astronomy has entered a transformative era. Large-scale surveys, such as APOGEE, *Gaia* and GALAH, are providing millions of high-quality spectroscopic and astrometric measurements of stars across the Milky Way (Majewski et al. 2017; De Silva et al. 2015; Gaia Collaboration et al. 2018a). Future large-scale surveys, which will release even more high-quality data, are on the horizon (de Jong et al. 2016; Kollmeier et al. 2017; Bonifacio et al. 2016).

In this landscape of high-volume high-quality stellar astronomy, fully extracting the scientifically relevant information from stellar spectra remains a difficult problem. Classically, this has been done by comparing observations to synthetic spectra generated from theoretical models (e.g. García Pérez et al. 2016). However, the precision with which stellar labels can be derived under such an approach is ultimately limited by the faithfulness with which synthetic spectra reproduce observations. Because of computational constraints and gaps in knowledge, synthetic spectra do not perfectly match observations,

something sometimes referred to as the "synthetic gap" (O'Briain et al. 2021). Computational models used to generate synthetic spectra use incomplete stellar line lists and usually must make simplifying assumptions. This includes for example that stellar atmospheres are one-dimensional, in hydrostatic equilibrium, and in local thermodynamic equilibrium. In addition, even beyond these issues, observations are affected by further systematics such as telluric lines introduced by the earth's atmosphere (e.g. Holtzman et al. 2015) and telescope imperfections/aberrations.

Ultimately this synthetic gap limits our ability to extract information from stellar spectra. In Ting et al. (2017); Ting & Weinberg (2021) it was shown that stellar spectra contain more chemical information than is captured in bulk metallicity and α-enhancement alone. The precision of derived individual stellar abundances from large surveys, however, may be limited by an inability to fully extract information given approximate models, rather than by the signal to noise of observations.

This is problematic because a lot of interesting science requires measuring chemical similarity between stars with a precision beyond that currently delivered by large stellar survey's modelling pipelines. In particular, high precision chemical measurements are needed for strong chemical tagging (Freeman & Bland-Hawthorn 2002). This is an ambitious Galactic archaeology endeavour aiming to identify stellar siblings - stars born from the same molecular cloud - using chemical information derived from spectroscopy long after clusters gravitationally dissipate. In practice, whether such chemical tagging is theoretically possible at scale is still an open question, but may be answered with large-scale surveys like GALAH (De Silva et al. 2015; Buder et al. 2021). For this form of chemical tagging to be successful, stellar siblings must share a near-identical chemical composition with sufficient variability in chemical compositions between clusters. Even if strong chemical tagging reveals itself to be impossible at large scale, precise chemical similarity measurements would still be useful in reconstructing the broad nature of our galaxy's evolution (e.g. Coronado et al. 2020; Kamdar et al. 2020).

These issues motivate the development of methods capable of extracting information from stellar spectra. and overcoming the synthetic gap between observations and theoretical models. Several data-driven methods have been developed for this purpose. Methods such as those proposed by Ness et al. (2015); Casey et al. (2016); Leung & Bovy (2018); Ting et al. (2019); O'Briain et al. (2021); Das & Sanders (2019) allow for improving precision of stellar labels through leveraging data-driven interpolators between stellar spectra

and labels, reducing the impact of noise and systematics on derived parameters. However, as such approaches still rely on synthetic spectra they do not fully alleviate issues with systematic errors from mismatching theoretical spectra. Recently, methods for finding chemically similar stars directly from stellar spectra without reliance on synthetic spectra have been developed Bovy (2016b); Price-Jones & Bovy (2017b); Cheng et al. (2021); de Mijolla et al. (2021). This category of method works by removing the effect of non-chemical parameters on stellar spectra, thus isolating the chemical information within the spectra. Such approach are not without drawbacks. Although they remove the dependency on synthetic models, they still require a comprehensive and precise determination of all non-chemical factors of variation. Additionally, they must make simplifying assumptions regarding the cross-dependencies between chemical and non-chemical factors of variations which may have an impact on accuracy.

In this thesis chapter, we present a new approach for identifying chemically similar stars from spectroscopic data which we name "Relevant Scaled Component Analysis" (RSCA) because of its similarities with Relevant Component Analysis (Shental et al. 2002). Our approach is grounded in the machine learning subfield of metric learning. Instead of estimating individual chemical abundances, we project spectra directly into a lower dimensional subspace in which distances between spectra are made to encode a useful notion of chemical similarity between stars. Crucially, as our approach for transforming stellar spectra does not rely at any stage on synthetic spectra or quantities derived from these, its performance is not hindered by inaccuracies in stellar modelling.

A novelty of our work is that instead of using synthetic spectra to learn this notion of chemical similarity, we make use of spectra from known open clusters with open cluster membership information. Open clusters are groups of stars born together that remain gravitationally bound after birth and up to the present day. They are relatively rare, as most stellar clusters dissipate rapidly after birth (Portegies Zwart et al. 2010). However, they are extremely useful tools in modern Galactic astronomy. In particular, open clusters, which can be identified using astrometry, display near-identical chemical abundances although small scatter may exist at the 0.01 to 0.02 dex level and up to <0.05 dex level for some elements (e.g. Bovy 2016b; Liu et al. 2019; Ness et al. 2018; Cheng et al. 2021). Open clusters have found many uses in modern astronomy, for example to obtain high-precision measurements of the radial abundance gradients in the Milky Way (Friel 1995; Magrini, L. et al. 2017) or to benchmark and calibrate stellar survey abundance measurements (García

Pérez et al. 2016). Here, we use open clusters as a gold standard for learning a notion of chemical similarity. In our approach, we take the viewpoint that if open clusters are indeed chemically homogeneous, then a successful metric for encoding chemical similarity will be one in which open cluster stellar siblings are highly clustered.

RSCA bears some similarities with the disentanglement approach introduced in the previous chapter. Both methods identify chemically similar stars without reliance on synthetic spectra through a form of representation learning in which chemical factors of variation are identified. However, unlike the method presented in the previous chapter which needed all non-chemical factors of variation to be manually specified, RSCA can automatically identify chemical factors of variation using open-clusters. This makes RSCA more practical, but this advantage does however comes at a cost. Open-cluster datasets contain only a few datapoints and so the method was purposely made linear to avoid overfitting.

Our algorithm, RSCA, has several properties that make it suitable for the task at hand, of measuring chemical similarity between stars:

- It is fully *data-driven*. Chemical similarity is measured without any reliance or dependency on theoretical models. This offers a measure of chemical similarity that is independent from the systematics introduced in traditional stellar modelling (e.g. Jofré et al. 2017), and offers a means of validating existing discoveries.

- It is *computationally efficient*. As the method is linear, processing spectra from the full APOGEE stellar survey can be done in minutes. The most computationally intensive step of the approach is a Principal Component Analysis decomposition.

- It is *interpretable*. In its current formulation, measuring chemical similarity using our method amounts to evaluating Euclidean distances between stars projected on a hyperplane of the stellar spectra space.

- It is *precise*. We find the method, using spectra, to be more effective at identifying stellar siblings from open clusters than is possible using stellar abundance measurements. We believe this to be in large part because our method bypasses the synthetic gap introduced by spectral modelling. Furthermore, our experiments suggest that the performance could be further improved, for example, with a larger dataset of open cluster stars or by taking into account the error on the flux which we do not currently do.

The chapter is organized as follows. In Section 5.2.1, we outline the conceptual ideas behind our approach for measuring chemical similarity. We then briefly introduce Principal Component Analysis in Section 5.2.2, which is a core component of our algorithm. In Section 5.3, we dive deeper, and present our algorithm, RSCA. This is implemented using open clusters observed by the APOGEE survey in Sections 5.4, and evaluated in light of the field distribution of stars. Its trade-offs and implications are discussed in Section 5.5.

5.2 Concepts and Assumptions

5.2.1 Chemical similarity as metric learning

The variability amongst stellar spectra are caused by the interplay of many factors of variation. These include chemical and physical parameters of the star and the instrumental systematics associated with the telescope, as well as interstellar dust along the line of sight. Measuring chemical similarities requires disentangling the imprint left on the spectra by chemical factors of variation from that left by the other non-chemical factors of variation. Our goal is to identify chemically similar stars from their spectra, for stars that span a range of physical stellar parameters (i.e. effective temperatures and surface gravities). We approach this task from a data-driven perspective, and build an algorithm for identifying stars that are as chemically similar as birth siblings, using open cluster spectra.

For our method, we assume that open clusters are close to chemically homogeneous because of their common birth origin (Ness et al. 2018) but are not special in any other way (at least in terms of their spectra). That is to say, we assume that the only information within spectra useful for recognizing open clusters are the chemical features of the spectra, and so that a model which identifies open clusters from the spectra will need to do so by extracting the chemical information within spectra

We frame the task of building such a model recognizing open clusters as a metric learning task. That is to say, we build a data-driven model converting stellar spectra into a representation in which Euclidean distances convey the uncalibrated probability of stars originating from a shared open cluster. To accomplish this, the training objective of our data-driven algorithm can be understood as transforming stellar spectra into a representation in which the distance between intra-cluster stars is minimized and the distance between inter-cluster stars is maximized.

Distances in the representation resulting from such an optimizing procedure will or-

ganically quantify the chemical similarity of stars. Non-chemical factors of variation, such as stellar temperatures and instrumental systematics, will not contribute to the representation as their presence would make distances between stellar siblings larger. Instead, such a representation will only contain those factors of variation of spectra that are discriminative of open clusters, ie the chemical factors of variation. Crucially, chemical factors of variation will contribute to distances in the representation in proportion to how precisely they can be estimated from stellar spectra. Stronger chemical features will be more strongly amplified than weaker chemical features.

The utility of this data-driven approach is that it is independent of imperfect model atmosphere approximations and other issues affecting synthetic spectra. This provides a high fidelity technique to turn to specific applications within Galactic archaeology, such as chemical tagging of stars that are most chemically similar (Freeman & Bland-Hawthorn 2002).

The assumption underpinning our work is that this chemical information will be the only information within stellar spectra useful for distinguishing open clusters and so will be the only information captured by our model. If this assumption is true, then the representation induced by the model will measure a form of chemical similarity between stellar spectra.

However, since open clusters, in addition to sharing a common age and near-identical birth abundances, are also gravitationally bound, they can be identified from their spatial proximity if such information is available in the spectra. As such spatial information does not robustly transfer towards identifying dissolved clusters, it mustn't be captured by our model. In this work, we apply our algorithm to pseudo-continuum normalized spectra with diffuse interstellar bands masked, which we assume not to contain any information about spatial location so as for our representation after training to only contain chemical information. Assuming pseudo-continuum normalized spectra do not encode any spatial information is plausible, since after continuum normalization, the spectrum should not contain significant information about stellar distance. With the impact of reddening removed and diffuse interstellar bands masked, a spectrum should also not contain any information about the stellar extinction and interstellar medium along the line of sight of the star. We examine the validity of these assumptions in latter sections. Ultimately, it is worth emphasizing that our method only exploits features proportionally to their discriminativeness at recognizing open clusters. Therefore, we can expect our model to not

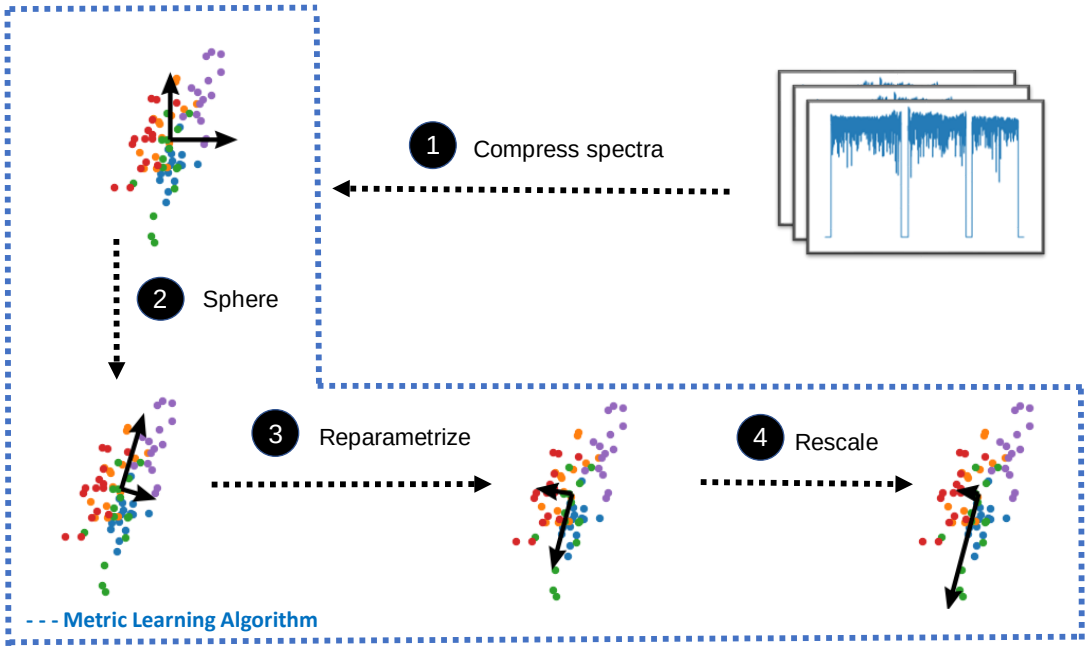

Figure 5.1: *Schematic depiction of RSCA. The algorithm proceeds by first encoding stellar spectra into a lower dimensional representation made two-dimensional for illustrative purposes. In this representation, stellar siblings - which are represented by same-coloured dots - are not initially identifiable by their Euclidean distance in the basis (represented by black arrows). The objective of the metric-learning algorithm (dashed blue) is to find a new basis in which distances are informative about which stars are stellar siblings. This objective is realized through three linear steps: a sphering transformation on the dataset, a reparametrization to a suitable basis, and a scaling of the basis vectors.*

heavily rely on non-robust features (provided that these are significantly less informative than robust features). This also relies on our open cluster training data being representative of the parameter space that should be marginalized out; i.e. our model does not learn to associate inter cluster stars via $\log g$ and $T_{\rm eff}$, which could happen if the evolutionary state of observed cluster stars was similar within clusters and different between clusters.

5.2.2 Principal Component Analysis

Our metric-learning algorithm, RSCA, first uses principal Component Analysis (PCA) to transform the data into a (lower-dimensional) basis that represents the primary variability of the ensemble of spectra we work with.

As discussed in Section 1.4.4, the principal components of a dataset X, of shape $N_D \times N_F$ containing N_D data points and N_F features (i.e. pixels in the context of this thesis chapter), are an ordered orthogonal basis of the feature space with special properties.

In the principal component basis, basis vectors are ordered by the amount of variance they capture. They have the property that for any k, the hyperplane spanned by the first k-axes of the basis is the k-dimensional hyperplane, which maximally captures the data variance. In PCA, the number of principal components used, k, is a hyperparameter controlling the trade-off between the amount of information preserved in the dataset X after compression and the degree of compression.

The principal component basis corresponds to the unit-norm eigenvectors of the covariance matrix of X ordered by eigenvalue magnitude. This can be obtained through diagonalization of the covariance matrix. The principal component basis can also be formulated as the maximum likelihood solution of a probabilistic latent model which is known as Probabilistic Principal Component Analysis (PPCA) (see Bishop (2006)). This probabilistic formulation is useful in that it enables one to obtain the principal components for a dataset containing missing values by marginalizing over these.

As we will make use of further in this thesis chapter, the principal components also allow for generating a sphering transformation. This is a linear transformation of the dataset to a new representation, in which the covariance matrix of the dataset X is the identity matrix. Sphering using PCA is carried out by performing a change-of-basis to a modified principal component basis in which the principal components are divided by the square root of their associated eigenvalues.

5.3 Relevant Scaled Component Analysis Algorithm

The input to RSCA is individual stellar spectra, some of which belong to open clusters, and some of which are a reference field sample. The output of RSCA is, for each spectrum, a N_k vector, in which dimensions are scaled such that distances between N_k vectors of pairs of stars encode chemical similarity between those stars. We step through this in detail below.

5.3.1 Overview

Let us define $X_{\rm clust}$ as the matrix representation of a dataset containing the spectra of known open cluster stars. Analogously, let us define X_{pop} as the matrix representation of a larger dataset of stellar spectra in the field (with unknown cluster membership). These matrices are respectively of shapes $N_{\rm d_{clust}} \times N_{\rm b}$ and $N_{\rm d_{pop}} \times N_{\rm b}$, where $N_{\rm d_{clust}}$ is the number

of open cluster stars, $N_{\mathrm{d_{pop}}}$ the number of of stars in the large dataset and N_b the number of spectral bins (7514 bins in the case of APOGEE). For our purposes, we assume access to only a limited number of open cluster stars such that $N_{\mathrm{d_{clust}}} << N_{\mathrm{d_{pop}}}$. We also assume that the spectra in these matrices are pseudo-continuum normalized spectra, with diffuse interstellar bands masked in a process following that described in Appendix B.1. Pseudo-continuum spectra are normalized rest-frame spectra in which the effects of interstellar reddening and atmospheric absorption are removed, in a process described in Majewski et al. (2017).

