UNIVERSITY COLLEGE LONDON

Faculty of Mathematical and Physical Sciences

Department of Physics & Astronomy

A STATISTICAL AND MACHINE LEARNING APPROACH TO THE STUDY OF ASTROCHEMISTRY

Thesis submitted for the Degree of

Doctor of Philosophy

by

Johannes Nasim Friedrich Heyl

Supervisors:

Prof. Serena Viti

Prof. Jonathan Tennyson

Examiners:

Prof. Wendy Brown

Prof. Ingo Waldmann

September 20, 2023

To my family

I, Johannes Nasim Friedrich Heyl, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis uses a variety of statistical and machine learning techniques to provide new insight into astrochemical processes. Astrochemistry is the study of chemistry in the universe. Due to the highly non-linear nature of a variety of competing factors, it is often difficult to understand the impact of any individual parameter on the abundance of molecules of interest. It is for this reason we present a number of techniques that provide insight.

Chapter 2 is a chemical modelling study that considers the sensitivity of a glycine chemical network to the addition of two H_2 addition reactions across a number of physical environments. This work considers the concept of a "hydrogen economy" within the context of chemical reaction networks and demonstrates that H_2 decreases the abundance of glycine, one of the simplest amino acids, as well as its precursors.

Chapter 3 considers a methodology that involves utilising the topology of a chemical network in order to accelerate the Bayesian inference problem by reducing the dimensionality of the parameters to be inferred at once. We demonstrate that a network can be simplified as well as split into smaller pieces for the inference problem by using a toy network.

Chapter 4 considers how the dimensionality can be simplified by exploiting the physics of the underlying chemical reaction mechanisms. We do this by realising that the most pertinent reaction rate parameter is the binding energy of the more mobile species. This significantly reduces the dimensionality of the problem we have to solve.

Chapter 5 builds on the work done in Chapters 3 and 4. The MOPED algorithm is utilised to identify which species should be prioritised for detection in order to reduce the variance of our binding energy posterior distributions.

Chapter 6 introduces the use of machine learning interpretability to provide better insights into the relationships between the physical input parameters of a chemical code and the final abundances of various species. By identifying the relative importance of various parameters and quantifying this, we make qualitative comparisons to observations and demonstrate good agreement.

Chapter 7 uses the same methods as in Chapters 4, 5 and 6 in light of new JWST observations. The relationship between binding energies and the abundances of species is also explored using machine learning interpretability techniques.

Impact statement

This thesis takes a multidisciplinary approach to the study of astrochemistry. Both Bayesian statistics and machine learning are used to provide novel insights into astrochemistry. The astrochemical community is the target audience for this thesis, which seeks to demonstrate how these powerful techniques can be used to generate a more sound understanding of observations and modelling.

In Chapter 2 we perform a sensitivity analysis in which we investigate the effect of adding the H_2 addition reactions $C + H_2 \longrightarrow CH_2$ and $CH + H_2 \longrightarrow CH_3$ to a glycine grain network. By demonstrating that we observe changes in the abundances of glycine, one of the simplest amino acids, and its precursors, we help explain why glycine has not been detected as well as how these addition reactions allow the network to tap into a previously unused hydrogen reservoir on the grains. This work has been published in Monthly Notices of the Royal Astronomical Society.

Chapter 3 focuses on exploiting the topology of a chemical network to reduce the computational expense of performing Bayesian inference to estimate reaction rates. By showing that one can reduce a toy network and split into smaller pieces on which the inference can be run in parallel, we develop a methodology that could be applied to larger grain-surface networks. This work has been published in The Astrophysical Journal.

Chapter 4 makes use of the physics of the grain-surface diffusion mechanism to demonstrate that the important parameters to be estimated are the binding energies of the most mobile species on the grains. Instead of estimating all the reaction rates, one can instead consider the binding energies, which significantly reduces the dimensionality of the reaction rate parameter estimation problem. This work has been published in The Astrophysical Journal.

Chapter 5 follows on from Chapter 4 in attempting to address the data paucity problem. Here, the MOPED algorithm is leveraged to identify species that should be priority targets for future ice observations. This work has been published in Monthly Notices of the Royal Astronomical Society.

Chapter 6 presents the first use of machine learning interpretability to understanding astrochemistry. By using SHapley Additive exPlanations (SHAP) values, a framework is created by which one can better understand the relationship between physical parameters and the abundances of species. Furthermore, a quantitative way of determining relative feature importance is presented. While this chapter presents a proof-of-concept, this methodology is ideal for detailed deep-dives and sensitivity analyses into individual astronomical objects that are typically the subject of modelling exercises. This work has been published in Monthly Notices of the Royal Astronomical Society.

Chapter 7 uses the same methods as Chapters 4, 5 and 6 to consider new JWST observations. The increased precision of these observations highlights the need for more species to be detected. SHAP values are used to demonstrate the relationship between binding energies and the abundances species, demonstrating its utility in another non-linear application. This work has been published in Faraday Discussions.

This PhD was done as part of the Centre for Doctoral Training in Data-Intensive Science. Alongside my astrochemistry research, I attended a number of machine learning and data science lecture courses. I also participated in a group project with the Office of National Statistics in order to apply my data science skills. The results of this project resulted in the publication "Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK" (Cheng et al. 2021) on which I was a joint-first author. I completed my mandatory data science placement at the Getting It Right First Time (GIRFT) programme, a division of NHS England. There I worked on applying the machine learning interpretability techniques that I use in this thesis to various medical problems. This resulted in a number of publications listed below:

- "Frailty, Comorbidity, and Associations With In-Hospital Mortality in Older COVID-19 Patients: Exploratory Study of Administrative Data" (Heyl et al. 2022)
- "Data quality and autism: Issues and potential impacts" (Heyl et al. 2023)
- "Data consistency in the English Hospital Episodes Statistics database" (Hardy et al. 2022), on which I was the second author

- "Estimating nosocomial infection and its outcomes in hospital patients in England with a diagnosis of COVID-19 using machine learning" (Hardy et al. 2023) on which I was the second author
- "Role of hospital strain in determining outcomes for people hospitalised with COVID-19 in England" (Gray et al. 2023).

One final first-author paper on spinal surgery has been submitted for publication.

Acknowledgements

This thesis is the product of a number of years of work that, while very stressful, have been some of the most intellectually stimulating in my life. I have thoroughly enjoyed my PhD and the reason that I will look back fondly on these years is due to my interactions and collaborations with a number of people who it would be amiss not to thank.

First and foremost, I would like to thank my supervisor, Serena Viti, for all that she has done to make the PhD a thoroughly enjoyable experience. Throughout the PhD, she has provided me with the intellectual freedom to pursue my interests within academia as well as in industry, whilst always being available when I was struggling with something. I would also like to thank Jonathan Holdship, who was a second supervisor in all but name, for his infinite patience with my sometimes inane questions about astrochemistry and UCLCHEM as well as for all his advice about life after academia. I would also like to thank my other collaborators who provided me with intellectually stimulating discussions on our projects: Stephen Feeney and Elena Sellentin for all things Bayesian, Thanja Lamberts for many chats on the intricate details of astrochemistry, Gijs Vermariën for often-times mind-blowing chats on how machine learning could revolutionise astrochemistry and Joshua Butterworth for his insight on molecular tracer ratios. I am indebted to all of you for your help with our various projects.

I would also like to thank the wider Viti group, both in London and in Leiden, for making the past few years so enjoyable. In particular, I am grateful to the Leiden group for always making me feel so welcome when I used to come to visit!

The Centre for Doctoral Training in Data-Intensive Science has provided a fantastic programme that I have thoroughly enjoyed. A big thank you to the management team for all their help and support when I had questions.

My time working at the Getting It Right First Time (GIRFT) programme was inspirational. I would like to thank Jeremy Yates for helping organise the placement as well as for the pastoral support. A big thank you as well to the GIRFT team including William Keith Gray, Katie Tucker and Adrian Hopper who showed me just how interesting the world of healthcare was and provided fantastic mentorship. This placement not only provided me with new ideas for my doctoral work, with the machine learning interpretability chapters in this thesis stemming from GIRFT projects, but also has motivated me to pursue health data science after my doctorate.

Thank you to my friends in London for some of the best years of my life. While the PhD process was oftentimes very stressful, you were all incredibly supportive and helped me achieve a better balance throughout. Thank you also to my old mathematics teacher and friend, Bryan Landmann, for being a massive source of inspiration and support when I began my academic journey all those years ago.

Last but certainly not least, I would like to thank my family. My parents have been incredibly supportive and encouraging to me in all my endeavours which has provided me with the platform to always seek challenges and believe in myself. I will forever be indebted to them. I would also like to thank my dog, Ruby, who always made sure I took her for a walk when the PhD work became a bit overwhelming!

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Investigating the impact of reactions of C and CH with molecular hydrogen on a glycine gas-grain network

   (b) **Please include a link to or doi for the work:**
   https://academic.oup.com/mnras/article/520/1/503/6987286

   (c) **Where was the work published?** Monthly Notices of the Royal Astronomical Society

   (d) **Who published the work?** Oxford University Press

   (e) **When was the work published?** 13 January 2023

   (f) **List the manuscript's authors in the order they appear on the publication:** Johannes Heyl, Thanja Lamberts, Serena Viti, and Jonathan Holdship

   (g) **Was the work peer reviewed?** Yes.

   (h) **Have you retained the copyright?** Yes.

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**
   https://arxiv.org/abs/2301.04324
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'? If 'Yes', please please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**
   Johannes Heyl: Conceptualization, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing.
   Thanja Lamberts: Conceptualization, Validation, Writing – review & editing.
   Serena Viti: Conceptualization, Formal analysis, Writing – review & editing.
   Jonathan Holdship: Conceptualization, Writing – review & editing.
4. **In which chapter(s) of your thesis can this material be found?** Chapter 2

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Johannes Heyl
Date: 2 August 2023
Supervisor/Senior Author signature (where appropriate)**:** Serena Viti
Date: 2 August 2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Exploiting network topology for accelerated Bayesian inference of grain surface reaction networks

   (b) **Please include a link to or doi for the work:**
   https://iopscience.iop.org/article/10.3847/1538-4357/abbeed/meta

   (c) **Where was the work published?** The Astrophysical Journal

   (d) **Who published the work?** IOP Publishing

   (e) **When was the work published?** 4 December 2020

   (f) **List the manuscript's authors in the order they appear on the publication:** Johannes Heyl, Serena Viti, Jonathan Holdship, and Stephen M. Feeney.

   (g) **Was the work peer reviewed?** Yes.

   (h) **Have you retained the copyright?** Yes.

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**
   https://arxiv.org/abs/2010.02877
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'? If 'Yes', please please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**
   Johannes Heyl: Conceptualization, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing.
   Serena Viti: Conceptualization, Formal analysis, Writing – review & editing.
   Jonathan Holdship: Conceptualization, Formal analysis, Validation, Writing – review & editing.
   Stephen M. Feeney: Formal analysis, Validation, Writing – review & editing.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 3 and Appendix A

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Johannes Heyl
Date: 2 August 2023

Supervisor/Senior Author signature (where appropriate)**:** Serena Viti
Date: 2 August 2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Using Statistical Emulation and Knowledge of Grain-surface Diffusion for Bayesian Inference of Reaction Rate Parameters: An Application to a Glycine Network

   (b) **Please include a link to or doi for the work:** https://iopscience.iop.org/article/10.3847/1538-4357/ac6606/meta

   (c) **Where was the work published?** The Astrophysical Journal

   (d) **Who published the work?** IOP Publishing

   (e) **When was the work published?** 20 May 2022

   (f) **List the manuscript's authors in the order they appear on the publication:** Johannes Heyl, Jonathan Holdship, and Serena Viti

   (g) **Was the work peer reviewed?** Yes.

   (h) **Have you retained the copyright?** Yes.

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**
   https://arxiv.org/abs/2204.08347
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'? If 'Yes', please please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**
   Johannes Heyl: Conceptualization, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing.
   Jonathan Holdship: Conceptualization, Formal analysis, Validation, Writing – review & editing.
   Serena Viti: Conceptualization, Formal analysis, Writing – review & editing.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 4 and Appendix B

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Johannes Heyl
Date: 2 August 2023

Supervisor/Senior Author signature (where appropriate)**:** Serena Viti
Date: 2 August 2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Identifying the most constraining ice observations to infer molecular binding energies

   (b) **Please include a link to or doi for the work:** https://academic.oup.com/mnras/article-abstract/517/1/38/6702733

   (c) **Where was the work published?** Monthly Notices of the Royal Astronomical Society

   (d) **Who published the work?** Oxford University Press

   (e) **When was the work published?** 17 September 2022

   (f) **List the manuscript's authors in the order they appear on the publication:** Johannes Heyl, Elena Sellentin, Jonathan Holdship, and Serena Viti

   (g) **Was the work peer reviewed?** Yes.

   (h) **Have you retained the copyright?** Yes.

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**
   https://arxiv.org/abs/2209.09347
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'? If 'Yes', please please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

(e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**
   Johannes Heyl: Conceptualization, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing.
   Elena Sellentin: Validation, Writing – review & editing.
   Jonathan Holdship: Conceptualization, Formal analysis, Writing – review & editing.
   Serena Viti: Conceptualization, Formal analysis, Writing – review & editing.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 5

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Johannes Heyl
Date: 2 August 2023

Supervisor/Senior Author signature (where appropriate)**:** Serena Viti
Date: 2 August 2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Understanding Molecular Abundances in Star-Forming Regions Using Interpretable Machine Learning

   (b) **Please include a link to or doi for the work:** https://doi.org/10.1093/mnras/stad2814

   (c) **Where was the work published?** Monthly Notices of the Royal Astronomical Society

   (d) **Who published the work?** Oxford University Press

   (e) **When was the work published?** 13 September 2023

   (f) **List the manuscript's authors in the order they appear on the publication:** Johannes Heyl, Joshua Butterworth, and Serena Viti

   (g) **Was the work peer reviewed?** Yes.

   (h) **Have you retained the copyright?** Yes.

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi** https://arxiv.org/abs/2309.06784
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   □ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'? If 'Yes', please please give a link or doi:** No.

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**
   Johannes Heyl: Conceptualization, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing
   Joshua Butterworth: Validation, Writing – review & editing
   Serena Viti: Conceptualization, Formal analysis, Writing – review & editing.
4. **In which chapter(s) of your thesis can this material be found?** Chapter 6 and Appendix C

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Johannes Heyl
Date: 14 September 2023
Supervisor/Senior Author signature (where appropriate)**:** Serena Viti
Date: 14 September 2023

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** A statistical and machine learning approach to the study of astrochemistry

   (b) **Please include a link to or doi for the work:**
   https://pubs.rsc.org/en/content/articlelanding/2023/fd/d3fd00008g/unauth

   (c) **Where was the work published?** Faraday Discussions

   (d) **Who published the work?** Royal Society of Chemistry

   (e) **When was the work published?** 13 June 2023

   (f) **List the manuscript's authors in the order they appear on the publication:** Johannes Heyl, Serena Viti, and Gijs Vermariën

   (g) **Was the work peer reviewed?** Yes.

   (h) **Have you retained the copyright?** Yes.

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi**
   https://arxiv.org/abs/2306.05790
   If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

   ☐ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

   (a) **What is the current title of the manuscript?**

   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'? If 'Yes', please please give a link or doi:**

   (c) **Where is the work intended to be published?**

   (d) **List the manuscript's authors in the intended authorship order:**

   (e) **Stage of publication:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:**
   Johannes Heyl: Conceptualization, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing.
   Serena Viti: Conceptualization, Data curation, Formal analysis, Writing – original draft, Writing – review & editing.
   Gijs Vermariën: Writing – review & editing
4. **In which chapter(s) of your thesis can this material be found?** Chapter 7

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)**:**

Candidate: Johannes Heyl
Date: 2 August 2023

Supervisor/Senior Author signature (where appropriate)**:** Serena Viti
Date: 2 August 2023

Hobbes: What a clear night! Look at all the stars. Millions of them!

Calvin: Yes, we're just tiny specks on a planet particle, hurling through the infinite blackness.

Calvin: Let's go in and turn on all the lights.

Calvin and Hobbes

Contents

This page was intentionally left blank

List of Figures

List of Tables

This page was intentionally left blank

Chapter 1

Introduction

1.1 Overview

It is often assumed that space consists of a near-vacuum. From this, it is difficult to imagine how such conditions could eventually lead to the formation of life, let alone the formation of galaxies, stars and planets. However, this assumption is false. The interstellar medium (ISM) is a multi-faceted set of environments with a wide range of physical conditions. It is this variety that leads to the many molecules that observations have detected. If we are to understand how and why these molecules form, it is paramount that we have a good understanding of the chemistry.

When we speak about the chemistry, what we really are interested in is:

- The chemical network, that is the set of all species and reactions between these species
- The chemical mechanisms that allow these reactions to happen

The interstellar medium consists of a gas phase as well as solid phase of interstellar dust. The dust particles are typically composed of silicates and range in size from 10-500 nm (Herbst and van Dishoeck 2009). These dust particles are crucial to the process of ISM chemistry. At the low temperatures of pre-stellar cores, it is only dust-grain chemistry

Region	n_H (cm^{-3})	T (K)
Coronal gas	$<10^{-2}$	5×10^5
HII regions	>100	1×10^4
Diffuse gas	100 - 300	70
Molecular clouds	10^4	10
Prestellar cores	$10^5 - 10^6$	10-30
Star-forming regions	$10^7 - 10^8$	100-300
Protoplanetary disks	10^4(outer)-10^{10}(inner)	10(outer)-500(inner)
Envelopes of evolved stars	10^{10}	2000-3500

Table 1.1: *A summary of the various regions in the ISM along with their typical densities and temperatures taken from Williams and Viti (2013).*

that is efficient as opposed to gas-phase chemistry, which is why this is the primary focus of this thesis.

1.2 The Interstellar Medium

1.2.1 Introduction to the Interstellar Medium

The interstellar medium is the space between stars and is mostly gaseous. However, it is not homogenous in nature, which is why it is the birthplace of stars. In fact, it is home to many different environments which correspond to the various stages of star formation. We summarise these in Table 1.1. The interstellar medium in our galaxy is 99% gas and 1% dust-grains. In terms of atomic composition, the ISM is 70% hydrogen, 28% helium and 2% other, heavier elements by mass. Heavier elements are produced via stellar nucleosynthesis within stars and supernovae.

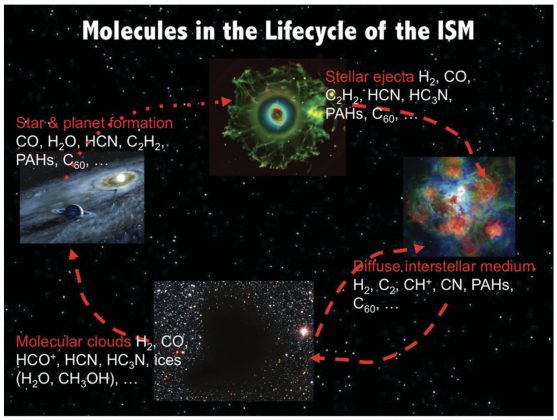

Figure 1.1: *Diagram of the star formation cycle in the ISM taken from Tielens (2013).*

1.2.2 The Star Formation Life Cycle

The star formation cycle is crucial to the chemistry of the ISM, as the heavier elements produced in stars and supernovae are returned back to the ISM via stellar winds. We briefly describe the parts of the star formation life cycle that are pertinent to this thesis.

Pre-Stellar Cores

The star formation process begins with diffuse gas clouds that undergo isothermal gravitational collapse. This collapse is caused by the gravitational force exceeding the thermal and magnetic pressure of the cloud. This is typically modelled by considering the critical mass of a Bonnor-Ebert sphere (Bonnor 1956) and equating this with the critical mass of a magnetically-supported core (McKee 1989) which is formulated as:

$$M_* = \frac{1.18 c_s^4}{\sqrt{G^3 P_0}} + 2 k_\phi \phi, \tag{1.1}$$

where c_s is the speed of sound, P_0 the surface pressure, ϕ the magnetic flux, G Newton's gravitational constant and k_ϕ a constant. If a core has a mass greater than the critical mass, then it is able to overcome its own thermal and magnetic pressure in the absence of turbulence.

At pre-stellar temperatures (typically around 10 K), there is significant adsorption of species and molecules to the dust surface. Due to the relatively large abundances of H, C and O, the dust particles begin to form "ices" composed of combinations of these species, such as H_2O, CO, CO_2, H_2CO and CH_3OH. This process is typically referred to as "freeze-out". During this time, the chemistry will be dominated by grain-surface chemistry and there are a number of processes that can occur on these grains. Figure 1.2 demonstrates the types of processes that can take place on the grains.

The Warm-Up Phase

The core initially begins to collapse isothermally. However, this does not occur indefinitely, as the gas begins to heat up and form a protostar. This increase in the temperature results in an increase of the internal pressure which balances out the gravitational pressure, thereby halting the collapse phase. The increase in the temperature results in the desorption of many of the molecules that are on the grain-surface ices into the gas-phase

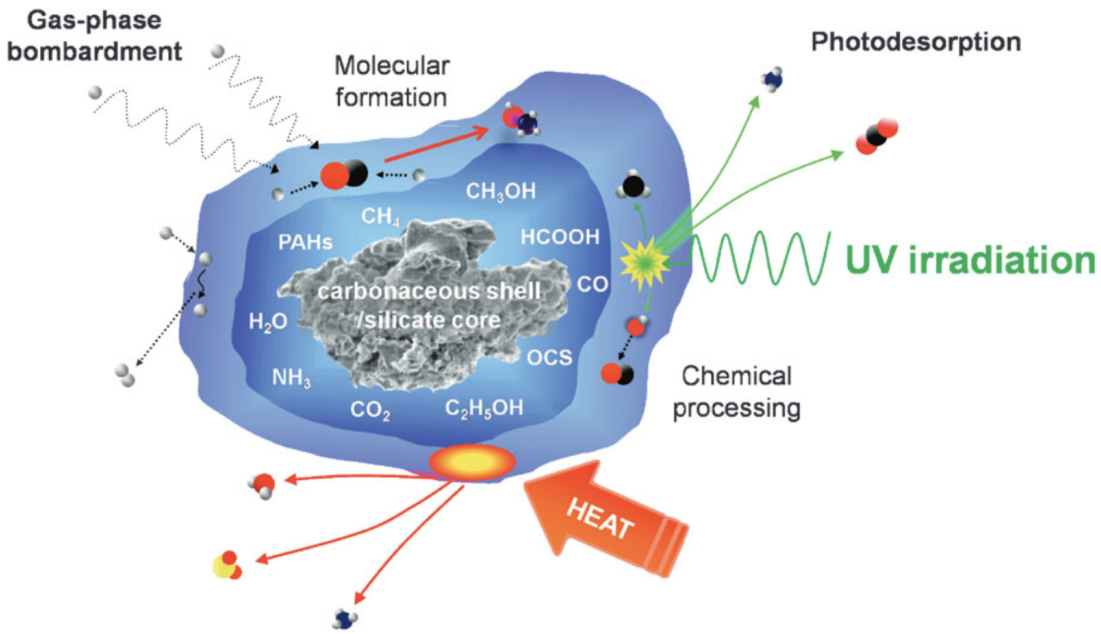

Figure 1.2: *Illustration of the types of processes that can take place on an interstellar dust grain. Figure taken from Burke and Brown (2010).*

once the temperatures approach their respective evaporation temperatures (Awad et al. 2010). This forms a "stellar gas-phase" which is distinct from the general ISM gas-phase due to these desorbed molecules being the result of grain-surface chemistry. Unlike the previous collapse phase in which most of the chemistry takes place on the interstellar dust grains, here we find that the higher temperatures allow for gas-phase chemistry to dominate (Viti et al. 2004). This is not to say that there is no grain-surface chemistry. Grain-surface reactions will increase in efficiency as the temperature increases, but there is added competition due to evaporation. This results in a very different chemistry which we now discuss.

1.3 Interstellar Chemistry

1.3.1 Dust-Grain Surface Chemistry

Interstellar dust is a crucial part of the interstellar medium. The dust grains are responsible for much of the rich chemistry observed, as they act as an energy sink, allowing reactions between adsorbed atoms and molecules (van Dishoeck 2014). There is strong evidence to suggest that complex-organic molecules (COMs), which are defined as carbon-based molecules with six or more atoms, form on these interstellar dust grains (Herbst and van

Dishoeck 2009; Jin and Garrod 2020). In fact, this rich chemistry is thought to take place long before stars have formed, during the dark cloud phase. In the last couple of decades many experiments have been performed to determine the surface reactions occurring on the icy mantles (e.g. see review by (Linnartz et al. 2015)). However, surface reactions in the laboratory are typically only investigated within a narrow range of physical parameters constrained by what is available with the experimental techniques employed and thus often not fully representative of the ISM conditions. Hence the available experimental data for interstellar ices are limited. This issue is further compounded by the lack of available observation data. Boogert et al. (2015) provides a review of the observed ices, but these abundance measurements had large relative uncertainties. However, towards the end of this PhD the James Webb Space Telescope (JWST) provided the first set of ice observations from a dedicated ice observation mission (McClure et al. 2017) which has now provided a new set of abundance measurements (McClure et al. 2023). Prior to JWST, there had been insufficient resolution of the absorption band profiles of molecules from previous missions such as the Infrared Space Observatory (ISO) and Spitzer. JWST, which observes in the infrared wavelength range of 0.6 - 28 μm, has up two magnitudes greater spectral resolution, especially in the 5-8 μm range. This range is important because it is likely to contain the vibrational modes of several molecules of interest (Boogert et al. 2015; Boogert 2016). Greater resolution is important so as to ensure that various absorption band profiles can be better attributed to the relevant molecules, especially as various functional groups may differ by species but can have similar values (Boogert 2016).

In fact, it is widely believed that complex-organic molecules (COMs) form on interstellar dust (Herbst and van Dishoeck 2009; Caselli and Ceccarelli 2012). In certain cases, grain-surface reactions are more efficient than gas-phase reactions. However, there exists much debate about the stage of star formation during which these molecules are produced. While modelling has shown that molecules such as glycine can be formed during the warm-up phase of star formation (Garrod 2013), there is also evidence that suggests that dark interstellar cloud conditions would suffice (Ioppolo et al. 2020).

Grain Surface Diffusion

An understanding of the actual grain surface mechanisms will prove crucial in this work.

According to the Langmuir-Hinshelwood mechanism which was developed in Hasegawa et al. (1992), the rate at which two species A and B react via diffusion is given by:

$$k_{AB} = \kappa_{AB} \frac{(k_{hop}^{A} + k_{hop}^{B})}{N_{site} n_{dust}}, \tag{1.2}$$

where N_{site} is the number of sites on the grain surface and n_{dust} is the number density of dust grains. k_{hop}^{X} is the thermal hopping rate of species X on the grain surface defined as:

$$k_{hop}^{X} = \nu_0 \exp\left(-\frac{E_b}{T_{gr}}\right), \tag{1.3}$$

where E_b is the diffusion energy of the species, T_{gr} is the grain temperature and ν_0 is the characteristic vibration frequency of species X. The diffusion energy is typically taken to be a fraction of the species binding energy, E_D. The characteristic vibration frequency, ν_0, is defined as:

$$\nu_0 = \sqrt{\frac{2 k_b n_s E_D}{\pi^2 m}}, \tag{1.4}$$

where k_b is the Boltzmann constant, n_s is the grain site density and m is the mass of species X.

Finally, κ_{AB}, which provides the reaction probability is taken to be:

$$\kappa_{AB} = \max\left(\exp\left(-\frac{2a}{\hbar}\sqrt{2\mu k_b E_A}\right), \exp\left(-\frac{E_A}{T_{gr}}\right)\right), \tag{1.5}$$

where $\hbar$ is the reduced Planck constant, μ is the reduced mass, E_A is the reaction activation energy, k_b is Boltzmann's constant and $a = 1.4$ Å is the thickness of a quantum mechanical barrier. The reaction probability is effectively a competition between the first term, which is the quantum mechanical probability of a tunnelling through a rectangular barrier of thickness a and the thermal reaction probability, the second term.

A correction needs to be made to κ_{AB} to account for the fact that species might diffuse or evaporate instead of reacting with each other. This correction is the reaction-diffusion competition (Chang et al. 2007; Garrod and Pauly 2011). The reaction probability is defined to be:

$$\kappa_{AB}^{final} = \frac{p_{reac}}{p_{reac} + p_{diff} + p_{evap}}, \tag{1.6}$$

where p_{reac}, p_{diff} and p_{evap} represent the probabilities of species A and B reacting, diffusing and evaporating per unit time, respectively. These quantities are defined as:

$$p_{reac} = \max(\nu_0^A, \nu_0^B)\kappa_{AB}, \tag{1.7}$$

$$p_{diff} = k_{hop}^A + k_{hop}^B \tag{1.8}$$

and

$$p_{evap} = \nu_0^A \exp\left(-\frac{E_D^A}{T_{gr}}\right) + \nu_0^B \exp\left(-\frac{E_D^B}{T_{gr}}\right). \tag{1.9}$$

In equation 1.2, κ_{AB} is replaced with κ_{AB}^{final}.

Another important grain-surface reaction mechanism is the Eley-Rideal mechanism. In this scenario, the reaction rate that is modelled is the reaction between an incident species from the gas-phase reacting with an adsorbed species on the grain-surface. The consequence of this is that the product is either desorbed into the gas-phase or the grain-surface is heated up, depending on the exothermicity of the reaction. The reaction rate, R_{ij}, between gas-phase species i and adsorbed species j via this mechanism is calculated as:

$$R_{ij} = \eta_j \sigma_d \langle v(i) \rangle n(i) n_d, \tag{1.10}$$

where η_j is the adsorbed species's average density on the grain-surface, σ_d is the grain's cross-section, $\langle v(i) \rangle$ is the thermal velocity of the incident species, $n(i)$ the abundance of the incident species and n_d is the grain number density (Ruaud et al. 2015).

For most reactions, this mechanism is typically negligible due to the dependence on

the number density of the incident species.

1.3.2 Gas-Phase Chemistry

While gas-phase chemistry is not the focus of the work done in this thesis, we briefly describe it for the sake of completeness. The only reactions that are efficient in the molecular clouds, prestellar cores and star-forming regions we consider in this work are 2-body reactions with 3-body reactions requiring densities of about 10^{13} cm^{-3} (Puzzarini 2022). The reaction rate between any two species is parameterised by an Arrhenius equation of the form:

$$k = \alpha \left(\frac{T}{300K} \right)^{\beta} e^{-\frac{-\gamma}{T}}, \tag{1.11}$$

where T is the gas-phase temperature and α, β and γ are reaction-dependent constants that are determined experimentally. The constant α can be interpreted as a frequency of collision between the two reacting species, β provides a temperature-dependent fudge-factor to this collisional frequency and γ is the reaction activation energy. In fact, the final exponential term, effectively represents the fraction of collisions that result in a reaction. These constants are summarised in reaction databases such as UMIST (McElroy et al. 2013) and KIDA (Wakelam et al. 2015).

1.4 Astrochemical Modelling

In order to better understand the interstellar medium and make predictions, we must make use of numerical codes. If our codes are to be able to produce reasonable approximations of observations, then we can be assured that we have a decent understanding of the underlying physics and chemistry. Many of these codes rely on the integration of systems of ordinary differential equations (ODEs) to describe the time-evolution of molecular abundances. On the most fundamental level, any time-dependent chemical model that aims to solve for the abundance of species i, which we denote by n_i, must solve a system of equations of the form:

$$\frac{dn_i}{dt} = \sum production - \sum destruction, \tag{1.12}$$

where we encode all the production mechanisms for species i in the first term on the right-hand side and all the destruction mechanisms in the second term on the right-hand side (Wakelam et al. 2013). This is a general framework for astrochemical models with the individual production and destruction mechanisms being dependent on whether the species is in the gas phase or on the grains as well as the implemented chemical mechanisms of the code in question. Note that in parallel to the chemical evolution of the system there can also be a physical evolution such as an increase in density of the object, such as with the pre-stellar phase collapse or a change in temperature, as is the case with the warm-up phase. These have an influence on the densities of various molecules as well as on the reaction rates of both grain-surface and gas-phase reactions.

1.4.1 UCLCHEM

In this thesis, we will be using the UCLCHEM code which is an open-source time-dependent gas-grain astrochemical code[1]. UCLCHEM models two phases of star formation. Phase 1 corresponds to the isothermal collapse of a diffuse cloud of gas to a fixed density. Phase 2 begins at this density and models the warm-up phase of star evolution. Throughout both phases, the model calculates the abundances of all the species in the chemical network via a system of ODEs. The ODE for each species can be written as:

$$\frac{dn_i}{dt} = \sum_{l,m} k^i_{lm} n_l n_m - n_i \sum_{i \neq r} k_r n_r - k^{des}_i n_i + k^{ads}_i n_{i,gas}, \tag{1.13}$$

where $k^i_{lm}$is the reaction rate of the reactions between species l and m to produce species i, n_i is the abundance of species i, k_r represents the reaction rates of all reactions where species i is consumed as a reactant and k^{des}_i and k^{ads}_i represent the desorption and adsorption rates of the species. The coupled differential equations represent the formation and destruction mechanisms for all the relevant species. The diffusion mechanism described in Hasegawa et al. (1992) was implemented in UCLCHEM in Quénard et al. (2018). More details about UCLCHEM's implementation can be found in Viti et al. (2004); Roberts et al. (2007); Holdship et al. (2017) as well as on the GitHub. More details will be provided in the individual Chapters as necessary.

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

1.4.2 Chemical Reaction Networks

Equation 1.13 represents the destruction and formation of every species in the network. It couples every species to the reactions that produce and deplete its abundance. This is effectively a means of representing the chemical reaction network. However, there is considerable uncertainty when it comes to the choice of network that is used in astrochemical modelling. While many chemical codes use one of the many available gas-phase networks, such as UMIST or KIDA, for the gas-phase modelling, there does not exist a standard grain-surface network. This is due to the fact that establishing the best networks for the formation of most species is still an active area of research.

1.5 Bayesian Inference

An important part of the chemical modelling procedure is the choice of physical and chemical parameters one includes. These range from the choice of final temperature or density at the end of UCLCHEM's Phase 2 to the binding energies of various grain-surface species in the network. In order to best model the final abundances of, say, a protostar, one must correctly choose these parameters. This is, however, easier said than done, as the observations that one wishes to calibrate against may have significant relative uncertainties associated with their values. This makes the choice of parameters difficult.

In order to estimate the best choice of parameters, $\boldsymbol{\theta}$, one can make use of the Bayesian school of thought which is based on Bayes' theorem. Bayes' theorem provides a rigorous means of updating the parameter probability distributions on the basis of new data, d. Given the data, the probability distributions of the binding energies of interest are given by:

$$P(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{d})}, \tag{1.14}$$

where $P(\boldsymbol{\theta}|\boldsymbol{d})$ is the posterior probability distribution, $P(\boldsymbol{\theta})$ is the prior, $P(\boldsymbol{d}|\boldsymbol{\theta})$ is the likelihood and $P(\boldsymbol{d})$ is referred to as the evidence. The prior distribution encodes the initial parameters' probability distribution functions. The likelihood tells us how likely the data is as a function of the parameter values. It is within the likelihood function that the generative physical model is encoded. The evidence is not just a normalising factor but also

represents the marginalised likelihood. The posterior distribution represents the updated probability distribution of reaction rates based on the data, the prior distribution, and the physical model. It is often the case that determining $P(\boldsymbol{\theta}|\boldsymbol{d})$ is unfeasible analytically, which is why one must resort to computational sampling algorithms. In this work, we made use of various implementations of sampling algorithms including Metropolis-Hastings, the affine invariant Markov Chain Monte Carlo as well as nested sampling.

1.6 Machine Learning

1.6.1 Introduction to Machine Learning

Machine learning is a field of computer science that focuses on algorithms that learn from data and that can then be used to make predictions. In this thesis, we only focus on the realm of supervised learning in which we have a pre-defined outcome of interest. This is in contrast to unsupervised learning in which there is no target variable to predict and which is also more exploratory in nature. We also do not consider reinforcement learning in which a computational agent learns to maximise a reward function by its actions in a specific environment (Molnar 2022).

The idea of supervised learning is that by learning from past data the algorithm can determine patterns to ensure it identifies the relationship between the co-variates or features of the data and the response variable of interest. A classic example of this is considering patient characteristics to predict the mortality risk associated with a specific disease. There exist two broad categories of supervised learning: classification and regression. The former is applied when the target variable of interest is categorical whereas the latter is used when it is continuous.

In a typical supervised machine learning task, we will split our data set into three subsets. The first is typically the training set on which the algorithm is meant to learn patterns from. This subset is typically the largest, comprising at least 70-80% of the total original data set. We do not report the performance of our algorithm on training data set, as we wish to ensure that it is generalisable to unseen data, that is data that it was not trained on. The second subset of data is the validation dataset on which we fine-tune what are known as the hyperparameters of our dataset. Hyperparameters refer to algorithm-specific quantities that control how the algorithm "learns" from the training data. By performing model comparison between models with different hyperparameters,

we can further improve the model's predictive power. Finally, the test dataset is the unseen dataset on which we evaluate the model.

1.6.2 Generating the data for machine learning algorithms

An important consideration when using machine learning algorithms is to ensure that we utilise a dataset that is representative of the problem space we are considering. For any problem with multiple dimensions in a continuous domain, it will not be possible to sample all possible combinations of parameter values. As such, there are a number of methods one could employ. The most obvious one is simply random sampling from the domain of interest. While intuitive, this has the disadvantage of points being independently sampled without any guarantee of good coverage of the domain. A superior method is the Latin Hypercube sampling scheme (McKay et al. 1979) in which each dimension is cut into a user-defined number of sub-sections and each sub-section is occupied by a single sample from the data set. By making sure that there is only one point per hyperplane axis, one ensures that the data sampled is near-random. We utilise the Latin Hypercube sampling scheme in this work through the implementation provided in the Surrogate Modelling Toolbox Python module (Bouhlel et al. 2019).

We now introduce two specific machine learning algorithms that we will use in this work.

1.6.3 Feedforward Neural Networks

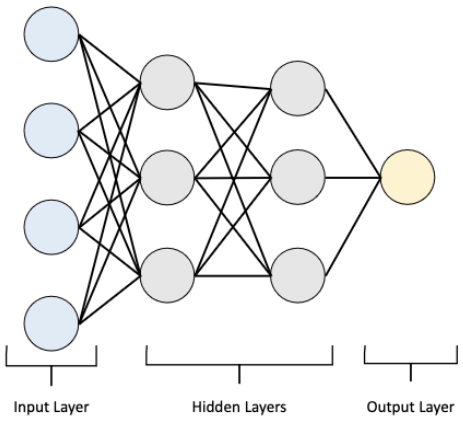

Figure 1.3: *Illustration of a deep feedforward network taken from de Mijolla et al. (2019).*

Neural networks are a class of machine learning algorithms that consist of multiple

layers of interconnected nodes that are chained together in order to approximate a function (Goodfellow et al. 2016). While the field of neural networks remains an active one, we will limit ourselves to considering deep feedforward networks which are also known as multilayer perceptrons. A depiction of a deep feedforward network is provided in Figure 1.3. The network is characterised by layers of nodes that are interconnected via edges. Each edge i in layer j is associated with a learnable parameter, w_{ij}, which is often termed the weight. The first layer takes a single value for each parameter in our input vector and the final layer provides the output, though note that the output can have multiple nodes if we choose to approximate a function with a multivariate range. The layers in between are referred to as the hidden layers and the number of layers as well as nodes per layer can be tuned to optimise performance. The value of every neuron a_j is first calculated by taking a weighted sum of the outputs of the previous layers, weighted by their w_{ij} to obtain:

$$z_j = \sum_i (w_{ij} a_i + b_j), \tag{1.15}$$

where b_j is a tunable parameter referred to as the bias.

Finally, to obtain the output of the neuron, we compute a_j by feeding it into a non-linear activation function: $a_j = g(z_j)$. The ability of the feedforward neural network to approximate a non-linear function stems from the presence of these activation functions (Goodfellow et al. 2016). The "feedforward" aspect of these neural networks stems from the fact that each layer takes computed values from the previous layer of neurons. The input layer is the first layer to provide outputs to the rest of the network.

We now briefly discuss how the parameters of the neural network are "learned". This typically takes the form of an unconstrained optimisation problem. The goal of "learning" is to find the set of parameters, θ, such that we minimise a cost function $J(\theta)$. In this case θ corresponds to the aforementioned weights of the neural network. This cost function typically measures the average loss on the training set. That is, we consider the discrepancy between the predicted value for a training example input x_i such that $\hat{Y}_i = f(x_i; \theta)$ and the true value Y_i, where f represents our parameterisation between the input and the output that is dependent on the parameters θ we are trying to optimise. Written more formally, our cost function takes the form:

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{d} L(f(x_i;\theta), Y_i), \tag{1.16}$$

where L represents the per-example loss function (Goodfellow et al. 2016). There are many choices of loss functions including the mean-squared error or the mean absolute error.

1.6.4 Decision Trees

Decision trees are a set of models that continuously divide the data based on cut-off values in the domain of each feature into subsets with each subset corresponding to a specific value of the target variable (Molnar 2022). Decision trees are simple to visualise and the fact that they rely on simple cuts makes them easy to interpret. However, they have a number of limitations which make them unsuitable for more complex tasks. In particular, decision trees are known to fit very complex trees to the training data that do not generalise well to the test data (Plaia et al. 2022). This is what is known as overfitting and effectively means that the algorithm is learning noise in the data that is simply not generalisable to the wider dataset. Recall that the purpose of supervised learning is to learn patterns from the training data in order to make good predictions on unseen data.

There are two main means of addressing this issue with decision trees: bagging and boosting. Bagging (booststrap aggregation) involves training individual decision trees on subsets of the data. This provides us with an ensemble of models over whose predictions we can average to obtain a more robust and stable prediction. The random forest algorithm (Breiman 2001) takes this one step further and only utilises a subset of the features for each decision tree. Boosting is a different approach. Instead of training decision trees in parallel, as we do in bagging, they are instead trained sequentially. A series of "weak learners", which are shallow decision trees, are consecutively trained on the data. At each stage, the errors made by the previous decision tree are assigned a larger weight so that the new decision tree is more likely to correctly predict (Freund and Schapire 1996). The XGBoost algorithm is an example of a boosted decision tree that is optimised for parallelisability and speed (Chen and Guestrin 2016). XGBoost computes the first and second derivatives of the cost function with respect to the predictions made by the boosted trees and then fits a new tree in order to amend the errors made by the previous ones.

