UNIVERSITY COLLEGE LONDON

Faculty of Mathematics and Physical Sciences

Department of Physics & Astronomy

STATISTICAL ANALYSES OF ASTROCHEMICAL BIG DATA

Thesis submitted for the Degree of Doctor of Philosophy of the University of London

by

Marcus Keil

Supervisors:

Prof. Serena Viti

Dr. Jeremy Yates

Examiners:

Prof. Benjamin Joachimi

Prof. Chris Lintott

May 2025

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

Abstract

This thesis presents original research and development of statistical tools for astrochemical inferences of observations using modelling tools. Funded through the AstroChemical Origin (ACO) Innovative Training Network (ITN), the work of this thesis specialises in the intersection of statistical software engineering and astrochemistry to complement the interdisciplinary research being conducted in the ACO ITN, which includes work in radio instrumentation, astronomical observations, and astrochemical modelling, as well as experimental and theoretical chemistry.

The work in chapter 2 describes the statistical inference tool UCLCHEMCMC, developed so that observers could perform advanced modelling of their molecular observations without needing detailed knowledge on how to model. The approach uses a database in order to reduce computational time while also presenting future opportunities to train machine learning algorithms with the collected data.

Chapter 3 describes the work done using UCLCHEMCMC in order to study which physical parameters influence the observable sulphur in the interstellar medium the most. This work aimed to not only study why sulphur is not observed at the expected abundances but also to showcase the broad applications of the UCLCHEMCMC tool beyond the standard inferences.

The penultimate chapter, chapter 4, describes the ACO outreach project, which was made to be a virtual reality experience. The experience is aimed at encouraging secondary school students to learn more about astrochemistry, specifically the journey of water, in order to help inspire the next generation of researchers. The development of this project follows more closely to that of a video game being developed, as education through entertainment will create a desire for learning in students.

The last chapter describes future projects already planned, some of which are already being worked on. It starts by describing the changes that need to be made to UCLCHEMCMC in order to allow an online version to be available to the public,

rather than just downloading the source code. The second part of the last chapter describes the prototype dashboard being developed for astronomical archival data, its potential challenges, and uses.

Appendix A describes work that is not directly related to astrophysics but has applications to astrophysics. This work was performed in a healthcare research group with the aim to set up a data storage solution and data dashboard that would aid a healthcare research group to improve their methods for research. The development of a data storage solution and dashboard in a healthcare setting inspired work to be completed in the future to create a dashboard for astronomical survey data.

Impact Statement

The work performed for this thesis presents useful developments and understanding both for the astrochemical community, for educational purposes and for professional development outside of academia.

The work in chapter 2 provides the astrochemical community with a useful, open-source statistical inference tool called UCLCHEMCMC. This tool allows observers to leverage the power of astrochemical modelling developments. It is supported by a SQL database in such a way that it also becomes more efficient with use. This kind of approach presents itself in a way that is novel and useful for future academic uses or developments. Approaching modelling in a way so as to store models, rather than discard them, will aid in future training of machine learning algorithms.

Chapter 3 describes work to further the understanding of sulphur bearing species and what physics impacts their abundances the most. Showing how UCLCHEMCMC can be fit with other modelling software to suit the needs of various users based on what modelling codes they prefer, how UCLCHEMCMC can be used this way, and studying the sulphur problem using it, provides the astrochemical community with the opportunity to understand the tool better and to aid in understanding what is impacting the sulphur abundance the most.

The VR project, described in chapter 4, provides the astrochemical community with a fun, interactive experience for secondary school students, to help inspire future scientists to study astrophysics, or chemistry. The impact of this project can have far reaching effects, as the VR headset was developed to be easy to transport, and can be set up in minutes at any conventions, outreach events or educational experiences where secondary school students would participate. Combining this with researchers that are well versed in the topics discussed in the experience can provide with an excellent opportunity for the inquisitive mind of a young student to find a new passion.

As appendix A describes work in a health care research group, which is separate from the rest of the work in this thesis, its impact is different from the rest. The database and dashboard created for this research group are aiding in helping future researchers start doing research in their field faster than previously, as they have easy access to a database containing already processed data, whereas previously they would have to create their own processed data documents containing the information they need. This means that it will save initial preparatory time, allowing future researchers to start producing useful work faster. Additionally, the dashboard provides an easy to understand interface that allows researchers to easily make arguments to support funding redistribution requests to improve health care outcomes for patients.

Acknowledgements

First and foremost, I wish to thank my supervisor and second supervisor. My supervisor, Serena Viti for her support, guidance and patience as I learned what it meant to be a researcher. The guidance from Serena on how to do research and the freedom she gave me to explore what I would like to do, while still aiding me to stay the course of the research proposed for the position has aided me in finding my own way of being a productive member of the research community. I am also very grateful for the role my second supervisor, Jeremy Yates, has played in my work. From moral and emotional support throughout my time at UCL, to career and research advice and opportunities such as getting to apply my data intensive skills in the healthcare field. My thesis would not have come to completion without either of them.

The support from my friends, both new and old, and that of my family helped me to stay positive throughout this experience and to push forward until the very end. I am also grateful to my good friend Bea, for her support in these trying years and for knowing just how to help me overcome some of the most stressful parts of being a PhD student, I will never forget what you meant to me during this time.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 811312 for the project "Astro-Chemical Origins" (ACO) as well as from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme MOPPEX 833460.

Abbreviations

ACO - AstroChemical Origins Network

ACT - Asthma Control Test

AR - Augmented Reality

CR - Cosmic Ray

CYP - Children and Young People

CYPHP - Children and Young People's Health Partnership

EHR - Electronic Health Record

GP(s) - General Practitioner(s)

GSTT - Guy's and St. Thomas' NHS Foundation Trust

HST - Hubble Space Telescope

IDAC(I) - Income Deprivation Affecting Children (Index)

ISM - Interstellar Medium

ITN - Innovative Training Network(s)

json - JavaScript Object Notation

KCH - Kings College Hospital

KPI(s) - Key Performance Indicator(s)

LAS - Largest Angular Scale

LSOA - Lower Layer Super Output Area

LTE - Local Thermal Equilibrium

MCMC - Markov Chain Monte Carlo MCMC

ODE(s) - Ordinary Differential Equation(s)

PDF - probability Density Function

POEM - Patient-Oriented Eczema Measure

SQL - Structured Query Language

UI - User Interface

UCL - University College London

VR - Virtual Reality

CONTENTS

List of Figures **16**

List of Tables **17**

I Introduction **18**

I.1 Overview 18

I.2 AstroChemical Origins Network 19

I.3 Interstellar Medium 19

I.3.1 Star Formation 24

I.3.2 Astrochemical modelling 27

I.3.3 Radiative transfer modelling 30

I.4 Statistics 33

I.4.1 Bayesian Statistics 33

I.4.2 Markov Chain Monte Carlo 34

I.5 Data Management 38

I.6 This Thesis 40

II UCLCHEMCMC **41**

II.1 Introduction 41

II.2 UCLCHEMCMC 43

II.2.1 Forward modelling 46

II.2.2 MCMC Inference 47

II.2.3 Database 48

II.2.4 Interface 51

II.3 Application 52

II.3.1 Mock Data Inference 55

II.3.2 Inferring parameters of L1544 63

II.4 Summary 65

III SVS 13-A **67**

III.1 Introduction 67

III.2 Methods 70

III.2.1 Observations 70

III.2.2 Astrochemical Modelling 72

III.2.3 Radiative Transfer 73

III.2.4 Comparison of observations and model predictions 74

III.3 Results 77

III.3.1 Verifying UCLCHEMCMC 81

III.4 Summary 83

IV Virtual Reality Experience **85**

IV.1 Introduction 85

IV.1.1 Outreach 85

IV.1.2 Virtual Reality 86

IV.2 Purpose of the VR project 88

IV.3 Tools 89

IV.3.1 Blender 89

IV.3.2 Unity 91

IV.4 The Experience 92

IV.4.1 Creating Water 93

IV.4.2 Building Planets 95

IV.4.3 Bringing Water 97

IV.5 Evaluation 100

IV.5.1 Public Interaction 100

IV.5.2 Reflection 101

IV.5.3 Lessons Learned 106

IV.6 Conclusion 109

V Future Work **110**

V.1 UCLCHEMCMC Improvements 110

V.2 Astronomical Archival Data Dashboards 112

V.2.1 Archival Data Studies 113
V.2.2 Exploration of New Surveys 114

A Healthcare Dashboard **127**

A.1 Overview 127
A.1.1 Healthcare data 128
A.1.2 Dashboards 130
A.2 CYPHP Health Check Data 131
A.3 Data Products by this work 133
A.3.1 Raw CYPHP Data 134
A.3.2 Data Pipeline: Raw Data to Semi-Structured Database 135
A.3.3 Healthcare Dashboard 136
A.4 Conclusion 141

LIST OF FIGURES

I.1 91α and 92α transition for hydrogen, helium and carbon from an HII region. Image from Quireza et al. (2006) 20
I.2 ISM life cycle from Groppi et al. (2009)..... 21
I.3 Simple emcee chains fitting a linear function to a dataset. 35

II.1 Flow Chart of the various processes that happen in UCLCHEMCMC. Green ovals indicate parts that the User interfaces with. Diamonds indicate parts of UCLCHEMCMC where the next step is dependent on options the user specifies. The SQL Database is represented with a cylinder and has been labeled this way for clarity. Arrows to and from the database represent a query of the SQL which then returns the models that match the query. The emulator part, marked with a green box, highlights where emulator codes could sit relative to the rest of the workflow of UCLCHEMCMC; implementation of such emulators is beyond the scope of this work. 42
II.2 Small flow chart showing which parameters go into UCLCHEM and which parameters are taken from UCLCHEM and given to RADEX. 45
II.3 Posterior distribution function of the evaluation run performed on mock data. The histograms represent the PDF of volume density, kinetic temperature, UV field factor, R_{out} and the cosmic ray ionisation rate factor, the colour bar shows the value ranges of the joint distribution functions. The white dashed lines in the joint distributions, and the black dashed lines in the PDFs represent the true value used to create the mock data. 56

II.4 The radiation temperature,T_R, calculated by RADEX against the Energy of the upper state, for the mock data and errors given to UCLCHEMCMC in black, and the data created when using the most likely parameter values from the 1D distributions from the inference of the mock data in blue. Red represents the peak in the joint distribution of the CR ionisation rate and UV radiation field while keeping the remaining parameters as they are for the previous model, while green and fuchsia represent two additional points with values for the CR ionisation rate and UV radiation field values in the elongated distribution of likely values to show why the inference still gave some importance to these values. 57

II.5 Posterior distribution function of the evaluation run performed on the emission lines from OCS only. The histograms represent the PDF of volume density, kinetic temperature, R_{out} and the cosmic ray ionisation rate factor, the colour bar shows the value ranges of the joint distribution functions, while the red dashed line in the PDF is the value with the highest probability. 59

II.6 T_R over T_R of OCS 6-5 against the upper state energy, for emission lines of OCS found in table II.4, compared to the data of the best fit model after running an inference using only the OCS lines. All of the lines fit the observed line ratios quite well. Black represents the real data with error bars, while blue is the best fit model. 60

II.7 T_R against the Energy of the upper state, for all emission lines in table II.4, compared to the data of the best fit model after running the stress test inference. While a couple lines almost fit their observed counterparts, it is clear that UCLCHEMCMC is unable to match all lines at once, which made it settle for a set of parameters, that allow each line to at least get somewhat close to the observations. Black represents the real data with error bars, while blue is the best fit model. 61

II.8 $\log(\chi^2)$ grid for kinetic temperature and volume density using column density from Vastel et al. (2018) for the six species that are used for the MCMC Inference. The lower value of the $\log(\chi^2)$ is fixed at 0 to allow for better comparison between each species while allowing a flexible upper end, as the large ranges of $\log(\chi^2)$ values make it difficult to make an informative Figure with a single range of values. 62

III.1 Visualisation of the way in which we perform the chemical modelling. In blue, is the range within which the log_{10}(Density) can be. In red is the kinetic Temperature of the gas. The x-axis represents the time in years that were modelled by UCLCHEM, but does not represent the predicted age of an object. As each model can model for a various amount of time, depending on how long the collapse takes, as well as the time between collapses we intentionally forgo marking any numbers on the x-axis of this plot. 72

III.2 χ^2 grid for each molecule using RADEX. The y-axis is the temperature and x axis is the log of the number density of H. The high temperature low density white corner is a model where RADEX failed to converge. The colour range is the log of χ^2 as the values were too high to be able to be shown without a log plot. 76

III.3 χ^2 grid for each molecule using GRELVG. The y-axis is the temperature and x axis is the log of the number density of H. The colour range is the log of χ^2 as the values were too high to be able to be shown without a log plot. 77

III.4 Corner plot of the results from UCLCHEMCMC when using RADEX as the radiative transfer code. The colour-bar represents the value of the normalised posterior of the MCMC inference. In order, the parameters are: Density after the first collapse (D_M); Final density (D_F); Initial temperature (T_I); Final temperature (T_F); Time between collapses (t_m); Cosmic Ray ionisation factor; Radius of cloud (R_{out}); Fractional abundance of S (Frac.$_S$), O (Frac.$_O$), and C (Frac.$_C$). 78

III.5 Plot of the walker chains for the RADEX inference. The initial steps were walkers are rapidly varying are discarded when plotting figure III.4. 79

III.6 Corner plot of the results from UCLCHEMCMC when using GRELVG as the radiative transfer code. The colour-bar represents the value of the normalised posterior of the MCMC inference. In order, the parameters are: Density after the first collapse (D_M); Final density (D_I); Initial temperature (T_F); Final temperature (T_F); Time between collapses (t_m); Cosmic Ray ionisation factor; Radius of cloud (R_{out}); Fractional abundance of S ($Frac._S$), O ($Frac._O$), and C ($Frac._C$). 80

III.7 χ^2 grid using RADEX with column density being varied. Each row is a different column density. The colour range maximum was set to one hundred, but is shown in log scale in order to make it easier to distinguish when models have values that would show them to be good fits or when they are not good fits..... 82

IV.1 Screenshot of the blender interface. The central window is the scene, showing the vertices of a cube marked with orange points, with connected lines to each other forming the surfaces that are the faces of a cube. The right hand windows show at the top the hierarchy of all the components in the scene, and the bottom shows the options, and settings for the scene or the selected object, depending on the selected tab. 90

IV.2 Screenshot of the Unity user interface, showing the main menu of the ACO VR Experience as seen in development mode. The largest central window shows the scene being edited. The left shows the hierarchy of all objects within the scene. The window at the bottom is the project window showing all available files created or imported for the project. The right hand window is the inspection window which shows the details of any object selected in the scene, or project window. 91

IV.3 Picture of the first scene from the development view inside the Unity engine..... 93
IV.4 Screenshot of the video scene between the first and second entertainment sections, depicting a molecular cloud as just a black area blocking the light of the stars behind the cloud. 96
IV.5 Screenshot of the video scene between the first and second entertainment sections, depicting the early stage star with its thick dusty disk surrounding it. This is shown directly before the user is allowed to dive into the disk to form a planet..... 97
IV.6 Screenshot of the second entertainment section where a user is allowed to form a planetoid or planet using dust-grains found within a very dusty disk. The visual impairment caused by the haze is intended to visualise a thick, dense dusty disk. 98
IV.7 Screenshot of the third entertainment section where a user is shown how the dusty disk that did not accrete onto the star or a planet, is blown away. Leaving just the planet and the star. 99
IV.8 Screenshot of the third entertainment section where a user is allowed to create comets and throw them at the planet they created in order to bring enough water to the planet to form oceans like we would have here on earth. 100

A.1 Example image of a dashboard from the Elasticsearch guide for Kibana (elasticsearch, 2015) 131
A.2 Screenshot of the first part of the CYPHP Health check dashboard, using mock data. The first two rows show number of patients as a function of time for A&E visits, out-patient services and admission to hospital, but from separate sources, the top showing GSTT data, and the lower showing KCH data. To the right of that is the number of patient records currently available with any applied filters. The map shows the location of residence for patients. This figure is shown using mock data that does not follow true statistical distributions of healthcare data..... 137

A.3 Screenshot of the second part of the CYPHP Health check dashboard, using mock data. The map is explained in the caption of figure A.2, below that, from left to right, is the distribution of patient ages, the demographic distribution and followed by the distribution of income deprivation affecting children (IDAC) index. On the bottom row is then a distribution showing the distribution of patients suffering of any combination of constipation, asthma or eczema. To the right of that is then a distribution of prescribed medications, and the details of the most prescribed medications. This figure is shown using mock data that does not follow true statistical distributions of healthcare data. 138

A.4 Screenshot of the third part of the CYPHP Health check dashboard, using mock data. The top row is described in the caption of figure A.3. Below that, is a clear representation of one of three nearly identical plots representing the values for asthma eczema and constipation. The pie-chart on the left shows the score for asthma (constipation, eczema) patients received based on the ACT (POEM, eczema) questionnaire given to them. To the right of this, is a heat-map of how many patients responded with what score to which question. This figure is shown using mock data that does not follow true statistical distributions of healthcare data..... 139

A.5 Screenshot of the fourth part of the CYPHP Health check dashboard, using mock data. The pie-chart and heat-map are described in figure A.4. The last row of the dashboard indicated the parental questions that were posed to patients. This bar chart shows how many patients responded to the questions and colour codes them according to the answer they gave. This figure is shown using mock data that does not follow true statistical distributions of healthcare data. 140

LIST OF TABLES

II.1 Inputs and options per page..... 44
II.2 Parameter Ranges..... 52
II.3 Mock Data used for Evaluation 53
II.4 Observations used for evaluation..... 54

III.1 Observations from Codella et al. (2021) that are used as the inputs for the UCLCHEMCMC inferences. (a) Frequencies have been obtained from the Cologne Database for Molecular Spectroscopy (Endres et al., 2016). (b) Errors in the intensity include uncertainties due to calibration of $\leq 10\%$ for the lines with frequencies between $80.2 - 103.9$GHz and $\leq 15\%$ for the remaining emission lines..... 71
III.2 The lower and upper bounds of the parameter range used as inputs to UCLCHEM, as well as how many grid points each parameter has. 71

A.1 Example list of some errors that can arise in patient records, and some examples on how they could be managed..... 134

CHAPTER I

Introduction

I.1. OVERVIEW

The study of the Interstellar Medium (ISM) includes a very large range of potential physical parameters. Depending on the region of the ISM that is being studied, it can be cold ($\sim 10\,\mathrm{K}$), hot ($\sim 10^5\,\mathrm{K}$), diffuse ($< 1\,\mathrm{n_H/cm^{-3}}$) and can reach fairly high densities ($\sim 10^{10}\,\mathrm{n_H/cm^{-3}}$) (Williams & Viti, 2013). These regions have different physical conditions that are changed by many processes, such as gravity, supernova explosions and radiation, to name just a few. All of these environments allow for different chemical reactions to occur, the study of which is called astrochemistry.

The primary way of studying these regions and the molecules that we find in them, has been through observations. This is because recreating an interstellar environment on earth is very difficult and often limited to matching just a few of the physical parameters at a time. In doing so, several molecules have become helpful tracers of the underlying physics. The CO 1-0 rotational transition for example has a low excitation temperature (5.5 K) while CO has a relatively stable abundance with respect to H_2 (Liu et al., 2013). This allows us to use this transition as a tracer for cold gas. If the intensity is higher, that would mean there is more CO, and that would allow us to then infer the relative abundance of H_2.

In the following sections, we introduce the AstroChemical Origins (ACO) Innovative Training Network (ITN) as the work in this thesis contains deliverable products for the network, as well as the necessary components of astrochemistry, statistics and data management as they relate to the work in this thesis.

I.2. ASTROCHEMICAL ORIGINS NETWORK

The ACO ITN[1] received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 811312. This interdisciplinary ITN seeks to support academic work and provide non-academic training and experience for all researchers involved in it. The academic research is split into four work packages which include: i) radio receiver development; ii) astronomical observations; iii) laboratory and theoretical chemistry; and iv) model and tool development. The work of this thesis is contained within the fourth work package, which seeks to develop tools and models to be used in understanding astrochemical observations while being informed by the chemical research conducted in the ITN. This thesis in particular focuses on enabling users to complete full forward modelling easily without starting from particular observations. The network also intends to aid students in gaining non-academic experience through placements at various ACO associated private companies and interdisciplinary experience for the researchers, either by requiring work between the academic fields that ACO spans or by encouraging members to interact with research groups outside of their field of expertise. In order to help inspire the next generation of researchers the ACO ITN also created a virtual reality (VR) product. This VR project is intended to tell a simplified version of the story of water and forming planets while allowing it to be played as if it were a small game.

I.3. INTERSTELLAR MEDIUM

Gas and dust in the ISM undergo a complex life cycle of forming dense molecular clouds which can then form stars. Both as these stars evolve and when they die, they disrupt the surrounding medium and enrich it with heavier elements they created in the process called nucleosynthesis (Johnson, 2019). The disturbed and enriched environment heats up, forming warm, neutral, and ionised gas. Emission driven cooling then allows the gas to become denser, allowing the formation of cold neutral hydrogen clouds. These hydrogen clouds can then further cool and condense to form molecular clouds again. This cycle can then repeat to form ever

[1] https://aco-itn.oapd.inaf.it/home

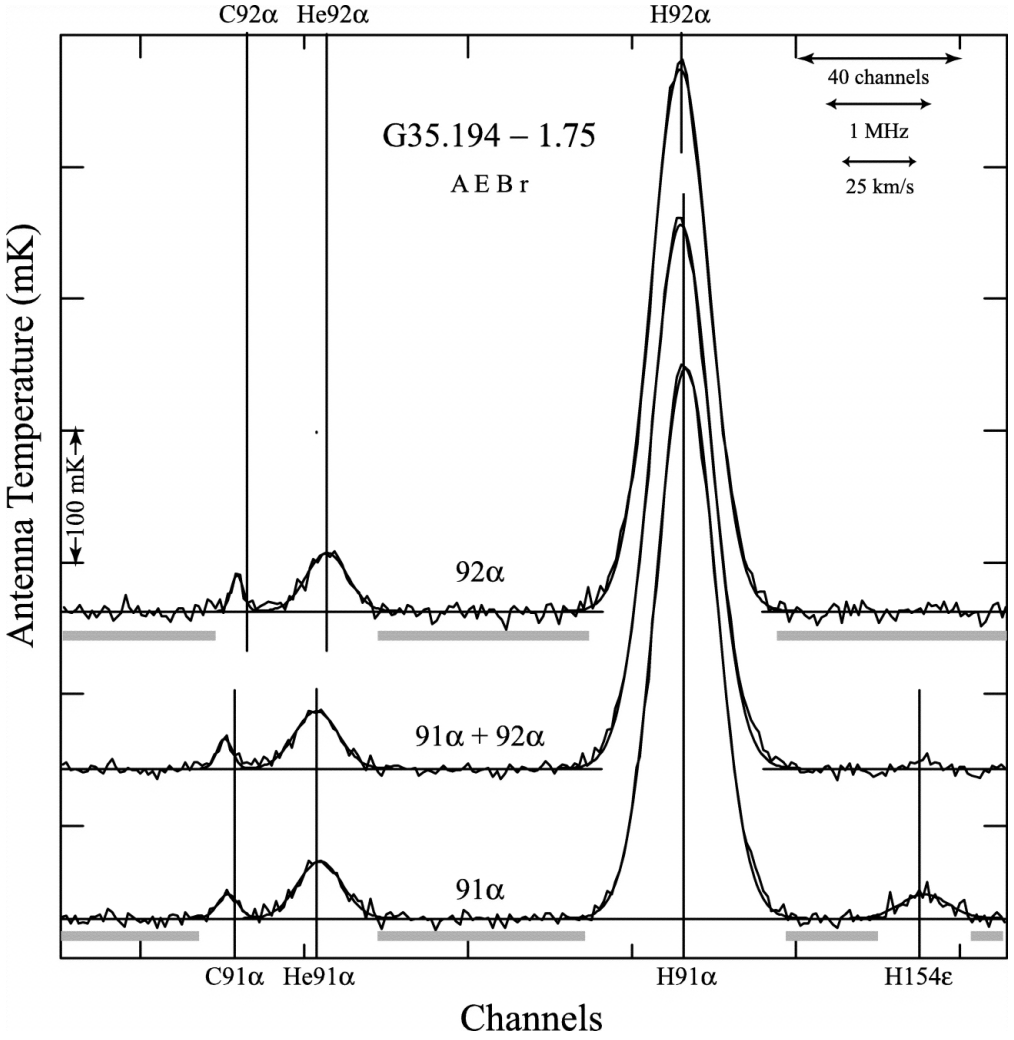

Figure I.1 91α and 92α transition for hydrogen, helium and carbon from an HII region. Image from Quireza et al. (2006)

more metal enriched stars. For an illustration see figure I.2.

With the diverse physical conditions found within the ISM gas and dust life cycle, it is useful to separate it into different objects or regions that we can study independently, even if the boundaries between them are blurry. In doing so we take the diverse ISM and separate it out into manageable components that we can study, model, and analyse. A slew of chemical reactions can occur within the different components, which adds to the molecular diversity that we can then observe.

Molecules are observed either in emission or absorption. The wavelength of the observed transitions is a unique identifier for atoms and molecules, which then can also give us information on the physical conditions they are in (Carroll & Ostlie, 2017). In radio astronomy a radiometer measures the noise coming from radio telescopes over a defined frequency range. Spectrometers split up the incoming

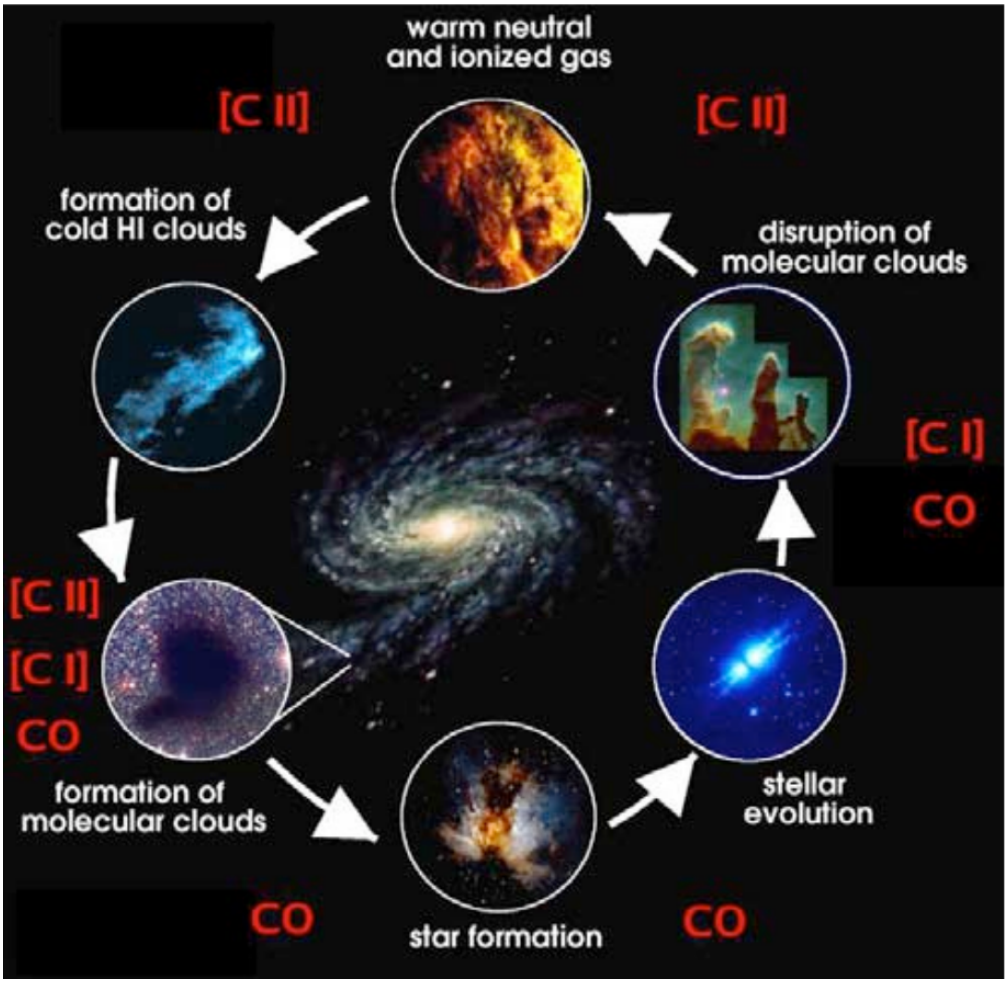

Figure I.2 ISM life cycle from Groppi et al. (2009)

frequency range into narrower frequency ranges and simultaneously measures the power of each one in order to measure spectral lines (Condon & Ransom, 2016). An example of what a spectrum can look like can be seen in figure I.1. The shape of the line profile is affected by several mechanisms. For example, the lifetime of an excited state relates to an uncertainty of the energy of the state, leading to natural broadening. A high number of collisions can lead to pressure broadening. Both of these types of broadening can be fit with Lorentzian profiles (Appenzeller, 2012). Another type of broadening that does not lead to a Lorentzian is broadening caused by the movement of individual atoms or molecules when observing a group of emitters or absorbers, which leads to a Gaussian profile (Condon & Ransom, 2016). When multiple types of broadening act independently on the observed line, then the line profile is a convolution of each of the mechanisms, which leads to a Voigt profile (Voigt, 1912).

Within all of the objects in the ISM, we find gas and dust, of which gas composes roughly 99% of the mass, while only 1% of the mass is in grains (Savage & Mathis, 1979). Of the gas, the majority is hydrogen, followed by helium and lastly oxygen before reaching all remaining heavier elements. While the dust grains do not contribute the majority of the mass, they play an important role in the chemistry. These grains can catalyse various chemical reactions. As the ISM has such low densities, the likelihood of gas-based reactions is low, but not negligible. Having a surface on which atoms and molecules can sit and diffuse increases the chances of reactions. Beyond this, grains also provide an energetic potential which can be used by atoms such as hydrogen to bond together to create molecules such as molecular hydrogen (H_2) (Hollenbach & Salpeter, 1971; Pirronello et al., 1997).

Grain surface formation of H_2 is important as it is difficult to form this species in the gas-phase, and yet H_2 plays a crucial role in gas-phase chemistry in the ISM (Williams & Viti, 2002). In order for two colliding atoms to stabilise into a bound molecule, they must be able to release energy greater than or equal to the centre of mass energy of the colliding pair so as to transition into a stable quantum state (Latter & Black, 1991). If there is no pathway to release this energy, the system can not transition into a stable state, and the pair splits apart. In gas-phase reactions, one method of releasing this energy is called radiative association, in which case the pair of atoms follow selection rules available to the system in order to release the energy in the form of a photon. Selection rules are a way of quantifying how systems can change from one quantum state to another and the energy that needs to be added or removed from the system in order to transition. These rules rely on approximations, and the transitions that can occur when using the electric dipole approximation for the interaction between the atom and a photon are called allowed transitions and have a high transition probability. Selection rules that rely on higher level approximation but are not allowed when using only the electric dipole approximation are labelled as forbidden. That does not mean they cannot happen, just that the likelihood of them occurring is considerably lower than those using just the electric dipole approximation (Bunker et al., 2006).

When looking at a system composed of two ground state hydrogen atoms colliding, the formed H_2 molecule is in a rotational-vibrational (rovibrational)

excited state which only has forbidden transitions available as the H_2 molecule is a homonuclear diatomic molecule with no permanent dipole. This means that the likelihood of a transition occurring during the collisional lifetime is very low, even negligible (Latter & Black, 1991). Because of this, the most dominant pathways to form H_2 in the gas-phase require multiple reactions to occur. The first pathway is

$$H + e^- \rightarrow H^- + h\nu \tag{I.1}$$

$$H^- + h\nu \rightarrow H_2 + e^- \tag{I.2}$$

and the second pathway is

$$H + H^+ \rightarrow H_2^+ + h\nu \tag{I.3}$$

$$H_2^+ + H \rightarrow H_2 + H^+ \tag{I.4}$$

(Latter & Black, 1991). Despite being the dominant gas-phase pathways, the reaction rates of these pathways can not account for the H_2 abundance in molecular and diffuse clouds (Williams & Viti, 2002). The abundance can be explained when considering dust grains as catalysts to H_2 formation rate (Hollenbach & Salpeter, 1971). When two hydrogen atoms interact on a dust grain, they can use the energetic potential of the dust grain to dissipate the energy needed to form stable H_2 molecules, which means that most, if not all, hydrogen atoms that interact with each other on dust grains should form H_2 (Hollenbach & Salpeter, 1971).

The nature of H_2 rovibrational transitions being forbidden means that even in situations where H_2 does transition from one rovibrational state to another, the signal we observe is faint and difficult to observe (Habart et al., 2005). There are allowed electronic transitions that have stronger signal strength for H_2. Electronic transitions happen when the electrons of the molecule transition between energy levels. However, these transitions absorb or emit ultraviolet photons, which are effectively blocked by the atmosphere of Earth requiring space based telescopes in order to observe them (Wakelam et al., 2017b). This combination of factors makes it difficult to observe and measure the abundance of H_2 outside of some special cases. Because of this, we have to turn to using other molecules as tracers of H_2 (Dickman, 1978).

One such tracer is carbon monoxide. CO is ubiquitously found in the ISM (Wilson et al., 1970). In addition, CO is also found in various types of objects within the ISM (Bash & Peters, 1976; Dickman, 1975; Knapp & Jura, 1976). This allowed previous studies, such as Dickman (1978) to estimate a linear relation between CO and H_2 column densities.

I.3.1. Star Formation

The process of star formation has been studied for over a century now. The first idea that resembles modern understanding of star formation can be traced back to Ritter in 1883 (Ritter, 1898), when it was first suggested that a diffuse mass could contract and heat through gravitational compression until it becomes hot enough to shine like the sun. While this original idea did not include the fusion of hydrogen into helium as the source of fuel for stars, instead suggesting that emission is driven by the heat from the gravitational compression, it was still closer to modern day understanding than most other theories at the time (Arny, 1990). Despite our understanding having grown significantly since 1883, research in star formation is still an important field in contemporary astrophysics with outstanding questions. These questions broadly cover both the physics of individual stars forming as well as the physics that governs the formation of clusters of stars. Examples of the latter include determining stellar mass functions which describe the distribution of stellar masses we should expect to see when a molecular cloud collapses (McKee & Ostriker, 2007). We begin our discussion of star formation by looking at a general description of molecular clouds with a simple description of how star formation occurs.