RSCA takes as inputs X_{clust} and X_{pop}. Through a series of linear-transformation to these matrices, RSCA maps these matrices into new matrices of shape $N_{\mathrm{d_{clust}}} \times N_{\mathrm{K}}$ and $N_{\mathrm{d_{pop}}} \times N_{\mathrm{K}}$ whose entries are the stellar spectra transformed to a metric-learning representation of dimensionality K. Euclidean distances in this new metric-learning representation can then be used to measure chemical similarity between spectra. It is worth mentioning that, as all the steps of RSCA are linear transformations, the mapping converting from spectra to the metric learning representation can be parameterized by an $N_{\mathrm{K}} \times N_{\mathrm{bins}}$ matrix and used to convert unseen spectra (or used for interpretability purposes).

We provide in Figure 5.1 a graphical depiction of the linear transformations involved in the RSCA algorithm. RSCA works by first projecting the spectra onto a set of basis vectors with PCA (Step 1). For visualization purpose, this basis is made two-dimensional although it would normally be higher dimensional. After this PCA compression (Step 1), stellar siblings are represented as same-coloured dots whose xy coordinates correspond to coordinates in the PCA basis. Once in the PCA basis a series of linear transformation are applied to the data. For improved clarity, we keep the data fixed throughout our algorithm visualization and represent linear transformations as change-of-basis (black arrows). Step 2 and 3 of the algorithm, find a new basis which more aptly captures spectral variability amongst stellar siblings and Step 4 of the algorithm rescales basis vectors of this basis based on a comparison between their spectral flux variance amongst stellar siblings and amongst field stars. The outcome of the RSCA algorithm is a new representation of the spectra in which dimensions that are unhelpful for discriminating stellar siblings are minimised in amplitude (through a stretching-out of basis vectors). Conversely, dimensions that are helpful in recognizing stars within the same open clusters are made larger (through a squeezing of basis vectors). The K vector of principal components for each star can be collapsed into a measure of chemical similarity through a Euclidean distance measure

between stars of their scaled representation, output from RSCA, where $d = \sqrt{(n_K - n'_K)^2}$ for any pair of stars n, n'.

We now walk step-by-step through the successive linear-transformations involved in the RSCA algorithm. For following along, the pseudo-code for RSCA is provided in Appendix B.4 and the full source code of our project which contains a Python implementation is made available online [1].

5.3.2 Step 1: Compress the spectra with PCA to reduce the risk of overfitting

In the first step of our approach, denoted as "1) Compress Spectra", in Figure 5.1, we apply PCA to X_{pop} to convert the population of stellar spectra into a lower dimensional representation. This dimensionality reduction step serves to make the algorithm more data-efficient which is crucial given the risk of overfitting from the small number of open clusters within our dataset.

As some stellar bins are flagged as untrustworthy, we use Probabilistic PCA (PPCA), a variant of the PCA algorithm which can accommodate missing values. After finding principal components of X_{pop}, we compress the data by discarding all but the K largest principal components, where K is a hyperparameter requiring tuning. Then, datasets X_{pop} and X_{clust} are each transformed using the K basis vectors, which we call Z_{pop} and Z_{clust}; the representation of the spectra in the PCA basis of X_{pop}. These have shapes of $N_{\mathrm{d_{pop}}} \times N_{\mathrm{K}}$ and $N_{\mathrm{d_{clust}}} \times N_{\mathrm{K}}$.

5.3.3 Metric Learning: Sphering, Reparameterization and Rescaling

Step (1) in our procedure of our PCA compression is a pre-processing step. The steps that follow fall into the realm of a general-purpose metric-learning algorithm. These rely on assumptions about the PCA-compressed spectra being satisfied. Performance should be robust to within small departures from these assumptions, but will still ultimately be tied to how well these assumptions are respected. We lay out our assumptions for steps (2)-(4) below.

[1] https://github.com/drd13/RSCA

Assumptions

First, we assume that the data (i.e. spectra) in Z_{pop} are well approximated as being drawn from a multivariate Gaussian distribution. That is to say that if we define μ_{pop} and Σ_{pop} as the mean and covariance of Z_{pop}, then the stars within Z_{pop} can be assumed as being samples drawn from $z_{pop} \sim N(\mu_{pop}, \Sigma_{pop})$.

Next, we make the assumption that individual clusters are themselves approximately Gaussian in the PCA-compressed space. That is to say we posit that the members of open clusters are well approximated as being samples drawn from a distribution $z_{clust} \sim N(\mu_{clust}, \Sigma_{clust})$. Crucially, we make the assumption that all open clusters share the same covariance matrix Σ_{clust} and only differ in their mean μ_{clust}. This is perhaps our strongest and most important assumption. That is, that the stars within different clusters are distributed following a shared covariance matrix (i.e. clusters have the same shape irrespective of their location in the representation). It is this assumption which allows a linear transformation i.e. a transformation that acts the same across the whole representation, to be an effective approach for measuring chemical similarity. This assumption of what is effectively cluster translation invariance, can be interpreted as assuming that the scatter amongst stellar spectra in physical and chemical parameters should be the same for all clusters irrespective of the clusters parameters which is a sensible assumption. Connecting these assumptions back to Figure 5.3, each step 2-4 requires that stars in any population, within clusters and within the field, follow a multivariate Gaussian with an invariant covariance matrix for each individual cluster.

Step 2: Sphering to transform the population covariance matrix into the identity matrix

Together, the sphering and reparametrization steps of RSCA serve to transform Σ_{pop} and Σ_{clust} to a vector representation of stellar spectra in which Σ_{pop} and Σ_{clust}, the covariance matrices of the field stars and clusters, respectively, are diagonal matrices. As there are then no off-diagonal terms in the covariance matrices, this ensures that the variances along basis vectors fully capture the covariance information amongst stellar siblings and amongst field stars.

In Step **"2) Sphere"** of our algorithm, in Figure 5.1, we linearly transform the vector representation of spectra such that the dataset Z_{pop} has a covariance matrix of identity after transformation. This linear transformation takes the form of a sphering transforma-

tion applied to Z_{pop}, where in our experiments we use the PCA-sphering scheme. This has the utility of fully capturing the variability amongst stellar spectra and amongst stellar siblings.

Step 3: Reparameterization to diagonalize the cluster covariance matrix

In steps 1 and 2 of RSCA we do operations on the full field population and full cluster population, respectively. In steps 3 and 4, we transform and scale to recognise stellar sibling likeness compared to the field, so consider stellar variability of the stars within individual clusters.

In the next step of our algorithm after the sphering transform, **"3) Reparametrize"** in Figure 5.1 represents a change of basis. In this step, the covariance matrix Σ_{clust}, is diagonalized. Since we do not have direct access to Σ_{clust}, as an intermediary step in doing so, a dataset approximately distributed according according to $N(0, \Sigma_{clust})$ is created from the open cluster dataset. This is done by subtracting each star's vector representation from the mean representation of all stars belonging to the same cluster $\widehat{\mu}_{clust}$, such that each cluster becomes zero-centered. It is worth noting that as $\widehat{\mu}_{clust}$ is estimated from a limited number of samples, it will not exactly match with the true μ_{clust} and so the resultant population will only approximately be distributed according to $N(0, \Sigma_{clust})$.

As PCA basis vectors correspond to (unit-norm) eigenvectors of covariance matrices, the PCA basis obtained by applying PCA to the zero-centered open cluster dataset parametrizes a transformation to a representation in which Σ_{clust} is (approximately) diagonal. A change-of-basis to this PCA basis thus parametrizes the desired diagonalization of the cluster covariance matrix. Because the basis vectors of the PCA basis have by construction unit-norm, the covariance matrix Σ_{pop} will still be the identity matrix after this change-of-basis and hence both Σ_{clust} and Σ_{pop} will be diagonal matrices as desired.

Step 4: Scaling to maximise discriminative power in identifying chemically similar stars

In the final step of the metric learning algorithm, denoted by **"4) Rescale"** in Figure 5.1, basis vectors are scaled proportionally to their dimension's usefulness at recognizing open clusters. This is done by applying separately to each dimension of the representation a scaling factor. Here, the independent scaling of dimensions is justified by Σ_{clust} and Σ_{pop} being diagonal covariance matrices.

We use the separate individual clusters within X_{clust} and field populations X_{pop}, to measure variance in each dimension to determine our scaling factor. We design our scaling factor such that, after transformation, distances between pairs of random stars quantify the ratio between the probability that pairs of stars originate from the same (open) cluster and the probability that they do not. To put it another way, we seek to scale dimensions such that pairs of stars that are more likely to originate from the same cluster as compared to originating from different clusters have a smaller separation (i.e. Euclidean distance) in the representation than less likely pairs.

Under our set of assumptions, along a dimension i amongst the K dimensions in the PCA-compressed representation, random stellar siblings (intra-cluster stars) are distributed as $z_{clust} \sim N(\mu_{clust_i}, \sigma_{clust_i})$ where μ_{clust_i} and σ_{clust_i} are the mean and standard deviation along dimension i (at this stage in the algorithm). Accordingly, using the standard formula for the sum of normally distributed variables, the one-dimensional distance along i between pairs of random stellar siblings $z_{clust_{1i}}$ and $z_{clust_{2i}}$ follows a half-normal distribution $d_{clust_i} \sim |z_{clust_{1i}} - z_{clust_{2i}}| = |N(0, 2\sigma_{\mathrm{clust_i}})|$. Likewise, the distance d_{pop_i} between pairs of random field stars follows a similar half-normal distribution $d_{pop_i} \sim |N(0, 2\sigma_{\mathrm{pop_i}})|$ where $\sigma_{\mathrm{pop_i}}$ is the standard deviation amongst field stars along dimension i.

For a pair of stars observed a distance d_i away from each other along a dimension i, the ratio between the probability of the pair originating from the same cluster (intra-cluster) and the probability of the pair not originating from the same cluster (inter-cluster) is:

$$r_i(d_i) = \frac{p(d_{clust_i} = d_i)}{p(d_{pop_i} = d_i)} = \left| \frac{N(0, 2\sigma_{\mathrm{clust_i}})}{N(0, 2\sigma_{\mathrm{pop_i}})} \right| \tag{5.1}$$

which evaluates to (as distances d_i are by design greater than 0)

$$r_i(d_i) = A_i e^{\frac{-d_i^2}{2\sigma_{r_i}^2}} \tag{5.2}$$

where

$$A_i = \frac{\sigma_{pop_i}}{\sigma_{clust_i}} \tag{5.3}$$

and

$$\sigma_{r_i} = \frac{\sigma_{\mathrm{clust_i}}\sigma_{\mathrm{pop_i}}}{\sqrt{\sigma^2_{\mathrm{pop_i}} - \sigma^2_{\mathrm{clust_i}}}} \tag{5.4}$$

As dimensions are assumed to be independent, the probability ratio accounting for all dimensions is the product of the probability ratio of the separate dimensions:

$$r = \prod_{i=0}^{K} \frac{N(0, 2\sigma_{\mathrm{clust_i}})}{N(0, 2\sigma_{\mathrm{pop_i}})} = Ce^{-\frac{1}{2}\sum_{i=0}^{D}(\frac{d_i}{\sigma_{r_i}})^2} \tag{5.5}$$

where $C = \prod_{i=0}^{K} A_i$.

From this expression, it can be seen that multiplying dimensions by a scaling factor of $\frac{1}{\sigma_{r_i}}$ leads to a representation in which the probability ratio r is a function of Euclidean distance and where pairs of stars with smaller Euclidean distance have a higher probability of originating from the same open cluster as compared to their probability of originating from different cluster than pairs with larger Euclidean.

It is clear that dividing dimensions by a scaling factor of $\frac{1}{\sigma_{r_i}}$ induces a representation where distances d which are measured between the scaled reconstructed representation directly encode the probability of stars originating from the same cluster, as desired for our metric-learning approach. However, using this expression as a scaling factor requires evaluating the σ_{clust_i}'s and σ_{pop_i}'s along all dimensions. Because the representation has been sphered, the population's standard deviation is unity along all directions ($\sigma_{pop_i} = 1$). We estimate the intra-cluster standard deviations using a pooled variance estimator:

$$\sigma^2_{\mathrm{clust_i}} = \frac{\sum_{j=1}^{k}(n_j - 1)\,\sigma^2_{ji}}{\sum_{j=1}^{k}(n_j - 1)} \tag{5.6}$$

where σ^2_{ji} refers to the sample variance along dimension i for the sample of stars belonging to an open cluster j containing n_j stars in X_{clust}. To make the algorithm more robust to the presence of any outliers in the dataset, such as misclassified stellar siblings, we use the median absolute deviation (MAD) as an estimator for the sample standard

deviation σ_{ji}. That is

$$\text{MAD} = \text{median}\left(\left|X_i - \tilde{X}\right|\right) \tag{5.7}$$

where X_i and $\tilde{X}$ are the data values and median along a dimension, respectively.

To better understand the effect of our scaling factor on the representation it is applied to, it is instructive to look into the impact it has on the distances between stars along dimensions of a representation. Dimensions along which open clusters have a similar standard deviation to the full population of stars (i.e. $\sigma_{clust} \approx \sigma_{pop}$) carry little useful information for recognizing open-clusters. Accordingly, after application of RSCA, such dimensions will have $\sigma_{r_i} \to \infty$ and so be strongly suppressed. On the other hand, for dimensions where the population's standard deviation σ_{pop} is significantly larger than the cluster standard deviation σ_{clust}, the population's standard deviation is no longer relevant and $\sigma_r \approx \sigma_{clust}$. That is to say the scaling devolves into measuring distances relative to the number of standard deviations away from the cluster standard deviation.

5.4 Experiments on APOGEE Data

We validate our approach for encoding chemical similarity by testing its performance on real data obtained by the APOGEE survey data release 16 (Ahumada et al. 2020). The APOGEE survey (Majewski et al. 2017) is an infrared, high resolution (R∼22500), high signal-to-noise spectroscopic survey. The APOGEE survey uses a 300-fiber spectrograph (Wilson et al. 2019) installed at the Sloan Digital Sky Survey telescope located at Apache Point Observatory (Gunn et al. 2006) and a similar spectrograph on the DuPont telescope at Las Campanas Observatory (Bowen & Vaughan 1973).

5.4.1 Dataset Preparation

For our experiments we use spectra from the public APOGEE data release DR16 (Ahumada et al. 2020) to create X_{clust} and X_{pop}, our datasets of field and open cluster stars. Our field dataset X_{pop} contains spectra for 151,145 red-giant like stars matching a set of quality cuts on the 16th APOGEE data release described below. Our open cluster dataset X_{clust} contains spectra for 185 stars distributed across 22 open clusters, obtained after further quality cuts using the OCCAM value-added catalogue Donor et al. (2020), a cat-

alogue containing information about candidate open cluster observed by APOGEE. We also create baseline datasets Y and Y_{clust} containing stellar abundances for the stars in X and X_{clust}. We include abundances for 21 species in Y and Y_{clust}: C, CI, N, O, Na, Mg, Al, Si, S, K, Ca, Ti, TiII, V, Cr, Mn, Fe, Co, Ni, Cu, Ce. These abundances are derived from the X_H entry in the allStar FITS file.

To create the dataset of field stars X_{pop}, we make the following dataset cuts. With the intention of only preserving red-giant stars, we discard all but those stars for which $4000 < T_{\rm eff} < 5000$ K and $1.5 < \log g < 3.0$ dex, where we use the $T_{\rm eff}$ and $\log g$ derived by the ASPCAP pipeline. In addition, we further exclude any stars for which some stellar abundances of interest were not successfully estimated by the ASPCAP pipeline by removing any star containing abundances set to -9999.99 for any of our 21 species of interest. We also exclude all spectra for which the STAR_BAD flag is set in ASPCAPFLAG. The pseudo-continuum spectra of those remaining stars, as found in the AspcapStar FITS file, were used to create the matrix X_{pop} in which each column contains the spectrum of one star.