1.6.5 Machine Learning Interpretability

While machine learning models have demonstrated considerable success in learning from data and making accurate predictions, there remains concern that we do not properly understand the inner workings of the algorithm in terms of why it makes a certain prediction. This has lead to these models being described as "black boxes". A lack of interpretability as to why the algorithm outputs a certain value is becoming increasingly problematic with machine learning applications in medicine and law (Rudin 2019).

While the stakes to human life are considerably lower in the case of astrochemistry, it is still important to have a good understanding of why a certain output is produced. As previously discussed, astrochemistry typically models molecular abundances by integrating coupled systems of ordinary differential equations. As such, the relationship between our chosen input parameters (which we will also call "features" throughout this thesis) and our output abundances is often non-linear and unintuitive. We wish to obtain a sense of three things from any machine learning interpretability analysis:

- which features are the most important in predicting our outputs
- a means of quantifying the relative feature importances and
- the exact nature of the relationship between the inputs and the outputs.

Within the field of machine learning interpretability, much research has been done to devise methods to determine the impact of a parameter in a machine learning model making a particular prediction, such as permutation feature importance or Local Surrogate Models. Molnar (2022) provides an overview of the different methods.

We elected to utilise Shapley values. These are taken from game theory and allow us to quantify the importance of the features of interest. Shapley value explanations are taken to be linear models (Molnar 2022). We are looking to explain the output of our function $f(x)$ which for the j^{th} data point in our test set x^j outputs $f(x^j)$. We define a feature explanation model, $\hat{g}$ in the following way:

$$f(x^j) = \hat{g}(x') = \phi_0 + \sum_{i=1}^{n} \phi_i x_i'^j, \tag{1.17}$$

where $\phi_0 = \mathbb{E}[f(x)]$ is the value of the average prediction, ϕ_i is the explained feature effect

of the ith feature, n is the number of features and $x_i'^j$ is an element of the "coalition vector", x', where $x' \in \{0,1\}^n$. The coalition vector states if a specific feature value is present (1) or absent (0).

We define the Shapley value of the ith feature, ϕ_i, as the marginal contribution of that feature in producing the model output. This is averaged over all possible coalitions, i.e. over all subsets of the set of features. This is calculated as follows for a data point:

$$\phi_i = \sum_{S \subseteq N} \frac{|S|!(n - |S| - 1)!}{n!} (g(\mathbf{x'}_i) - \hat{g}(\mathbf{x'}_{-i})), \tag{1.18}$$

where N is the set of features, $n = |N|$, S is the subset, $g(x_i')$ is the explanatory model evaluated when the feature is included and $g(\mathbf{x'}_{-i})$ the explanatory model evaluated when the feature is not included. When we say that a feature is not included it means we replace the value of the feature for that data point with a random number drawn from the feature's distribution of values.

1.7 This Thesis

The work in this thesis is focussed on using statistical and machine learning techniques to better understand astrochemistry. Chapter 2 is a chemical modelling exercise that considers the impact of adding two H_2-based addition reactions to a glycine network. This Chapter takes a more traditional approach to analysing astrochemistry, which is why it prefaces the rest of the thesis. Chapter 3 investigates how the network topology of a chemical network can be used to accelerate the Bayesian inference of the reaction rates. Chapter 4 considers a different approach to accelerating the Bayesian inference by making use of the grain-surface diffusion formalism to reduce the dimensionality of the inference problem. Chapter 5 involves the utilisation of the MOPED algorithm to make recommendations for which molecules should be prioritised for future observations to reduce the variance of the binding energy distributions. Chapter 6 considers the effect of varying physical parameters on the molecular abundances. This is connected to the use of astrochemical tracers using machine learning interpretability. Chapter 7 uses machine learning interpretability to probe the impact of various binding energies on molecular abundances. Finally, Chapter 8 summarises the results presented in this thesis and provides suggestions

for further study.

This page was intentionally left blank

Chapter 2

Investigating the impact of C and CH reacting with H_2

The work presented in this Chapter is based on the paper Heyl et al. (2023a), in collaboration with Thanja Lamberts, Serena Viti and Jonathan Holdship.

2.1 Introduction

As detailed in Chapter 1, interstellar dust plays a significant role in the rich chemistry that takes place in the interstellar medium. It is widely believed that complex-organic molecules (COMs) form on interstellar dust (Herbst and van Dishoeck 2009; Caselli and Ceccarelli 2012) since for certain molecules, grain-surface reactions are more efficient than gas-phase reactions. This is particularly important in cold astronomical environments where some gas-phase reactions may be highly inefficient, because a "third body" is needed to take up the excess heat of an exothermic reaction. Dust grains thus act as an energy sink allowing the chemistry to thrive and this can lead to the formation of more complex organic molecules.

Both experimental work and modelling has shown that one such molecule, namely the amino acid glycine can be formed through energetic processing of the ices during the warm-up phase of star formation (Bernstein et al. 2002; Woon 2002; Lee et al. 2009; Bossa et al. 2009; Ciesla and Sandford 2012; Garrod 2013; Sato et al. 2018), although there is evidence

to suggest that glycine would undergo destruction under increased irradiation (Pernet et al. 2013; Maté et al. 2015). In addition, in a joint experimental and modeling effort, Ioppolo et al. (2020) suggested that non-energetic mechanisms such as atom-addition reactions might be a promising route for glycine formation.

A new grain-surface reaction, inserting C atoms in H_2 to form CH_2 via $C + H_2 \longrightarrow CH_2$, was recently proposed to be barrierless by Simončič et al. (2020), based on earlier lab work by Krasnokutski et al. (2016). They included this reaction in their network and found a far more rapid conversion of C to CH_4. Subsequently, Lamberts et al. (2022) performed a combined experimental and computational work to investigate the importance of reactions with molecular hydrogen for the formation of methane. It was found that while the former reaction might not be fully barrierless, and the barrier likely depends on the binding site, the reaction $CH + H_2 \longrightarrow CH_3$ does in fact proceed without a barrier. The reason these 'dihydrogenation' reactions might be of interest is that they make H_2 more chemically active, the importance of which was recognized already by Hasegawa et al. (1992) and by Meisner et al. (2017) in the context of water formation. Typically, H_2 has one of the lowest binding energies of grain-surface species, lower than even atomic H (Al-Halabi and van Dishoeck 2007; Wakelam et al. 2017; Molpeceres and Kästner 2020), which allows the molecule to diffuse readily on the surface. Moreover, the molecular hydrogen abundance in molecular clouds and pre-stellar cores is much higher than that of atomic hydrogen (van Dishoeck and Black 1988; Goldsmith and Li 2005).

By including these reactions in chemical models, one might first of all expect changes in the CH_4 abundance, but it is equally interesting to consider the effect on downstream species such as complex organic molecules, whose typical abundances are far lower. At low temperatures, one would not expect methane to desorb in large quantities. Instead, it is likely to remain chemically active on the grains. Their sensitivity to new reactions should be considered, as their more abundant precursors might see changes in their abundances.

In this Chapter, we look to build on the work by Simončič et al. (2020) and Lamberts et al. (2022) to investigate the impact of the dihydrogenation reactions of C and CH on our gas-grain chemical network. In particular, we are interested in observing the effect these reactions have on the production of glycine and its precursors. Our glycine network is based on the kinetic Monte Carlo network used in Ioppolo et al. (2020), using in part updated rate constants from recent literature, as indicated in Table 2.1.

We start by describing the astrochemical model, our choice of parameters and how we

Reaction No.	Reaction	Reference
1	$CO + OH \longrightarrow HOCO$	Arasa et al. (2013)
2	$HOCO + H \longrightarrow H_2 + CO_2$	Goumans et al. (2008)
3	$HOCO + H \longrightarrow HCOOH$	Goumans et al. (2008); Ioppolo et al. (2011b)
4	$CH_4 + OH \longrightarrow CH_3 + H_2O$	Lamberts et al. (2017)
5	$NH_2 + CH_3 \longrightarrow NH_2CH_3$	Ioppolo et al. (2020)
6	$NH_3 + CH \longrightarrow NH_2CH_2$	Balucani et al. (2009)
7	$NH_2CH_2 + H \longrightarrow NH_2CH_3$	Ioppolo et al. (2020)
8	$NH_2CH_3 + H \longrightarrow NH_2CH_2 + H_2$	Oba et al. (2014)
9	$NH_2CH_3 + OH \longrightarrow NH_2CH_2 + H_2O$	Ioppolo et al. (2020)
10	$NH_2CH_2 + HOCO \longrightarrow NH_2CH_2COOH$	Woon (2002)
11	$H_2 + OH \longrightarrow H_2O + H$	Meisner et al. (2017)
12	$O_2 + H \longrightarrow HO_2$	Lamberts et al. (2013)
13	$HO_2 + H \longrightarrow OH + OH$	Lamberts et al. (2013)
14	$HO_2 + H \longrightarrow H_2 + O_2$	Lamberts et al. (2013)
15	$HO_2 + H \longrightarrow H_2O + O$	Lamberts et al. (2013)
16	$OH + OH \longrightarrow H_2O_2$	Lamberts et al. (2013)
17	$OH + OH \longrightarrow H_2O + O$	Lamberts et al. (2013)
18	$H_2O_2 + H \longrightarrow H_2O + OH$	Lamberts and Kästner (2017)
19	$N + O \longrightarrow NO$	Ioppolo et al. (2020)
20	$NO + H \longrightarrow HNO$	Fedoseev et al. (2012)
21	$HNO + H \longrightarrow H_2NO$	Fedoseev et al. (2012)
22	$HNO + H \longrightarrow NO + H_2$	Fedoseev et al. (2012)
23	$HNO + O \longrightarrow NO + OH$	Ioppolo et al. (2020)
24	$HN + O \longrightarrow HNO$	Ioppolo et al. (2020)
25	$N + NH \longrightarrow N_2$	Ioppolo et al. (2020)
26	$NH + NH \longrightarrow N_2 + H_2$	Ioppolo et al. (2020)
27	$C + O \longrightarrow CO$	Ioppolo et al. (2020)
28	$CH_3 + OH \longrightarrow CH_3OH$	Qasim et al. (2018)
29	$C + H_2 \longrightarrow CH_2$	Simončič et al. (2020); Lamberts et al. (2022)
30	$CH + H_2 \longrightarrow CH_3$	Lamberts et al. (2022)

Table 2.1: *Table of the reactions added to the standard UCLCHEM network.*

evaluate the network sensitivity in Section 2.2. We then discuss the results as well as the astrochemical implications in Section 2.3 and summarize our conclusions in 2.4.

2.2 Methodology

2.2.1 The Astrochemical Model

In this Chapter, the gas-grain chemical code UCLCHEM was used (Holdship et al. 2017)[1]. UCLCHEM makes use of a rate equation approach to modelling the gas and grain-surface

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

Parameter	Values
Final Density of Phase 1/Initial Density of Phase 2	10^5 cm^{-3}, 10^6 cm^{-3}, 10^7 cm^{-3}
Efficiency for barrierless $C + H_2 \longrightarrow CH_2$	0, 0.05, 1
Cosmic Ray Ionisation Rate	$\zeta, 10\zeta$

Table 2.2: *The parameters that were varied in this Chapter to assess the effect of the two reactions. Note that the density of Phase 1 is the same as the initial density of Phase 2. An efficiency of 0 is equivalent to reaction being excluded. ζ is the standard cosmic ray ionisation rate of* 1.3×10^{-17} s^{-1}

and bulk abundances. The gas-phase reaction network is taken from the UMIST database (McElroy et al. 2013). The grain-surface network used was the default one as available on GitHub.

Various reaction mechanisms are implemented in UCLCHEM. The grain-surface reaction mechanisms that exist in UCLCHEM include the Eley-Rideal mechanism as well as the Langmuir-Hinshelwood diffusion mechanism, which were implemented in Quénard et al. (2018), as was the competition formula from Chang et al. (2007) and Garrod and Pauly (2011). The binding energies that are used to calculate the diffusion reaction rate are taken from Wakelam et al. (2017). We also included an updated version of the glycine grain-surface network from Ioppolo et al. (2020), also including both the reactions $C + H_2 \longrightarrow CH_2$ and $CH + H_2 \longrightarrow CH_3$ as summarized in Table 2.1. Note that the reaction $OH + H_2 \longrightarrow H_2O + H$ had been already included, based on previous work by, e.g., Meisner et al. (2017). The code also includes thermal and non-thermal desorption, such as due to H_2 formation, cosmic ray ionisation as well as UV-induced desorption. Note that the astrochemical model used in Ioppolo et al. (2020) makes use of the non-diffusive grain-surface chemistry that is described in Garrod and Pauly (2011) and Jin and Garrod (2020). This is not used in UCLCHEM. The implications of this will be discussed later in this Chapter.

UCLCHEM is used to model two distinct phases of the star formation process. Phase 1 is the free-fall collapse phase of a dark cloud for a default value of 5 million years, whereas Phase 2 models the warm-up phase immediately following Phase 1, with the initial density of Phase 2 equal to the final density of Phase 1. Phase 2 runs for 1 million years. Further details of the code can be found in Holdship et al. (2017).

2.2.2 Parameter Selection

To assess the importance of the two proposed reactions to the network under various interstellar conditions, three parameters were varied, as listed in Table 2.2. The standard cosmic ray ionisation rate in UCLCHEM is $\zeta = 1.3 \times 10^{-17}$ s^{-1}. This is in line with typical values that are of the order 10^{-17} s^{-1} in diffuse ISM conditions (O'Donnell and Watson 1974; Black et al. 1978; Hartquist et al. 1978; Indriolo and McCall 2013). However, there exist observations of higher cosmic ray ionisation rates (Indriolo et al. 2007; Indriolo and McCall 2012), which is why we also include analysis of a region with cosmic ray ionisation rate of 10ζ. Cosmic ray ionisation is typically expected to break larger molecules into smaller radicals. We did not consider lower values of the cosmic ray ionisation rate, as these are typically not observed. The cosmic ray dependency on column density in O'Donoghue et al. (2022) covered a range of values that were, however, already covered by the factor of 10 we consider here. While they found differences for lower densities during the collapse phase, these were ironed out once the collapse reached larger final densities, which is why here we do not include this dependency on column density.

Three different astronomical regions were modelled:

1. a dark cloud with a final density of 10^5 cm^{-3}
2. a low-mass protostar with a final density of 10^6 cm^{-3}
3. a high-mass protostar, with a final density of 10^7 cm^{-3}

The heating profiles during Phase 2 for the last two cases are based on Viti et al. (2004) and differ for each astronomical object. The dark cloud simulation was only run for Phase 1, but was allowed to run for a further million years to allow the chemistry to settle.

Another parameter that was varied was the efficiency, α, of the extent to which the reaction $C + H_2 \longrightarrow CH_2$ is barrierless. While Simončič et al. (2020) considered the reaction to be fully barrierless, Lamberts et al. (2022) found that the reaction barrier likely depends on the binding site. As such, our grid of models considers efficiencies for the reaction of 0 (the reaction is not included), 0.05 (5% of binding sites lead to a barrierless reaction and 95% of the binding sites have an infinitely high barrier) and 1 (the reaction is fully barrierless). What this means practically is that the reaction rate is multiplied by the efficiency. The reaction $CH + H_2 \longrightarrow CH_3$ was included as only being barrierless, based on Lamberts et al. (2022).

2.2.3 Evaluating the network sensitivity

We quantify the effect of the new reactions on the model by considering the change in abundances of the species that are the most affected when taking the ratio of the abundances of the modified and original models. The modified model is the chemical network which has $\alpha = 1$, whereas the original model was taken to be the network which had neither of the dihydrogenation reactions. These two scenarios were taken to be the extremes of the parameter range in terms of including these reactions. The ratio is most sensitive to strong deviations in the molecular abundances as a result of the dihydrogenation reactions.

This ratio is defined for each species i as:

$$\delta_i(t) = \frac{x_i^M(t)}{x_i^O(t)}, \tag{2.1}$$

where $x_i^M(t)$ is the abundance of species i in the modified model at time t and $x_i^O(t)$ is the abundance of the same species in the original model at time t.

We only considered species which had a value above a "threshold of detectability". This was to ensure that we did not look at species whose original and changed abundances were below what can be observed from an astronomical point-of-view. For grain-surface species this threshold was set to 10^{-8} with respect to hydrogen whereas for gas-phase species this threshold equalled 10^{-12} with respect to hydrogen. We took 10^{-8} as a lower-limit threshold for grain-surface species, as this was the order of magnitude of the lowest reported abundances in Boogert et al. (2015). Similarly, the gas-phase threshold was taken based on the abundances of COMs typically observed in the gas-phase, such as in Jiménez-Serra et al. (2016, 2021).

We can also define a quantity that tracks the absolute change in the abundance of species:

$$\Delta_i(t) = x_i^M(t) - x_i^O(t) = x_i^O(t)[\delta_i(t) - 1], \tag{2.2}$$

This value indicates how species with relatively large abundances, such as elemental species or their hydrogenation products, are re-distributed.

Dark Cloud			**Low-Mass Star**			**High-Mass Star**		
Species	δ	$x_i^O(t_{final})$	Species	δ	$x_i^O(t_{final})$	Species	δ	$x_i^O(t_{final})$
#CH_2	2.8	4.1×10^{-7}	#CH_2	2.8	4.1×10^{-7}	#CH_2	2.8	4.1×10^{-7}
#CH_3	2.3	2.6×10^{-7}	#CH_3	2.3	2.6×10^{-7}	#CH_3	2.3	2.6×10^{-7}
#CH_4	1.3	4.0×10^{-6}	#CH_4	1.3	3.8×10^{-6}	#CH_4	1.3	3.8×10^{-6}
#NH_3	1.1	3.8×10^{-6}	#NH_3	1.1	3.7×10^{-6}	#NH_3	1.1	3.7×10^{-6}
#H_2CS	1.1	2.4×10^{-8}	#H_2CS	1.1	2.4×10^{-8}	#H_2CS	1.1	2.4×10^{-8}
#CH_3OH	1.04	1.5×10^{-5}	#CH_3OH	1.04	1.3×10^{-5}	#CH_3OH	1.04	1.3×10^{-5}
#HNC	1.03	2.3×10^{-8}	#HNC	1.04	2.3×10^{-8}	#HNC	1.04	2.3×10^{-8}
#H_2SiO	1.03	3.3×10^{-7}	#H_2SiO	1.03	1.1×10^{-8}	#H_2SiO	1.03	3.4×10^{-7}
#HCN	1.02	1.7×10^{-7}	#HO_2	1.03	2.3×10^{-7}	#HO2	1.03	2.3×10^{-7}
#O_2	1.02	1.8×10^{-6}	NO	1.03	1.0×10^{-10}	#HCN	1.02	1.6×10^{-7}
#CH	1.1×10^{-15}	7.2×10^{-7}	#CH	2.0×10^{-15}	7.2×10^{-7}	#CH	2.1×10^{-15}	7.2×10^{-7}
#C	2.4×10^{-13}	1.4×10^{-6}	#C	2.5×10^{-13}	1.4×10^{-6}	#C	2.5×10^{-13}	1.4×10^{-6}
#NCH_4	3.7×10^{-13}	1.5×10^{-7}	#NCH_4	3.4×10^{-13}	1.5×10^{-7}	#NCH_4	3.4×10^{-13}	1.5×10^{-7}
#NH_2CH_3	8.0×10^{-13}	1.9×10^{-7}	#NH_2CH_3	8.3×10^{-13}	2.0×10^{-7}	#NH_2CH_3	8.3×10^{-13}	2.0×10^{-7}
NH_2CH_3	1.5×10^{-12}	8.7×10^{-10}	#Si	0.98	5.6×10^{-8}	#Si	0.98	5.6×10^{-8}
CH	0.96	9.3×10^{-10}	#SiH	0.99	2.5×10^{-8}	#SiH	0.99	2.5×10^{-8}
CH_3	0.98	1.5×10^{-9}	#SiH_2	0.99	1.3×10^{-8}	#SiH_2	0.99	1.3×10^{-8}
#Si	0.98	5.7×10^{-8}	#O	0.99	7.8×10^{-5}	#Si	0.99	6.7×10^{-5}
#SiH	0.99	2.6×10^{-8}	#H_3CO	0.99	1.7×10^{-6}	#H_3CO	0.99	1.7×10^{-6}
#SiH_2	0.99	1.4×10^{-8}	#HNO	0.99	1.2×10^{-5}	#HNO	0.99	1.2×10^{-5}

Table 2.3: *Summary of the species that experienced the greatest increases (top section) and decreases (bottom section) for each of the three astronomical objects in Phase 1. Species with a "#" are grain-surface species. All other species are gas-phase.*

2.3 Results and Astrochemical Implications

We find that even though the amounts by which various species are affected differs for each stage of star formation, the general trends are broadly similar. As such, we group our analysis per phase. Tables 2.3 and 2.4 summarise the changes in terms of δ. The effect of the enhanced cosmic ray ionisation rate is discussed in Section 2.3.1.

Our results differ from Ioppolo et al. (2020) in that, while glycine does form on the grains, it does not do so in Phase 1, as UCLCHEM does not utilise non-diffusive grain-surface mechanisms. Instead, glycine forms on the grains as the temperature increases in Phase 2.

2.3.1 Impact of the Parameters

In this sub-section we consider the role that the physical and chemical parameters at play. Tables 2.3 and 2.4 show the changes in abundance when we compare the original network without the dihydrogenation reactions with the $\alpha = 1$ case. Figures 2.1 and 2.2 show the

Low-Mass Star			**High-Mass Star**		
Species	δ	Original Abundances	Species	δ	Original Abundances
HOCO	3.7	9.3×10^{-10}	HOCO	2.1	4.3×10^{-8}
H_2O_2	2.6	4.3×10^{-9}	CH_3OH	2.0	1.8×10^{-9}
CH_3CHO	2.2	1.0×10^{-7}	CH_3CHO	2.0	1.5×10^{-7}
CH_3OH	2.1	3.7×10^{-9}	C_2H_4	2.0	2.5×10^{-9}
CH_3CN	1.7	1.0×10^{-9}	CH_2CO	1.9	1.8×10^{-10}
C_4H	1.6	3.2×10^{-10}	H_2CO	1.7	9.3×10^{-9}
C_3H_2	1.5	5.6×10^{-9}	CH_3	1.7	1.1×10^{-10}
CH_3CCH	1.5	2.4×10^{-8}	NH_3	1.6	1.3×10^{-8}
NH_3	1.5	2.7×10^{-7}	CH_3CN	1.5	7.1×10^{-10}
NH_2CHO	1.4	2.7×10^{-7}	C_2H_2	1.5	1.1×10^{-8}
NH_2CH_2	3.8×10^{-5}	9.2×10^{-7}	NH_2CH_2	4.7×10^{-5}	8.3×10^{-7}
NH_2CH_3	2.4×10^{-3}	1.6×10^{-7}	NH_2CH_3	2.5×10^{-3}	1.7×10^{-7}
NH_2CH_2COOH	6.0×10^{-2}	6.3×10^{-9}	NH_2CH_2COOH	6.3×10^{-3}	7.2×10^{-8}
H_2S	0.88	2.0×10^{-9}	NO	0.82	4.0×10^{-6}
SO_2	0.92	4.4×10^{-8}	NCCN	0.96	3.9×10^{-7}
Mg^+	0.93	8.0×10^{-8}	O_2	0.96	7.1×10^{-6}
O	0.95	1.3×10^{-5}	HCOO	0.96	1.9×10^{-10}
CH_2OH	0.95	6.4×10^{-8}	C_2N	0.97	3.5×10^{-8}
O_2	0.96	4.2×10^{-5}	O	0.97	3.6×10^{-8}
SO	0.97	1.9×10^{-6}	CO_2	0.97	7.6×10^{-6}

Table 2.4: *Summary of the species that experienced the greatest increases (top section) and decreases (bottom section) for each of the three astronomical objects in Phase 2. All species listed are gas-phase.*

time series of the abundances for glycine and its precursors.

Final Density

The final density of the collapsing cloud had a minor effect on the final abundances of the species in Phase 1. For all three astronomical objects modelled in Phase 1, we observe a significant decrease of grain-surface CH and C when the reactions are included and see an enhancement of grain-surface CH_2, CH_3 and CH_4. However, the values of δ as well as their original abundances seem to be independent of the density, suggesting a saturation effect.

In Phase 2, we observe that the final density of the collapsing cloud does affect the extent to which the added reactions influence the final abundances. We notice that several hydrogenation-based species have greater abundances at lower densities, including species such as HOCO, H_2O_2, CH_3CCH and H_2CO.

Efficiency

For more abundant species, such as H_2O and CH_3OH, we find that the results obtained from using a branching fraction of 0.05 for the barrierless dihydrogenation of C are essentially the same as using a efficiency of 1 (the reaction is fully barrierless).

We do find that the efficiency parameter plays a role in the final abundances of glycine and its precursors during the warm-up phase of low and high-mass stars. This can be seen in Figures 2.1 and 2.2. For Phase 1, the species are not detectable except for the original configuration. However, we still observe that for the other three configurations an increasing value of α corresponds to an increased level of depletion. In Phase 2, the configurations are all detectable and this same hierarchy remains in the gas-phase.

Cosmic Ray Ionisation Rate

The degree of cosmic ray ionisation is found to play an important role in enhancing or counteracting the role of the dihydrogenation reactions. The cosmic ray destruction routes we include in our standard network are from Garrod et al. (2008). These consist of hydrogen abstraction reactions and reactions that produce radical-radical pairs of products. An enhanced cosmic ray ionisation rate leads to the destruction of many of the hydrogenated species, such as CH_4, NH_3, H_2O and CH_3OH, as well as their precursors. This leads to further hydrogen reservoirs being released and radicals being formed which can go on to form glycine and its precursors. Because no cosmic ray destruction mechanisms for these complex, larger, species are included, we find that these are more abundantly produced.

This is important to consider in the context of glycine. In Figures 2.1 and 2.2, we plot the time dependence of the abundance of glycine precursors for eight different parameter sets, including the enhanced cosmic ray ionisation rate. In Phase 1, we find that on the grains, the enhanced cosmic ray ionisation rate depletes the species. In Phase 2, the effect varies by configuration and species. The original configuration consistently leads to a decrease of all plotted species in the presence of enhanced cosmic ray ionisation. The $\alpha = 0$ configuration is depleted for the methylamine radical and glycine, but enhanced for methylamine. The $\alpha = 0.05$ and $\alpha = 1$ configurations are depleted for methylamine and glycine, but enhanced for the methylamine radical.

2.3.2 General Implications

As can be seen in Tables 2.3 and 2.4, the inclusion of reactions with molecular hydrogen affects the hydrogen economy of the reaction network. Previously, the reaction network had a significant amount of H_2 being adsorbed or produced on the surface with no chemical destruction mechanisms. The H_2 molecules are a previously untapped hydrogen reservoir that is now being utilised (Hasegawa et al. 1992). Because one H_2 frees up two H atoms on the surface, other atomic hydrogenation reactions can take place more easily. Therefore, we observe the increase in the abundances of species in Phases 1 and 2 that are the products of hydrogenation. While for many of the more common species, the relative increase, i.e., δ is small, the abundance increases in absolute terms. There are large relative and absolute changes in the network of less abundant species, such as NH_2CH_2, NH_2CH_3 and NH_2CH_2COOH and there are fairly large absolute changes in the network of highly abundant species, such as C and its hydrogenation products.

We can also comment on the carbon budget. The previously defined Δ parameter allows us to consider how carbon is redistributed as a result of the new reactions being included. For instance, for the dark cloud during Phase 1, the total Δ for the main carbon-based grain-surface species that increase

$$\Delta_{\text{total}}(\#CH_2 + \#CH_3 + \#CH_4 + \#H_2CS + \#CH_3OH) = 2.9 \times 10^{-6} \, .$$

is nearly equal to that of the total decrease Δ of main grain-surface species:

$$\Delta_{\text{total}}(\#C + \#CH + \#NCH_4 + \#NH_2CH_3) = 2.5 \times 10^{-6} \, .$$

From this we can see that the dihydrogenation reactions redistribute the carbon between the aforementioned species. The remaining carbon is redistributed to other species in the network in smaller amounts. We also observe that besides the methyl radical, also species that contain the CH_3 group, such as CH_3OH and CH_3CN see increases in their abundances, via the reactions $CH_3 + OH \longrightarrow CH_3OH$ and $CH_3 + CN \longrightarrow CH_3CN$.

In a similar fashion, nitrogen is redistributed throughout the network. The grain-surface ammonia abundance increases by 10%, i.e., $3.8{\times}10^{-7}$. The decrease in $\#NCH_4$ and $\#NH_2CH_3$ accounts for $3.4{\times}10^{-7}$ or $\sim 90\%$.

2.3.3 Implications for Simple Grain-Surface Species

In the light of the recent ice observations with the James Webb Space Telescope, both published (Yang et al. 2022) and upcoming missions (such as the one detailed in (McClure et al. 2017)), it is important to consider the effect on the main ice constituents. Figure 2.3 shows the time-evolution of the abundances of grain-surface H_2O, CO, CO_2, CH_3OH, H_2CO, NH_3 and CH_4 in Phase 1 of a dark cloud. These are species that have been securely or likely identified in the ices (Boogert et al. 2015). The shaded areas in the plots indicate the 68% confidence interval for the measured abundances, taken from Boogert et al. (2015). In Boogert et al. (2015), the abundances were given in terms of the median value as well as the upper and lower quartiles. It was assumed that the spread in the measurements was Gaussian, which meant that the interquartile range represented 1.36σ. This spread in measurements is due to both observational error and source-to-source variation. We observe that we recover the measured abundances for most of the species within the uncertainty, with the exception of grain-surface CO_2. The inclusion of the dihydrogenation reactions does not change how well the models agree with the abundance measurements, however, for all hydrogenation products we observe that the inclusion of reactions with molecular hydrogen increases their abundance, as a result of the additional atomic hydrogen on the surface. In short, despite uncertainties surrounding activation energies, networks and binding energies, we are able to recover observational abundances reasonably well when we include the reactions with molecular hydrogen and this gives us confidence that the predictions we make for glycine and its precursors are accurate.

2.3.4 Implications for Glycine and its Precursors

In Tables 2.3 and 2.4, we observe that the abundances of glycine and its precursors decreases if molecular hydrogen is part of the reaction network. We can also explain why the abundance of precursors of glycine, gas and grain NH_2CH_3 and NH_2CH_2 decrease. The former is formed through the reaction $NH_2 + CH_3$, but since more atomic H is present on the grains, both radical species are preferentially hydrogenated. The inclusion of H_2 as a reacting species, not just in the context of the two reactions we consider in this Chapter, introduces greater competition for radicals that are needed for the formation of complex organic molecules. This results in the lower abundances of NH_2CH_3 and NH_2CH_2.

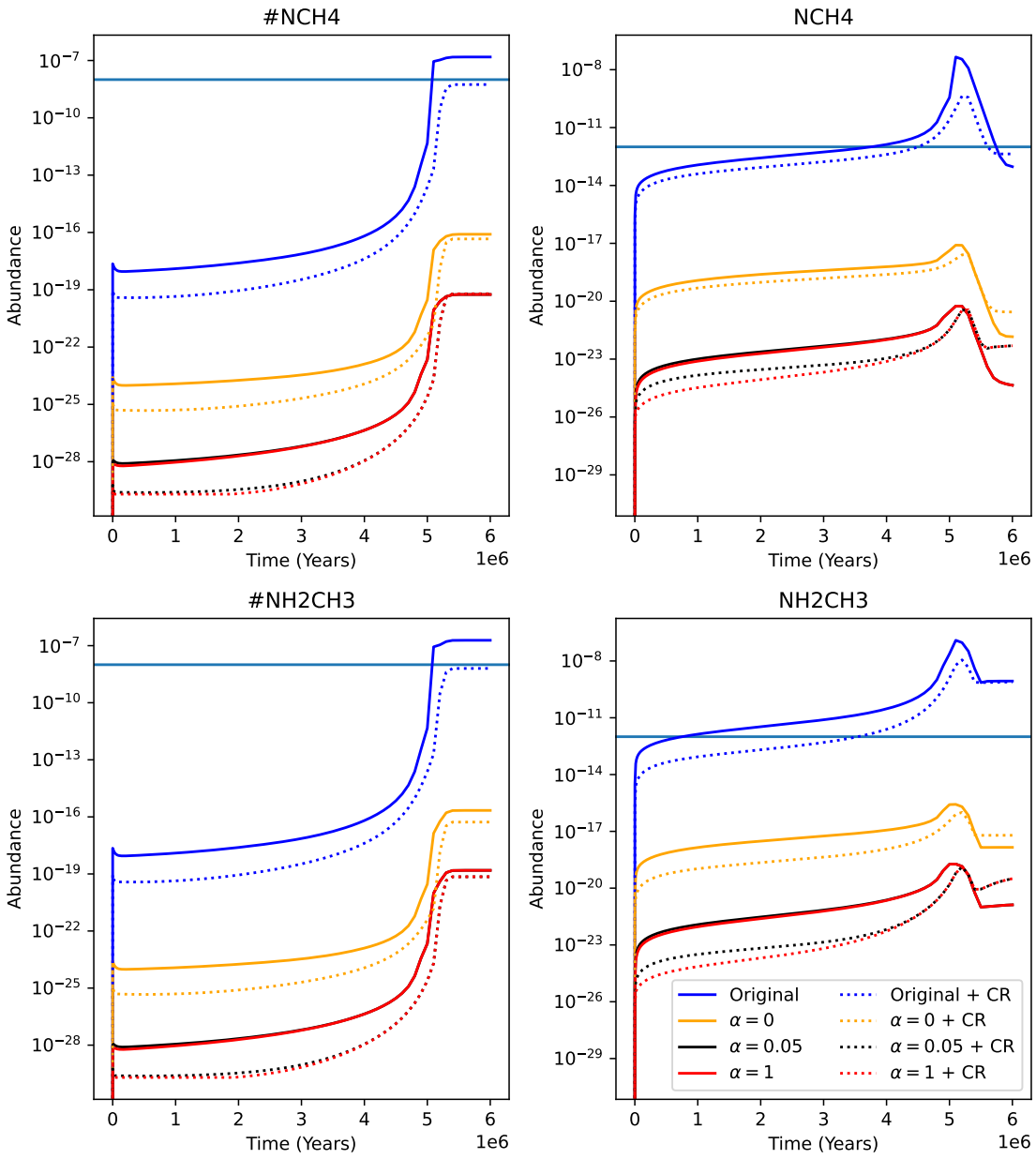

Figure 2.1: *Time series of the abundances of grain-surface and gas-phase NH_2CH_2 and NH_2CH_3 in Phase 1 of a dark cloud. Furthermore, we observe that the inclusion of the dihydrogenation reactions, regardless of efficiency α severely depletes the abundances of the glycine precursors in both phases relative to the original model which did not include either of the dihydrogenation reactions. Also plotted are the limits of detectability we have used for gas and grain-surface species. We do not plot glycine, as it is not formed at all in Phase 1. We observe that only the original model is capable of producing 'detectable' levels of methylamine and the methylamine radical. For the other configurations, an increase in α results in increased depletion of the species relative to the original model. We also observe that enhanced cosmic ray ionisation depletes the abundances on the grains but not in the gas.*

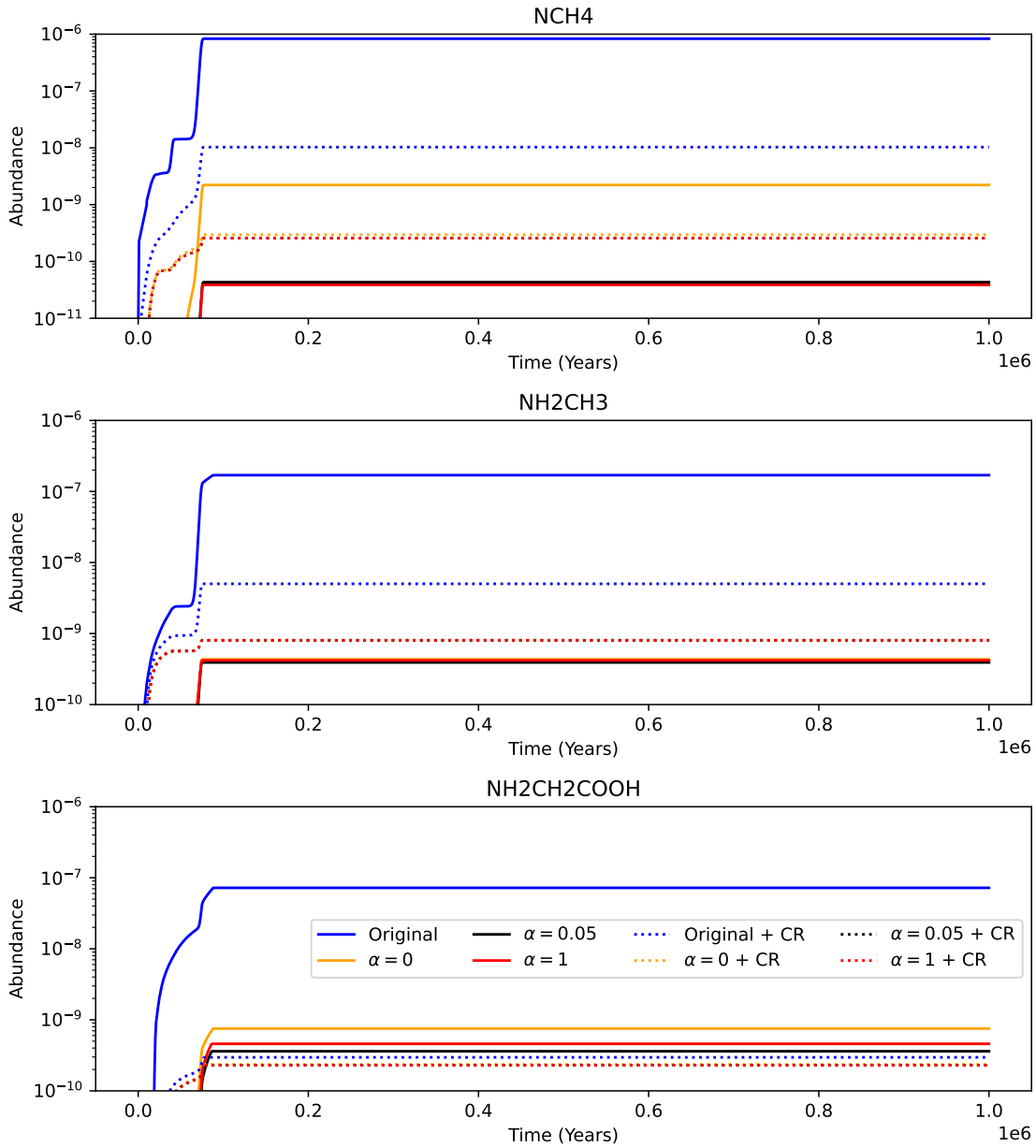

Figure 2.2: *Time series of the abundances of gas-phase NH_2CH_2, NH_2CH_3 and NH_2CH_2COOH in Phase 2 of a high-mass star. We observe that glycine is produced in the warm-up phase. The enhanced cosmic ray ionisation rate is found to significantly deplete all three species in the gas-phase for the original model. For NH_2CH_2 and NH_2CH_3, when $\alpha = 0$, $\alpha = 0.05$ or $\alpha = 1$, the enhanced cosmic ray ionisation rate results in an increase of their abundances. For glycine, the enhanced cosmic ray ionisation rate seems to decrease its gas-phase abundance.*

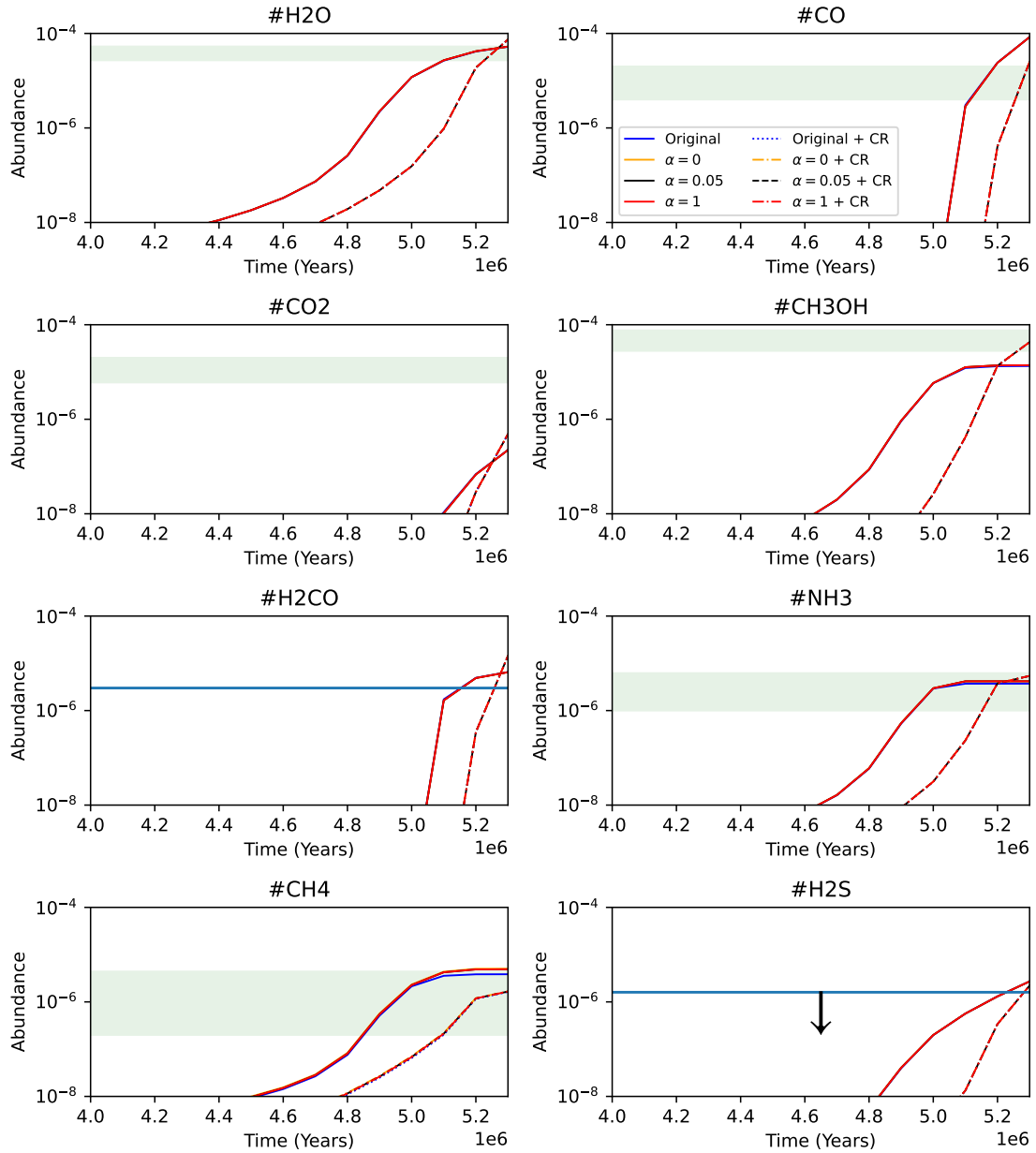

Figure 2.3: *Time series of the abundances of grain-surface H_2O, CO, CO_2, CH_3OH, H_2CO, NH_3, CH_4 and H_2S in Phase 1 of a dark cloud. We include the species that have securely identified or likely identified. The abundances were adapted from Boogert et al. (2015). The shaded areas include the 1σ region of abundances. In the case of H_2CO, no uncertainty was provided in the original source, so there is no shaded area. Grain-surface H_2S only has an upper limit on its abundance. For both normal and enhanced cosmic ray ionisation rates, the time-series differ very little, which is why it is difficult to distinguish them visually.*

We can also use this to justify the impact of the efficiency. Figures 2.1 and 2.2 plot the time series for the various efficiencies as well as with enhanced cosmic ray ionisation in Phase 1 and 2, respectively. We previously remarked that the original configuration produced the most of glycine and its precursors. For the other configurations, the greater the value of α, the greater the depletion of these species. This makes sense when one considers that an increasing value of α results in more H_2 being consumed and therefore more atomic H becoming available to hydrogenate precursors.