Within the ISM, molecular cloud regions have, on average, temperatures of around 10K with a density of about $10^4 \mathrm{cm}^{-3}$. These clouds can be small ($\sim$ 1pc with masses of $\sim 5 - 500\,\mathrm{M}_\odot$), or giant cloud structures ($\sim$ 100pc with masses up to $\sim 10^6\,\mathrm{M}_\odot$) (Williams & Viti, 2013). In such cold, dusty clouds, atoms and molecules can freeze out of the gas phase onto dust grains. The molecular cloud structures have a large visual extinction (Lynds, 1962) which allows us to detect them by locating areas where we observe fewer stars than expected. Originally, this was done in the optical regime of observations but is now commonly done in

near-infrared instead (Lombardi et al., 2006). It is such dense cold regions which have the potential to collapse into stars.

These clouds are inhomogeneous, lacking uniform physical properties and are held together by their own gravity and magnetic fields (Elmegreen & Scalo, 2004). The inhomogeneity is observed in substructures, such as filaments, bundles, and cores (André et al., 2010; Hacar et al., 2013). Star formation happens along the filaments of these clouds, with evidence of larger clusters of stellar formation at the knots where filaments collide (Schneider, N. et al., 2012). In order to allow for any star formation, the gravitational force of the mass of the cloud must exceed the internal pressures that prevent collapse. The mass needed to exert enough gravitational force to cause collapse is known as the critical mass. This critical mass can be described by using various physical parameters. For example, the Bonnor-Ebert mass considers a molecular cloud with a negligible magnetic field, supported only by thermal pressure (Bonnor, 1956). In such a case, the critical mass is equal to:

$$M_{BE} \approx \frac{225 c_s^4}{32\sqrt{5\pi p_0}(aG)^{3/2}} \tag{I.5}$$

where a varies based on the density distribution within the cloud, equalling one in a uniform density distribution, c_s is the isothermal sound speed in the medium, p_0 is the gas pressure, and G is the gravitational constant. The temperature is encoded within the sound speed, and the makeup of the material is also contained in the gas pressure and sound speed. Therefore, if the information for all three of these variables of a given cloud is known, we can derive the mass that could be supported by the thermal pressure exerted by this cloud. This is a simplified critical mass, and other types exist which additionally add physical parameters such as magnetic field support. If the mass of the cloud exceeds the critical mass in such a way that it overcomes all internal pressures, then it should start collapsing (McKee, 1989; Bonnar, 1956).

Collapsing cores that have surpassed the critical mass previously mentioned will evolve into stars by undergoing several stages as prestellar clouds. They start with an isothermal collapse until the density becomes large enough to create an optically thick medium, which would then also lead to heating of the core. While no observations of this very early stage have been fully reported, it is suspected

that such a sphere can find an equilibrium between the gravitational and internal pressure called a hydrostatic core (Enoch et al., 2010). As the cloud heats up further, it can reach a state where the molecular hydrogen dissociates, which then changes the specific heat, which in turn changes the internal pressure in such a way that collapse can continue until it becomes a prestellar core (Larson, 1969).

Once a core is prestellar, we classify it into one of three stages based on its morphology, and for each class, we can estimate an age which should be used as a guideline for the ages of each class: (i) Class 0 ($\leq 10^5$ yr); (ii) Class I ($> 10^5$ yr); (iii) Class II ($> 10^6$ yr). Class 0 objects are contained within an envelope with more mass than the core that they accrete from to grow larger (Barsony, 1994). Class I objects are protostars that still accrete from their envelopes, but the envelope has lost a significant amount of mass to the core and from outflows. The envelopes of Class I objects are undergoing the process of forming a thick dusty disc. As the envelope clears away, leaving just a thick disc, and the outflows of the core become weaker, we reach the regime of Class II objects. From here, the cores will continue to erode their discs and lose their outflows further before beginning their evolution towards being a main sequence star. During this process the temperature of the dust in the cloud also increases, which leads to various molecules being released from the surface of the back into the gas phase that were previously bound to grains. Depending on how quickly the heating occurs, it will affect the chemistry that can happen in the gas phase and on grains, as slow heating would first release weakly bound molecules from the grain and strongly bound ones later, while sudden heating would release both groups simultaneously (Williams & Viti, 2002).

This description of stellar formation, while not incorrect, is not complete and overlooks many details, including ones that are still being investigated today. One example of such a detail is angular momentum. The simplest equation for the angular momentum of an orbiting body is as follows:

$$L = 2\pi M f r^2 \tag{I.6}$$

where M is the mass of the object, f is the frequency of the orbit and r is the radius of the orbit. Unless it is removed from a system, angular momentum is conserved. This means that the individual elements that compose the collapsing cloud need to

increase the frequency of their orbits as their radius of orbit decreases if the angular momentum cannot be efficiently removed, thereby preventing full collapse of the cloud. One way in which this can be done is through magnetic fields. With the exception of the densest parts, molecular clouds are weakly ionised by the influence of cosmic rays as well as UV photons from external sources (McKee & Ostriker, 1977) leading to them being influenced by the galactic magnetic field (Mestel & Spitzer, 1956). This influence has many effects on star formation, both aiding and hindering star formation. As suggested earlier, one way in which magnetic fields can aid in star formation is by removing angular momentum. As the weakly ionised cloud collapses and spins, the movement of charged matter concentrates and twists the magnetic field lines affecting the cloud, leading to the flattening of the collapsing object into a disc (Ray, 2012). The twisting magnetic field is also capable of accelerating some of the material in the disc to flow out in jets parallel to the rotational axis, taking with it angular momentum and energy (Pudritz & Norman, 1983).

I.3.2. Astrochemical modelling

As star forming molecular clouds evolve physically, the molecules present change as well. This is due to a variety of effects; for example, the increasing density will cause various species to interact with each other at different rates than they would at a lower density. This leads to the destruction of some species and the formation of others. In order to study the molecular abundances and connect them to the physical conditions of star forming molecular clouds, we turn to modelling. In order to do so, modelling codes rely on integrating a system of ordinary differential equations (ODEs) numerically. These ODEs describe how the abundances of the various chemical species evolve at each time step for the object being modelled. The construction of such ODEs relies on understanding the various reactions that can occur and how they link to each other in a reaction network. To create these networks, we need information from databases such as KIDA (Wakelam et al., 2012) and UMIST (McElroy, D. et al., 2013), which can provide the gas-phase reactions. This is only part of the puzzle though, as there are also grain-grain reactions that need to be understood, as well as various processes that can cause

species to freeze-out onto grains (Rawlings et al., 1992) as well as how they can desorb from grains (Roberts et al., 2007).

For the sake of this thesis, UCLCHEM has been used as the chemical modelling code of choice. UCLCHEM is a gas-grain, time dependent chemical modelling code that allows for gas-gas, gas-grain and grain-grain chemical reactions in the reaction networks that are configurable by any user. This code was originally created in Viti & Williams (1999), before being updated in Viti et al. (2004) and Holdship et al. (2017). As of the time of writing, UCLCHEM can be found as an open-source project with ongoing updates for usability, performance and scientific use cases on github[2].

UCLCHEM allows for chemical evolution in two distinct phases. Like all UCLCHEM parameters, the initial conditions are free parameters, the default value for density is $100\,[\mathrm{cm}^{-3}]$ containing some molecules, and the default starting temperature is $10\,[\mathrm{K}]$. Historically, in order to save on computational time the density was set to be around $100\,[\mathrm{cm}^{-3}]$ as the collapse from a significantly lower density would take considerably longer to calculate at fixed time steps. Additionally, initial molecular abundances were set based on the given initial atomic abundances irrespective of the initial physical conditions as it was assumed that the chemistry in these clouds is fast enough to mitigate the inaccuracy of this assumption. In UCLCHEM, the phase I setting is most frequently used in order to collapse a diffuse cloud. In phase II, UCLCHEM starts with the fractional abundances of molecules from the phase I model, before injecting an energetic source which increases the temperature of the cloud. This second phase is meant to more closely resemble the specific observed environments.

Both evolutionary phases use DVODE (Brown et al., 1989) in order to solve the ODEs of each species in the chemical network that UCLCHEM is working with. DVODE solves the ODEs by using a linear multistep method for numerically solving ODEs in order to approximate solutions of initial value problems. In the most basic form, initial value problems look like the following

$$\dot{y} = f(t, y),\ y(t_0) = y_0 \tag{I.7}$$

[2]https://uclchem.github.io/

with the resulting approximation looking like

$$y_i \approx y(t_i), \tag{I.8}$$

where $t_i = t_0 + i\Delta t$ with Δt being the time step and i being an integer. In order to calculate the approximation of y_i linear multistep methods use the values of y_i and $f(t_i, y_i)$ of previous step in order to calculate the next desired step. So then the form of the multistep method becomes

$$\sum_{s=0}^{i} a_s y_{n+s} = \Delta t \sum_{s=0}^{i} b_s f(t_{n+s}, yn + s), \tag{I.9}$$

where n represents the last step for which an approximation has been calculated, i is the amount of steps to be taken from n, $a_i = 1$ while all remaining $a_0, ..., a_{s-1}$ as well as all b_s are determined by the method that will be used (Byrne & Hindmarsh, 1975). In the case of UCLCHEM the equations of the ODEs are different for the different chemical reaction environments such as gas-phase, and grain surface reactions as well as for physical processes that transition species from one environment to the other. For example, if a species is to freeze out onto the grain surface, then UCLCHEM requires a separate grain version of the given species and then the ODEs can treat this freeze out like any other chemical reaction turning a gas phase species into the equivalent grain surface species. This means that other transitions such as desorption, both thermal and non-thermal, have to be considered as their own reactions as well.

Each type of reaction has a corresponding equation associated to it; to give an example of how these equations look, we show the gas phase reaction of two species which has a rate equation calculated through the Arrhenius equation:

$$R_{AB} = \alpha \left(\frac{T}{300[\mathrm{K}]} \right)^{\beta} e^{\frac{-\gamma}{T}}, \tag{I.10}$$

where A and B represent the reactants, α, β, and γ are rate constants derived from theory and experiments, and T is the temperature at which the rate is being calculated. The results of this equation give out a value in units of $\mathrm{cm}^3\, s^{-1}$. We can use this reaction rate in order to set up the gas phase reaction ODE for a given

product from the reactants A and B in the following way:

$$\dot{Y}_{\text{product}} = R_{\text{AB}} Y_{\text{A}} Y_{\text{B}} n_{\text{H}}, \tag{I.11}$$

where Y_x represents the fractional abundance of x, $\dot{Y}_x$ is the rate of change for the fractional abundance of x and n_{H} is the number density of hydrogen. For the various types of reactions, there is a corresponding rate equation. This can be repeated using an equation for grain phase reactions, cosmic ray (CR) proton reactions, CR induced photon reactions and UV ionisation rate. The detailed descriptions for each process can be found in Holdship et al. (2017).

As the understanding of astrochemical reactions and processes is incomplete, so too is the ability to model this chemistry based only on the current best information. The nature of the objects being modelled with their extreme conditions and long time-scales at which they exist, makes them extremely difficult to study in laboratories (Collings et al., 2004; Fulvio et al., 2017) as well leaving the uncertainties on these chemical reactions extremely high (Linnartz et al., 2015). The limitations this poses, as well as the time needed in order to conduct laboratory experiments to constrain the uncertainties that are needed for astrochemical modelling, means that alternative approaches to at least reduce uncertainties are being explored using data driven methods, such as those in Holdship et al. (2018).

I.3.3. Radiative transfer modelling

In order to relate the astrochemical models to observations a second step is needed. Astrochemical models such as UCLCHEM only give out physical parameters and abundances of different chemicals, they do not provide observable information. As the gas temperature and the molecular excitation levels are not usually in thermodynamic equilibrium with each other in star forming regions, we must turn to radiative transfer models that calculate the intensities at different frequencies from an environment that does not have local thermal equilibrium (LTE). In order to model LTE and non-LTE environments, models solve the radiative transfer equation (van der Tak et al., 2007):

$$\mathrm{d}I_\nu \mathrm{d}s = j_\nu - \alpha_\nu I_\nu, \tag{I.12}$$

where the local emission and extinction coefficients are represented by j_ν and α_ν respectively, and I_ν is the specific intensity at a given frequency ν (van der Tak et al., 2007). The emission and extinction coefficients can be combined into one source function which is defined as:

$$S_\nu \equiv \frac{j_\nu}{\alpha_\nu}, \tag{I.13}$$

which when this is used in the integral form of the radiative transport equation and combined with the optical depth along a light path of infinitesimal width, defined as:

$$\mathrm{d}\tau_\nu \equiv \alpha_\nu \mathrm{d}s, \tag{I.14}$$

where τ_ν is the spectral optical depth in frequency, we get the equation:

$$I_\nu(\tau_\nu) = I_\nu(0)e^{-\tau_\nu} + \int_0^{\tau_\nu} S_\nu(\tau'_\nu)e^{-(\tau_\nu - \tau'_\nu)}\mathrm{d}\tau'_\nu. \tag{I.15}$$

In this, $I_\nu(\tau_\nu)$ is the intensity originating from the medium being modelled, and $I_\nu(0)$ is the intensity of any background emission that did not originate from the medium, but would be observed by someone if the object being modelled was not present. The limit of τ_ν is the total optical depth along the line of sight. This equation holds for radiation that is from a continuum over a large band-with as well as for spectral lines which can have more drastic changes over small frequencies. In order to calculate the value of the source function, we need to take a closer look at bound-bound transitions of multi-level molecules (van der Tak et al., 2007).

Multi-level molecules have N levels of Einstein coefficients for spontaneous emission rates (A_{ul}) from upper (u) to lower (l) levels, Einstein coefficients for radiatively stimulated emission (B_{ul}) and excitation (B_{lu}), as well as collisional excitation (C_{ul}) and de-excitation rates (C_{lu}). Spontaneous emission is the process of an electron in an upper level, u, jumping to a lower level, l, while releasing the energy in the form of a photon of a wavelength dependent on the difference between the two energy levels. Stimulated emission is the process of a photon from the surrounding environment inducing an electron to jump from a higher energy level to a lower one while releasing a photon again dependent on the difference between the energy levels. Stimulated absorption describes the process of an electron absorbing

a photon from the surrounding medium in order to jump from a lower level to a higher level. The energy of the photon that is needed is dependent on the difference between the two energy levels. Collisional excitation is the process of colliding molecules losing kinetic energy to electrons that use the energy to jump to higher energetic levels, while collisional de-excitation is the reverse process if one of the collisional partners had an excited electron already. In order to calculate the source function for such molecules, we use these rates to calculate the emission and extinction coefficients:

$$j_{\mathrm{ul}} = n_{\mathrm{u}} A_{\mathrm{ul}} \tag{I.16}$$

$$\alpha_{\mathrm{ul}} = n_{\mathrm{l}} B_{\mathrm{lu}} - n_{\mathrm{u}} B_{\mathrm{ul}}, \tag{I.17}$$

where n_i represents the number density of either the upper or lower states. If we were to be observing an environment in LTE, then having the kinetic gas temperature would allow for both n_{u} and n_{l} to be determined, which would then allow us to solve the radiative transfer equation explicitly. However, as stated most star forming regions are usually not in LTE. It is often safe to assume statistical equilibrium, at which point we have:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = 0 = \sum_{j \neq i}^{N} n_j P_{ji} - n_i \sum_{j \neq i}^{N} P_{ij}, \tag{I.18}$$

where P_{ij} represents the rate of formation for level j by destroying level i. This value is given by the equation:

$$P_{ij} = \begin{cases} A_{ij} + B_{ij}\bar{J}_\nu + C_{ij} & , \text{if } i > j \\ B_{ij}\bar{J}_\nu + C_{ij} & , \text{if } i < j. \end{cases} \tag{I.19}$$

In this equation $\bar{J}_\nu$ is the number of induced radiative transitions from level i to j calculated using the specific intensity integrated over solid angle and averaged across all directions. Both cases together allows us to reformulate equation I.18 to:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = \sum_{i<j}^{N} (n_j A_{ji} + (n_j B_{ji} - n_i B_{ij})\bar{J}_\nu) - \sum_{i<j}^{N} (n_i A_{ij} + (n_i B_{ij} - n_j B_{ji})\bar{J}_\nu) + \sum_{j \neq i}^{N} (n_j C_{ji} - n_i C_{ij}) \tag{I.20}$$

which still equals zero for this discussion in order to allow the radiative transfer problem to be solved independently of any assumption made for the chemical processes. Simultaneously solving equations I.15, I.20, and I.13 with use of I.16, is called the radiative transfer problem and is iteratively solved by radiative transfer models. As population levels and mean intensity are dependent on the position within a modelled object, it is vital to either understand the source structures being modelled, or to assume one. If such an assumption is not or cannot be made, then it becomes necessary to solve the radiative transfer problem through the use of grid points within the modelled source (van der Tak et al., 2007).

I.4. STATISTICS

In astronomy, the study and application of statistics is very important. The science of taking in, studying and interpreting the meaning of data is a corner stone to much of astronomical research. Through the study of large number of data-points, it becomes possible to discover trends between similar classes of objects. While understanding how to draw conclusions from a very limited amount of observations, such as only having one universe to observe for astronomy, is also important. In this way, statistics provides both a way to understand large amounts of observed data, and how to handle large sets of models to fit limited observations.

A common tool to quantify a fit is using a Bayesian likelihood. As the complexity of models increases, and the potential parameter space to explore grows, it becomes ever more difficult to fully evaluate all possible models. Because of this, it is important to use and understand sampling algorithms, such as Markov Chain Monte Carlo (MCMC) Foreman-Mackey et al. (2013), which we can use in conjunction with our method of evaluating fits.

I.4.1. Bayesian Statistics

Bayesian statistics focuses on handling probabilities and defining a degree of belief about the events being studied (Murphy, 2012). To do this, it needs to know how to handle uncertainties. The basis to the degree of belief lies in the available knowledge, and the uncertainties are in what is yet unknown to the current inference. In Bayesian statistics, we therefore form a probability function, not from repeated

experiment as frequentists would, but rather by describing the likelihood of making the observations given our model and previous knowledge.

In order to describe the likelihood, Bayesian statistics uses the Bayes' theorem (Bayes & Price, 1763; Joyce, 2021). This equation produces a posterior likelihood of a given event, using previous information or data. If for example, we wanted to know how well the parameters p in a model, describe the observable data D, then we could do so by calculating the probability of p, given D. To calculate this, we turn to Bayes' theorem, which looks like this:

$$P(p|D) = \frac{P(D|p)P(p)}{P(D)}, \tag{I.21}$$

where $P(p|D)$ is called the posterior and describes the likelihood of parameters p being able to describe what we observe in data D. $P(D|p)$ is the likelihood of observing data D given the parameters p. The function $P(p)$ is the prior belief of the chosen parameters p, and $P(D)$ is called the evidence, it describes the likelihood of having observed data D, but is often used as a normalisation factor as it does not depend on the parameters p. Often, the posterior does not have a closed form, and depends heavily on the observations and parameter space within which we sample for p. As these spaces can be quite large in dimension, we must approximate the posterior. To do so, we sample the parameter space. One important method to do so, is MCMC sampling (Foreman-Mackey et al., 2013).

I.4.2. Markov Chain Monte Carlo

The MCMC method is comprised of two components, the Markov chain describing a sequence of possible events, and the Monte Carlo method Foreman-Mackey et al. (2013). Markov chains move from one state to another using a likelihood of performing the transition. In doing so, the next step should be described by the current state only, and be independent to the previous states (Gagniuc, 2017). When combining this with Bayes' theorem we can describe the likelihood of switching from one state to the next, using the likelihood of switching from the current state to the next by comparing the likelihood of each state. The higher the likelihood of state A relative to that of state B, the more likely that the walker

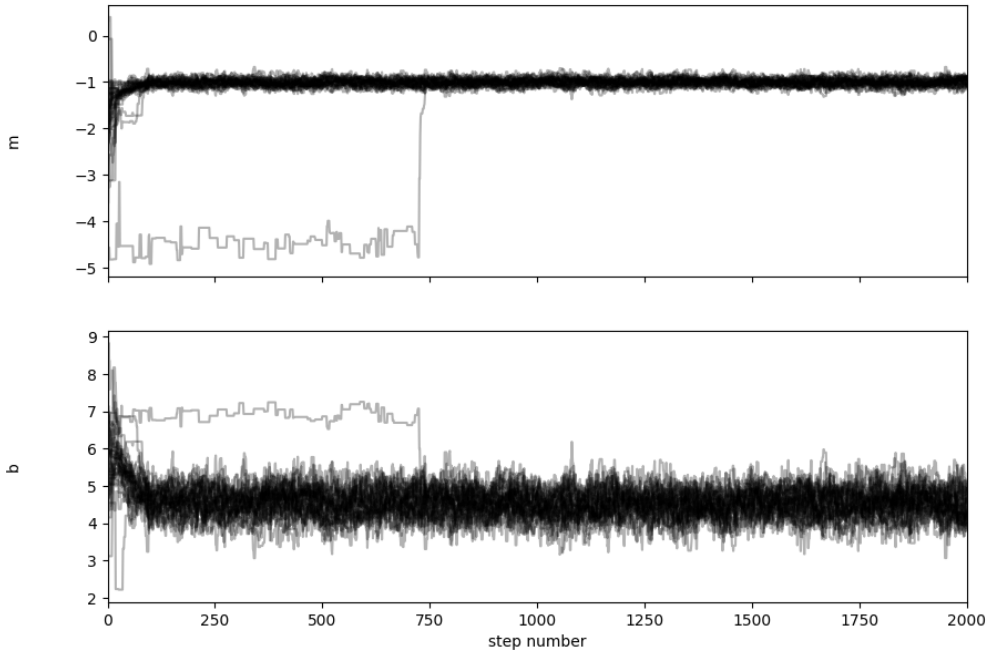

Figure I.3 Simple emcee chains fitting a linear function to a dataset.

will switch to, or remain at state A depending where it started.

Monte Carlo methods are a form of algorithm to draw random samples in order to obtain results for a given problem.

When combined, these two methods can boost each others performance. Starting with a set number of walkers which each represents a Markov chain, and where each walker has a state assigned to them by the Monte Carlo method, we can begin the process of mapping the most likely parameters in the sampled space. The Monte Carlo method samples the space for potential future states that the walkers could move towards, at which point the Markov chain walkers can determine the chance of switching to the new state, or staying where they were.

The collection of steps taken by an individual walker is called a chain, and when analysing the results of an MCMC inference, we use the chains of the walkers, not the final steps. The initial starting parameters of the MCMC walkers and the steps the walkers take towards the parameter space with the highest likelihood are important for the inference to succeed but should be cut from the chains prior to plotting inference results, as they do not include information for the final result. To explain this, we turn to a toy example of fitting a linear function to a generated data set. To do this, the slope, m, and y-axis intercept, b, need to be determined, for which we will use an MCMC inference. The resulting chain of the walkers

can be seen in figure I.3. Here, the impact of random or naive starting positions can be seen as the walkers slide from an area of lower likelihood towards an area with higher likelihood. If a better informed starting position had been chosen, this behaviour could potentially be limited or removed. If this run were allowed to progress until each combination of parameters had been sampled at least once, then the amount of steps at any given point relative to any other point would be dependent on the likelihood of each step describing the data. In such a scenario a likelihood plot of the full parameter space would simply be the number of times a walker was at any given combination of parameters. However, the benefit of using MCMC inferences is not to map the full space but to find areas of high local likelihood without having to map the full parameter space, with the hope of finding the absolute maxima in the covered parameter space. Therefore, when analysing the results of the inference, we could cut the initial steps, removing the sliding in, in this case, potentially cutting the first 200 steps, leaving only the parts of the chain describing areas of higher likelihood. Doing so would be referred to as burn-in (Johansen, 2010).

MCMC modelling can be found in many fields of astrophysics. In cosmology the use of MCMC has led to the development of multiple packages aimed at simplifying the process (Lewis & Bridle, 2002; Akeret et al., 2013), aiding researchers to understand their observations faster without having to recreate work that has already been done by others. One example of such work is by Hildebrandt et al. (2016) which used CosmoMC with a weak gravitational lensing analysis of observations from the KiDS survey (de Jong et al., 2015) in order to find constraints on cosmological parameters. In astrochemistry, MCMC can be used in order to infer reaction rates of grain surface chemistry, such as was done in Holdship et al. (2018). While these are just two examples, the vast difference in the research topic should highlight that MCMC inference has wide applications in astrophysics.

A remaining question is: how do we know that the MCMC inference has reached a steady solution? In order to validate the convergence of walkers we turn to convergence tests. Convergence tests, as the name implies, aim to validate that the chains of walkers have reached a stationary solution. That does not mean the walkers themselves no longer change their parameters, but that the space

being explored stays relatively steady. We refer again to figure I.3 to explain this point. If we look at the distribution of the chains after step 1000, we see that the distribution, while noisy, continuously stays within a range of values both for the slope and the y-intercept. This evaluation could be considered a by-eye convergence test, which can be useful in a qualitative way but does not provide a quantitative understanding of convergence (Johansen, 2010). A quantitative approach of such a test can be seen in the Gelman-Rubin statistic, which estimates the variance within a chain (intra-chain variance) and the variance between chains (inter-chain variance)(Gelman & Rubin, 1992). Starting with calculating the posterior mean of a parameter, θ, for each chain i of I total chains ($I > 2$), with a total number of steps S:

$$\bar{\theta}_i = \frac{1}{S}\sum_{s}^{S}\theta_s^i. \tag{I.22}$$

The intra-chain variance, σ_i^2, is then:

$$\sigma_i^2 = \frac{1}{S-1}\sum_{s}^{S}(\theta_s^i - \bar{\theta}_i)^2. \tag{I.23}$$

From here we can calculate the average intra-chain variance W:

$$W = \frac{1}{I}\sum_{i}^{I}\sigma_i^2. \tag{I.24}$$

To get the inter-chain variance we calculate the mean of all chains for parameter θ:

$$\bar{\theta} = \frac{1}{I}\sum_{i}^{I}\bar{\theta}_i. \tag{I.25}$$

Which is then used with the posterior mean of each chain to calculate the inter-chain variance:

$$B = \frac{S}{I-1}\sum_{i}^{I}(\bar{\theta}_i - \bar{\theta})^2. \tag{I.26}$$

We can then estimate the target variance using a weighted average of W and B:

$$\hat{\sigma}^2 = \frac{S-1}{S}W + \frac{1}{S}B. \tag{I.27}$$

This potentially overestimates the actual variance, σ^2, depending on the starting values of the walkers. In the limit of S going to infinity, $\hat{\sigma}^2$ is an unbiased estimate of the true variance. For any S that is finite, W should be less than the true variance but will approach the true variance as S goes to infinity. Because of this we can use the Gelman-Rubin statistic R:

$$\sqrt{R} = \frac{\frac{S-1}{S}W + \frac{1}{S}B}{W}. \tag{I.28}$$

This value should be approximately 1 in the case of the chains having convergence (Gelman & Rubin, 1992). One thing of note for this convergence test, as well as the qualitative by-eye analysis, is that they do not guarantee that the solution is the correct solution, instead only indicating that the walkers have found the same area in parameter space which could be a local maximum in likelihood. It is possible that other, higher likelihood areas exist within the parameter ranges described by the prior but that no walkers approached them.

I.5. DATA MANAGEMENT

Data management is a specialisation that focuses on how we store and collect our data. This means that it focuses on how to organise, collect and protect data while making it available for those that have permission to use it. This specialisation has many techniques it uses to function well. One technique is the data pipeline. This is a series of processes, where each process takes as input, the output of the previous process. Data pipelines transport data from one location to another and/or perform a series of tasks on the data. An example of this is the data pipelines used to work on observations, like those for the Hubble Space Telescope (HST) (Zieba & Kreidberg, 2022). Some of the HST pipelines look for sink pixels, or correct for charge transfer efficiency. Sink pixels refer to pixels that are capable of producing electrons when interacting with photons, but some or all of the charge becomes trapped in the pixel rather than being read out for the observer (Anderson & Baggett, 2014). Another technique is data storage solutions. Data storage solutions take aggregated data, and place it in an accessible format so that it can be used as needed.

Depending on the type of data being stored, different solutions may be used.

In order to understand which to use, it is important to look at the type of data that will be stored and which category of storage solution will fit it. The three broadest categories of data storage solutions are: (i) Structured; (ii) Semi-Structured (Buneman, 1997); (iii) Unstructured (Deckler, 2022).

Structured databases are often also called relational databases and come in the form of tables which are connected to each other. Many such databases use the Structured Query Language (SQL) in order to relate data-points between tables, and so users can query the database to find the data they want to use. The structured format allows for data to be split up by the information each data-point contains into separate tables that are linked to each other. By assigning a key to each individual data-point we can then maintain the relation between tables. Each table then has columns and rows representing the information we stored. Beyond allowing data-points to be split up, the relational component of structured databases also allows for extra, external information to be stored, and associated to different fields in tables. A basic example could be storing the value of individual ingredients that a kitchen would require to cook the dishes listed on the menu. The different data-points, in this case dishes, could have the information of what ingredients are needed to make but the cost of each ingredient would be additional external information, not decided on by the data-point itself. By storing this information in the database, we can perform a quick query in order to ascertain the cost of ingredients for each dish stored in the database.

Semi-Structured databases rely on each data-point having their own self describing tags (Buneman, 1997). In doing so, each data-point will not need to be sorted into a table, but rather can simply be placed into the format of how it is being stored. Each data-point of the same class may have separate components to describe it. This allows for a greater diversity of types of data-points, but comes at the cost of losing relational links. A great example use case of this type of data could be astronomical observations. A single object can have many different molecules that are observable, but not each molecule is found within each object. By storing the data in a semi-structured format, it is convenient to add individual elements which describe the molecules that were found, and at which abundance. In a Structured database, this would lead to either having too many tables, or too

many columns that would be empty for a large number of observed objects that we stored.

Unstructured data refers to data which does not have a defined structure with either self-describing tags, or in a form so that it could be used like structured data without additional work. One example of such data is form filled data. One example from the medical field is when patients have to type in the exact symptoms they have rather than make a selection from a list. In doing so we have created Unstructured data. This data needs to either be stored as it is, or needs to be evaluated through a data pipeline in order to change it from unstructured to structured or semi-structured. If stored as it is, it will likely not be useful for any larger analysis beyond what the one patient requires. Without additional components, this may not sound like a useful format to store data, however with additional tools to work on the data, it can be very important to have systems in place to identify and store data that is unstructured. This then allows systems to work through and structure the unstructured data.

I.6. THIS THESIS

In this thesis, we discuss the complexity of modelling molecular clouds and protostellar cores, as well as how it is important to understand general trends between objects. In order to do this, we implement statistical methods to study the large quantities of data that have been gathered.

This thesis showcases the work of creating a statistical inference software code in Chapter 2 which performs a Markov chain Monte Carlo (MCMC) inference with a Bayes' theorem based likelihood equation, while also having a user interface (UI). In Chapter 3 we discuss using this software tool to study sulphur bearing molecules in the class I protostellar core, SVS 13-A. The penultimate chapter describes the creation of a Virtual Reality (VR) product created as part of the Horizon 2020 Astro-Chemical Origins (ACO) Innovative Training Network, created for outreach purposes. Work in appendix A diverges from astronomy by describing the creation of a data storage solution and dashboard for medical research and how a similar project could be created for astronomical data in archives to help observers find desired fits files, or to help in proposing future surveys.

CHAPTER II

UCLCHEMCMC

The work presented in this chapter is based on the paper by Keil et al. (2022), in collaboration with S. Viti and J. Holdship, with the work being conducted by M. Keil with supervision and guidance by both S. Viti and J. Holdship.

II.1. INTRODUCTION

Throughout the interstellar medium (ISM), chemical reactions impact the environments that we observe. In turn, the physics of a molecular cloud greatly affects the chemistry. For example, at high densities ($\gtrapprox 10^5\,\mathrm{cm}^{-3}$) and low temperatures ($\lessapprox 30\,\mathrm{K}$), atoms and molecules freeze out onto dust grains where they can react through many pathways (for a review see Allodi et al. (2013)). On the other hand shocks from protostellar outflows impacting the surrounding medium can lead to desorption of many molecular species stored on dust grains (Caselli et al., 1997). Hence, emission and absorption lines of different species allow us to study the physics of the objects we observe.

In order to interpret the observations that are made, radiative transfer codes can be used to calculate the expected intensities that should be observed with a given set of physical parameters. One example of such a code is RADEX (van der Tak et al., 2007) which focuses on non-LTE analysis. These types of codes require parameters that describe the condition of the gas as well as the column density of a species. These are connected by chemistry but treated as free parameters in many radiative transfer models. Modelling tools such as UCLCHEM (Holdship et al., 2017) or GRAINOBLE (Taquet et al., 2012), provide the fractional abundances of species which can be used to calculate the column density. The fractional abundances can be combined with estimates of the total gas column density of an

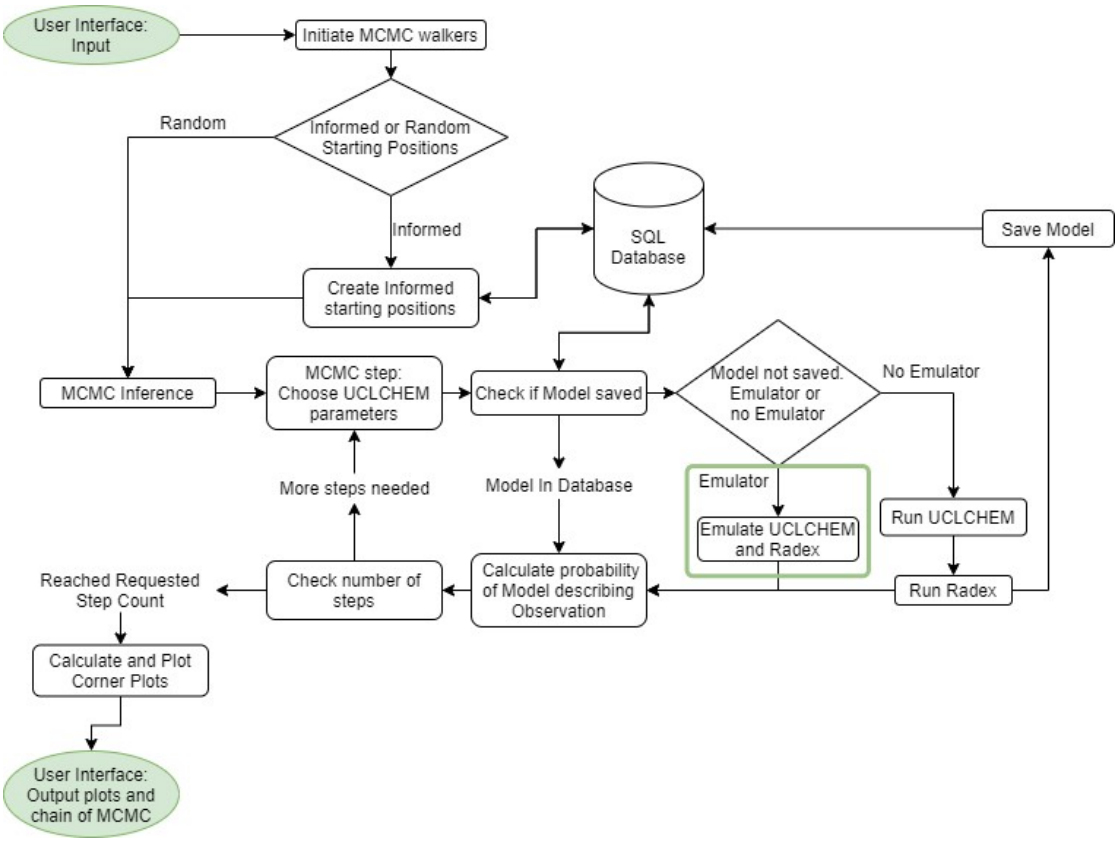

Figure II.1 Flow Chart of the various processes that happen in UCLCHEMCMC. Green ovals indicate parts that the User interfaces with. Diamonds indicate parts of UCLCHEMCMC where the next step is dependent on options the user specifies. The SQL Database is represented with a cylinder and has been labeled this way for clarity. Arrows to and from the database represent a query of the SQL which then returns the models that match the query. The emulator part, marked with a green box, highlights where emulator codes could sit relative to the rest of the workflow of UCLCHEMCMC; implementation of such emulators is beyond the scope of this work.

object in order to calculate the column density of an individual species.