To create the dataset of open cluster member stars X_{clust}, we cross-match our filtered dataset with the OCCAM value-added catalogue Donor et al. (2020) so as to identify all candidate open clusters observed by APOGEE. We only keep those spectra of stars with open cluster membership probability CG_PROB>0.8 (Cantat-Gaudin, T. et al. 2018). After this cross-match, we further filtered the dataset by removing those clusters containing only a single member star as these are not useful for us. Additionally, we further discard one star with Apogee ID "2M19203303+3755558" found to have a highly anomalous metallicity. After this procedure, 185 OCCAM stars remain, distributed across 22 clusters. We do not cut any stars based on their signal-to-noise ratio. The stars in X_{pop} have a median signal-to-noise ratio of 157.2 and interquantile range of 102.0-272.4 while those in X_{clust} have a median signal-to-noise ratio of 191.4 and interquantile range of 117.7-322.6.

Because of cosmic rays, bad telluric line removal or instrumental issues, the measurements for some bins of stellar spectra are untrustworthy. We censor such bad bins to prevent them from impacting our low-dimensional representation. Censored bins are treated as missing values in the PPCA compression. In this work, we have censored any spectral bin for which the error (as found in the AspcapStar FITS file error array) exceeds a threshold value of 0.05. Additionally, we censor for all stars in the dataset, those wavelength bins located near strong interstellar absorption features. More detail about

the model-free procedure for censoring interstellar features can be found in Appendix B.1.

5.4.2 Measuring Chemical Similarity

Evaluating how good a representation is at measuring the chemical similarity of stars requires a goodness of fit indicator for assessing the validity of its predictions. We use the "doppelganger rate" as our indicator. This is defined as the fraction of random pairs of stars appearing as similar or more similar than stellar siblings according to the representation, where similarity is measured in terms of distance d in the studied representation. It is worth noting that this procedure for estimating doppelganger rates is related to but different from the probabilistic approach presented in Ness et al. (2018) and from the approach used in the previous chapter to compare stellar twins (Chapter 4).

We estimate doppelganger rates on a per-cluster basis by measuring distances between pairs of stars in the RSCA output representation. For each cluster in X_{clust}, the doppelganger rate is calculated as the fraction of pairs composed of one cluster member and one random star whose distance in the studied representation, $d_{inter-family}$ is $\leq$ than the median distance amongst all cluster pairs $d_{intra-family}$, that is, $d_{intra-family}$ are pairs composed of two confirmed cluster members within X_{clust} and $d_{inter-family}$ are pairs with one random field star selected from X_{pop} and one cluster member from the studied cluster in X_{clust}. When calculating $d_{inter-family}$, we only consider pairs of stars with similar extinction and radial velocity, that is to say with ΔAK_TARG < 0.05 and ΔVHELIO_AVG < 5. By only comparing stars at similar extinction and similar velocity, we ensure that any model being investigated cannot reduce its doppelganger rate through exploiting extinction or radial velocity information in the spectra.

So as to facilitate comparisons between different representations, we aggregate the per-cluster doppelganger rates into a "global" doppelganger rate which gives an overall measurement of a representation's effectiveness at identifying open clusters. The global doppelganger rate is obtained by averaging the per-cluster doppelganger rates through a weighted average in which clusters are weighted by their size in X_{clust}.

There is an added subtlety to assessing a representation through its global doppelganger rate. There are very few open cluster stars in the dataset. Therefore, RSCA as a data-driven procedure applied to open clusters, is susceptible to overfitting to the open cluster dataset. To prevent overfitting from affecting results, we carry out a form of cross-validation in which clusters are excluded from the dataset used for the derivation of

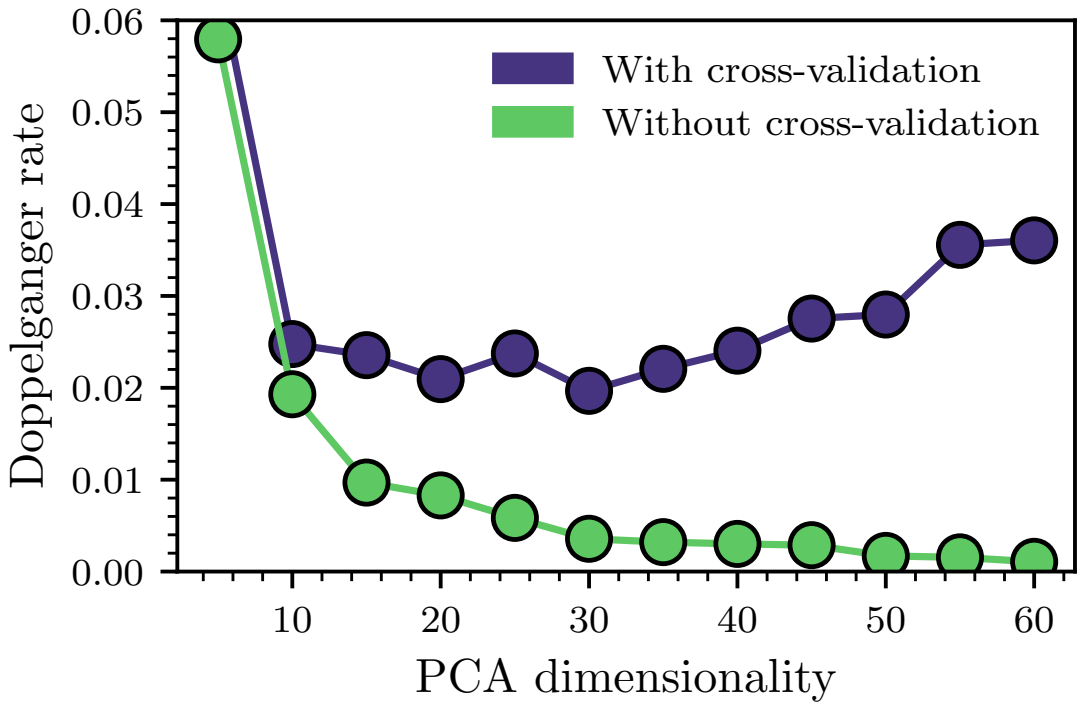

Figure 5.2: *Global doppelganger rates as a function of the number of PCA components used to encode spectra. Performance with cross-validation is shown in blue while performance without cross-validation is shown in green.*

their own doppelganger rate. In this scheme, calculating the global doppelganger rate of an RSCA representation requires repeated application of our algorithm, each time on a different subset with one cluster removed, as many times as there are open clusters.

We caution that our cross-validation approach has some implications on the derived doppelganger rates. Because, every cluster's doppelganger rate is evaluated on a slightly different data subset, the quoted distances and doppelganger rates are not comparable from cluster to cluster.

5.4.3 PCA Dimensionality

The number of principal components used in the compression, or encoding stage of RSCA (Step 1) is an important hyperparameter requiring tuning. In Figure 5.2, we plot the doppelganger rate against the number of principal components, both with and without using the cross-validation procedure described in section 5.4.2. Results without cross-validation display significant overfitting and are mostly shown in an effort to highlight the importance of the cross-validation procedure.

This Figure illustrates how, unsurprisingly, RSCA's performance is strongly dependent on the PCA dimensionality. Doppelganger rates decrease with increasing PCA dimensionality, up until a dimensionality of size 30, but largest gains can be associated to the first 10 PCA components. At K > 30, doppelganger rates start increasing because the marginal improvements in performance from adding additional dimensions are offset by overfitting. Since on our studied dataset, RSCA reaches its peak performance for 30 PCA components, in order to maximize the chemical content in our PCA representation whilst avoiding excessive overfitting, all further quoted results and figures will accordingly use the first 30 PCA components.

That RSCA's performance still improves up to a dimensionality of 30 is interesting. This demonstrates that a hyperplane of at least size 30 is required for capturing the intrinsic variations of APOGEE stellar spectra, a number noticeably larger than the 10 dimensional hyperplane found in Price-Jones & Bovy (2017b). Methodological differences between studies may partially explain such differences. For example, the PCA fit in Price-Jones & Bovy (2017b) was applied on spectra displaying limited instrumental systematics and with non-chemical imprints on the spectra preliminary removed. It should also be noted that this does not mean that the chemical space is 30 dimensional. Some PCA dimensions may capture instrumental systematics or non-chemical factors of variation, such as residual sky absorption and emission and interstellar dust imprints. Also, since chemical species leave non-linear imprints, because of the linearity of PCA, each non-linear chemical dimension may require multiple PCA components to be fully captured.

5.4.4 RSCA interpretability

We now study which spectral features are leveraged by the RSCA algorithm when recognizing open clusters. In the RSCA rescaled basis (Step 4), the dimensions are scaled proportionally to their perceived usefulness at measuring chemical similarity. Therefore, factors of variation judged most important by RSCA will correspond to the most strongly scaled dimensions of the representation (ie with the largest $\frac{1}{\sigma_{r_i}}$). Figure 5.3 shows the relationship between the three features with scaled dimensions with the largest amplitudes and [Fe/H] for a representation obtained by running the RSCA algorithm with a PCA dimensionality of 30.

As seen from the leftmost panel, there is a close relationship between "Feature #1" - the RSCA dimension with the largest associated scaling factor and the ASPCAP [Fe/H]

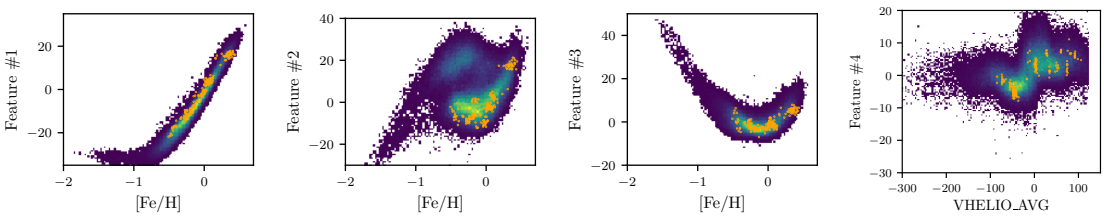

Figure 5.3: *Three features judged most important by metric learning approach plotted against [Fe/H] for the 151,145 stars in X_{pop} and the fourth most-important feature plotted against VHELIO_AVG (radial velocity ASPCAP label). Location of the 185 stars in X_{clust} (the open cluster dataset used to train the metric learning model) are shown by orange markers.*

label. The relationship is close to a one-to-one mapping, which illustrates how this feature traces the metallicity content of the stellar spectra. Because "Feature #1" (as a direction in a hyperplane of stellar spectra) is a linear function of the stellar spectra and metallicity is a non-linear feature, some degree of scatter in the relationship is expected.

The relationship between "Feature #2" and [Fe/H] (second panel) exhibits the same bimodality as observed when plotting alpha enhancements [α/Fe] against [Fe/H] (Leung & Bovy 2018). This indicates that "Feature #2" captures α-element enhancements. It is particularly noteworthy that we are able to recover the α-element bimodality when the open clusters in our dataset (orange markers) are located only in the low-α sequence. This shows that RSCA was able to extract a feature resembling α-elements which generalizes to stars in the field dataset which are dissimilar to those in the open cluster sample. This demonstrates the metric model's capacity to extrapolate to abundance values outside the range of values covered in the open cluster training dataset. This provides evidence that the model may still be effective for stars atypical of those in the open cluster dataset provided that similar stars exist in the field dataset.

The relationship between "Feature #3" and metallicity (third panel) is not as easily interpreted. Given the non-linear nature of metallicity, it is possible that it encodes residual metallicity variability not captured by "Feature #1" but it is also possible that it contains some further independent chemical dimension.

This figure illustrates a nice property of RSCA. Because dimensions of the RSCA representation correspond to eigenvectors of the covariance matrix Σ_{clust}, the RSCA algorithm, at least to first order, assigns the distinct factors of variation within spectra to separate dimensions of the representation. That is, to say that the dimensions of the RSCA representation capture distinct factors of variation, such as the metallicity or the

alpha-element abundance, rather than a combination of factors of variation. Additionally, the most important factors of variation for recognizing open clusters occupy the dimensions with the largest scaling factors. This property makes RSCA particularly versatile. For example RSCA can be used to separate out high and low α-abundance stars in the disk, or to select low-metallicity stars. Additionally, it is likely that because of this property RSCA could be used to search for hidden chemical factors of variation within stellar spectra, although this has not been attempted in this thesis chapter.

We found that some of the dimensions of the RSCA representation showed trends with radial velocity despite our model operating on the AspcapStar spectra which are shifted to the rest-frame. An example of a dimension showing a trend with radial velocity is shown in the last panel of Figure 5.3 and an investigation into the detailed causes of the radial velocity trends is presented in Appendix B.2. The existence of such trends indicates that even after our pseudo-continuum normalization procedure, RSCA is still capable, at least weakly, to exploit radial velocity information in the spectra to recognize stellar siblings. Because only a subset of the dimensions show such trends, a representation tracing only chemistry can be obtained by only keeping those dimensions which show no trends with radial velocity. In this work, we propose to only keep the first three dimensions of the representation. While this choice may appear particularly stringent, as we will show in coming sections, these three dimensions contain the bulk of the discriminative power of the representation (see Table 5.1).

5.4.5 Comparison of using RSCA versus measured abundances in calculating chemical likeness

In this section, we compare the effectiveness of measuring chemical similarity using a data-driven approach on the spectra, with that achievable from using measured stellar abundances. To do so, we compare the doppelganger rates that are obtained by RSCA to those from using stellar abundance labels. The results of such a comparison are shown in Figure 5.4. For this Figure, doppelganger rates are measured consistently from abundances and the RSCA approach (see Figure caption for more detail), such that any differences in performance can be attributed to underlying differences in the information content of the representations. For this comparison, we omit the PCA dimensionality reduction step when working with abundances, but otherwise apply the exact same algorithmic approach to abundances and to the spectra. We also compare the performance of the RSCA metric

learning approach (shown in red) to that of alternative simpler representation rescaling approaches applicable for recognizing open-clusters from abundances (shown in blue and green). We remind the reader that the doppelganger rates are evaluated for pairs of stars at the same extinction and radial velocity. This guarantees that the doppelganger rate cannot be artificially reduced through our model exploiting information relating to radial velocity or extinction. Per-cluster doppelganger rates are also provided in Appendix B.5.

From this Figure, we see that executing all steps of the RSCA algorithm is crucial for obtaining low doppelganger rates when working directly with spectra. This is seen from how the low doppelganger rates are only obtained with full application of our metric-learning approach (red). On the other hand, when working with stellar abundances, the RSCA approach appears to bring only limited benefits - as seen from the only slight difference in doppelganger rates between measuring distances in the raw abundance space (blue) and in the transformed space (red) that is obtained by application of our full-metric learning approach minus the PCA compression. This result is not surprising and reflects how most steps of the RSCA approach are designed with the aim of generating a representation comparable to stellar labels, that is to say one where all factors of variation other than chemical factors of variation are removed. The RSCA approach does however still bring some benefits when working with abundances as seen from the lower doppelganger rates.

This Figure also shows that excluding chemical species improves the doppelganger rate. This is seen from how the doppelganger rate is lower when using a carefully chosen subset of species (right) than when using the full set of abundances (center). This can appear counter-intuitive as it implies that more data leads to worsened performance but in this specific case, where the uncertainties on abundances are not accounted for, it can be justified by the low intrinsic dimensionality of chemical space. Since many species contain essentially the same information, adding species with higher uncertainty into the representation adds noise into the representation. It does this without contributing any additional information beneficial for recognizing open clusters. We expect that such an effect would disappear when accounting for uncertainties on stellar labels, but it is still a good illustration of the brittleness of abundance-based chemical tagging.

The combination of species shown in red is the set of species which were found, after manual investigation, to yield the lowest doppelganger rates. This is the combination of stellar individual element abundance labels Fe,Mg,Ni,Si,Al,C,N (with respect to Fe with

the exception of Fe which is with respect to H). We caution that there is a possibility that some of the chosen species have only lowered the doppelganger rates because of spurious correlation with the open-cluster sample (i.e. overfitting) and not because they carry geniunely useful information for recognizing open-clusters. The doppelganger rate from this combination of species, of 0.023, despite being the smallest doppelganger rate achieved from stellar labels, is higher than the doppelganger rate obtained from stellar spectra of 0.020 (2 percent). That our method is able to produce better doppelganger rates from spectra than from stellar labels highlights the existence of information within stellar spectra not being adequately captured by stellar labels. While stellar labels are derived from synthetic spectra which only approximately replicate observations, our fully data-driven model makes direct use of the spectra translating into lower doppelganger rates.