We now look to compare our results with observations. We do this separately for glycine and its precursors. We also discuss the implications of not using non-diffusive grain-surface mechanisms in our code, such as the ones discussed in Garrod and Pauly (2011) and Jin and Garrod (2020).

Methylamine and the methylamine radical

Methylamine (NH_2CH_3) and the methylamine radical (NH_2CH_2) are important precursors of glycine. The hydrogen abstraction of methylamine to form the methylamine radical is crucial, as there is growing evidence to suggest that the reaction NH_2CH_2 + HOCO $\longrightarrow$ NH_2CH_2COOH is a feasible glycine formation route (Ramesh and Yuan-Pern 2022). Confirmed detections of methylamine in high-mass star forming regions are summarised in Table 2.5. We observe improved level of agreement between our model outputs and observations when the reactions are included with $\alpha = 1$. We observed significant enhancement when the cosmic ray ionisation rate was increased. This suggests that if dihydrogen is chemically active on the grains, one would need to consider regions of high cosmic ray ionisation rate to detect these precursors of glycine, as these reactions reduce the abundance of methylamine. In the case of the Bøgelund et al. (2019) observation, we have confidence in the value of our ratio, as the chemical network for methanol is well-established.

However, the entirety of the above discussion regarding the agreement of our results with observations is incomplete without discussing the effect of the nondiffusive reaction mechanisms being absent in our modelling. These mechanisms are of particular use when considering reactions between reactants which are likely to react very slowly via the Langmuir-Hinshelwood diffusion mechanism, such as the reaction between CO and OH to form CO_2. Methylamine and the methylamine radical are formed via reactions 6 and 7, which involve species with high binding energies, thereby making their formation at

Reference Molecule	Reference	Abundance Measurements	Original Ratio	New Ratio
CH_3OH	Bøgelund et al. (2019)	$8 \times 10^{-3} - 0.1$	37	0.02
H_2	Ohishi et al. (2019)	$1.5 \pm 1.1 \times 10^{-8}$	3.5×10^{-7}	3.9×10^{-10}

Table 2.5: *Table of methylamine abundance measurements relative to reference molecules for high-mass stars. Also included are the corresponding ratios obtained in this Chapter for high-mass stars with the standard cosmic ray ionisation rate for both the original model and the new model.*

10K inefficient via diffusion. As a result, the fact that we do not include the non-diffusive mechanisms means that methylamine and its radical are under-produced.

Glycine

While there may be no confirmed detection for glycine in the literature, various estimates exist. In Gibb et al. (2004), an upper limit of 0.3% with respect to water was determined, whereas in Jiménez-Serra et al. (2014), this was estimated to be around 0.1%. In this Chapter, we find that when the dihydrogenation reactions are not included this value is 0.07% and when we include both reactions then it is 2×10^{-4}%. We should note that in the absence of experimentally-motivated gas-phase glycine destruction reactions the values derived in this Chapter are only upper limits, if one neglects non-diffusive mechanisms. In the previous sub-section, we discussed that methylamine and its radical are underproduced. This will result in glycine being underproduced as well, not just due to the underproduction of its precursors, but also because reaction 10 is less efficient if assumed to be diffusion-only.

2.4 Conclusion

In this Chapter, we considered the effect of including the reactions of H_2 with C and CH in our grain-surface network. We ran a grid of 12 models that vary the final density of the collapsing cloud, the efficiency for the 'barrier' of $C + H_2 \longrightarrow CH_2$ as well as the cosmic ray ionisation rate.

Making molecular hydrogen chemically active unlocks a previously untapped reservoir of hydrogen, and therefore freeing up the use of atomic hydrogen for hydrogenation reactions. A particularly interesting consequence of this is that making H_2 more chemically active decreased the abundances of glycine and its precursors. This may aid in explaining glycine has remained undetected so far.

We note that we do not have a comprehensive gas-phase network for glycine and its precursors. That is likely to be a limitation. While it is still likely that glycine and its precursors form on the grains and then evaporate into the gas-phase, it is possible that there would be gas-phase destruction routes as well. Additionally, cosmic-ray ionisation destruction routes on the grains and in the gas-phase are likely also needed, as these typically break large molecules down into smaller radicals which are then recycled for further gas-phase reactions. As such, the abundances we obtain for glycine and its precursors are likely to only be upper limits.

An additional limitation is the absence of the non-diffusive reaction mechanisms discussed in Garrod and Pauly (2011) and Jin and Garrod (2020). The consequence is that glycine and its precursors do not form efficiently on the grains at 10 K, which is different to what was found in Ioppolo et al. (2020). As such, they are under-produced in our models, whereas diffusion-efficient reactions overproduce certain species. However, without implementing this formalism in the code, it is difficult to assess the relative impacts of these mechanisms on the final abundances.

This page was intentionally left blank

Chapter 3

Exploiting Network Topology for Inference of Surface Reaction Networks

The work presented in this Chapter is based on the paper (Heyl et al. 2020), in collaboration with Serena Viti, Jonathan Holdship and Stephen M. Feeney.

3.1 Introduction

Bayesian inference has become a standard tool in astrophysics for determining model parameters from observations. In recent years, it has also become a tool in astrochemistry (Holdship et al. 2018; Makrymallis and Viti 2014; de Mijolla et al. 2019). However, by considering increasingly rich chemistry one must ultimately consider more complicated reaction networks. This results in an increased computational cost. There exists much in the literature regarding chemical network reduction as a means of reducing the computational complexity of the problem being solved (Ayilaran et al. 2019). An understanding of chemical reaction networks and how to simplify them has become increasingly crucial in astrochemistry (Xu et al. 2019; Grassi et al. 2012). However, these methods have primarily focused on simplifying the network for the forward problem. For example, Xu et al. (2019) adopted an iterative approach, where they evaluated the importance of each species at each

timestep. The advantage of Bayesian inference is that it provides probability distributions for the parameters of interest conditioned on the available data, thereby allowing us to quantify the uncertainties on these parameters. However, there is the issue that not all the parameters of interest can be determined. The iterative approach mentioned above only focuses on the reactions to which species are the most sensitive. This allows for the reduction in computational expense, unlike for Bayesian inference. In this Chapter we look to use various features of the network topology to reduce the computational expense of the inference process.

In this Chapter, we build upon the work done by Holdship et al. (2018) (hereafter H18) and use the same reaction network to highlight how aspects of the geometry of the network can be exploited to determine the model parameters at reduced computational expense. It should be noted that this reaction network only has 24 reactions with four constraints, but this technique should generalise to larger networks. It is hoped this will prove particularly useful when considering reaction networks of complex organic molecules (COMs), where the number of reactions is large and the number of constraints small.

We begin by first presenting the chemical network used in Section 3.2. We introduce the concept of Bayesian inference as applied in this Chapter in Section 3.3. Following this, we argue why we can reduce the network with its constraints to a simpler one before presenting explanations for how the positions of particular constraints in the networks are crucial in Section 3.4. We then go on to discuss how specific aspects of the topology of the reaction network are useful to consider and how these influence our choices when we reduce the network in Section 3.5. Finally, we look at how we can separate a reaction network into smaller sub-networks in Section 3.6.

3.2 The Chemical Network

We use the same chemical network as in H18, which we set out pictorially in Figure 3.1. We list all reactions in Table 3.1, assigning them numbers which we use throughout the paper for brevity. For the sake of simplicity, the hydrogen abundance is not a conserved quantity, as it is typically $\sim 10^4$ times more abundant than any other molecule, so its loss in this reaction network is negligible. We also emphasise that this network is a toy model and is not meant to properly reflect a complete grain surface network but simply serves as a proof-of-concept.

A modified version of the open source gas-grain code UCLCHEM was used (Holdship et al. 2017). This version only considered the surface chemistry of a collapsing dark cloud. The cloud is modelled as collapsing from a density of 10^2 cm^{-3} to 2×10^4 cm^{-3} over 10 million years. The collapse takes place isothermally at 10 K. As chemistry only occurs on the grain surface, freeze-out rates are required. Freeze-out rates are the rates at which atomic and molecular species stick to the dust grains and build up ices (Hocuk et al. 2014; Fraser et al. 2003). The freeze-out rates were found in H18 by running a single-point model of the full UCLCHEM code. The species that were given freeze-out rates were: CO, CS, O, H, OH and S. More details can be found in H18.

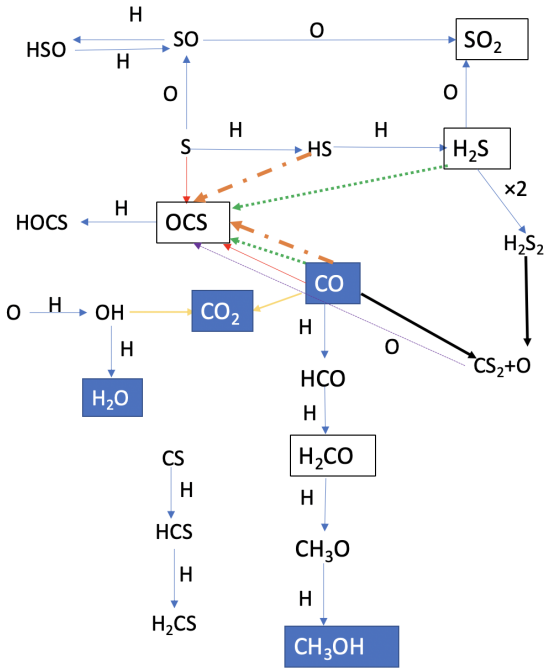

Figure 3.1: *A diagram of the chemical reaction network considered. For the sake of simplicity, any reactions with hydrogen and oxygen are represented with H and O next to the arrow. For the case where a molecule can be formed in multiple reactions, such as for OCS, the arrow colours pointing to that molecule indicate the reactants. For example, the dash-dotted orange arrows that point from HS and CO to OCS indicate that these two molecules form OCS. Molecules in blue boxes have constraints on their final abundances. Molecules in white boxes have upper limits on their abundances.*

As already mentioned above, the small chemical network we use for this Chapter is not meant to represent a comprehensive surface network. Nevertheless, the choice of most of the reactions was based on the results of experimental studies. For example, the successive hydrogenation of CO to form CH_3OH has been studied in detail and is considered to be the dominant reaction pathway (Chuang et al. 2016). Similarly, CO_2 has been found to be

Reaction No.	Reaction
1	$O + H \longrightarrow OH$
2	$OH + H \longrightarrow H_2O$
3	$CO + OH \longrightarrow CO_2$
4	$S + H \longrightarrow HS$
5	$HS + H \longrightarrow H_2S$
6	$H_2S + S \longrightarrow H_2S_2$
7	$CS + H \longrightarrow HCS$
8	$HCS + H \longrightarrow H_2CS$
9	$CO + S \longrightarrow OCS$
10	$OCS + H \longrightarrow HOCS$
11	$H_2S + CO \longrightarrow OCS$
12	$H_2S + H_2S \longrightarrow H_2S_2$
13	$H_2S_2 + CO \longrightarrow CS_2 + O$
14	$H_2S + O \longrightarrow SO_2$
15	$CS_2 + O \longrightarrow OCS + S$
16	$CO + HS \longrightarrow OCS$
17	$S + O \longrightarrow SO$
18	$SO + O \longrightarrow SO_2$
19	$SO + H \longrightarrow HSO$
20	$HSO + H \longrightarrow SO$
21	$CO + H \longrightarrow HCO$
22	$HCO + H \longrightarrow H_2CO$
23	$H_2CO + H \longrightarrow H_3CO$
24	$H_3CO + H \longrightarrow CH_3OH$

Table 3.1: *Table of the reactions used in this Chapter taken from Holdship et al. (2018)*

efficiently formed when CO and OH react (Ioppolo et al. 2011a). Beside a small network representing the main routes of carbon- and oxygen-bearing species on the ices, we chose to include a small network producing sulphur-bearing species, since there is still much unknown about the form that ultimately sulphur takes on the ices during the cold phase of the star formation process (Woods et al. 2015; Laas and Caselli 2019).

In this Chapter we consider a number of variants of the chemical network shown in Figure 3.1. These configurations, which differ in terms of the reactions and/or constraints used, are enumerated in Table 3.2: the configuration numbers will be used throughout the work. The combination of the full reaction network shown in Figure 3.1 with the abundance measurements as listed in Table 3.3 is Configuration 1.

Configuration No.	Reactions Used	Molecules with Constraints
1	1-24	CO, CH_3OH, CO_2, H_2O
2	1-6, 9-24	CO, CH_3OH, CO_2, H_2O
3	1-3, 21-24	CO, CH_3OH, CO_2, H_2O
4	1-24	CH_3OH, CO_2, H_2O
5	1-24	CO, CO_2, H_2O
6	1-24	CO, CH_3OH, CO_2, H_2O, SO_2, OCS, H_2S
7	4-20	CO, SO_2, OCS, H_2S
8	4-20	SO_2, OCS, H_2S
9	1-3, 21-24	CH_3OH, CO_2, H_2O

Table 3.2: *A table listing all the various network configurations used and referred to throughout this Chapter.*

3.3 Bayesian Inference

3.3.1 Introduction to Bayesian Inference

As in H18, our aim is to determine the 24 reaction rates, which we represent as a parameter vector, $\boldsymbol{\theta} = (k_1, k_2 \ldots, k_{24})$. We are therefore faced with a 24-dimensional inference problem. The grain code used took in these reaction rates and produced the corresponding abundances, which are represented by a vector $\mathbf{Y} = (Y_1, Y_2 \ldots, Y_{23})$. Henceforth we refer to the "forward model" when we input a particular value of $\boldsymbol{\theta}$ to obtain some $\mathbf{Y}$.

We know the abundances of four of the molecules in the ices of the network shown in Figure 3.1, which are in blue boxes. These are taken from Boogert et al. (2015) and listed in Table 3.3. We are looking to solve the "inverse problem", i.e. what values of $\boldsymbol{\theta}$ yield values of $\mathbf{Y}$ that match the observations best? Such a problem naturally lends itself to a Bayesian approach. The inherent degeneracy of this problem should be noted. Observations only exist for 4 of the 24 molecules. This suggests that the rates of the reactions that do not influence the abundances of these four molecules will be poorly constrained (if at all), and there will be many values of these rates that give the same observations. For a discussion of the degeneracies, please refer to H18. Exploiting the low number of constraints in order to speed up the inference process is a crucial point in this Chapter.

We use Bayes' Law to determine the probability distribution of the values of the reaction rates as described in Section 1.5. The evidence is a normalising factor and is typically difficult to evaluate. However, as it is independent of $\boldsymbol{\theta}$, we can instead just consider

$$P(\boldsymbol{\theta}|\boldsymbol{d}) \propto P(\boldsymbol{d}|\boldsymbol{\theta})P(\boldsymbol{\theta}), \tag{3.1}$$

which is just the unnormalised posterior.

3.3.2 Implementation

Evaluating the reaction rate posterior requires specification of a prior on the reaction rates and a likelihood. We chose the same prior as in H18: a log-uniform distribution between 10^{-30} and 10^{-5}. From chemical considerations, we know that these reaction rates are ultimately very fast. Typically, we might therefore select a prior that favours higher reaction rates. We chose, however, to ignore this information, instead following the traditional approach of using a log-uniform prior, which equally weights rates over a range of orders of magnitude.[1] It is important to note that, despite our motivation for this prior, we realise that we are in a prior-dominated regime, which we demonstrate in Appendix A with very different prior assumptions. Unlike in the data-dominated regime, these prior-dominated posteriors differ significantly among themselves. This, however, does not detract from the analysis we conduct in this Chapter.

A Gaussian likelihood function was used, which takes the form

$$P(\mathbf{d}|\boldsymbol{\theta}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(d_i - Y_i)^2}{2\sigma_i^2}\right), \tag{3.2}$$

where n_d is the number of observations and σ_i is the uncertainty of the ith observation. We only multiply over the species which have observed abundances. We refer to these observed abundances as constraints as they constrain the parameter space of our reaction rate posteriors.

In order to determine the posterior, the PyMC3 Python package was utilised (Salvatier et al. 2016). The PyMC3 package includes a range of samplers. Here, we used the Metropolis-Hastings algorithm, a simple Markov Chain Monte Carlo (MCMC) method. A Gaussian proposal distribution was used. Before each run, 500 tuning steps were initially

[1] It is worth noting that the use of a single prior to represent complete ignorance (Walley 2000; Norton 2008) has received criticism. Complete ignorance can be represented by repeating the analysis using several priors that significantly differ from one another (see Fischer (2019) for a straightforward application to a simple problem of chemical kinetics).

taken to determine the optimal covariance of the proposal distribution. These tuning steps were not included in our analysis and were discarded. For the relatively low number of dimensions (< 50), a point-based sampler such as this one is suitable. 50 chains of length 10^6 were used to sample the posterior probability space. We created a Python wrapper of the grain code using F2Py that was then fed values of $\boldsymbol{\theta}$ during the sampling process (Peterson 2009).

Though simple sampling methods suffice for the dimensionality of the posteriors considered here, the same will not be true for more complex networks. A key point that needs to be considered is the limitations of many MCMC methods as the number of dimensions increases. Brewer and Foreman-Mackey (2016) discuss how some widely-used samplers struggle to give sensible results in higher dimensions. The popular emcee Python package that was used in H18 is discussed to be useful for when the number of dimensions is fewer than 50, with it struggling in higher dimensions (Huijser et al. 2015). Even for the case where the sampler does not struggle as the number of dimensions increases, the time taken to run the inference process will still increase. This increase in computation time will eventually become prohibitive. In Section 3.6 we will argue that we can split our reaction network up into sub-networks on which we can perform Bayesian inference. By carefully placing the "cut" on the reaction network, we then show that we can reproduce the results of the full reaction network inference with these sub-networks. Each sub-network has lower dimensionality than the original network, meaning its rates can be inferred more quickly and by simpler samplers. Additionally, the inference process on all the sub-networks can be run in parallel.

3.3.3 Constraints

Species	Abundances relative to H
H_2O	$(4.0 \pm 1.3) \times 10^{-5}$
CO	$(1.2 \pm 0.8) \times 10^{-5}$
CO_2	$(1.3 \pm 0.7) \times 10^{-5}$
CH_3OH	$(5.2 \pm 2.4) \times 10^{-6}$

Table 3.3: *The abundances and uncertainties for the molecules with observed values taken from Boogert et al. (2015).*

It is essential to include constraints in order to formulate a likelihood function. In H18, four constraints for molecules in the reaction network were taken. Table 3.3 shows the

abundances of the molecules taken from Boogert et al. (2015). These ice abundances are derived from ice band profiles. The column densities can be calculated using the integrated optical depth as well as the integrated band strength. The latter were determined from laboratory experiments by Boogert et al. (2015). From the table, it is clear that the strongest constraint is on H_2O, which is known to exist at the $3\,\sigma$ level, whereas the other molecules' abundances differ from zero at only $1.5 - 2.2\,\sigma$. H18 also reformulated the likelihood function to include upper bounds on the abundances of OCS, H_2S, SO_2 and H_2CO. This was not found to have a significant effect on the reaction rates determined and so we do not include these upper bounds in the following work. For the rest of this Chapter we use a likelihood function of the form in Equation 3.2.

3.4 Network Reduction Methods

3.4.1 Overview

Galagali and Marzouk (2019) considered a similar problem to the one discussed in this Chapter in the context of systems biology. There, they considered the case of a reaction network with a single observation. The main differences between the networks they considered and the one being considered here is the absence of enzymes as well as the absence of reversible reactions.

They defined the "effective reaction network" as the subset of reactions that must be kept in order to produce the same values of the observable. This is a useful concept to consider, especially in the context of network connectivity. Some subsets of reactions will evolve completely independently of one another, with there being no competition for chemical species. While this may seem unlikely in the context considered here, it is true when one assumes that hydrogen's abundance is significantly higher than that of any other species. This can be seen in figure 3.1, where the successive hydrogenation of CS to form H_2CS is clearly independent of the rest of the reaction network. This reaction chain will be referred to as the "H_2CS chain" from now on.

3.4.2 Network Reduction of Non-Connected Networks

We consider the "H_2CS chain" in greater depth. Under the assumption that there is no competition for hydrogen, we should expect the "H_2CS chain" and the other reaction network with the remaining 22 reactions (Configuration 2) to evolve completely indepen-

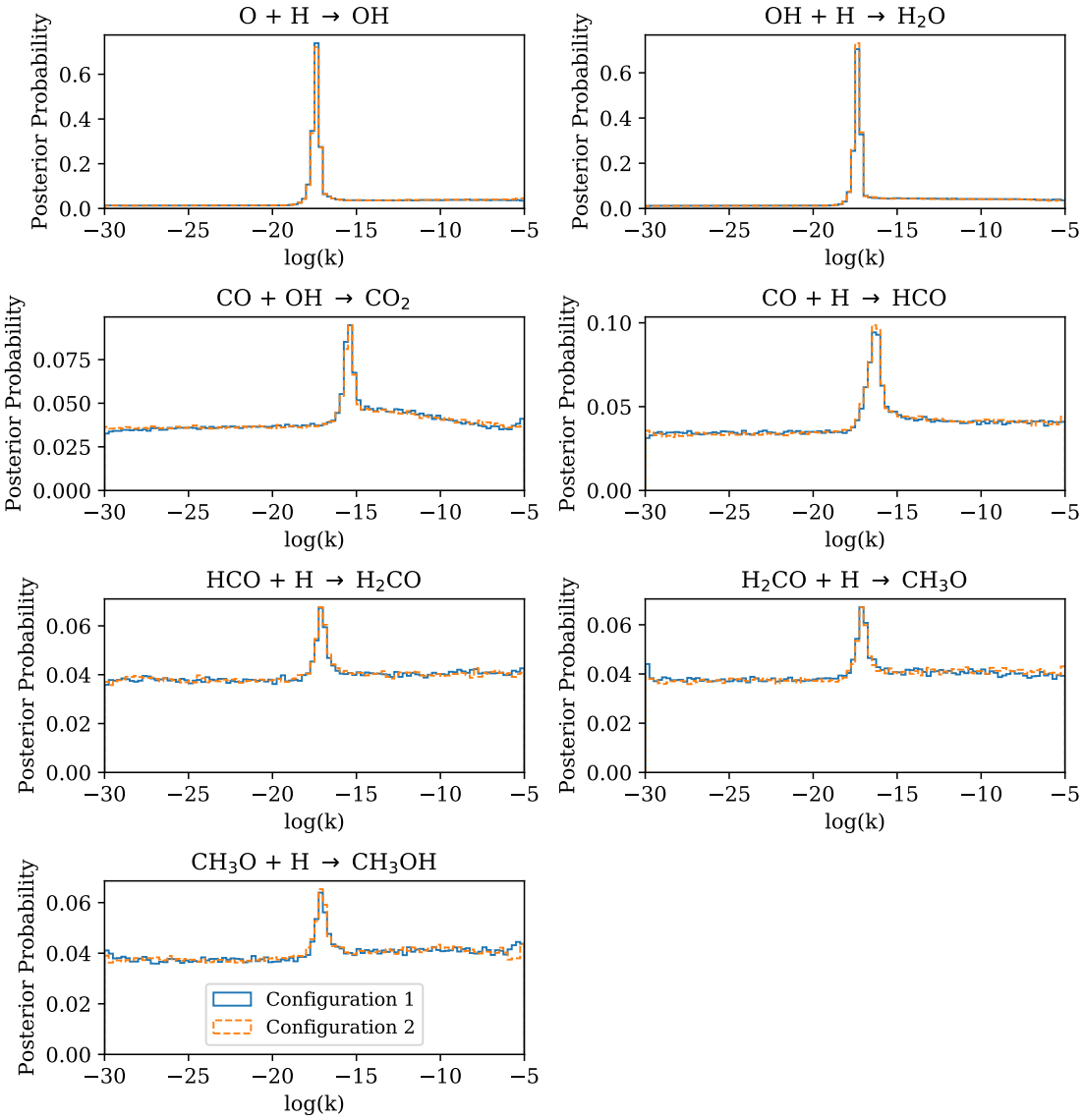

Figure 3.2: *Plots of the posterior probability distribution for the original reaction network considered in Holdship et al. (2018) as well as the 22-dimensional effective network by removing the "H_2CS chain". We observe good agreement in the shapes of the posterior distributions, with any differences due to specific samples drawn from the MCMC chains. The configuration 1 posteriors match those from H18.*

dently, as the species do not interact. We infer the reaction rate posteriors for Configurations 1 and 2. Figure 3.2 shows the posteriors on the common reactions to be essentially identical. Any differences are due to the specific samples drawn in the MCMC chains. This intuitively makes sense as the ordinary differential equations that govern the two sub-networks will evolve independently of one another. Since none of the reactions in the "H_2CS chain" are constrained in Configuration 1, it stands to reason that these reaction rates are nuisance parameters. As such, removing these two reactions from the inference should make no difference.

While the example given might seem trivial, one needs to consider under what circumstances one might have a reaction network with a disconnected segment. In surface grain chemistry, the molecules must contend with both an activation energy barrier as well as a diffusion energy barrier. Only if both of these can be overcome, is the reaction likely to happen efficiently. If either one of these barriers is too high, then one can in fact approximate that reaction as not happening and "cut" off that reaction. This might lead one to separating a reaction chain from the rest of the network. In this example, one could conceivably imagine a hypothetical reaction being possible between H_2CS and any other molecule in Configuration 2, but the activation energy barrier is too high to overcome at 10 K. It is therefore simpler to exclude it.

3.5 Further Network Reduction

In the previous section, we observed that the network connectivity of a chemical reaction network can allow us to discard reactions that do not influence the values of observed abundances. In this section, we develop this idea further by arguing that the locations of the constraints in the reaction network allow us to discard more reactions.

We wish to emphasise once again that we are not seeking to make quantitative predictions about the reaction rates. Instead, we are looking to develop a qualitative understanding of the kinetics as well as develop an intuition for how the methods that will be discussed in this Chapter can be applied to other astrochemical modelling scenarios.

3.5.1 Reducing the Network

We begin by briefly returning to the posteriors in Figure 3.2. The uncertainties on the rates of reactions 1 and 2 are significantly smaller than for reactions 3 and 21-24. This

relates to the uncertainties on the abundances of the molecules, as discussed in Section 3.3.3. In the limit that the abundances of all molecules involved in a reaction are perfectly known, one would expect the posterior distribution of the reaction rate to approach a Dirac delta function. The greater confidence level in water's presence is therefore responsible for the tighter constraints on the rates of Reactions 1 and 2. Improving the precision of the abundance measurements of CO, CO_2 and CH_3OH would in turn tighten the constraints on their reaction rates. In Figure 3.1, we observe that all the reactions whose reaction rates are constrained have a constraint at the end of their respective chain. Consider the successive hydrogenation of carbon monoxide to form methanol, henceforth referred to as the "methanol chain". The fact that there is a constraint present at the end is significant. By constraining the amount of methanol, one effectively constrains the reaction rates of its precursors, CH_3O and H_2CO from below, as the existence of methanol requires its precursors to have been produced. If the reaction rate of these reactions is too high, then too much methanol will be produced. However, there is an inherent degeneracy in the rates of the intermediate reactions, so it is unclear how the rates are partitioned between the two reactions. What one finds is that these intermediate reaction rates are coupled, by observing their joint probability distributions. One reaction will serve as the rate-limiting reaction, with the other compensating for this by being significantly faster to produce a sufficient amount of the final product, in this case methanol. This is discussed in more depth for the specific reactions in H18.

Additionally, the constraint on carbon monoxide constrains the abundances of these molecules from above. One can then qualitatively say that as there are now constraints on their abundances, there are therefore constraints on their reaction rates, as the reaction rates reflect how much of these molecules forms over time.

In the network shown in Figure 3.1, none of the sulphur-bearing molecules have constraints on them. The reaction rates of these reactions can be treated as unconstrained parameters, despite the fact that carbon monoxide, a central molecule in the sulphur-centric network, is constrained. This suggests that the sub-network consisting of the 10 non-sulphur bearing species can be treated independently. We exclude these sulphur-bearing reactions for now and only consider reactions 1-3 and 21-24. This is Configuration 3, which is a "sub-network" of Configuration 1 as it contains a subset of the reactions.

To investigate the impact of excluding the sulphur-bearing reactions, we re-run our Bayesian inference on this configuration, producing the posterior probability distribution

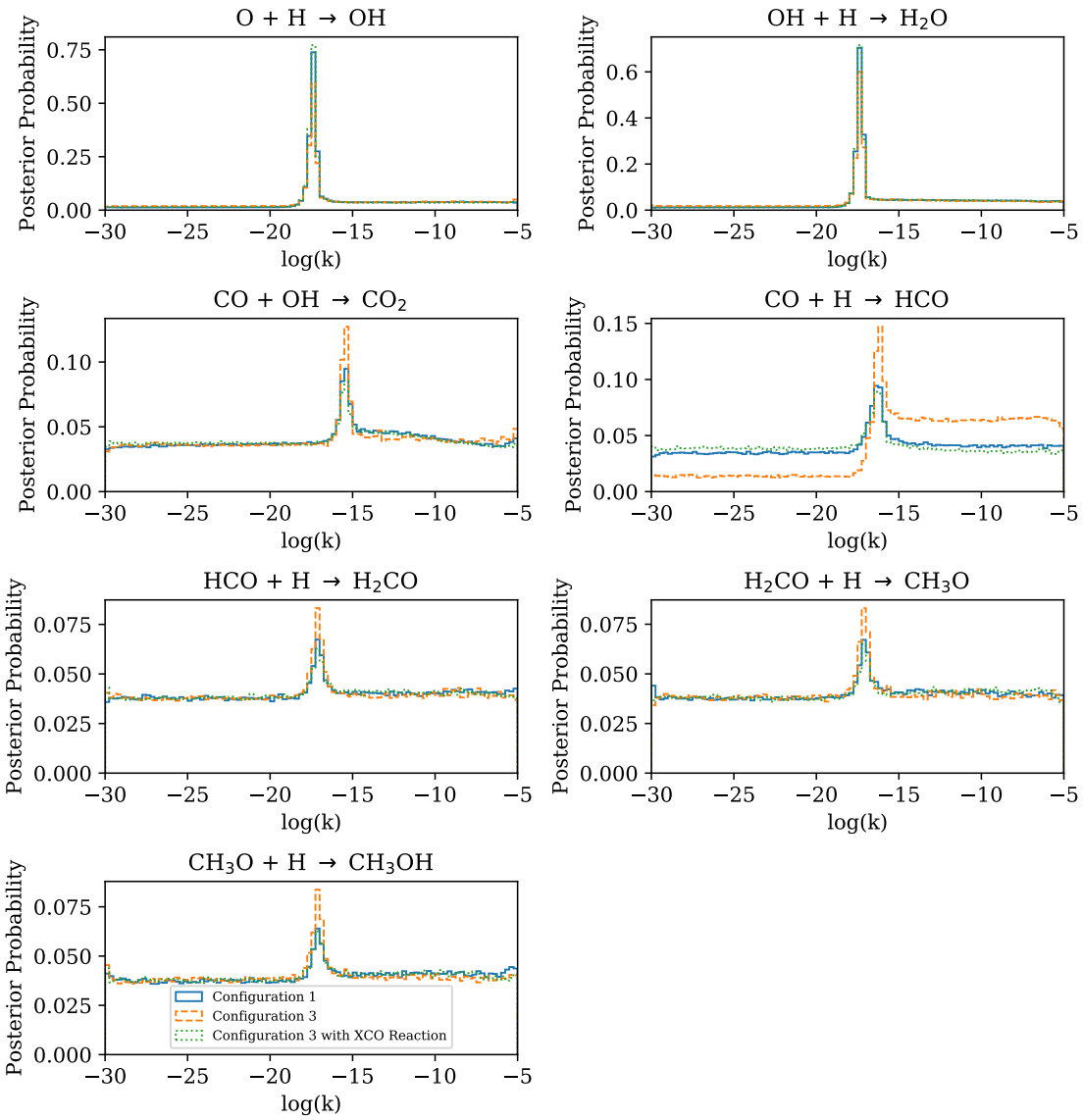

Figure 3.3: *Plots of the posteriors of Configuration 1, Configuration 3 and Configuration 3 with the dummy reaction $X + CO \rightarrow XCO$. We observe that the inclusion of this additional dummy reaction provides a better approximation to the Configuration 1 posterior than the Configuration 3 posterior does.*

functions in Figure 3.3. With the exception of $H + CO \rightarrow HCO$, we recover the maximum-posterior reaction rates obtained previously. However, we see that the variances have decreased for reactions 3 and 21-24, resulting in more peaked distributions, and reactions 1 and 2 have posteriors that are less peaked than for Configuration 1.

We would like to emphasise that this is a purely artificial effect. By eliminating the sulphur-bearing reaction rates, which were essentially nuisance parameters, the variance of reactions 3 and 21-24 have decreased. It should be noted that these reactions compete for CO with the removed sulphur sub-network. Removing the sulphur-based reactions means the non-sulphur reactions have to use up more CO. We observe that of all the reactions, the reaction $H + CO \rightarrow HCO$ sees the greatest change between Configuration 1 and Configuration 3. In fact, we observe that a significant portion of the posterior mass is shifted from the reaction rates below the peak to the reaction rates above the peak. This, coupled with the increase in the maximum-posterior of the rate parameter, suggests that the excess CO, that would normally be consumed by the sulphur reactions, is stored in the methanol chain. In particular, the fact that only the hydrogenation of CO experiences a significant change in the posterior suggest that the excess CO is stored as HCO.

At this point, it is unclear why Reactions 1 and 2 see increases in the variance of their posteriors. For these reactions, the decrease in the posterior mass under the peaks is compensated for by an increase in the posterior masses for reaction rates slower than the maximum posterior-rate. This suggests that the reactions can proceed at slower rates and still produce sufficient OH that goes on to produce H_2O and CO_2.

The MCMC runs for Configuration 3 are 2.3 times shorter than for Configuration 1, with the time taken for the runs decreasing from 30 to 13 hours. By excluding the unconstrained reactions, we are able to reduce walltime drastically, at the cost of moderate changes to the posterior. In the next sub-section, we discuss a method for recovering the full posterior. Finally, it should be emphasised that this dimension reduction must only be considered when solving the inverse problem. For a full picture of the chemistry one must include all reactions in the forward-model.

3.5.2 Recovering the Full Variance

We noted previously that the decrease in the variance of the posteriors was an artificial effect. Recovering as much of this original variance as possible is critical, as without it the precision of the inferred rates will be overstated. A consideration of the reactions

that were removed is a good starting point. Looking at Figure 3.1, one can see that from the perspective of Configuration 3, the removal of the reactions affected the CO depletion. In other words, Configuration 3 only sees that we removed CO-depleting reactions. The products of these CO-depleting reactions are not constrained, and these reactions are therefore responsible for the larger uncertainty in the posteriors of Configuration 1 compared to Configuration 3.

In order to recover this variance, one must account for the artificially altered CO depletion. As a first, simple approximation we add a fake reaction $X + CO \rightarrow XCO$, where X is meant to encompass all the removed reactions that consumed CO, and XCO is simply the network of products. The abundances of both X and XCO remain unconstrained. Even though X is not meant to represent a particular molecule, it still requires a freeze-out rate, as this grain-code only considers reactions that take place on the grain surface. The freeze-out rate for sulphur was used. The posteriors are shown in Figure 3.3. Adding a single unconstrained reaction clearly yields a good (though still imperfect) approximation to the full set of sulphur-consuming reactions in this setting, matching the variance of the full network's rate constants more closely and removing the bias on the inferred rate of $H + CO \rightarrow HCO$. Further work should be done to investigate whether increasing the number (and architecture) of "dummy reactions" aids in recovering the full posterior.

3.5.3 Network Topology Considerations

We now discuss how the placement of the constraints in the network can be significant. From the above analysis, it is not entirely clear what constraints are the most "essential". To shed light on this problem, we focus on the methanol chain, which has a constraint on molecules at both ends.

We perform Bayesian inference on the full reaction network twice:

- once without the CO constraint (Configuration 4)
- once without the CH_3OH constraint (Configuration 5)

The resulting posterior probability distribution functions for Reactions 21-24 are shown in Figure 3.4 alongside the distributions for the original network. The posteriors for Reactions 1-3 are not included, because these did not change significantly.

We observe that removing the constraint on CO has no effect on the reaction rates. The reaction rate posteriors for reactions 21-24 are broadly identical. This can be easily

explained. In this case, one knows that a fixed amount of methanol is produced. As such, over the period of 10 Myr, a certain amount of CO has to be consumed. This results in the successive hydrogenations being constrained, which is why the reaction rate posteriors do not change.

However, removing the constraint on methanol results in the loss of constraints on 3 of the 4 reaction rates of the methanol chain, with only the hydrogenation of CO being recovered. This appears to suggest that there is a notion of "distance" between a constraint and the reaction rate of interest. Information about the subsequent reaction rates in the methanol chain is lost due to methanol's abundance being unconstrained. In this model, we know that CO (as one of the adsorbed species) is present on the grains, but there is no information about how much of it goes into making methanol. However, it is interesting that reaction 21 remains constrained. One possible interpretation is that we know how much CO is used in reaction 3 and this possibly helps constrain how much CO is used in reaction 21. However, using this reasoning, one cannot explain why reactions 9 and 13 are not constrained, as these are also CO depletion reactions.

This suggests that in any reaction chain, some knowledge of the abundance of the end-products is required, which might be problematic when the species are undetected. However, one can still provide theoretical predictions for these abundances that could be used.

3.6 Application to the Network with Artificial Sulphur Constraints

As a proof of concept, we wish to apply the insight from the previous section to a new grain surface network. However, this is difficult due to the limited number of observations that exist for grain-surface molecules. Boogert et al. (2015) provided upper limits for several molecules in the ice. We chose to artificially transform the upper limits on OCS, H_2S and SO_2, into weak measurements by taking their abundances to be half the respective upper limit with an uncertainty of one-quarter of the upper limit. We would like to emphasise again here that we are simply trying to demonstrate how the location of these three additional constraints provides us with more knowledge of $\mathbf{k}$. We do not claim this to be an accurate representation of sulphur chemistry on the ices in a dark cloud. Many theoretical and modelling studies have been recently performed investigating the sulphur

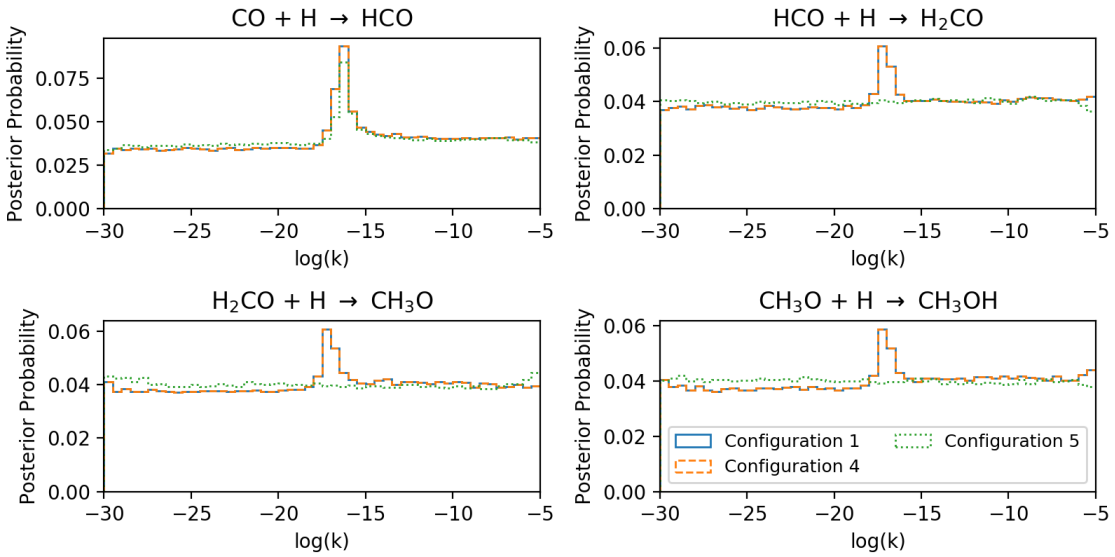

Figure 3.4: *The posterior probability distributions for reactions 21-24 when CO and methanol are separately removed. The original distributions are also included for comparison. We observe that for reactions 21-24, removing CO neither affects the position of the peak of the distribution nor the shape of the distribution. Removing methanol does not change reaction 21's maximum-posterior rate, but removes all information about the reaction rates of reactions 22-24. We do not include reactions 1-3, as their posteriors are unchanged when the constraints are removed.*

depletion and the reactions on surfaces involving sulphur-bearing species and we refer the reader to such studies for a comprehensive review on the subject (e.g. Jiménez-Escobar et al. (2014); Woods et al. (2015); Vidal and Wakelam (2018); Laas and Caselli (2019)).

3.6.1 The Full Network

Bayesian inference was performed for the full network with the new artificial constraints. This is Configuration 6. Despite adding these constraints elsewhere in the network, the maximum-posterior reaction rates of reactions 1-3 and 21-24 (none of which involve sulphur compounds) were found to be unchanged and the posteriors were largely similar. This fact strongly implies that the sulphur-based and non-sulphur-based reactions can be separated into sub-networks, whose reaction rates can be inferred independently, even when constraints on the sulphur-based products become available.