Chemical models calculate fractional abundances by considering the rate of change of many species as they interact through a network of reactions. These reaction networks usually include a gas phase database such as KIDA (Wakelam et al., 2012) or UMIST (McElroy, D. et al., 2013), as well as gas-grain and grain surface processes such as freeze out, non thermal desorption and surface reactions. Additional processes such as thermal desorption or sputtering of ice mantles are often included depending on the chemical code and its intended purpose. The complexity involved in determining what should be included in these models in order to maximise the accuracy while minimising computational cost of creating a

model, is an aspect that requires significant expertise.

The best modelling approaches combine the chemical modelling codes with the radiative transfer codes. One benefit in doing this is that the column densities can be calculated with the chemical model which can then be combined with the set of physical parameters calculated by the chemical code to use as parameters for the radiative transfer model. There are additional parameters for the radiative transfer codes, such as line width, which are not directly calculated by a chemical code but can be treated as free parameters or derived from observations. The outputs from the radiative transfer code can then be compared to spectroscopic observations (Harada et al., 2019; Punanova et al., 2018; Viti, 2017).

In order to assist in the inference of physical parameters of an observation, we present the open-source, MCMC inference tool UCLCHEMCMC[1]. The intended use of UCLCHEMCMC is to infer the probability distribution of key physical parameters given some observed data. The following section will describe the code in detail, starting with the forward modelling approach and what tools it uses in section II.2.1, followed by the work flow of the code and how it stores models to allow future inferences to be more efficient in section II.2.3 followed by a brief description of the chosen interface in section II.2.4. After that, we will examine an example case by providing UCLCHEMCMC with mock observations, and then perform a stress test inference using observations of the prestellar core L1544 in section II.3, before going on to discuss caveats for the use of this tool and summarising in section II.4.

II.2. UCLCHEMCMC

UCLCHEMCMC infers physical parameter values from molecular observations using chemical and radiative transfer models. First, we use a chemical model, in order to obtain abundances for a user-defined list of species UCLCHEMCMC has been configured for. These abundances, can then be used with a radiative transfer model, to calculate the intensities for the emission lines of those species. For a single model, this process can take several minutes to be calculated on a standard computer.

[1] `https://zenodo.org/badge/latestdoi/334982976`

Table II.1. Inputs and options per page

Page	Input/Option	Description
Parameter input	Final volume Density (cm^3)	Hydrogen volume density at which the model stops collapsing
	Kinetic temperature (K)	Kinetic temperature of the gas
	Cosmic ray ionisation rate	Multiplicative factor of the galactic rate of ionisation caused by Cosmic rays ($1.3 \times 10^{-17}\,s^{-1}$)
	UV radiation field strength	Strength of the external UV radiation field strength acting on the cloud
	R_{out} (pc)	Radius of the modelled cloud
	Line width ($km\,s^{-1}$)	RADEX Line width of observation
Observation input	Species list	List of the species that have been configured
	Transition list	List of transition lines that have been configured for a given species
	Observation inputs	Space to fill in the observations, errors and choice of units for the observation
Options (Page 3)	Grid type	Choice on whether to use coarse or fine grid for the parameters being inferred
	Informed starting position	Choice on whether to use informed starting positions or random starting positions
	Session name	Back-end session name to allow the session to be reloaded later
	Number of walker	Number of walkers the MCMC should use (It is recommended by the package emcee to use twice as many walkers as parameters being inferred)
	Number of steps	choice of the number of steps an inference should take before stopping and loading evaluation corner plots
Inference	Start inference	to start the MCMC inference or continue it if a previous session is loaded
	MCMC corner plot	MCMC corner plots of previous steps, if previous steps exist, or the starting positions if a new inference is being started

Note. — Options of inputs and outputs per page for UCLCHEMCMC

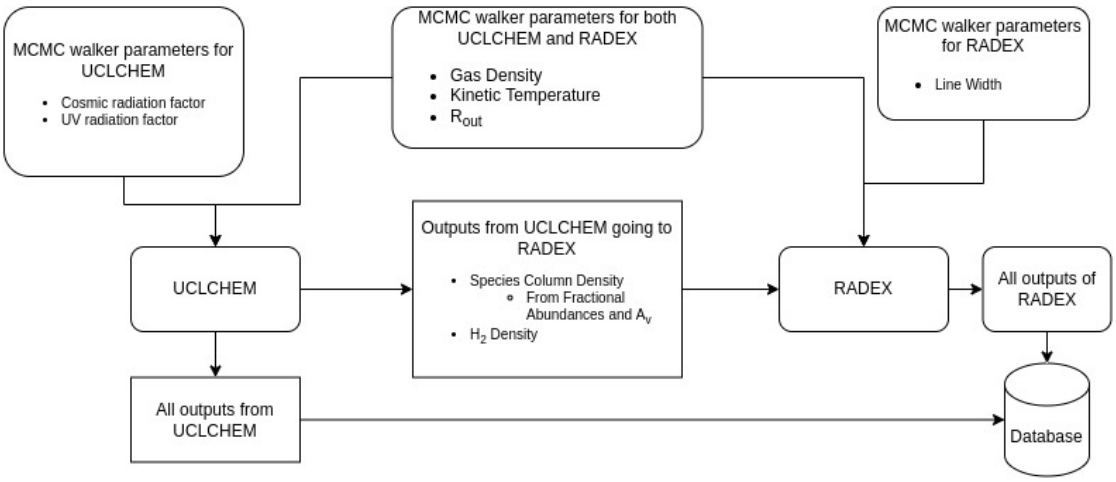

Figure II.2 Small flow chart showing which parameters go into UCLCHEM and which parameters are taken from UCLCHEM and given to RADEX.

In order to infer the physical parameter values, we use the affine-invariant Markov Chain Monte Carlo (MCMC) Ensemble algorithm of Goodman & Weare (2010) as implemented in the python package emcee (Foreman-Mackey et al., 2013). This kind of sampling initiates walkers with a set of physical parameters which are used to calculate the chemical and radiative transfer models as just described. During each step, the walkers calculate the likelihood value for that set of parameters using the likelihood function given by the user (see section II.2.2 for the details on the likelihood). After calculating the likelihood of the current values, the walkers choose a new set of values in parameter space, for which the likelihood is also evaluated. At this point, the walker must decide if it will remain stationary for this step, discarding the new set of values and keeping the one it already has, or if it will discard the old values and keep the new ones. This decision is dependent on the type of "move" function that is chosen (for details, please refer to Foreman-Mackey et al. (2013)). The default function for UCLCHEMCMC is a combination of a differential evolution proposal (Nelson et al., 2013) and a snooker proposal using differential evolution (ter Braak & Vrugt, 2008). Mixing like this is recommended by Foreman-Mackey et al. (2013) when dealing with multi-modal problems as the differential evolution proposal can allow for large enough step sizes to cross between peaks in probability. However, the standard version strongly recommends that at least $N = 2d$ walkers are used, where d is the number of parameters over which to infer, and performs better with higher N. The snooker

proposal using differential evolution allows for improved performance, compared to the basic version, when a smaller number of walkers is used. The process of sampling parameter space in this way requires thousands of models to be calculated before a meaningful posterior is produced. The calculations of each model using UCLCHEM and RADEX can take a minute or more, which can result in several hours of computing time prior to the parameter space being sampled enough to produce a usable posterior.

Our aim is to improve the efficiency without decreasing the accuracy. To do this UCLCHEMCMC manages a database of previously calculated models from which it can retrieve values when required. The curated database contains the input and output from both the chemical and radiative transfer models, and is used to perform an inference of the physical parameter space of an observed object without repeating any calculation. Anytime a new step is taken, it can check if this combination of parameters has been used before, and if it has, use the old output rather than perform a new calculation. A full flowchart of the processes done can be found in Figure II.2. In this section we will start by briefly describing the forward modelling method we use and the software UCLCHEMCMC requires in order to create the models that it stores, followed by the details of the MCMC inference and how UCLCHEMCMC manages the database, before detailing the interface it has.

II.2.1. Forward modelling

To create the simulations that are stored in the database, we use UCLCHEM combined with RADEX as the chemical model and radiative transfer code respectively. Physical parameters describing the gas conditions are passed to UCLCHEM to generate abundances and then a subset of these values are passed to RADEX in order to get a list of transition lines. Beyond the outputs from UCLCHEM, additional free parameters are required for RADEX such as the line width. A detailed flow chart on how the modelling tools interact with each other can be found in Figure II.2, for clarity. The inputs for UCLCHEM and RADEX, listed in the flow chart, can be changed according to the needs of the inference to be run, but requires changes to be made to configuration files.

By default, UCLCHEMCMC is configured to use RADEX for the radiative transfer calculations; we use the fractional abundance, calculated by UCLCHEM, to approximate the column density by using the visual extinction calculated from the given R_{out} and gas volume density. This value is used alongside the other inputs UCLCHEM was given, that RADEX needs as well, such as gas volume density and kinetic temperature, in order to run the radiative transfer model to produce observables which can be compared to the data. The observable values RADEX produces are: radiation peak temperature (T_R) in K which is comparable to the observed antenna temperatures; integrated surface brightness in $\mathrm{K\,km\,s^{-1}}$; the isotopic flux emitted in all directions in ergs $(\mathrm{s\,cm^2})^{-1}$ (van der Tak et al., 2007). Any of these can be sent to the likelihood function depending on the user's data.

Both UCLCHEM and RADEX can be replaced, as UCLCHEMCMC is only designed to perform an MCMC inference and manage an SQL database. As long as inputs and outputs are carefully tailored to a given project, the code could be simply modified to be used with any chemical modelling or radiative transfer codes. To begin, we use a limited list of chemical species whose collisional data is available in the Leiden Atomic and Molecular Database (LAMDA Schöier et al. (2005)), as RADEX requires such data.

II.2.2. MCMC Inference

For the MCMC inference, UCLCHEMCMC uses the python package emcee (Foreman-Mackey et al., 2013) in order to calculate the posterior probability density function (PDF) of the physical parameters for the desired observation. We assume the errors on the data are Gaussian and that our model provides the true intensities for any given parameters. The initially configured likelihood of observing our data given some parameters that was used for the example cases in section II.3 is therefore,

$$\mathcal{L}(d|\,\vec{\theta}) = \exp\left[-\frac{1}{2}\sum_i \frac{(d_i-\theta_i)^2}{\sigma_{\mathrm{d}_i}^2}\right], \tag{II.1}$$

where d is the collection of I observations, with the ith element being represented with d_i, θ is the collection of I RADEX modelled lines, which used physical

parameters from UCLCHEM, with the ith element being represented with θ_i , and σ_{d_i} is the error in the observed input, each for line i. This is then combined with a prior on the physical parameters, $\mathrm{P}(\theta)$ and the Bayesian evidence, $\mathrm{P}(\mathrm{d})$, in Bayes theorem to get the full PDF in the form of

$$\mathrm{P}(\theta|\,\mathrm{d}) = \frac{\mathrm{P}(\theta)\,\mathcal{L}(\mathrm{d}|\,\theta)}{\mathrm{P}(\mathrm{d})}. \tag{II.2}$$

By default, the prior is a uniform top hat function in grid space, on the ranges designated by the end user but can be altered by end users. As the prior is applied in grid space, it will be a log uniform prior if the physical parameter has a log spaced grid applied to it. The evidence is treated as a normalisation factor, and the values that are used for our example application on prestellar core will be discussed in section II.3. This approach to parameter inference has been used before (Holdship et al., 2019) and UCLCHEMCMC makes such inference problems simple.

II.2.3. Database

The core of UCLCHEMCMC is the managing of a database of models and running an MCMC inference which is supplemented by that database. After giving the inputs, which will be detailed in section II.2.4, UCLCHEMCMC initiates a string of operations in order to start calculating the posterior PDF of the physical parameter values of an object using an MCMC sampler. A detailed flow chart of the processes can be seen in figure II.2. The code starts by initiating the walkers that will be used for the MCMC inference. If informed starting positions were requested, the SQL database will be searched for models which are similar to the given observations based on a simple top-hat function with a configurable distance on either side of the observed intensities. By searching and retrieving the parameters of models that have intensities similar to the given observation, we can construct a function by calculating the mean and standard deviation for each parameter to produce a normal distribution from which we can sample starting positions. Each parameter will have its own distribution from which the starting positions for the MCMC walkers are sampled. If random starting positions are chosen, then a uniform

distribution of the parameter space is sampled in order to create the starting positions for each walker. Upon creating the starting positions, we then invoke the MCMC inference which will need to be able to access the SQL database as described in section II.2.3.

For the sake of storing the inputs and outputs from UCLCHEM and RADEX, we use an SQL database using the SQLite implementation. This is chosen as it is a light weight, widely used, and easy to implement solution for storing large volumes of data in such a way that it can easily be queried. The main advantage of having access to an SQL search method is that it can quickly check if the combination of parameters to be calculated has been previously stored. This is complemented by a grid based parameter space for UCLCHEMCMC, which limits the potential values of parameters which are set by the user. When checking the database, the grid cell of the parameter and the corresponding central value are used as the representative value for the model. If the requested combination of parameters is in the database, then the program goes directly to evaluating the likelihood using equation II.1, which takes an almost negligible amount of time to perform. If the combination of parameters is not in the database, then a model can be created and stored for that set of parameters. This means that the calculation speed of the MCMC inference is dependent not only on the number of walkers and desired steps, but also on how many models are stored in the database. The SQL tables store the parameters given to UCLCHEM in one table with an associated ModelID, which is a unique integer value. The remaining tables store the output from individual molecules, where each row has the ModelID of the corresponding model that was used to calculate the inputs to RADEX. By separating the tables in this way, a single table with few columns is searched to identify if a model already exists in the database. If this is the case, the ModelID is used to then query the tables of molecules relevant for the current inference. This process takes less than a second to be performed when using SQLite. If chemical and radiative transfer modelling emulators were to be added, they could take the combination of parameters not already present in the database and calculate results more quickly instead. The implementation of such emulators goes beyond the scope of this work.

Based on that, UCLCHEMCMC will become faster as more inferences are

performed. The calculation of a single model can take around one minute when using UCLCHEM and RADEX. While this can be parallelised, it is still a limiting factor when thousands of models have to be calculated to get a reasonable estimation of a PDF. On the contrary, the action of submitting a query to an SQL database and evaluating the probability of the stored models matching the observations takes less than a second. Improving efficiency this way has the advantage over techniques such as emulation (de Mijolla, D. et al., 2019) as no approximation is made. To quantify the improvement, we measure the time it took ten walkers, to perform one hundred steps, at three different times: (i) When no database is being used; (ii) When around half of the models the inference wants to use are retrieved from the database; (iii) When nearly all models the inference is using are retrieved from the database. We use this type of measurement for the performance, as the minimum time, the time when every model can be retrieved from the database, should be identical irrespective of the chemical and radiative transfer model that is used. For the three cases, the mean time and standard deviation are: (i) $5185.33\,\mathrm{s} \pm 1041.96\,\mathrm{s}$; (ii) $4834.67\,\mathrm{s} \pm 843.24\,\mathrm{s}$; (iii) $68.89\,\mathrm{s} \pm 45.39\,\mathrm{s}$. We emphasise that the times found for case (i) and (ii) are strongly dependent on the chemical and radiative transfer models that are used, while case (iii) should only be weakly dependent on which models are used, with the dependency disappearing if all models the inference wants to use are within the database. The relatively small decrease in performance time between case (i) and case (ii) arises from the walkers waiting for each other to finish prior to taking another step; this means that the time of taking a step is dependent only on the walker that takes the longest to calculate its likelihood.

The database can be accessed both by the code, and by a user who wishes to use the models stored within it for other purposes. At the time of the first release, we store all inputs that are given to the chemical model when it is run, as well as all outputs that are produced by it. This can include the output of intermediate time steps which UCLCHEMCMC can be set up to store if requested. UCLCHEMCMC then takes the given line width, either as a free parameter of the MCMC or as a constant value given by the user, as well as the kinetic temperature, volume density of H_2, and the fractional abundances of the atomic or molecular species from the chemical code in order to run the radiative transfer code. The outputs from this

code are then stored in the SQL database. From here, the emission lines given by the radiative transfer code can be compared to the observations to evaluate the likelihood as discussed previously.

II.2.4. Interface

The User Interface (UI) for UCLCHEMCMC is browser based, in order to give a simple usable interface that should be compatible with most operating systems. A further advantage of this is that an online, publicly available version can be more easily created in the future such that end users will not need to change the workflow.

The inputs that are requested from a user of UCLCHEMCMC are separated onto three pages within the UI. The first page requests the ranges of the physical parameters over which the inference should be performed. At the time of the first release, the configured parameters are: (i) the volume density of the gas in cm^{-3} at which point the model should stop collapsing; (ii) the kinetic temperature of the gas in K; (iii) the cosmic ray (CR) ionisation rate in units of the galactic CR ionisation rate (ζ_0); (iv) UV radiation field strength in units of Habing; (v) the radius of the assumed spherical cloud being modelled in parsec (R_{out}); (vi) the line width to be used with RADEX in units of $km\,s^{-1}$. Upon supplying the desired ranges and which parameters should be kept constant, the next set of inputs is the observations. Here, a list of species can be selected to be added to the current inference. Once a species has been selected, the compatible lines will be shown and can be selected, after which it is possible to add the values of the observations, errors and the observed quantity. As of the first release, UCLCHEMCMC is configured to allow for the units that RADEX has as outputs, detailed in section II.2.1.

The penultimate page contains the options for the MCMC inference. The options are: the MCMC details, walker starting positions, and grid type. There are three options for the MCMC algorithm that an end user can easily change and they are: (i) the number of walkers that the inference should have; (ii) the number of steps the inference should perform before saving; (iii) the name of the session. Naming the session allows for an inference to be started again at a later time without having to re-enter all the previous parameters and observations.

Table II.2. Parameter Ranges

Parameter (Units)	Lower Bound	Upper Bound
Volume Density (cm^{-3})	5.0×10^4	1.0×10^7
Kinetic Temperature (K)	5	20
UV radiation field (Habing)	0.1	10
R_{out} (pc)	0.0001	0.1
CR ionisation rate ($1.3 \times 10^{-17} s^{-1}$)	0.1	10

Note. — Physical Parameter range the inference is allowed to explore for both the mock and observational inference.

This was added, in case the code crashes, or if after evaluating the results it was determined that the MCMC walkers could benefit from more steps to ensure the walkers converged. The starting positions of the MCMC walkers can either be randomly determined or set to inform starting positions depending on the end users preference. The details of how informed starting positions are calculated are given in section II.2.3. The grid type option allows an end user to choose which physical parameter space grid they want to use for the inference they are going to run, and are intended to be created and managed by the end user. By default, there is a coarse and fine grid provided to give an example of how they are meant to be created. The discretisation of the parameter space to grids was implemented as chemical models with physical parameters that differ only to a small degree would produce nearly indistinguishable outputs but would be considered separate models by the code that retrieves and stores models in the SQL database. This would lead to models with minor differences in parameters space being calculated despite producing indistinguishable outputs.

II.3. APPLICATION

In order to give an example of UCLCHEMCMC, we run three inferences. One inference is on mock data which were created by using UCLCHEM and RADEX, as these two codes are used in UCLCHEMCMC to perform the inference. The second and third inference are for the prestellar core L1544 (Caselli et al., 2002), once

Table II.3. Mock Data used for Evaluation

Species	Transition	Freq. (GHz)	Line width ($km\,s^{-1}$)	RADEX Value	Mock Data (T_{MB})	Units
CS	2,0 - 1,0	97.98095	1.0	2790.9	2412.4 ± 558.2	mK
SO	2,2 - 1,1	86.09395	1.0	1918.6	2553.0 ± 383.7	mK
	2,3 - 1,2	99.29987	1.0	450.3	367.6 ± 90.0	mK
	3,1 - 2,1	109.2522	1.0	185.5	222.3 ± 37.1	mK
o-H_2CS	$3_{1,3} - 2_{1,2}$	101.4778	1.0	130.6	151.6 ± 26.1	mK
	$3_{1,2} - 2_{1,1}$	104.6170	1.0	85.5	83.8 ± 17	mK

Note. — Values of the mock data created for evaluation of UCLCHEMCMC using UCLCHEM and RADEX. The RADEX Value column contains the value given by RADEX, while the Mock Data column contains the same values with added Gaussian noise, and corresponding error values.

Table II.4. Observations used for evaluation

Species	Transition	Frequency (GHz)	Line width ($km\,s^{-1}$)	Observation (T_{mb})	Units
CS	2,0 - 1,0	97.98095	0.64±0.07	1226.5 ± 0.1	mK
SO	2,2 - 1,1	86.09395	0.42±0.01	223.7 ± 5.9	mK
	2,3 - 1,2	99.29987	0.45±0.01	1422.5 ± 40.6	mK
	3,1 - 2,1	109.2522	0.39±0.01	176.1 ± 5.2	mK
HCS+	2-1	85.34789	0.43±0.01	246.8 ± 6.1	mK
OCS	6-5	72.97678	0.36±0.01	106.3 ± 6.8	mK
	7-6	85.13910	0.38±0.01	87.4 ± 7.2	mK
	8-7	97.30121	0.37±0.01	70.5 ± 5.4	mK
	9-8	109.4631	0.34±0.04	49.4 ± 6.4	mK
o-H_2CS	$3_{1,3} - 2_{1,2}$	101.4778	0.44±0.01	558.4 ± 11.7	mK
	$3_{1,2} - 2_{1,1}$	104.6170	0.44±0.01	514.3 ± 12.0	mK
p-H_2CS	$3_{0,3} - 2_{0,2}$	103.0405	0.45±0.02	536.9 ± 16.5	mK

Note. — Observations collected from (Vastel et al., 2014) and (Vastel et al., 2018), "o-" and "p-" represent ortho- and para- version of species respectively.

considering the emission from only one molecular species and once with all sulfur bearing species. The data we used for L1544 can be found in table II.4 and were used to infer the kinetic temperature, volume density, CR ionisation rate, and R_{out}. This object is a very well studied prestellar core located at R.A. = $05^{h}01^{m}11^{s}.0$, Dec=25°07′00″ (Caselli et al., 2002) (Vastel et al. (2014), Punanova et al. (2018) and Vastel et al. (2018)).

We use the same input parameter space for all inferences. The exception is the UV radiation field which we hold constant for the inferences on L1544 as we expect the visual extinction to be sufficiently high for changes in the UV to be negligible. The ranges for the physical parameters can be found in table III.2.

II.3.1. Mock Data Inference

First, we verify that UCLCHEMCMC performs as intended by creating mock data using the same modelling codes that UCLCHEMCMC uses to perform an inference. We add Gaussian noise to the data with a standard deviation of five percent for each emission line as this would make the mock data errors slightly higher than the average of the uncertainties on the L1544 observational data. We do this because running an inference where the true values are known and the data are model generated allows us to test whether UCLCHEMCMC performs as intended when the models are appropriate for to the data. As this is just an example case, and many different combinations of chemicals and transition lines could be picked, we choose a subset of emission lines from the observations we use for the inferences on L1544. In order to create the mock data, we randomly chose physical parameters that resulted in all emission lines having an observable flux. The parameters we chose are as follows: Final volume density $1.0 \times 10^5\,\mathrm{cm}^{-3}$, a kinetic gas temperature of 13 K, radiation field of 3 Habing, a cloud radius of 0.08 pc, and a CR ionisation rate value of $2.6 \times 10^{-17}\,\mathrm{s}^{-1}$. The emission lines and corresponding mock data values are found in table II.3, which contains the exact values of each line given by UCLCHEM and RADEX prior to adding noise, as well as the data with Gaussian noise added to it and the corresponding uncertainties.

We run the inference, monitoring the chain of steps that each walker has taken. The likelihood of accepting a new set of parameters decreases as a function of

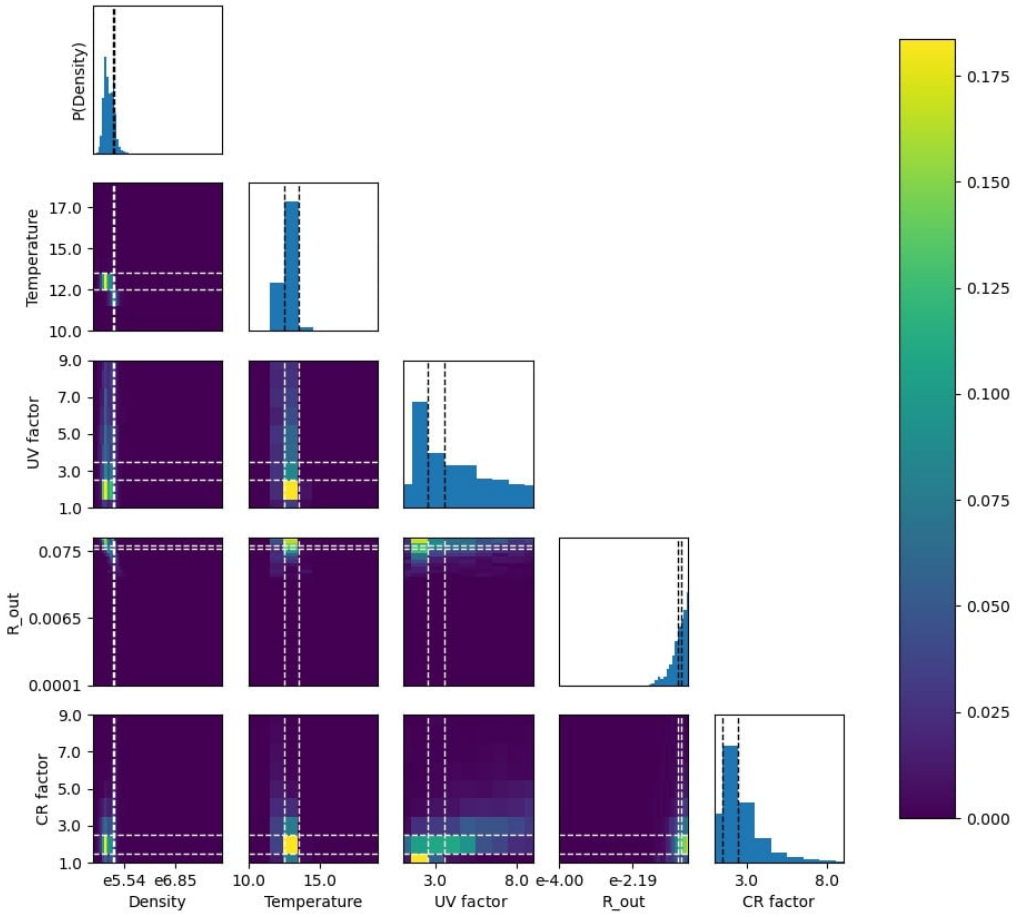

Figure II.3 Posterior distribution function of the evaluation run performed on mock data. The histograms represent the PDF of volume density, kinetic temperature, UV field factor, R_{out} and the cosmic ray ionisation rate factor, the colour bar shows the value ranges of the joint distribution functions. The white dashed lines in the joint distributions, and the black dashed lines in the PDFs represent the true value used to create the mock data.

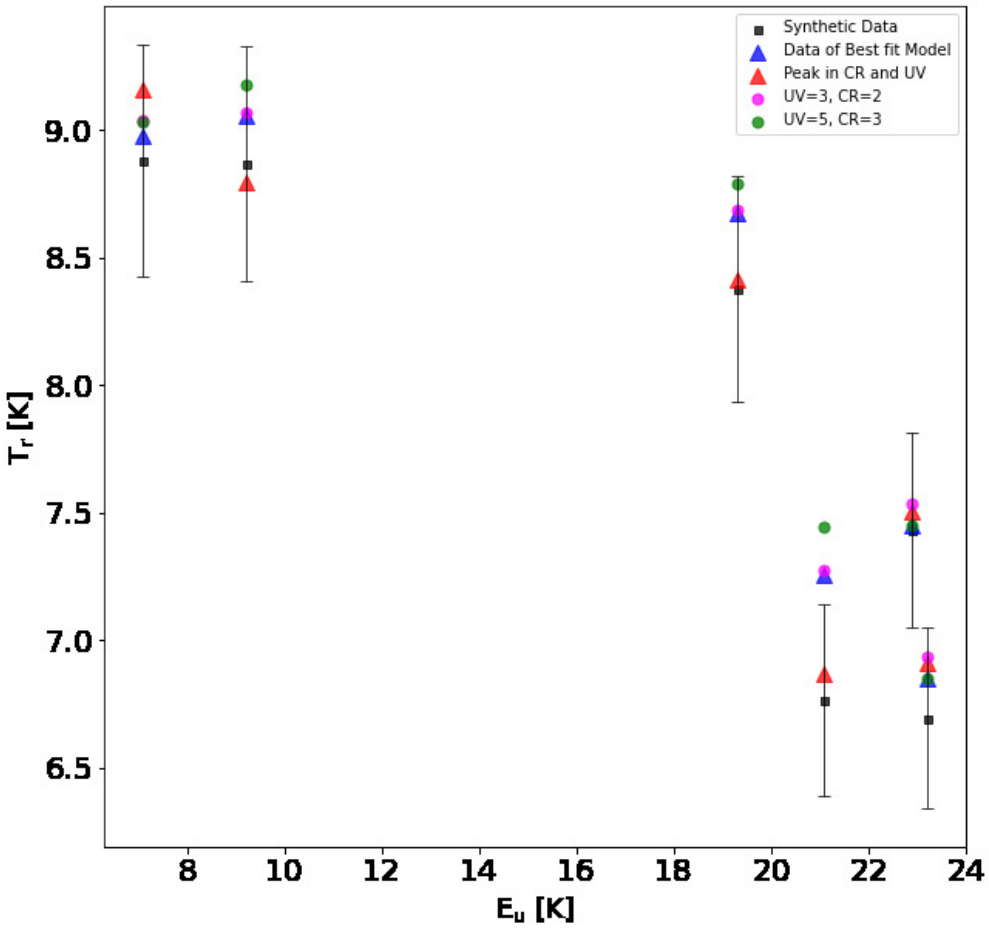

Figure II.4 The radiation temperature,T_R, calculated by RADEX against the Energy of the upper state, for the mock data and errors given to UCLCHEMCMC in black, and the data created when using the most likely parameter values from the 1D distributions from the inference of the mock data in blue. Red represents the peak in the joint distribution of the CR ionisation rate and UV radiation field while keeping the remaining parameters as they are for the previous model, while green and fuchsia represent two additional points with values for the CR ionisation rate and UV radiation field values in the elongated distribution of likely values to show why the inference still gave some importance to these values.

the difference between the likelihood of the current model and the new model. This means that over time, the parameter space that is being traversed by all walkers, will decrease as the walkers find areas of parameter space where the set of parameters produce models with a higher likelihood. Once the walkers stop reducing this parameter space, we stop the inference. Using these chains from the inference, we then create the posterior, shown in Figure II.3. The distributions contain the true values, which indicates that UCLCHEMCMC works as expected. In order to validate this, we plot all mock observation lines against the upper state energy, seen in Figure II.4, and do the same for the model values that are produced from UCLCHEMCMC's parameters with the highest likelihoods in the 1D distributions. When we do this, we see that all but one line lies within the uncertainties of the mock data.

We note that in the posteriors, there is a considerable degeneracy between the UV field and the CR ionisation rates. Additionally, there is a clear peak in the joint distribution of CR ionisation rate and UV radiation field. This peak is at a CR ionisation rate equal to the galactic value (CR=1) and at a UV radiation field strength of two Habing, however this peak does not have the same value of the CR ionisation rate as the parameter set used to generate Figure II.4 as that takes the most likely value from the 1D marginalised distributions rather than the overall most likely parameter set. To see how the emission lines change along this extended distribution, and how the observations look at this peak in the joint distribution, we include the emission lines of this peak, and two additional combinations of CR ionisation rate and UV radiation field strength in Figure II.4, while holding all other parameters constant. In looking at how the antenna temperature of the emission lines compare between the models and the mock data, it becomes quite clear that the values of the observations in this distribution all show significant agreement with the mock data but that the peak in the CR ionisation rate and UV radiation field strength distribution produces observations that fit better than the model created using the most likely parameter values in the 1D distribution.