5.4.6 Dimensionality of chemical space

That the PCA representation is 30 dimensional does not mean that all 30 dimensions carry information useful for recognizing open clusters. To get a grasp of the dimensionality of the chemical space captured by RSCA, we calculated the doppelganger rates for RSCA representations in which only the dimensions with the largest scaling factors are kept (ie other dimensions are excluded from the doppelganger rate distance calculations). We calculate doppelganger rates multiple times, each with a different number of dimensions preserved. The results of this investigation are shown in Table 5.1. From this Table, we see that the dimensionality of spectra appears to be, at least to first degree, extremely low. The top two dimensions of the RSCA model (as shown in Figure 5.3) are capable of matching the performance obtained from stellar labels whilst the top four dimensions exceed the performance from using the full representation. The four-dimensional representation is even more effective at recognizing chemically identical stars than the full RSCA representation which itself was more effective than stellar labels. It is not a new result that the dimensionality of chemical space probed by APOGEE disk stars is low. Recent research suggest that, at the precisions captured by APOGEE labels, chemical abundances live in a very low-dimensional space, for stars of the disk. For example, it was found in Ness et al. (2019); Ting & Weinberg (2021); Weinberg et al. (2021) that [Fe/H] and stellar age or [Fe/H] and [Mg/Fe] could predict all other elemental abundances to within or close to measurement precision (nonetheless, Ting & Weinberg (2021) and Weinberg et al.

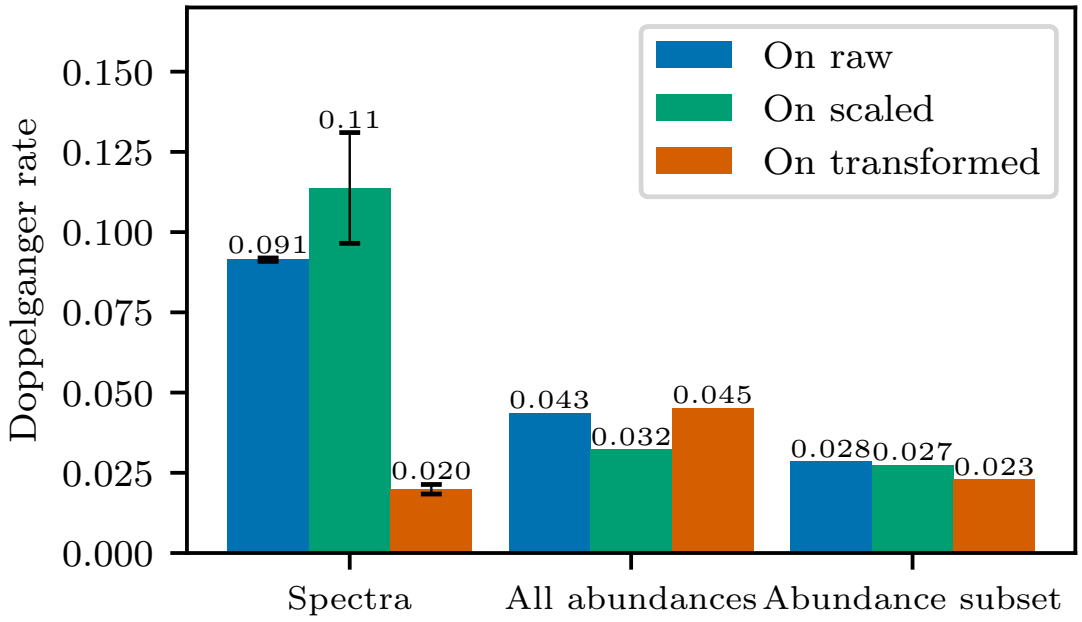

Figure 5.4: *Global doppelganger rates estimated for varying metric-learning approaches and representations. On the x-axis, "spectra" refers to doppelganger rates obtained from spectra X after dimensionality reduction with PPCA to a 30-dimensional space, "all abundances" to doppelganger rates obtained from a representation formed from the full set of APOGEE abundances in Y, "abundance subset" to doppelganger rates obtained using a representation formed only from the abundances for the following species: Fe,Mg,Ni,Si,Al,C,N. Global doppelganger rates "on raw" (blue) are obtained by measuring distances in the raw representation without any transformation of the representation, "on scaled" (green) are obtained by applying the scaling transform on the raw representation without preliminary application of the sphering and reparametrization transform (Steps 1 and 4 for spectra and only Step 4 for abundances which do not need dimensionality reduction), "on transformed" are obtained by applying all steps of the proposed metric learning approach (Steps 1,2,3 and 4 for spectra and Steps 2,3,4 for abundances). As the implementation of the PPCA algorithm used in this thesis chapter yielded stochastic PCA components, doppelganger rates from spectra correspond to the mean across 10 runs with error bars corresponding to the standard deviation amongst runs.*

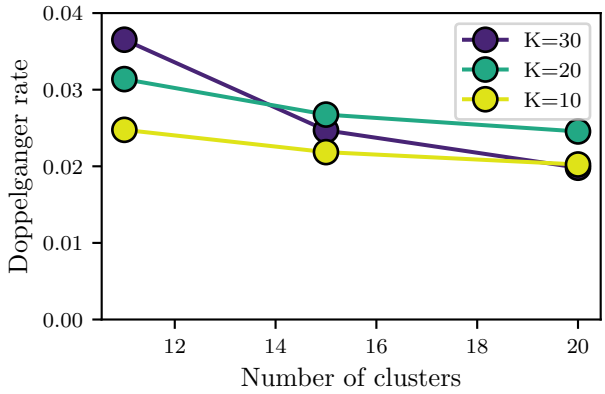

Figure 5.5: *Expected global doppelganger rates when training a metric-learning model on only a subset of all open clusters in X_{clust} with a number of clusters given by the x-axis. Results for different PCA dimensionalities used for compressing stellar spectra are represented by different colored lines. Clusters used in the expectated doppelganger rate calculations were chosen randomly from X_{clust}, and quoted results are for the average of 50 repeated trials.*

(2021) argue that correlations of abundance residuals imply underlying intrinsic structure in abundance space, even if the scatter of these residuals is only moderately larger than the per-star observational uncertainties). However, while previous analysis have depended on abundances to show this, here we can do this directly from spectra. As our methodology directly picks up on factors of variation and, if not controlled for, is capable of picking up on weak factors of variation such as diffuse interstellar bands, we can be confident that any remaining chemical factors of variation are either i) highly non-linear for our model to not be capable of picking up on them, ii) very weak spectral features, or iii) not particularly discriminative of open clusters as would be the case for chemical variations arising from internal stellar processes for example induced by accretion of planetary materials or internal stellar processes.

5.4.7 Impact of Dataset Size

Our method learns to measure chemical similarity directly from open clusters without reliance on external information. Because of this, its performance will be tightly linked to the quality and quantity of data available. Figure 5.5 attempts to estimate our method's dependency on the size of the open cluster dataset. In this figure, for varying PCA dimensionalities, we plot the expected doppelganger rate for an open cluster dataset containing a given quantity of open clusters, whose number is given by the x-axis. We estimate the expected doppelganger rate for a given number of open clusters by estimating and averag-

N	Doppelganger rate
1	0.0962 ± 0.0212
2	0.0219 ± 0.0025
3	0.0198 ± 0.0028
4	0.0182 ± 0.0021
5	0.0180 ± 0.0021
6	0.0188 ± 0.0017
7	0.0184 ± 0.0017
30	0.0199 ± 0.0015

Table 5.1: *Global doppelganger rate obtained by the RSCA model applied to stellar spectra in which all but the N most strongly scaled dimensions of a 30 dimensional RSCA representation are discarded. As the implementation of the PPCA algorithm used in this thesis chapter yielded stochastic PCA components, doppelganger rates from spectra correspond to the mean across 10 runs with error bars corresponding to the standard deviation amongst runs.*

ing the doppelganger rates for all data subsets containing that number of clusters. From this figure, we see that the larger PCA dimensionalities still benefit from the addition of open clusters. This is suggestive that performance would likely further improve with access to additional open clusters. Such larger datasets may also enable the usage of more complex non-linear metric-learning approaches which may yield further improvements not captured in this figure.

5.4.8 Impact of stellar sample

Because RSCA is a data-driven method, the chemical representation it extracts will inexorably be linked to the stellar sample used. As a consequence of this, our findings on the low-dimensionality of the chemical space are only valid in the context of our stellar sample and may not generalize to spectra dissimilar to this sample. In Figures 5.6, 5.7 and 5.8, in an effort to help elucidate the domain of validity of our results, we show ages, locations, metallicity ([M/H]) and alpha-element enhancement ([α/M]) for our APOGEE stellar samples X_{pop} and X_{field}.

We wish to draw attention to the fact that as APOGEE primarily probes disk stars (Ahumada et al. 2020), conclusions about the low-dimensionality of chemical space may not generalize to other galactic components such as the halo or the bulge. Similarly, as the APOGEE spectral window does not probe all nucleosynthetic pathways and in particular does not contain many spectral lines associated to r-process and s-process elements, conclusions about the dimensionality of chemical space may change when spectral windows

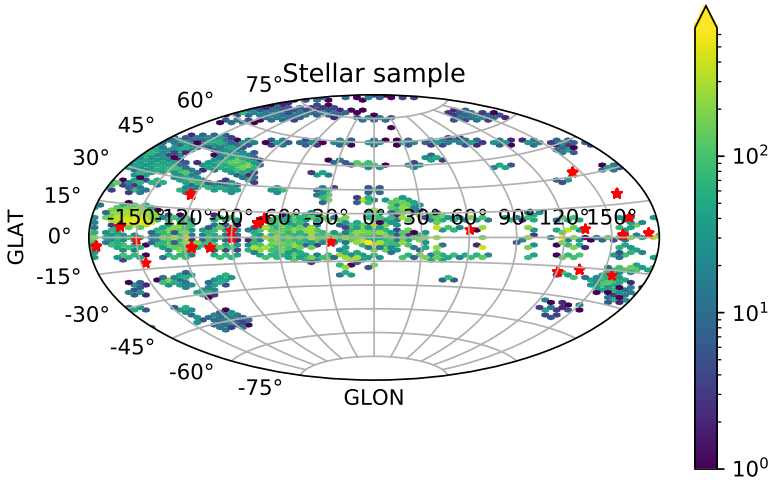

Figure 5.6: *Galactic longitude and latitude of our sample of stars. Heatmap shows density of field stars (X_{pop}) whilst red markers denote location of open-clusters (X_{field}).*

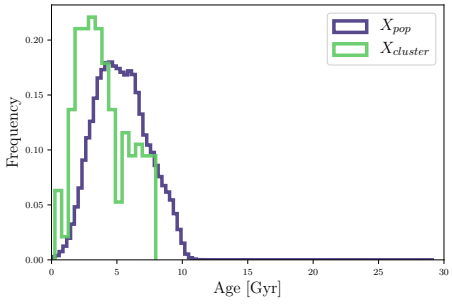

Figure 5.7: *Histogram showing distribution of ages for our field sample (X_{pop}) and open cluster sample (X_{clust}). We use ages from the AstroNN catalogue (Mackereth et al. 2019) derived using a neural network. We caution that ages in this catalogue for the oldest stars (≈ 10 - 11 Gyr) are likely to be underestimated.*

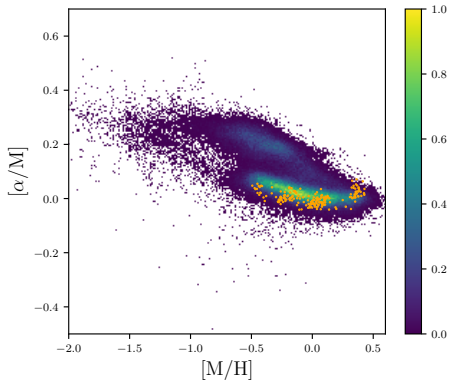

Figure 5.8: *[α/M] plotted against [M/H] for the 151,145 stars in X_{pop} Location of the 185 stars in X_{clust} (the open cluster dataset used to train the RSCA algorithm) are shown by orange markers.*

containing strong lines from these families of elements are considered. We also further caution that the performance of RSCA may deteriorate for older stars, as the algorithm learns to measure chemical similarity from open-clusters which are typically younger than the general population of stars. Although, without further work, it is unclear whether this would be an issue.

5.5 Discussion

We have presented a novel approach for identifying chemically similar stars from spectroscopy based on training a metric-learning model on the spectra of open cluster stars. This approach has several appealing properties. It is end-to-end data-driven in the sense that it does not rely on synthetic models nor labels derived from synthetic models. This method only makes use of open cluster members, which can themselves be identified with minimal reliance on theoretically-derived quantities (Gao 2014; Agarwal et al. 2021; Castro-Ginard, A. et al. 2018). This makes the method insensitive to the domain gap of synthetic spectra. Additionally, where traditional spectral fitting approaches require instrumental systematics to be fully suppressed, lest they further exacerbate the domain gap, our fully data-driven approach, at least in theory, automatically learns on its own to ignore most variations due to instrumental systematics.

We expect that the approach that we have developed will perform particularly well in a number of regimes. For example, we expect it to perform well on low resolution spectra, where blended features lead to compounding model inaccuracies and on M-type stars where molecular features complicate the retrieval process. In general, our method will likely be an efficient and effective approach where theoretical models are inaccurate or the observed spectra itself is plagued by complex systematics. This is for example the case for M dwarfs (Birky et al. 2020; Behmard et al. 2019).

Although our data-driven algorithm shows excellent performance, one may wonder whether there is still room for further improvements, particularly since strong chemical tagging, if ever possible, will require improvements in our chemical similarity measurements (Ting & Weinberg 2021). There are reasons to be hopeful here.

Firstly, our data-driven model comes with clear limitations which are not inherent to the approach, but rather imposed by our modelling choices. There are many algorithmic choices for building a metric learning model optimized at distinguishing open clusters (e.g.

Weinberger & Saul 2009; Goldberger et al. 2004; Shental et al. 2002; Murphy 2021) and ours is only one of many choices. In particular, the ability of RSCA to extract the chemical content of stellar spectra is constrained by its linear nature. Although this linearity is convenient for avoiding over-fitting, enabling better out-of-distribution performance and facilitating cross-validation, it also artificially limits the precision with which "pure" features containing only chemical signal can be learned. One can be hopeful that a suitably regularized non-linear metric learning models, such as for example twin neural network (Chopra et al. 2005; Murphy 2021), could surpass our model. However, building such a model from our limited number of open clusters would present its own unique challenges.

Secondly, by being entirely data-driven, the performance of this approach is inexorably linked to the quality and quantity of data available. This makes it poised to benefit from new open cluster discoveries and/or deliberately targeted observations. Improvements may also be possible by leveraging other source of chemically similar stars such as wide binaries.

Our method also comes with caveats. Data-driven methods do not extrapolate well outside of their training dataset. As such, performance may be lower for clusters that are atypical compared to those in the open cluster reference set. Since open clusters are typically younger stars (Portegies Zwart et al. 2010), this means performance may be decreased on older cluster stars. However, given the tight relationships between RSCA dimensions and chemical parameters for stars in X_{pop}, such an effect is likely to be small. Additionally, our model makes no use of the error information in spectra, which is valuable information that could likely be squeezed-out for even better performance.

Another downside of our approach is its coarse-grained nature. While stellar labels provide a fine-grained view into chemical similarity with a breakdown into chemical composition of individual species, our approach only provides a coarse-grained measurement of chemical similarity. This limits the types of scientific problems the approach can be used to answer. However, it may very well be possible to extend the method from measuring overall chemical similarity to measuring individual elemental abundances. A way in which this could possibly be done is through applying it to windows centered on the locations of stellar lines instead of to the full spectrum. Also, it is not always clear what exact information is captured by the representation, and in particular there is always a risk, despite all of our checks, that the model is acting on non-chemical information within the spectra.

5.6 Conclusion

Large-scale Galactic surveys, like APOGEE, LAMOST and GALAH, have collected hundreds of thousands of high-quality stellar spectra across the galaxy. These surveys are vastly broadening our understanding of the Milky Way. How best to analyse these spectra however still remains an open-question. One limitation is that traditional spectral fitting methods currently do not make full use of the information in stellar spectra. This is largely because our stellar models are approximations.

In this thesis chapter we developed a fully data-driven, linear metric learning algorithm operating on spectra for extracting the chemical information within stellar families. Through experiments on APOGEE, we demonstrated that our metric learning model identifies stars within open clusters more precisely compared to using stellar labels which indicates an improved ability to discriminate between chemically identical stars. We further found that our model's capacity to distinguish open clusters could largely be attributed to a two-dimensional subspace of our final representation which was found to approximately coincide with metallicity and α-elemental abundances. That our model's capacity at recognizing open clusters plateaus at N$\sim$ 4 supports the idea that the dimensionality of chemical space probed by APOGEE is, for Galactic archaeology purposes, low, in the disk. However, we do find hints of further dimensions potentially containing chemical information and could expect different conclusions had further population of stars and families of elements been considered.