We also identify eight new reactions for which the marginalised posterior probability distributions deviate from uniformity. These are all reactions that involve several of the molecules whose abundances have now been constrained. The posterior probability

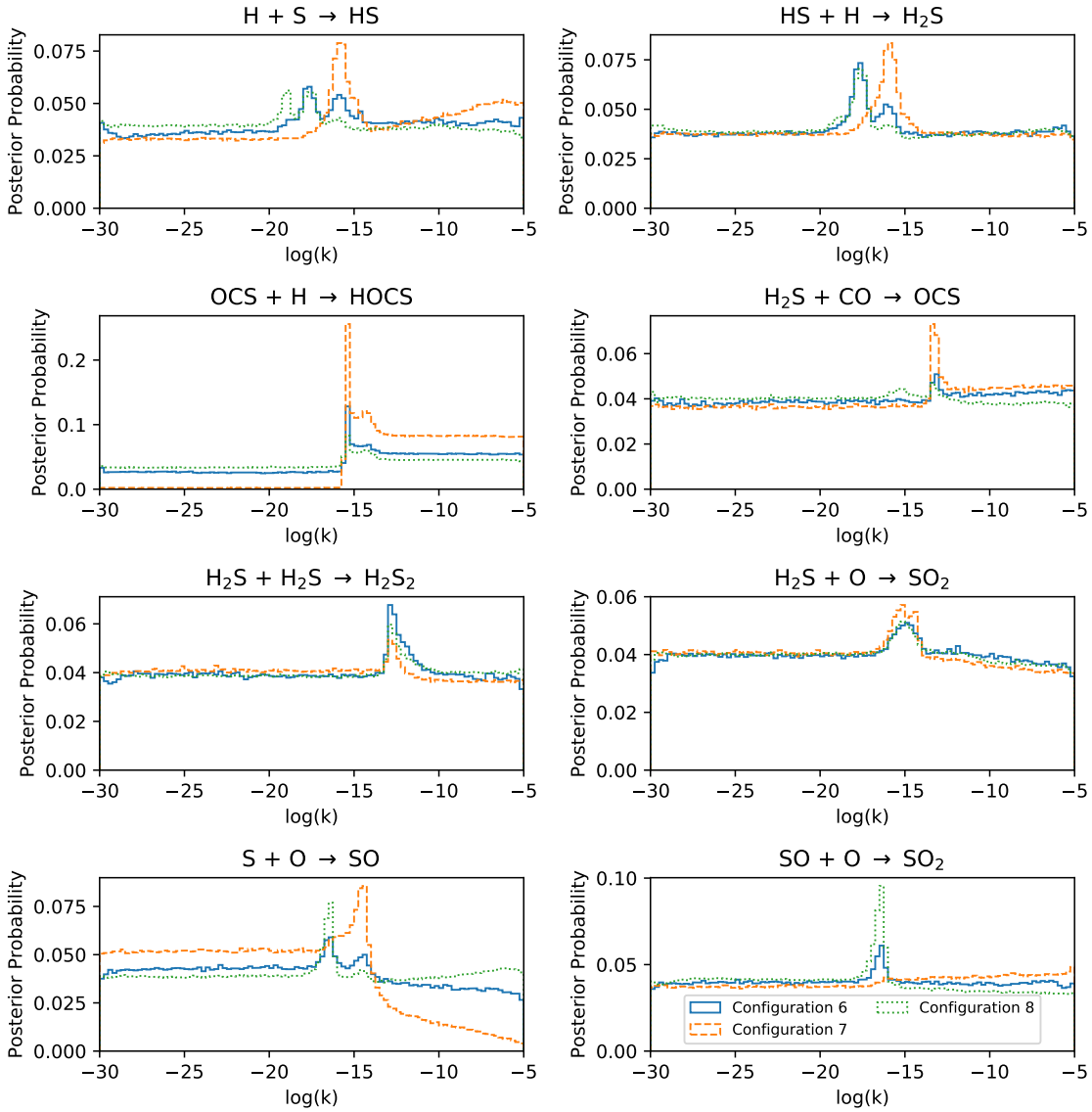

Figure 3.5: *Plots of the posterior probability distribution which deviate from uniformity for the expanded reaction network. We compare the posterior distributions of Configuration 6 with those of Configurations 7 and 8. We observe better agreement of the sulphur sub-network when we leave CO's abundance as a free parameter, which corresponds to Configuration 8.*

Species	Abundances
H_2O	$(4.0 \pm 1.3) \times 10^{-5}$
CO	$(1.2 \pm 0.8) \times 10^{-5}$
CO_2	$(1.3 \pm 0.7) \times 10^{-5}$
CH_3OH	$(5.2 \pm 2.4) \times 10^{-6}$
OCS	$(6.0 \pm 3.0) \times 10^{-8}$
SO_2	$(2.0 \pm 1.0) \times 10^{-6}$
H_2S	$(8.0 \pm 4.0) \times 10^{-7}$

Table 3.4: *The abundances and uncertainties taken for the network with artificial sulphur constraints. For the first four species, the abundances were taken in their present form from Boogert et al. (2015). Boogert et al. (2015) provided upper bounds for the listed sulphur-based species. For the analysis in this section, the abundances of the sulphur-based species were taken to be half the upper bound value. Their uncertainties were taken to be 50%.*

distributions are shown in Figure 3.5.

3.6.2 Including the CO constraint 1

In the following, we investigate the optimal way of splitting the full network into sulphur- and non-sulphur-based sub-networks. The two sub-networks compete over CO, one of the molecules with a constraint. We know from Section 3.5 that we can include the full CO constraint in the non-sulphur sub-network without significantly biasing the inferred reaction rates. To test if this is the case for the sulphur sub-network, we consider two cases, performing Bayesian inference on the sulphur sub-network with the full CO constraint (Configuration 7) and leaving the CO abundance as a free parameter (Configuration 8). The time taken for Configuration 6 is about 30 hours, whilst the runs for Configurations 7 and 8 took about 23 hours.

In Figure 3.5 we compare the rate posteriors for the artificially constrained sulphur-based reactions obtained with Configurations 6-8. For the eight new rate posteriors obtained in Figure 3.5, we observe better agreement when the CO constraint is not included. The main explanation for this is that as several CO depletion reactions in Configurations 7 and 8 have been discarded, the other reactions in the network are required to produce more molecules that will react to deplete the CO abundance. This can be seen by the fact that the reaction rates for the successive hydrogenation of sulphur to produce H_2S are greater when the CO constraint is included. This is because H_2S reacts with CO to produce OCS. To deplete the excess CO, more H_2S must be produced.

It is interesting to note that one can apply the full CO constraint in the non-sulphur

sub-network, but not in the sulphur sub-network. This is due to the relative sizes of the abundances of the constrained species. As shown in Table 3.4, the abundances of the sulphur-based molecules that are added are between 2 and 3 orders of magnitude less abundant than the constrained species in the other network. It should be noted that it is assumed that CO is already present to begin with and can only be consumed. No CO-formation reactions are present. As such, the contribution of the sulphur-network in depleting the CO is small compared to the non-sulphur network.

To get a better idea of the amount of CO that is used up by the sub-networks, we ran the forward models of the grain-code. By setting the reaction rates to zero, the only rates left were the freeze-out rates. These gave an idea of the total amount of CO available. This was found to be 4.0×10^{-5}. From Table 3.3, we know that the amount of CO that should be left is $(1.2 \pm 0.8) \times 10^{-5}$. Running the forward model of the sulphur-only network with CO as a free parameter, shows that the final CO abundance is about $(3.3 \pm 1.1) \times 10^{-5}$. This suggests that the non-sulphur network consumes four times as much CO as the sulphur-centric network does. By considering how the two sub-networks rely on the common constrained molecule, CO, we find that we can easily separate them.

Leaving the CO abundance unconstrained reduces the bias in the the maximum-posterior reaction rates, but this does not perfectly reproduce the full network's posterior. This is an issue, because it means that the variance of the full network is not preserved. An additional reaction of the form $X + CO \rightarrow XCO$ could potentially be used, just as was done in section 3.5.2. However, unlike before this reaction will be replacing a reaction sub-network with constraints. This makes the problem more complicated than before, as one might need to consider how to combine the constraints to create a "constraint" for XCO. One could simply give XCO an abundance equal to the sum of all the constraints that have been replaced. One might also need to provide several "dummy" reaction chains of varying length to best recover the original posterior. Considerations of the architecture will be discussed in future work.

3.6.3 Including the CO constraint 2

In the previous subsection, we argued that by virtue of the fact that the constraints in one sub-network were orders of magnitude greater than in the other sub-network, we could simply take the full CO constraint and use it in the former. Specifically, we want the CO constraint to be comparable to that of OCS, a molecule which depletes CO. However,

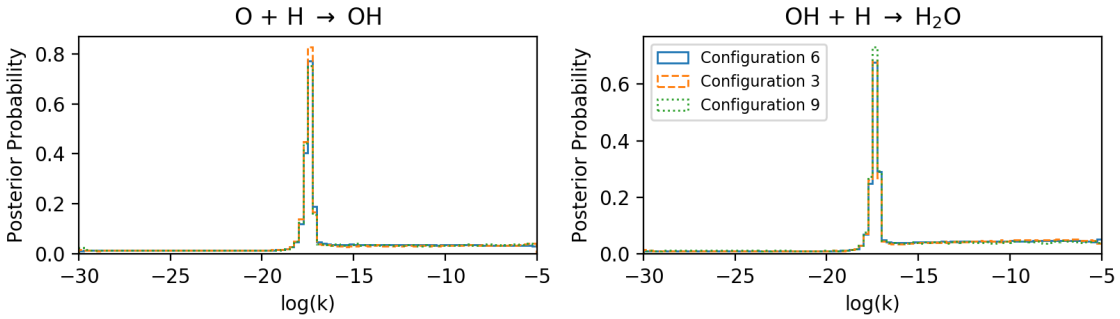

Figure 3.6: *The posterior distributions for Reactions 1 and 2 when the initial CO abundance is reduced by a factor of* 10^4*. We compare the posterior distributions of Configuration 6 with those of Configurations 3 and 9. We observe the best agreement between Configurations 6 and 9, suggesting that for this sub-network, it is better to exclude the reduced CO constraint.*

we now consider the case where the constraints in each sub-network are comparable in nature. To do this, we artificially reduce the abundance of CO in the system. This is done by reducing the freeze-out rate by a factor of 10^4. As only grain surface reactions are considered, this reduces the amount of CO available for grain-reactions by the same factor.

Obviously changing the amount of CO present in the model has an effect on reactions obtained. With so little CO, it is impossible to match the abundances of methanol and carbon dioxide. As a result of this, the posteriors of these reaction chains are close to uniform. However, the point of these simulations is not to model the chemistry accurately. We want to know what we should do with the CO constraint in each sub-network. To aid understanding, we will consider the two sub-networks separately. We would like to note that even though the CO constraint has been reduced, the configuration for the full-network still corresponds to Configuration 6.

The non-sulphur sub-network

Recall that the non-sulphur sub-network consists of reactions 1-3 and 21-24. The question is whether or not one wishes to include the reduced CO constraint to recover the posterior of the full-network (Configuration 6). Including the CO constraint gives Configuration 3 and excluding it gives Configuration 9. The posterior distributions are shown in Figure 3.6. We observe that the only posteriors that deviate from non-uniformity are those for the reaction rates of Reactions 1 and 2. We observe that Configuration 9, which corresponds to

excluding the reduced CO constraint, matches the posterior distribution of Configuration 6 the best for Reaction 1, not only in terms of the location of the maximum-posterior but also in terms of the posterior shape. Configuration 3 is found to recover the posterior better for Reaction 2.

We also observe that the maximum posterior reaction-rates for Reactions 1 and 2 match those obtained previously in the work, even though the CO abundance was greatly reduced. This adds support to the idea that Reactions 1 and 2 form their own sub-network and are ultimately independent.

The sulphur sub-network

Figure 3.7 shows the non-uniform posteriors of the reaction rates for the sulphur sub-network. We find that the maximum-posterior rates for these reactions do not change when we discard Reactions 1-3 and 21-24, regardless of whether we include CO's new abundance constraint. As before, however, the precise forms of some of the posteriors are very different. This did not appear to be as much of an issue in Figure 3.6, which might be related to the relative levels of uncertainty on the relevant species, as discussed in section 3.4.

3.6.4 Comments on the Topology of the Sulphur Sub-Network

We notice that adding the artificial constraints on sulphur-based species results in the rate of the reaction between H and OCS being constrained. This is interesting, as the abundance of the product, HOCS, is not constrained. Instead it is the penultimate molecule in the reaction chain, OCS, that is constrained. This suggests that it is not necessary for the end of a reaction chain to be constrained to constrain the reaction rates, as was observed with methanol in section 3.5.2. It seems that having a constraint on the penultimate molecule is sufficient. Constraining an earlier molecule would not do the trick, as was demonstrated when methanol's constraint was removed but CO's was kept. There is a notion of distance that needs to be considered. However, this would need to all be reconsidered for the case where there is more than one depletion mechanism for OCS. It is likely that having two depletion mechanisms, each of whose end product is unconstrained, would have a different effect, as there will be uncertainty about the branching ratio of each depletion route.

3.7 Conclusion

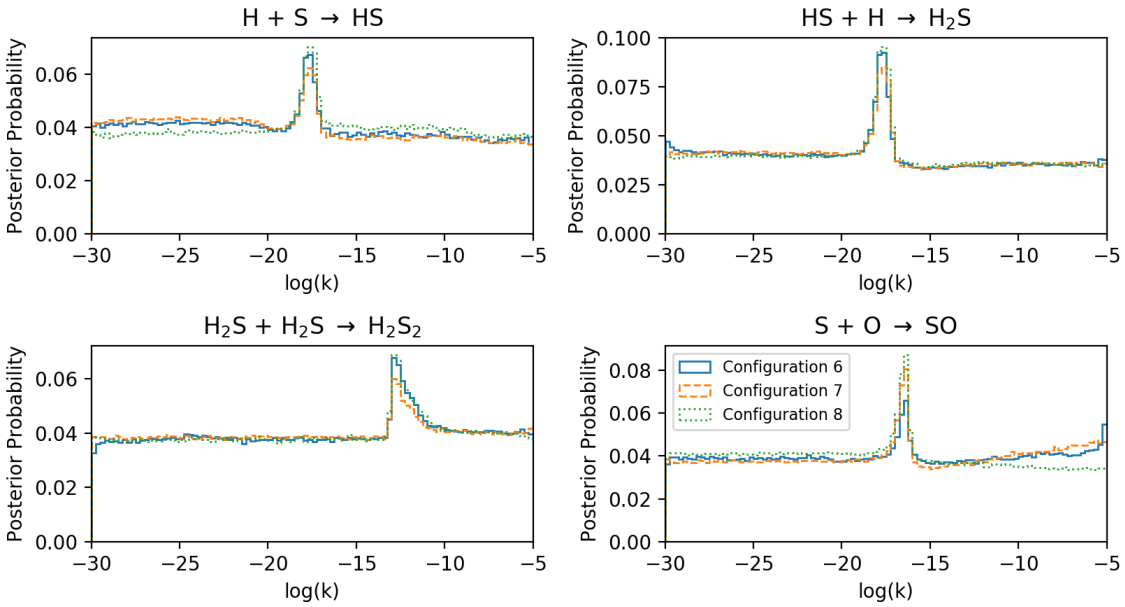

Figure 3.7: *Plots of the obtained posteriors for Configurations 6,7 and 8 when the CO abundance is reduced by a factor of* 10^4*. We observe that for this case, where the CO constraint is not several orders of magnitude greater than the abundances of the molecules in the sulphur sub-network, that it does not matter whether we include the CO abundance or not. Both cases allow us to recover the reaction rate with a very small bias.*

In this Chapter, we have proposed new methods for performing Bayesian inference on chemical reaction networks that have very few constraints. We find that reducing the reaction network to just the reaction chains whose ends are constrained allows us to greatly reduce the computational expense. Despite the simplification, our most likely reaction rate values are mostly unchanged. We also find that we can separate chemical reaction networks into sub-networks, which can be analysed in parallel. We believe that such network reduction techniques will be of great use when looking at grain surface reaction networks, where there are few constraints on the molecules. However, it should be noted that the results of such a simplified chemical model can only provide a qualitative understanding of the chemistry.

We briefly summarise some general observations we have made that might prove useful in reducing reaction networks for Bayesian inference:

- Reducing the network reduces the computational expense of the inference process. The time taken for our inference runs scales roughly linearly with the number of

reactions. However, the network connectivity is also likely to be a significant factor. This warrants further investigation.

- Reducing the network comes at the cost of artificially changing the variance of the posteriors. However, this variance can be partially recovered by adding a dummy reaction where the product is unconstrained, though there is the risk that the joint posterior distribution becomes more unrealistic. This needs to be investigated in further depth.

- When considering a reaction chain, it is important to include constraints for the final or penultimate molecule produced, as this ensures (for the case of a linear reaction chain) that the reaction rates of intermediate reactions are constrained. This might provide a general idea for future observations in terms of which molecules in the ices to look for. However, for more complex reaction networks, the intermediate reactions may play a more significant role.

- A network can be "separated" into sub-networks. This is a potentially very useful tool, especially when looking at the grain surface chemistry of complex organic molecules, where the networks themselves are very large. For example, Garrod (2013) provides a potential surface reaction network with around 200 reactions. In principle, this network could be split up into smaller sub-networks and Bayesian inference could then be performed on each sub-network in parallel. A potential general strategy would be to perform the network splitting at the point in the network with the highest network connectivity. For the network considered in this Chapter, this was the CO molecule. By making appropriate arguments about the relative magnitudes of the sub-networks and placing appropriate cuts in the networks, one could repeat the procedure as above. In order to decide what to do with a constrained molecule that is shared by the two sub-networks, one can make arguments about the relative abundances of the molecules in each sub-network. There are two cases to consider:

  - For the case where the shared constrained molecule has a significantly larger abundance than the molecules in one of the sub-networks, one can include its constraint in the higher-abundance sub-network.

  - For the case where the shared constrained molecule is roughly the same as the abundances in either of the two sub-networks, one can choose to include it in

either or both sub-networks.

For the case of a more interconnected networks with linear components (see Figure 11 in Linnartz et al. (2015) for an example of such a network), it makes sense to separate out the linear reaction chains first, as we found in section 3.5.3 that their topology is easy to understand intuitively. By removing these linear reaction chains, one could treat the interconnected sub-networks separately. Depending on their topology, one can employ the separation strategies discussed in this Chapter.

Further work will need to focus on recovering the posterior distributions better. A quantitative approach is needed to better compare the posteriors inferred with full networks and sub-networks as well as explore how to use "dummy reactions" of the form $X + CO \rightarrow XCO$ to recover the variance for the case where the reaction replaces a sub-network with constraints.

Future work will also need to to look reaction networks with more complex geometries. The example considered here is fairly simple, with a relatively low degree of connectivity. As the complexity of the reaction networks considered increases, there will need to be more well-defined notions of how the position of a constraint influences the inference of related reaction rates. The guidelines we have presented are applicable to simple networks. Investigation of more interconnected networks is the focus of ongoing work. We aim to come up with a set of criteria to determine how to best separate more complex networks.

Chapter 4

Bayesian Inference of Reaction Rate Parameters of a Glycine Network

The work presented in this Chapter is based on the paper Heyl et al. (2022b), in collaboration with Jonathan Holdship and Serena Viti.

4.1 Introduction

In Chapter 3 we discussed how we could use the topology of the network to reduce the computational time to estimate the reaction rates. However, this involved removing some reactions and/or splitting the network, both of which are manual processes that are likely to get more complicated as the network complexity increases. In this Chapter, we wish to come up with another way of reducing the dimensionality of the problem without having to make changes to the network.

As previously discussed, Bayesian inference can be used to estimate reaction rate parameters using observations. While this tool has become a staple in many areas of astrophysics, it is only recently that it has found use cases in astrochemistry (Makrymallis and Viti 2014; Holdship et al. 2018; de Mijolla et al. 2019). In Holdship et al. (2018), reaction rates were inferred using a toy network. In Chapter 3, the topology of this network was

also considered, specifically the placement of constraints within the network. Both of these works considered the rates of the reactions, without considering the actual, underlying reaction mechanisms. However, it was noted in both works that the paucity of grain-surface species abundances means that many of the reaction rates will remain undetermined, due to the high levels of degeneracy. This Chapter seeks to circumvent this issue by using the physics of the grain-surface diffusion mechanism to reduce the number of free parameters and therefore break this degeneracy.

To better understand the importance of various reactions, it is important to have knowledge of the binding energies on dust grains of the species involved. Molecular binding energies provide an upper temperature limit at which the species is still active on the grain surface before it desorbs into the gas phase (Penteado et al. 2017). As such, having accurate molecular binding energy values is crucial when modelling grain-surface chemistry, as Penteado et al. (2017) showed that the grain-surface chemistry was very sensitive to the values of the binding energy. A variety of approaches have been taken to determine the binding energies, ranging from experimental approaches (He et al. 2016) to density functional theory (Ferrero et al. 2020).

However, despite the various approaches used to estimate binding energies, there is still significant uncertainty when it comes to their values. In this Chapter, we use the Bayesian framework to estimate the binding energies of species. This is an important quantity, as it represents the mobility of the species on a dust grain. The values of the binding energies of species differ significantly across the literature (Penteado et al. 2017; McElroy et al. 2013; Wakelam et al. 2017). This high level of disagreement may be due to differing modelling and/or experimental approaches which cannot necessarily be reconciled. By using measured abundances of some grain-surface species, we are looking to provide estimates of binding energies with uncertainties.

However, Bayesian inference typically has a long run-time, that is dependent on both the number of dimensions that are being explored as well as the time taken per forward model evaluation. A higher dimensionality means that the Bayesian inference sampler requires more samples to converge to a stationary posterior distribution. We reduce the dimensionality of our problem by utilising physical considerations of the reaction mechanism. This also reduces the total time taken for the inference. We also use statistical emulation to reduce the time further by decreasing the time taken per forward model evaluation. This is particularly relevant when performing the inference multiple times,

given that each inference run calls the forward model tens of thousands of times.

We begin by first explaining the chemical code and network that will be used in this Chapter in Section 4.2. Additionally, we describe the grain-surface diffusion mechanism that lies at the heart of our investigation. We will explain how we can make approximations regarding a species' mobility to estimate the binding energy of said species. In Section 4.3, we will discuss how statistical emulation can be leveraged to accelerate the running of the forward model, before describing how Bayesian inference will tie all of this together in Section 4.4. Following this, we will present the resulting binding energies estimates using this method Section 4.5. We then consider the binding energy of atomic hydrogen in more depth in Section 4.6. In Section 4.7, we then look to see how well we are able to recover abundances when we run a full gas-grain chemical code using the estimated values.

4.2 The Chemical Code and Network

4.2.1 The Chemical Model and Code

The code that was used was based on the gas-grain chemical code UCLCHEM (Holdship et al. 2017). The surface chemistry is modeled through the rate equation approach. The code has to solve a system of coupled ordinary differential equations of the form given in Equation 1.13.

However, in order to reduce the runtime of the inference process, some changes had to be made to reduce the time taken for UCLCHEM to run. These are described in detail in Holdship et al. (2018), but are outlined briefly here. The code that was used considered only grain-surface chemistry to reduce the complexity of the system of coupled ordinary differential equations. However, it was important to still include the key processes that couple the gas and grain chemistry. It should be noted that the final two terms in Equation 1.13 represent the net flux of gas-phase molecules adsorbing to the grain surface. As such, if one only wishes to consider grain-surface chemistry, then one just needs to parameterize this net "freeze-out". The net freeze-out was found by running a single point-model of the full gas-grain version UCLCHEM. The net movement of each species between the gas and grain phases as a function of time was then extracted. Only the species which were deposited in abundances relative to n_H greater than 10^{-7} on the grains were included. These species were: H, O, OH, C, CO, N, CH and CH_3. These were all species which would form in the gas-phase and were involved in the reactions listed in Table 4.1. The freeze-out

rates were inserted as source terms into the grain-surface models. The freeze-out of the more complex species was not considered, as these species were unlikely to form in the gas-phase at 10 K. The advantage of doing this is that one avoided needing to consider the system of ODEs for gas-phase reactions, thereby significantly reducing the computational complexity.

The code models the surface chemistry of a collapsing dark cloud from a density of 10^2 cm^{-3} to 10^6 cm^{-3} over 10 million years at 10 K. As in Holdship et al. (2018), the model reaches its final density at 6 Myr, but the chemistry continues to evolve at constant velocity until the age of the cloud reaches 10 Myr. The grains start off as bare grains, with the freezeout of the gas-phase species acting as source terms for the grain-surface chemistry.

4.2.2 The Chemical Network

Our network is composed of radicals that react to form glycine, the simplest amino acid. The reactions that make up this chemical network are shown in Table 4.1. The grain-surface network used in this Chapter is based on the one used in Ioppolo et al. (2020) with the final two reactions being taken from Linnartz et al. (2015). In Ioppolo et al. (2020), laboratory and chemical modelling found that the first 47 reactions were able to produce glycine in dark interstellar conditions, long before the warm-up phase of star formation, without requiring any energetic input (Ioppolo et al. 2020). This is in contrast to previous work that assumed that the formation of glycine required an increased temperature as well as energetic processing (Garrod 2013). Based on Ioppolo et al. (2020), it was expected that this network would be sufficient to learn about COMs in the pre-stellar phase with the help of observed abundances. Reactions 48 and 49 were included, as they involved species already present in the network. Furthermore, one of the end-products, NH_4^+, had a constraint on its abundance that could be used for the Bayesian inference to further constrain the parameters.

4.2.3 Grain Surface Chemistry

Grain Surface Diffusion

Recall the discussion of the grain-surface diffusion mechanism in Section 1.3.1. Within this formalism, the diffusion energy is typically taken to be a fraction of the species binding

Reaction No.	Reaction	$\frac{E_b}{E_D}$
1	$H + H \longrightarrow H_2$	0.6
2	$O + H \longrightarrow OH$	0.6
3	$OH + H \longrightarrow H_2O$	0.6
4	$CO + H \longrightarrow HCO$	0.6
5	$HCO + H \longrightarrow H_2CO$	0.6
6	$HCO + H \longrightarrow H_2 + CO$	0.6
7	$H_2CO + H \longrightarrow H_3CO$	0.6
8	$H_2CO + H \longrightarrow HCO + H_2$	0.6
9	$H_3CO + H \longrightarrow CH_3OH$	0.6
10	$CO + OH \longrightarrow HOCO$	0.5
11	$CO + OH \longrightarrow CO_2$	0.5
12	$HOCO + H \longrightarrow H_2 + CO_2$	0.6
13	$HOCO + H \longrightarrow HCOOH$	0.6
14	$N + H \longrightarrow NH$	0.6
15	$NH + H \longrightarrow NH_2$	0.6
16	$NH_2 + H \longrightarrow NH_3$	0.6
17	$C + H \longrightarrow CH$	0.6
18	$CH + H \longrightarrow CH_2$	0.6
19	$CH_2 + H \longrightarrow CH_3$	0.6
20	$CH_3 + H \longrightarrow CH_4$	0.6
21	$CH_4 + OH \longrightarrow CH_3 + H_2O$	0.6
22	$NH_2 + CH_3 \longrightarrow NH_2CH_3$	0.5
23	$NH_3 + CH \longrightarrow NCH_4$	0.5
24	$NCH_4 + H \longrightarrow NH_2CH_3$	0.6
25	$NH_2CH_3 + H \longrightarrow NCH_4 + H_2$	0.6
26	$NH_2CH_3 + OH \longrightarrow NCH_4 + H_2O$	0.5
27	$NCH_4 + HOCO \longrightarrow NH_2CH_2COOH$	0.5
28	$OH + H_2 \longrightarrow H_2O$	0.35
29	$O + O \longrightarrow O_2$	0.6
30	$O_2 + H \longrightarrow HO_2$	0.6
31	$HO_2 + H \longrightarrow OH + OH$	0.6
32	$HO_2 + H \longrightarrow H_2 + O_2$	0.6
33	$HO_2 + H \longrightarrow H_2O + O$	0.6
34	$OH + OH \longrightarrow H_2O_2$	N/A
35	$OH + OH \longrightarrow H_2O + O$	N/A
36	$H_2O_2 + H \longrightarrow H_2O + OH$	0.6
37	$N + N \longrightarrow N_2$	0.6
38	$N + O \longrightarrow NO$	0.6
39	$NO + H \longrightarrow HNO$	0.6
40	$HNO + H \longrightarrow H_2NO$	0.6
41	$HNO + H \longrightarrow NO + H_2$	0.6
42	$HNO + O \longrightarrow NO + OH$	0.6
43	$HN + O \longrightarrow HNO$	0.6
44	$N + NH \longrightarrow N_2$	0.6
45	$NH + NH \longrightarrow N_2 + H_2$	0.5
46	$C + O \longrightarrow CO$	0.6
47	$CH_3 + OH \longrightarrow CH_3OH$	0.6
48	$NH + CO \longrightarrow HNCO$	0.5
49	$NH_3 + HNCO \longrightarrow NH_4^+ + OCN^-$	0.35

Table 4.1: *Reactions taken from Ioppolo et al. (2020) and Linnartz et al. (2015). The values of $\frac{E_b}{E_D}$ used for the more mobile species of each reaction are given.*

energy, E_D. There is debate surrounding the value of the fraction $\frac{E_b}{E_D}$, though there is agreement that it should be in the range 0.3 to 0.8, with lower values in this range being more appropriate for stable molecules (Penteado et al. 2017). However, it has been found that for O and N atoms, a ratio of 0.55 is more suitable (Minissale et al. 2016a). In this Chapter, we follow the convention adopted by Jin and Garrod (2020) where the ratio was set to equal 0.6 for atomic species and 0.35 for stable species. For all other species, a value of 0.5 was used. For each reaction, the value of $\frac{E_b}{E_D}$ for the more mobile species is given in Table 4.1 with the exception of Reactions 34 and 35 as these reactions are not assumed to take place via diffusion due to OH being widely reported as being an immobile molecule with a large binding energy. The reason we only consider the value of this ratio for the more mobile species is given in Section 4.4.5.

As most of the reactions in Table 4.1 are radical-radical, it was assumed that their activation energies were 0K. Even for reactions involving a known reaction barrier, such as reaction 5, it was found that $p_{reac} \gg p_{diff}$, which means the activation energy barrier is lower than the diffusion barrier. As such, $\kappa_{AB}^{final} \simeq 1$. This "diffusion-limited regime" corresponds to the situation where the diffusion process is the rate-limiting step and is due to the fact that the temperature being considered is 10 K. At 10 K, we also observe that the rate of evaporation is far lower than the rates of diffusion and reaction, so will be neglected throughout this Chapter.

4.3 Statistical Emulation

Statistical emulation involves fitting a statistical function to match the inputs and outputs of a forward model (Grow and Hilton 2018). The advantage in doing so is that one replaces the slow-to-evaluate forward model with the fitted emulator in order to save time. This becomes particularly significant when multiple evaluations of the forward model are required, such as in Bayesian inference which typically involves calling the forward model hundreds of thousands of times. Statistical emulators have primarily been used in the past in cosmology (Auld et al. 2007; Wang et al. 2020; Rogers et al. 2019; Schmit and Pritchard 2017), but have also recently found use in astrochemistry (de Mijolla et al. 2019; Holdship et al. 2021). In our case, the forward model requires solving a coupled system of ODEs of the form given in Equation 1.13. The evaluation of the forward model can be time-consuming, especially if this has to be repeated multiple times as would be the case for

Bayesian inference. This proves to be particularly important for the analysis we do in Appendices B.1 and B.2. In this case, a statistical emulator would be particularly useful, as it can interpolate within the range of input values considered. This is computationally faster than making use of the original forward model for evaluation.

There are a number of algorithms that can be used for the purposes of emulation. One particularly popular one is the Gaussian process emulator, which has found widespread usage (Kennedy and O'Hagan 2001; Pellejero-Ibañez et al. 2020; Rogers et al. 2019). An inherent advantage is the ability of this sort of emulator to quantify the uncertainty associated with the regression. This allows for the use of acquisition functions that iteratively improve the emulator approximation by sampling points in areas of high uncertainty (Pellejero-Ibañez et al. 2020; Rogers et al. 2019). However, a disadvantage is that the emulation process scales badly as the cube of number of training points (Pellejero-Ibañez et al. 2020). This is in contrast to neural network emulators, which will be used in this Chapter. Neural networks aim to fit the relationship between the inputs and outputs of the model without considering the uncertainty of the approximation. Neural networks do not struggle as drastically with an increase in training points. A higher number of training points will ensure better model performance as the emulator, which is the reason that we elected to use neural networks.

4.3.1 Training the Emulator

In order to be able to use the emulator, it must first be trained on some data. It is important that the sampling is done in such a way that the entire parameter space is explored. One cannot simply use random uniform sampling, as each point is drawn independently of the others. This can result in the training points being clustered. This has the consequence of the emulator attempting to match the training data more in these regions, thereby introducing bias in other less well-covered regions of the parameter space. A Latin Hypercube Sampling Scheme was used (McKay et al. 1979) and implemented using the Python surrogate modelling toolbox (Bouhlel et al. 2019). As both the input and output parameters span several orders of magnitude, the emulator was trained to learn the mapping between the logarithm of these two. The training dataset spanned the prior range for each parameter. Given that a log-uniform prior between 10^{-15} and 10^{0} was used for the Bayesian inference (see Section 4.4.2 for details), this ensured that any conceivable input to the emulator from the inference was within the prior range, as outside that range

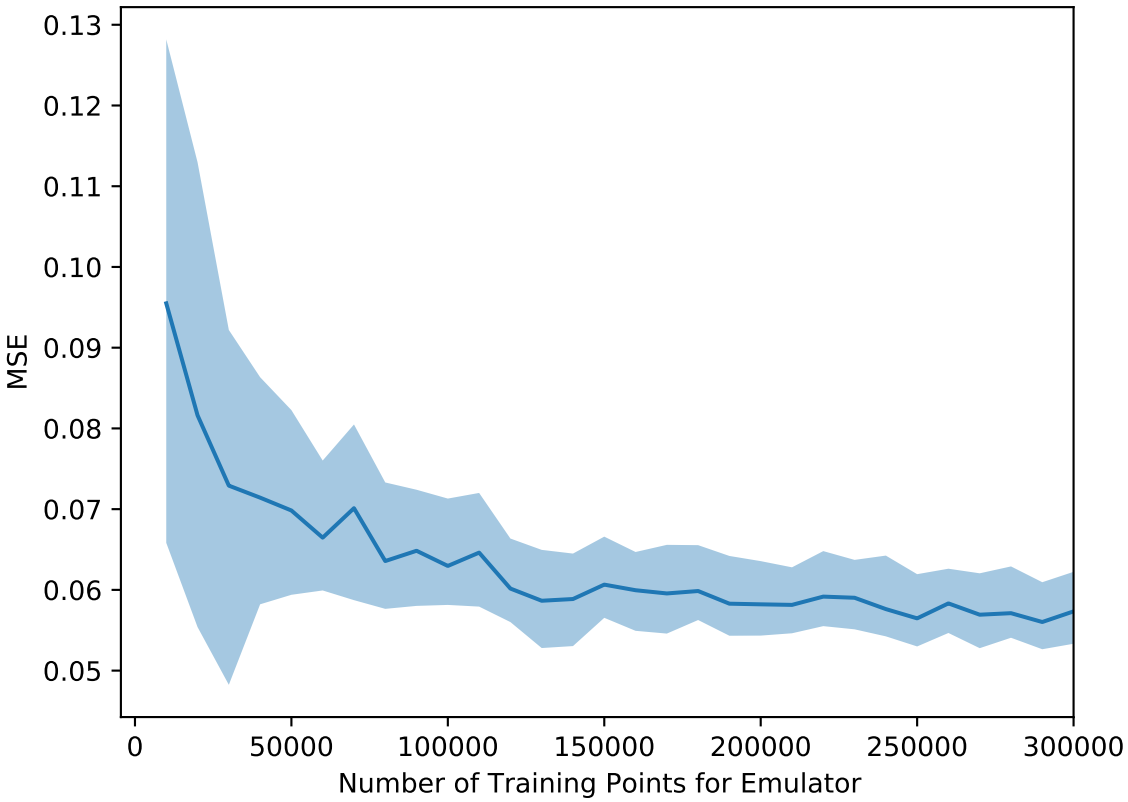

Figure 4.1: *A plot of the mean-squared error of the emulator as a function of the number of training points used to train the emulator. The shaded area represents the 95% confidence interval around the mean-squared error.*

the posterior is zero due to the prior being zero. The parameter ranges defined the range of values over which the emulator could interpolate. The emulator was not needed to extrapolate, as the range of the prior was covered.

Choosing the number of training points is a crucial parameter. It is clear that increasing the amount of training data will improve the emulator performance. However, this will also result in the time taken for training increasing. As such, a balance needs to be struck. Figure 4.1 shows the mean-squared error (MSE) on a test set as a function of the number of training points. It was found that using 150,000 training points was sufficient. By evaluating these points on a single Research Capital Infrastructure Funds (RCIF) node with 40 cores, the training time was about 30 minutes.

4.3.2 The Neural Network

In this Chapter, an artificial neural network was used as the emulator. To improve the neural network's performance, the input log-rates were scaled to lie between zero and one. A five-layer neural network was used with the three hidden layers containing 512, 256 and 128 neurons, respectively. The hyperbolic tangent was used as the activation function. The scikit-learn package was used to train the emulator (Pedregosa et al. 2011). To avoid over-fitting to the training data, the training process was terminated when the validation error stopped decreasing by at least 0.01.

4.4 Bayesian Inference

4.4.1 Introduction to Bayesian Inference

The aim of this Chapter is to deduce the reaction rates of the reactions in this network, which we represent as a vector, $\boldsymbol{\theta} = (k_1, k_2 \ ... \ k_{49})$, and use these inferred reaction rates to determine the binding energies of diffusive species. This is initially a 49-dimensional inference problem. The code used takes this vector as an input and outputs the abundances of all the species in this network, which is represented by the vector $\mathbf{Y} = (Y_1, Y_2 ... Y_{35})$. There exist measurements for the abundances of a subset of the molecules in this network. These form the data $\mathbf{d}$, which are listed in Table 4.2. We once again use Bayes' Law to determine the probability distribution of the reaction rates given the data.

4.4.2 Implementation

To obtain the posteriors of the reaction rates, a prior must be specified. As has been done previously, a log-uniform prior was chosen, so as to equally weight rates over different orders of magnitude. However, a different range is chosen compared to Holdship et al. (2018) and Chapter 3, to accommodate the fact that the reaction rates, $\boldsymbol{\theta}$, are normalised by the cloud density. Additionally, it was found in Holdship et al. (2018) and Chapter 3 that the probability density is very low in the range $10^{-30} - 10^{-15}$. As such, a log-uniform prior between 10^{-15} and 10^{0} was used.

We assume that the measurements are Gaussian based on the fact the distribution of reported measurements such as in Whittet et al. (2011) are not strongly skewed but instead are reasonably well fit by Gaussians with the parameters we include in our data

table. A Gaussian likelihood function was used:

$$P(\mathbf{d}|\boldsymbol{k}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(d_i - Y_i)^2}{2\sigma_i^2}\right), \tag{4.1}$$

where n_d is the number of observations and σ_i is the uncertainty of the ith observation. Only the species for which there are abundances are multiplied over. Table 4.2 contains species for which we have abundances with Gaussian uncertainties. Observed abundances will be referred to as constraints in this Chapter as they constrain the prior parameter space of reaction rate posteriors.

Boogert et al. (2015) also contains upper limits for the abundances of some species of interest. The upper limits for O_2, N_2, H_2O_2 and glycine are also included in Table 4.2. Equation 4.1 can be rewritten to account for these upper limits, as was done in Holdship et al. (2018).

$$P(\mathbf{d}|\boldsymbol{\theta}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\delta_i \frac{(d_i - Y_i)^2}{2\sigma_i^2}\right) (1 - S(C_i))^{1-\delta_i}, \tag{4.2}$$

where δ_i is 1 for observed species and 0 for species with upper limits. Notice that in this case that n_d is the number of observations as well as upper limits. C_i is the upper limit of that species and $S(C_i)$ is the survival function, which is defined as

$$S(C_i) = 1 - \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{C_i - Y_i}{\sigma_i^{UL}}\right)\right), \tag{4.3}$$

where erf is the error function and σ_i^{UL} is taken to be one-third of the upper limit. The value of σ_i is to account for the fact that there might be some level of uncertainty on the value of the upper limit.

In order to sample the posterior, the PyMultiNest Python package was used (Buchner, J. et al. 2014), which is a wrapper for the MultiNest package (Feroz and Hobson 2008; Feroz et al. 2009; Feroz et al. 2019), which implements nested sampling (Skilling 2006). A Python wrapper of the UCLCHEM code was created using F2Py. The input was the vector of reaction rates $\boldsymbol{\theta}$.

4.4.3 Degeneracy Problem

Before performing the inference, it is important to consider the problem in more depth. There are 49 parameters to estimate, but there are only 12 measurements. As was observed in Holdship et al. (2018) and Chapter 3, having far more parameters than constraints introduces a significant amount of degeneracy into the problem. Some rates of reactions do not influence the abundances of species with constraints. As was observed in Holdship et al. (2018), this will result in the majority of reaction rate posteriors being uniform. Additionally, many posteriors will only deviate weakly from uniformity. This was found to be the case for linear reaction chains with successive hydrogenations, such as the successive hydrogenation of CO to form methanol. The degeneracy stemmed from the fact that the reactions were tightly coupled. Provided one rate took a minimum value and acted as the rate-limiting step, the other reaction rate was free to vary above this minimum rate. The high level of degeneracy inherent to this problem meant that despite running a sampler for several weeks, it never converged.

4.4.4 Degeneracy Solution

To reduce the degeneracy of the problem, one can exploit information about the underlying grain-surface diffusion mechanism. Ultimately, the reaction rate is strongly dependent on the hopping rates of the reactant species, which, assuming the grain temperature is constant, implies that the reaction rate is set by the binding energies. Given the strong dependence of the hopping rate on the binding energy, it is clear that a small difference in the binding energy between two species will mean that the hopping rate of the more mobile species (the one with the lower binding energy) will dominate the reaction rate. In equation 1.2, this corresponds to $k_{hop}^{A} \gg k_{hop}^{B}$ and yields

$$k_{AB} = \kappa_{AB}^{final} \frac{k_{hop}^{A}}{N_{site} n_{dust}}, \tag{4.4}$$

where we see that this equation only depends on the hopping rate of species A. Recall that $\kappa_{AB}^{final} \simeq 1$ in the diffusion-limited regime.