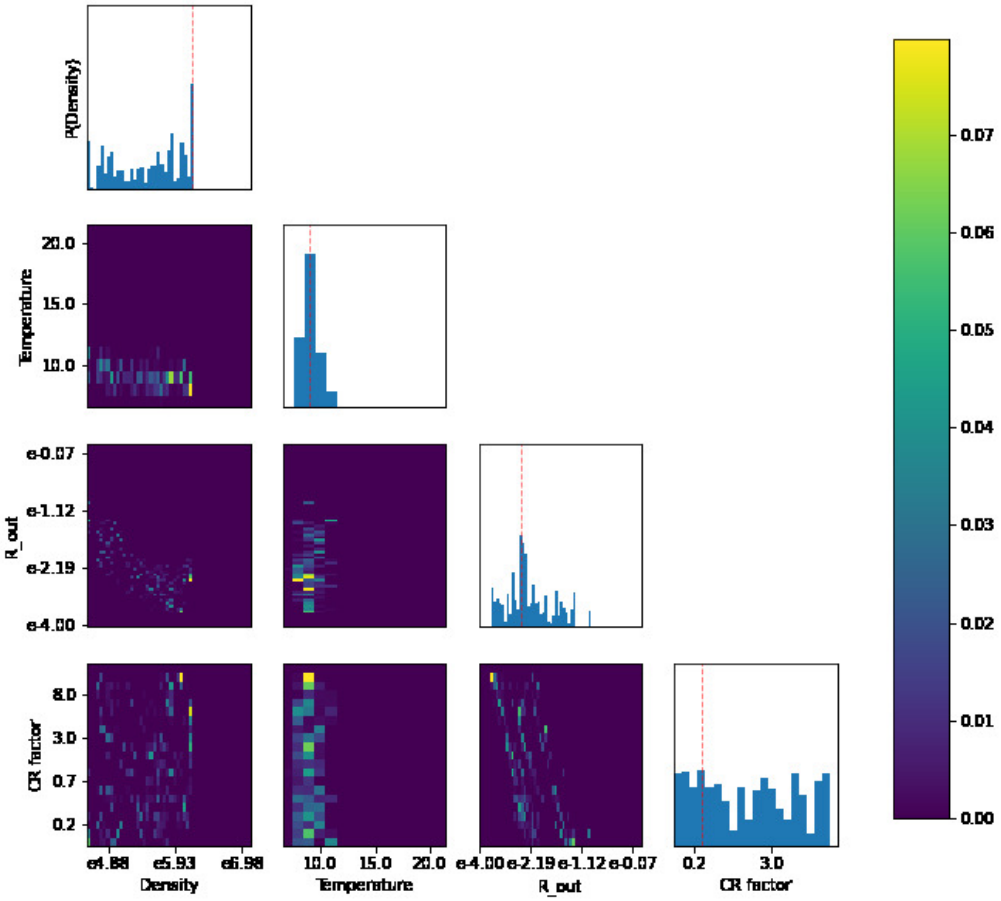

Figure II.5 Posterior distribution function of the evaluation run performed on the emission lines from OCS only. The histograms represent the PDF of volume density, kinetic temperature, R_{out} and the cosmic ray ionisation rate factor, the colour bar shows the value ranges of the joint distribution functions, while the red dashed line in the PDF is the value with the highest probability.

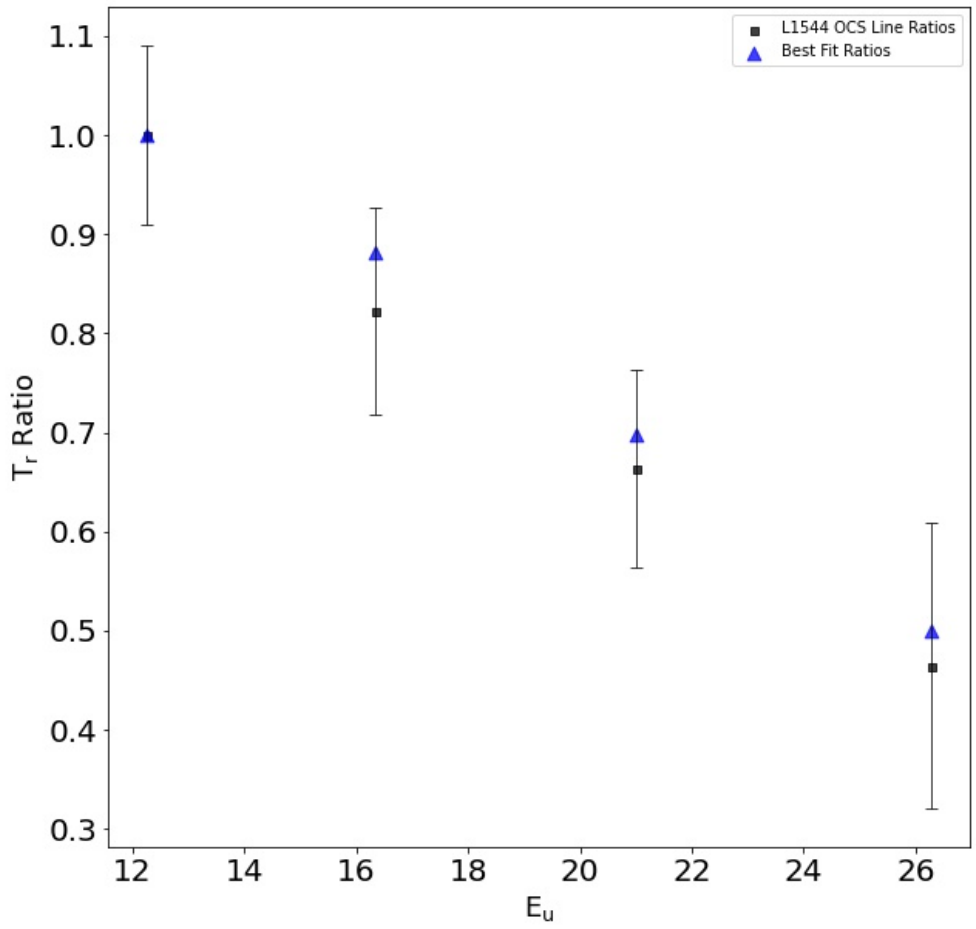

Figure II.6 T_R over T_R of OCS 6-5 against the upper state energy, for emission lines of OCS found in table II.4, compared to the data of the best fit model after running an inference using only the OCS lines. All of the lines fit the observed line ratios quite well. Black represents the real data with error bars, while blue is the best fit model.

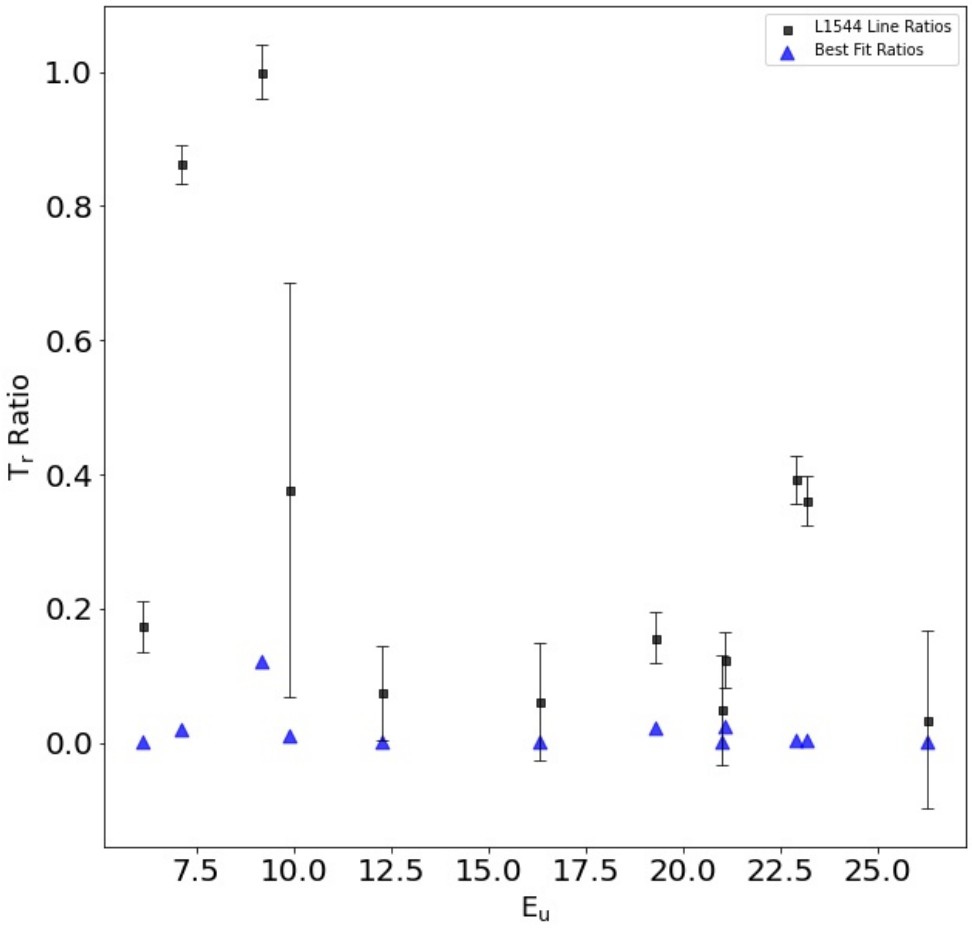

Figure II.7 T_R against the Energy of the upper state, for all emission lines in table II.4, compared to the data of the best fit model after running the stress test inference. While a couple lines almost fit their observed counterparts, it is clear that UCLCHEMCMC is unable to match all lines at once, which made it settle for a set of parameters, that allow each line to at least get somewhat close to the observations. Black represents the real data with error bars, while blue is the best fit model.

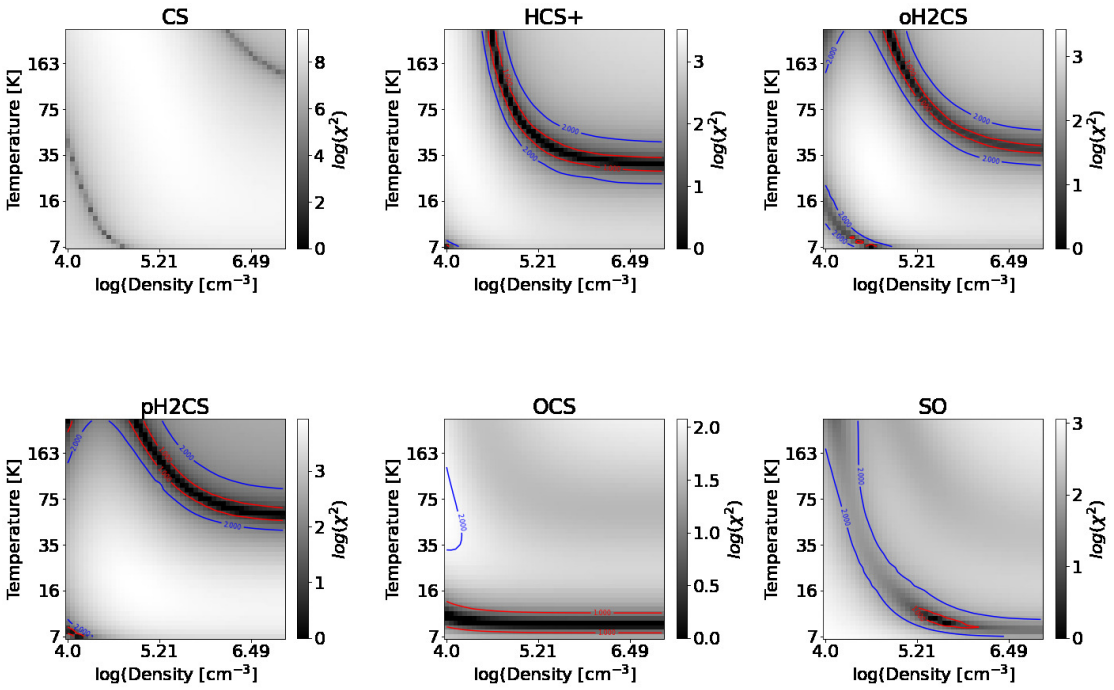

Figure II.8 $\log(\chi^2)$ grid for kinetic temperature and volume density using column density from Vastel et al. (2018) for the six species that are used for the MCMC Inference. The lower value of the $\log(\chi^2)$ is fixed at 0 to allow for better comparison between each species while allowing a flexible upper end, as the large ranges of $\log(\chi^2)$ values make it difficult to make an informative Figure with a single range of values.

II.3.2. Inferring parameters of L1544

With verification of how well UCLCHEMCMC can perform when using mock data, we now use real observations in order to run another inference. The observations we use are from Vastel et al. (2018), which present observations for two distinct regions in the L1544 object. One region is a methanol shell that is around 8000 AU (Vastel et al., 2014) from the core and in this shell UV photons can desorb methanol from the dust. The other region is a dust peak situated at the centre of the object. We perform two inferences on the dust peak as there is a collection of sulphur bearing species where we start by using only the emission lines of OCS. In this first inference, we start with very broad priors and do not include additional information on the priors to show how UCLCHEMCMC can perform. In an inference that is not designed to showcase the performance, any prior knowledge would be used to inform the ranges of the priors. We then follow that with an inference using all sulphur bearing species, to serve as a test to inform us of how well of an inference can be performed when UCLCHEMCMC is given unfavourable conditions. More chemical species and emission lines present potentially unfavourable conditions. Unless well understood, adding additional emission lines and chemical species adds to the risk that not all the species share the same physical parameter space, meaning UCLCHEMCMC may struggle to find a single set of parameters that fit all lines at once. For reference of values we expect the inference to estimate, we turn to Vastel et al. (2018) who model the kinetic temperature and the volume density of the dust peak to be around 7K and $2 \times 10^6 \mathrm{cm}^{-3}$ respectively.

As this is an example case of how to use UCLCHEMCMC, we leave a large physical parameter range for the volume density, $\mathrm{R_{out}}$, and CR ionisation rate value. We limit the kinetic temperature to be between 5 and 30 K, as at this temperature, a single degree can make a difference to the diffusion and desorption rates of various species, impacting the fractional abundances of different species. The only two exceptions we make on limiting parameters is leaving the UV radiation field strength at the default value for UCLCHEM, as it is not a parameter of interest for the dust peak of a pre-stellar core, and setting the line width to the error weighted mean of $0.37\,\mathrm{km\,s^{-1}}$ for the OCS only inference.

We follow this, by using all of the chemical lines found in table II.4 for a second inference of the dust peak. We intentionally use all of these lines as a stress test. For the second inference, we make the assumption that these species trace the same substructure, which we emphasise in the inference by setting all line widths to $1.0\,\mathrm{km\,s^{-1}}$ in UCLCHEMCMC. This could lead to an inference that is unable to fit the observations as UCLCHEMCMC could struggle to find one set of parameters that lead to emission lines that match the observations.

The posterior of the limited inference, can be found in Figure II.5. The distribution has a very broad range in volume density and the CR ionisation rate value, while having a strong peak in kinetic temperature and peaked area in R_{out}. This suggests that there is a wide range of possible volume density and CR ionisation rate values that can describe the observations well.

In order to choose a good fit, we use previously modelled values of the total column density along with the fractional abundance of hydrogen to constrain the gas volume density. Caselli et al. (2002) modelled a total column density of $4.4 \times 10^{22}\,\mathrm{cm^{-2}}$ which we combine with the peak value of the R_{out} posterior to obtain a likely volume density of $10^{6}\,\mathrm{cm^{-3}}$.

To validate that this is a good fit, we plot the emission line ratios, with respect to the most intense line OCS 6-5, against the upper state energy, to get Figure II.6 to create a diagram analogous to a rotation diagram. This diagram shows that the modelled line ratios, are within the estimated error bars of the observed line ratios, which supports the accuracy of the inference performed. A useful next step would be to remove volume density as a free parameter, and use the measured column density to calculate it from R_{out} during another round of inference, with a finer grid in parameter space. Since this is just an example case, we will instead move on to the stress test of UCLCHEMCMC.

Prior to running this test, we calculate a χ^2 fit of volume density and kinetic temperature using only the radiative transfer code RADEX, to serve as a baseline comparison to the performance of UCLCHEMCMC on the observed data, as this is a more common approach. To perform this fit, we take the column densities, determined by Vastel et al. (2018) through radiative transfer modelling, for each species in table II.4, and run RADEX on a grid of kinetic temperatures and volume

densities. Results of this fit can be seen in Figure II.8. The χ^2 fit is unable to find one set of parameters that agree with each other for all lines. It also accepts a very large area of parameter space as potential fits to each individual species, severely limiting how helpful this fit is to any modelling effort. Beyond that, this method requires that we either provide a column density estimate, or that we include a grid of column densities over which to calculate, which would significantly increase the calculation time as it would be adding an additional dimension to the parameter space. We note however, that the speed at which these calculations was performed is at least three orders of magnitude lower than a traditional MCMC inference that does not use an SQL database.

We perform the stress test inference with the observational data by using all species and emission lines found in table II.4. As is the case with the χ^2 fit, the stress test inference is unable to match all lines at once. The area onto which the MCMC inference converges is a delta-like distribution on a single set of parameters that we then use to model the emission lines. We again plot the emission line ratios against the upper state energy, this time with respect to the SO (2,3)-(1,2) line as it is the strongest line, resulting in Figure II.7. In this Figure it is clear that while one or two of the line ratios fit the observed ratios, the vast majority of lines from the best fit model do not match the observations at all. This failure is expected as UCLCHEMCMC assumes a simple homogeneous model should fit the observations. As more species and transitions are added, the assumption of a simple homogeneous model will be broken. We include this example of a failed fit, to assist users of UCLCHEMCMC in understanding some of the limitations and more importantly as a cautionary note when trying to fit multiple molecular transitions with one single gas component.

II.4. SUMMARY

The publicly available MCMC inference and SQL database managing tool, UCLCHEMCMC, is capable of inferring physical parameters of astrochemical observations. This paper presents the details necessary to understand the use, strengths and shortcomings of this tool. The management of the database, using SQLite, increases the efficiency of parameter inference as the tool is used. Using

the MCMC inference package, emcee, as well as having decoupled the chemical code and radiative transfer code from the inference, also makes UCLCHEMCMC capable of handling any other chemical modelling or radiative transfer modelling tool.

We showed the outputs of UCLCHEMCMC when inferring the physical parameters of mock data created using UCLCHEM and RADEX, detailing just how well this recovered the physical parameters of the mock data. The use of the SQL database in the inference has showed that, once most models the inference looks for are in the database, UCLCHEMCMC goes from taking 5185.33 ± 1041.96 s for ten walkers to take one thousand steps to needing 68.89 ± 45.39 s which is a significant decrease in computational time. When inferring the physical parameters of actual observations, we detailed some of the issues that must be taken into consideration when running this tool. Users should be aware that if they keep the physical parameter ranges too small, then the inference may not be able to find matching parameters, resulting in non physical answers. We intend to add a fast prior predictive checking functionality to UCLCHEMCMC. Additionally, giving too many emission lines, without taking into consideration that they may arise from separate structures, or that the combination of emission lines require physical parameters outside of the inference range, can lead to UCLCHEMCMC being unable to find physical parameters that match all lines. Because of this, we advise caution on using a long list of emission lines without first studying if those lines come from regions within the object that have similar physical parameters, as this tool will assume they all come from one structure with one set of physical parameters. When taking these factors into consideration, UCLCHEMCMC can be a great asset in inferring physical parameter ranges in which to start modelling astrochemical environments.

CHAPTER III

SVS 13-A

The work in this chapter is based on a paper that is currently being circulated to co-authors. The paper is authored by M. Keil, in collaboration with S. Viti, C. Ceccarelli, E. Bianchi, and C. Codella, with the work that is the focus of this chapter having been conducted by M. Keil, with supervision and guidance by S. Viti and C. Ceccarelli and assistance, access and detailed descriptions of the data used for this work being provided by E. Bianchi and C. Codella.

III.1. INTRODUCTION

As star forming regions evolve, the chemical richness of the gas, initially mainly atomic, becomes more complex. This complexity stems from the chemical reactions forming a large variety of molecules as the density and the temperature of the gas increase (van Dishoeck & Blake, 1998). Different chemical reactions occur in various regions of the Interstellar Medium (ISM) as some environments may prove hospitable to some reactions, while rendering other reactions unlikely (Chuang et al., 2017). In this context, studying different environments and understanding the abundance of various species is important. To do this, we turn to molecular line emission. These lines, and their ratios with respect to each other, provide us with a wealth of information on the observed environments. Common targets for these types of observations, are star forming regions within molecular clouds.

The densest areas of molecular clouds can host star forming regions, known as prestellar clouds, which will go through several evolutionary stages. The abundance and presence of specific molecules can aid in tracing the different evolutionary stages (van Dishoeck & Blake, 1998; Evans, 1999; Herbst & van Dishoeck, 2009; Caselli & Ceccarelli, 2012).

One of the least well understood atomic species that should be relatively abundant, is that of sulphur (Martín-Doménech, R. et al., 2016). The observed abundances of sulphur in dark molecular clouds is far from the solar abundance of $S_\odot/H \;\; 1.5 \times 10^{-5}$ Asplund et al. (2009), and the cosmic sulphur abundance estimates. Examples of observations that have shown the [S]/[H] to be below solar can be found in Tieftrunk et al. (1994) and Phuong, N. T. et al. (2018). This has been referred to as "The sulphur depletion problem" by Martín-Doménech, R. et al. (2016), which states that the currently observable species do not account for the amount of sulphur that we expect to be in dense dark-clouds.

However, due to their high reactivity, sulphur bearing molecules can act as clocks in star forming regions (Charnley, 1997; Hatchell et al., 1998). One suggested such clock is the ratio of SO_2 to SO, and SO to H_2S which can be used to trace protostellar evolution as the H_2S on grain mantels will desorb as hot cores form. In the gas phase, this H_2S undergoes endothermic reactions to form either atomic sulphur or SO which can then be turned into SO_2. It has been shown in modelling these ratios, that they are also heavily dependent on the physical conditions, such as density and temperature of the protostellar core not just age, making them imperfect tracers for the age of a core (Wakelam, V. et al., 2004). However, suggestions to update the clock to instead use the OCS/SO_2 ratio have been made and supported with observations (Taquet et al., 2020), which gives additional importance to understanding where sulphur is stored in order to further our understanding of the molecular composition of star formation regions. The effects of the sulphur depletion problem previously mentioned should be minimal as the ratio of the molecular abundance is used, and each molecule would roughly equally affected by the depletion of sulphur.

Many observations of sulphur bearing species have been performed (see e.g. Vastel et al. (2014, 2018); Codella et al. (2021). These observations highlight that while sulphur bearing molecules are observed, they do not account for the abundance of sulphur that is predicted based on solar abundances (Martín-Doménech, R. et al., 2016). Many modelling efforts have addressed the depletion problem (Wakelam, V. et al., 2004; Laas, Jacob C. & Caselli, Paola, 2019; Woods et al., 2015) however, more observations are needed to validate the proposed solutions.

A particularly well studied nebula containing prestellar clouds and protostellar cores where sulphur bearing molecules have been observed, is NGC1333. One of the protostellar cores is SVS13-A which is categorised as a Class I prototypical source. This means it has an estimated age of $\sim 10^5$[yr] and has a bolometric luminosity of $\sim 50 L_{bol}$ (Chini et al., 1997). SVS13-A is also a binary source (Anglada et al., 2000; Tobin et al., 2016). The NGC1333 cluster itself is estimated to be at a distance of 299 ± 14 pc (Zucker et al., 2018), while SVS13-A is near to another object, VLA3, which sit together in a large molecular envelope (Lefloch et al., 1998). Various observations in different frequencies have been performed which iteratively contributed to our knowledge of SVS13-A (Chini et al., 1997; Bachiller et al., 1998; Looney et al., 2000; Chen et al., 2009). First detected in the infrared by Strom et al. (1974) the object that would later become SVS13 was first denoted to be near the region of HH 7-11, a reflection nebula near young stars (Herbig, 1974). Later, Looney et al. (2000) suggested that SVS13 may be embedded in the nebula linking the two objects. Observations using the Very Large Array (VLA) by Anglada et al. (2000) found evidence that SVS13 is a binary protostar system through the use of radio observations. Dhabal et al. (2019) suggests that SVS13-A is part of a group of protostellar objects that are forming because of the influence of an elongated system of turbulence events and De Simone et al. (2022) suggests these turbulent events are a train of shocks within the NGC1333 cluster.

Towards SVS13-A, observations of heavy water have revealed there to be a hot corino (Codella et al., 2016) which was later imaged by De Simone et al. (2017) using emissions from $HCOCH_2OH$ (glycolaldehyde). The term hot corino refers to the core of a low-mass protostar, in order to differentiate these types of cores from those found around high-mass protostars (Bottinelli et al., 2004).

In this work we use SVS13-A as a test-bed to understand which physical parameters affect the abundances of sulphur bearing species as many such molecules were detected in this object as part of the SOLIS (Seeds Of Life In Space)[1] (Ceccarelli et al., 2017) IRAM-NOEMA program. The SOLIS project initially aimed to image five organic molecules in six solar-like star-forming regions at various different stages in their evolution. The molecules of choice were methoxy (CH_3O),

[1] `https://solis.osug.fr`

methanol (CH_3OH), dimethyl ether (CH_3OCH_3), methyl formate ($HCOOCH_3$) and Formamide (NH_2HCO) all of which were in the SOLIS definition of interstellar complex organic molecules (iCOMs) which represents C-bearing molecules with six or more atoms, where interstellar was added in order to help differentiate between complex organic molecules in the ISM (Ceccarelli et al., 2017). The detection of additional molecules beyond the targetted iCOMs were included in the analysis of the observed targets. Examples of additional molecules detected within SOLIS include, but are not limited to, SO, CS, NS, OCS, H_2CS, (Codella et al., 2021), and cyclopropenylidene (C_3H_2)(Favre et al., 2018). We will first describe the previously published observations, the modelling tools, and inference method we use in order to perform this study in section III.2. After this, we describe the results of the modelling and inference effort we performed using the observed emission lines in section III.3. At the end, we conclude with a summary in section III.4.

III.2. METHODS

III.2.1. Observations

The observations we use to make the inferences are from Codella et al. (2021) and were taken using the IRAM-NOEMA interferometer as part of the SOLIS program. We used some of the emission lines of the following molecules: (i) SO; (ii) OCS; (iii) H_2CS; (iv) CS. Details of which lines we took and their observed intensities are shown in table III.1. The observations were taken at two different times using different configurations. The first observations were taken in March 2018 and covered frequency ranges of 80.2-88.3 GHz and 95.7-103.9 GHz. Here, the field of view was ≈ 60", with the Largest Angular Scale (LAS) being $\simeq 8$" and a spectral resolution of 2 MHz. This set of observations includes all of the emission lines for CS as well as the majority of the observed lines for SO. The second set of observations focused on the frequency range of 204.0-207.6 GHz, with a field of view of ≈ 24" and LAS off $\simeq 9$". These observations contained all the remaining lines. For more details of the observations, we refer to Codella et al. (2021). We chose these lines as they come from molecules that are sulphur bearing, as well as the fact that the emission maps, and non local thermodynamic equilibrium(LTE) large velocity gradient (LVG) analysis in Codella et al. (2021) suggest that they

Table III.1. Observations from Codella et al. (2021) that are used as the inputs for the UCLCHEMCMC inferences. (a) Frequencies have been obtained from the Cologne Database for Molecular Spectroscopy (Endres et al., 2016). (b) Errors in the intensity include uncertainties due to calibration of $\leq 10\%$ for the lines with frequencies between $80.2 - 103.9$GHz and $\leq 15\%$ for the remaining emission lines.

Species	Transition	Freq. [GHz][(a)]	$I_{\rm int}$ [K km/s]	Errors [K km/s][(b)]
SO	2_2-1_1	86.09395	27.69	0.23
SO	2_3-1_2	99.29987	44.3	1.99
SO	4_5-4_4	100.02964	9.55	0.27
SO	4_5-3_4	206.17601	422.0	7.38
OCS	7-6	85.13910	28.39	0.34
OCS	8-7	97.30121	40.73	1.18
OCS	17-16	206.74516	308.66	4.89
o-H_2CS	$3_{1,3}$-$2_{1,2}$	101.47781	17.12	4.50
o-H_2CS	$6_{3,3}$-$5_{3,2}$	206.05224	108.84	24.75
p-H_2CS	$3_{0,3}$-$2_{0,2}$	103.04045	10.49	2.26
p-H_2CS	$3_{2,1}$-$2_{2,0}$	103.05187	3.79	1.15
p-H_2CS	$6_{0,6}$-$5_{0,5}$	205.98786	57.35	0.56
p-H_2CS	$6_{4,3}$-$5_{4,2}$	206.00188	116.65	1.84
p-H_2CS	$6_{2,4}$-$5_{2,3}$	206.15860	48.87	0.49
CS	2-1	97.98095	38.72	0.62

Table III.2. The lower and upper bounds of the parameter range used as inputs to UCLCHEM, as well as how many grid points each parameter has.

Physical Parameter [Unit]	Lower Bound	Upper Bound	Grid Points
Density between collapses [cm^{-3}]	1000	100000	6
Final density [cm^{-3}]	1000000	100000000	5
Initial temperature [K]	6	14	9
Final temperature [K]	20	190	10
Time between collapses [yrs]	5000	1000000	8
Cosmic ray factor [$1.3 \times 10^{-17} s^{-1}$]	0.1	10	7
Radius of cloud [pc]	0.001	0.05	7
Fractional abundance of S	1×10^{-7}	1×10^{-5}	17
Fractional abundance of O	1×10^{-4}	5×10^{-4}	6
Fractional abundance of C	1×10^{-6}	3×10^{-4}	9

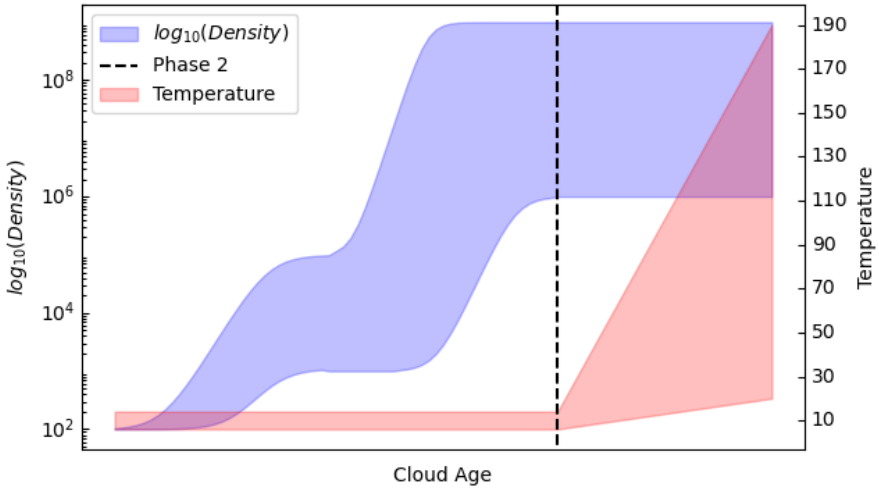

Figure III.1 Visualisation of the way in which we perform the chemical modelling. In blue, is the range within which the log_{10}(Density) can be. In red is the kinetic Temperature of the gas. The x-axis represents the time in years that were modelled by UCLCHEM, but does not represent the predicted age of an object. As each model can model for a various amount of time, depending on how long the collapse takes, as well as the time between collapses we intentionally forgo marking any numbers on the x-axis of this plot.

should originate from a similar area, and not be from separate components within SVS13-A.

III.2.2. Astrochemical Modelling

For this work we use the public code UCLCHEM (Holdship et al., 2017), which is a time dependent gas-grain chemical modelling code[2]. UCLCHEM uses rate equations in order to estimate abundances of different species in the gas and on the surface of grains. The default gas-phase reaction network is taken from the UMIST database (McElroy, D. et al., 2013), while the default grain-surface network is supplied with UCLCHEM. The latter network includes Eley-Rideal and Langmuir-Hinshelwood diffusion reactions as implemented by Quénard et al. (2018), competition reactions (Chang, Q. et al., 2007; Garrod & Pauly, 2011) as well as using reactions whose rates are determined by the binding energies Wakelam et al. (2017a). Beyond this, UCLCHEM also includes thermal and non-thermal desorption induced by cosmic rays, chemically induced desorptions and UV radiation induced desorption.

[2] `https://uclchem.github.io/`

UCLCHEM functions best by running distinct evolutionary phases. The first phase usually forms the molecular cloud by collapsing a parcel of gas from a diffuse cloud. The way in which it collapses is by default free-fall but can be altered. After this, more stages can be added using the chemical abundances estimated in the first phase as initial abundances.

In this work, we use UCLCHEM to first model the collapse of a molecular cloud. We follow this with a phase of "stagnation" where the gas experiences a time of no collapse which can have varying lengths before modelling a second collapse phase. At the end we then perform a phase of heating, mimicking the birth of the star. To illustrate the behaviour of the density and temperature as a function of modelled cloud age see figure III.1. The age in this figure does not represent the real age of the cloud, just that of the modelled time.

III.2.3. Radiative Transfer

In order to allow us to compare the outputs of UCLCHEM with the observations, we need to compute the line intensities from the various species. To this end, we use the radiative transfer, non-LTE LVG (Large Velocity Gradient) codes RADEX (van der Tak et al., 2007) and GRELVG (Ceccarelli, C. et al., 2003), respectively. The major difference between the two codes is how they deal with large line optical opacities τ. Specifically, GRELVG finds the solution by gradually increasing the species column density, which allows it to converge no matter how large τ is. Both codes, as any non-LTE LVG code, require as input the species column density, temperature and densities of the colliders, and the source size. All these quantities are obtained using the outputs of UCLCHEM. Finally, we use the collisional coefficients between the considered species and H_2 provided by the data base LAMBDA van der Tak (2011), which collected the ab initio quantum mechanic computations of the S-bearing molecules used for this work (SO: Lique et al. (2006), OCS: Green & Chapman (1978); Schöier, F. L. et al. (2005), H2CS: Wiesenfeld & Faure (2013), CS: Denis-Alpizar et al. (2018); Endres et al. (2016)).

III.2.4. Comparison of observations and model predictions

In order to perform the inference for this work, we use UCLCHEMCMC (Keil et al., 2022). This is an open source tool that uses Bayes' theorem (Bayes & Price, 1763; Joyce, 2021) for the likelihood function in order to infer physical parameters of an object for which spectral observations are available. The tool performs this inference using a Markov Chain Monte Carlo (MCMC) method, as implemented by the package Foreman-Mackey et al. (2013), while also providing an easy to use interface. It uses a full forward modelling approach by starting with a chemical modelling code (UCLCHEM) before passing the outputs to the radiative transfer code (RADEX or GRELVG). It is the outputs from the radiative transfer code that UCLCHEMCMC uses in combination with the given observations to calculate the likelihood used by the MCMC inference. Beyond this, UCLCHEMCMC stores previously calculated models so that it can retrieve previous models rather than recalculate them which is done to decrease calculation times. Future development plans for UCLCHEMCMC include adding automatic convergence testing, as well as the exploration of how non-detections could be flagged by UCLCHEMCMC.

For this work, UCLCHEMCMC was adapted to be able to use either RADEX or GRELVG as a radiative transfer modelling code in separate inferences, in order to test the performance of both radiative transfer codes compared to each other as well as to ensure that UCLCHEMCMC performs as intended when the radiative transfer code is changed from the default RADEX.