There are several reasons why our metric-learning approach could be favoured over using stellar labels. It can be applied to spectra of stars for which we do not yet have very good synthetic spectra and so would otherwise not be able to analyze well. It is completely independent of our theoretical knowledge of stellar atmospheres and so could be used to validate existing astronomical results in a way which is independent of any biases thatf may exist in our synthetic spectra. Finally and perhaps most importantly, whereas the traditional derivation of stellar labels is fundamentally limited by our inability to generate faithful synthetic spectra, our metric learning approach does not suffer from such a limitation. This means that by improving the quality of the training dataset and the metric-learning approach used, performance may be further improved.

Chapter 6

Conclusion and future prospects

This thesis has focused on the use of machine learning algorithms to better constrain the chemical information within astronomical spectra. This has involved the study of interstellar spectra, as presented in Chapter 2 and 3, and of stellar spectra in Chapter 4 and 5.

In the case of interstellar spectra, molecular abundances are difficult to constrain from observations. The radiative transfer inverse problem that must be solved in order to map observed molecular intensities back to the parameters and abundances of the interstellar medium gas they trace is typically degenerate, especially when more than one gas phase is observed. Because astrochemical models carry information about which molecular species coexist together, they contain valuable information for breaking such degeneracies. However, the relatively long running time of astrochemical models hinders their usage alongside radiative transfer models and poses a barrier to their widespread usage.

In Chapter 2 of this thesis we present an effort at incorporating astrochemical codes into the radiative-transfer modelling task. We developed and publicly released an emulator for the UCLCHEM astrochemical code, which enables the estimation of UCLCHEM model outputs in a fraction of the computational time otherwise required. This speedup enables the incorporation of astrochemical codes into the radiative transfer inverse problem. We then demonstrated, through examples on synthetic and real observations, how, by calculating Bayesian posteriors on the joint model of chemistry and radiative transfer, the

emulator could be used to break radiative-transfer degeneracies and also estimate physical parameters, such as cosmic-ray ionization rate, that would not be accessible without knowledge of the astrochemical networks.

There are many possible applications of our UCLCHEM astrochemical emulator beyond those shown in this thesis. The experiments carried out aimed at showing how physical conditions of the gas probed by molecular line observations could be constrained by the chemical knowledge implicit in astrochemical models. However, conversely, provided a large enough dataset containing observations of molecular line intensity for many observations of gas covering a wide range of physical conditions, it should be possible to make inference in the other direction and use the emulator to constrain chemical knowledge from observations. This could, for example, be accomplished by estimating Bayesian posteriors for held-out molecular transitions from all other transitions and comparing to observed intensities. Assuming sensible priors, the quality of the agreement between posteriors and observed intensities would give an idea of the accuracy of chemical networks for studied molecules. Alternatively, the Bayes factor between different chemical networks for a large enough dataset of observations and carefully chosen priors could be used to identify which chemical networks, from a set of networks of interest, most closely matched observations.

In Chapter 3 of this thesis, we investigated the use of NMF - a matrix-factorization approach - for separating out the molecular emission originating from different gas components with overlapping emission. We showcased the strengths but also the weaknesses of such an approach through experiments on synthetic datasets. We found the approach capable of recovering meaningful gas components resembling the ground-truth components present within observations. However, we also found that irreducible degeneracies in the NMF matrix factorization task prevented the algorithm from exactly recovering the ground-truth gas phases in many practical situations involving overlapping gas phases.

Chapter 4 and Chapter 5 of this thesis present two complimentary approaches for extracting the chemical content of stellar spectra. The work in these chapters is motivated by the inaccuracies introduced by reliance on synthetic stellar spectra when determining stellar labels. Chapter 4 presented the proof-of-concept for a method aiming at removing the imprint of nuisance parameters on stellar spectra. The idea behind the approach is to learn, using a neural network, a representation in which known factors of variation are removed. By removing the stellar physical parameters and instrumental systematics,

such a technique could then be used to directly measure chemical similarity without using synthetic spectra, or alternatively, used for extending differential analysis to stars with highly different temperatures and surface gravities.

A weakness of the work presented in Chapter 4 is that it is only capable of removing factors of variation which are known and so may struggle to remove unknown or poorly estimated systematics. The method presented in Chapter 5 addresses such weakness. Instead of specifying factors of variation to be removed, it automatically removes all variations within spectra which are unhelpful for recognizing open-clusters, leaving behind a compact representation containing only the chemical information of the spectra.

We applied this new method to stellar spectra from the APOGEE survey and found it more effective at recognizing open-clusters than the stellar labels derived by the official survey pipeline, providing evidence that the approach is more effective at extracting the chemical information contained in the spectra than the stellar labels. Furthermore, an investigation into the internals of the RSCA model found that that the information used by the algorithm when recognizing open-clusters was, for the most part, low-dimensional with the two primary dimensions corresponding to metallicity and alpha-abundances.

Although stellar spectra contain line transitions for a wealth of atomic elements, it is becoming increasingly clear to the research community, that the abundances of many elements in the disk are strongly coupled such that the chemical space probed by stellar surveys is in actual fact rather low-dimensional with only a few degrees-of-freedom. Constraining and measuring the dimensionality of this chemical space in the disk has become a goal of modern galactic archaeology. Our works offers a model-free alternative to abundance labels for finding the chemical dimensionality of spectra. In agreement with existing approaches applied to stellar labels, we find that birth chemical abundances, as probed by APOGEE, live in a low-dimensionality chemical space. However, we also find some weak evidence of further chemical dimensions.

Beyond its uses for constraining the dimensionality of spectra, this work lays the foundations for a new, entirely data-driven, approach of analysing stellar spectra that has the advantages of being free from the biases introduced by synthetic spectra and applicable to spectra for which accurate synthetic spectra may not exist. A further appeal of such a data-driven approach is that its performance is not static and can be improved through larger datasets and methodological improvements.

There are many possible extensions to this line of work on model-free analysis of

stellar spectra. One avenue of work would be to extend the RSCA approach to new surveys. For example, applying RSCA to the GALAH survey stellar sample (De Silva et al. 2015) could help shed light on the relationship between chemical abundances across different nucleosynthetic channels. As the RSCA algorithm requires a large number of open clusters for good performance, the GAIA-RVS survey (Randich et al. 2013), which will observe a large number of open-clusters, could also be particularly suitable for further analysis. Another avenue of work would be to improve upon the developed algorithms, for example through replacing the metric-learning in RSCA with a twin neural network approach (Chopra et al. 2005) or through exploiting the sparse contribution of chemical abundances to spectra (Moran et al. 2022). Finally, the introduced approaches could be adapted to be used as tools for preprocessing stellar spectra. For example, the methods could be adapted to remove telluric lines or radial velocity systematics from spectra (as discussed in Chapter 5).

Appendix A

Supplementary material for Chapter 4

A.1 Neural Network Training Details

We briefly review some implementation details useful for reproducing the results in Chapter 4. We have made our repository open-source to aid in making our research reproducible and encourage readers to refer to the code for additional details.

Dataset processing: We process the continuum-normalized spectra by first multiplying the spectra by 4 and then substracting by 3.5. This makes the spectra roughly occupy the [-1,1] range.

Neural network training: All quoted results use feedforward neural networks with self-normalized rectified units (SELU) activation functions (Klambauer et al. 2017). All results are obtained using the ADAM optimizer (Kingma & Ba 2015) with a learning rate of 10^{-5}. In the following, $n_{\rm bins} = 7751$ refers to the number of spectral bins used, $n_{\rm conditioned} = 2$ or 3 the number of parameters the encoder is conditioned on, and $n_{\rm z} = 20$ the size of the autoencoder latent.

FactorDis architecture: Our FactorDis neural network has the following architecture (including input and output layers). Results can be reproduced with loss weighting

term $\lambda = 10^{-4}$

$$\text{encoder dimensions} = \{n_{\text{bins}} + n_{\text{conditioned}}, 2048, 512, 128, 32, n_z\} \tag{A.1}$$

$$\text{decoder dimensions} = \{n_{\text{z}} + n_{\text{conditioned}}, 512, 2048, 8192, n_{bins}\} \tag{A.2}$$

$$\text{discriminator dimensions} = \{n_{\text{bins}} + n_{\text{conditioned}} + n_{\text{z}}, 4096, 1024, 512, 128, 32, 1\} \tag{A.3}$$

FaderDis architecture: Our FaderDis neural network has the following architecture (including input and output layers). Results can be reproduced with loss weighting term $\lambda = 10^{-5}$. When training the auxiliary network, each disentangled parameter was split into 10 discrete values creating 100 equal sized bins when disentangling two parameters and 1000 equal sized bins when disentangling three.

$$\text{encoder dimensions} = \{n_{\text{bins}} + n_{\text{conditioned}}, 2048, 512, 128, 32, n_{\text{z}}\} \tag{A.4}$$

$$\text{decoder dimensions} = \{n_{\text{z}} + n_{\text{conditioned}}, 512, 2048, 8192, n_{\text{bins}}\} \tag{A.5}$$

$$\text{auxilary dimensions} = \{n_{\text{z}} + n_{\text{conditioned}}, 512, 256, 10^{n_{\text{conditioned}}}\} \tag{A.6}$$

Non-linear Chemical Estimation: In Figure 4.4, neural networks, taking latents z as inputs, are used as non-linear estimators of abundances. A separate neural network with the following structure was trained for every chemical specie.

$$\text{non-linear dimensions} = \{n_z, 512, 256, 128, 1\} \tag{A.7}$$

Appendix B

Supplementary material for Chapter 5

B.1 Interstellar masking

Regions containing interstellar absorption features are identified from the APOGEE data using a data driven-procedure as described below.

The method makes use of two datasets: one containing spectra of stars at low extinction X_{low} which should not contain interstellar features and one of high extinction stars X_{high} which should contain strong interstellar features. X_{low} is formed through a dataset cut on X_{pop} in which only stars with AK_TARG$<$ 0.005 are kept, X_{high} only preserves stars with AK_TARG$>$ 0.5.

We apply PCA with 30 principal components to the dataset of low extinction stars X_{low} to obtain a PCA basis capturing the natural variations amongst stellar spectra at low extinction. Since this PCA basis only captures the variations amongst low extinction stars, features in high-extinction stars associated to the interstellar medium will be poorly reconstructed by the projection onto this low-extinction PCA hyperplane.

In Figure B.1, we plot the mean residual per wavelength between stellar spectrum and their projection on the low-extinction PCA hyperplane averaged over all stars in the high-extinction dataset X_{high}. High-residual regions in this fit will correspond to regions poorly captured by the low-extinction PCA hyperplane. Comparing the high-residual

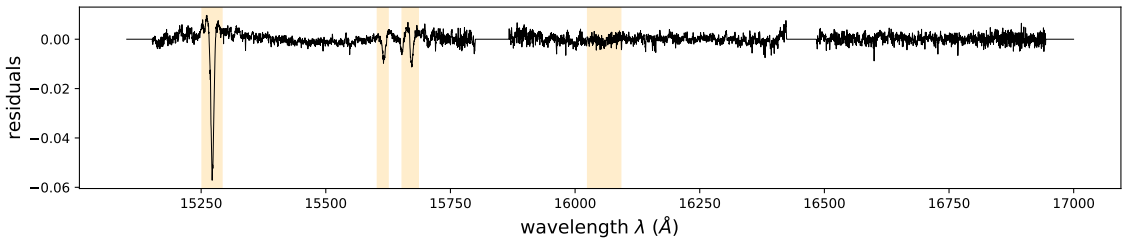

Figure B.1: *Mean residual per wavelength between stellar spectrum and their projection on the low-extinction PCA hyperplane averaged over all stars in the high-extinction dataset X_{high}. High residual spectral bins correspond to wavelengths where interstellar extinction strongly affects stellar spectra. Region highlighted in yellow are the regions that were chosen to be censored to suppress interstellar features from the spectra.*

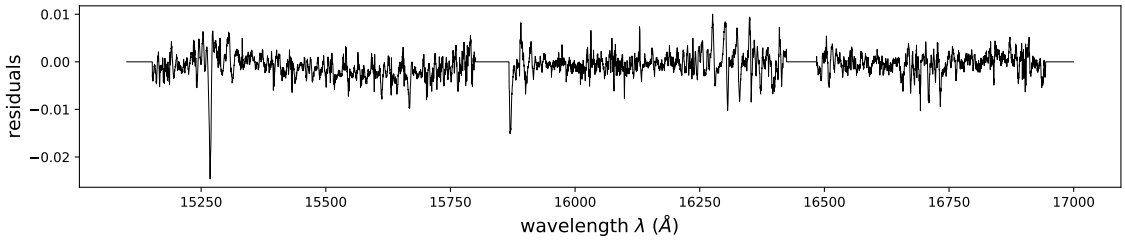

Figure B.2: *Mean residual per wavelength between stellar spectrum and their projection on the low radial-velocity PCA hyperplane averaged over all stars in the high radial-velocity dataset X_{high}. High residual spectral bins correspond to spectral regions with strong dependence on radial velocity.*

regions with the locations of known diffuse interstellar bands reveals excellent agreement (Elyajouri, M. et al. 2017; Elyajouri et al. 2016). For the results presented in Chapter 5, those regions that should be censored were selected manually with the final choice of regions overlain in yellow in Figure B.1.

B.2 Visualizing radial velocity instrumental systematics

We apply a similar approach to that taken for extinction in Appendix B.1 to radial velocities. That is to say we create a dataset of low radial-velocity stars by selecting only stars for which |VHELIO_AVG| < 5 kms^{-1}. We train a PCA model on the low-velocity spectra and visualize the PCA model's residuals on a dataset of high-velocity spectra with VHELIO_AVG > 80 kms^{-1}. As in the previous section we use a 30 dimensional PCA model.

In Figure B.2, we plot the mean absolute residual per wavelength between stellar spectrum and their projection on the low-velocity PCA hyperplane averaged over all stars in the high radial-velocity dataset X_{high}. High-residual regions in this fit will correspond

to regions poorly captured by the low-velocity PCA hyperplane. In this plot, in addition to the diffuse interstellar bands, there appears to be other regions with weak instrumental systematics correlated with radial velocity.

It is worth mentioning that a variant of RSCA can be used for removing radial velocity imprints on the spectra. By applying the RSCA algorithm to same radial velocity stellar groups instead of open clusters, one can identify a hyperplane of the stellar spectral space capturing solely spectral features correlated to radial-velocity. Substracting variations within this hyperplane from stellar spectra then yields spectra in which features correlated with radial-velocity are selectively suppressed. Here, we do not apply such a preprocessing procedure as it complicates the analysis while not improving over the simpler procedure of only keeping the first three dimensions.

B.3 Checking for instrumental systematics

It is worthwhile to ascertain that our model, when identifying open clusters, is only relying on chemical features within the spectra and not on instrumental systematics that happen to be predictive of open clusters and would not transfer towards identifying dissolved clusters. This is especially important as any dependency on instrumental systematics would lead to overly optimistic doppelganger rates.

Because the stars in open clusters are gravitationally bound, they are often part of the same telescope field of view and so observed simultaneously on nearby fibers of a plate. Being observed together could plausibly introduce systematics (due to instrumental imperfections or telluric residuals) which could then be actioned on by the metric learning-model when identifying open clusters. Here, we run an experiment to ascertain that this is not an issue for our model.

Since any such shared instrumental systematic will only affect stars observed simultaneously, we can validate that our model is not exploiting shared instrumental systematics by comparing the similarity distributions for stellar siblings that were observed together on the same plate and for stellar siblings that were not. The idea being that a model exploiting instrumental systematics would have a lower doppelganger rate on the pairs of stars observed together than on the pairs of stars observed separately.

To separate stellar siblings into pairs of stars observed together and pairs observed separately, we went through all pairs of stellar siblings in the open cluster dataset. Using

the individual observation dates of exposures that comprise the combined spectra, as provided by the VISITS allStar field, we categorized pairs of siblings into two groups: those pairs composed of stars observed together i.e. with the same dates of visits for exposures, and those pairs of stars observed separately. When doing this analysis we discarded the small fraction of stellar pairs for which visit dates only partially overlapped.