Based on this, one can separate reactions into various classes, depending on which of their reactants is more mobile. Even though the actual values of the binding energies will differ across the literature (see Penteado et al. (2017) for a discussion on this), most works

agree on the "hierarchy" of mobility, that is which of the two species is more mobile. By making an assumption or by considering literature values, one can make a decision on which species should be treated as more mobile. In this Chapter, the more mobile species was assumed to be the one with the lower binding energy in at least two of Penteado et al. (2017), UMIST and Wakelam et al. (2017). The groupings used are shown in Table 4.3. For reactions 34 and 35, one does not expect diffusion to be the dominant reaction mechanism. However, since this is the diffusion-limited regime, one can assume that these two reactions will have the same reaction rate.

The major implication is that this now allows for the calculation of a species' binding energy. In fact, provided that species is far more mobile, one can calculate that species' binding energy. What one finds is that the reaction rates of many reactions are effectively only dependent on the binding energy of the same species. As such, the dimensionality of the problem is significantly reduced, as one simply needs to determine the binding energies of the more mobile species.

4.4.5 Deriving the Binding Energies

Binding energy values vary greatly across the literature. Their values can determine whether or not a reaction can occur efficiently via diffusion. For example in Ioppolo et al. (2020), it is stated that 10 K is too low a temperature for any species other than atomic hydrogen to diffuse. However, a statement such as this one assumes a value for the binding energy of hydrogen and that it is far lower than the binding energies for other species. While many works state the binding energy of H to be 650K, many others find that species such as O and N have comparable binding energies (Penteado et al. 2017).

Our goal is to determine the binding energies of various species. In this Chapter, we will be inferring reaction rates for the various reactions and use these to solve for the binding energies. Each reaction rate varies as a function of time, as seen in equation 1.2 due to the dependence on the total hydrogen number density. However, by multiplying by n_H on both sides, one obtains

$$k'_{AB} = k_{AB} n_H = \kappa_{AB} \frac{(k^A_{hop} + k^B_{hop})}{N_{site} \frac{n_{dust}}{n_H}}, \tag{4.5}$$

where $\frac{n_{dust}}{n_H}$ is a constant. Note now that the expression on the right-hand side only consists

of constants. This implies that k'_{AB} is constant with respect to the density of the cloud. While k'_{AB} can still be interpreted as a reaction rate, it has units of s^{-1}.

Due to the exponential dependence of the hopping rates on the binding energies, one finds that in most cases, one species dominates the reaction rate. If species A has a binding energy of, say, 500K and species B has a 10% higher binding energy, then A's hopping rate is almost 150 times greater, due to the low grain temperature of 10 K. This difference will only get larger as the binding energies under consideration increase. Hence, we can state that $k^A_{hop} \gg k^B_{hop}$ and then determine the binding energy of the species by substituting equation 1.3 into equation 4.4:

$$k'_{AB}\frac{n_{dust}N_{site}}{n_H\kappa_{AB}}\sqrt{\frac{\pi^2 m}{2k_b n_s}} = \sqrt{\frac{E_b^A}{f}}\exp\left(-\frac{E_b^A}{T_{gr}}\right), \tag{4.6}$$

where the corresponding value of $f = \frac{E_b}{E_D}$ is used, depending on the species under consideration. This equation cannot be solved analytically, so has to be solved numerically.

4.4.6 Constraints

The final component required to perform Bayesian inference is the data, which in this case would be measured abundances of species. A number of constraints for molecules in this network can be found in Boogert et al. (2015), which provides the median abundance as well as lower and upper quartile. As in Holdship et al. (2018), we assume the measurements are Gaussian-distributed, which implies the median is the mean. Additionally, the upper and lower quartiles are $0.68\,\sigma$ from the mean. Using this information, the abundances used in this Chapter are listed in Table 4.2. We combine measured molecular ice abundances from dark, quiescent cloud as well as Large Young Stellar Objects (LYSOs). We observe that the species CO, CO_2, H_2O, CH_3OH and NH_4^+ have similar abundances in quiescent clouds and LYSOs. Using this, we assume that other species, which have only been detected in LYSOs, will have broadly similar dark cloud abundances. We argue that while chemistry is expected to happen during the warm-up phase for LYSOs, this will be relatively short-lived and any abundances will likely have been built up during the cold phase of star formation. However, even though the warm-up phase will be shorter, the chemical time scales will decrease due to the reaction rate's dependence on temperature. Overall, while there is justification for using LYSO abundances for dark cloud conditions,

Species	Abundances relative to H	Source
H_2O	$(4.0 \pm 1.3) \times 10^{-5}$	Cloud
CO	$(1.2 \pm 0.8) \times 10^{-5}$	Cloud
CO_2	$(1.3 \pm 0.7) \times 10^{-5}$	Cloud
CH_3OH	$(5.2 \pm 2.4) \times 10^{-6}$	Cloud
NH_3	$(3.6 \pm 2.6) \times 10^{-6}$	LYSOs
CH_4	$(2.3 \pm 2.1) \times 10^{-6}$	LYSOs
HCOOH	$(2.4 \pm 1.3) \times 10^{-6}$	LYSOs
NH_4^+	$(3.8 \pm 1.5) \times 10^{-6}$	Cloud
O_2	$< 60 \times 10^{-6}$	Comet
N_2	$< 0.1 - 28 \times 10^{-6}$	Comet
H_2O_2	$< 0.6 - 8 \times 10^{-6}$	Comet
NH_2CH_2COOH	$< 0.1 \times 10^{-6}$	Comet

Table 4.2: *The abundances and uncertainties taken for the network adapted from Boogert et al. (2015). There were two distinct values for the upper limit on the abundance of O_2, so the higher one was selected.*

Grouping	**Reactions in Group**
Hydrogenations	1-9, 12-20, 24, 25, 30-33, 36, 39-41
Oxygenations	29, 42, 43
Nitrogenations	37, 38, 44
CO-based reactions	10, 11, 48
OH+OH	34, 35
CH3-based reactions	22, 47

Table 4.3: *The main reaction groupings, separated by the molecule that the literature suggested was more dominant. Any reaction not included in this table had its reaction rate inferred separately.*

it should be noted that we are adding additional uncertainty into our analysis.

4.5 Results

4.5.1 Highest Density Regions

Parameter estimates are typically quoted by considering the marginalised posterior distributions. The important quantities to estimate are typically the mean and variance. However, one must be careful when estimating these quantities, as depending on how broad and asymmetric the posterior space is around the maximum-posterior value, these might not be meaningful quantities. To determine useful estimators, one can choose to only consider the highest density region (HDR) of the posterior.

Species	BE 1 (K)	BE 2 (K)	Penteado (K)	Wakelam (K)	UMIST (K)
H	1099^{+33}_{-57}	1016^{+65}_{-68}	650 ± 100	650	600
O	824^{+180}_{-109}	805^{+88}_{-97}	1660 ± 60	1600	800
N	894^{+326}_{-202}	932^{+102}_{-130}	715 ± 358	720	800
C	1336^{+136}_{-160}	1361^{+124}_{-256}	715 ± 360	10000	800
CO	1009^{+158}_{-123}	1018^{+91}_{-135}	1100 ± 250	1300	1150
CH	1160^{+130}_{-240}	1107^{+228}_{-162}	590 ± 295	925	925
CH_3	1088^{+375}_{-242}	1133^{+350}_{-288}	1040 ± 500	1600	1175
CH_4	1343^{+765}_{-200}	1327^{+153}_{-139}	1250 ± 120	960	1090
H_2	1719^{+377}_{-356}	1976^{+184}_{-283}	500 ± 100	440	430
NH	1172^{+307}_{-325}	1115^{+351}_{-254}	542 ± 270	2600	2378
NCH_4	1265^{+206}_{-354}	1046^{+326}_{-222}	-	-	-
NH_2CH_3	1694^{+486}_{-355}	1581^{+544}_{-427}	-	-	-

Table 4.4: *The binding energies obtained for various species obtained through the use of Bayesian inference as well as values from Penteado et al. (2017), McElroy et al. (2013) and Wakelam et al. (2017). The first set of predicted binding energies come from performing Bayesian inference on the standard network, while the second set of predictions stem from including the dummy reaction $H + X \longrightarrow HX$. With the exception of H, most of the other binding values match at least one literature value. For most of the species, the uncertainty on the binding energy values is lower compared to the spread of literature values. No values for the binding energies of NCH_4 and NH_2CH_3 were found in the literature.*

For a probability density function $f(x)$ for some random variable X, the $100(1-a)\%$ HDR is the subset $R(f_a)$ of values in X such that

$$R(f_a) = x : f(x) \geq f_a, \tag{4.7}$$

where f_a is the largest constant that ensures that the probability of being in $R(f_a)$ is greater than 1-a (Hyndman 1996). In other words, the HDR allows one to only consider a subset of the posterior density function that has a value greater than some threshold f_a.

4.5.2 Reaction Rate Marginalised Posteriors

We find that all 14 parameter distributions are non-uniform. As such, this means that we have gained information about the entirety of our 49-D reaction network. By exploiting our knowledge of the grain-surface diffusion mechanism and assuming that reaction rates are dominated by the diffusion rates of a subset of molecules, we have been able to significantly reduce the dimensionality of our problem, therefore making it computationally tractable for the sampler. Figures 4.2 and 4.3 show the marginalised posterior distributions for the

reaction rates with the 65% HDR being the shaded regions when we use the likelihoods expressed in Equations 4.1 and 4.2. For all of the posteriors, the 65% HDR lies away from the boundaries of the uniform distribution, implying that our choice of prior was appropriate. We choose not to consider 2-D marginalised posterior distributions, due to the fact that the parameters correspond to groups of reactions as opposed to individual reactions. We consider the frequentist properties of the estimators in Appendix B.1.

There are some noticeable differences in the marginalised posterior distributions when the upper limits are included. The fact that the oxygenation and nitrogenation reaction rate distributions do not significantly change with the inclusion of the upper limits on O_2 and N_2 is surprising. One would expect that the reactions $O + O \longrightarrow O_2$ and $N + N \longrightarrow N_2$ would be the dominant formation mechanisms. As such, it is possible that the upper limits on the abundances of these species may not be constraining enough to affect the obtained posterior distributions. In Appendix B.2 we explore the distribution of the maximum-posterior binding energy as we vary the weak constraints for the aforementioned four species with upper limits. We also consider how the relative uncertainty on these four abundance measurements affects the obtained values.

We observe that the posterior for the reaction rate of hydrogenation is the most constrained in that it rules out more of the prior parameter space than any of the other posteriors do. In Chapter 3, the lower uncertainty on hydrogen's posterior was related to the size of the constraints on the species formed by hydrogenation, in particular the constraint on water, which is known to have an abundance greater than 0 at the $3.1\,\sigma$ level. It would make sense that this low level of uncertainty on the constraint drives the low uncertainty on the hydrogenation reaction rate posterior, as it penalises the likelihood function more. In the limit of the uncertainties on the molecular abundances going to zero, one would expect the the posterior distribution of the relevant reaction rate to look like a Dirac delta function.

4.5.3 Binding Energy Posteriors

The advantage of inferring the reaction rates as opposed to directly inferring the species binding energies is that the reaction rate posteriors make no assumption about the exact nature of the reaction mechanism. One can then select specific reactions which one believes occur via diffusion, thereby reducing the dimensionality of the problem. The list of species that were thought to diffuse are listed in Table 4.4. These are calculated from the reaction

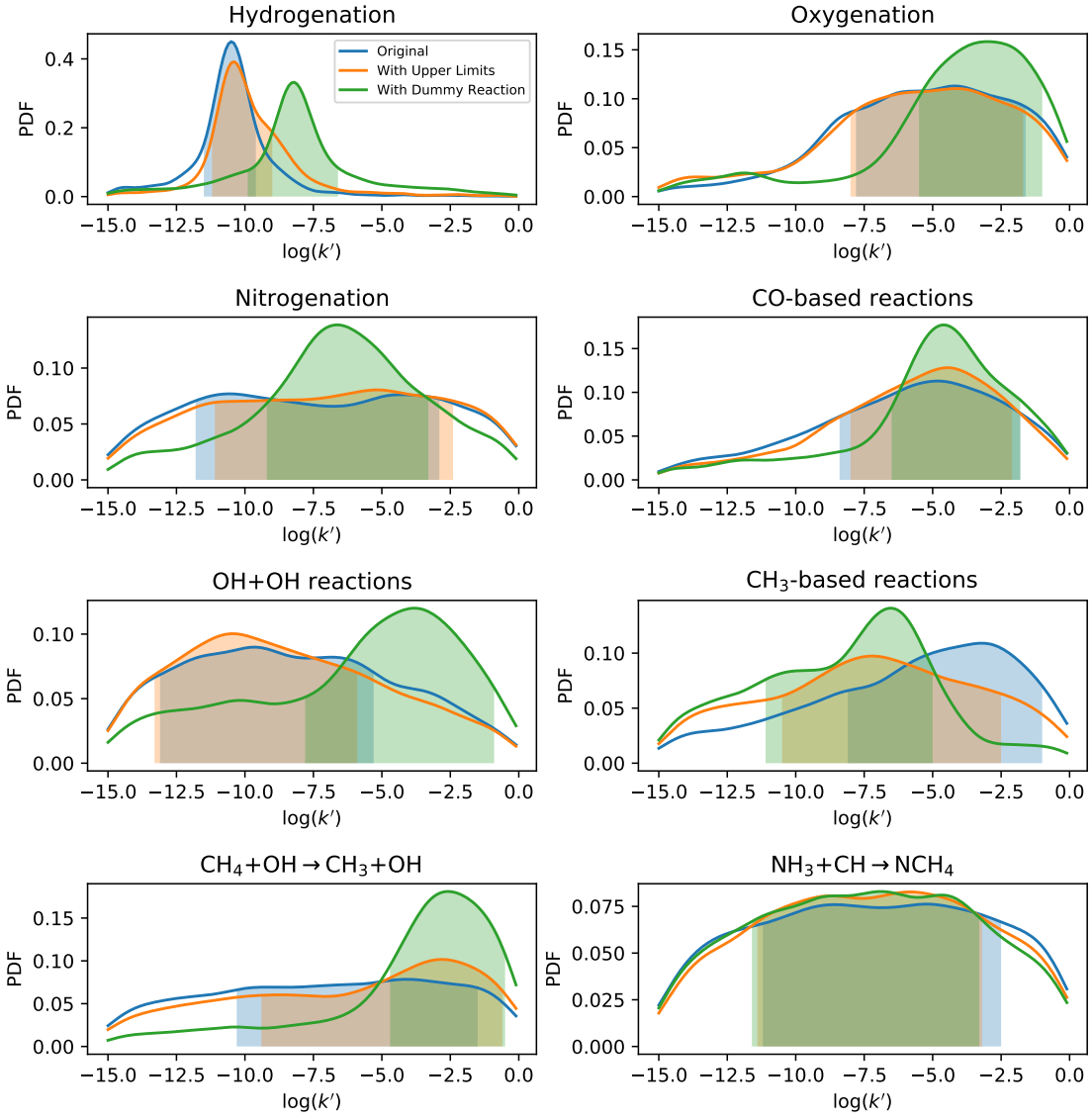

Figure 4.2: *Marginalised posterior distributions for the first eight reaction rate parameters.*

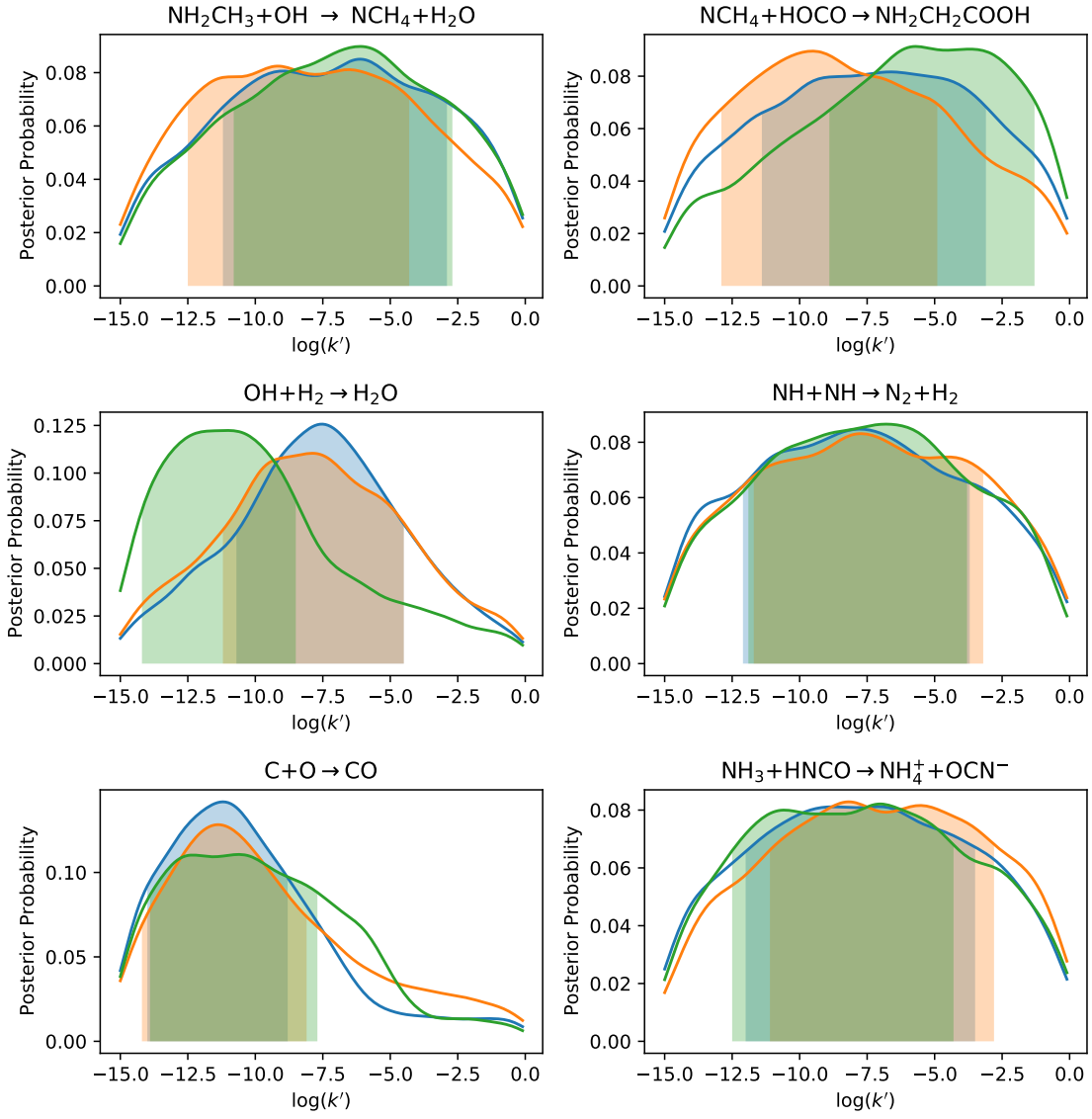

Figure 4.3: *Marginalised posterior distributions for the remaining six reaction rate parameters.*

rate posteriors by solving equation 4.6, with the posteriors shown in Figure 4.4. In Table 4.4, these binding energies are compared to the values used in McElroy et al. (2013), Penteado et al. (2017) and Wakelam et al. (2017). We make use of the posteriors obtained using Equation 4.2. This first round of inference is referred to "Binding Energy 1". We observe that there are no significant differences in the binding energy distributions for most species when we include the upper limits, with the except of CH_3, for which we see that the inclusion of the upper limits results in a significantly decreased estimated binding energy.

For most of the species for which there are literature binding energies, there is agreement with at least one literature value and the uncertainty on the values is lower than the spread of literature values. No values for the binding energies of NCH_4 and NH_2CH_3 were found in the literature. The binding energies for O and N were both found to be lower than that of H. This is surprising as the reactions of the species with H were classified as hydrogenations in Table 4.3.

However, the binding energies of H and H_2 were found to differ greatly from the literature binding energies. For the latter, this is related to the fact that there is only a single reaction that H_2 is consumed in: $OH + H_2 \longrightarrow H_2O$. The production of water is likely to be dominated by hydrogenation, due to the fact that H is so much more abundant. Furthermore, for this reaction H_2 must compete with many other molecules to react with OH. As such, the amount of water produced through this pathway is less than the amount produced through hydrogenation, which means its reaction rate will be lower than it should be. This results in the high binding energy.

For some of the species, there is a large variance in the posteriors. This can be attributed to the lack of enough constraints in the network. To demonstrate this, the upper limits on the species N_2, O_2, H_2O_2 and glycine were replaced with weak constraints that were derived by halving the upper limit with a 50% relative uncertainty. It was found that the uncertainties for most species substantially decreased. This could be attributed to the fact that most of the constrained species were formed through hydrogenation, hence why hydrogen's binding energy is so much more well-constrained. This appeared to suggest that the inclusion of these constraints of species not formed solely through hydrogenation would help reduce the variance.

It should be noted that even amongst the literature values, there is not always agreement on the values of the binding energies. The tension in the values can be attributed

to varying assumptions made about the grains, such as the ice composition. Additionally, recent work by Bovolenta et al. (2020) and Grassi et al. (2020) suggests that it might be more appropriate to consider binding energy distributions that vary as functions of the individual binding site. In our work, we have assumed that there is a single binding energy value, which implies that the grains are uniform in nature. In reality, this is unlikely to be true and will need to be accounted for in future work.

4.6 The Binding Energy of Hydrogen

We observe that, despite the high precision of the hydrogenation rate estimate, the binding energy of hydrogen is inaccurate and does not match any of the literature values within the error. We now look to address this.

The rate equation approach does not consider positional dependence of species, i.e. it assumes everything can react with everything else on the grain. This might be problematic for H, as there is so much of it, but only a small amount is on the grain mantle. This will not be considered here, as it is outside the scope of the work.

A rigorous solution would be to account for the formation and subsequent chemical desorption of H_2. This would take H out of the system and might be more physically realistic. Most of the products of hydrogenation in this network are species for which we have abundances. This means that in order to satisfy all these constraints, the hydrogenation rate posterior will be unrealistically over-constrained given that the reaction network is not complete. We can choose to add a "dummy reaction" of the form $\mathrm{H} + \mathrm{X} \longrightarrow \mathrm{HX}$ to represent all the possible reactions involving hydrogen. Recall that we first did this in Chapter 3. Notice that these will not necessarily all involve hydrogenation of grain species, but will also include desorption of the produced species. This is why the dummy reaction is not assumed to have the same reaction rate as all the hydrogenations. By leaving the reaction rate of the dummy reaction as an additional free parameter, we can increase the variance of the posterior distribution of hydrogenation and therefore its binding energy posterior.

4.6.1 Including Chemical Desorption of H_2

The energy released in the reaction of $\mathrm{H} + \mathrm{H} \longrightarrow \mathrm{H_2}$ can cause the product to desorb into the gas phase. An estimate for the fraction of H_2 released was determined in Minissale

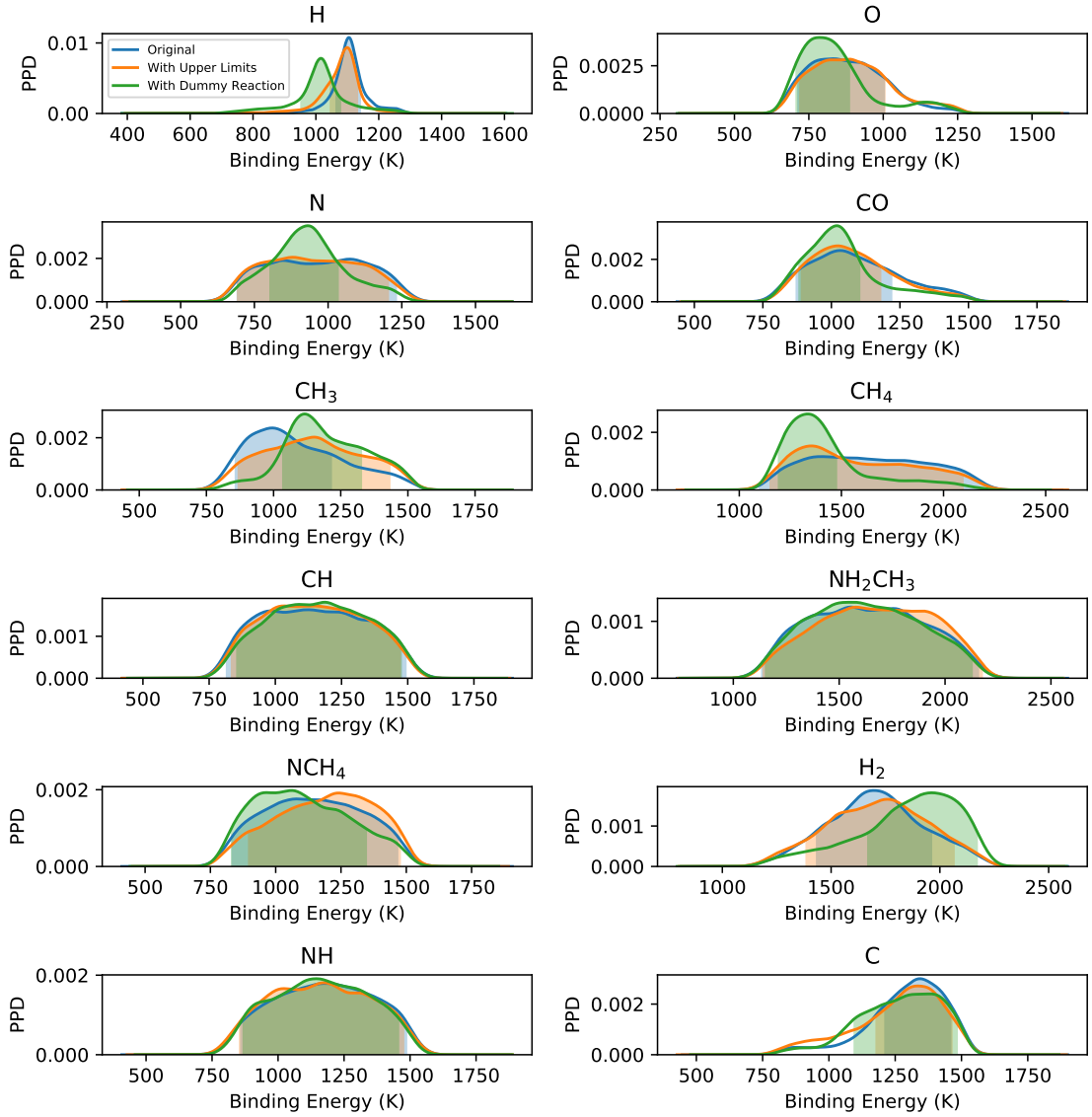

Figure 4.4: *Marginalised posterior probability distributions (PPDs) for the binding energies of the species of interest. The marginalised posterior distributions are also plotted for the case where a dummy reaction for hydrogen is included in the network.*

et al. (2016b) to be:

$$\eta_{CD} = \exp\left(-\frac{E_D N_{dof}}{\epsilon_{CD}\Delta H_R}\right), \tag{4.8}$$

where E_D is the desorption energy of the reacting species ΔH_R is the enthalpy of the reaction, $N_{dof} = 3 \times n_{atoms}$ and ϵ_{CD} is the fraction of kinetic energy the product has as it bounces off the grain surface to escape the potential well. The latter is defined as:

$$\epsilon_{CD} = \frac{(m-M)^2}{(m+M)^2}, \tag{4.9}$$

where m is the mass of the product and M is the effective mass of the grain surface, which is taken to be 120 amu in this Chapter.

For the chemical desorption of H_2, η_{CD} was found to be roughly 0.9 and this additional loss term due to desorption was included in the differential equation for H_2. However, it was found to not have a significant impact on the reaction rate and binding energy posteriors. This was a surprising result, but was attributed to the fact that there is far more H in the system than any other species, including H_2.

It is also possible that H_2 formation via this reaction is dominated by the Eley-Rideal mechanism, in which a gas-phase molecule reacts with a grain-surface species (Jin and Garrod 2020; Ruaud et al. 2015). This would indicate a weakness of the computational model used which decouples the gas and grain chemistries. While this was done in order to significantly reduce computational runtime and therefore significantly reduce the runtime for the Bayesian inference, highly abundant species such as H and H_2 are likely to not be accurately described by a decoupled model as they are likely to move between these two phases quite a bit.

4.6.2 Including a Dummy Reaction in the Network

We consider the effect of including the dummy reaction $H + X \longrightarrow HX$ on the entire network. The posteriors for the reaction rates are shown in Figures 4.2 and 4.3 with the corresponding binding energy posteriors being shown in Figure 4.4 and listed in Table 4.4 as "Binding Energy 2".

Hydrogen is a unique species, as it is so much more abundant than any other species. Combining this with its higher mobility means that it can react with a wide range of species on the grain. As such, it is important to try and accurately model its behaviour on the grain by accounting for all the possible reactions it can participate in. This dummy reaction acts as a sink for all the excess reactions by accounting for all the other reactions it can participate in.

There are some differences between the previous posteriors and the ones produced using the dummy reaction. As expected, the hydrogenation reaction rate's posterior sees an increase in its variance. This then translates to an increase in the estimated variance of hydrogen's binding energy. This is not unexpected. Since both X and HX are unconstrained, the amount of hydrogen that is consumed by this reaction is also unconstrained. Therefore, this places a significant uncertainty on the amount of hydrogen that is available for other reactions, thereby inflating the uncertainty. However, even with this increased variance, the binding energy from Penteado et al. (2017) is not matched within the error. For many of the other parameters, we observe a decrease in the variance of the posteriors through the inclusion of the dummy reaction, with CH_4 seeing its HDR size shrink significantly through this hydrogen sink. A similar observation can be made for the binding energy posteriors of O, N, CH_3 and CO.

4.7 Application to a Gas-Grain Chemical Code

In this section, we will look to use the binding energies obtained in this Chapter in a full version of the gas-grain chemical code UCLCHEM. This is a form of model-checking. To do this, we sample from the binding energy posteriors in Figure 4.4 and input these into UCLCHEM. We aim to determine how well the abundances of species of interest are recovered when the binding energies obtained in this Chapter are inputted into the full gas-grain version of UCLCHEM. Figure 4.5 shows the time series evolution of the fractional abundances of H_2O, CO, CO_2, CH_3OH, NH_3, CH_4 and HCOOH, with the 95% confidence interval for the time series also shown. These are species from Table 4.2 that have observed abundances, not simply upper limits that have been converted into weak measurements. The only species with an observed value that has not been included is NH_4^+, but this is not expected to form via diffusion. The purpose of this section is to see how well the inferred binding energy values help in recovering the abundances in a

general gas-grain network. We observe that at late times, the final abundances of most of the species become less affected by the binding energies than for earlier times, suggesting that we approach an equilibrium point at a temperature of 10 K. If we were to consider a warmer core, it is likely that our final abundances would be different.

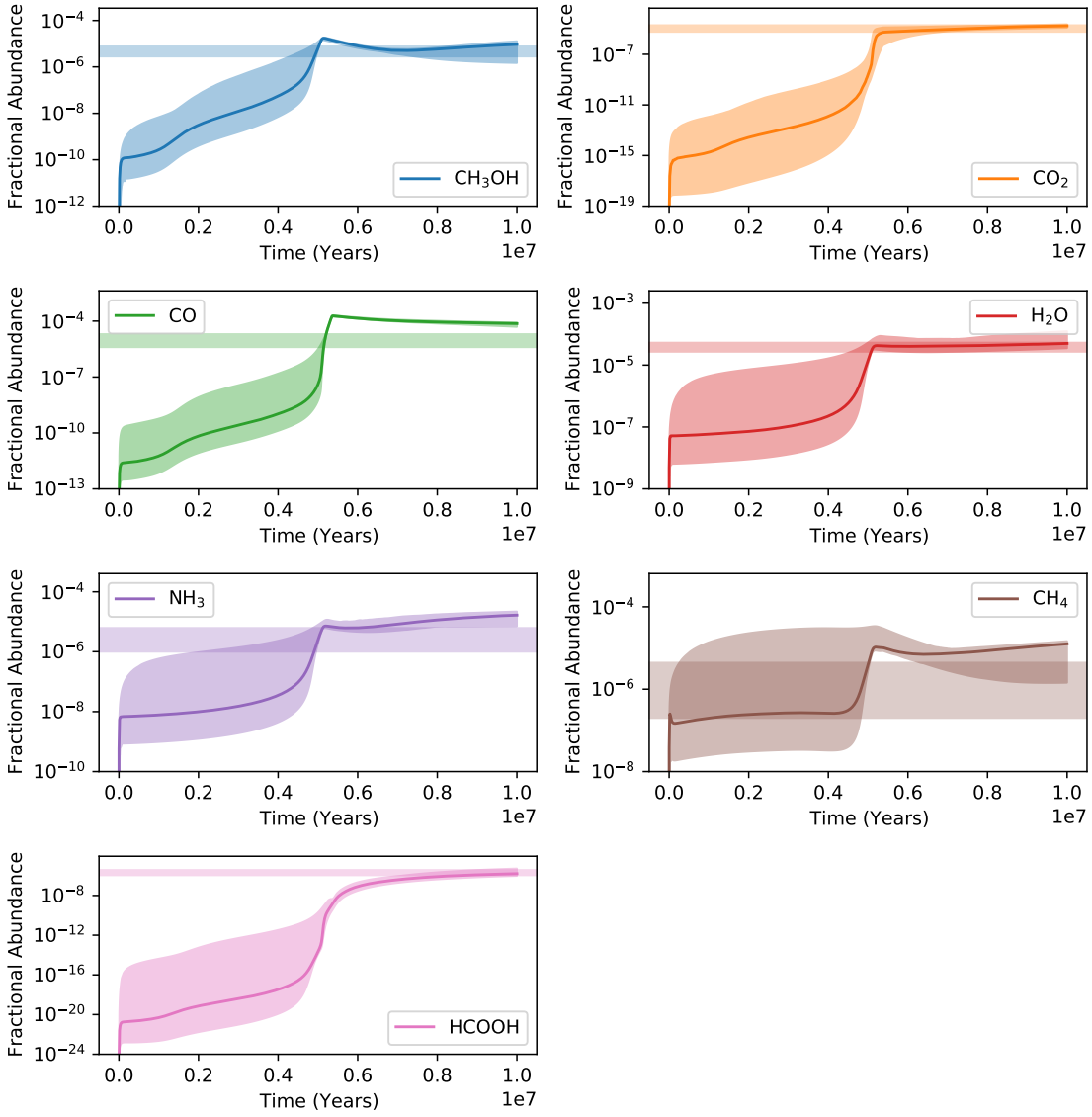

Figure 4.5: *Time series of the fractional abundances for H_2O, CO, CO_2, CH_3OH, NH_3, CH_4 and HCOOH. The binding energies for each species were sampled from the marginalised posterior distributions and inputted into the full UCLCHEM code. The horizontal shaded regions are the corresponding measured molecular abundances with their 67% confidence interval. The time series are plotted with their 95% confidence intervals.*

The final abundances of H_2O, CO_2, CH_3OH, CH_4 and HCOOH match the measured values, within the $1\,\sigma$ error. This is not the case for NH_3 and CO. The final abundance

for NH_3 is $1.2\,\sigma$ from the mean, which is a relatively small discrepancy. On the other hand, CO's final abundance is $4.6\,\sigma$ from the mean. One possible reason for this is that all the other molecules which are constrained are stable molecules that are unlikely to be depleted sufficiently at 10 K. In contrast, CO is a radical and is likely to react with other radicals. The fact that CO appears to be overproduced here suggests that the network being employed is incomplete. A more complete network would have more CO-based reactions that would lower the CO abundance. Despite, an incomplete network, uncertain elemental abundances or the gas-grain decoupling all contributing to this systematic error, the final abundance for CO is still a sensible value. Overall, knowledge of the diffusion mechanism has allowed us to not only reduce the dimensionality by grouping reactions, but also recover the observed values more precisely compared to Holdship et al. (2018), where each reaction rate was inferred separately.

However, despite the discrepancy with CO, the binding energies obtained using the decoupled code provide reasonable results when input into the full gas-grain chemical code. The next step would be to infer the binding energies directly from the full gas-grain code, though this is complicated by the fact that each evaluation of UCLCHEM, which models the full evolution of the cloud, takes of the order of a minute, suggesting a statistical emulator might need to be used to perform the inference in a reasonable amount of time.

4.8 Conclusion

In this Chapter, we used the diffusion mechanism formalism to significantly reduce the dimensionality of the inference problem, reducing the number of reaction rates to be estimated from 49 to 14. A statistical emulator was trained to further reduce the time taken per forward model evaluation. It was found that the reaction rate of many reactions is ultimately driven by the hopping rate of the more mobile species, thereby allowing us to group several reactions into classes. In doing so, the reaction rate posteriors obtained could be converted into binding energy posteriors for the corresponding mobile species driving the reaction.

This approach yielded binding energy values that were consistent with literature values. The notable exceptions were the binding energies of H and H_2, whose binding energy values were found to be significantly higher than other literature values. This discrepancy

was attributed to issues relating to the chemical model used, which decoupled the gas and grain chemistries in the interest of reducing the time taken for the evaluation of the forward model and therefore the time taken for the inference process. While chemical desorption was found to not have a significant effect on the discrepancy, using a dummy reaction of the form H+X $\longrightarrow$ HX to account for all the possible other reactions involving H somewhat reduced the discrepancy, but not enough.

This Chapter has developed an important step in estimating reaction rate parameters using Bayesian inference. It was seen that dimensionality will scale slower than the number of reactions. This reduces the number of samples that are needed to reach a stationary posterior. This approach can be trivially expanded to include more complex reaction networks. This will prove particularly important in the context of considering the formation chemistry of glycine or other amino acids. The formation routes are likely to contain a large number of diffusion reactions. However, inferring the reaction rates will not become unfeasible, due to how the dimensionality scales with the number of reactions.

In Section 4.7, we sampled from the obtained binding energy posteriors and input these binding energies into the full gas-grain version of UCLCHEM. We found some agreement between the obtained molecular abundances and the observed values. However, if one wished to infer from UCLCHEM directly, one would need to account for the fact that the inference process would take longer, on account of one evaluation of the full version of UCLCHEM taking of the order of a minute compared to the 0.5 seconds that is typical of the simplified code used in this Chapter. Future work will look to employ statistical emulation to the full version of UCLCHEM to circumvent this problem. Alternative sampling techniques that are adaptive could be utilised to this Chapter's emulator (Gramacy and Lee 2008).

Further work will need to consider larger grain-surface networks and include gas chemistry. Additionally, one should look to consider other non-diffusive grain-surface reaction mechanisms. Expressions for these reaction rates have been formulated in Jin and Garrod (2020). Including these would ensure that a more accurate picture of the chemistry would be obtained. It would also be interesting to investigate the validity of the claim that the measurements are Gaussian-distributed, as this has a direct impact on the formulation of the likelihood function. This assumption would have an impact on the posteriors obtained. There exist various methods to perform Bayesian inference without requiring the specification of a likelihood function, such as Approximate Bayesian Computation that

are the focus of current work.

Further work would also need to address the lack of sufficient abundance data. It is clear that the abundances of more species need to be known in order to better constrain the reaction rate posteriors as well as the binding energy posteriors. We propose a means of doing this in Chapter 5 by making recommendations about which molecules would reduce the variance of our posteriors. Recommendations of this sort can now be made in light of the James Webb Space Telescope's Ice Age mission (McClure et al. 2017).

This page was intentionally left blank

Chapter 5

Identifying the most constraining ice observations

The work presented in this Chapter is based on the paper Heyl et al. (2022a), in collaboration with Elena Sellentin, Jonathan Holdship and Serena Viti.

5.1 Introduction

As was discussed in Chapters 3 and 4, in order to better understand how grain surface chemistry proceeds, it is important to have good estimates of the reaction rate parameters. For grain-surface reactions, these parameters may not necessarily be the rates themselves, but rather parameters that are more specific to the reaction rate mechanism. For diffusion-based reactions, which are typically taken to be the dominant grain-surface reaction mechanism, the reaction rate parameters of relevance are the binding energies of the reacting species and reaction activation energy barriers (Hasegawa et al. 1992). Much experimental work has been done to determine these, but there are often significant disagreements, due to differing laboratory conditions (see Penteado et al. (2017) for a survey of binding energy values).

There exist a variety of methods to estimate the binding energies, ranging from experimental approaches (He et al. 2016) to density functional theory (Ferrero et al. 2020) to machine learning approaches (Villadsen et al. 2022). However, in our work to estimate

these reaction rate parameters given observed abundances, Bayesian inference is typically employed. Bayesian inference has become a ubiquitous tool in astrophysics and has recently found more use within the field of astrochemistry. Holdship et al. (2018) and the work presented in Chapter 3 have considered the rate-parameter estimation problem and have shown that the paucity of available grain-surface species abundances inhibits precise estimates of these rate parameters. The problem due to the lack of sufficiently constraining data has been somewhat ameliorated by considering the network structure in Chapter 3 or the underlying chemical mechanisms to reduce the dimensionality of the problem as was done in Chapter 4. However, it remains the case that many binding energies cannot be constrained to the point that they would be useful in chemical codes. This is clear from a survey of the literature which shows quite significant disagreements for some binding energy values (McElroy et al. 2013; Wakelam et al. 2017; Quénard et al. 2018).

Observations of the ices have typically considered the molecular vibration transitions in the infrared region (Boogert et al. 2015). A number of space telescopes such as the Infrared Space Observatory (ISO) and Spitzer have provided observations of ice band profiles that have been used to determine molecular abundances. However, until now there has been insufficient resolution of the absorption band profiles. The James Webb Space Telescope (JWST) observes in the infrared wavelength range of 0.6 - 28 μm. It provides higher spectral resolution observations of up two magnitudes, especially in the 5-8 μm range which potentially contains the vibrational modes of several molecules of interest (Boogert et al. 2015; Boogert 2016). This is particularly important as infrared spectroscopy reveals the features of various functional groups which differ by species but can have similar values (Boogert 2016). As such, having greater resolution will ensure that the various absorption band profiles can be disentangled.

Despite the fact that we massively reduced the dimensionality of our inference problem in Chapter 4, we still have posterior distributions with large variances. We wish to reduce these in order to obtain more precise estimates of binding energies. In this Chapter, we wish to provide recommendations of which species should be prioritised for future ice observations in order to reduce the uncertainties on the binding energy values. To achieve this, we make use of the "Massive Optimised Parameter Estimation and Data compression" (MOPED) algorithm (Heavens et al. 2000, 2017; Heavens et al. 2020). A key output of the MOPED algorithm is a measure of how strongly knowledge of a species ice phase abundance would constrain the binding energies.