UCLCHEMCMC relies on a grid based physical parameter space which is configurable to fit the needs of the work which is being performed. In order to explore the physical parameters of interest for the sulphur bearing species of SVS13-A, the default parameters and parameter space of UCLCHEMCMC were altered to allow for a higher dimensional parameter space. As this would drastically increase the potential combination of models, we opted to use fewer grid points. The ranges were chosen so they would still be within acceptable ranges of Class I objects while allowing for large variations.

The parameters that we explore are: (i) Density after first collapse; (ii) Final density; (iii) Initial kinetic temperature; (iv) Final kinetic temperature; (v) Time

between the first and second collapse; (vi) Cosmic ray ionisation rate; (vii) Size of molecular cloud; (viii) Initial sulphur fractional abundance; (ix) Initial oxygen fractional abundance; (x) Initial carbon fractional abundance. All of these parameters are used as inputs to UCLCHEM. These were chosen so as to study which of these parameters will have a larger impact on the sulphur bearing species.

The number of grid points were chosen for each parameter, and they are detailed in table III.2. With the exception of the fractional abundance of sulphur, the grid points of each parameter were intentionally chosen to be coarse so as to reduce the number of potential models in the ten dimensional parameter space in this problem, while aiming to have enough grid points to resolve potential fits, which could be used to inform finer grid inferences in the future. For example, for the final density parameter, we initially chose to have two grid points per order of magnitude. However, we chose to remove the second to last grid point from the final density as at such high densities, the order of magnitude is the important value and removing a single grid point in such a high dimensional grid space reduces the combination of potential models significantly. For initial temperature we choose to have one grid point per Kelvin in the predefined range. The final temperature was split into steps of twenty Kelvin with exception of the first step going from twenty to thirty Kelvin. The seven grid points of the cosmic ray factor were chosen to cover coarsely the order of magnitudes from 10^{-18} s^{-1} up to 10^{-16} s^{-1} while still exploring a range similar to calculated cosmic ray ionisation rates of high-mass star-forming regions done by Sabatini et al. (2020). As the main focus of the study is sulphur, we chose the initial fractional abundance of sulphur to have the largest amount of grid points.

The two ways in which we use modelling to study the previously published emission lines in this work, are through a χ^2 grid, and the previously detailed tool UCLCHEMCMC. Both forms use the emission lines, and the same physical parameter ranges for temperature and final density. However, the χ^2 analysis only uses the radiative transfer codes and each grid is for individual molecules. All other parameters of the radiative transfer codes are kept constant for all molecules. The only exception is the column density, which has been set individually for each molecule to be the median value of the estimated ranges found in table 2 of

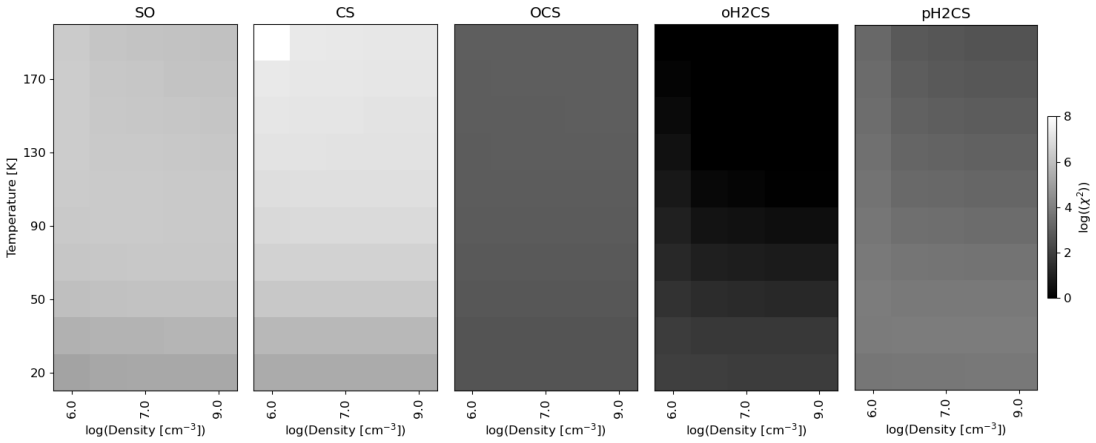

Figure III.2 χ^2 grid for each molecule using RADEX. The y-axis is the temperature and x axis is the log of the number density of H. The high temperature low density white corner is a model where RADEX failed to converge. The colour range is the log of χ^2 as the values were too high to be able to be shown without a log plot.

Codella et al. (2021) for each molecule. By separating the molecules into their own plots we can see if there is a clear distinction between each species on which temperature and density are most "preferred" by them. If there are clearly distinct areas with values of χ^2 close to one, but the parameter spaces of those low values are different in each molecule then that would be an indication that the molecules and by extension, the emission lines stem from different regions within SVS13-A, with different temperature and/or density. If the plot of a single molecule has no areas where the value of χ^2 is close to one, then this would mean that there is no model, within the parameter ranges we have chosen, where all modelled lines of that molecule have intensities close to the observed values. In this case, it could indicate that the different emission lines of the same molecule could come from different regions, or that they can not be described within the parameter ranges used to calculate the models. In order to differentiate the two cases, the model outputs of each individual line could be inspected or a χ^2 plot of each emission line could be created. If the χ^2 of each molecule has a distinct area with a low χ^2 value that overlaps with the areas of the other molecules χ^2 plots, then this would mean that there are some models within our chosen parameter that fit the full observations. This approach follows what was previously done in chapter II.

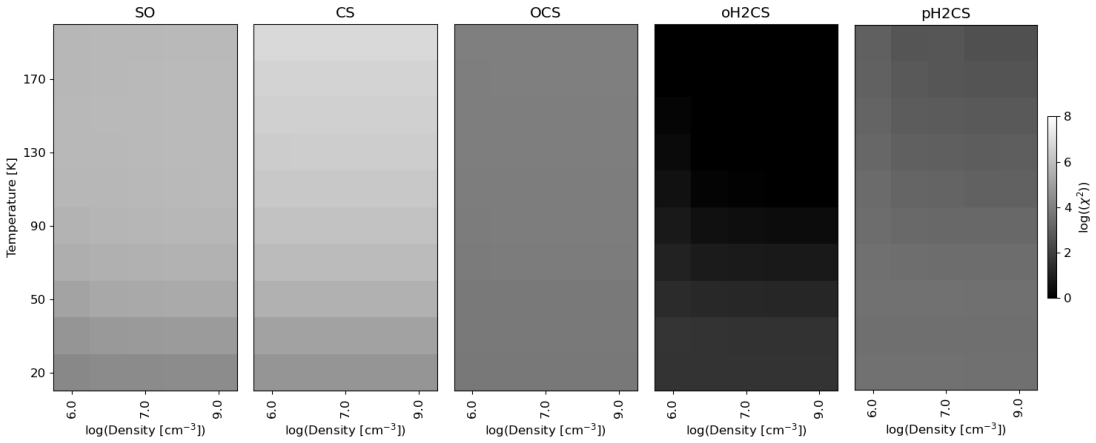

Figure III.3 χ^2 grid for each molecule using GRELVG. The y-axis is the temperature and x axis is the log of the number density of H. The colour range is the log of χ^2 as the values were too high to be able to be shown without a log plot.

III.3. RESULTS

Figure III.2 and figure III.3 show the results of the χ^2 analysis for RADEX and GRELVG respectively. They show that there is no unique solution for any of the molecules. This means that within the chosen physical parameter ranges no model can fit all emission lines of an individual molecule simultaneously. In the case of CS we only had one emission line, therefore, this would indicate that the one CS emission line intensity can not be described by the models in the given density and temperature ranges when we use the median of the estimated range of column densities found by Codella et al. (2021). For the remaining molecules, the lack of any area with low χ^2 value indicates that the individual emission lines either stem from different areas within the observed object or that the chosen parameter range is unable to reproduce the observed emission line intensities. This result shows that UCLCHEMCMC is unlikely to find a fit for all emission lines simultaneously as UCLCHEMCMC only uses a single point for chemical modelling with UCLCHEM. In order to verify if this is the case when we perform the chemical and the radiative transfer modelling, we use all of the lines from table III.1 for both inferences.

The UCLCHEMCMC inference using RADEX produced the posterior shown in figure III.4, with the GRELVG posteriors shown in figure III.6. Neither posterior shows a clear area of higher likelihood with a smooth decrease in likelihood. To understand more about the distributions, we turn to a chain plot of the walkers

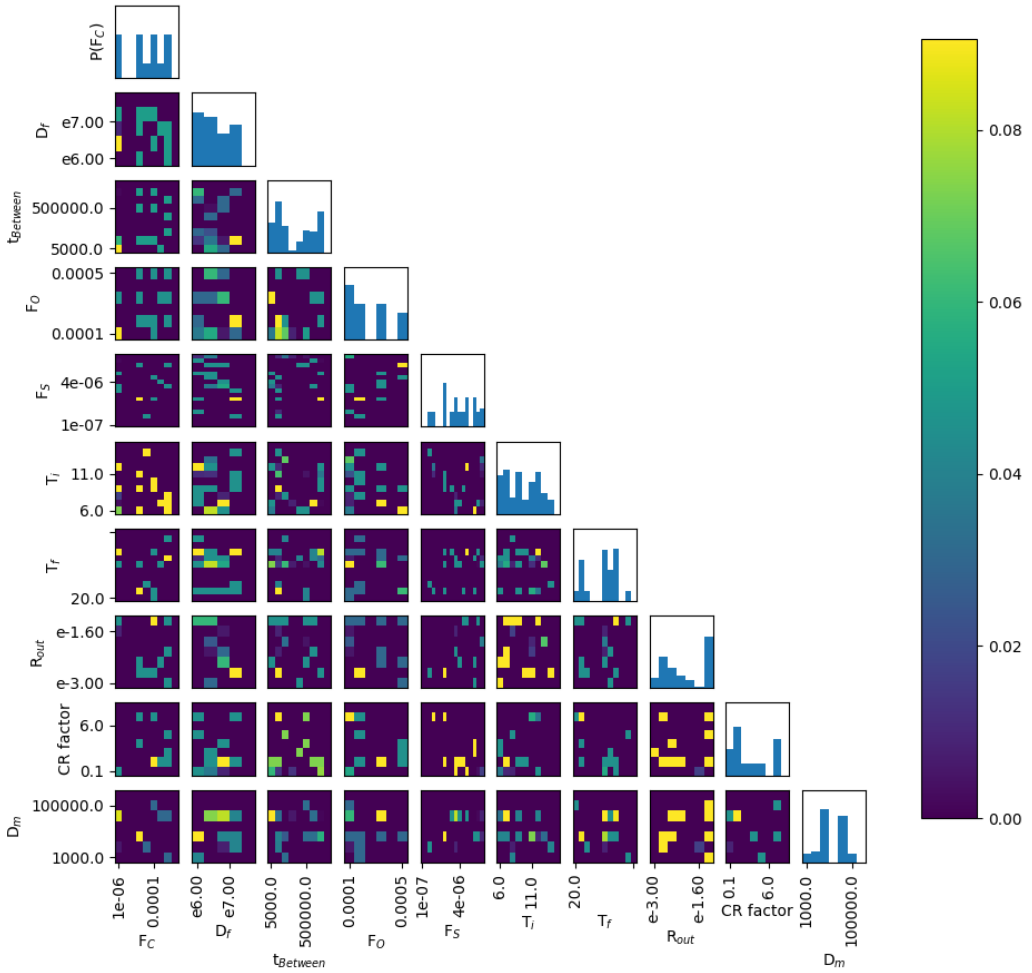

Figure III.4 Corner plot of the results from UCLCHEMCMC when using RADEX as the radiative transfer code. The colour-bar represents the value of the normalised posterior of the MCMC inference. In order, the parameters are: Density after the first collapse (D_M); Final density (D_F); Initial temperature (T_I); Final temperature (T_F); Time between collapses (t_m); Cosmic Ray ionisation factor; Radius of cloud (R_{out}); Fractional abundance of S (Frac.$_S$), O (Frac.$_O$), and C (Frac.$_C$).

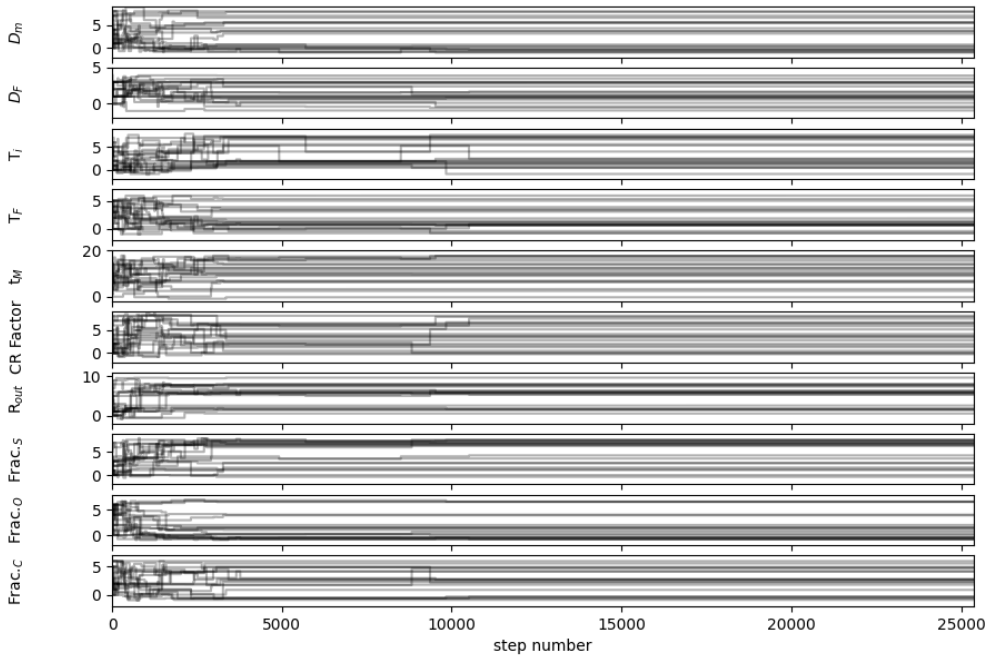

Figure III.5 Plot of the walker chains for the RADEX inference. The initial steps were walkers are rapidly varying are discarded when plotting figure III.4.

for the RADEX model in figure III.5. In this plot we can see that beyond step 12000, the walkers no longer change, which indicates that each walker has found a local likelihood maximum that is surrounded by values low enough that the walker is unlikely to take a step away from its current combination of parameters. This is the same for the GRELVG models. As this occurred with repeated runs of UCLCHEMCMC, we turned to inspecting the database of models created by UCLCHEMCMC. Inspecting the database of models and calculating the χ^2 for all models shows that the models with the lowest χ^2 value still have values of $> 10^3$ clearly showing that all models that have been created by the walkers are a very poor fit for the given observations. This is the case for both the RADEX and GRELVG model databases.

The inability to find any fitting models could stem from multiple reasons. On the technical side of UCLCHEMCMC, the grid based parameter space severely limits potential models and heavily relies on setting of a good parameter space and choice of grid points. While the work in chapter II showed that UCLCHEMCMC is able to fit mock data and find models with fitting line ratios in a case with few

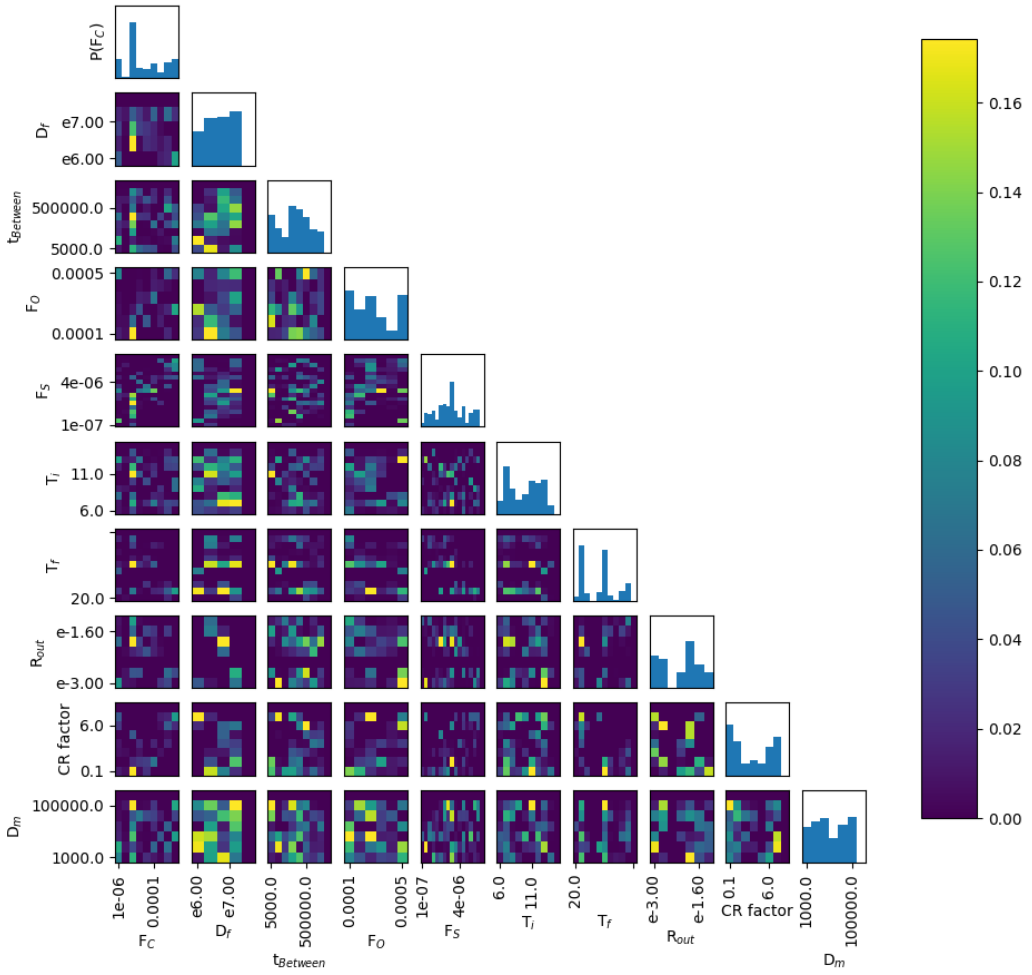

Figure III.6 Corner plot of the results from UCLCHEMCMC when using GRELVG as the radiative transfer code. The colour-bar represents the value of the normalised posterior of the MCMC inference. In order, the parameters are: Density after the first collapse (D_M); Final density (D_I); Initial temperature (T_F); Final temperature (T_F); Time between collapses (t_m); Cosmic Ray ionisation factor; Radius of cloud (R_{out}); Fractional abundance of S (Frac.$_S$), O (Frac.$_O$), and C (Frac.$_C$).

emission lines, the problems in that chapter were significantly simpler cases both in terms of chemical modelling and in number of parameters. This allowed each parameter dimension to have significantly more grid points.

It is also possible that the grid for this case was not fine enough to find a model that fit the observations. This could also explain the χ^2 plots in figures III.2 and III.3, which already suggested UCLCHEMCMC would be unable to fit all lines simultaneously with the given grid size.

III.3.1. Verifying UCLCHEMCMC

In order to verify if UCLCHEMCMC failed because of how coarse the grid was chosen to be, we turn back to calculating χ^2 values for all models. We do this using RADEX with a finer grid and for multiple column densities, rather than constraining the column density to previously published values. The choice to only use RADEX is done as the original χ^2 plots of the outputs of the two radiative transfer codes were near identical. We take the temperature range of 10 K to 300 K with a grid point every 10 K, expanding the range by 110 K. For the gas density, we refine the steps by splitting each order of magnitude into four grid points going from 10^5 to $e10^8\,\mathrm{cm}^{-3}$. The column densities are chosen to be between 10^{14} and $10^{18}\,\mathrm{cm}^{-2}$, where each magnitude is represented by three grid points with 3×10^n, 6×10^n, and $9 \times 10^n\,\mathrm{fm}^{-3}$ except for the final magnitude which only has the first. In doing so, we aim to provide a finer and wider grid in temperature and density while also adding column density as a parameter that changes. This broader and finer parameter space with an additional dimension should either find models that have overlapping low χ^2 which would support the idea of the grid limitations being the problem for UCLCHEMCMC, or it should show no model exists that matches all emission lines.

The new χ^2 plots can be seen in figure III.7. The scaling of the colours has been shifted from one to one hundred in log scale. This range makes it easier to highlight models that are a good fit from those that are not good fits. As previously seen, the ortho-H_2CS emission lines have the lowest χ^2 values, in many cases being close to one. As the column density diverges from $3 \times 10^{15}\,\mathrm{cm}^{-2}$, we see that less of the parameter space provides good fitting models, while some areas remain around one.

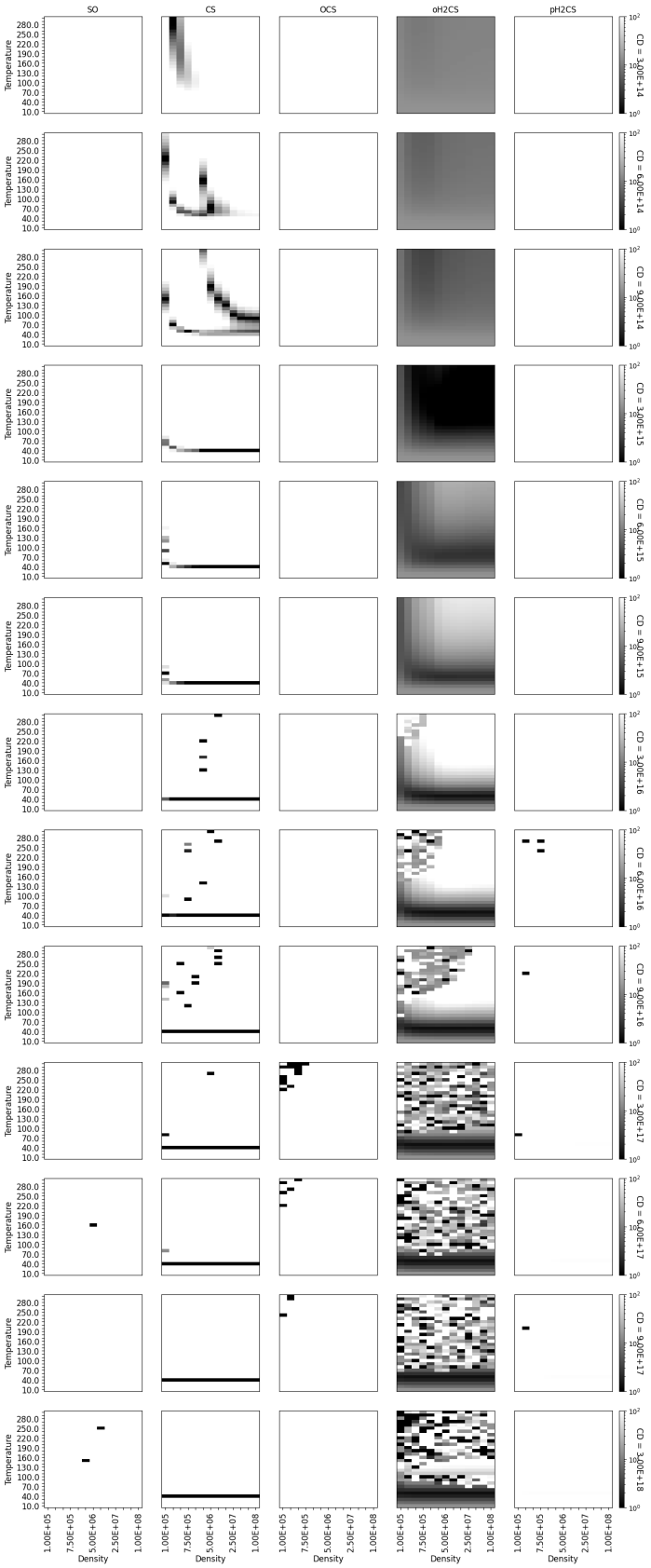

Figure III.7 χ^2 grid using RADEX with column density being varied. Each row is a different column density. The colour range maximum was set to one hundred, but is shown in log scale in order to make it easier to distinguish when models have values that would show them to be good fits or when they are not good fits.

For CS we can also see some areas with values closer to one, while not being as widely spread as those of ortho-H_2CS. Of note are the models at temperature 40 K which can be seen when the column densities for CS approach $1 \times 10^{15}\,\mathrm{cm}^{-2}$, which was previously not captured in the coarse grid that skipped 40 K. OCS shows some scattered points in parameter space where values are close to one while OCS, SO, and para-H_2CS gain individual points with low χ^2 value but these have minimal overlap with areas of good fit in other molecules, mostly just overlapping with ortho-H_2CS which has distributions so broad that they contain little information.

The results of the refined χ^2 parameter grid plots suggest that the parameter grid for UCLCHEMCMC could be refined, but that it is not the main cause of the UCLCHEMCMC walkers getting stuck. The lack of models that can fit all emission lines simultaneously in the refined RADEX grid suggests that even with this refined grid UCLCHEMCMC would struggle to find suitable models to match the observations. Instead the χ^2 values suggest that either the ranges of parameters are not broad enough or the combination of emission lines for OCS, pH_2CS and SO stem from the different regions within SVS13-A that have different densities and temperatures.

The lack of a good fit from UCLCHEMCMC and the additional χ^2 plots to verify that the chosen coarse grid is not the main cause of the UCLCHEMCMC fit failure indicates that the combination of emission lines cannot be fit with a single point using the combination of parameters we used for this study. Instead, more work needs to be done in order to identify which molecules and emission lines stem from the same regions within SVS13-A. Such a collection of emission lines could then be used for a new inference with UCLCHEMCMC to study the the impacts of different physical parameters on the sulphur abundance.

III.4. SUMMARY

Using the inference tool UCLCHEMCMC, we attempted to constrain the history and physical parameters of the source SVS13-A using sulphur-bearing species as our "tools". We performed two sets of inferences, one with RADEX and one with GRELVG as the radiative transfer codes. Both of these inferences lead to the MCMC walkers stopping at local maxima, that had very high χ^2 values. This

indicated a failure to fit the observations. Additional χ^2 calculations of a finer and expanded physical parameter space verified that the failure is not solely caused by the coarse grid, but instead suggests that the molecules and or individual emission lines, must be located in different regions within SVS13-A.

These results lead to the inability to constrain the effects of individual physical parameters on the sulphur abundance, instead encouraging further study into verifying the origin of various emission lines within SVS13-A.

CHAPTER IV

Virtual Reality Experience

The virtual reality (VR) project was created as an outreach deliverable for the ACO network. The initial planning of the project was done by all early stage researchers of the ACO network. The development, production and director work was done as part of this thesis, with contributions to the first video by Ross O'Donoghue[1] *and contributions to the story board and script by Lucy Evans* [2].

IV.1. INTRODUCTION

IV.1.1. Outreach

In their most basic form outreach projects, such as the work in this chapter, focus on bringing scientific knowledge to a target audience. However, in order to do so more effectively, outreach projects should consider not just the expansion of knowledge for the target audience but also how to grow the science capital of those they wish to reach. Science capital is a concept first suggested by Archer et al. (2015) intended to be used as a "lens" with which it should be possible to understand why some engage in science and some do not. The science capital of an individual is the collection of three components: i) amount of knowledge of science they posses; ii) their attitude towards science; iii) connections to the scientific community. The science capital of students can be used as an indicator for their likelihood on choosing to study a science, technology, engineering, or mathematics (STEM) degree (Sini Kontkanen & Havu-Nuutinen, 2025). Because of this, outreach projects aimed at students should focus on expanding science capital. One way in which this can be done is by making the knowledge sharing component

[1]ORCiD: 0000-0002-5317-6304
[2]ORCiD: 0009-0006-1929-3896

entertaining which, if done well, should not only increase the knowledge of science for an individual, but also positively impact the attitude individuals have towards science.

Educating through games has been shown to be a useful tool to enhance engagement and motivation (Boyle et al., 2016) while still being a field that requires more research (Videnovik et al., 2023). Games in general, not just video games, are often designed to maintain engagement and drive forward progress to specific goals usually determined by the designers. The application of games in education has been studied by many different researchers and spans large age ranges, from primary school up to learning through games as an adult (Jordan, 2009; Lee & Hammer, 2011; Walker, 1987; Zapalska et al., 2012; Videnovik et al., 2023). Most research finds that well designed games for learning can leverage the engagement driving nature of games, as is desirable by educators and can help students learn, and retain, the lessons (Boyle et al., 2016). It is also possible to take a step back from the creation of a full game and instead only take some elements of games in order to incorporate them into lessons. The process of taking only some elements and lessons from games is called gamification which still shows improved engagement from the target audience (Lee & Hammer, 2011). Gamification attempts to tap into the same drivers that cause gamers to invest vast quantities of hours into improving their own skills for a game (Jordan, 2009) by providing similar feedbacks without needing to develop a full game. For example, gamification could be the addition of badges, a point system, objectives, or even some rewards for achieving specific goals.

IV.1.2. Virtual Reality

While fairly common today, the concept of VR, is relatively new, with the first recorded use that reflects what we still mean today, dating back to "The Judas Mandala" in 1982 (Broderick, 1990). In that context and in today's definition VR refers to being able to experience a three dimensional computer generated environment that can be interacted with in a way that is similar to reality. This often includes having headsets which block out most if not all of the real world to a user's eyes, while then giving a user controllers or other forms with which

they can then interact with the virtual world. In instances where the real world is augmented with virtual features, rather than visual being overridden, we refer to the technology as augmented reality (AR) rather than VR.

The levels of immersion that can be achieved with VR, means that it lends itself extremely well for educational experiences, as it is engaging to multiple senses that would lead to better memory retention in students (DeMarinis et al., 2018), while also being exciting in a way that has a larger potential of maintaining excitement in students. This led to the desire to create an educational VR experience for the Astro-Chemical Origins (ACO) Innovative Training Network. The development of the ACO VR experience is focused on use with a Meta Quest 2 VR headset, as it was easy and affordable to attain several such headsets, while not requiring expensive computers in order to run the experience.

The idea of using VR for astronomy and outreach is not novel and the application and styles can vary largely. For astronomy, some VR outreach is done in the form of open events, which are similar to the in person talks held at the Royal Institution, but in VR, where users can attend talks and ask questions of scientists. Platforms where that can happen are "the Future of Meetings"[3] and the "Metavisionary Academy"[4] both of which are platforms that have been used and contributed to by Elisabeth Tasker[5] for outreach. It is also possible to use VR to create virtual exhibits such as the work by the University of Arizona to showcase the astronomy and space science work they do (Impey, 2024). Another example of VR for outreach is apps such as the Gleamscope VR[6] app which shows the night sky in various wavelengths as it would be seen from western Australia. The WorldWide Telescope from the American Astronomical Society (Rosenfield et al., 2018) allows users to view the night sky through a large array of wavelengths and surveys while also providing a platform for researchers to view observations. There has even been research on using VR for data visualisation more focused on aiding researchers in understanding complex data (Jarrett et al., 2021)

[3]https://thefutureofmeetings.wordpress.com/
[4]https://learn.metavisionaries.io/
[5]https://www.elizabethtasker.com/
[6]https://gleamoscope.icrar.org/gleamoscope/trunk/src/

IV.2. PURPOSE OF THE VR PROJECT

The main goal of the ACO VR project was to create an experience for school students around the age of 12-18. This age range was chosen collectively by the Early Stage Researchers (ESRs) and a senior member of the outreach work package of the ACO ITN. At the time of choosing this age range, the justification was that this age range would know enough about the topic to make the experience interesting without being overwhelming. We will discuss an evaluation of this choice in section IV.5.3. This meant that there were two objectives for the development. The first objective was to create an experience that is at an educational level so that most students would be able to learn something without being overwhelmed. The second objective was to make sure that the experience was exciting or entertaining enough for students to want to experience it and so they would remember it by gamifying parts of the VR experience.

In exploring the desired lessons to teach prospective students, it was decided to focus on water. Our focus was on teaching two things about water. The first point was that the water molecule is formed in stages, not in a single collision. The first step is to combine oxygen and hydrogen to form OH before adding another hydrogen to form water, H_2O. The second point that we wanted to emphasise was how large amounts of water could have come to Earth, namely through impacts from objects on the outer edges of the solar system such as comets. We focused on these points as we thought they would be easy to convey and interesting to students. Depending on the level of science knowledge they had, students may have already known these facts. For those students, the aim was to engage them through an entertaining way and encourage them to ask questions to the ESRs that would be available at the events where the experience would be shown. The experience is cut up into four parts. The first part sees the students explore water on a molecular level, letting them play with oxygen and hydrogen atoms. After that, a molecular cloud is shown and briefly explained. This cloud collapses into a star and then lets students dive into the disk that forms in order to create a planetoid from the dust in the protoplanetary disk. Between creating the planetoid and the last experience, students are shown a small animation of the disk being blown away by

the stellar winds, leaving just the star, and the planet that have formed thanks to the students help in forming it. In the last experience, students can press a button to let comets appear near them, so they can then attempt to throw the comet at the orbiting planet they created. This is to show a simplified way in which water can be brought to an earth like planet. This depiction was chosen as it was simple and described one of the major concepts of how water could have come to earth. We do acknowledge that it does not address the ongoing scientific debate on whether this is actually the dominant source of water on earth, something that has recently been called into question by initial data of the deuterium to hydrogen ratio in comet 67P/Churyumov-Gerasimenko, and observed by the Rosetta mission. It was first determined that the deuterium ratio is three times higher than terrestrial deuterium ratios (Altwegg et al., 2015), which was contested more recently by analysis of more data from Rosetta that found the deuterium ratio to be closer to terrestrial values(Mandt et al., 2024). Each experience is accompanied by minor explanations as well as small animations between experiences to give extra context to the science of what they are experiencing. The interactive experiences are the elements of the outreach project which have been gamified. By providing small, short term goals, some of which show scores, we aimed to drive up engagement with the project, increasing the attitude towards science, while also increasing the knowledge of students.

IV.3. TOOLS

In order to create the ACO VR experience we mainly used two tools, Blender (Community, 2018) which is an open source 3D modelling software, and Unity (Haas, 2014) a commercial game engine. In order to understand how the experience was created, we will give a brief explanation of each tool, and how it was used for the ACO VR experience.