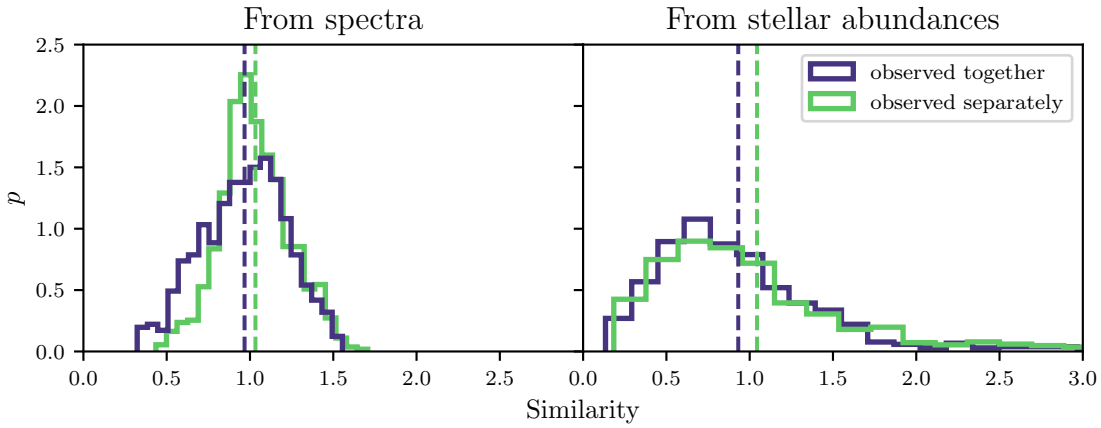

Figure B.3: *Investigation into metric-learning models dependency on instrumental systematics. "From masked spectra" refers to distances derived from a metric-learning model applied to masked stellar spectra. "Stellar abundances" refers to distances derived from a core set of abundances (see Section 5.4 for full details).*

In Figure B.3, we show the distributions of chemical similarities for pairs of open cluster stars observed simultaneously compared to pairs observed separately. At left, we show our metric learning approach applied to spectra (with a 30 dimensional latent). At right, we show this applied to abundances. For both metric-learning models, stars observed together are predicted to be slightly more chemically similar than stars observed separately. Since, both approaches, applied to spectra and to abundances, provide similar behaviours with visit overlap, we conclude that our method is not making strong use of instrumental systematics when recognizing open clusters. It is nonetheless interesting that both approaches seem to marginally favour stars observed together as being more chemically similar, although given the small number of open clusters may be due to small sample sizes.

B.4 RSCA Pseudocode

The Pseudocode for the RSCA algorithm.

Algorithm 1: RSCA Algorithm

Data: X_{clust}, X_{pop}

compressor = PPCA(data = X_{pop}, $\text{n}_{\text{components}} = \text{N}_{\text{K}}$) ; // Step 1

$Z_{\text{pop}} = \text{compressor.transform}(\text{X}_{\text{pop}})$;

$Z_{\text{clust}} = \text{compressor.transform}(\text{X}_{\text{clust}})$;

spherer = sphere(data = Z_{pop}) ; // Step 2

$Z_{\text{pop}} = \text{spherer.transform}(\text{Z}_{\text{pop}})$;

$Z_{\text{clust}} = \text{spherer.transform}(\text{Z}_{\text{clust}})$;

$Z_{\text{intra-cluster}} = \text{ZeroCenterClusters}(\text{Z}_{\text{clust}})$; // Step 3

reparametrizer = PCA(data = $\text{Z}_{\text{intra-cluster}}$, $\text{n}_{\text{components}} = \text{N}_{\text{K}}$) ;

$Z_{\text{pop}} = \text{reparametrizer.transform}(\text{Z}_{\text{pop}})$;

$Z_{\text{clust}} = \text{reparametrizer.transform}(\text{Z}_{\text{clust}})$;

for i = 1 to N_K **do**

$\sigma^2_{\text{clust}_\text{i}} = \frac{\sum_{j=1}^{k}(n_j-1)\sigma^2_{ji}}{\sum_{j=1}^{k}(n_j-1)}$; // Step 4

$\sigma^2_{\text{pop}_\text{i}} = 1$; // because of sphering

$\sigma_{\text{r}_\text{i}} = \frac{\sigma_{\text{clust}_\text{i}}\sigma_{\text{pop}_\text{i}}}{\sqrt{\sigma^2_{\text{pop}_\text{i}} - \sigma^2_{\text{clust}_\text{i}}}}$;

$Z_{\text{pop}_\text{i}} = Z_{\text{pop}_\text{i}} \div \sigma_{\text{r}_\text{i}}$;

$Z_{\text{clust}_\text{i}} = Z_{\text{clust}_\text{i}} \div \sigma_{\text{r}_\text{i}}$

end

B.5 Per-cluster Doppelganger Rates

The doppelganger rates for all open clusters in X_{clust} (Figures B.4 & B.5)

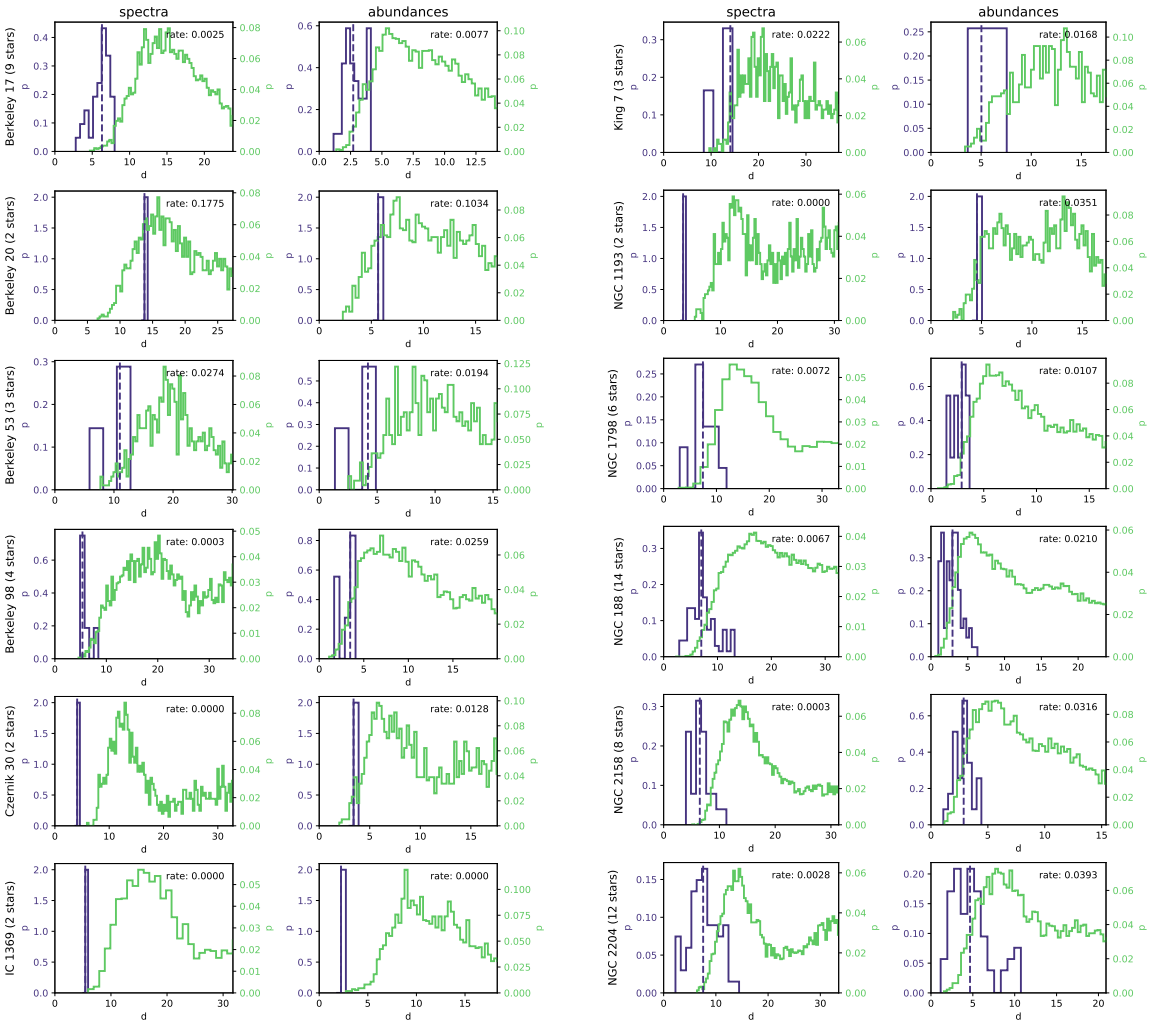

Figure B.4: *Histograms of the chemical similarity between open cluster pairs of stars as predicted by the metric-learning approach. For each open cluster, the distribution of inter-cluster similarities, calculated as the distribution of similarities between pairs of stars composed of one random cluster member and a random field star, is shown in green and the distribution of intra-cluster similarities - similarity between pairs of stellar siblings - is shown in blue. The median intra-cluster similarity, as used in doppelganger rate calculations, is marked by dashed vertical line. The leftmost panel displays the histograms derived from applying the metric-learning approach to stellar spectra. The rightmost panel displays the histograms derived from applying the metric-learning approach to the "abundance subset" as defined and described in Section 5.4.5. Doppelganger rates for individual clusters are shown in top-left corner of every panel.*

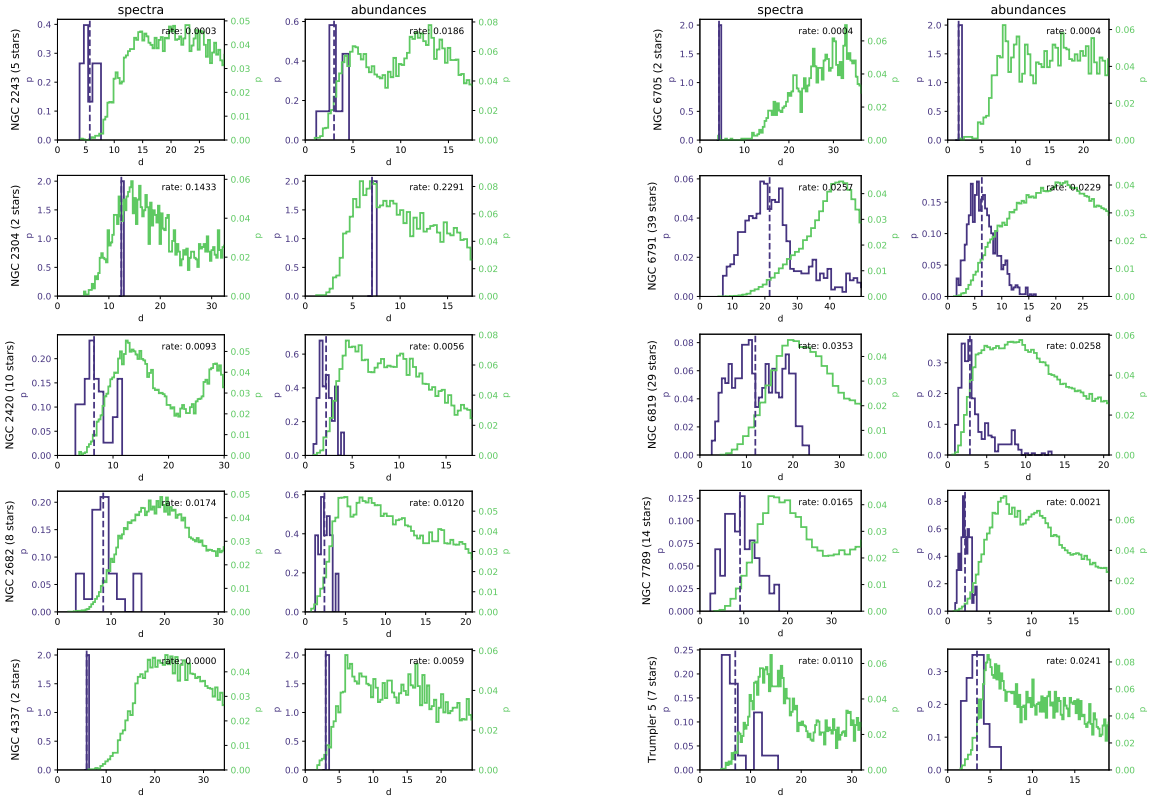

Figure B.5: *Continuation of Figure B.4.*

This page was intentionally left blank

Bibliography

Adibekyan, V. Z., Sousa, S. G., Santos, N. C., et al. 2012, A&A, 545, A32, doi: 10.1051/0004-6361/201219401

Agarwal, M., Rao, K. K., Vaidya, K., & Bhattacharya, S. 2021, Monthly Notices of the Royal Astronomical Society, 502, 2582, doi: 10.1093/mnras/stab118

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJSS, 249, 3, doi: 10.3847/1538-4365/ab929e

Arjovsky, M., Chintala, S., & Bottou, L. 2017, in Proceedings of Machine Learning Research, Vol. 70, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, ed. D. Precup & Y. W. Teh (Pmlr), 214–223. http://proceedings.mlr.press/v70/arjovsky17a.html

Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, Annual Review of Astronomy and Astrophysics, 47, 481, doi: 10.1146/annurev.astro.46.060407.145222

Beane, A., Ness, M. K., & Bedell, M. 2018, The Astrophysical Journal, 867, 31, doi: 10.3847/1538-4357/aae07f

Beck, J., & Guillas, S. 2016, SIAM/ASA Journal on Uncertainty Quantification, 4, 739, doi: 10.1137/140989613

Bedell, M., Bean, J. L., Meléndez, J., et al. 2018, ApJ, 865, 68, doi: 10.3847/1538-4357/aad908

Behmard, A., Petigura, E. A., & Howard, A. W. 2019, The Astrophysical Journal, 876, 68, doi: 10.3847/1538-4357/ab14e0

Belghazi, M. I., Baratin, A., Rajeswar, S., et al. 2018, in Proceedings of Machine Learning Research, Vol. 80, Proceedings of the 35th International Conference on Machine Learn-

ing, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, ed. J. G. Dy & A. Krause (Pmlr), 530–539. http://proceedings.mlr.press/v80/belghazi18a.html

Benedettini, M., Viti, S., Codella, C., et al. 2013, Monthly Notices of the Royal Astronomical Society, 436, 179, doi: 10.1093/mnras/stt1559

Bengio, Y., Courville, A. C., & Vincent, P. 2013, IEEE Trans. Pattern Anal. Mach. Intell., 35, 1798, doi: 10.1109/tpami.2013.50

Berné, O., Joblin, C., Deville, Y., et al. 2007, A&a, 469, 575, doi: 10.1051/0004-6361:20066282

Bertelli Motta, C., Pasquali, A., Richer, J., et al. 2018, Monthly Notices of the Royal Astronomical Society, 478, 425, doi: 10.1093/mnras/sty1011

Birky, J., Hogg, D. W., Mann, A. W., & Burgasser, A. 2020, The Astrophysical Journal, 892, 31, doi: 10.3847/1538-4357/ab7004

Bisbas, T. G., Schruba, A., & van Dishoeck, E. F. 2019, Monthly Notices of the Royal Astronomical Society, 485, 3097, doi: 10.1093/mnras/stz405

Bishop, C. M. 2006, Pattern Recognition and Machine Learning (Information Science and Statistics) (Berlin, Heidelberg: Springer-Verlag)

Blanco-Cuaresma, S., & Fraix-Burnet, D. 2018, Astronomy & Astrophysics, 618, A65, doi: 10.1051/0004-6361/201832815

Bolatto, A. D., Wolfire, M., & Leroy, A. K. 2013, Annual Review of Astronomy and Astrophysics, 51, 207, doi: 10.1146/annurev-astro-082812-140944

Bonaca, A., Conroy, C., Cargile, P. A., et al. 2020, ApJL, 897, L18, doi: 10.3847/2041-8213/ab9caa

Bonifacio, P., Dalton, G., Trager, S., et al. 2016, in Proceedings of the annual meeting of the French Society of Astronomy I& Astrophysics Lyon, June 14-17, 2016, ed. C. Reylé (Société Francaise d'Astronomie et d'Astrophysique (SF2A)), 267–270

Boulais, A., Berné, O., Faury, G., & Deville, Y. 2021, A&a, 647, A105, doi: 10.1051/0004-6361/201936399

Bovy, J. 2016a, The Astrophysical Journal, 817, 49, doi: 10.3847/0004-637x/817/1/49

—. 2016b, The Astrophysical Journal, 817, 49, doi: 10.3847/0004-637x/817/1/49

Bovy, J., Rix, H.-W., Liu, C., et al. 2012, The Astrophysical Journal, 753, 148, doi: 10.1088/0004-637x/753/2/148

Bowen, I. S., & Vaughan, A. H. 1973, Appl. Opt., 12, 1430, doi: 10.1364/AO.12.001430

Bower, R. G., Vernon, I., Goldstein, M., et al. 2010, Monthly Notices of the Royal Astronomical Society, 407, 2017, doi: 10.1111/j.1365-2966.2010.16991.x