We start by explaining the chemical code and network we will use throughout this Chapter in Section 5.2. Section 5.3 will be dedicated to explaining the approach we take in this Chapter, specifically our use of Bayesian inference and the MOPED algorithm. We follow this up in Section 5.4 by showing the results of the Bayesian inference and the MOPED algorithm as well as by discussing the observational implications of our findings. We briefly conclude in Section 5.5.

5.2 The Chemical Code and Network

5.2.1 The Chemical Code

In this Chapter, the gas-grain astrochemical code UCLCHEM (Holdship et al. 2017) was used to model the chemistry of a collapsing dark cloud. The cloud was taken to collapse isothermally at 10 K from 10^2 cm^{-3} to 10^6 cm^{-3} over a period of 5 million years. By the end of this collapse, we expect the ice phase abundances to be representative of a dark cloud. The code utilises the grain-surface diffusion mechanism described in Section 1.3.1.

Overall, we find that Equations 1.2-1.9 show that the key quantities are ν_0, k_{hop}^X, E_b and E_A. The first three are all functions of the binding energies of the reacting species, indicating the binding energies are the crucial parameters. We assume that the activation energies in Equation 1.5 are well-known. This is reasonable, as these should be independent of the ice composition (unlike the binding energies) and can be determined theoretically or experimentally. Many of the reactions would also be expected to have zero activation energy as they are radical-radical reactions (Quénard et al. 2018).

5.2.2 The Chemical Network

The chemical network consists of a gas-phase network taken from UMIST12 (McElroy et al. 2013) and a grain-surface network based on Quénard et al. (2018) and expanded to include the reactions from Garrod et al. (2008); Minissale et al. (2016b); Quan et al. (2010); Fedoseev et al. (2016); Belloche et al. (2017); Song and Kästner (2016); Garrod and Herbst (2006).

We believe the gas phase network is comprehensive and sufficiently accurate that any deficiencies in the network will not have a great effect on our results. The gas-phase network was benchmarked against observations in McElroy et al. (2013). The abundances of species freezing out from the gas phase are likely to be approximately correct and we

therefore only need to be concerned by the accuracy and completeness of the grain surface network. We operate under the assumption that the gas-phase network is complete.

Our grain surface network is less comprehensive but we argue it is sufficient to reproduce the abundance of major species, given the results of Makrymallis and Viti (2014), Holdship et al. (2018) and Chapters 3 and 4 which used smaller networks. The network includes the freeze out of all species, hydrogenation reactions of all species up to their saturated forms, and radical-radical reactions that have been shown to be efficient in laboratory experiments, as well as other diffusion reactions from the literature (see above). By including all reactions known to be the main routes through which species like H_2O and CH_3OH are formed on the grain surfaces, our network is sufficient to produce accurate ice phase abundances of these species. Therefore, we can properly predict how important the binding energies of those species are to the surface chemistry.

5.3 Analytical Approach

5.3.1 Parameters

The aim of this Chapter is to determine the binding energies of the chemically reactive species. While it would be ideal to determine the binding energies of all species in the network, the reality of the situation is that this is not strictly necessary. In Chapter 4, it was demonstrated that at 10 K, a moderate difference in binding energies between two species results in a significant difference in reaction rates. As such, one can significantly reduce the dimensionality of the problem one is trying to solve by only considering the most diffusive species. These are those species that will be the more reactive species with the greater hopping frequency for at least one reaction in the network. The more reactive species were determined by considering the literature. Even though there is widespread disagreement about the values of the binding energies, there is less disagreement about the hierarchy of binding energy values. This can be seen by considering the values given in Wakelam et al. (2017), McElroy et al. (2013) and Penteado et al. (2017). For reactions where the literature was not definitive in specifying which species had the lower binding energy, both species' binding energies were included as parameters. The binding energies we considered as parameters were the binding energies of H, H_2, C, CH, N, CH_3, NH, CH_4 and O.

Species	Abundances relative to H	Source
H_2O	$(4.0 \pm 1.3) \times 10^{-5}$	Cloud
CO	$(1.2 \pm 0.8) \times 10^{-5}$	Cloud
CO_2	$(1.3 \pm 0.7) \times 10^{-5}$	Cloud
CH_3OH	$(5.2 \pm 2.4) \times 10^{-6}$	Cloud
NH_3	$(3.6 \pm 2.6) \times 10^{-6}$	LYSOs
CH_4	$(2.3 \pm 2.1) \times 10^{-6}$	LYSOs
HCOOH	$(2.4 \pm 1.3) \times 10^{-6}$	LYSOs

Table 5.1: *The abundances and uncertainties taken for the network adapted from Boogert et al. (2015).*

5.3.2 Bayesian Inference

Introduction to Bayesian Inference

The goal is to estimate the binding energies of the most diffusive species in this network. We represent these parameters of interest as a vector, $\boldsymbol{\theta} = (\mathrm{E}_{b,H}, \mathrm{E}_{b,H_2}, \mathrm{E}_{b,C}, \mathrm{E}_{b,CH}, \mathrm{E}_{b,N}, \mathrm{E}_{b,CH_3}, \mathrm{E}_{b,NH}, \mathrm{E}_{b,CH_4}, \mathrm{E}_{b,O})$. UCLCHEM was modified so that it took these values as an input and output all the final abundances of grain-surface abundances. We represent the 72 grain-surface abundances as a vector $\mathbf{Y} = (Y_1, Y_2 ... Y_{72})$. The mapping between $\boldsymbol{\theta}$ and $\mathbf{Y}$ is simply UCLCHEM and we can write this as $\mathbf{Y} = f(\boldsymbol{\theta})$.

In order to solve the inverse problem, we require abundance measurements of grain-surface species, $\mathbf{d}$. These are listed in Table 5.1. These are taken from Boogert et al. (2015). These form the data $\mathbf{d}$ in Equation 1.14 that we use for the Bayesian inference as described in Section 1.5.

Implementation

The prior for all binding energies was specified as uniform distribution between 400 K and 2000 K. The abundance measurements in Table 5.1 were assumed to be Gaussian which allowed for the specification of a Gaussian likelihood function:

$$P(\mathbf{d}|\mathbf{E}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(d_i - Y_i)^2}{2\sigma_i^2}\right), \tag{5.1}$$

where n_d is the number of observations and σ_i is the uncertainty of the ith observation. Only the species for which there are abundances are indexed over.

The UltraNest Python package (Buchner 2021) was used for the Bayesian inference,

which is based on the MLFriends algorithm (Buchner 2016, 2019). The package also outputs the maximum likelihood-estimator, $\boldsymbol{\theta_{ML}}$. We will use this later for the MOPED algorithm.

5.3.3 The MOPED Algorithm

The aim of the MOPED algorithm is to determine which of the M species in our chemical network need to be prioritised for future ice observations in order to best constrain the posteriors for our p parameters. In our situation, $p = 9$ and $M = 72$. In other words, we wish to determine which species will provide us with the most information upon its detection.

Recall that we wish to determine a set of parameters $\boldsymbol{\theta}$. The species that are found to be important may include the species already listed in Table 5.1, in which case we would aim to improve the uncertainties surrounding their values. However, it is also possible that we would need to detect species that have not been detected yet.

All of our future measurements will be have some instrumental uncertainty. For our purposes, we assume the uncertainty on each measurement will be the same. We define a covariance matrix to summarise this: $\mathbf{C} = diag(\sigma_1^2, \sigma_2^2, ...\sigma_M^2)$. By operating under this assumption that we can measure any species to the same level of abundance uncertainty, we are aiming to determine which species would be the most useful to detect. In general, it might be the case that different species have different levels of uncertainty.

It is likely that some species will be significantly more impactful in providing information about the parameters of interest. As such, we need to identify the species in question. To this end, we will use a filtering technique developed by Heavens et al. (2000, 2017); Heavens et al. (2020) who propose using a linear combination of the final abundances of network, $\boldsymbol{Y}$, to compress data points. Such a compression would be of the form:

$$c_\alpha = \boldsymbol{b_\alpha^T Y}, \tag{5.2}$$

where α ranges from 1 to p and $\boldsymbol{b_\alpha}$ is a set of orthonormal linear filters, such that each one contains as much information about that parameter that is not contained in any other $\boldsymbol{b_\alpha}$. $\boldsymbol{Y}$ represents a vector containing the final abundances for some arbitrary value of $\boldsymbol{\theta}$. As a fiducial model, we typically take $\boldsymbol{\theta} = \boldsymbol{\theta_{ML}}$, which we can determine using the

Bayesian inference discussed in Section 5.3.2. Using the maximum-likelihood parameters as a fiducial model has been found to be sufficient (Heavens et al. 2000, 2017). The value of each c_α will ultimately be more strongly influenced by the components $\boldsymbol{b_\alpha}$ that are larger in magnitude. As there is one species for each component, this means that if a component has a greater magnitude then it contains more information about that parameter.

The vectors $\boldsymbol{b_\alpha}$ are given by

$$\mathbf{b}_1 = \frac{\mathbf{C}^{-1}\mathbf{Y}_{,1}}{\sqrt{\mathbf{Y}_{,1}^T\,\mathbf{C}^{-1}\mathbf{Y}_{,1}}} \tag{5.3}$$

and

$$\mathbf{b}_\alpha = \frac{\mathbf{C}^{-1}\mathbf{Y}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\mathbf{Y}_{,\alpha}^T\mathbf{b}_{,\beta})\mathbf{b}_{,\beta}}{\sqrt{\mathbf{Y}_{,\alpha}^T\mathbf{C}^{-1}\mathbf{Y}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\mathbf{Y}_{,\alpha}^T\,\mathbf{b}_{,\beta})^2}}, \tag{5.4}$$

where $\mathbf{Y}_{,\alpha}$ is the partial derivative of $\mathbf{Y}$ with respect to the parameter α. The equations for b_α were derived in Heavens et al. (2000) through a Lagrange multiplier procedure. The iterative process of determining each linear filter $\boldsymbol{b_\alpha}$ from previous ones is akin to the Gram-Schmidt orthogonalisation. This ensures that all the filters are orthonormal, that is

$$\boldsymbol{b_\alpha^T C b_\beta} = \delta_{\alpha\beta}, \tag{5.5}$$

which is important because it means that all the filter vectors are uncorrelated. Note also that each component of b_α is weighted towards species which are low in noise, as measured by the inverse covariance matrix, as well as species with a greater impact on the parameter, as determined by the values in $\mathbf{Y}_{,\alpha}$.

Ultimately, we find that vector of abundances of all species x which has dimensionality M has been reduced to p numbers, where $p < M$. This data compression is lossless, which means the same information is included in the p values of c_α. This was originally stated in Tegmark et al. (1997) and proven in Heavens et al. (2000).

Recall that the magnitude of each component of $\boldsymbol{b_\alpha}$ gives a weighting for that species'

influence on the parameter α. To determine the best species to prioritise detection for, we simply add the absolute values of the components of $\boldsymbol{b_\alpha}$ for species across all α. That is, we perform the sum over our linear filters

$$\sum_{\alpha=1}^{p}[|b_\alpha^1|, |b_\alpha^2|..., |b_\alpha^M|]. \tag{5.6}$$

We now have a "filter sum" for each of the M species in our network. We can rank the species by their filter sum in order to determine which ones have the greatest impact on our parameters.

5.4 Results

5.4.1 Results of the Bayesian Inference

Figure 5.1 shows the marginalised posterior distributions for the binding energies of interest. The marginalised prior distribution is also plotted for comparison. It is clear that, with the exception of atomic hydrogen's binding energy, the marginalised posterior distributions differ very little from the prior suggesting a lack of sufficiently constraining data. It is for this reason that we now use the MOPED algorithm to identify species we need to detect to better constrain our posterior distributions.

5.4.2 Using MOPED

We now look to use the MOPED algorithm to allow us to make predictions about which grain-surface species need to be detected in order to better constrain the posterior distribution. The maximum-likelihood estimate (MLE) from the inference was taken and partial derivatives taken around this point. It was found that near the MLE the partial derivatives of $\mathbf{Y}$ with respect to the binding energies of C, NH, CH_4 and O were equal to the zero vector. This implies that for binding energies near the MLE, the reaction rates of the network are not sensitive to changes in the binding energies of these species. As such, these parameters were not included when calculating the filter values in the MOPED algorithm.

Figure 5.2 shows the sum of the filters for all grain-surface species. The greater the filter sum, the more important it is to detect that molecule. Additionally, one must also

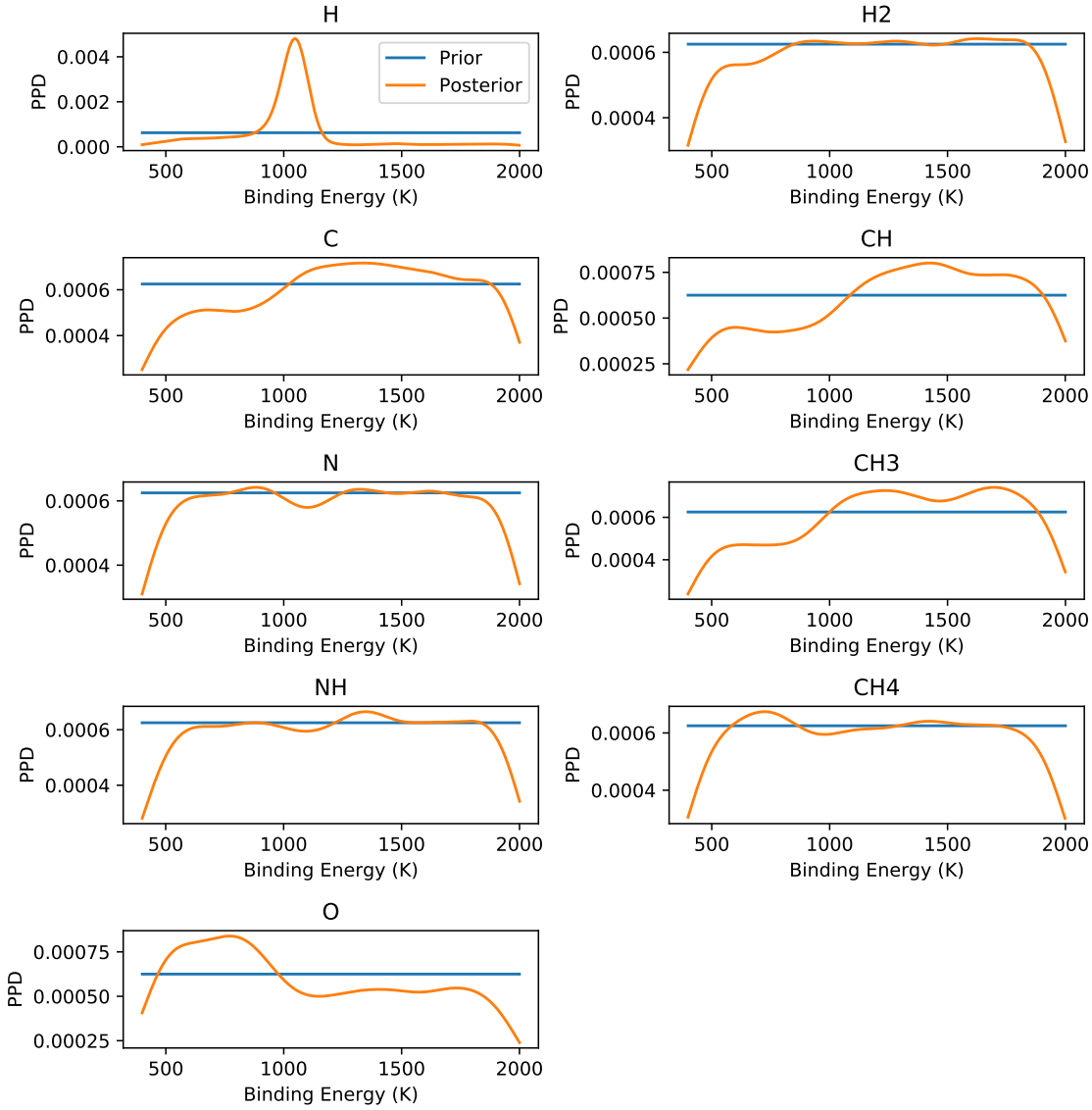

Figure 5.1: *Marginalised posterior distributions of the binding energies of the diffusive species of interest. Also plotted is the prior distribution on the binding energies. With the exception of H, most binding energy distributions differ very little from the prior distribution. This is due to the lack of enough sufficiently constraining data. This motivates the need for further ice observations to reduce the variance of the distributions.*

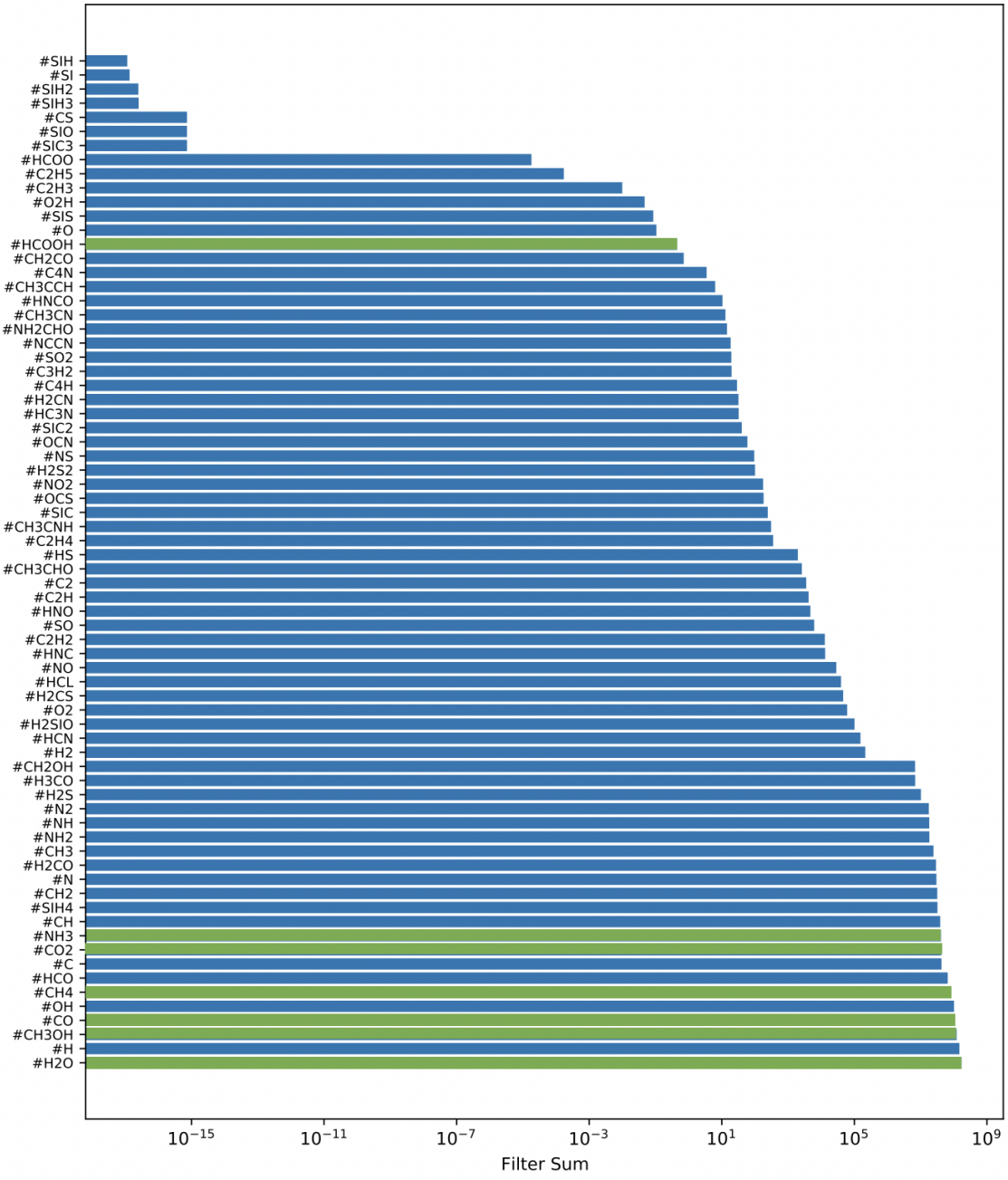

Figure 5.2: *Bar chart showing the filter sums for each species in ascending order. Species with a larger filter sum should be prioritised for detection. Species with green bars are previously detected species. Many of the species we observe are the intermediate species formed during the creation of the saturated species in Table 5.1. This indicates that understanding these intermediate products is essential to better constraining the binding energies of interest. We also note that many of the highest-ranked species have already been detected. This suggests that future observations should aim to improve the level of precision of these abundance measurements.*

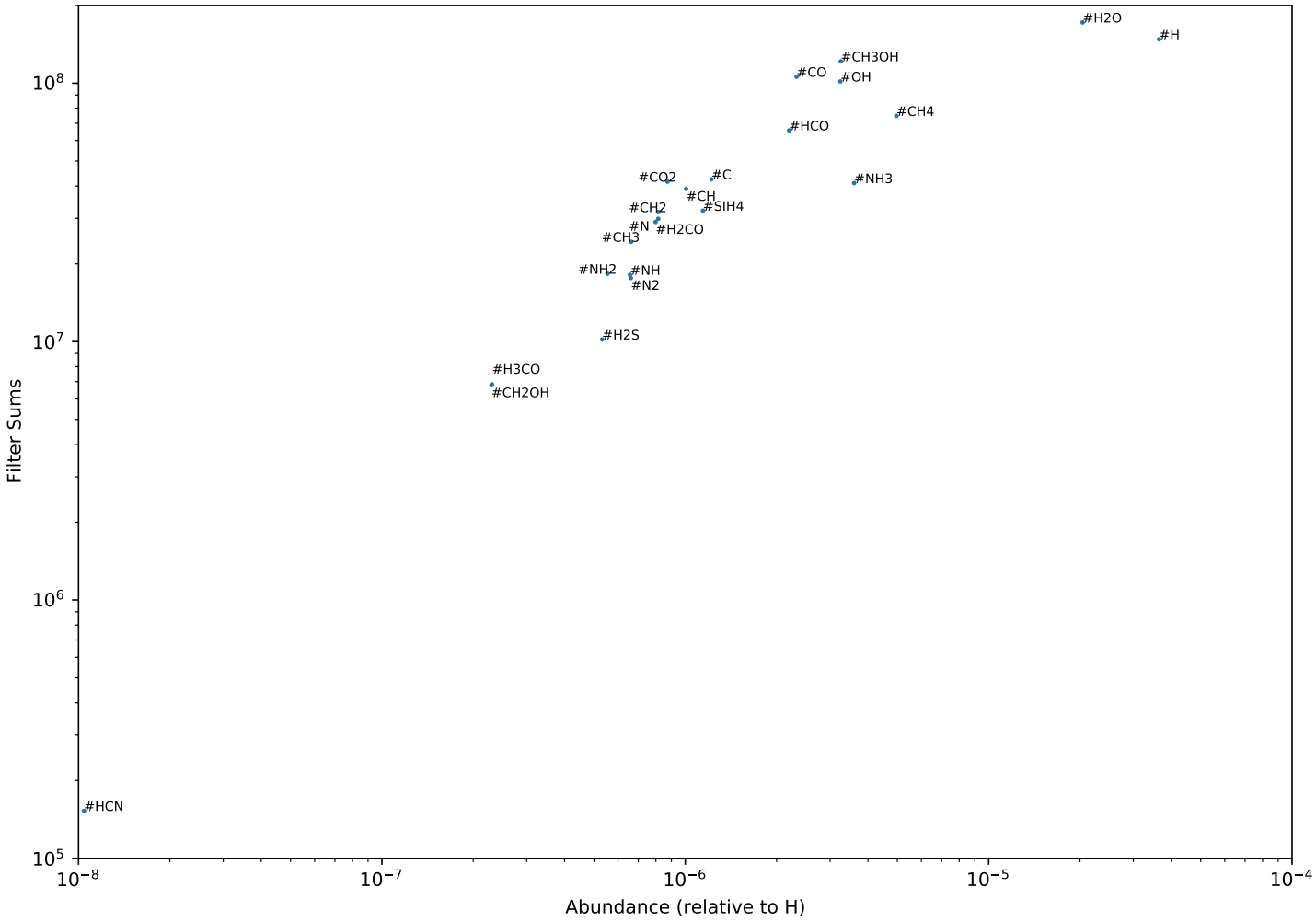

Figure 5.3: *Scatter plot depicting filter sum against the predicted abundances when the MLE for binding energies are inserted into UCLCHEM. Given constraints on instrumental uncertainties, we should look to prioritise species that are not only important, as determined by their filter sums, but that can also be realistically detected. These include saturated species such as #CH4, #NH3, #CO2 and #H2O, but also their precursors. All abundances are relative to gas-phase atomic H.*

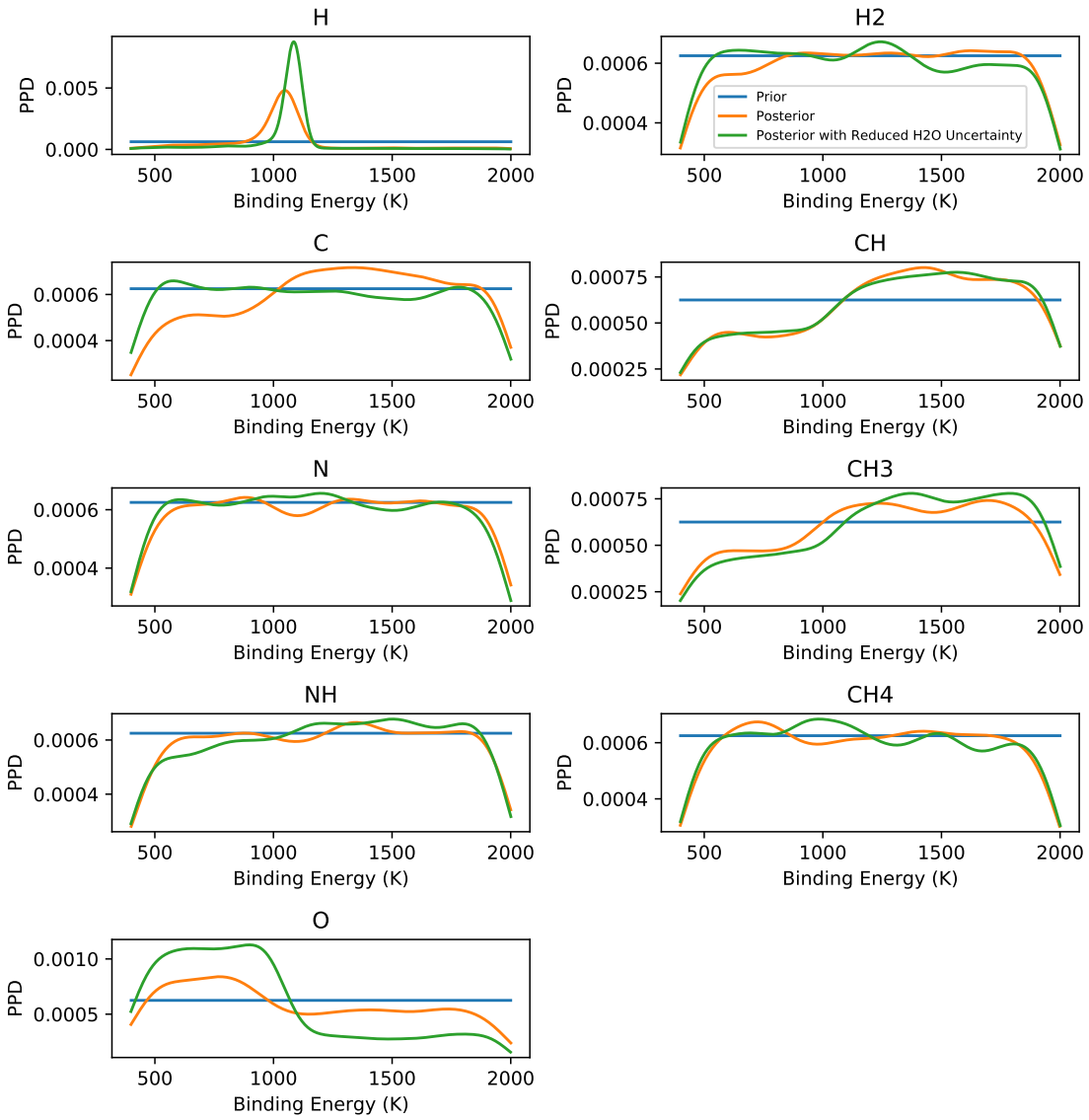

Figure 5.4: *Marginalised posterior distributions of the binding energies of the diffusive species of interest. We also plot the prior distribution and the posterior distributions when when the uncertainty on water's abundance is reduced to* 10^{-6}*. We observe that this has a significant effect on the marginalised posterior distributions of H and O, indicating that there is promise in improving the abundance measurements for species that have already been detected.*

consider the likely abundance of each species, as the species will only be observable in the ices if its abundance is above some minimum threshold. We therefore believe that future ice observations should prioritise species that have a high filter sum as well as a high abundance. In order to provide estimates of the abundances, we inserted the maximum-likelihood estimator values for the binding energy, $\boldsymbol{\theta_{ML}}$, into UCLCHEM and obtained the fitted abundances for all the species. Figure 5.3 is a scatter plot of the filter sum values against the abundances for each species. From this plot, we are able to identify high-importance species that are also likely to be detectable in the ices. However, one needs to also account for which species are realistic targets from a chemical point of view. This is discussed in the next subsection.

5.4.3 Observational Implications

The MOPED analysis has resulted in a clear ranking of which species should be targeted in future ice observations. This ranking is shown in Figure 5.2. Of course we note that many of these species have very low abundances and others are difficult to detect in absorption. Diatomic molecules, atomic species and all radicals except CO will be neglected in our considerations of which species to consider.

We briefly return to the issue of the network's reliability which was first discussed in Section 5.2.2. Whilst one can be confident in the abundances of CH_4, H_2CO, CH_3OH and H_2O as their networks are experimentally derived (Fuchs et al. 2009; Ioppolo et al. 2011a; Chuang et al. 2016; Qasim et al. 2020), other species should be viewed more skeptically. This is particularly the case for sulphur. Many works indicate sulfur may primarily be locked in other forms (Vidal et al. 2017; Woods et al. 2015). It may be that the sulphur reaction network is incomplete. Most concerning is H_2S which the model suggests is the primary sulphur reservoir on the grains. Observations of ices have never detected H_2S but have instead provided upper limits of $\sim 10^{-6}$(Boogert et al. 2015). The most likely value of the H_2S abundance derived here is lower than this limit and so it may be correct. However, there are other species in the network such as CS whose surface chemistry is not well-understood (Woods et al. 2015). Taking this into consideration, it could be argued that observers should instead target species such as H_2CO or HCN which have similar filter sums and more reliable networks despite their lower predicted abundances.

There is much to be gained from obtaining more precise measurements for the abundances of species listed in Table 5.1. All of these species except for HCOOH and NH_4^+

have high filter sums and high abundances in the fitted model. However, the uncertainties on the measured abundances are often 50% of the measured value. Our MOPED analysis shows that it would actually be much more valuable to determine these abundances to a smaller degree of uncertainty than it would be to measure the abundance of new species. To demonstrate, the effect of reducing the uncertainties on the abundances, we redid the Bayesian analysis, but reduced the uncertainty on water's abundance to 10^{-6}. Figure 5.4 shows the resulting binding energy posteriors. We observe significant changes in the posterior distributions for H and O. This suggests that there is much promise in improving the measured ice abundances for those molecules. Many of the absorption band profiles for these species are in the wavelength range of JWST, but especially in the 5-8 μm range that will have higher resolution compared to Spitzer (Boogert et al. 2015). This is promising as it is certain that H_2O and the other abundant species can be observed and telescope time simply needs to be dedicated to further constraining their abundances.

The infrared absorption profile of HCN has been studied recently in a laboratory setting (Gerakines et al. 2022). Values for selected IR absorptions of amorphous HCN at 10 K were given including the C-H stretch (3.19 μm), the C≡N stretch (4.75 μm) and the HCN bend (12.12 μm). These as well as the combination and overtone features are well within the range of wavelengths that JWST will consider. As such, this would be a viable target molecule.

While there might be some uncertainties relating to the sulphur network, H_2S has indeed a high fitted abundance as well as a high filter sum, hence it could potentially remain a target. There currently only exists an upper limit for the abundance of H_2S which was noted in Smith (1991). This Chapter identified a an S-H stretch mode at 3.925 μm, with Fathe et al. (2006) identifying an S-H stretching overtone mode at 1.982 μm.

SiH_4 is known to have several modes in the range 2.21 - 11.32 μm range (Kaiser and Osamura 2005a,b). These are all within the range that will be considered by JWST.

H_2CO has its C=O stretching mode at around 5.8 μm, but this region is also host to other species with a C=O bond such as acetaldehyde, formic acid and formamide (Keane et al. 2001; Terwisscha van Scheltinga et al. 2021). It is thought to have another feature at 3.46 μm, which is however considerably weaker (Keane et al. 2001). It is for this reason that JWST's increased resolution in the 5-8 μm region would prove useful in separating out the various components.

5.5 Conclusion

In this Chapter, we have utilised the MOPED algorithm to identify the species that would best constrain binding energies. Bayesian inference was found to result in poorly-constrained marginalised posterior distributions for the binding energies. This was due to the lack of enough sufficiently constraining data. The MOPED algorithm allowed us to determine which ice species should be prioritised for future ice observations in such a way that they would further constrain the posteriors. By then considering which species in the fitted model have the highest filter sums as well as the largest abundances, we come up with a list of species that should be targeted. These species are H_2O, CO_2, NH_3, CH_4, CO, CH_3OH, H_2CO, HCN, H_2S. While some of these species have not been detected, some of them have, which suggests that more precise measurements of these species is necessary. We also comment on which features of each species are likely to appear in the wavelength range considered by JWST.

There are some limitations to this Chapter. While our chemical network is for the most part reliable and reflects the current understanding in the literature, there are still some uncertainties relating to particular species, such as sulphur. As such, if detecting sulphur species were a priority for future observations, then more work would need to be done to be completely confident of the sulphur network.

Finally, one assumption that is made is that any species that will be detected will have the same level of uncertainty. This might not necessarily be true. The MOPED algorithm will favour species that have a strong dependence on the parameters, but also those which are low in variance. We have made use of the former, but not the latter in this Chapter. For now, the results of this Chapter are a proof-of-concept of the utility of the MOPED algorithm for this task.

This page was intentionally left blank

Chapter 6

Interpretable Machine Learning in Astrochemistry

The work presented in this Chapter is based on the paper Heyl et al. (2023c), in collaboration with Joshua Butterworth and Serena Viti.

6.1 Introduction

Modelling the interstellar medium and star formation is often a complex matter. This is normally done using computational codes that take in a number of physical parameters and use these to integrate the system of coupled ordinary differential equations (ODEs) that represent a chemical network (Taquet et al. 2012; Ruaud et al. 2016; Holdship et al. 2017). However, due to the non-linear nature of the chemistry, it is often unclear what the exact relationship is between the initial parameters and the output chemical abundances of the molecules of interest. This is often complicated by the fact that the various parameters have differing effects on the output abundances for different ranges.

It has been customary in astrochemistry to consider grids of models in which the various parameters are varied (Taquet et al. 2012; Tunnard and Greve 2016; Viti 2017; Bianchi et al. 2019; James et al. 2020; Holdship and Viti 2022). The time-consuming and computationally expensive nature of many computational codes often limits the total number of model evaluations possible. This makes drawing conclusions about the importance of

various parameters difficult. In this Chapter, we look to address both of these issues. We make use of SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) to help improve our understanding of a chemical code. SHAP provides us with a means of understanding why a machine learning model outputs a particular value. By considering various combinations of inputs and outputs, these techniques will tell us what the relationship is. This has found use in astrophysics recently (Machado Poletti Valle et al. 2021; Ansari et al. 2022) in the context of interpreting the outputs of machine learning models.

To improve the efficiency of this process, we employ statistical emulation. The process of statistical emulation involves fitting a statistical function to model the relationship between the inputs and outputs of a forward model (Grow and Hilton 2018). A significant amount of work has been done in recent years in applying statistical emulation to astrochemistry. de Mijolla et al. (2019) used a feed-forward neural network to accelerate the Bayesian inference process, while Grassi et al. (2011) used these to accelerate the forward modelling. Branca and Pallottini (2023) considered how a physics-informed neural network could be used to reduce the computational cost of predicting the time evolution of chemistry. Holdship et al. (2021) utilised autoencoders to model temperature and abundance time evolution.

We adopt the approach taken by de Mijolla et al. (2019) and Grassi et al. (2011) in this Chapter by using an emulator to simulate the final outputs of a chemical code and then evaluate these a number of times for the purposes of the machine learning interpretability algorithm. Work has been done in this area to simplify chemical networks to improve interpretability (Hoffmann et al. 2019; Grassi et al. 2022), but this is not an approach we wish to consider. Instead we build on the work done in de Mijolla et al. (2019) and look to use the interpretability techniques on these emulators. The purpose of using an emulator is that it accurately predicts the output of the forward model it is emulating in a fraction of the time. Furthermore, if the emulator is an accurate approximation for the forward model output, then it stands to reason that it accurately captures the mapping between the input parameters and the output. By using machine learning interpretability algorithms, we can identify these.

In Section 6.2, we introduce the chemical code that we will be looking to emulate. In Section 6.3 we introduce statistical emulation and machine learning interpretability. Section 6.4 is dedicated to discussing the results of the analysis.

6.2 The Chemical Code and Network

In this Chapter, we use the open source publicly available time-dependent astrochemical code UCLCHEM (Holdship et al. 2017). This astrochemical code has been developed with several updates (Viti et al. 2004; Roberts et al. 2007; Holdship et al. 2017). UCLCHEM is a time-dependent gas-grain astrochemical code. It utilises a rate equation approach to modelling the abundances of the gas phase and surface species. The initial elemental abundances are listed in Table C.1. The default values in the code for the radiation field and the cosmic ray ionisation rate are $\psi = 1$ Habing and $\zeta = 1.3 \times 10^{-17}$ s^{-1}. Radiation is attenuated by the visual extinction. Extensive documentation on the inner workings of UCLCHEM can be found on the GitHub page[1].

In this Chapter, we use UCLCHEM in two phases of modelling. Phase 1 corresponds to the isothermal gravitational collapse of a diffuse gas cloud modelled as a Bonnor-Ebert sphere. However, this stops once the internal pressure begins to balance out the gravitational pressure. This increase in internal pressure is accompanied by an increase in temperature which is when Phase 2 begins, which models a protostar. At this point, the temperature continues to increase and grain-surface species begin to evaporate as the temperatures near their respective evaporation temperatures.

In Phase 1, the gas cloud collapses isothermally at 10 K from 100 cm^{-3} to some final density, which is left as a free parameter. Phase 2 starts off at this density and begins to heat up. It is Phase 2 that has a number of physical parameters that can be varied in order to model various star-forming scenarios. There are a number of free parameters that we vary in this Chapter, which are the same as in de Mijolla et al. (2019). These are:

- Final density of Phase 1 or initial density of Phase 2 (cm^{-3})
- Metallicity (a scaling factor of all abundances)
- Radiation Field (Habing)
- Cosmic ray ionisation rate (in units of 1.3x10^{-17} s^{-1})
- Final Temperature of Phase 2 (in Kelvin)

The ranges over which we vary the parameters are summarised in Table 6.1.

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

Parameter ranges				
Parameter	Minimum	Maximum	Unit	Scale
Density (n)	10^4	10^7	cm^{-3}	Logarithmic
Cosmic Ray Ionisation Rate (ζ)	1	1000	$1.3x10^{-17}$ s^{-1}	Logarithmic
Temperature (T)	10	200	K	Linear
Metallicity (m_z)	0	2	Z	Linear
Radiation Field (ψ)	1	10^3	Habing	Logarithmic

Table 6.1: *The range of values used for each parameter as well as their units and scales. In the context of the machine learning application in this Chapter, we refer to these parameters as the features of the model.*

The grain-surface network we utilise is the default one in the GitHub repository that has been able to reproduce the abundances of the main observed grain-surface species for example in Holdship et al. (2017) and Chapter 2. The grain-surface reaction mechanisms that are used in UCLCHEM include the Eley-Rideal mechanism as well as the Langmuir-Hinshelwood grain-surface diffusion mechanism. These were implemented into the code in Quénard et al. (2018), along with the competition formula from Chang et al. (2007) and Garrod and Pauly (2011). The binding energies that are required in order to calculate diffusion reaction rates are taken from Wakelam et al. (2017). The gas-phase network is taken from UMIST (McElroy et al. 2013). While the grain network has undergone minor modifications since de Mijolla et al. (2019), the gas network has remained the same. Since we are only considering gas-phase species, minor modifications to the grain network are unlikely to be influential.

6.3 Machine Learning Interpretability and Statistical Emulation

6.3.1 Machine Learning Interpretability

It is often unclear why a model provides a certain output for a given input. This is not exclusive to machine learning algorithms, but can also be an issue with computational codes that integrate systems of differential equations, such as UCLCHEM. As a result, identifying the effect that a specific physical parameter, which we refer to as a feature in this Chapter, has on the output becomes difficult. The concept of feature importance refers to the size of the contribution of a specific feature in determining the model output. There exist many methods by which one can interpret the effect of a parameter in making

a certain prediction value, such as permutation feature importance or Local Surrogate Models. For an overview of the various methods, see Molnar (2022).

We use Shapley values, a method from game theory, to quantify the importance of the features (Shapley 2016). This is the first such application in the area of astrochemistry. While we provide an overview of the method we use in this Chapter below, we refer the reader to Shapley (2016), Lundberg and Lee (2017) and Lundberg et al. (2018) for further details.

The Shapley value of the ith feature, ϕ_i^j,[2] is defined as the marginal contribution of that feature in mapping the j^{th} data point in our dataset, x^j, to its corresponding output $f(x^j)$ averaged over all possible coalitions. A coalition is defined as a subset of the set of features. Notice that in this case, the function f corresponds to UCLCHEM and x^j corresponds to a particular input vector consisting of one entry for each feature in Table 6.1 that we modify. Each Shapley value, ϕ_i^j, is specific to each parameter of each data point.

Shapley value explanations are given as a linear model (Molnar 2022). We define a feature explanation model, $\hat{g}$ in the following way:

$$\hat{g}(x'^j) = \phi_0 + \sum_{i=1}^{n} \phi_i x_i'^j, \tag{6.1}$$

where $\phi_0 = \mathbb{E}[f(x)]$ is the value of the average prediction in our dataset, ϕ_i is the explained feature effect of the ith feature, n is the number of features and $x_i'^j$ is an element of the "coalition vector", x'^j, where $x'^j \in \{0,1\}^n$. The coalition vector is a vector consisting of zeros and ones with a zero indicating that a feature is "absent" and a one indicating it is "present".