IV.3.1. Blender

The Blender 3D modelling software has a large list of available features to allow for various mediums of artistic expressions. For the sake of the ACO VR experience, we focus on Blender's ability to create 3D models and animate them. Blender

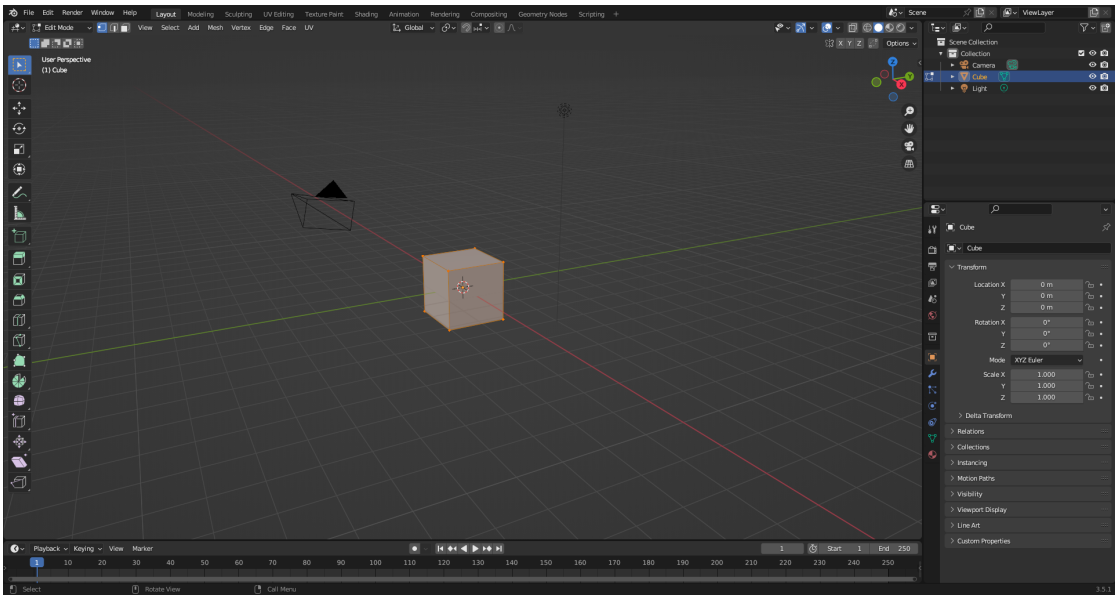

Figure IV.1 Screenshot of the blender interface. The central window is the scene, showing the vertices of a cube marked with orange points, with connected lines to each other forming the surfaces that are the faces of a cube. The right hand windows show at the top the hierarchy of all the components in the scene, and the bottom shows the options, and settings for the scene or the selected object, depending on the selected tab.

allows for the manipulation of points, called vertices, that can be connected to each other in order to create lines and two-dimensional surfaces. In turn, these surfaces connect to form a mesh in three dimensions. Beyond this, Blender also allows for these meshes to be manipulated within the time dimension to form an animated object. An example of the user interface, with a default cube in the rendering scene can be seen in figure IV.1.

To create an entertaining and immersive experience, we used blender to create all of the objects that are interacted with, as well as the representation of a robot that gives the explanations to a student. This robot seems like a minor addition as it is simply a monitor with hands. However, having an object that has been personified by having the voice emanate from it, and giving it hands which move, gives the experience an additional anchor for memories to be created. All of this gives the VR experience a greater potential to create a lasting memory for the user (Smith, 2019). The animations of the robot are created using key-frame animation in blender which allow designers to specify where objects need to be spatially, and how they need to be rotated at given times.

IV.3.2. Unity

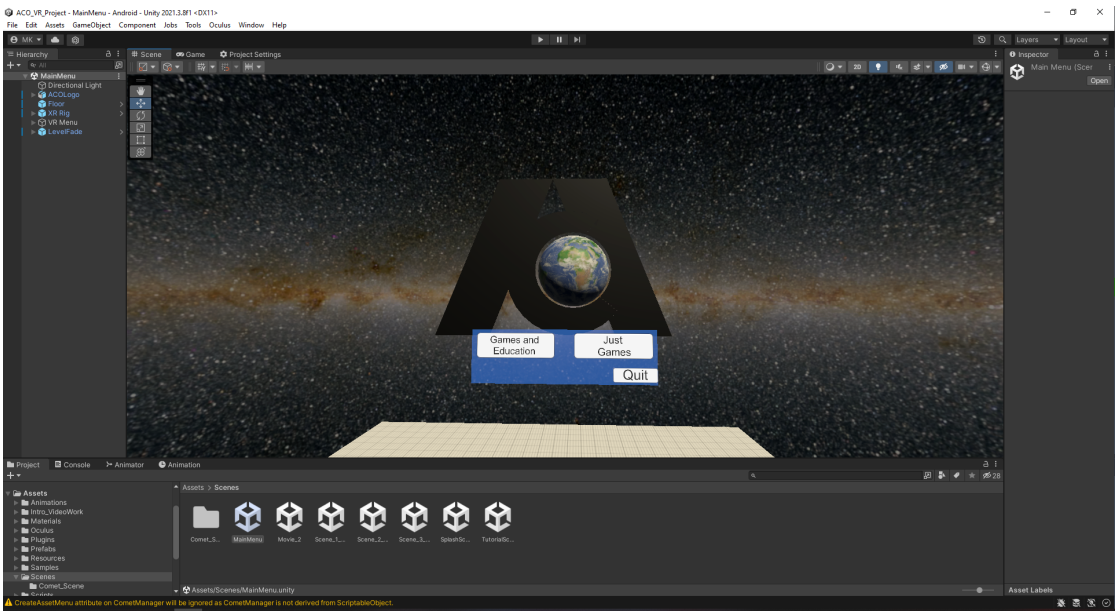

Figure IV.2 Screenshot of the Unity user interface, showing the main menu of the ACO VR Experience as seen in development mode. The largest central window shows the scene being edited. The left shows the hierarchy of all objects within the scene. The window at the bottom is the project window showing all available files created or imported for the project. The right hand window is the inspection window which shows the details of any object selected in the scene, or project window.

Unity is a game engine, that has been created to allow for cross-platform designing to be performed. This aspect applies to both developmental environments, as Unity is available to be used on Windows, Mac and Linux machines, as well as to the environments for which you can develop. Unity allows for creation of games for mobile devices, consoles, computers and VR headsets (Haas, 2014). An example of what the Unity interface can look like is shown in figure IV.2. Unity was chosen for this project because of its ease of use, and beginner friendly interface that allowed the project to be completed at a high standard with the aim of using the least amount of time to acquire that standard, allowing us to focus on the educational components.

Unity has a number of advantages over creating the ACO VR Project from scratch. For one, it has a built in physics engine, that allows for efficient collision detection and calculations to allow games to act more in line with how we would expect them to behave. As Unity is a large development platform there is a

packages that have been developed either specifically for it, or made compatible with it. One such package, is the OpenXR standards package [7]. This package is a standardisation of the VR and AR development fields that is designed to create a coherent platform from which experiences can be created. For the sake of the ACO VR project, it allowed for quick utilisation of the six-degrees of freedom that the Meta Quest 2 headset has, as well as making the integration of the controllers easy.

As just discussed, the work done for this experience relied on existing programs and libraries as well as the creation of many codes and geometric meshes. The functions not developed for this work, but provided internally by Unity, were those that allowed for interactions with controllers, scene transition, particle systems used for creating artistic representations of molecular clouds, as well as the functions that allowed for retrieving and altering the location, rotation and velocity of objects and the functions that allowed for collision detection. We relied on the OpenXR library combined with the MetaXR software development kit (SDK) for tracking the headsets, moving users, controller tracking and button interactions. The SDK included the geometry mesh for the headset controllers used in the experience. The background images found in each section of the experiences are publicly available images from NASA. Beyond these components, and the already mentioned contributions by Ross O'Donoghue and Lucy Evans, the remaining code, geometry designs, voice over recording, story board and script editing were done as part of this thesis. This includes, but is not limited to, the coding of the intra- and inter- molecular interactions, the coding to determine what happens when dust grains collide, the creation of the geometry meshes, particle system design, and animations.

IV.4. THE EXPERIENCE

The ACO VR experience is split into three distinct playable scenes, with some explanations or animations interspersed. It begins with an introduction video setting the scene for what to expect. This leads into the first experience which allows users to create water molecules from hydrogen and oxygen atoms. After this there is a brief video on molecular clouds forming protostellar cores and dusty

[7] https://www.khronos.org/openxr/

disks, which then leads to the second experience where users make a planetoid from the dust-grains in the dusty disk. The final scene then shows the dusty disk be blown away by the fully formed star, leaving just the planet that was created in the second experience orbiting the star. A user is then asked to throw comets at the planet until they can cause the comet to come close enough to break apart, leaving water on the planet. In the following parts, we will go through more details of each experience as well as the videos that precede the experiences.

IV.4.1. Creating Water

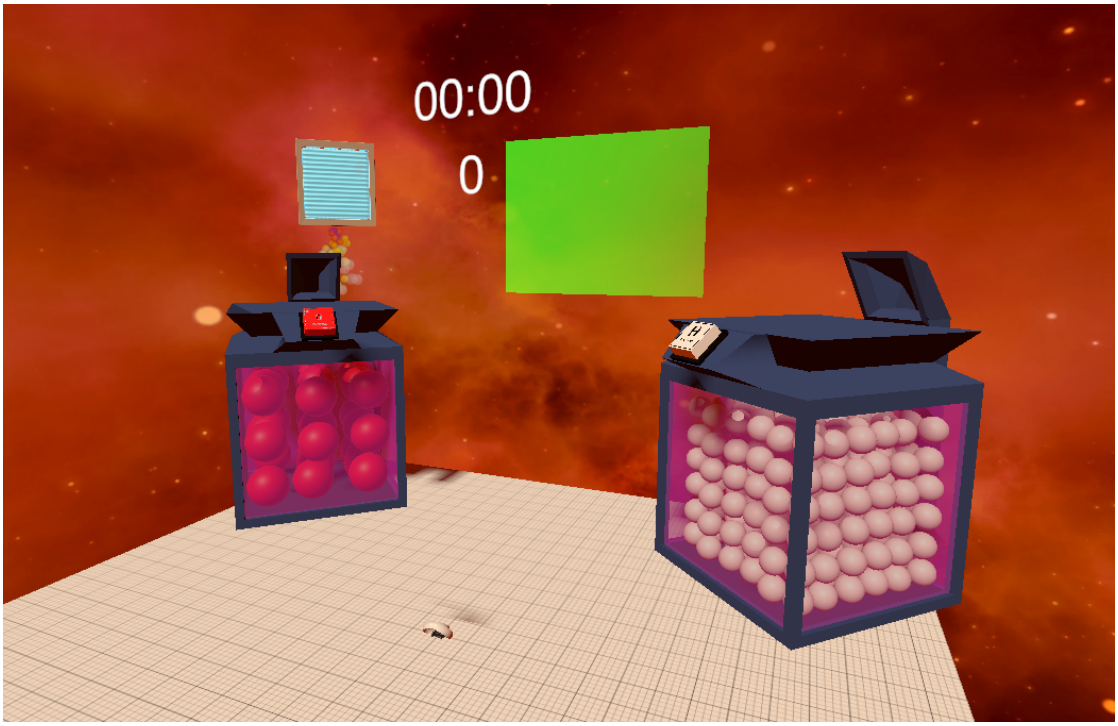

Figure IV.3 Picture of the first scene from the development view inside the Unity engine.

The introduction video does not provide any scientific explanations to a user, but instead simply welcomes them with the robot that will guide them through the experience, and tries to find a footing from where to continue the educational journey. As the focus of the experience is on water and how it reaches planets, we set the scene with the questions: "What is water, and where does it come from?". As this is being posed, an artistic representation of a star forming region is shown to the user in VR which is being zoomed in on. As the user gets ever closer, a grain of dust grows in size, implying that the user is shrinking down to an atomic

size. The user is then met with the scene of the first experience. How the scene looks from the Unity development scene can be seen in figure IV.3.

We start the first scene with a small tutorial on how to interact with the world. We then briefly explain what water is. The main objective of the first scene is to familiarise the user with the concept that water is composed of two hydrogen atoms and one oxygen atom, and to give them a physical representation of it using spheres similar to how it may be done in a class room. However, rather than the spheres being attached to each other with sticks, or magnets, in this environment the molecules use equations to govern how they should act with respect to each other. To maintain simplicity we still represent each atom as a hard sphere; however, each atom calculates the forces it should exert on surrounding atoms. Depending on the distance to each other, each atom will attract or repel other atoms until it forms molecular hydrogen, molecular oxygen or water. At which point, for simplicity of the experience and to prevent performance issues, the molecule will cease to attract other atoms and will instead exist in their bonded state and vibrate.

Each atom, whether in a molecule or not, vibrates at a constant energy level in random directions to make it feel like it is an energetic object, rather than just a hard sphere. The strength of the vibrations is inversely proportional to the size of the molecule, meaning the oxygen atoms almost seem like they do not vibrate, while the hydrogen atoms act as if they had a mind of their own. The attraction that each atom feels to each other, is modelled after a Lennard-Jones potential (Lennard-Jones, 1931). Meaning that at far distances, they will experience minor attractions, at a certain distance they will experience no force towards each other, before they start repelling each other at close range. We chose the Lennard-Jones potential despite it not being the most accurate representation of intermolecular interactions (Fischer & Wendland, 2023) because it describes them in a computationally expedient way, which was an important consideration considering the limitations posed by the Meta Quest 2 headset. As this is an equation not shown to the users we use it to make the experience feel more realistic. For simplicity, once an atom is bonded to another atom or molecule, we make it so the Lennard-Jones potential no longer affects the inter-molecular interactions, and instead use Hooks law to calculate the way the atoms should behave with

respect to each other within the molecule. We still let users stretch the atoms apart and let the atoms vibrate independently from each other but significantly reduce computational expenses by prohibiting this newly formed molecule or any of its components from being affected by the Lennard-Jones potential. An additional component that has been added to the water molecules is the angle between the hydrogen atoms. It is set to always be 104.49 degrees (Hoy & Bunker, 1979), and when one atom moves, the other will follow suit on the 2 dimensional interaction surface of the oxygen atom. Users can play with the individual atoms to experience this rather than having to rotate a physical object in a classroom.

The process of bonding has intentionally been made permanent and simplified, as the entertainment goal of the first scene is to create water molecules and to throw them through a green rectangular goal. In bonding the molecules permanently we make background calculations less computationally expensive by treating the molecule as one object, rather than the atoms as individuals, but additionally simplify the check to determine if an object passing through the goal is a water molecule or something else.

IV.4.2. Building Planets

Following the water experience, we show the user an artistic interpretation of a molecular cloud, see figure IV.4 for a screenshot of the scene. This was created by having black disks, with some transparency, randomly be generated and destroyed over time within a sphere. The effect of this is the creation of a sphere that seems to let some of the light behind it through, but significantly reduces the amount of stars that are visible towards the location of the sphere. In essence this mimics how some of the earliest observations of molecular clouds were seen.

At the beginning of this scene, we give the explanation of why we made users create water. We chose to give this explanation after the first experience in order to allow users to get into an interactive session faster. The user was emulating a dust grain, allowing the oxygen and hydrogen to have a shared area where they can easily and quickly interact with each other to form OH and then H_2O. We then describe where we can find dust, where these reactions can occur, which is when we point out the molecular cloud in front of the user. We skip over the details of

Figure IV.4 Screenshot of the video scene between the first and second entertainment sections, depicting a molecular cloud as just a black area blocking the light of the stars behind the cloud.

what dust grains are and what they are made out of in order to keep the experience short and since it was not one of our learning goals to leave users with knowledge of what dust grains are. . Explaining that they can collapse to form prestellar cores and then protostars we let the clouds collapse in a time-lapsed style to a star with a dusty disk. This is not meant to represent real time, but instead supposed to show how such a system can evolve over time.

Once the disk has been formed, we name the disk a protoplanetary disk where dust grains can collide together leading either to the destruction of those grains, or formation of larger objects like asteroids or even planets. We then let the user dive into the disk to help speed up the planet formation in the second entertainment scene of the experience.

In the second experience, dust-grains spawn around the user continuously. Once the timer starts, the user is allowed to grab individual grains and guide them towards other grains. Once they collide, they form one joint object that has a volume equal to the sum of the individual components. The user is allowed to continue growing grains in this way, as grains continue to be created around them. As a grain grows larger, it starts to affect the surrounding grains with a gravitational

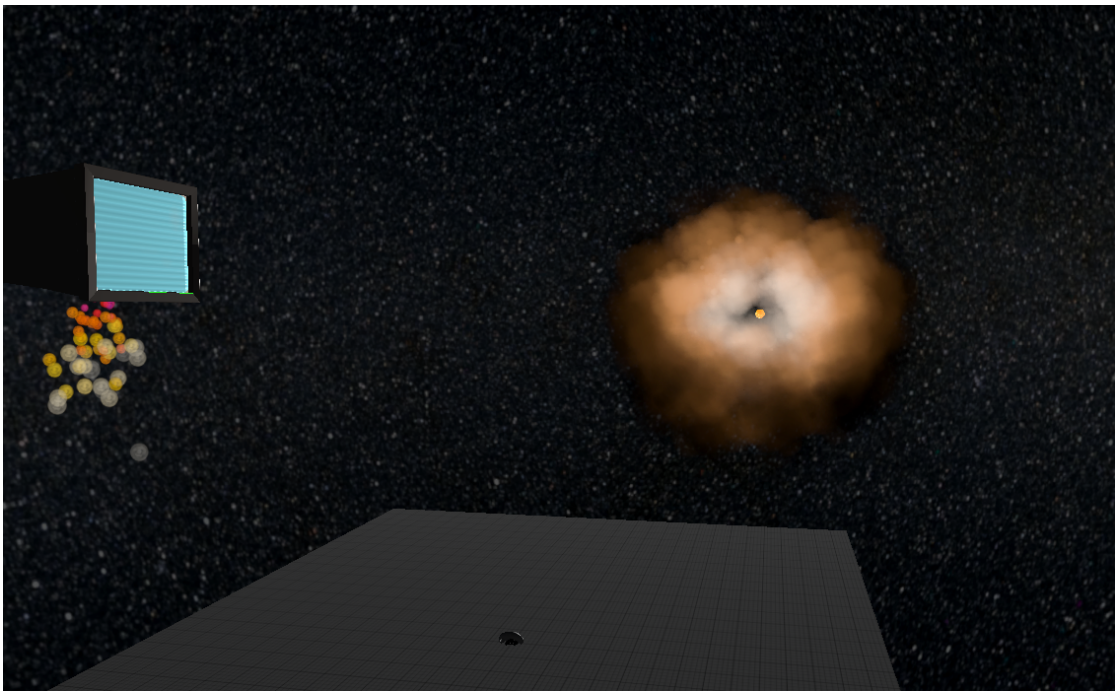

Figure IV.5 Screenshot of the video scene between the first and second entertainment sections, depicting the early stage star with its thick dusty disk surrounding it. This is shown directly before the user is allowed to dive into the disk to form a planet.

pull dependent on the mass of the object. This is calculated using the mass of the individual grains that contributed into making this one large grain. If the user lets go of their large grain, it will tend towards one spot in the scene in-front of the user, and pull other grains towards it, mimicking how a planetoid would carve out a cavity in a planetary disk during its formation.

At the end of the second scene, we explain that we will now zoom back out of the dusty disk, stating that the object created by the user will continue to grow as it creates an ever larger cavity in the disk. In order to make it feel like a user created the planet, we use the same model object for the planet in the final scene as we do for the planetoid that is created by a user in this second scene. This helps tie in the two, and makes a user feel like they have contributed to the experience which helps in memory retention for the experience.

IV.4.3. Bringing Water

The final Scene is preceded by one last explanation video, where we once again show the stellar system with a dusty disk. A brief explanation is given on how

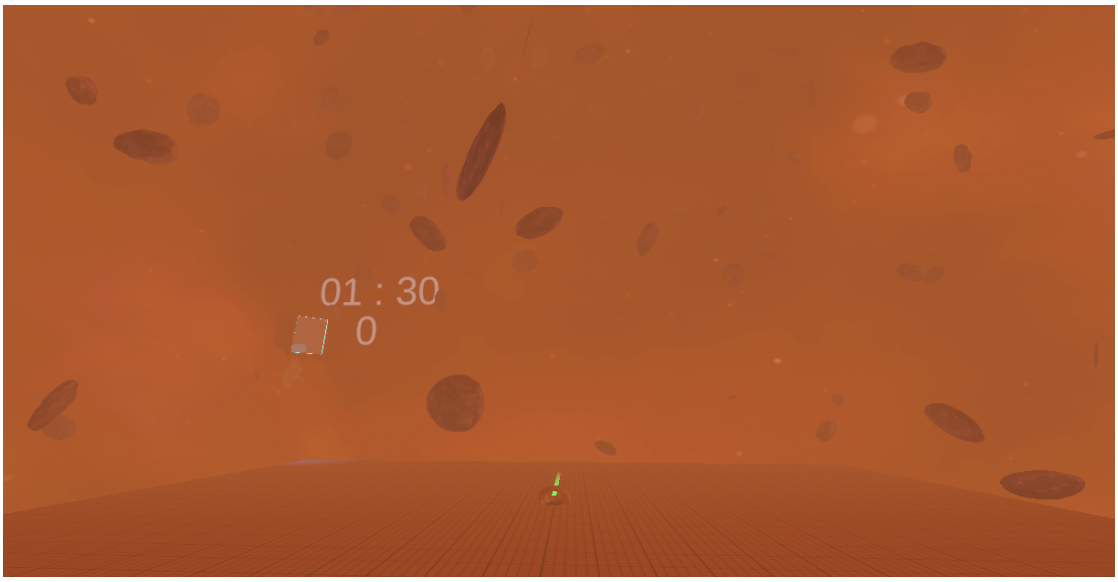

Figure IV.6 Screenshot of the second entertainment section where a user is allowed to form a planetoid or planet using dust-grains found within a very dusty disk. The visual impairment caused by the haze is intended to visualise a thick, dense dusty disk.

the parts of a disk which are not accreted onto either the star, or an object bound by its own gravity will be blown away as the star continues to produce more and more energy as it tends towards the main-sequence. Visually, the dusty disk is then blown away for the user, revealing just the star and the planet created by the user. This planet orbits the star with a near circular orbit and in the plane of the dusty disk.

While some water will be present on the planet, current research suggests that it would not be enough to account for the oceans we have on planet earth, and therefore would require external sources of water. The origin could come from comets as they are mostly ice, asteroids or other planetoids that formed further out than the earth but were set on a collision course with a younger earth for various reasons are also viable options as they would also carry water with deuterium ratios more closely resembling those found on earth (Pepin, 1991). As this is meant as entertainment and education, we let users spawn in comets with the push of a button so they can grab and throw them at the planet they created.

The physics of this scene is very simple. There are two objects, a star and a planet, which gravitationally affect each other. The planet has an extremely small influence compared to the star, but it is not negligible for a comet that passes

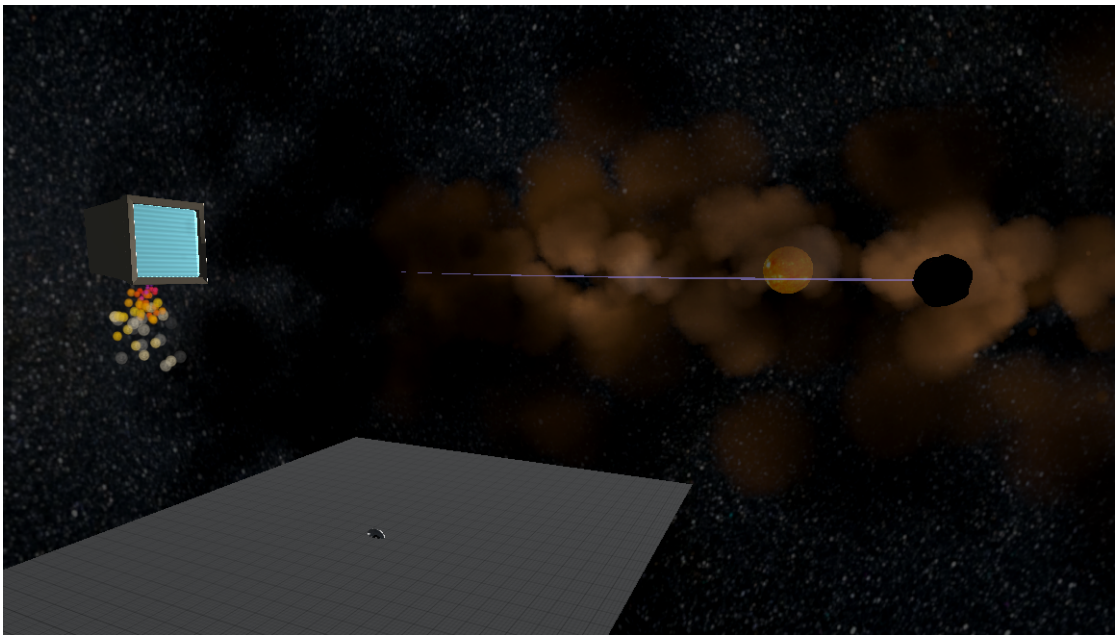

Figure IV.7 Screenshot of the third entertainment section where a user is shown how the dusty disk that did not accrete onto the star or a planet, is blown away. Leaving just the planet and the star.

closely by the planet. Additionally, the planet has been given an initial velocity to allow it to orbit around the star continuously in a near circular orbit. That is to say, that this scene does not just depict a two body system, but actually models it even if the size of the star and planet have been adjusted for the sake of a more pleasant and illustrative representation in a VR experience. When the comet is released by the user, it now also follows the laws of gravity in this system. If no initial energy is imparted into the comet it will simply be pulled towards the star. If enough energy is given by the user, it will create an eccentric elliptical orbit equivalent to that of a comet seen in our solar system. Additionally, the tail of the comet has been created taking into account that it should grow as it approaches the star, and always emits in such a way that it points directly away from the star it is orbiting. If a user were to throw the comet and simply watch, they would observe the comet continuously orbit the star, while passing the planet at different distances and closest approaches with each orbit.

Once the comet approaches the planetoid in a close enough approach, a small window appears for the user, allowing them to see the comet impact close up. We chose the distance of this approach based on user-experience design rather than physical reasons. In development we tested the experience with multiple users and

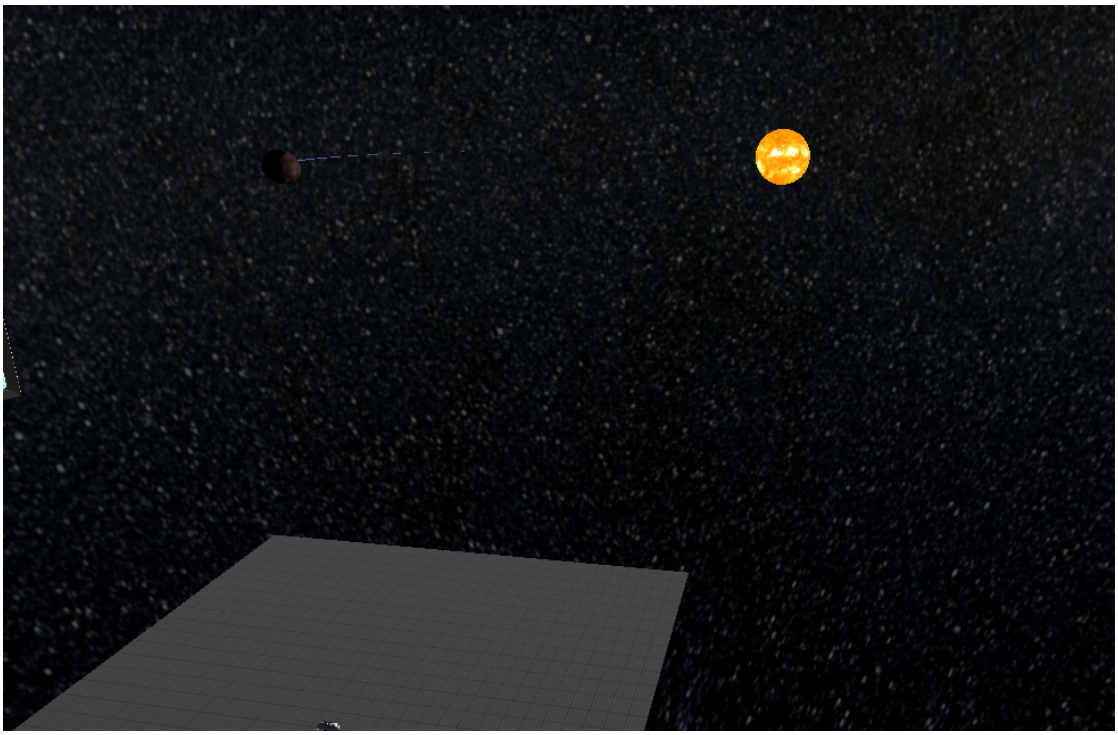

Figure IV.8 Screenshot of the third entertainment section where a user is allowed to create comets and throw them at the planet they created in order to bring enough water to the planet to form oceans like we would have here on earth.

tuned in the approach distance to allow the experience to be easier to complete rather than feel frustrating for users. The comet then breaks down, and oceans form on the planet that the user created. As this is meant to be a short experience, we opted for this fast approach, rather than explaining all of the details that would happen to the planet. We intentionally skip explanations of the planet being hot, needing to cool and so on to prevent the experience from becoming too long in such a way that we would run the risk of users not retaining any of the information we attempted to teach them. This then concludes the ACO VR experience.

IV.5. EVALUATION

IV.5.1. Public Interaction

The ACO VR project was taken to the European Researchers' Night 2022 in Turin and Perugia, as well as to the Science Festival of Genoa, in 2022. Italian research nights were chosen as a focus point for using the VR experience as the headsets procured for the ACO project were stored with the head of the outreach work-package situated in Italy, and transportation of them would have added

additional expenses. To the benefit of the ACO ITN, several of the early stage researchers (ESR) were native Italian speakers who had taken at least some of their graduate studies in Italian, making them well suited to interacting with the public that attended the research nights. The ACO ESRs that were present at the public events stated that the experience was well received by the public. Many participants enjoyed the games and listened to the translations of the explanation provided by the ESRs. In addition, the ESRs reported an increased curiosity for the science by members of the public, as well as interest in the ITN project after using the VR headset.

IV.5.2. Reflection

After showcasing the VR experiences at research nights and presenting the VR experience as a deliverable for the ACO ITN project, a post project reflection has been completed in order to internally review the experience. In order to do this, I have played through the experience with a focus on understanding what works well to achieve our learning goals, and what things could have be done better. I will first discuss a couple comments that affect the entire experience.

When going through the experience, the first thing I notice, is the feeling of educational components and gaming components being distinctly separate things. The experience takes users on a journey, letting them form water molecules, play with dust grains to form a planetoid and then throw comets at a planet to bring water. Done with more experience, the three experiences, any gamified components and any games within them, could be nearly seamlessly integrated with the goals of teaching the audience about water formation in space, and how large amounts of water can get to an earth like planet. Instead they feel choppy, or staggered, between education and gaming, and as if each lesson section is independent from the rest. I note that I can see the attempts to bridge the individual topics between the three components, but no such attempt is seen between the playable experiences and the educational parts. This all leads to the experience lacking a smooth flow, which has the potential of decreasing user engagement and motivation to finish the experience. Another alternative would be to intentionally segment each lesson section so they fully stand on their own. In doing this, users would be able to select

which lesson they want to learn and do them in any order. This would not fix the issue of the educational and playable components feeling separated, but presents a viable alternative which gives users a choice of what to learn, increasing users autonomy which improves engagement. This would however require reworking the animations and voice overs so that each lesson section is a self contained experience, not requiring or providing any knowledge from previous or subsequent sections.

The second thing of note is the recorded voice overs during the non playable components. In some of the voice overs the tone is considerably more serious with some detailed explanations while some parts are very casual, in one place even glancing over a potentially interesting concept. While not present in all voice overs, the extremes still present a potential issue. Even though water formation is one of the focuses for the learning goals, the actual steps of monoxide formation and then water formation are explained in potentially too much detail after the game, which could lead to losing engagement from users as parts of the post explanation overlap with the explanation given prior to forming water molecules. On the other extreme, the mechanisms that could start the collapse of molecular clouds are not even mentioned, potentially leaving users with questions that do not get answered by the experience. While explaining what can cause cloud collapse was not a learning goal, the presence of the concept in the experience, but lack of explanation, leaves room for improvement. The last explanation of how some water can get to planets through comets, which was a learning goal, is talked about in a more appropriate tone if the intended age range was more focused on younger students, not the whole age range of 12-18, a topic that I will discuss in IV.5.3. In addition, the switching between a more serious and casual tone leads to some parts sounding condescending such as the explanation of how to use the controls. While it is meant to be educational and entertaining, the tone through out the voice overs was intended to reflect that the experience is meant to be light educational content with some game components. This is better reflected in the initial voice overs and those in the comet section. The floating monitor with hands, which is intended to reflect the guide that is speaking to the users via the voice overs, is simple but effectively represents the guide. Despite knowing the technical components of the experience, and that the monitor does not actually emit any sound, I find ourselves looking at

the avatar as though I were listening to an actual person when I played through the experience for our post project reflection. This was the desired response of users. Having reflected on the overarching components of the VR experience I will now look at the individual playable scenes and the animations that precede them.

The first animated section contains a nice representation of a colourful object in space. While it is never explained that it is supposed to be an artistic representation of a nebula, it does capture and maintains the users attention over time as it moves and changes, while not being so distracting that it disrupts the voice over. Experiencing it again, I find that the centrepiece should have been discussed, or at least acknowledged, and that users should have been told that they would be scaled down to atomic scales in order to play with oxygen and hydrogen atoms as users are transitioned to the first playable experience. The scaling down is hinted at by showing the user a dust grain covered in ice right before the transition, but it is not actually discussed until the start of the first playable experience. The first playable experience starts with how to interact with the controllers, which could have been shortened or sped up significantly. It is in the explanation of how to interact that the voice overs first sounded potentially condescending. This explanation could have been shorter, preventing potentially loss of engagement from users. The brief science explanation of forming water is also considerably longer than it needs to be, and part of this could have been in the first animation as users flew towards the dust grain. Doing so would have helped with the flow of the first playable experience. As the first experience stands now, it feels like an explanation, followed by a simple game, rather than a gamified learning experience.