Brown, T. B., Mann, B., Ryder, N., et al. 2020, in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, ed. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, Astronomy & Astrophysics, 564, A125, doi: 10.1051/0004-6361/201322971

Buckle, J. V., Hills, R. E., Smith, H., et al. 2009, MNRAS, 399, 1026, doi: 10.1111/j.1365-2966.2009.15347.x

Buder, S., Sharma, S., Kos, J., et al. 2021, MNRAS, 506, 150, doi: 10.1093/mnras/stab1242

Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&a, 618, A93, doi: 10.1051/0004-6361/201833476

Casali, G., Spina, L., Magrini, L., et al. 2020, A&A, 639, A127, doi: 10.1051/0004-6361/202038055

Casey, A. R., Hogg, D. W., Ness, M., et al. 2016, arXiv e-prints, arXiv:1603.03040. https://arxiv.org/abs/1603.03040

Casey, A. R., Hawkins, K., Hogg, D. W., et al. 2017, ApJ, 840, 59, doi: 10.3847/1538-4357/aa69c2

Casey, A. R., Ho, A. Y. Q., Ness, M., et al. 2019, ApJ, 880, 125, doi: 10.3847/1538-4357/ab27bf

Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&a, 618, A59, doi: 10.1051/0004-6361/201833390

—. 2020, A&a, 635, A45, doi: 10.1051/0004-6361/201937386

Chen, J., Konrad, J., & Ishwar, P. 2019, CoRR, abs/1906.09313. https://arxiv.org/abs/1906.09313

Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: ACM), 785–794, doi: 10.1145/2939672.2939785

Chen, X., Duan, Y., Houthooft, R., et al. 2016, in Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, ed. D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett, 2172–2180. https://proceedings.neurips.cc/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html

Cheng, C. M., Price-Jones, N., & Bovy, J. 2021, Monthly Notices of the Royal Astronomical Society, 506, 5573, doi: 10.1093/mnras/stab2106

Chopra, S., Hadsell, R., & LeCun, Y. 2005, in Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, Cvpr '05 (Usa: IEEE Computer Society), 539–546, doi: 10.1109/cvpr.2005.202

Cichocki, A., & Phan, A.-H. 2009, IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, 92, 708, doi: 10.1587/transfun.E92.A.708

Collings, M. P., Anderson, M. A., Chen, R., et al. 2004, Monthly Notices of the Royal Astronomical Society, 354, 1133, doi: 10.1111/j.1365-2966.2004.08272.x

Colombo, D., Sanchez, S. F., Bolatto, A. D., et al. 2020, A&a, 644, A97, doi: 10.1051/0004-6361/202039005

Coronado, J., Rix, H.-W., Trick, W. H., et al. 2020, Monthly Notices of the Royal Astronomical Society, 495, 4098, doi: 10.1093/mnras/staa1358

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Research in Astronomy and Astrophysics, 12, 1197, doi: 10.1088/1674-4527/12/9/003

Cybenko, G. 1989, Mathematics of Control, Signals, and Systems (MCSS), 2, 303, doi: 10.1007/bf02551274

Das, P., & Sanders, J. L. 2019, MNRAS, 484, 294, doi: 10.1093/mnras/sty2776

de Jong, R. S., Barden, S. C., Bellido-Tirado, O., et al. 2016, in Ground-based and Airborne Instrumentation for Astronomy VI, ed. C. J. Evans, L. Simard, & H. Takami, Vol. 9908, International Society for Optics and Photonics (Spie), 473 – 490, doi: 10.1117/12.2232832

de Mijolla, D., Frye, C., Kunesch, M., Mansir, J., & Feige, I. 2020, CoRR, abs/2010.07384. https://arxiv.org/abs/2010.07384

de Mijolla, D., & Ness, M. K. 2021, Measuring chemical likeness of stars with RSCA. https://arxiv.org/abs/2110.02250

de Mijolla, D., Ness, M. K., Viti, S., & Wheeler, A. J. 2021, The Astrophysical Journal, 913, 12, doi: 10.3847/1538-4357/abece1

de Mijolla, D., Viti, S., Holdship, J., Manolopoulou, I., & Yates, J. 2019, A&a, 630, A117, doi: 10.1051/0004-6361/201935973

De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604, doi: 10.1093/mnras/stv327

Donor, J., Frinchaboy, P. M., Cunha, K., et al. 2020, The Astronomical Journal, 159, 199, doi: 10.3847/1538-3881/ab77bc

Dotter, A., Conroy, C., Cargile, P., & Asplund, M. 2017, ApJ, 840, 99, doi: 10.3847/1538-4357/aa6d10

Draine, B. 2003, Annual Review of Astronomy and Astrophysics, 41, 241, doi: 10.1146/annurev.astro.41.011802.094840

Draine, B. T. 1978, The Astrophysical Journal Supplement Series, 36, 595, doi: 10.1086/190513

Draine, B. T., & Bertoldi, F. 1996, Astrophysical Journal v.468, p.269, 468, 269, doi: 10.1086/177689

Dyson, J. E. J. E., & Williams, D. A. 1997, The physics of the interstellar medium (Institute of Physics Pub), 165. https://books.google.co.uk/books/about/The{_}Physics{_}of{_}the{_}Interstellar{_}Medium{_}S.html?id=k7x3Es3OrV8C

Edwards, H., & Storkey, A. J. 2016, in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun. http://arxiv.org/abs/1511.05897

Elyajouri, M., Monreal-Ibero, A., Remy, Q., & Lallement, R. 2016, The Astrophysical Journal Supplement Series, 225, 19, doi: 10.3847/0067-0049/225/2/19

Elyajouri, M., Lallement, R., Monreal-Ibero, A., Capitanio, L., & Cox, N. L. J. 2017, A&a, 600, A129, doi: 10.1051/0004-6361/201630088

Feng, Y., & Krumholz, M. R. 2014, Nature, 513, 523, doi: 10.1038/nature13662

Feroz, F., Hobson, M. P., & Bridges, M. 2009, Monthly Notices of the Royal Astronomical Society, 398, 1601, doi: 10.1111/j.1365-2966.2009.14548.x

Feuillet, D. K., Frankel, N., Lind, K., et al. 2019, MNRAS, 489, 1742, doi: 10.1093/mnras/stz2221

Fluke, C. J., & Jacobs, C. 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349, doi: https://doi.org/10.1002/widm.1349

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 24, doi: 10.21105/joss.00024

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306, doi: 10.1086/670067

Foschino, S., Berné, O., & Joblin, C. 2019, A&a, 632, A84, doi: 10.1051/0004-6361/201935085

Frankel, N., Rix, H.-W., Ting, Y.-S., Ness, M., & Hogg, D. W. 2018, The Astrophysical Journal, 865, 96, doi: 10.3847/1538-4357/aadba5

Freeman, K., & Bland-Hawthorn, J. 2002, Ann. Rev. Astr. Astrophys., 40, 487, doi: 10.1146/annurev.astro.40.060401.093840

Friel, E. D. 1995, Ann. Rev. Astr. Astrophys., 33, 381, doi: 10.1146/annurev.aa.33.090195.002121

Frye, C., de Mijolla, D., Begley, T., et al. 2021, in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (OpenReview.net). https://openreview.net/forum?id=OPyWRrcjVQw

Fu, X., Huang, K., & Sidiropoulos, N. D. 2018, IEEE Signal Processing Letters, 25, 328

Fulvio, D., Góbi, S., Jäger, C., Kereszturi, Á., & Henning, T. 2017, 233, 14, doi: 10.3847/1538-4365/aa9224

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018a, A&A, 616, A1, doi: 10.1051/0004-6361/201833051

Gaia Collaboration, Babusiaux, C., van Leeuwen, F., et al. 2018b, A&A, 616, A10, doi: 10.1051/0004-6361/201832843

Ganin, Y., Ustinova, E., Ajakan, H., et al. 2016, J. Mach. Learn. Res., 17, 2096. http://dl.acm.org/citation.cfm?id=2946645.2946704

Gao, X.-H. 2014, Research in Astronomy and Astrophysics, 14, 159, doi: 10.1088/1674-4527/14/2/004

García-Burillo, S., Combes, F., Usero, A., et al. 2014, Astronomy & Astrophysics, 567, A125, doi: 10.1051/0004-6361/201423843

García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., et al. 2016, AJ, 151, 144, doi: 10.3847/0004-6256/151/6/144

Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25

Glorot, X., & Bengio, Y. 2010, in Proceedings of Machine Learning Research, Vol. 9, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, ed. Y. W. Teh & M. Titterington (Chia Laguna Resort, Sardinia, Italy: Pmlr), 249–256. https://proceedings.mlr.press/v9/glorot10a.html

Godard, B., Falgarone, E., Gerin, M., Hily-Blant, P., & De Luca, M. 2010, Astronomy and Astrophysics, 520, A20, doi: 10.1051/0004-6361/201014283

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. 2004, in Proceedings of the 17th International Conference on Neural Information Processing Systems, Nips'04 (Cambridge, MA, USA: MIT Press), 513–520

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (The MIT Press)

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al. 2014, in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, ed. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger, 2672–2680. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

Grassi, T., Merlin, E., Piovan, L., Buonomo, U., & Chiosi, C. 2011. https://arxiv.org/abs/1103.0509

Gratier, Pierre, Bron, Emeric, Gerin, Maryvonne, et al. 2017, A&a, 599, A100, doi: 10.1051/0004-6361/201629847

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. 2017, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, ed. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett, 5767–5777. https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, The Astronomical Journal, 131, 2332, doi: 10.1086/500975

Hadad, N., Wolf, L., & Shahar, M. 2018, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018 (IEEE Computer Society), 772–780, doi: 10.1109/cvpr.2018.00087

Harada, N., Nishimura, Y., Watanabe, Y., et al. 2019, The Astrophysical Journal, 871, 238, doi: 10.3847/1538-4357/aaf72a

Hawkins, K., & Wyse, R. F. G. 2018, MNRAS, 481, 1028, doi: 10.1093/mnras/sty2282

Hawkins, K., Lucey, M., Ting, Y.-S., et al. 2020, Monthly Notices of the Royal Astronomical Society, 492, 1164, doi: 10.1093/mnras/stz3132

Hayden, M. R., Bovy, J., Holtzman, J. A., et al. 2015, The Astrophysical Journal, 808, 132, doi: 10.1088/0004-637x/808/2/132

Haywood, Misha, Di Matteo, Paola, Lehnert, Matthew D., Katz, David, & Gómez, Ana. 2013, A&a, 560, A109, doi: 10.1051/0004-6361/201321397

Henghes, B., Pettitt, C., Thiyagalingam, J., Hey, T., & Lahav, O. 2021, Monthly Notices of the Royal Astronomical Society, 505, 4847–4856, doi: 10.1093/mnras/stab1513

Higgins, I., Matthey, L., Pal, A., et al. 2017, in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (OpenReview.net). https://openreview.net/forum?id=Sy2fzU9gl

Ho, A. Y. Q., Ness, M. K., Hogg, D. W., et al. 2017, ApJ, 836, 5, doi: 10.3847/1538-4357/836/1/5

Hochreiter, S., & Schmidhuber, J. 1997, Neural Computation, 9, 1735

Hogg, D. W., Casey, A. R., Ness, M., et al. 2016, The Astrophysical Journal, 833, 262, doi: 10.3847/1538-4357/833/2/262

Holdship, J., Jeffrey, N., Makrymallis, A., Viti, S., & Yates, J. 2018, 866, 116, doi: 10.3847/1538-4357/aae1fa

Holdship, J., Viti, S., Jiménez-Serra, I., Makrymallis, A., & Priestley, F. 2017, The Astronomical Journal, 154, 38, doi: 10.3847/1538-3881/aa773f

Holdship, J., Jimenez-Serra, I., Viti, S., et al. 2019, The Astrophysical Journal, 878, 64, doi: 10.3847/1538-4357/ab1cb5

Holtzman, J. A., Shetrone, M., Johnson, J. A., et al. 2015, AJ, 150, 148, doi: 10.1088/0004-6256/150/5/148

Imanishi, M., Nakanishi, K., & Izumi, T. 2018, The Astrophysical Journal, 856, 143, doi: 10.3847/1538-4357/aab42f

Ioffe, S., & Szegedy, C. 2015, in Proceedings of Machine Learning Research, Vol. 37, Proceedings of the 32nd International Conference on Machine Learning, ed. F. Bach & D. Blei (Lille, France: Pmlr), 448–456. https://proceedings.mlr.press/v37/ioffe15.html

James, T. A., Viti, S., Yusef-Zadeh, F., Royster, M., & Wardle, M. 2021, The Astrophysical Journal, 916, 69, doi: 10.3847/1538-4357/abfd99

Jha, A. H., Anand, S., Singh, M., & Veeravasarapu, V. S. R. 2018, in Lecture Notes in Computer Science, Vol. 11207, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III, ed. V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Springer), 829–845, doi: 10.1007/978-3-030-01219-9_49

Jofré, P., Das, P., Bertranpetit, J., & Foley, R. 2017, MNRAS, 467, 1140, doi: 10.1093/mnras/stx075

Jofré, P., Heiter, U., & Soubiran, C. 2019, Annual Review of Astronomy and Astrophysics, 57, 571, doi: 10.1146/annurev-astro-091918-104509

Johnson, J. A. 2019, Science, 363, 474, doi: 10.1126/science.aau9540

Johnston, M. D. 1989, Knowledge based telescope scheduling, ed. A. Heck & F. Murtagh (Berlin, Heidelberg: Springer Berlin Heidelberg), 33–49, doi: 10.1007/3-540-51044-3_15

Jönsson, H., Prieto, C. A., Holtzman, J. A., et al. 2018, The Astronomical Journal, 156, 126, doi: 10.3847/1538-3881/aad4f5

Jönsson, H., Holtzman, J. A., Prieto, C. A., et al. 2020, The Astronomical Journal, 160, 120, doi: 10.3847/1538-3881/aba592

Juvela, M., Lehtinen, K., & Paatero, P. 1996a, MNRAS, 280, 616, doi: 10.1093/mnras/280.3.616

—. 1996b, MNRAS, 280, 616, doi: 10.1093/mnras/280.3.616

Kalberla, P. M. W., & Kerp, J. 2009, Ann. Rev. Astr. Astrophys., 47, 27, doi: 10.1146/annurev-astro-082708-101823

Kamdar, H., Conroy, C., Ting, Y.-S., et al. 2019, The Astrophysical Journal, 884, 173, doi: 10.3847/1538-4357/ab44be

Kamdar, H., Conroy, C., Ting, Y.-S., & El-Badry, K. 2020, arXiv e-prints, arXiv:2007.10990. https://arxiv.org/abs/2007.10990

Kamenetzky, J., Privon, G. C., & Narayanan, D. 2018, The Astrophysical Journal, 859, 9, doi: 10.3847/1538-4357/aab3e2

Kauffmann, Jens, Goldsmith, Paul F., Melnick, Gary, et al. 2017, A&a, 605, L5, doi: 10.1051/0004-6361/201731123

Kingma, D. P., & Ba, J. 2015, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun. http://arxiv.org/abs/1412.6980

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. 2017, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 971–980. http://papers.nips.cc/paper/6698-self-normalizing-neural-networks

Klessen, R., & Glover, S. 2014, 43, doi: 10.1007/978-3-662-47890-5_2

Kollmeier, J. A., Zasowski, G., Rix, H.-W., et al. 2017, arXiv e-prints, arXiv:1711.03234. https://arxiv.org/abs/1711.03234

Kollmeier, J. A., Zasowski, G., Rix, H.-W., et al. 2017, SDSS-V: Pioneering Panoptic Spectroscopy

Krips, M., Martín, S., Eckart, A., et al. 2011, The Astrophysical Journal, 736, 37, doi: 10.1088/0004-637x/736/1/37

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, ed. P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger, 1106–1114. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Krumholz, M. R., McKee, C. F., & Bland -Hawthorn, J. 2019, Ann. Rev. Astr. Astrophys., 57, 227, doi: 10.1146/annurev-astro-091918-104430

Kwan, J., Heitmann, K., Habib, S., et al. 2015, The Astrophysical Journal, 810, 35, doi: 10.1088/0004-637x/810/1/35

Lahav, O., Nairn, A., Sodré, L., J., & Storrie-Lombardi, M. C. 1996, Monthly Notices of the Royal Astronomical Society, 283, 207, doi: 10.1093/mnras/283.1.207

Lample, G., Zeghidour, N., Usunier, N., et al. 2017, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, ed. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett, 5967–5976. https://proceedings.neurips.cc/paper/2017/hash/3fd60983292458bf7dee75f12d5e9e05-Abstract.html