One can imagine that this feature explanation model gives us an understanding of what happens when we choose to remove certain features, that is set a particular $x_i'^j$ to equal zero. If we want to be able to calculate the feature importance of a specific feature, then we need to be able to selectively "remove" features and see how this impacts our model output. When we say that we "remove" a feature, what we effectively mean is that we replace that value in the input vector by a random value from the dataset for that feature. The logic behind Shapley values is that we wish to see the contribution of a

[2]Unless otherwise specified, superscripts are used throughout this Chapter as labels, not for exponentiation.

specific feature when we include or exclude it from our data point for varying coalitions of features.

More formally, we can calculate the feature value importance as follows for a data point:

$$\phi_i^j = \sum_{S \subseteq N} \frac{|S|!(n - |S| - 1)!}{n!} (g(x_i'^j) - \hat{g}(x_{-i}'^j)), \tag{6.2}$$

where N is the set of features, $n = |N|$, S is the subset, $g(x_i'^j)$ is the explanatory model evaluated when the feature is included and $g(x_{-i}'^j)$ the explanatory model evaluated when the feature is not included. We refer to ϕ_i^j as a Shapley value.

We can specifically make a connection between the function we are trying to explain, $f(x)$, and the explanation function by noting that $\phi_0 = \mathbb{E}[f(x)] = \frac{1}{d}\sum_{j=1}^{d} f(x_j)$, where d is the number of data points. By setting all x_i' equal to 1 we obtain:

$$f(x^j) = \hat{g}(x'^j) = \mathbb{E}[f(x)] + \sum_{i=1}^{n} \phi_i^j, \tag{6.3}$$

which implies that the value of a function at a given data point is equal to the global average of the function (i.e. $\mathbb{E}[f(x)]$) plus the feature value importances we calculate for that data point.

We now explain what this entails practically. Say that we have a data point of the form (n, ζ, T, m_z, ψ) = (10^3, 500, 50, 1, 500) and we are interested in determining the contribution of the temperature being 50 K in producing an abundance of, say, 10^{-6}. What this entails is taking all subsets of the set of features. Two of these subsets might be:

- All of the original features
- All of the original features except the density

For the first of these subsets, we consider the change in the value of the explanatory model, $\hat{g}$, when we include and exclude the temperature value of $x_3 = 50$. "Excluding" simply means that we replace the 50 K with a randomly drawn temperature value from our dataset of temperatures. We then compute the feature explanation model when this

temperature value is included and take the difference as seen in Equation 6.2. For the second sample subset, we repeat this process except we always take a random value for the density as this is excluded from this subset. This is done for all subsets to calculate the feature importance for temperature.

However, observe that the calculation across all the subsets becomes computationally unfeasible as the number of features grows, with the number of coalitions growing exponentially. We employ SHAP (Lundberg and Lee 2017) to allow us to address this issue. SHAP is particularly useful, as it approximates the Shapley values, greatly reducing the time taken to compute them. SHAP has been found to be the theoretically optimal means of calculating feature attribution (Lundberg and Lee 2017; Lundberg et al. 2018). This is done through the use of the TreeSHAP algorithm (Lundberg et al. 2018). TreeSHAP is an algorithm that exactly computes the SHAP values for tree-based algorithms, such as XGBoost or random forests. One drawback of TreeSHAP is that it can give unintuitive explanations when the features are related (Molnar 2022). This is unlikely to be the case in this Chapter, as we work with five physically unrelated physical features that we sample independently when we generate our data set.

We can also provide a ranking of the various features in terms of global feature importance. As Shapley values can be negative, this can be achieved by averaging the absolute value of all Shapley values for each feature across all datapoints. Formally, this is defined as:

$$I_i = \frac{1}{d}\sum_{j=1}^{d}|\phi_i^j|, \tag{6.4}$$

where d is the number of data points and I_i is the average absolute value of the ith feature.

In principle, if we wished to compute the relative importances of the features we can do this by taking the above-mentioned average of the absolute values for a single feature and normalising this by the sum of the average of absolute values for all the features. We can then define the "relative importance" for a feature i, $\hat{I}_i$, as:

$$\hat{I}_i = \frac{\sum_{j=1}^{d}|\phi_i^j|}{\sum_{m=1}^{n}\sum_{j=1}^{d}|\phi_m^j|}, \tag{6.5}$$

where n is the number of features. This quantity effectively gives us a fractional contribution of each feature to the average behaviour of the model. We summarise the relative importance of each parameter in predicting the outputs we consider in this Chapter in Table C.2.

6.3.2 Implementation

While the use of SHAP greatly reduces the time taken to obtain the Shapley values relative to calculating the Shapley values in full, this process is still likely to take long due to the time taken per evaluation of the forward model, i.e. UCLCHEM. Each evaluation of the forward model takes on the order of 1-2 minutes. This makes considering an ensemble of models with 100000 runs or more unfeasible. To circumvent this, we elect to train a statistical emulator to reproduce the results of UCLCHEM. If the emulator has a sufficiently high accuracy, then it is safe to assume it is able to capture the internal workings of the original code, which we wish to probe. We now discuss the emulator and how we build it.

To train the emulator, we generated 120,000 points in parameter space using a Latin Hypercube sampling scheme (McKay et al. 1979), which was implemented using the Python surrogate modelling toolbox (Bouhlel et al. 2019). Data points in parameter space were generated such that all values were in the ranges given in Table 6.1. For those features that spanned several orders of magnitude, we elected to sample in log-space.

Each species' final log-abundance was used as the output of the algorithm. This was to ensure that all orders of magnitude were treated equally. All abundances less than 10^{-12} were set equal to 10^{-12} to ensure that the emulator was not being trained to learn what was effectively numerical noise. The input parameters were then scaled to be in the range 0 to 1. This limit was chosen because this is typically the lowest observed gas-phase abundance in the literature. We summarise the range of outputs for each species and ratio we consider in this Chapter in Table C.3.

An XGBoost regressor was trained for the emulation process (Chen and Guestrin 2016). XGBoost is a gradient-boosted decision tree regressor. We used the Python implementation for XGBoost to train our model[3]. It was found that better performance was obtained if a separate emulator was trained for each species, as opposed to having one network trained to predict the final abundances of all 239 species in the network. While

[3]https://xgboost.readthedocs.io/en/stable/index.html

we trained an emulator for every species in the network, we only present the results of a handful of molecules in this Chapter. We elected to train an XGBoost model instead of using a neural network as in de Mijolla et al. (2019), as XGBoost has been found to perform better on tabular datasets such as the one we consider, while also requiring less tuning (Shwartz-Ziv and Armon 2022).

In order to find the best set of hyperparameters for each emulator, we utilised Bayesian Hyperparameter optimisation. Under this procedure, we tune the hyperparameters on a validation set and find the best combination of parameters that minimise the L2 loss. Unlike a grid-search approach to hyperparameter tuning, Bayesian optimisation uses the model performance on previous hyperparameter combinations to choose a next best option, thereby saving a considerable amount of time compared to a grid-search approach. XGBoost has five tunable hyperparameters that we varied using the Bayesian Optimisation Python library (Nogueira 2014). We list the ranges over which we varied these in Table 6.2. For integer hyperparameters, we would round to the nearest integer.

When evaluating the accuracy of each trained emulator, we considered both the L2 loss obtained obtained from the performance of the emulator on the test dataset as well as the R^2 coefficient, which is often referred to as the coefficient of determination. The L2 loss is defined as:

$$L2 = \sum_{i=1}^{n} (y_n - \hat{y}_n)^2, \tag{6.6}$$

where n is the number of data points in the test set, y_n is the true value of the n^{th} test data point and $\hat{y}_n$ is the predicted value of the n^{th} test data point.

The R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_n - \hat{y}_n)^2}{\sum_{i=1}^{n} (y_n - \bar{y})^2}, \tag{6.7}$$

where $\bar{y}$ is the average value of y. All emulators in this Chapter had R^2 scores greater than 0.98.

Hyperparameter	Range of values	Data type
Maximum Depth	(3, 100)	Integer
Maximum features	(0.8, 1.0)	Float
Learning Rate	(0.01, 1.0)	Float
Number of Estimators	(80, 150)	Integer
Sub-sample	(0.8, 1)	Float

Table 6.2: *Table of the hyperparameter ranges used when tuning the XGBoost regressor.*

6.4 Results

We now look to consider a number of molecules of interest and explore how machine learning interpretability adds to our understanding of their equilibrium abundances in Phase 2. Note that all the molecules we will be considering will be gas-phase molecules, as they evaporate during the warm-up phase. We only considered a small number of molecules as a proof-concept for this method and provide the figures for these here. Figures for other molecules can be found in a dedicated repository[4].

6.4.1 Molecules

We begin by first considering individual molecules of interest to demonstrate what can be done with machine learning interpretability. We elect to consider three molecules: H_2O, CO and NH_3. CO is considered as it is the most abundant molecule besides H_2 and also plays a role in molecular gas cooling (Goldsmith 2001; Shi et al. 2015). H_2O is of interest due to its high abundance in planetary systems and of course its importance in the area of astrobiology (Gensheimer et al. 1996). NH_3 is speculated to be one of the main carriers of nitrogen and it often used as a tracer molecule of cold, dense clouds (Benson and Myers 1989; Caselli et al. 2019).

H_2O

We now investigate the importance of the various physical parameters on the value of the abundance of H_2O. Figure 6.1 is a beeswarm plot which is meant to serve as an information-dense qualitative summary of feature importances. Each point in the beeswarm plot represents a data point from our test set. The features are arranged from top to bottom in decreasing order of importance to the model output, which is measured by Equation 6.4. Recall that the SHAP value measures the impact of each feature on the value of the

[4] https://github.com/Bamash/MLinAstrochemistry

prediction, relative to some baseline value, which is simply the global average, i.e. the average logarithm of the abundance. Along the horizontal axis, individual predictions are plotted in terms of their SHAP value. The points are colour-coded according to their value with the colour bar indicating the value relative to the range of values of that feature. A single colour bar is used for all the features in order to qualitiatively show the relationship between the feature value and the SHAP value. It is for this reason that the colour bar range is from "Low" (indicating the lowest value of the respective feature) to "High" (indicating the highest value that the feature can take) Furthermore, the vertical clustering of the points indicates the density of the points in a manner akin to a violin plot. We emphasise again that this plot is meant to help provide easy-to-use qualitative explanations for observers.

We also consider a more quantitative plot to delve deeper into some of the finer points of the beeswarm plot. This is useful if one wishes to consider the nature of the relationship between each feature and the log-ratio. While one can deduce that for temperature and metallicity the relationship is monotonic and increasing, this might still not be enough. It is for this reason that we can plot dependence plots such as Figure 6.2 which plots the SHAP value for all the variables as a function of the individual variable. Notice that in the plot for each feature i, the SHAP value corresponds to the importance of only that feature, ϕ_i^j, for a point j. Effectively, these dependence plots gives us the marginal contribution of each feature i to the output.

We can also consider the relationship between the abundance of water (instead of the SHAP value) as a function of each of the features. This is plotted in Figure 6.3. Notice that in order to compute the abundance, we must utilise Equation 6.3. This means that to compute the abundance we must add the mean log-abundance of water, $\phi_0 = \mathbb{E}[f(x)]$, to the SHAP values of each of the features for that data point. As a result of the explanatory model being linear in nature, we do not see the same relationships in Figure 6.3 and in fact observe that there is no relationship between all the parameters besides metallicity and the log-abundance. This is because many of the SHAP importances cancel each other out. Only metallicity still has a noticeable relationship with the log-abundance when we add up the importances of all the parameters.

However, in the interest of better understanding the impact of each parameter's individual relationship with the log-abundance, we consider the marginal effects in Figure 6.2. We would like to emphasise that it is still useful to consider the marginal effects.

While we consider a wide range of physical conditions, many observational and modelling exercises relating to tracers will be far more restrictive in their parameter ranges as well as the number of varying parameters. A typical observational environment will not contain the parameter ranges we consider here. It is precisely for these tasks that this methodology will be useful. We observe that the relationships between these variables and the log-abundance of H_2O are mostly monotonic. However, we will only consider the impact of metallicity, as this has the strongest impact on the model output. The SHAP values for the other features range between -0.4 and 0.4 in log-abundance space which corresponds to factors of 2.5 relative to the average water abundance. Throughout this Chapter, we will only consider features whose SHAP values exceed 1 in log-abundance space. It is clear that metallicity will play a significant role in the abundance of water. While there exists some debate as to what fraction of the ISM oxygen abundance is present in water (van Dishoeck et al. 2021), a decrease in the metallicity will result in a decrease in the amount of oxygen, which in turn will mean that less water will be formed, due to greater competition for the little oxygen present. On the other hand, a large amount of oxygen will result in the opposite effect, to an extent. Water has several destruction pathways that impose an upper limit on how much of it is formed in the gas-phase, regardless of how much oxygen is present.

CO

Carbon monoxide is an important molecule to consider in astrochemistry. Not only is it an important molecule in the context of grain-surface chemistry and the formation of various complex organic molecules, but it also plays a significant role in gas-phase chemistry. In particular, it is often considered a molecular gas coolant at low temperatures and densities (Goldsmith 2001; Shi et al. 2015). We are interested in considering how the various parameters we are changing influence its abundance. Figure 6.4 is a beeswarm plot of the various features and shows that only the metallicity plays a strong role in determining the final CO abundance, which has an $\hat{I}_i$ of 0.91. In order to investigate the exact nature of the relationship, we plot the SHAP dependence plots in Figure 6.5. We observe an interesting relationship between the metallicity and the CO abundance that is monotonic in nature. We do not observe any notable relationships between its log-abundance and the other parameters, so we only focus on metallicity for now. We observe that for very low metallicities the CO abundance ends up being almost 2 orders of magnitude lower

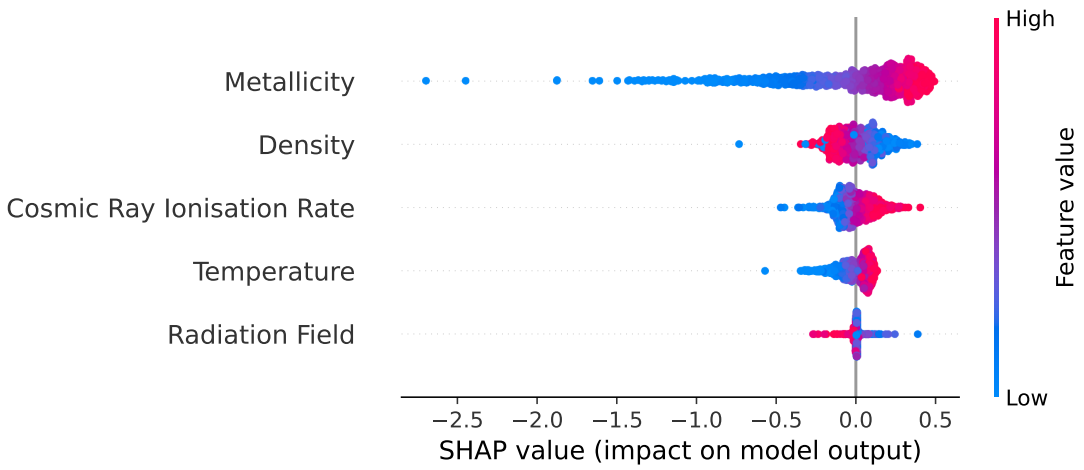

Figure 6.1: *A beeswarm plot of the various physical parameters demonstrating their relative importance in predicting the log-abundance of H_2O. The features are arranged from top to bottom in decreasing order of importance to the model output, which is measured by the mean of the absolute value of the SHAP value averaged across all predictions. Individual predictions are plotted along the horizontal axis according to their SHAP value, which indicates the difference in the value of the model output for that prediction relative to the global average. Furthermore, the points are colour-coded in terms of the size of the feature value relative to the range of values that the respective feature takes. We observe that metallicity has the greatest impact followed by density, cosmic ray ionisation rate, temperature and radiation field.*

than the 'average' value due to the marginal effect of the metallicity. In Figure 6.6 we plot the log-abundance of CO as a function of each of the parameters. As we discussed for H_2O before, to compute the CO abundance we must add the contributions of all the features. As a result of this, only the metallicity appears to have a strong effect on the log-abundance.

Work has been done to consider the impact of metallicity on CO. In Shi et al. (2015), this was considered in the context of metal-poor galaxies. Here they wished to consider to what extent CO, known to be a coolant in metal-rich galaxies, could also serve the same role in metal-poor ones. It was found that there was significant CO depletion in metal-poor galaxies due to photodissociation. This is unlikely to be the case here as the radiation field is found to be the least influential parameter. The radiation field is only likely to be effective in photodissociation when the density is very low and the radiation field itself is high, which will only be the case for a small number of parameter combinations. We must consider other reasons for the importance of metallicity.

Within the UCLCHEM code, the metallicity parameter is a scale factor that scales all elemental abundances of elements heavier than helium by the same factor. This means

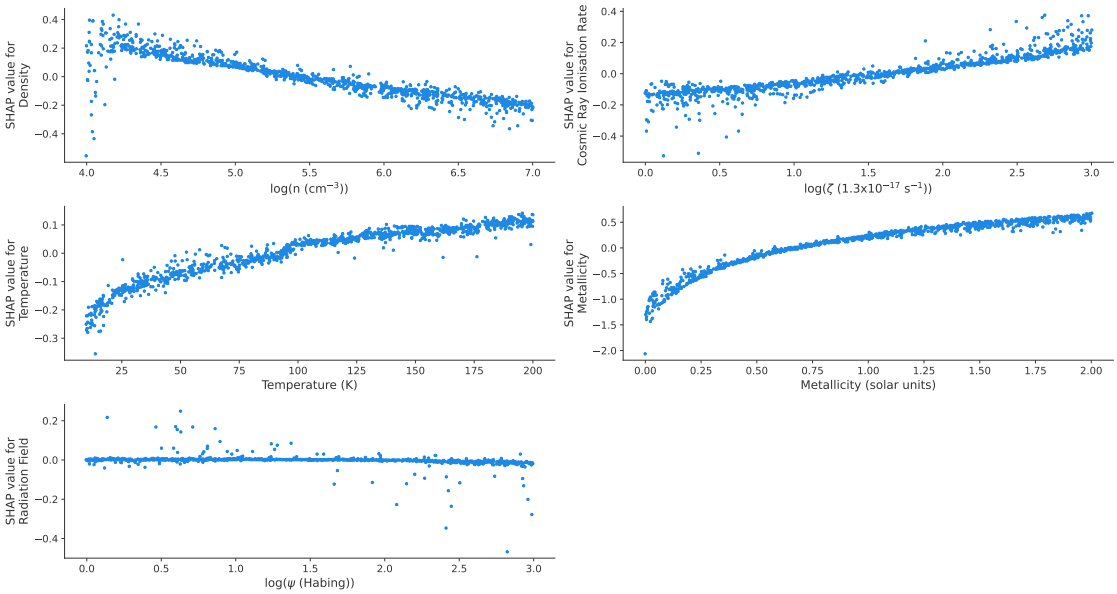

Figure 6.2: *A plot of the SHAP values as a function of the feature values used to predict the log-abundance of H_2O. Unlike the beeswarm plot, these SHAP dependence plots allow us to see the exact nature of the relationship between the feature value and SHAP value. Recall that the SHAP value tells us the difference in value between the average output value (log-abundance of the water). We see that the logarithms of density and the cosmic ray ionisation rate are roughly linear with respect to the SHAP value with the same being true for the temperature. For metallicity, we observe a significant decrease in the SHAP value for low metallicities, but this seems to level off for values greater than 1.*

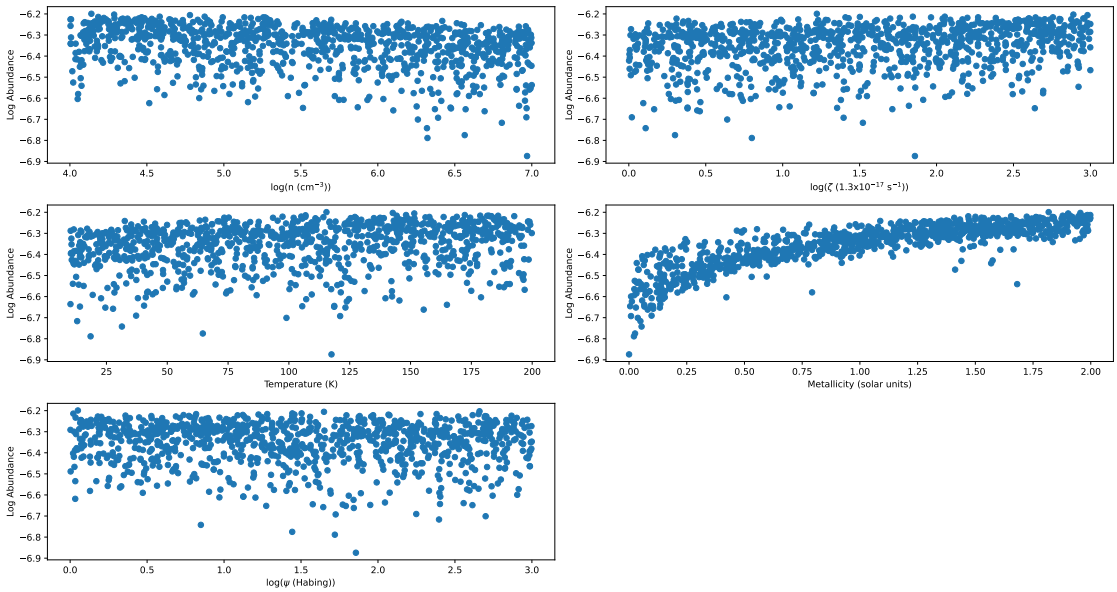

Figure 6.3: *A plot of the log-abundance of H_2O as a function of the various features. To calculate the log-abundance for a given data point, we needed to sum up the importance values of each feature for that data point. We observe that only metallicity maintains a clear trend. For the other features, we have no discernible trend which can be attributed to the feature importances nullifying each other.*

that as the metallicity parameter is reduced, the abundances of some elements will be reduced so there will be greater competition for them. This results in the abundances of species dropping, as there is simply less of their constituent elements. In the case of CO, we know from Table C.1, that there is less C than O which means reducing the metallicity results in C becoming more scarce. It is for this reason that at metallicities close to zero the final abundance of CO drops by 2 orders of magnitude.

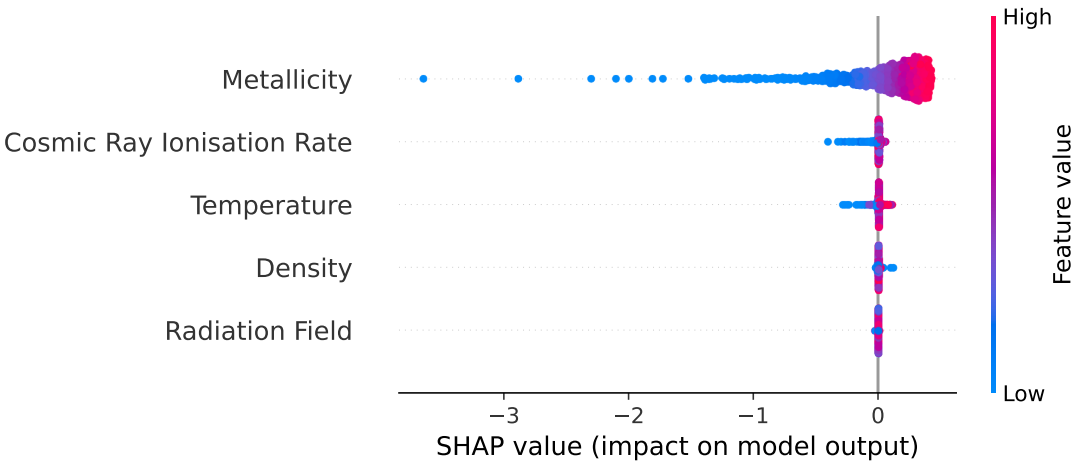

Figure 6.4: *A beeswarm plot of the various physical parameters demonstrating their relative importance in predicting the log-abundance of CO. We observe that metallicity is the only parameter with a significant influence on the value with the other parameters not being very useful predictors.*

NH_3

We now consider ammonia, which is considered one of the significant sources of nitrogen in the interstellar medium. The beeswarm plot in Figure 6.7 shows the ranking of the 5 features in terms of their relative importance. We consider the nature of the relationship through the use of the SHAP dependence plots in Figure 6.8 with only temperature being found to have a consistently significant relationship with the log-abundance. Parameters such as density and cosmic ray ionisation rate may have individual points with large SHAP values but these are low in frequency compared to the tens of thousands of points plotted, which is why we do not discuss them further.

The dependence on temperature is quite interesting, as we notice that there are two separate temperature ranges over which the abundance takes a different constant value, with the cutoff temperature being 100 K. This is also seen in Figure 6.9 which is a plot of the abundances as a function of the individual parameters seems to indicate two regimes.

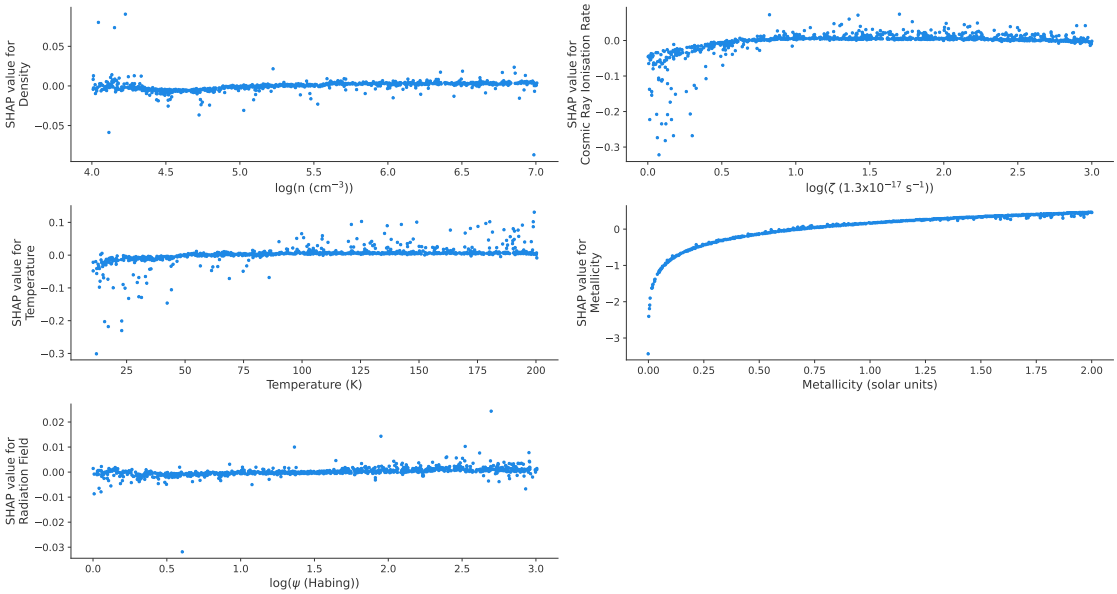

Figure 6.5: *A plot of the SHAP values for the various features (besides the radiation field) as a function of the feature values used to predict the log-abundance of CO. As was observed in the beeswarm plot, only metallicity has a significant effect on the abundance. For low metallicities, we observe a large decrease in the SHAP value. The SHAP value monotonically increases with metallicity, eventually levelling off for values greater than 1.*

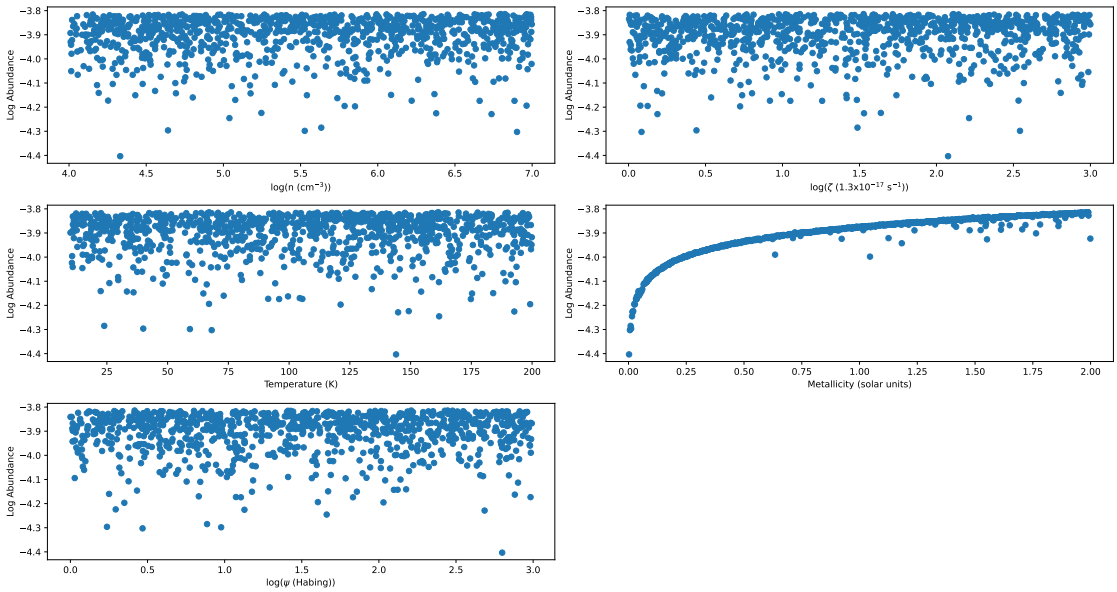

Figure 6.6: *A plot of the log-abundance of CO as a function of the various features. To calculate the log-abundance for a given data point, we needed to sum up the importance values of each feature for that data point. Only metallicity maintains a clear trend compared to Figure 6.5. For the other features, we have no discernible trend. This is due to the marginal feature importances nullifying each other.*

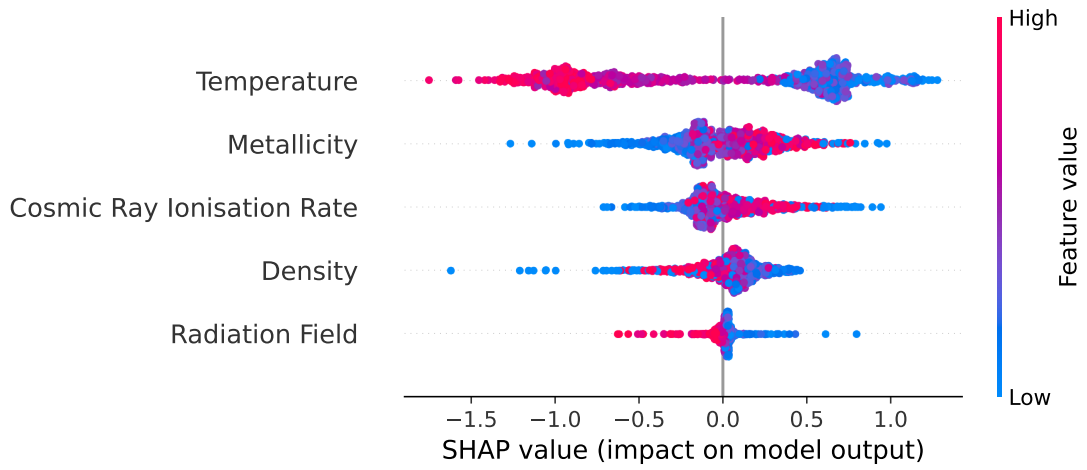

Figure 6.7: *A beeswarm plot of the various physical parameters demonstrating their relative importance in predicting the log-abundance of NH_3. We observe that temperature has the largest impact on the model output with SHAP values ranging from -1.5 to 1.0. The temperature relationships does not seem to be monotonic. The next most important features are metallicity, followed by the cosmic ray ionisation rate, density and the radiation field, with the first three also not having monotonic relationships with the SHAP value.*

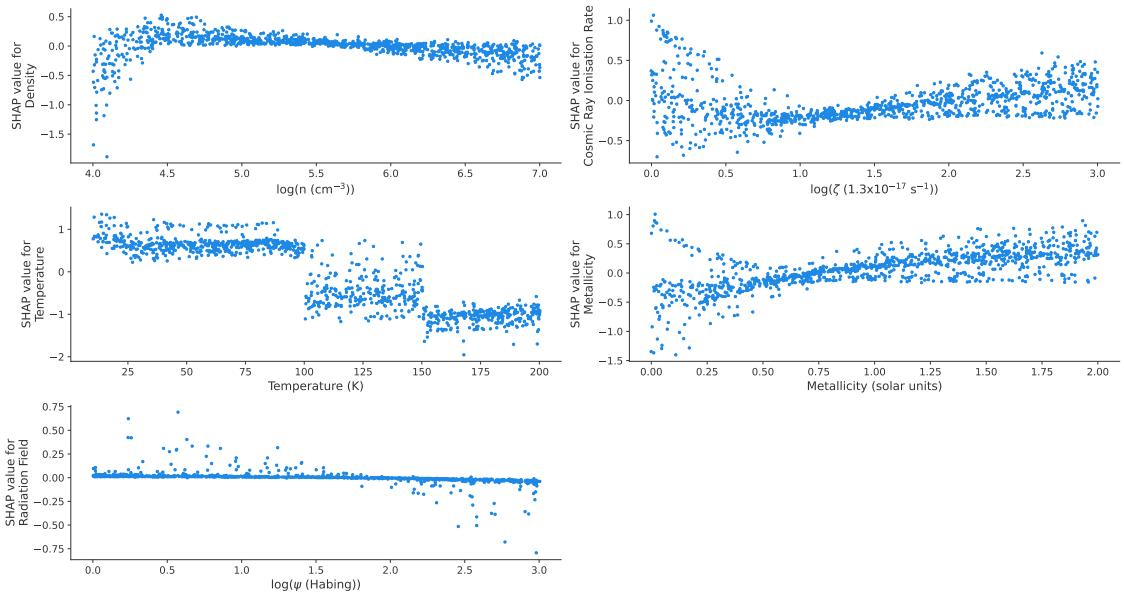

Figure 6.8: *A plot of the SHAP values for the various features (besides the radiation field) as a function of the feature values used to predict the log-abundance of NH_3. We observe that temperature has an interesting relationship with the SHAP value. What we observe is that there exist three separate temperature regimes under which the final abundance is relatively constant. The abundance does show some non-monotonic variance with respect to the other features, but most of these are within 0.5 of the average value (or a multiplicative factor of 3).*

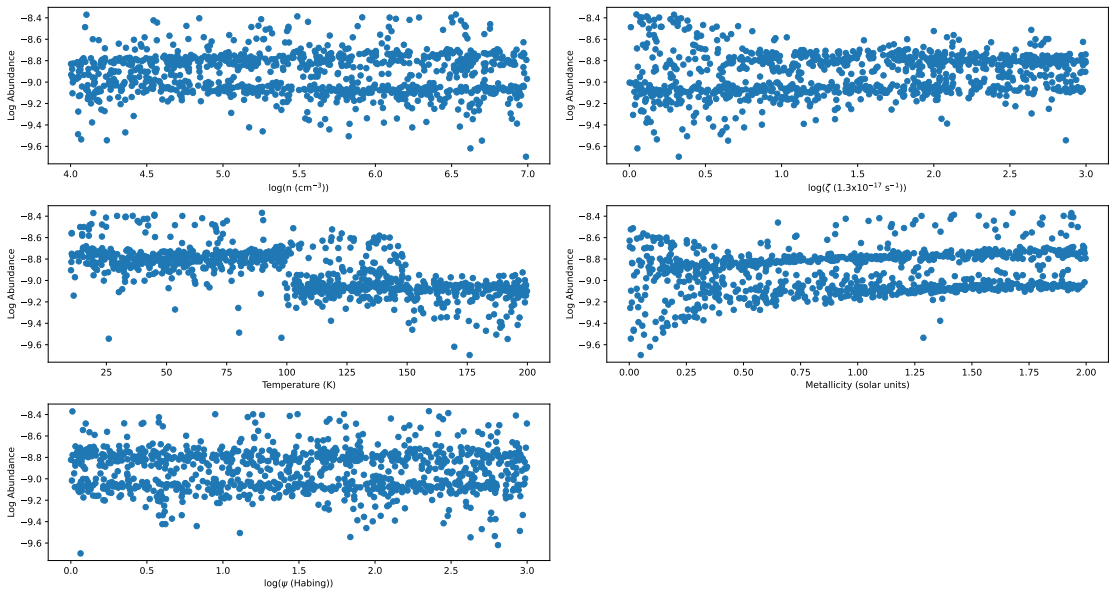

Figure 6.9: *A plot of the log-abundance of NH_3 as a function of the various features. To calculate the log-abundance for a given data point, we needed to sum up the importance values of each feature for that data point. We observe that only temperature maintains a clear trend relative to what we observed in Figure 6.8. However, we now appear to have something closer to a two-temperature regime rather than a three-temperature one. For the other features, we have no discernible trend which can be attributed to the feature importances nullifying each other.*

We deduce that these different regimes are related to the chemistry surrounding NH_3 being very different at these stages. The other parameters are of less importance as the deviation from the average NH_3 log-abundance is within about 0.5, or a factor of 3 in actual abundance. The non-temperature parameters cancel each other out in terms of their contributions when these are added together. This might explain why the transition in the log-abundance as a function of the temperature is less sudden than the transition in the SHAP value. The SHAP value potentially has abrupt jumps at specific temperatures due to the heating profile utilised by UCLCHEM for the warm-up phase that is potentially triggered at 100 K.

We can investigate the temperature dependence by considering the relative rates of formation and destruction of ammonia at specific points in time. Figure 6.10 plots the fractional contributions of the various formation and destruction routes of NH_3 for an instance where the peak temperature is 160 K. We only considered the top reactions that contributed to 99% of the creation or destruction of NH_3. The main NH_3 formation routes are:

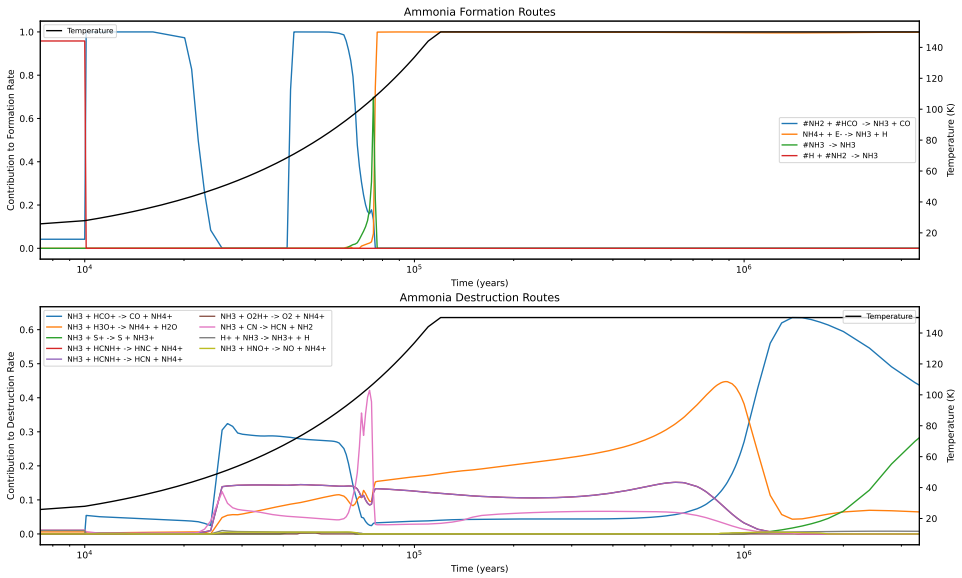

Figure 6.10: *Top: Plot of the fractional contribution of various ammonia formation routes that contribute to 99% of the NH_3 formation at each time. The temperature as a function of time is also plotted. Bottom: Plot of the fractional contribution of various ammonia destruction routes that contribute to 99% of the NH_3 at each moment in time. We only considered the top reactions that contributed to 99 % of the creation or destruction to limit the number of lines we would have to plot.*

$$\#\mathrm{H} + \#\mathrm{NH_2} \longrightarrow \mathrm{NH_3} \tag{6.8}$$

$$\mathrm{NH_4}^+ + \mathrm{e}^- \longrightarrow \mathrm{NH_3} + \mathrm{H} \tag{6.9}$$

$$\#\mathrm{NH_3} \longrightarrow \mathrm{NH_3}\,(\mathrm{UV\ and\ CR\ desorption}) \tag{6.10}$$

$$\#\mathrm{NH_2} + \#\mathrm{HCO} \longrightarrow \mathrm{NH_3} + \mathrm{CO} \tag{6.11}$$

Some of the main destruction mechanisms throughout the hot core phase are:

$$H_3^+ + NH_3 \longrightarrow NH_4^+ + H_2 \tag{6.12}$$

$$NH_3 + HCO^+ \longrightarrow CO + NH_4^+ \tag{6.13}$$

$$NH_3 + H_3O^+ \longrightarrow NH_4^+ + H_2O \tag{6.14}$$

$$NH_3 + CN \longrightarrow HCN + NH_2 \tag{6.15}$$

$$NH_3 + HNO^+ \longrightarrow NO + NH_4^+ \tag{6.16}$$

$$H^+ + NH_3 \longrightarrow NH_3^+ + H \tag{6.17}$$

$$NH_3 + HCNH^+ \longrightarrow HCN + NH_4^+ \tag{6.18}$$

$$NH_3 + HCNH^+ \longrightarrow HNC + NH_4^+ \tag{6.19}$$

$$NH_3 + S^+ \longrightarrow NH_3^+ + S \tag{6.20}$$

When the peak temperature is reached, the only formation reaction left is the gas-phase electron addition reaction. This is due to high temperature making grain-surface chemistry untenable, as most of the available grain-surface material has evaporated. In the gas-phase, the destruction routes are still active and recycle some of the gas-phase NH_3 and turn it back into NH_4^+, but some of it goes on to form HCN and other species, resulting in the eventual decrease in the NH_3 abundance. This is more severe for higher final hot-core temperatures as these destruction reactions see their rates increase, resulting in even lower final NH_3 gas-phase abundances.

6.4.2 Molecular ratios

While species may serve as useful tracers for specific energetic processes under certain density and temperature conditions, it is often more useful to consider intensity ratios between different molecules, espeially in extragalactic environments (Viti 2017; Imanishi et al. 2019; Butterworth et al. 2022). Tracer ratios are often considered in observations to cancel out the beam filling factor. The two tracer ratios we consider are HCN/HNC and HCN/CS, both of which have been extensively studied in the literature. The former is considered a good tracer of temperature and the latter a dense gas tracer.