If the first playable experience would be recreated now, I would change it to give users a more guided experience of creating water, rather than explaining the concept and then playing. By creating event triggers and changing the voice overs, it would be possible to explain what monoxide is and then ask users to create it. This alone could change the feeling of the experience from "lesson" and then "play", to learning through experimenting and playing, which was the intended format. Such triggers could also allow us to add voice overs in the case that users do something unintentional, or different, such as forming H_2, adding the possibility of additional learning that can be framed as a reward for performing a

secret task. Such additions would raise the potential for maintaining engagement and motivation to continue the experience beyond the VR technology by itself. Another change to the experience would be to give players a dust grain to which they need to throw oxygen and hydrogen, which could then be shown to diffuse across the surface of the grain and react. By doing this it would naturally lead to a visual example of the formation of water on dust grains, rather than rely only on a vocal explanation. If this was done, then part of the voice over from the second animated section could be pulled into the first playable experience which would further improve the flow of the overall experience. Despite the issues raised about the first experience and relying on parts of the second animation, I do find that the VR experience does teach one of the desired lessons of how water forms in space through stages.

As just stated, the second animated section starts with more details about the formation of water on dust grains through the voice overs, but dose not provide any visual of this to the user. It then continues to address the black cloud in-front of the users, intended to represent a molecular cloud. While no explanation is given as to what mechanisms initiate collapse, I find that the rest of the explanation of how molecular clouds collapse once the process starts is accurately reflected and the voice over is in the tone that was intended for the experience. The visuals that accompany the explanation are simple but effective. The ignition of fusion and growth of the central object neatly align with the explanation leading to a satisfying rendering and explanation. At the end of the second animation, the voice overs transition the discussion away from water and towards forming asteroids and planetoids in the protoplanetary disk. This information was not initially a learning goal and could instead be replaced with an explanation of the frost line. The explanation is not inaccurate or poorly done, but discussing the frost line would more closely align with the learning objectives by explaining where water is on solid grains in a solar system. This approach would tie the first and the last playable experience together better than is currently done. Users would have learned how water can form on dust grains, then learn why water would be unlikely to stay frozen onto a grain when that grain is too close to a star, which could lead into the explanation of why objects beyond the frost line, such as comets, should have

more water frozen onto them than objects closer to the star than the frost line.

Beyond potentially changing the topic of the second lesson, I also find that the second playable experience, like the first, feels more like a game related to, but separate from, the given lesson. If this playable experience were to be recreated, without changing the topic of the second section, I would more smoothly transition from the explanation section to the playable section. This could again be done with event triggers. The addition of event triggers would be beneficial in the same way as they would be in the first playable experience. If the experience allowed voice overs to start when specific activities are performed, such as causing two objects to collide, then it would reduce the feeling of jumping from education to game and back. This could still be followed by the same game currently in the experience, but would make the actual explanation part more engaging. On top of that, I think it would be beneficial to add the chance that two colliding dust grains do not merge and instead break one or both dust grains into smaller pieces. By adding the chance of collisions breaking apart objects and making it velocity and relative size dependent, I would then be able to interweave more explanations in a way that feels explorative for the users. This would enhance engagement as users would potentially feel like their interactions play a key role in the virtual environment, rather than just aiming for a high score in the game. As mentioned when discussing the animation of the second section, the discussion of planet formation was not an intended learning objective and because of that, this entire section does not further any of the learning objectives beyond the part of the voice over that should have been in the first section. However, this section was intended as a bridge between how water forms and how large amounts of water can come to a planet. With respect to that goal, this section does work well in providing a transition between the two sections.

The last lesson and playable experience section does stand on its own quite well, leading simply with the star blowing away the disk, leaving behind one planet large enough to maintain an orbit. The visuals of this are again simple, but effective in making it feel like a small solar system is in-front of the user. I find that the explanation component of the last experience more accurately reflects the desired tone the experience was intended to have, but unnecessarily explains part of what

will happen in the playable part. Again like the other two sections, I find that interweaving the explanation with events that the player triggers, would allow for the experience to be more engaging.

If this last section would be recreated, most of the voice over could be kept the same with some parts saved for specific event triggers. The part of the voice over that explains the game components could be shortened, and simplified or even removed in favour of visual queues. There is a highlighting function showing which button spawns a comet in place but this could be improved upon by adding a floating tag with text more clearly pointing to the button and labelling what it does or by providing a floating image of the controllers with markings on the buttons that show how to play rather than using the voice over to do so. As for the visuals, I find that the effect of a planetary system could be more convincing if there was an asteroid belt or more planets. By adding other planet closer to and further away from the star I could add additional micro lessons about the habitable zone of stars when the comets interact with the planets too close or too far away to have liquid water. By providing the micro lesson when players perform specific actions, it feels like the lesson is provided in a more organic form based on the players actions. This could also allow for better interplay with the proposed changes to the second lesson of discussing the frost line. As for how this section helps in furthering the learning goals, I find that it does a good job of showing how water can reach planets through comets, but does not talk about the fact that some water and some hydrogen are already present in planetoids as they were formed, which is a concept that should have been mentioned in order to provide a more rounded educational experience.

IV.5.3. Lessons Learned

The ACO VR experience created as part of this thesis has met the goals of what was requested as part of the ITN deliverables. Beyond the creation of the project, there has also been a lot of lessons that should be shared with other groups that are considering taking on similar challenges. What I consider to be potentially the most important lessons are the required knowledge of how to plan and do outreach as well as the time, and technical knowledge required in order to create a VR

experience.

The ACO VR experience relied on using the Unity software and the VR interaction libraries of OpenXR and MetaXR in order to be developed. All three of these tools aim to simplify the technical components creating a lower bar for entry into making a VR project. However, they still require a degree of familiarity and some knowledge of the coding languages C++ or C# in order to be used effectively. At the time of starting this project, most of the ACO ESRs had not had any experience with either coding language and had not used these tools. This meant that the project relied on a subset of the ESRs that had some familiarity with those languages and tools in order to produce the VR experience within the scope of the ITN project.

I strongly suggest that any group(s) that wish to produce a similar outreach VR project consider the following two options. The first option is to do an investigation into scientific outreach, as well as the technical and artistic load of creating a VR project. By performing an investigation into scientific outreach, future projects would be better equipped to effectively plan their goals so as to increase the potential of expanding the scientific capital of their target audience. Investigating the technical and artistic load allows a group to scale the project to the amount of contributors they have and how much time they can afford to spend on the VR project away from regular work. This leads to setting more realistic goals for the technical production. The second option I recommend is considering a partnership with a game studio, VR development company or experienced developers that have made VR experiences in a professional capacity. Such external members could be asked to help create assets and code for the experience and/or to develop the skills of the members of the group trying to create the VR experience. This would allow contributors to learn from the experienced developers who would have a greater understanding of how to address the challenges a VR project will have, while also providing experienced developers that could be consulted for difficult to solve components of a project. If possible, I recommend both options be used.

For evaluation of the interaction with the public I find that collecting data from participants would have been key for understanding the impact of the experience. Understanding the distribution of ages of users, and how they viewed science as

a subject before using the VR headset, as well as asking for information about the experience in post. All of this type of information would have been helpful in understanding the impact that this work has had on users. In addition, it would give us external feedback on improvements that could have been made for this work as well as future VR outreach projects that will be created. Without such data, I have been forced to rely on the anecdotal data given by the ESRs that were present at the research nights in order to gain any form of feedback.

Another major point I would like to stress is the time spent in setting the learning objectives and selecting an age range for the target audience. Both of these things would be significantly easier to do with an investigation into the science of scientific outreach. For the ACO project, at the time that these decisions were made, none of the ESRs had a significant familiarity with outreach projects beyond giving talks at events like Astronomy on Tap[8] or through social media posts.

A prime example of a naive decision is the chosen age range of the target audience. We chose ages 12 to 18 which covers a wide field in terms of knowledge that would be expected of audience. This is reflected in anecdotal data from the ESRs that attended the research nights. They stated that older students seemed to either be more interested in the VR than the science or they asked questions that were more complex than the ESRs expected and the experience was intended to cover. The ESRs also stated that, on average, the younger students seemed to engage well with the explanations given to them while also enjoying the gamified experience. This leads us to conclude that the age range should have been considered more carefully. For the goals that we had of looking at water formation in space and how it gets to earth, the anecdotal evidence suggests that focusing more on younger students would have been beneficial. Had that been our goal from the start, rather than trying to appeal to the entire age group, we could have taken a closer look at the knowledge that would be expected of the smaller age range in order to more effectively tailor our lessons and goals to them. The flaws that exist in the naive approach were mitigated by the fact that some of the ESRs were present at the time that users were going through the experience. This meant that even when older students asked questions beyond the scope of the experience, they

[8]https://astronomyontap.org/

were able to get responses and even have positive discussions with the ESRs.

IV.6. CONCLUSION

The ACO VR experience takes users on a journey from what water is on the atomic level, to where it can form and how planets are formed. The final leg of the journey introduces one of the concepts of how we think water may have come to a planet like the earth. Throughout the journey, users get to experiment, and play with the experiences in such a way that we hope to create a memorable experience. The aim is, that if the experience is remembered in an episodic style of memory, then hopefully the information we attempted to teach them is also retained. Once the project was created, the ACO VR experience was shown off at various public outreach events throughout the European Union, with the aim of letting as many students experience it as time allowed. Additionally, with the short time of the experience, we were able to not only let more students try it, but also had time to answer questions posed to us by the most curious of students. Overall, the ACO VR experience has been rated a success as a deliverable for the ACO Innovative Training Network, despite the flaws that are present in the experience as a stand alone product.

CHAPTER V

Future Work

In order to continue building on the work that has been achieved in this thesis, there are a few projects that will be worked on beyond the time doing this PhD. The two main future goals that pertain to the work completed in this thesis, are to improve the online available version of UCLCHEMCMC and to create a prototype astronomical archival data dashboard.

V.1. UCLCHEMCMC IMPROVEMENTS

UCLCHEMCMC in its current form allows for the source code to be downloaded and run on systems if compiled by a user. A remote version with a dedicated database is available for use, but still requires a few key features before it can be advertised to a wider user base. Those improvements include a queuing system, a result returning system and a system to load old runs from remote files. Beyond this, additional improvements to UCLCHEMCMC should be explored, guided by the experience from working on the inference in chapter III. These improvements include deviating from a grid based parameter space, further abstracting the MCMC runs from the chemical and radiative transfer modelling, as well as optimisation improvements to the model execution and database storage.

The queuing system that will be implemented would need to take in the requests from remote users to the local facilities hosting UCLCHEMCMC and order them based on when they were submitted. At this point it would submit the requests directly to UCLCHEMCMC when instances of UCLCHEMCMC become available from finishing other requests. In doing so, not only can requests be asynchronously submitted with how they will be calculated, but also we can control how many resources assigned to UCLCHEMCMC. Having this control, allows

for UCLCHEMCMC to be run computationally efficiently and adapted as more resources become available for it, or reduced if UCLCHEMCMC use is low.

A further issue with the current deployment of UCLCHEMCMC online is that it requires a user to remain on the website in order to view the results and to then request manually the files they generated for the inference. Therefore, an additional upgrade that needs to be developed is a system that will generate an email containing the results of the inference. This would entail a feature to request emails, associate them to the correct inferences and generating the actual email to forward the results to the correct users. A system like this would make it easier for users to submit an inference, leave the site and simply wait for the email to arrive rather than having to make sure the website stays open at all times.

In order to facilitate hosting an MCMC inference tool, UCLCHEMCMC will also need one further upgrade in order to function well. Either a system to load previous inferences so that it can continue where the last step left off, or it would need a system to allow the MCMC to continue taking steps until pre-determined criteria are met. The latter option, would still benefit from being able to load previous inferences in order to adjust criteria without needing to restart an inference, the option to develop a loading system is preferred. UCLCHEMCMC already has code in place to allow for previous inferences to be continued, however, as storage space is not infinite on the remote host, it would be beneficial to let users to upload the file containing the steps of the previous inference which would be sent to users when they finished.

After the failed inference in chapter III, which prompted additional exploration of the gridded parameter space, it would be beneficial to improve UCLCHEMCMC to the point where it no longer needs a grid based parameter space. This could, for example, be done by allowing each parameter to have an additional value that determines how far away in that parameter space a model has to be from a model in the database in order for it to trigger the creation of a new model. This would mean that if a new run is started with initial temperature as a free parameter, a user could stipulate a value of 0.5 K for this third parameter. In the event of a new model having all parameters, aside from initial temperature, equal to the parameters of a model in the database new calculations would only be done if the

initial temperatures of the new model is greater than (less than) the stored models initial temperature plus (minus) 0.5 K. If the new models initial temperature is within that set range, then the stored model will be used as a proxy for the new model. In doing so, the grid system would no longer need to be updated whenever a new project is created, and instead would be adaptable by design, allowing users to refine specific parameters as they desire.

Another improvement inspired by the work from chapter III is to further abstract the chemical and radiative transfer modelling away from the MCMC inference components of the code. By further abstracting the model codes out, it becomes possible to more easily update the codes used for UCLCHEMCMC, without requiring significant changes to UCLCHEMCMC, as well as making it easier to add new modelling codes. In addition, it facilitates adding optimisations to UCLCHEMCMC without affecting the integration of modelling codes. For example, the methods of storing models could be expanded. Currently, the SQLite database system, while useful, does require rerunning chemical models when changing the radiative transfer model. If changes were made to further abstract the models out, making UCLCHEMCMC more independent from modelling codes, it would then also be easier to use other storage technologies more suitable for UCLCHEMCMC in such a way that chemical models and radiative models can be stored and checked independently. This could significantly improve inference times when comparing radiative transfer models as the chemical models would not have to be recreated.

V.2. ASTRONOMICAL ARCHIVAL DATA DASHBOARDS

While the healthcare application, discussed in Appendix A, has huge potential, it would also be useful to see where such dashboards could be useful in astrophysics. For this, we look to the potential of a dashboard for archival telescope and survey data. Many archival data sites already use simple forms of dashboards with limiters, in order to help in perusing data taken by specific telescopes. However, they tend to be limited to just that telescope. There are alternative options that allow for searching across different telescopes, such as the tools provided by the Strasbourg astronomical Data centre (Genova, 2013), but they lack dashboards and are limited

on how useful they are by mainly focusing on finding specific objects or finding objects by coordinates. While both of these styles are very helpful, they are very restrictive on what they can be used for. On the other hand, a collected dashboard across surveys and telescopes would facilitate both styles together, allowing for searches by specifications, name of object and/or location.

The two easiest advantages that can be found for this type of dashboards are for archival data studies as well assisting potential future surveys to identify areas where surveys would be potentially most effective. Let us explore how exactly these two advantages could work in a potential cross telescope dashboard starting with the archival data studies.

V.2.1. Archival Data Studies

Using a dashboard that collects data from across telescopes and archives would allow for searches across the data in various different ways. Work done by the international virtual observatory alliance (IVOA)[1], and the Mikulski archive for space telescopes (MAST)[2] already facilitate such searches. Both of these platforms have several tools available that improve their usability, such as allowing users to view spectra and publications related to specific objects. A natural next step for either or both of these services would be the addition of a dashboard similar to those found in Appendix A but tuned to fit astronomical researchers needs. Such a dashboard could enhance the search functionalities of the platforms further by continuously showing statistical distributions which would update as limitations to the search query are imposed. Limitations on the data could be placed in observed frequencies, bands or other constraints needed for the study being conducted exactly as they would be with a normal search, but through a more visual interface and with more statistical data being provided prior to looking at the individual observations that fit the limitations.

[1] https://ivoa.net/
[2] https://archive.stsci.edu/

V.2.2. Exploration of New Surveys

Another potential use of such a dashboard would be in aid of crafting future surveys. In collecting the data of different telescopes and surveys the data dashboard can show different statistical distributions. Using multiple types of information from observations, such as the objects observed, frequencies used, and parts of the sky that were observed all in one location we can create a dashboard to help find gaps in observational information. Much of the information is attainable through archival searches which already collect the data of various telescopes together, dashboards would simply provide quicker access to statistical distributions, with less work by individual researchers.

The simplest case that is also easy to study without such a dashboard, is looking at which frequency ranges have the least amount of observations. We can take this a step further though by limiting the frequency range we wish to show in the distributions, and then look at which classes of object were observed in those ranges. If certain classes of objects are under-represented without an understood explanation, then this could lead to research on finding why those classes are not observed in a given range. These are simple examples which can be done by hand with the current software provided by the IVOA or MAST but would be quicker to see if either of these tools had a data dashboard. The potential of such a dashboard will only grow as we continue to build more and more telescopes, because as the sources of data increase, so too does the challenge of keeping track of what types of observations are available.

Bibliography

Akeret, J., Seehars, S., Amara, A., Refregier, A., & Csillaghy, A. 2013, Astronomy and Computing, 2, 27–39, doi: 10.1016/j.ascom.2013.06.003

Allodi, M., Baragiola, R., Baratta, G., et al. 2013, Space Science Reviews, 180, 101, doi: 10.1007/S11214-013-0020-8

Altwegg, K., Balsiger, H., Bar-Nun, A., et al. 2015, Science, 347, 1261952, doi: 10.1126/science.1261952

Anderson, J., & Baggett, S. 2014, Sink Pixels and CTE in the WFC3/UVIS Detector, Instrument Science Report WFC3 2014-19, 12 pages

André, P., Men'shchikov, A., Bontemps, S., et al. 2010, , 518, L102, doi: 10.1051/0004-6361/201014666

Anglada, G., Rodríguez, L. F., & Torrelles, J. M. 2000, , 542, L123, doi: 10.1086/312933

Appenzeller, I. 2012, Historical Remarks, Cambridge Observing Handbooks for Research Astronomers (Cambridge University Press), 1–25

Archer, L., Dawson, E., DeWitt, J., Seakins, A., & Wong, B. 2015, Journal of Research in Science Teaching, 52, 922, doi: https://doi.org/10.1002/tea.21227

Arny, T. 1990, Vistas in Astronomy, 33, 211, doi: https://doi.org/10.1016/0083-6656(90)90021-Y

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

Bachiller, R., Guilloteau, S., Gueth, F., et al. 1998, Astronomy and Astrophysics, 339, L49

Barsony, M. 1994, in Astronomical Society of the Pacific Conference Series, Vol. 65, Clouds, Cores, and Low Mass Stars, ed. D. P. Clemens & R. Barvainis, 197

Bash, F. N., & Peters, W. L. 1976, , 205, 786, doi: 10.1086/154334

Bayes, T., & Price, n. 1763, Philosophical Transactions of the Royal Society of London, 53, 370, doi: 10.1098/rstl.1763.0053

Bell, S. K., Delbanco, T., Elmore, J. G., et al. 2020, JAMA Network Open, 3, e205867, doi: 10.1001/jamanetworkopen.2020.5867

Bonnar, W. B. 1956, Monthly Notices of the Royal Astronomical Society, 116, 351, doi: 10.1093/mnras/116.3.351

Bonnor, W. B. 1956, , 116, 351, doi: 10.1093/mnras/116.3.351

Bottinelli, S., Ceccarelli, C., Lefloch, B., et al. 2004, The Astrophysical Journal, 615, 354, doi: 10.1086/423952

Boyle, E. A., Hainey, T., Connolly, T. M., et al. 2016, Computers Education, 94, 178, doi: https://doi.org/10.1016/j.compedu.2015.11.003

Broderick, D. 1990, The Judas Mandala (Milsons Point, NSW, Australia: Mandarin)

Brown, P. N., Byrne, G. D., & Hindmarsh, A. C. 1989, SIAM Journal on Scientific and Statistical Computing, 10, 1038, doi: 10.1137/0910062

Buneman, P. 1997, in PODS '97 Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (ACM), 117–121, doi: 10.1145/263661.263675

Bunker, P., Jensen, P., Canada, N. R. C., & Program, N. R. C. C. M. P. 2006, Molecular Symmetry and Spectroscopy, NRC monograph publishing program (NRC Research Press). https://books.google.co.uk/books?id=FZEI7VmjNyMC

Byrne, G. D., & Hindmarsh, A. C. 1975, ACM Transactions on Mathematical Software, 1, 71

Carroll, B. W., & Ostlie, D. A. 2017, An Introduction to Modern Astrophysics, 2nd edn. (Cambridge University Press)

Caselli, P., & Ceccarelli, C. 2012, Astron. Astrophys. Rev., 20

Caselli, P., Hartquist, T. W., & Havnes, O. 1997, , 322, 296

Caselli, P., Walmsley, C. M., Zucconi, A., et al. 2002, The Astrophysical Journal, 565, 331, doi: 10.1086/324301

Ceccarelli, C., Caselli, P., Fontani, F., et al. 2017, Astrophys. J., 850, 176

Ceccarelli, C., Maret, S., Tielens, A. G. G. M., Castets, A., & Caux, E. 2003, A&A, 410, 587, doi: 10.1051/0004-6361:20031243

Chang, Q., Cuppen, H. M., & Herbst, E. 2007, A&A, 469, 973, doi: 10.1051/0004-6361:20077423

Charnley, S. B. 1997, Astrophys. J., 481, 396

Chen, X., Launhardt, R., & Henning, T. 2009, The Astrophysical Journal, 691, 1729

Chini, R., Reipurth, B., Sievers, A., et al. 1997, , 325, 542

Chuang, K.-J., Fedoseev, G., Qasim, D., et al. 2017, Monthly Notices of the Royal Astronomical Society, 467, 2552, doi: 10.1093/mnras/stx222

Codella, C., Ceccarelli, C., Cabrit, S., et al. 2016, Astron. Astrophys., 586, L3

Codella, C., Bianchi, E., Podio, L., et al. 2021, Astron. Astrophys., 654, A52

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

Community, B. O. 2018, Blender - a 3D modelling and rendering package, Blender Foundation, Stichting Blender Foundation, Amsterdam. http://www.blender.org

Condon, J. J., & Ransom, S. M. 2016, Essential Radio Astronomy, sch - school edition edn. (Princeton University Press). http://www.jstor.org/stable/j.ctv5vdcww

Data-Resources. 2023, Library guides: Data resources in the Health Sciences: Clinical Data. https://guides.lib.uw.edu/hsl/data/findclin#s-lg-box-1908462

de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., et al. 2015, , 582, A62, doi: 10.1051/0004-6361/201526601

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

De Simone, M., Codella, C., Testi, L., et al. 2017, Astron. Astrophys., 599, A121

De Simone, M., Codella, C., Ceccarelli, C., et al. 2022, Mon. Not. R. Astron. Soc.

Deckler, G. 2022, Learn Power BI, 2nd edn. (Birmingham, England: Packt Publishing)

DeMarinis, T., Calligaro, L., Harr, C., & Mariani, J. 2018, Real learning in a virtual world, Deloitte. https://www2.deloitte.com/us/en/insights/industry/technology/how-vr-training-learning-can-improve-outcomes.html

Denis-Alpizar, O., Stoecklin, T., Guilloteau, S., & Dutrey, A. 2018, Monthly Notices of the Royal Astronomical Society, 478, 1811, doi: 10.1093/mnras/sty1177

Dhabal, A., Mundy, L. G., Chen, C.-Y., Teuben, P., & Storm, S. 2019, Astrophys. J., 876, 108

Dickman, R. L. 1975, , 202, 50, doi: 10.1086/153951

—. 1978, , 37, 407, doi: 10.1086/190535

elasticsearch. 2015, elastic/elasticsearch. https://github.com/elastic/elasticsearch

Elmegreen, B. G., & Scalo, J. 2004, Annual Review of Astronomy and Astrophysics, 42, 211, doi: 10.1146/annurev.astro.41.011802.094859

Endres, C. P., Schlemmer, S., Schilke, P., Stutzki, J., & Müller, H. S. 2016, Journal of Molecular Spectroscopy, 327, 95, doi: https://doi.org/10.1016/j.jms.2016.03.005

Enoch, M. L., Lee, J.-E., Harvey, P., Dunham, M. M., & Schnee, S. 2010, The Astrophysical Journal, 722, L33, doi: 10.1088/2041-8205/722/1/l33

Evans, Neal J., I. 1999, , 37, 311, doi: 10.1146/annurev.astro.37.1.311

Favre, C., Ceccarelli, C., López-Sepulcre, A., et al. 2018, The Astrophysical Journal, 859, 136, doi: 10.3847/1538-4357/aabfd4

Fischer, J., & Wendland, M. 2023, Fluid Phase Equilibria, 573, 113876, doi: https://doi.org/10.1016/j.fluid.2023.113876

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306–312, doi: 10.1086/670067

Fulvio, D., Góbi, S., Jäger, C., Kereszturi, Á., & Henning, T. 2017, Astrophys. J. Suppl. Ser., 233, 14

Gagniuc, P. A. 2017, Markov chains (Nashville, TN: John Wiley & Sons)

Garrod, R. T., & Pauly, T. 2011, Astrophys. J., 735, 15

Gelman, A., & Rubin, D. B. 1992, Statistical Science, 7, 457. http://www.jstor.org/stable/2246093

Genova, F. 2013, Data Science Journal, 12, WDS56, doi: 10.2481/dsj.WDS-007

Goodman, J., & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65, doi: 10.2140/camcos.2010.5.65

Green, S., & Chapman, S. 1978, , 37, 169, doi: 10.1086/190523

Groppi, C., Walker, C., Kulesa, C., et al. 2009

Haas, J. K. 2014

Habart, E., Walmsley, M., Verstraete, L., et al. 2005, Molecular hydrogen (Springer), 71–91, doi: 10.1007/1-4020-3844-5_3

Hacar, A., Tafalla, M., Kauffmann, J., & Ková cs, A. 2013, Astronomy &amp Astrophysics, 554, A55, doi: 10.1051/0004-6361/201220090

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

Hatchell, J., Thompson, M. A., Millar, T. J., & MacDonald, G. H. 1998, , 338, 713

Herbig, G. H. 1974, Lick Observatory Bulletin, 658, 1

Herbst, E., & van Dishoeck, E. F. 2009, Annu. Rev. Astron. Astrophys., 47, 427

Hildebrandt, H., Viola, M., Heymans, C., et al. 2016, Monthly Notices of the Royal Astronomical Society, 465, 1454, doi: 10.1093/mnras/stw2805

Holdship, J., Jeffrey, N., Makrymallis, A., Viti, S., & Yates, J. 2018, Astrophys. J., 866, 116

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., Viti, S., Codella, C., et al. 2019, The Astrophysical Journal, 880, 138, doi: 10.3847/1538-4357/ab1f8f

Hollenbach, D., & Salpeter, E. E. 1971, , 163, 155, doi: 10.1086/150754

Hoy, A. R., & Bunker, P. R. 1979, Journal of Molecular Spectroscopy, 74, 1, doi: 10.1016/0022-2852(79)90019-5

Impey, C. 2024, Astronomy Education Journal, doi: 10.32374/AEJ.AECON.2023.081aepp

Jarrett, T., Comrie, A., Marchetti, L., et al. 2021, Astronomy and Computing, 37, 100502, doi: https://doi.org/10.1016/j.ascom.2021.100502

Johansen, A. 2010, in International Encyclopedia of Education (Third Edition), third edition edn., ed. P. Peterson, E. Baker, & B. McGaw (Oxford: Elsevier), 245–252, doi: https://doi.org/10.1016/B978-0-08-044894-7.01347-6

Johnson, J. A. 2019, Science, 363, 474

Jordan, T. 2009, Information, Communication & Society, 12, 291, doi: 10.1080/13691180802552890

Joyce, J. 2021, in The Stanford Encyclopedia of Philosophy, Fall 2021 edn., ed. E. N. Zalta (Metaphysics Research Lab, Stanford University)

Keil, M., Viti, S., & Holdship, J. 2022, The Astrophysical Journal, 927, 203, doi: 10.3847/1538-4357/ac51d0

Knapp, G. R., & Jura, M. 1976, , 209, 782, doi: 10.1086/154776

Laas, Jacob C., & Caselli, Paola. 2019, A&A, 624, A108, doi: 10.1051/0004-6361/201834446

Larson, R. B. 1969, Monthly Notices of the Royal Astronomical Society, 145, 271, doi: 10.1093/mnras/145.3.271

Latter, W. B., & Black, J. H. 1991, , 372, 161, doi: 10.1086/169961

Lee, J., & Hammer, J. 2011, Academic Exchange Quarterly, 15, 1

Lefloch, B., Castets, A., Cernicharo, J., & Loinard, L. 1998, Astrophys. J., 504, L109

Lennard-Jones, J. E. 1931, Proceedings of the Physical Society, 43, 461, doi: 10.1088/0959-5309/43/5/301

Lewis, A., & Bridle, S. 2002, Physical Review D, 66, doi: 10.1103/physrevd.66.103511

Linnartz, H., Ioppolo, S., & Fedoseev, G. 2015, Int. Rev. Phys. Chem., 34, 205

Lique, F., Dubernet, M. L., Spielfiedel, A., & Feautrier, N. 2006, , 450, 399, doi: 10.1051/0004-6361:20054520

Liu, A. H., Zeiger, R., Sorkness, C., et al. 2007, J. Allergy Clin. Immunol., 119, 817

Liu, T., Wu, Y., & Zhang, H. 2013, , 775, L2, doi: 10.1088/2041-8205/775/1/L2

Lombardi, M., Alves, J., & Lada, C. J. 2006, A&A, 454, 781, doi: 10.1051/0004-6361:20042474

Looney, L. W., Mundy, L. G., & Welch, W. 2000, The Astrophysical Journal, 529, 477

Lynds, B. T. 1962, Astrophys. J. Suppl. Ser., 7, 1

Mandt, K. E., Lustig-Yaeger, J., Luspay-Kuti, A., et al. 2024, Science Advances, 10, eadp2191, doi: 10.1126/sciadv.adp2191

Mansfield, A., Nathanson, V., Jayesinghe, N., & Roycroft, G. 2020, Growing up in the UK: Ensuring a healthy future for our children, British Medical Association. https://www.bma.org.uk/what-we-do/population-health/improving-the-health-of-specific-groups/growing-up-in-the-uk-ensuring-a-healthy-future-for-our-children

Martín-Doménech, R., Jiménez-Serra, I., Muñoz Caro, G. M., et al. 2016, A&A, 585, A112, doi: 10.1051/0004-6361/201526271

McElroy, D., Walsh, C., Markwick, A. J., et al. 2013, A&A, 550, A36, doi: 10.1051/0004-6361/201220465

McKee, C. F. 1989, , 345, 782, doi: 10.1086/167950

McKee, C. F., & Ostriker, E. C. 2007, Annual Review of Astronomy and Astrophysics, 45, 565, doi: https://doi.org/10.1146/annurev.astro.45.051806.110602

McKee, C. F., & Ostriker, J. P. 1977, , 218, 148, doi: 10.1086/155667

Mestel, L., & Spitzer, L., J. 1956, Monthly Notices of the Royal Astronomical Society, 116, 503, doi: 10.1093/mnras/116.5.503

Murphy, K. P. 2012, Machine Learning, Adaptive Computation and Machine Learning series (London, England: MIT Press)

Nathan, R. A., Sorkness, C. A., Kosinski, M., et al. 2004, J. Allergy Clin. Immunol., 113, 59

Nelson, B., Ford, E. B., & Payne, M. J. 2013, The Astrophysical Journal Supplement Series, 210, 11, doi: 10.1088/0067-0049/210/1/11

Pepin, R. O. 1991, Icarus, 92, 2

Phuong, N. T., Chapillon, E., Majumdar, L., et al. 2018, A&A, 616, L5, doi: 10.1051/0004-6361/201833766

Pirronello, V., Biham, O., Liu, C., Shen, L., & Vidali, G. 1997, Astrophysical Journal, 483, L131–L134, doi: 10.1086/310746

Pudritz, R. E., & Norman, C. A. 1983, , 274, 677, doi: 10.1086/161481

Punanova, A., Caselli, P., Feng, S., et al. 2018, The Astrophysical Journal, 855, 112, doi: 10.3847/1538-4357/aaad09

Quénard, D., Jiménez-Serra, I., Viti, S., Holdship, J., & Coutens, A. 2018, Mon. Not. R. Astron. Soc., 474, 2796

Quireza, C., Rood, R. T., Balser, D. S., & Bania, T. M. 2006, The Astrophysical Journal Supplement Series, 165, 338, doi: 10.1086/503901

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

Ray, T. 2012, Astronomy Geophysics, 53, 5.19, doi: 10.1111/j.1468-4004.2012.53519.x

Ritter, A. 1898, , 8, 293, doi: 10.1086/140534

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

Rosenfield, P., Fay, J., Gilchrist, R. K., et al. 2018, The Astrophysical Journal Supplement Series, 236, 22, doi: 10.3847/1538-4365/aab776

Sabatini, G., Bovino, S., Giannetti, A., et al. 2020, Astronomy amp; Astrophysics, 644, A34, doi: 10.1051/0004-6361/202039010

Satherley, R.-M., Green, J., Sevdalis, N., et al. 2019, BMJ Open, 9, e027302

Savage, B. D., & Mathis, J. S. 1979, Annu. Rev. Astron. Astrophys., 17, 73

Schneider, N., Csengeri, T., Hennemann, M., et al. 2012, A&A, 540, L11, doi: 10.1051/0004-6361/201118566

Schöier, van der Tak, F. F. S., van Dishoeck, E. F., & Black, J. H. 2005, A&A, 432, 369, doi: 10.1051/0004-6361:20041729

Schöier, F. L., van der Tak, F. F. S., van Dishoeck, E. F., & Black, J. H. 2005, A&A, 432, 369, doi: 10.1051/0004-6361:20041729

Sini Kontkanen, Teija Koskela, O. K. S. K. K. W. M. M.-E., & Havu-Nuutinen, S. 2025, Studies in Science Education, 61, 89, doi: 10.1080/03057267.2024.2388931

Smith, S. A. 2019, Psychon. Bull. Rev., 26, 1213

Strom, S. E., Grasdalen, G. L., & Strom, K. M. 1974, , 191, 111, doi: 10.1086/152948

Taquet, V., Ceccarelli, C., & Kahane, C. 2012, Astronomy Astrophysics, 538, A42, doi: 10.1051/0004-6361/201117802

Taquet, V., Codella, C., De Simone, M., et al. 2020, Astron. Astrophys., 637, A63

Team, P. B. 2023, What is a Data Dashboard: Microsoft power bi, Microsoft. https://powerbi.microsoft.com/en-us/data-dashboards/

ter Braak, C., & Vrugt, J. 2008, Statistics and Computing, 18, 435, doi: 10.1007/s11222-008-9104-9

Tieftrunk, A., Pineau des Forets, G., Schilke, P., & Walmsley, C. M. 1994, , 289, 579

Tobin, J. J., Looney, L. W., Li, Z.-Y., et al. 2016, , 818, 73, doi: 10.3847/0004-637X/818/1/73

van der Tak, F. 2011, Proc. Int. Astron. Union, 7, 449

van der Tak, F. F. S., Black, J. H., Schöier, F. L., Jansen, D. J., & van Dishoeck, E. F. 2007, Astronomy Astrophysics, 468, 627–635, doi: 10.1051/0004-6361:20066820

van Dishoeck, E. F., & Blake, G. A. 1998, Annual Review of Astronomy and Astrophysics, 36, 317, doi: 10.1146/annurev.astro.36.1.317

Vastel, C., Ceccarelli, C., Lefloch, B., & Bachiller, R. 2014, The Astrophysical Journal, 795, L2, doi: 10.1088/2041-8205/795/1/l2

Vastel, C., Quénard, D., Le Gal, R., et al. 2018, Monthly Notices of the Royal Astronomical Society, 478, 5514–5532, doi: 10.1093/mnras/sty1336

Videnovik, M., Vold, T., Kionig, L., Madevska Bogdanova, A., & Trajkovik, V. 2023, International Journal of STEM Education, 10, doi: 10.1186/s40594-023-00447-2

Viner, R. M., Blackburn, F., White, F., et al. 2018, Archives of Disease in Childhood, 103, 128, doi: 10.1136/archdischild-2017-313307

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

Voigt, W. 1912, Über das Gesetz der Intensitätsverteilung innerhalb der Linien eines Gasspektrums. https://publikationen.badw.de/de/003395768

Wakelam, V., Loison, J.-C., Mereau, R., & Ruaud, M. 2017a, Mol. Astrophys., 6, 22

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. 2017b, Mol. Astrophys., 9, 1

Wakelam, V., Caselli, P., Ceccarelli, C., Herbst, E., & Castets, A. 2004, A&A, 422, 159, doi: 10.1051/0004-6361:20047186

Walker, J. 1987, The Journal of Economic Education, 18, 51. http://www.jstor.org/stable/1182396

Wiesenfeld, L., & Faure, A. 2013, Monthly Notices of the Royal Astronomical Society, 432, 2573, doi: 10.1093/mnras/stt616

Williams, D. A., & Viti, S. 2002, Annu. Rep. Prog. Chem., Sect. C: Phys. Chem., 98, 87, doi: 10.1039/B111165P

—. 2013, Observational Molecular Astronomy: Exploring the Universe Using Molecular Line Emissions, Cambridge Observing Handbooks for Research Astronomers (Cambridge University Press), doi: 10.1017/CBO9781139087445

Wilson, R. W., Jefferts, K. B., & Penzias, A. A. 1970, , 161, L43, doi: 10.1086/180567

Wolfe, I., Cass, H., Thompson, M. J., et al. 2011, BMJ, 342, d1277

Wolfe, I., Thompson, M., Gill, P., et al. 2013, Lancet, 381, 1224

Woods, P. M., Occhiogrosso, A., Viti, S., et al. 2015, Mon. Not. R. Astron. Soc., 450, 1256

Zapalska, A. M., Brozik, D., & Rudd, D. P. 2012, US-China education review, 2, 164. https://api.semanticscholar.org/CorpusID:17163475

Zieba, S., & Kreidberg, L. 2022, Journal of Open Source Software, 7, 4838, doi: 10.21105/joss.04838

Zucker, C., Schlafly, E. F., Speagle, J. S., et al. 2018, Astrophys. J., 869, 83

APPENDIX A

Healthcare Dashboard

A.1. OVERVIEW

Due to the interdisciplinary nature of the Horizon 2020 project, ACO, this thesis involves work applying and learning more about the skills and knowledge involved in data processing, data storage and data evaluation on a separate field. The aim is to then return to the astrochemical field with extra insights to support future projects. For this work, the chosen field was data intensive healthcare research. Healthcare research was chosen as it contains very complex data. Healthcare data spans a large variety of sources and depending on the form of studies being conducted, groups either collect their own clinical data, which will vary in size and depth based on the type of study being conducted, or they are granted access to patient records, which can contain millions of data points with very shallow information. Additionally, these methods are not mutually exclusive, which means that data management is important regardless of the type of study being conducted. Once data is collected, researchers then need to be able to work with the data in order to glean any important information that could be contained within it. This requires the data to be processed, cleaned and stored in a secure environment as the data often includes sensitive information. In order to look at one such project, we choose the Children and Young People's Health Partnership (CYPHP) Health check data (Satherley et al., 2019) and the data dashboard created for it. From the learnings of this work, we propose the creation of a dashboard for all public domain telescope data in chapter V. To describe the work in the healthcare field we first look at the kind of information that healthcare data contains and what a dashboard is.