Langer, W. D., & Glassgold, A. E. 1990, ApJ, 352, 123, doi: 10.1086/168519

Lebouteiller, V., & Kunth, D. 2005, in Starbursts, ed. R. De Grijs & R. M. González Delgado (Dordrecht: Springer Netherlands), 247–250

LeCun, Y., Boser, B., Denker, J. S., et al. 1989, Neural Comput., 1, 541–551, doi: 10.1162/neco.1989.1.4.541

Lee, D. D., & Seung, H. S. 1999, Nature, 401, 788

Leistedt, B., & Hogg, D. W. 2017, The Astrophysical Journal, 838, 5, doi: 10.3847/1538-4357/aa6332

Leung, H. W., & Bovy, J. 2018, Monthly Notices of the Royal Astronomical Society, 483, 3255, doi: 10.1093/mnras/sty3217

Lezama, J. 2019, in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (OpenReview.net). https://openreview.net/forum?id=Hkg4W2AcFm

Linnainmaa, S. 1976, BIT Numerical Mathematics, 16, 146

Linnartz, H., Ioppolo, S., & Fedoseev, G. 2015, International Reviews in Physical Chemistry, 34, 205

Liu, F., Asplund, M., Yong, D., et al. 2019, A&A, 627, A117, doi: 10.1051/0004-6361/201935306

Lloyd, S. P. 1982, IEEE Transactions on Information Theory, 28, 129

Lo, N., Cunningham, M. R., Jones, P. A., et al. 2009, MNRAS, 395, 1021, doi: 10.1111/j.1365-2966.2009.14594.x

Locatello, F., Bauer, S., Lucic, M., et al. 2019, in Proceedings of Machine Learning Research, Vol. 97, Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, ed. K. Chaudhuri & R. Salakhutdinov (Pmlr), 4114–4124. http://proceedings.mlr.press/v97/locatello19a.html

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. S. 2016, in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, ed. Y. Bengio & Y. LeCun. http://arxiv.org/abs/1511.00830

Lu, Z., Pu, H., Wang, F., Hu, Z., & 0001, L. W. 2017, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, ed. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett, 6232–6240. http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width

Mackereth, J. T., Bovy, J., Leung, H. W., et al. 2019, MNRAS, 489, 176, doi: 10.1093/mnras/stz1521

Maffucci, D. M., Wenger, T. V., Gal, R. L., & Herbst, E. 2018, The Astrophysical Journal, 868, 41, doi: 10.3847/1538-4357/aae70c

Magrini, L., Randich, S., Kordopatis, G., et al. 2017, A&a, 603, A2, doi: 10.1051/0004-6361/201630294

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94, doi: 10.3847/1538-3881/aa784d

Martell, S. L., Shetrone, M. D., Lucatello, S., et al. 2016, ApJ, 825, 146, doi: 10.3847/0004-637x/825/2/146

Martig, M., Fouesneau, M., Rix, H.-W., et al. 2016, MNRAS, 456, 3655, doi: 10.1093/mnras/stv2830

Masseron, T., & Gilmore, G. 2015, MNRAS, 453, 1855, doi: 10.1093/mnras/stv1731

Matheron, G. 1963, Economic Geology, 58, 1246, doi: 10.2113/gsecongeo.58.8.1246

McElroy, D., Walsh, C., Markwick, A. J., et al. 2013, Astronomy & Astrophysics, 550, A36, doi: 10.1051/0004-6361/201220465

McKay, M. D., Beckman, R. J., & Conover, W. J. 1979, Technometrics, 21, 239, doi: 10.2307/1268522

Meijerink, R., Spaans, M., & Israel, F. P. 2007, Astronomy & Astrophysics, 461, 793, doi: 10.1051/0004-6361:20066130

Melchior, P., Moolekamp, F., Jerdee, M., et al. 2018, Astronomy and Computing, 24, 129, doi: 10.1016/j.ascom.2018.07.001

Mészáros, S., Allende Prieto, C., Edvardsson, B., et al. 2012, The Astronomical Journal, 144, 120, doi: 10.1088/0004-6256/144/4/120

Michiyama, T., Iono, D., Sliwa, K., et al. 2018, The Astrophysical Journal, 868, 95, doi: 10.3847/1538-4357/aae82a

Mondal, A., Das, R., Shaw, G., & Mondal, S. 2019, Monthly Notices of the Royal Astronomical Society, 483, 4884, doi: 10.1093/mnras/sty3361

Moran, G. E., Sridhar, D., Wang, Y., & Blei, D. M. 2022, Identifiable Deep Generative Models via Sparse Decoding. https://arxiv.org/abs/2110.10804

Murphy, K. P. 2013, Machine learning : a probabilistic perspective (Cambridge, Mass. [u.a.]: MIT Press). https://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8&qid=1336857747&sr=8-2

—. 2021, Probabilistic Machine Learning: An introduction (MIT Press). probml.ai

Nentwich, C., & Engell, S. 2016, in 2016 International Joint Conference on Neural Networks (IJCNN), 1291–1296, doi: 10.1109/IJCNN.2016.7727346

Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, The Astrophysical Journal, 808, 16, doi: 10.1088/0004-637x/808/1/16

Ness, M., Hogg, D. W., Rix, H. W., et al. 2016, ApJ, 823, 114, doi: 10.3847/0004-637x/823/2/114

Ness, M., Rix, H.-W., Hogg, D. W., et al. 2018, The Astrophysical Journal, 853, 198, doi: 10.3847/1538-4357/aa9d8e

Ness, M. K., Johnston, K. V., Blancato, K., et al. 2019, ApJ, 883, 177, doi: 10.3847/1538-4357/ab3e3c

Neufeld, D. A., Hollenbach, D. J., Kaufman, M. J., et al. 2007, ApJ, 664, 890, doi: 10.1086/518857

O'Briain, T., Ting, Y.-S., Fabbro, S., et al. 2021, The Astrophysical Journal, 906, 130, doi: 10.3847/1538-4357/abca96

Occhiogrosso, A., Vasyunin, A., Herbst, E., et al. 2014, A&a, 564, A123, doi: 10.1051/0004-6361/201322598

Papadopoulos, P. P. 2007, The Astrophysical Journal, 656, 792, doi: 10.1086/510186

Paszke, A., Gross, S., Chintala, S., et al. 2017, in NIPS Autodiff Workshop

Peraiah, A. 2001, An Introduction to Radiative Transfer: Methods and Applications in Astrophysics (Cambridge University Press), doi: 10.1017/cbo9781139164474

Petersen, K. B., & Pedersen, M. S. 2008, The Matrix Cookbook, Technical University of Denmark. http://www2.imm.dtu.dk/pubdb/p.php?3274

Plez, B. 2012, Turbospectrum: Code for spectral synthesis. http://ascl.net/1205.004

Polykovskiy, D., Zhebrak, A., Vetrov, D., et al. 2018, Molecular pharmaceutics, 15, 4398

Poovelil, V. J., Zasowski, G., Hasselquist, S., et al. 2020, The Astrophysical Journal, 903, 55, doi: 10.3847/1538-4357/abb93e

Portegies Zwart, S. F., McMillan, S. L., & Gieles, M. 2010, Annual Review of Astronomy and Astrophysics, 48, 431, doi: 10.1146/annurev-astro-081309-130834

Price-Jones, N., & Bovy, J. 2017a, Monthly Notices of the Royal Astronomical Society, 475, 1410, doi: 10.1093/mnras/stx3198

—. 2017b, Monthly Notices of the Royal Astronomical Society, 475, 1410, doi: 10.1093/mnras/stx3198

—. 2019, Monthly Notices of the Royal Astronomical Society, 487, 871, doi: 10.1093/mnras/stz1260

Price-Jones, N., Bovy, J., Webb, J. J., et al. 2020, arXiv e-prints, arXiv:2004.04263. https://arxiv.org/abs/2004.04263

Randich, S., Gilmore, G., & Gaia-ESO Consortium. 2013, The Messenger, 154, 47

Rawlings, J. M. C., Hartquist, T. W., Menten, K. M., & Williams, D. A. 1992, Monthly Notices of the Royal Astronomical Society, 255, 471, doi: 10.1093/mnras/255.3.471

Roberts, J. F., Rawlings, J. M. C., Viti, S., & Williams, D. A. 2007, Monthly Notices of the Royal Astronomical Society, 382, 733, doi: 10.1111/j.1365-2966.2007.12402.x

Rosenblatt, F. 1958, Psychological Review, 65, 386, doi: 10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, Nature, 323, 533, doi: 10.1038/323533a0

Salak, D., Tomiyasu, Y., Nakai, N., et al. 2018, The Astrophysical Journal, 856, 97, doi: 10.3847/1538-4357/aab2ac

Savage, B. D., & Mathis, J. S. 1979, Ann. Rev. Astr. Astrophys., 17, 73, doi: 10.1146/annurev.aa.17.090179.000445

Schiavon, R. P., Johnson, J. A., Frinchaboy, P. M., et al. 2017, MNRAS, 466, 1010, doi: 10.1093/mnras/stw3093

Schmidhuber, J. 1991, Learning Factorial Codes By Predictability Minimization, Tech. Rep. Cu-cs-565-91, Dept. of Comp. Sci., University of Colorado at Boulder

Schmidt, M. N., Winther, O., & Hansen, L. K. 2009, in Independent Component Analysis and Signal Separation, ed. T. Adali, C. Jutten, J. M. T. Romano, & A. K. Barros (Berlin, Heidelberg: Springer Berlin Heidelberg), 540–547

Schmit, C. J., & Pritchard, J. R. 2017, Monthly Notices of the Royal Astronomical Society, 475, 1213, doi: 10.1093/mnras/stx3292

—. 2018, Monthly Notices of the Royal Astronomical Society, 475, 1213, doi: 10.1093/mnras/stx3292

Schöier, F. L., van der Tak, F. F. S., van Dishoeck, E. F., & Black, J. H. 2005, Astronomy & Astrophysics, 432, 369, doi: 10.1051/0004-6361:20041729

Senior, A. W., Evans, R., Jumper, J., et al. 2020, Nature, 577, 706, doi: 10.1038/s41586-019-1923-7

Shematovich, V. I. 2012, Solar System Research, 46, 391, doi: 10.1134/s0038094612060068

Shental, N., Hertz, T., Weinshall, D., & Pavel, M. 2002, in Lecture Notes in Computer Science, Vol. 2353, Computer Vision - ECCV 2002, 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28-31, 2002, Proceedings, Part IV, ed. A. Heyden, G. Sparr, M. Nielsen, & P. Johansen (Springer), 776–792, doi: 10.1007/3-540-47979-1_52

Silver, D., Schrittwieser, J., Simonyan, K., et al. 2017, Nature, 550, 354. http://dx.doi.org/10.1038/nature24270

Simpson, J. D., Martell, S. L., Da Costa, G., et al. 2019, MNRAS, 482, 5302, doi: 10.1093/mnras/sty3042

Skilling, J. 2006, Bayesian Analysis, 1, 833, doi: 10.1214/06-ba127

Souto, D., Prieto, C. A., Cunha, K., et al. 2019, The Astrophysical Journal, 874, 97, doi: 10.3847/1538-4357/ab0b43

Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645, doi: 10.1086/506564

Storrie-Lombardi, M. C., Lahav, O., Sodré, L., J., & Storrie-Lombardi, L. J. 1992, Monthly Notices of the Royal Astronomical Society, 259, 8P, doi: 10.1093/mnras/259.1.8P

Tamura, N., Takato, N., Shimono, A., et al. 2016, in Ground-based and Airborne Instrumentation for Astronomy VI, ed. C. J. Evans, L. Simard, & H. Takami, Vol. 9908, International Society for Optics and Photonics (Spie), 456 – 472, doi: 10.1117/12.2232103

Tielens, A. G. G. M. 2013, Rev. Mod. Phys., 85, 1021, doi: 10.1103/RevModPhys.85.1021

Ting, Y.-S., Conroy, C., & Rix, H.-W. 2016, ApJ, 816, 10, doi: 10.3847/0004-637x/816/1/10

Ting, Y.-S., Conroy, C., Rix, H.-W., & Cargile, P. 2017, The Astrophysical Journal, 843, 32, doi: 10.3847/1538-4357/aa7688

—. 2019, The Astrophysical Journal, 879, 69, doi: 10.3847/1538-4357/ab2331

Ting, Y.-S., Freeman, K. C., Kobayashi, C., De Silva, G. M., & Bland-Hawthorn, J. 2012, MNRAS, 421, 1231, doi: 10.1111/j.1365-2966.2011.20387.x

Ting, Y.-S., & Weinberg, D. H. 2021, arXiv e-prints, arXiv:2102.04992. https://arxiv.org/abs/2102.04992

Tunnard, R., & Greve, T. R. 2016, The Astrophysical Journal, 819, 161, doi: 10.3847/0004-637x/819/2/161

Tunnard, R., Greve, T. R., Garcia-Burillo, S., et al. 2015, The Astrophysical Journal, 815, 114, doi: 10.1088/0004-637x/815/2/114

Valenti, J. A., & Fischer, D. A. 2005, ApJSS, 159, 141, doi: 10.1086/430500

van der Tak, F. F. S., Black, J. H., Schöier, F. L., Jansen, D. J., & van Dishoeck, E. F. 2007, Astronomy & Astrophysics, 468, 627, doi: 10.1051/0004-6361:20066820

van Dishoeck, E. F. 2017, Proceedings of the International Astronomical Union, 13, 3–22, doi: 10.1017/s1743921317011528

Vinod, N., & Hinton, G. 2010, in Proceedings of the 27th International Conference on International Conference on Machine Learning (Association for Computing Machinery), 170. https://dl.acm.org/citation.cfm?id=3104425

Viti, S. 2017, Astronomy & Astrophysics, 607, A118, doi: 10.1051/0004-6361/201628877

Viti, S., Collings, M. P., Dever, J. W., McCoustra, M. R. S., & Williams, D. A. 2004, Monthly Notices of the Royal Astronomical Society, 354, 1141, doi: 10.1111/j.1365-2966.2004.08273.x

Viti, S., & Williams, D. A. 1999, Monthly Notices of the Royal Astronomical Society, 305, 755, doi: 10.1046/j.1365-8711.1999.02447.x

Viti, S., García-Burillo, S., Fuente, A., et al. 2014, Astronomy & Astrophysics, 570, A28, doi: 10.1051/0004-6361/201424116

Wakelam, V., Herbst, E., Loison, J.-C., et al. 2012, The Astrophysical Journal Supplement Series, 199, 21, doi: 10.1088/0067-0049/199/1/21

Wakelam, V., Bron, E., Cazaux, S., et al. 2017, Molecular Astrophysics, 9, 1, doi: https://doi.org/10.1016/j.molap.2017.11.001

Walmsley, M., Smith, L., Lintott, C., et al. 2019, Monthly Notices of the Royal Astronomical Society, 491, 1554, doi: 10.1093/mnras/stz2816

Weinberg, D. H., Holtzman, J. A., Hasselquist, S., et al. 2019, ApJ, 874, 102, doi: 10.3847/1538-4357/ab07c7

Weinberg, D. H., Holtzman, J. A., Johnson, J. A., et al. 2021, Chemical Cartography with APOGEE: Mapping Disk Populations with a Two-Process Model and Residual Abundances. https://arxiv.org/abs/2108.08860

Weinberger, K. Q., & Saul, L. K. 2009, J. Mach. Learn. Res., 10, 207–244

Wheeler, A., Ness, M., Buder, S., et al. 2020, ApJ, 898, 58, doi: 10.3847/1538-4357/ab9a46

Wheeler, A. J., Hogg, D. W., & Ness, M. 2021, The Astrophysical Journal, 908, 247, doi: 10.3847/1538-4357/abd544

Whitmore, B. C. 1984, ApJ, 278, 61, doi: 10.1086/161768

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, Monthly Notices of the Royal Astronomical Society, 435, 2835, doi: 10.1093/mnras/stt1458

Williams, D. A., & Viti, S. 2013, Observational molecular astronomy: exploring the universe using molecular line emissions (Cambridge University Press)

Wilson, J. C., Hearty, F. R., Skrutskie, M. F., et al. 2019, Publications of the Astronomical Society of the Pacific, 131, 055001, doi: 10.1088/1538-3873/ab0075

Wootten, A., & Thompson, A. 2009, Proceedings of the IEEE, 97, 1463, doi: 10.1109/jproc.2009.2020572

Xu, R., & Bai, X.-N. 2016, The Astrophysical Journal, 819, 68, doi: 10.3847/0004-637x/819/1/68

Young, J. S., & Scoville, N. Z. 1991, Annual Review of Astronomy and Astrophysics, 29, 581, doi: 10.1146/annurev.aa.29.090191.003053