A.1.1. Healthcare data

Healthcare data comes from many different sources, which can have multiple types and formats for the data. The most common categories are electronic health records (EHR), health surveys, and clinical trial data (Data-Resources, 2023). The different types of data have uses for different studies, and can be used in combination with each other. This is especially useful if it is possible to link patients across the different sources of data so that one complete record can be created, rather than multiple instances of a single patient without knowing it is the same person.

EHR data includes information from primary care providers such as those from surgeries or general practitioners (GPs) as well as secondary care providers such as hospitals. Regardless of the source of the data, it will include patient details such as age and sex. Then, depending on where the data came from, it should contain information detailing the reason for the patient's visit to the facility. For example, this could be the symptoms that caused the visit, or the chronic condition that causes a regular scheduled appointment among other information. Additional information, such as the procedures performed on a patient, the prescriptions they were given or other optional self identification data, such as ethnicity, can also be included.

Health surveys are usually conducted by larger organisations such as governmental agencies and are conducted in order to attempt to understand the health and well being of various groups. These types of surveys can include information necessary to understand if general treatment practices for certain diseases are resulting in general improvement for the well being of people that suffer it. They can also include information on where expenditures are being made that are justifiable for the outcome of the procedures they pay for. In general, health surveys attempt to garner a small targeted amount of data from a large population.

Clinical trial data can often be described as narrow but deep data collection. That is to say they either have more specific, narrow categories for the participants, or a limit on how many they can admit to the study, but for that, attempt to collect as much information as they can about those participants, rather than wanting just

a few specific points about a large population. This can include, but is not limited to, health record information, detailed information about home environment, or education and professional backgrounds. Other additions can also be made to clinical trial data, such as asking patients to adhere to certain restrictions in order to study results in a more isolated environment. All of this together, means that the amount of information per participant in a clinical trial is high. As this data is often entered manually into computers, it means that data entry for such trials can become a bottleneck.

When collecting large amounts of data in a hugely variable field such as health care, data will inevitably have errors in it. The errors and uncertainties in astronomical data can be quantified using measures for completeness, noise levels, as well as from understanding the noise induced by the physics surrounding the observations such as the atmospheric noise, and instrumental noise. Healthcare data errors however, are not as easily quantified due to the source of the errors. One major part is the errors introduced by human influence in collecting the data. A study that surveyed patients that had seen their records reported that one in five found some mistake in their records which ranged from serious medical errors, simple misspelling of names or an incorrect date of birth in patient records (Bell et al., 2020). That means, that a lot of errors in patient records are made by those entering data rather than through processes that can be characterised such as those found in astrophysics. Additionally data standards and systems can differ between institutions, and standards do change over time which means that sometimes there will be categorisations in one system that do not exist in another system or may actually be a collection of categories in a third system. On the other hand, there is also the intrinsic noise of patient health. What is meant with that, is that it is difficult to ensure that all patients will fit into neat categories without splitting categories down so far that they could almost be individually defined per patient. This leads to the necessity to clean data in ways that maximise the amount of usable information, while minimising the impact of human errors and system inconsistencies.

Because of the privacy issues inherent in dealing with healthcare data, it is vital to choose carefully where and how data is stored, even if extra steps are taken to

shield the identity of a patient. If cyber security precautions are taken so that a computer or other device can act as a trusted environment, or can securely connect to a trusted environment that is hosted on a remote server, then the choice of software technology used to store data will depend on how it will be interacted with. For example, some records are stored in excel files because of the simplicity of interacting with them. If a software to store data has layers of encryption or other forms of protection, then these can provide additional layers of security. However, these types of additional protection layers do not override the necessity for a secure environment.

A.1.2. Dashboards

Data dashboards function similar to the dashboard in a car. They display relevant information in an easy to understand, compact format using a large amount of data, or by simplifying complex data. Additionally they can connect different data types and platforms from a variety of sources to provide the insight that is desired by those who built the dashboard. These could be the performance of sales, patient attendance rates, or any other form of key performance indicators (KPIs) that are needed. The power that can be leveraged from such a tool, is therefore dependent on the quality and amount of data. Dashboards have use cases in many, if not all, data driven fields as long as there is insight to be gleaned from the data (Team, 2023).

As dashboards condense and simplify data, there are two main aims for creating and using them. The first is to allow for easy to understand insights despite the potential complexity of the data. The second aim is to allow for automated real-time updates. The example dashboard in figure A.1 shows a basic version of a dashboard using mock financial data. The financial data that is used to create this dashboard is not necessarily complex, but contains a lot of data points. Because of this, the dashboard collects the data points into convenient and easy to understand segments. In this way, the data can be understood at a glance. Additionally, as more information comes in, this dashboard updates to continuously inform the business owner on what the current trends in their company are.

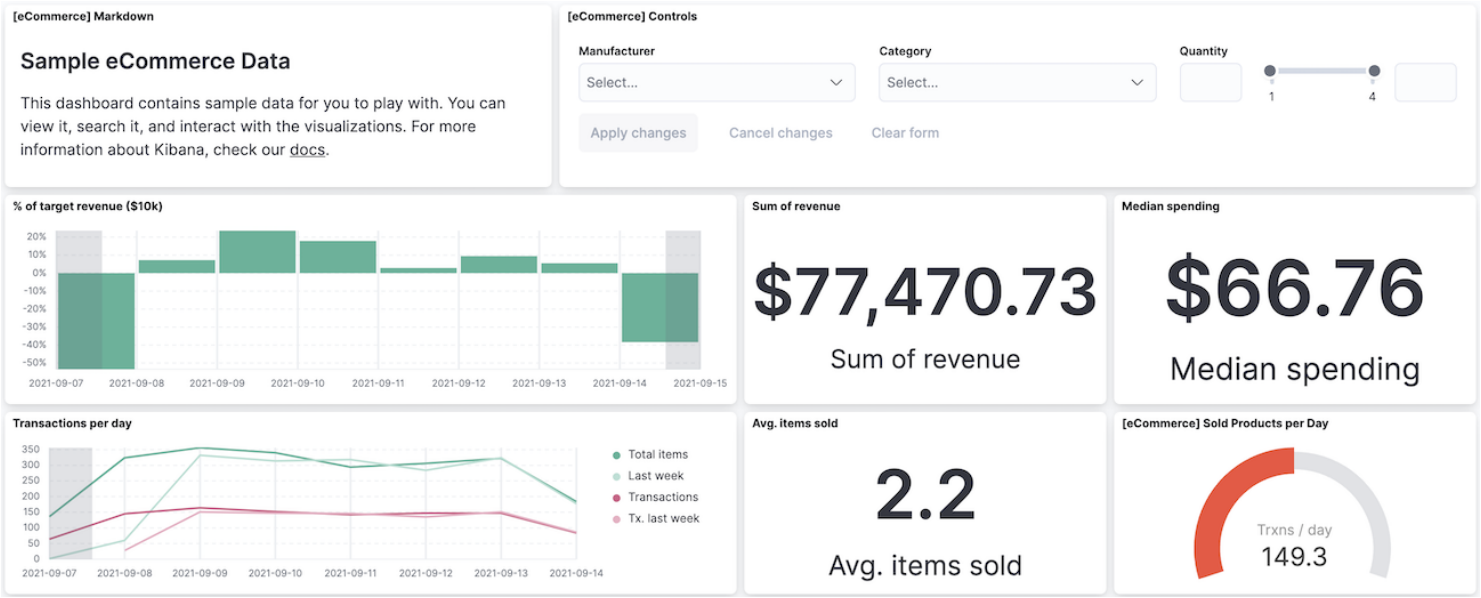

Figure A.1 Example image of a dashboard from the Elasticsearch guide for Kibana (elasticsearch, 2015)

A.2. CYPHP HEALTH CHECK DATA

Just like in a financial setting, dashboards can also be useful in a healthcare clinical trial setting. The data set we created a dashboard for in this work is the Children and Young People's health partnership (CYPHP) Health Check data. This data was collected from patients in the enhanced usual care (EUC, acting as a control) group during the time that the CYPHP Evelina London model of care (Satherley et al., 2019) was being evaluated. The aim of collecting the health check data, was to provide researchers with the extra data needed in order to quantify the evaluation of the Evelina London model of care. Additionally, as the dashboards are not limited only to the data stored for a single grouping, we can even explore additional external data where relevant. In order to showcase the impact that a dashboard can have we will first give some details on the health check data set we used.

In order to improve the outcome of healthcare for children and young people (CYP), the CYPHP continuously evaluates any developments they make to an integrated model of care as part of the health system they run. Some of these developments focus on proactive case finding, as well as training and education for professionals that work with CYP. One such evaluation is of the Evelina London

model of care. The Evelina London model integrates care for the CYP throughout various branches of healthcare. That is to say that primary care, such as a general physician (GP), and secondary care, such as a hospital, work closely, sharing more information on the patients that both are treating which leads to a clearer picture of the patient's health. Traditionally they would rely on requests for patient health records, or the patient supplying the records themselves. The Evelina London model treats both physical and mental health, while taking into account social context of the patient and their family. This model of care was proposed in order to address the outcome of patient health for CYP patients (Wolfe et al., 2011). At the time this model was proposed, CYP patients suffering from a long-term condition lacked long programs focusing on high-quality of care. Beyond this, because traditional care for CYP did not take into consideration the family needs of a patient, they were very ineffective leading to a reliance on acute treatments for long-term conditions (Wolfe et al., 2013; Mansfield et al., 2020; Viner et al., 2018).

In order to evaluate the CYPHP model of care, Satherley et al. (2019) had a group of general practices within which the trial would be conducted. Clustering the practices into 23 groups, they then randomly assigned each cluster to either receive the CYPHP model or the EUC model which served as a control group. The study was conducted over a two year period during which pseudonymised healthcare data was collected on CYP patients that were registered with the participating practices. CYP Patients that had asthma, eczema or constipation were asked to fill in an optional health check form, while the parents were asked to additionally fill out biopsychosocial questionnaires. The health check forms were comprised of multiple different types depending on the conditions they had. If a patient had asthma, an Asthma Control Test (ACT) score questionnaire (Nathan et al., 2004; Liu et al., 2007) was included, which asks questions regarding the patient's asthma in order to calculate a score on the severity. If a patient had eczema, then a Patient-Oriented Eczema Measure (POEM) questionnaire was included in order to monitor atopic eczema as experienced by the patient. Lastly, patients with constipation were given a questionnaire that was specifically created for the Satherley et al. (2019) study, in order to emulate the POEM and ACT questionnaires but for constipation.

As the CYPHP Evelina Model is designed to integrate secondary care data with

primary care data, as well as take the family into consideration with its treatment methods, additional data was needed for the EUC model so as to make a fair evaluation. In order to study the integration of secondary and primary care data, extra data from participating secondary care facilities was requested with consent of the patients or guardians. This means that hospital data was available for some of the patients that were in the EUC group. Data on the families was requested in optional forms with consent of the parents. The questions for parents focused on parental well-being such as mental health, as well as on the home environment of the patients. For example, some questions in the parental questionnaire asked if people smoked in the home, if parents were concerned about their own mental-well being, or if there was enough food in the home. Asking these types of questions supplements the EUC model with some of the information it needs in order to better understand which patients are or are not benefiting from the changes that the CYPHP Evelina Model makes to including family oriented issues into the healthcare treatment plan.

A.3. DATA PRODUCTS BY THIS WORK

There were 2 major requirements for this work, set by Dr Ingrid Wolfe (King's College London). The first was to find an alternative storage method to excel files, that would be easy to use and brows so that the data could still be used in the same way as before. The second was to create a dashboard that would allow researchers to quickly plot statistical information of the data while allowing researcher to put limits on what type of patients should be shown in those plots. For example, limiting the age range of the patients to include in plots. The dashboard that has been created for this work will be shown using mock data created to mimic the data from the CYPHP Health Checks from Satherley et al. (2019). Due to privacy regulations, we are not allowed to show the real data products and will instead show the dashboard using mock data. The mock data was not created in order to follow the statistical distribution of the real data, instead it was created only to show the functionality of the dashboard. Generation of mock data that follows statistical distributions of real healthcare data is still an emerging field and is not a subject of this thesis. The Dashboard has also been deployed to work with the

Table A.1. Example list of some errors that can arise in patient records, and some examples on how they could be managed.

Data error	Remedy Method example
Missing pseudonymised NHS number	Match patient record using combination of other markers together, else remove entry from data for statistical analysis.
Record with incorrect patient markers (e.g. Date of Birth)	use pseudonymised NHS number or other patient markers to match patient records.
Incorrect diagnosis code	If identified as true error, attempt to verify code using records of follow up visits, else remove entry from data for statistical analysis
Missing voluntary information (e.g. ethnicity)	If not present in any record, mark as not disclosed information.

full data, but is at time of writing only for internal use.

A.3.1. Raw CYPHP Data

As the data from the CYPHP Health Check comes in raw from those collecting the data, and the secondary care providers, the data has to be cleaned. A few examples of common patient record errors, and how they can be addressed for using the data in a research setting can be found in table A.1. Due to the complexity of the data, we have relied on the cleaning scripts previously validated on this data set for cleaning. The work here has focused on integrating them into a more streamlined processing pipeline. In order to verify how clean the data is post processing, we rely on cross referencing multiple entries of visits for individual patients in order to find the most frequent versions of data that should not change, such as date of birth, ethnicity, pseudonymised NHS number, or sex of patient. The first goal is then to create a data pipeline and storage solution, to allow incoming data to be cleaned and prepped automatically, so it can then be stored in a data storage solution that is secure. In doing this, we need to make sure that the data is easy to work with once it is in storage.

After that, the second goal is to build an additional Healthcare Dashboard as an example of how that technology can leverage the data in order to speed up

analysis, and improve potential future studies on such data sets. Both the pipeline and the dashboard need to be built in such a way that they can be maintained as incoming data standards and the needs of researchers change.

A.3.2. Data Pipeline: Raw Data to Semi-Structured Database

When the data is given to the research group conducting the study, it comes in from several surgeries, and currently from two different secondary care provides. There are many different standards in the medical field, which can either be used in isolation or in conjunction with each other. Because of this there are certain deviations between records that are to be expected. As time goes on, certain facilities may make changes to their data collection methods or standards, without altering previous records. Even when the same standard is used, the different facilities may abbreviate different columns in ways that disagree with each other. On top of this most of the data is filled in by members of staff from the respective facilities, which inevitably leads to mistakes being made. In order to address this we used the cleaning methods that were developed by the research team prior to this project. These methods address differences between the facilities formats, the differences between standards internally with historical data as well as between facilities, and to address the most common mistakes made by data entry personal and turned them into a data pipeline capable of being automated. The last step in the pipeline is to format the data so that it can easily be stored in an Elasticsearch Semi-structured database (elasticsearch, 2015). Elasticsearch is a search engine that can store records in JavaScript Object Notation (JSON). These JSON objects can be stored in an index, which is a grouping of similar objects. As each object is independent from the others in an index, each object can have its own field values. These fields are somewhat equivalent to a column in a table. This freely changeable, but still defined method of storing data is what classes Elasticsearch as a semi-structured database. As healthcare data contains many free form fields, with information which may or may not be sortable into categories by scripts, as well as information stored using standardisation which may change in the future, semi-structured data seemed the best fitting type of data structure for the health

check database. In using semi-structured data, if a free form field, filled in by a physician or clerk, can be categorised, then this can be done in post within the database, rather than needing to be done prior to storing the data. Additionally, in a semi-structured database, if a standardisation changes such that the entry type also changes, for example from integer to character, then this is easily changed unlike in a structured database.

To study the data from the health check, aforementioned scripts were created to clean the data of the potential patient information mistakes which according to Bell et al. (2020) can affect twenty percent of records. Beyond that, these scripts also match individual records if there are identifiers which indicate that the records stem from the same patient. However, these scripts are bespoke for different sources of data and need to be run manually. The data pipeline is able to take in files of data, match them to the source of origin before running the cleaning and matching scripts. This simplifies the procedure of cleaning the data as there is now no longer a need to match data to scripts manually. The last step then, is to take individual data points from the cleaned data and transform them into JSON format. This format is the standard that works with Elasticsearch. In doing this, we can give each data point a mapping so that Elasticsearch can store the data points in an easy to use and search format as well as add more information such as descriptions of what the various components are. This can then be stored in the Elasticsearch database so that it can be searched, and used in a dashboard.

A.3.3. Healthcare Dashboard

The data in this section and all figures within it are using mock data and do not follow true statistical distributions of healthcare data. In order to create an easy to use, yet potentially powerful way of leveraging the large quantity of data, we build a data dashboard using Kibana. The choice of Kibana is based on the fact that it is developed by the same group, and intended for the use with Elasticsearch. Kibana provides a browser based interface which allows administrators and users to search the Elasticsearch database it is connected to, as well as providing tools internally that can be used to develop and view dashboards. The potential of the dashboard is limited only by the data at hand, meaning that any number of distributions or

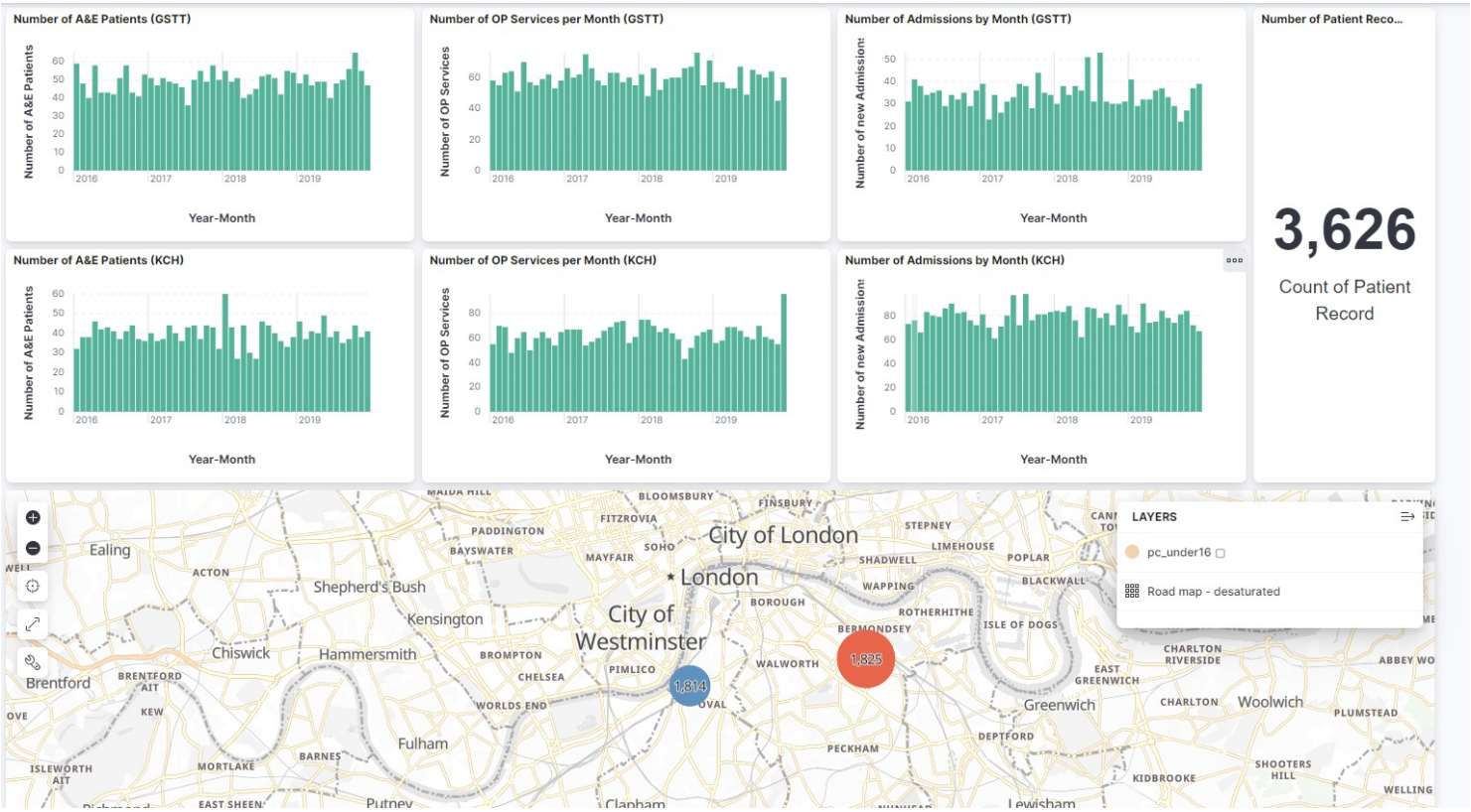

Figure A.2 Screenshot of the first part of the CYPHP Health check dashboard, using mock data. The first two rows show number of patients as a function of time for A&E visits, out-patient services and admission to hospital, but from separate sources, the top showing GSTT data, and the lower showing KCH data. To the right of that is the number of patient records currently available with any applied filters. The map shows the location of residence for patients. This figure is shown using mock data that does not follow true statistical distributions of healthcare data.

charts can be plotted. Therefore, it is important to know what a dashboard is meant to achieve, prior to building it. For the sake of this work, the main focus for the dashboard was to allow for an easy overview of the data, while maintaining the ability to create cohorts based on the Health Check questionnaires. Because of this, the dashboard is designed to show reduced down information of the secondary care data, with more detailed views of the Health Check responses. As this dashboard was designed to be used by both researchers and clinicians, the creation of it was performed iteratively with the director of the Institute of Women and Children's Health Dr Ingrid Wolfe (King's College London) along side Dylan Clarke (PhD at King's College London) and is already being used for planning purposes.

The dashboard takes in the cleaned data that is stored in the Elasticsearch database in order to create the plots in figures A.2, A.3, A.4, and A.5. These

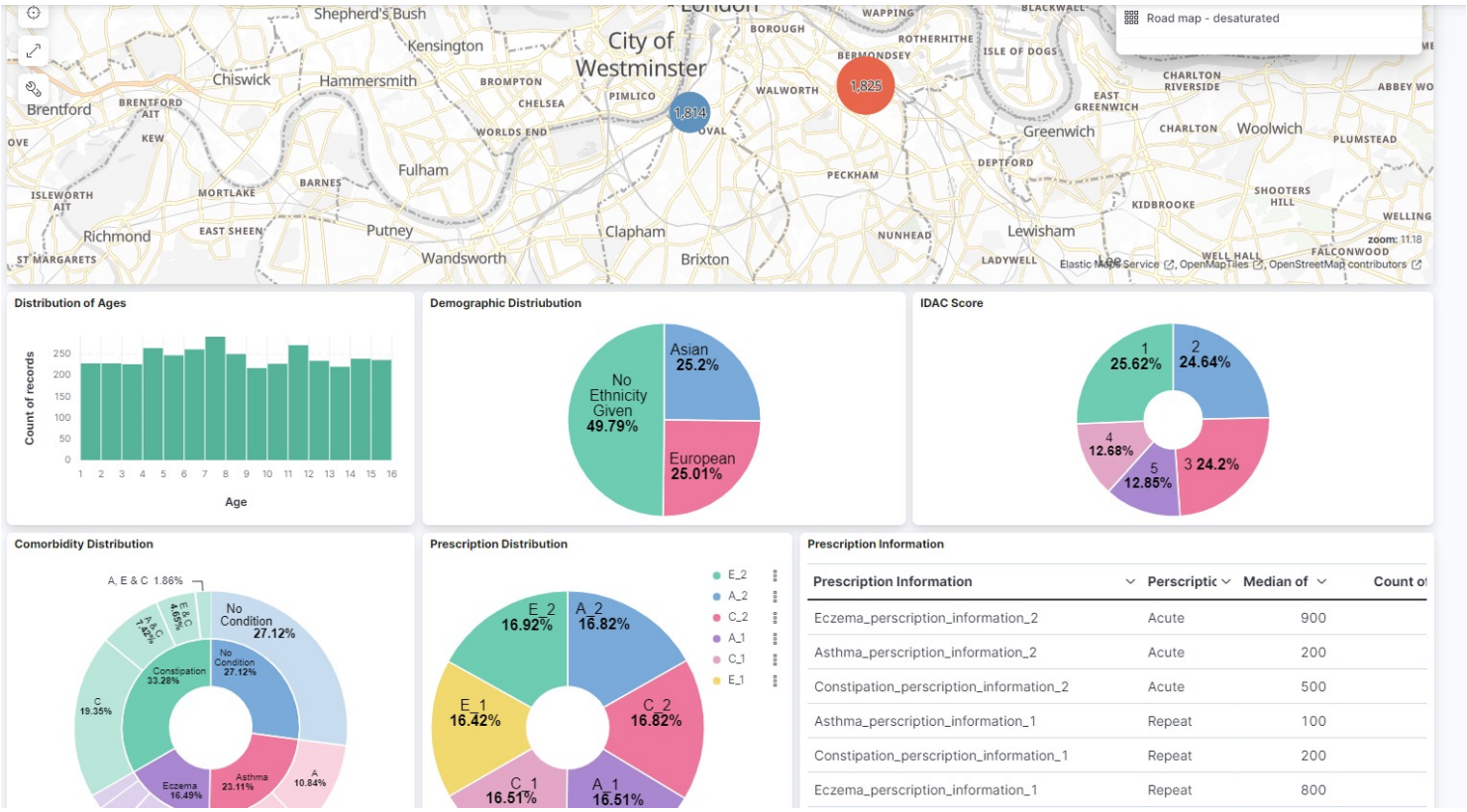

Figure A.3 Screenshot of the second part of the CYPHP Health check dashboard, using mock data. The map is explained in the caption of figure A.2, below that, from left to right, is the distribution of patient ages, the demographic distribution and followed by the distribution of income deprivation affecting children (IDAC) index. On the bottom row is then a distribution showing the distribution of patients suffering of any combination of constipation, asthma or eczema. To the right of that is then a distribution of prescribed medications, and the details of the most prescribed medications. This figure is shown using mock data that does not follow true statistical distributions of healthcare data.

figures are screenshots of the dashboard shown here with mock data, in order to avoid any potential privacy violations. The charts shown in the dashboard start with secondary care data from Guy's and St. Thomas' Trust (GSTT) and Kings College Hospital (KCH), showing the number of patients that used accident and emergency services, outpatient services and how many were admitted to hospital over night, each month. Below one can see a map of where patients live based on their Lower Layer Super Output Area (LSOA) derived from their postcodes. After the map is a distribution of the ages of patients, a pie-chart of the demographic that their patients claimed to belong to, as well as a pie-chart of the Income Deprivation Affecting Children (IDAC) Index derived from the LSOA codes of patients. The next row contains a donut chart which has two layers and the sum

Figure A.4 Screenshot of the third part of the CYPHP Health check dashboard, using mock data. The top row is described in the caption of figure A.3. Below that, is a clear representation of one of three nearly identical plots representing the values for asthma eczema and constipation. The pie-chart on the left shows the score for asthma (constipation, eczema) patients received based on the ACT (POEM, eczema) questionnaire given to them. To the right of this, is a heat-map of how many patients responded with what score to which question. This figure is shown using mock data that does not follow true statistical distributions of healthcare data.

of each component is equal to more than the sum of patients. The inner donut, is a count of how many patients had eczema, asthma, constipation or none of those. In this inner layer, if a patient marked all three, they would be represented once in each grouping. The outer layer of the donut shows how many patients indicated having more than one of the listed problems which one they have, or if they only have the first one they marked. After that is two charts showing prescription information. First a pie-chart of how often which medication is prescribed, and then a chart ranking the prescriptions by frequency of being distributed but showing more details of the individual medication. As this is mock data, we had also created random medications with no ties to current medications. After this is three rows with near identical charts. The charts in these three rows represent the responses

Figure A.5 Screenshot of the fourth part of the CYPHP Health check dashboard, using mock data. The pie-chart and heat-map are described in figure A.4. The last row of the dashboard indicated the parental questions that were posed to patients. This bar chart shows how many patients responded to the questions and colour codes them according to the answer they gave. This figure is shown using mock data that does not follow true statistical distributions of healthcare data.

patients gave to the questionnaires they were given if they had asthma, constipation or eczema. The pie-chart represents the number of patients in each score range for the given questionnaires. Next to that, is a heat map of the individual questions and given scores. The x-axis on these heat-maps is the questions, while the y-axis is the score given to the corresponding question. The colouring of individual boxes represents the number of patients that gave the corresponding answer to the given question. The last chart, is a bar chart of how many parents of patients responded to the parent questionnaire questions and colour codes the bars according to what the parents responded with to the questions.

Having this kind of dashboard with these plots, is one simple example of how data can be represented in a useful and easy to understand way. These plots, however, are not static. As more data is added to the database, the plots will automatically update. Additionally, if a user were to impose a limit on the data

they wish to view, then these plots would change accordingly. For example, if a user wished to only see information on patients younger than ten, then the plots would adjust to only use data for patients that fit that criteria. This allows a researcher using this dashboard, to quickly limit data to the cohorts for which they wish to perform an analysis while already providing some of the most common distributions that they may wish to look at for their study. Additionally, this dashboard can be expanded on by adding more charts, if that is desired.

A simple example of use cases that can be studied using just this dashboard is taking the income distributions for different areas, and seeing how the severity of various illnesses changes according to the deprivation of an area. For an example using external data such as pollution, we could overlap pollution maps with the patient location maps, and look at severity of asthma based on the ACT scores. By identifying if there is any correlations it could help support applications to conduct deeper studies. These are just two basic examples to showcase the potential of such a dashboard, as the traditional methods of getting information to support a grant application can be significantly more time consuming that having a dashboard like this.

The current capabilities of the dashboard allow users to plainly see trends in patient service use such as which areas are using services the most, comorbidity indicators such as which patients suffer from eczema and asthma at the same time. With applying a few filters to the distributions, it can become easy to see trends of children with asthma living with smokers, or healthcare service use based on deprivation scores of patients. The tool itself has already been used in order to support proposals for fund redistribution based on the trends observed within the dashboard.

A.4. CONCLUSION

Creating a data pipeline, storage solution and dashboard for a clinical healthcare dashboard has been a useful endeavour both for the research that can be done with it, and potentially for astronomical research, as discussed in chapter V. By incorporating the cleaning scripts previously created for the health check data with the data pipeline to reformat the data to fit into a semi-structured database, we

achieve an easily automated pipeline that can be used to work with new incoming data as well as the previously collected data. The dashboard shows the ease with which live data can be portrayed and worked with to improve research capabilities in the healthcare research setting.