HCN/HNC

We begin by considering the ratio of the abundances of HCN to HNC. The ratio of these two molecules has been extensively studied and has also been subject to a considerable amount of debate. These two molecules are of great interest, due to their high abundances, their excitation conditions the areas in which they form as well as the proximity of their transitions in frequency space (Pety et al. 2017; Hacar et al. 2020). Recently, this intensity ratio was suggested as a potential chemical thermometer for the ISM (Hacar et al. 2020).

In Figure 6.11, we observe that temperature is indeed the most important feature. We observe that only the temperature and metallicity have significant impacts on the value of the ratio with relative importance values of $\hat{I_T} = 0.7$ and $\hat{I_{m_z}} = 0.24$. The other parameters do not have much influence on the log-abundance, so will not be discussed. Looking at the dependence plot for temperature further in Figure 6.12, we observe that the log-ratio increases monotonically with temperature, with there appearing to be two different temperature regimes judging by the change in gradient throughout the curve, which is also evident in Figure 6.13 which is a plot of the log-ratio against the features.

Figure 6.14 is a plot of the ratio (as opposed to the log-ratio) against the temperature. We fit a two-part linear function to the data. The presence of two regimes is in agreement with the literature (Graninger et al. 2014; Hacar et al. 2020). In Hacar et al. (2020), the relationship between the temperature and the ratio was described with a two-part linear function, which is what we roughly observe. The two isomers are formed in roughly equal proportions through the dissociative recombination of $HCNH^+$ (Herbst et al. 2000). As such, any deviation in the ratio from a value of 1 can be attributed to the destruction routes. The main ones considered in the literature are:

$$HNC + H \longrightarrow HCN + H \tag{6.21}$$

$$HNC + O \longrightarrow NH + CO \tag{6.22}$$

The pre-established energy barriers for both of these reactions have been questioned (see Graninger et al. (2014) for a full discussion of this). We update these values in line with Hacar et al. (2020) and Graninger et al. (2014) to be 200 K and 20 K respectively. The first reaction is particularly dominant at high temperature, where we have a large abundance of atomic H, whereas the second reaction is more dominant at low temperatures.

However, the second reaction does not appear to be the dominant HNC reaction at low temperatures. This can be seen in Figure 6.15 where we plot the fractional contribution of the reactions that are responsible for creating and destroying 99% of the HNC at each time step alongside the temperature as a function of time. We see that it is in fact the reaction $H_3^+ + HNC \longrightarrow HCNH^+ + H_2$ as well as freeze-out responsible for this at low temperatures. As such, we still have an explanation for the two regimes observed, but the oxidisation reaction seems to not play as important a role in our model, suggesting further study might be required. However, the inflection point in Hacar et al. (2020) is observed to be at 40 K, whereas in this Chapter it is at 65 K. This can be explained by noting that we consider a wider variety of physical parameter combinations, whereas the other work considered the ones specific to the Orion A Cloud. As such, a quantitative comparison is difficult to make. However, it is reassuring to observe qualitative agreement. Similarly, we observe no real relationship between the cosmic ray ionisation rate and the log-ratio, which

is broadly in agreement with the modelling done in Meijerink et al. (2011). However, there is some disagreement with observations as seen in Behrens et al. (2022), though this can be attributed to them considering a larger range of cosmic ray ionisation rates. In that paper, the ratio was found to decrease as the cosmic ray ionisation rate increased, though this was only in the presence of mechanical heating which we do not consider here.

We observe that for metallicity, we have the same "tailing-off" effect that we have observed previously, though this is only in the marginal case in Figure 6.12. This is the case for metallicity values between 0 and 1. Again, we can attribute this to increased competition for the individual atomic species which results in the ratio decreasing.

Bayet et al. (2012) considered a gas density of 10^4 cm^{-3}, radiation field values of 1 Habing, a cosmic ray ionisation rate of 5.0 $\times 10^{-17}$ s^{-1} and metallicities between 1 and 5. For metallicities between 1 and 2, we see a roughly linear marginal increase in the log-ratio. This is in line with what was observed in Bayet et al. (2012) in which an increase in the metallicity results in a linear increase in the log-abundances of HCN and HNC with the HCN having a steeper increase with metallicity. This suggests that their ratio would also increase linearly.

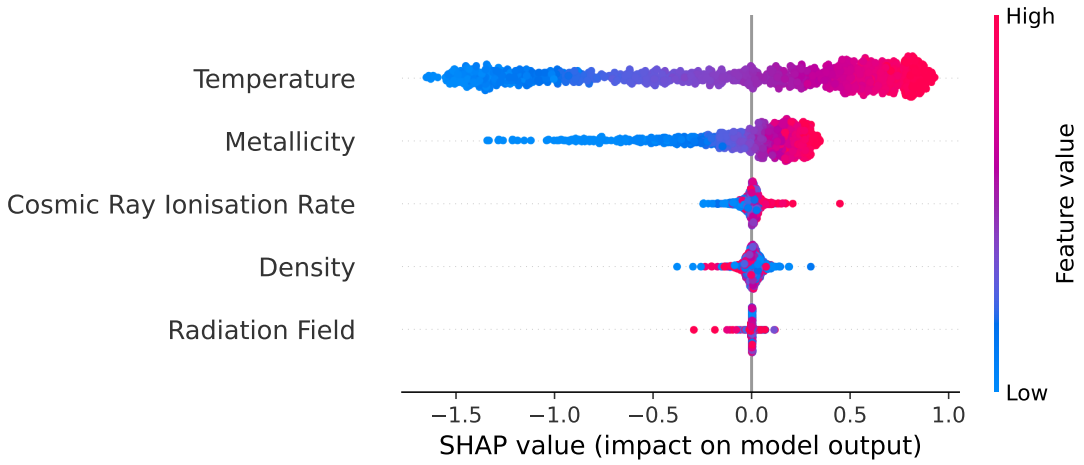

Figure 6.11: *A beeswarm plot of the various physical parameters demonstrating their relative importance in predicting the log-ratio of HCN to HNC. We observe that temperature has the largest impact on the model output with SHAP values ranging from -1.5 to 1.0. The fact that temperature is the most important feature is hardly surprising given that this ratio is seen as a thermometer. The next most important features are metallicity, followed by the cosmic ray ionisation rate, density and the radiation field, with the first three also not having monotonic relationships with the SHAP value.*

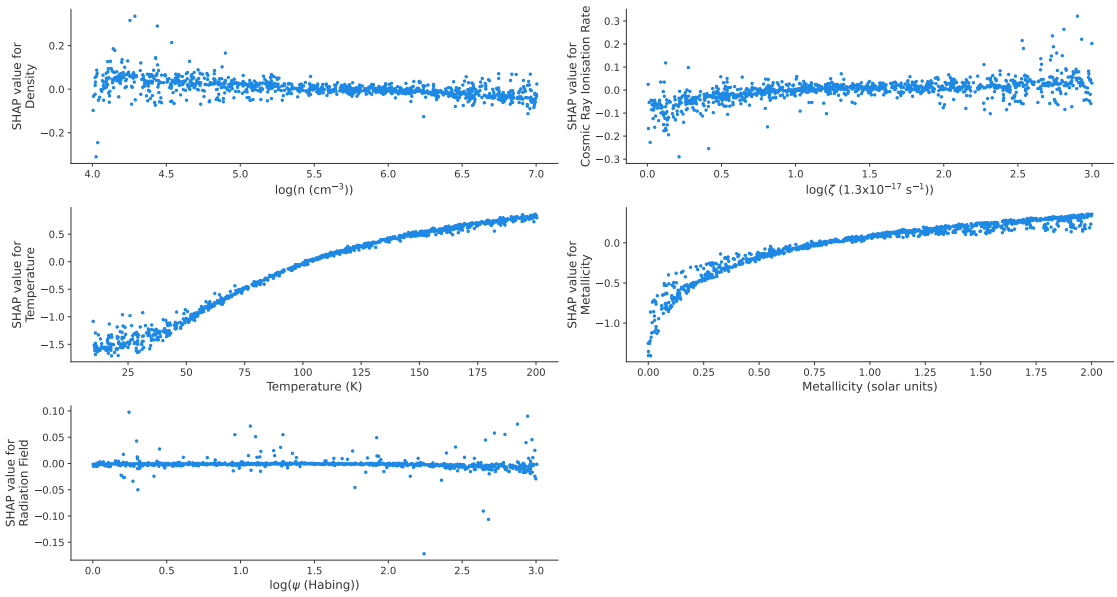

Figure 6.12: *A plot of the SHAP values for the various features (besides the radiation field) as a function of the feature values used to predict the log-ratio of HCN to HNC. We observe that temperature has an interesting relationship with the SHAP value with there being two regimes under which the ratio increases at different rates. This is in line with what was observed in Hacar et al. (2020) and was approximated there as a two-part linear function. The relationship between the SHAP value and metallicity is similar to what we observed in other molecules.*

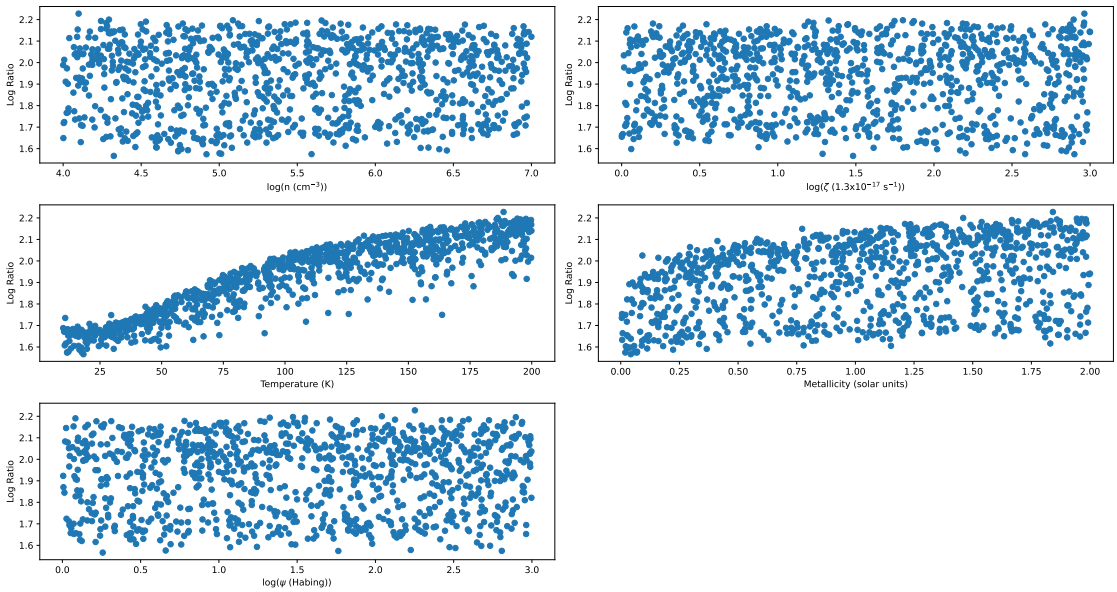

Figure 6.13: *A plot of the log-abundance of HCN/HNC ratio as a function of the various features. To calculate the log-ratio for a given data point, we needed to sum up the importance values of each feature for that data point. We observe that only temperature maintains a clear trend relative to what we observed in Figure 6.12. For the other features, we have no discernible trend which can be attributed to the feature importances nullifying each other.*

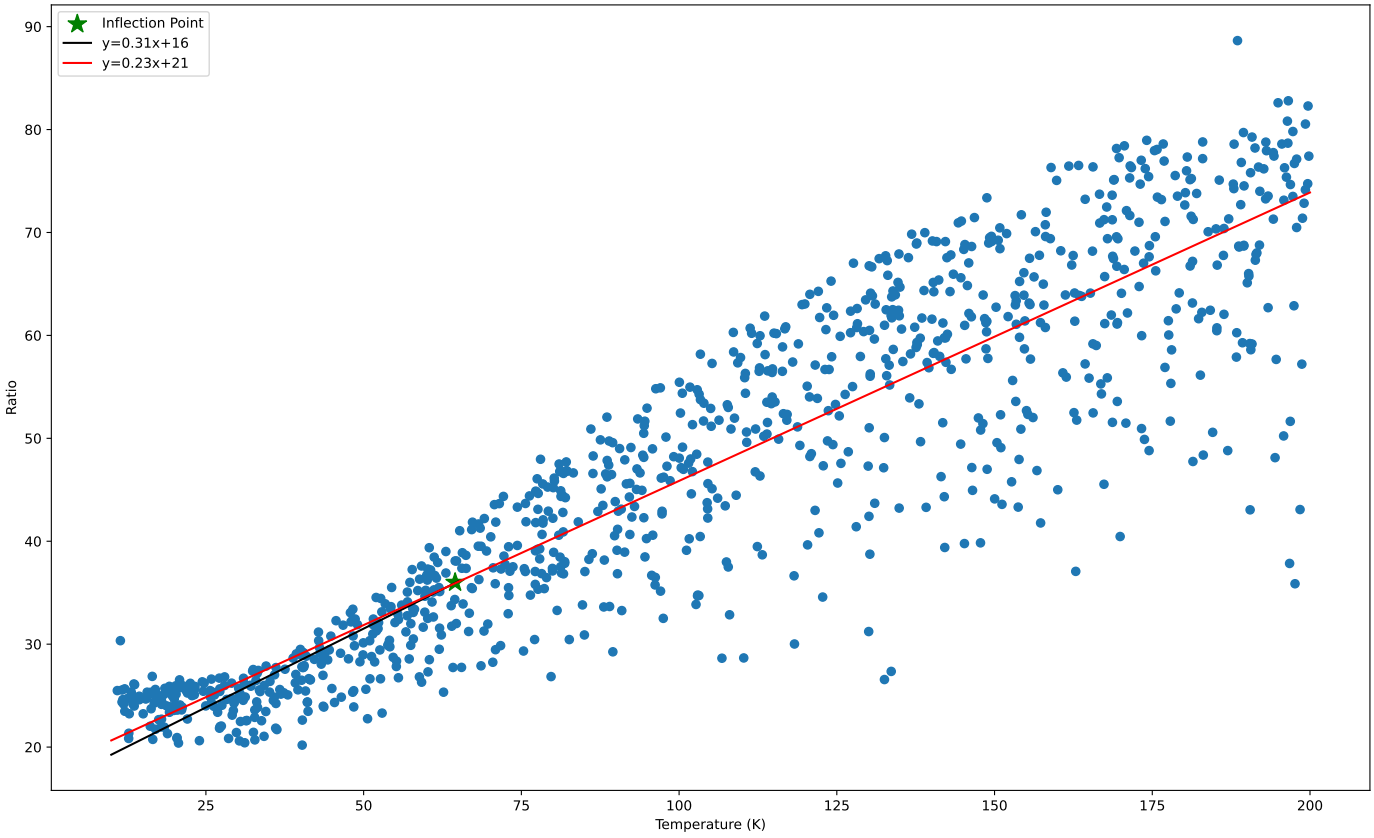

Figure 6.14: *Scatter plot of the HCN/HNC ratio (note: not the log-ratio) as a function of the temperature. We continue to observe an inflection point at 65 K and fit a two-part linear function. Below 65 K, the trend line is* $y = 0.31x + 16$ *(black) and above it is* $y = 0.23x + 21$ *(red). For the sake of clarity, we have included the entirety of the second part of the red linear function to make the change in gradient easier to see.*

HCN/CS

We now consider another tracer, the HCN to CS ratio. This ratio has received significant interest in recent years (Izumi et al. 2013, 2016; Butterworth et al. 2022), with one of the reasons being the fact that both HCN and CS are dense gas tracers (Viti 2017), with the HCN(4-3)/CS(2-1) ratio being a good tracer of active galactic nuclei (AGN) activity. Just as for the ratio of HCN to HNC, we now wish to obtain a sense of the relationship of the five features of interest with this ratio.

We begin by considering the relative importance of the five features. Figure 6.16 is a beeswarm plot demonstrating this. We observe that temperature is once again the most

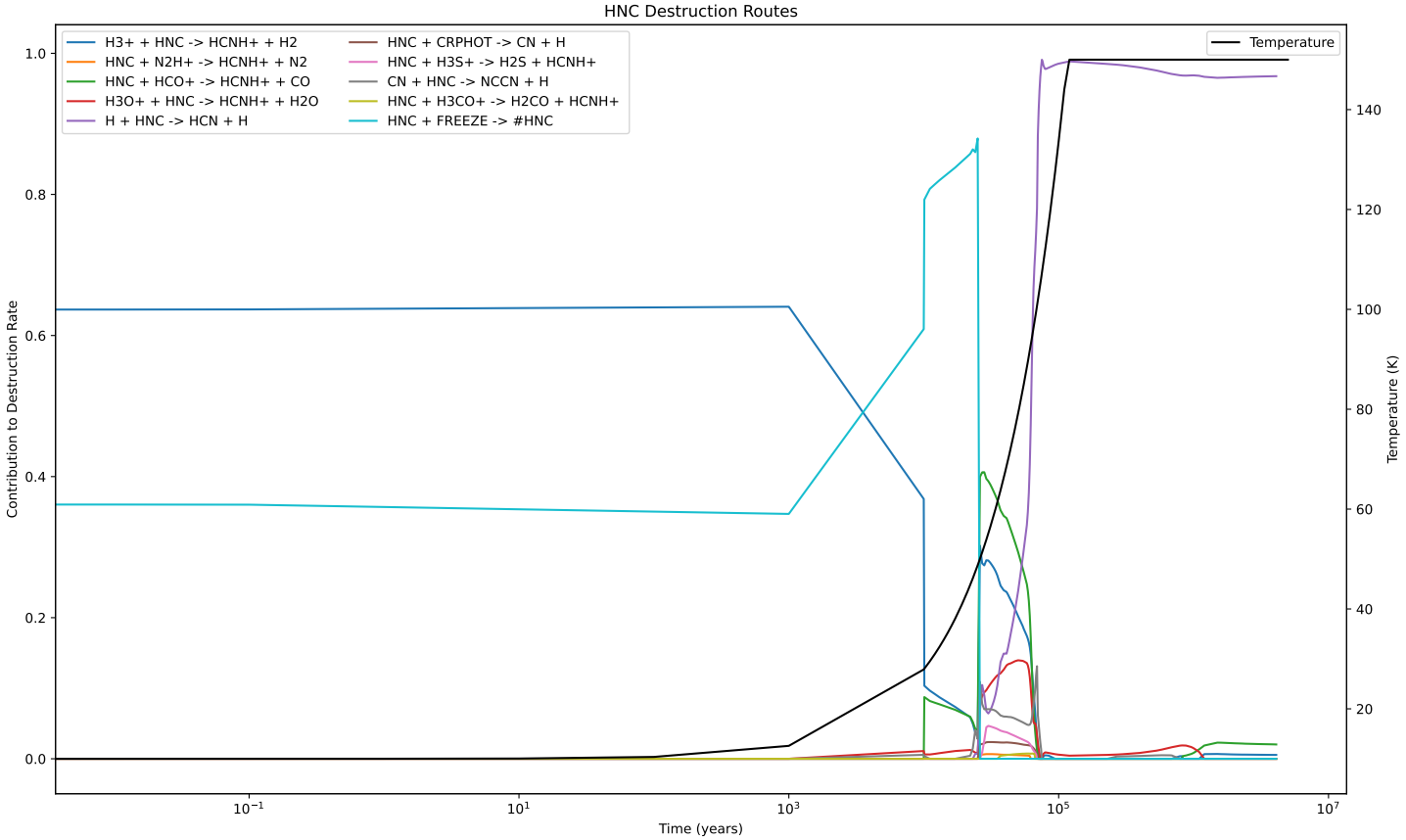

Figure 6.15: *Plot of the fractional contribution of various routes that contribute to 99% of the HNC destruction as a function of time. The temperature as a function of time is also plotted. We observe that for low temperatures, the main sources of gas-phase HNC destruction are* $H_3^+ + HNC \longrightarrow HCNH^+ + H_2$ *as well as freeze-out onto the grains, which runs contrary to our expectations of the reaction* $HNC + O \longrightarrow NH + CO$ *playing a dominant role. As the temperature increases we observe that the main destruction mechanism is the isomerisation reaction* $H + HNC \longrightarrow HCN + H$*. Note that the increase in the fractional contribution of the freeze-out reaction after* 10^3 *years is not due to the increase in temeprature, but rather simply numerical as the other destruction mechanisms become far smaller which leads to its fractional contribution to increase despite the absolute contribution being negligible. We only considered the top reactions that contributed to 99 % of the creation or destruction to limit the number of lines we would have to plot.*

relevant feature followed by density, cosmic ray ionisation rate, metallicity and radiation field.

We consider this more in Figures 6.17 and 6.18. The former shows the SHAP value as a function of the feature value, which means it shows the marginal effect of each feature. The latter considers the abundance as a function of each feature. As we discussed earlier, the abundances plotted are derived from summing the marginal effects of all the features. There is a clear quasi-linear relationship between the log-ratio and the log-density which supports the idea that the ratio could serve as a density tracer. The cosmic ray ionisation rate and the radiation field do not appear to have discernible relationships with the ratio. We find there is not a monotonic relationship with temperature. In fact, we once again seem to observe three separate temperature regimes. The abruptness in the changes in the SHAP values due to temperature could be attributed to specific physical changes that occur during the warming-up phase. This warrants further investigation, but would have to be done with a deep-dive that focusses specifically on temperature.

We observe that for the temperature variable there are three separate regimes of interest when it comes to the log-ratio: one for below 100 K, one for between 100 and 150 K and another for above 150 K. To start off with, we plot the temporal evolution of the abundances of the two molecules and the temperature in Figure 6.19 for three different values of the final temperature: 47 K, 105 K and 176 K. These were plotted using UCLCHEM. Note that these temperatures are not special in any way, but they are simply chosen as examples to illustrate the points we wish to discuss. Each of these temperatures falls within one of the three different regimes we observe in Figure 6.17 and were taken from the dataset. We also plot a time series of the ratio in Figure 6.20.

We observe that at 47 K, we initially have a large build-up of HCN until about 10^5 years. CS is also built-up, but not to the same extent. After this point, both abundances drop sharply, though the CS drops far more, leading to an increase in the value of the ratio. However, for 107 K the abundance of CS exceeds that of HCN leading to a smaller HCN/CS ratio. This is still true for 176 K, but CS approaches HCN's abundance much more closely.

In the low-temperature (< 100K) regime, the dominant destruction reaction of HCN is $H_3{}^+ + HCN \longrightarrow HCNH^+ + H_2$. Once the maximum temperature is reached, the main formation reactions are

$$HCNH^+ + e^- \longrightarrow HCN + H \tag{6.23}$$

$$CH + NO \longrightarrow HCN + O \tag{6.24}$$

$$NH_3 + CN \longrightarrow HCN + NH_2. \tag{6.25}$$

However, the NH_3-based reaction becomes less efficient over time at this temperature and is replaced by $N + HCO \longrightarrow HCN + O$.

In the mid-temperature regime (100 K-150 K), the major formation routes are:

$$NH_3 + CN \longrightarrow HCN + NH_2 \tag{6.26}$$

$$NH_3 + HCNH^+ \longrightarrow HCN + {NH_4}^+ \tag{6.27}$$

$$CH + NO \longrightarrow HCN + O \tag{6.28}$$

$$HCNH^+ + H_2CO \longrightarrow H_3CO^+ + HCN \tag{6.29}$$

$$N + HCO \longrightarrow HCN + O \tag{6.30}$$

$$CN + HCO \longrightarrow CO + HCN \tag{6.31}$$

with the major destruction routes being:

$$H_3O^+ + HCN \longrightarrow HCNH^+ + H_2O \tag{6.32}$$

$$HCN + CRPHOT \longrightarrow CN + H \tag{6.33}$$

$$H_3{}^+ + HCN \longrightarrow HCNH^+ + H_2 \tag{6.34}$$

$$HCN + H_3CO^+ \longrightarrow H_2CO + HCNH^+ \tag{6.35}$$

$$CH_3{}^+ + HCN \longrightarrow CH_3CNH^+ + PHOTON \tag{6.36}$$

with the final reaction becoming less efficient after about 2.3×10^5 years.

In the high-temperature regime (> 150K), the major HCN reserves are built up until 7.7×10^4 years via these reactions:

$$NH_3 + CN \longrightarrow HCN + NH_2 \tag{6.37}$$

$$CH + NO \longrightarrow HCN + O \tag{6.38}$$

$$HCNH^+ + H_2CO \longrightarrow H_3CO^+ + HCN \tag{6.39}$$

$$HCNH^+ + e^- \longrightarrow HCN + H \tag{6.40}$$

$$N + HCO \longrightarrow HCN + O \tag{6.41}$$

$$H + H_2CN \longrightarrow HCN + H_2. \tag{6.42}$$

The reactions primarily responsible for the destruction are:

$$H_3O^+ + HCN \longrightarrow HCNH^+ + H_2O \tag{6.43}$$

$$HCN + CRPHOT \longrightarrow CN + H \tag{6.44}$$

$$H_3{}^+ + HCN \longrightarrow HCNH^+ + H_2 \tag{6.45}$$

The aforementioned destruction mechanisms are more efficient in the mid-temperature range than in the high-temperature range. This explains why the value of the ratio drops between 100 and 150 K.

We observe a weak linear relationship between the SHAP value and the metallicity. This is in line with what has been observed previously (Davis et al. 2013). In that work, they considered galaxies with metallicities ranging from 0.1 - 0.6, temperatures between 90

and 220 K with the remainder of conditions not listed in the paper. With the exception of the 200 - 220 K range, the listed conditions overlap with the ones in this Chapter. What they found is that they were able to obtain a separate linear function fitting the log-ratio to the metallicity for each visual extinction value. We know that the greater the visual extinction, the greater the final density of the cloud. Furthermore, fixing the visual extinction and therefore the density fixes the final temperature that our cloud reaches during the warm-up phase. Cosmic ray ionisation rates and the radiation field are also taken to be constant in the observed galaxies. This means that each linear relationship provided in Davis et al. (2013) gives the relationship between the log-ratio and the metallicity when our other four parameters are fixed. As such, it is sensible to state that there is qualitative agreement between the linear marginal SHAP relationship for metallicity in Figure 6.17 and the relationships found in Davis et al. (2013), as both of these assume the other parameters are fixed. Once again, it makes little sense to compare the exact numbers as we consider a far wider range of conditions. However, the qualitative similarity lends support to the validity of this methodology.

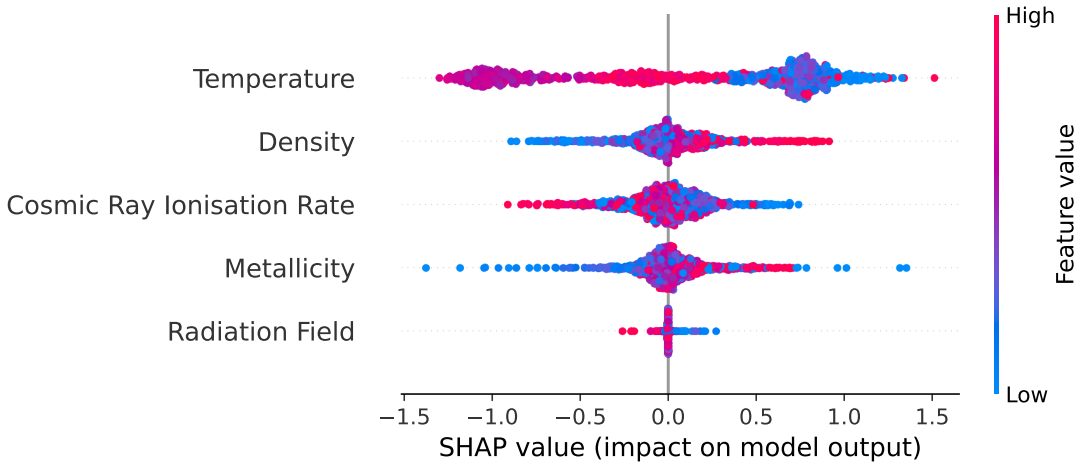

Figure 6.16: *A beeswarm plot of the various physical parameters demonstrating their relative importance in predicting the log-ratio of HCN to CS. We observe that temperature has the largest impact on the model output with SHAP values ranging from -1.5 to 1.5. Density is also found to have a significant impact, which makes sense as it is seen both HCN and CS are dense gas tracers. The next most important features are cosmic ray ionisation rate, followed by the metallicity and the radiation field.*

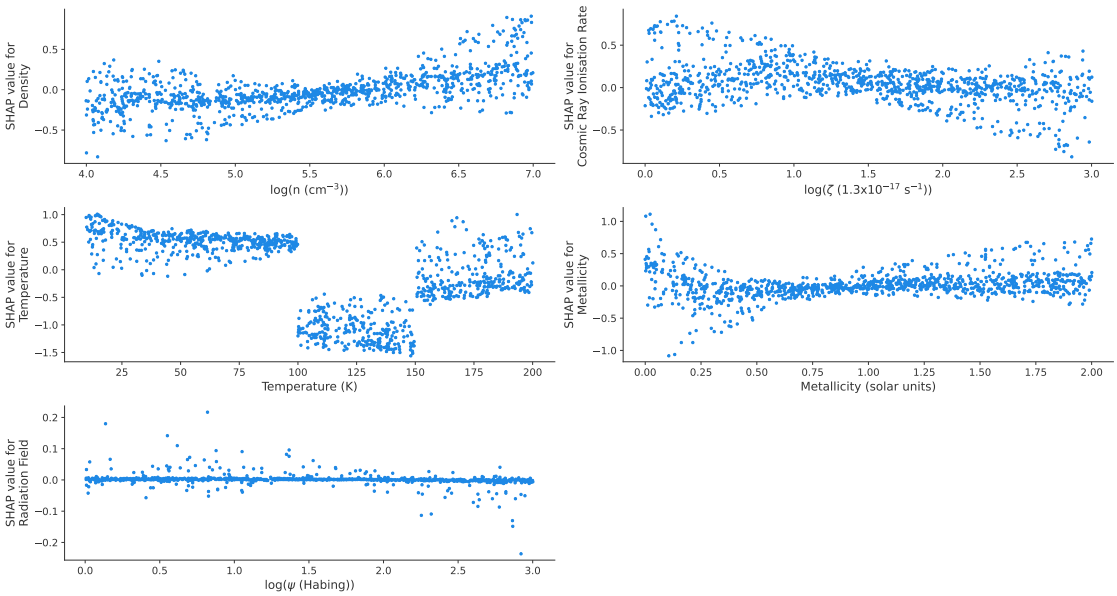

Figure 6.17: *A plot of the SHAP values for the various features (besides the radiation field) as a function of the feature values used to predict the log-ratio of HCN to CS. What we observe is that there exist three separate temperature regimes under which the final abundance is relatively constant. We also notice an increase in the SHAP value as the log-density increases.*

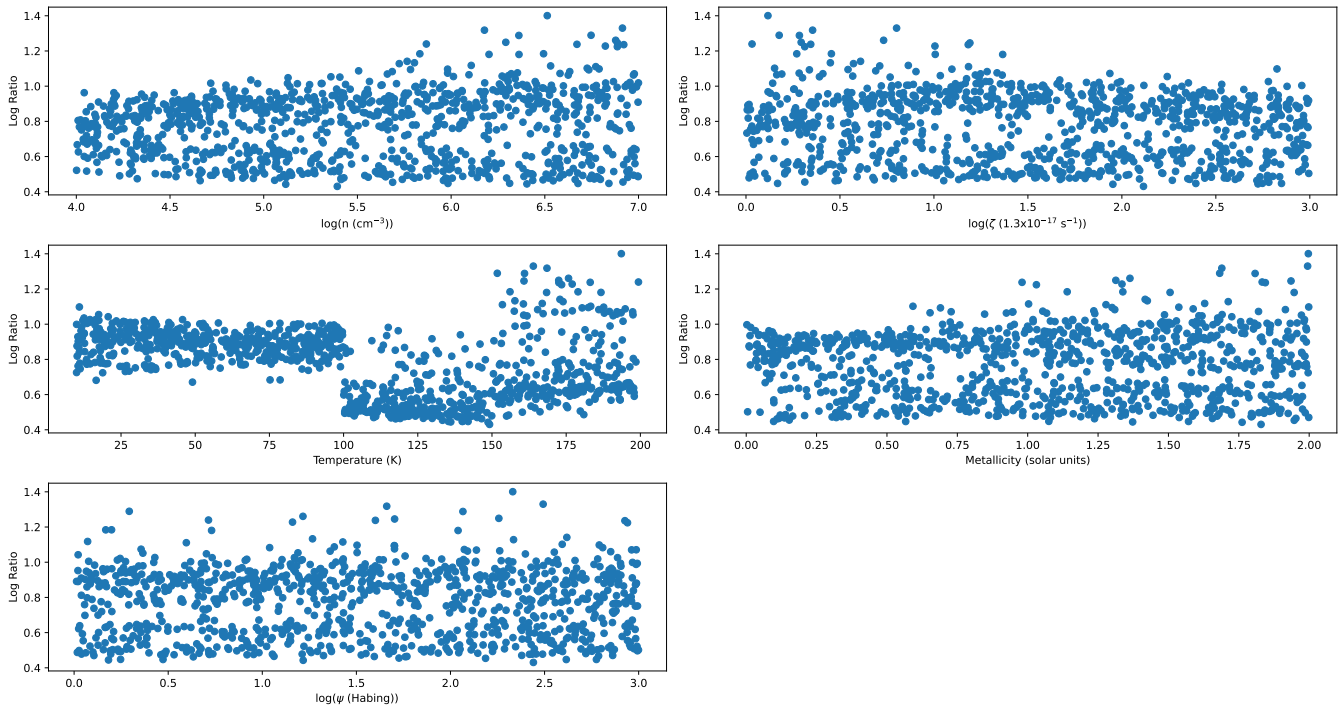

Figure 6.18: *A plot of the log-abundance of HCN/CS ratio as a function of the various features. To calculate the log-ratio for a given data point, we sum up the importance values of each feature for that data point. We observe that only temperature maintains a clear trend relative to what we observed in Figure 6.17. For the other features, we have no discernible trend which can be attributed to the feature importances nullifying each other.*

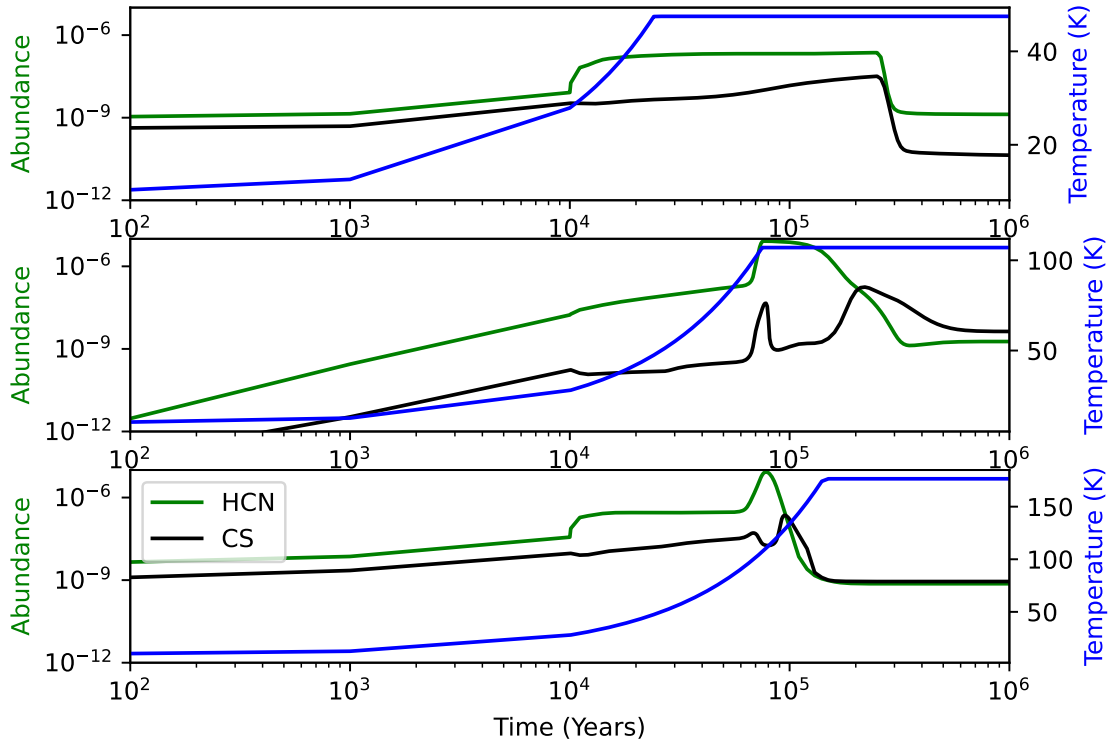

Figure 6.19: *A plot of the abundances of HCN and CS as a function of time for three different temperatures taken from the dataset: 47 K, 107 K and 176 K, each of which is within one of the three temperature regimes we observe in the dependence plots for the HCN/CS ratio.*

6.5 Conclusion

In this Chapter, we present the first application of machine learning interpretability techniques to better understand the effect of various physical parameters on molecular abundances. We trained an XGBoost statistical emulator to replicate the outputs of our chemical model, UCLCHEM. From this, we used SHAP to determine a relative ranking of feature importance as well as to identify the nature of the relationships between the input parameters and the output of interest. A quantitative measure for the relative feature importance was also presented.

This Chapter essentially presents a sensitivity analysis, but is different in many ways to previous studies. This is the first time that the concept of machine learning interpretability has been applied in astrochemistry to consider the impacts of various parameters on abundances. Our methodology offers a number of advantages. In the first instance, by training a statistical emulator to replace our forward model, UCLCHEM, we are able to

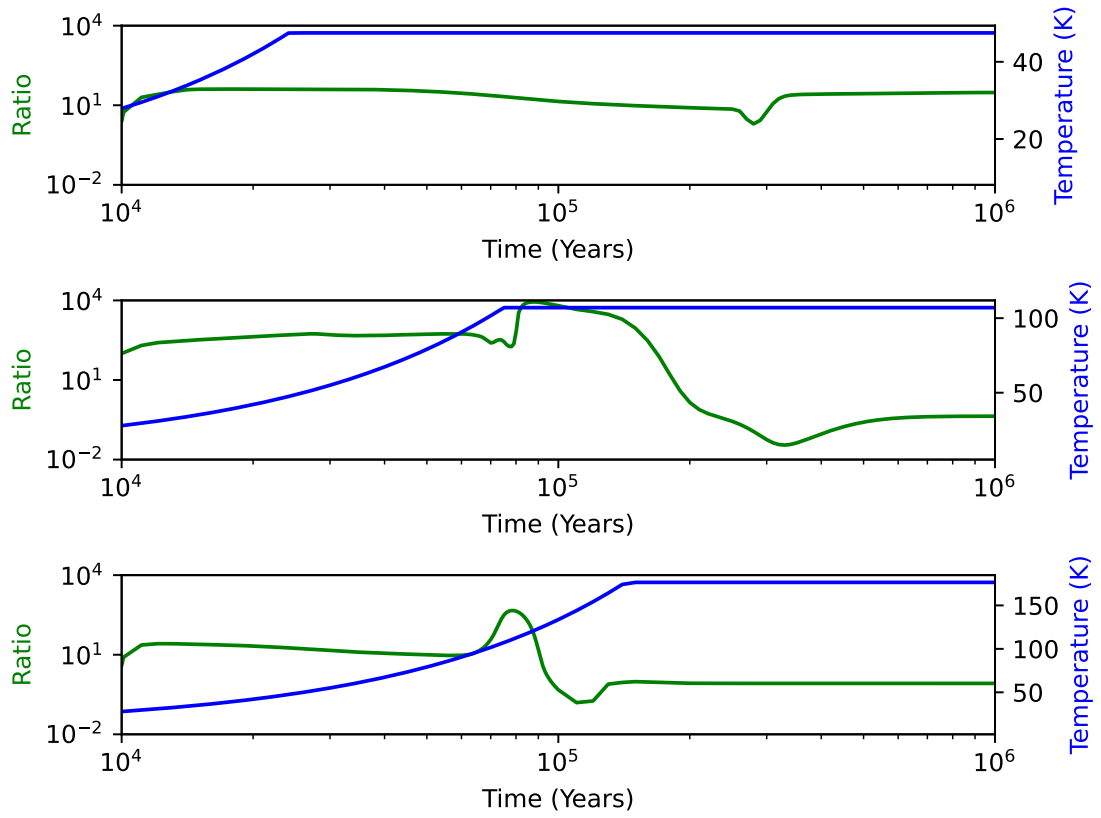

Figure 6.20: *A plot of the ratio of HCN to CS as a function of time for three different temperatures taken from the dataset: 47 K, 107 K and 176 K, each of which is within one of the three temperature regimes we observe in the dependence plots for the HCN/CS ratio.*

significantly reduce the time taken per forward model evaluation, therefore allowing for a much larger grid to be evaluated. Additionally, we are able to quantify the relative importances of the various features as well as comment on the marginal impacts of each of the features.

The main takeaways from this Chapter for the various outputs are as follows:

- H_2O and CO's gas phase abundances depend strongly on the metallicity, which we relate to the fact that a low metallicity results in the production of each molecule being constrained by the amount of the less abundant atomic element (O and C, respectively).

- NH_3 has a strong temperature dependence. There exist two temperature ranges (< 100K and > 100 K) for which the abundance is constant. We are able to relate this to the chemical reactions in our network and find that the increased temperature results in an increase in the destruction pathways.

- We are able to confirm that the HCN/HNC ratio can serve as a cosmic thermometer and find a two-part linear relationship with temperature as in Hacar et al. (2020). However, the dominant HNC destruction at low temperatures is found to be $H_3{}^+ + HNC \longrightarrow HCNH^+ + H_2$ instead of $HNC + O \longrightarrow NH + CO$. We also find a linear relationship between the metallicity and the log-ratio in the range 1-2 which matches what we find in Bayet et al. (2012).

- For the HCN/CS ratio, we observe that it serves as a density tracer, as expected. Furthermore, we once again observe three separate regimes for the temperature dependence, which we are able to relate to the chemistry.

Throughout this Chapter, we have observed similarities between our results and what has been discussed in the literature. This is encouraging. However, it is difficult to make direct quantitative comparisons, as we consider a wide range of physical parameter combinations. On the other hand, the literature we cited considered actual observations. A follow-up study would need to sample the training data for the machine learning model more precisely in order to be able to better model and understand the relationships between inputs and outputs for a specific astronomical object.

This page was intentionally left blank

Chapter 7

Insights from JWST Ice Observations

The work presented in this Chapter is based on the paper Heyl et al. (2023b), in collaboration with Serena Viti and Gijs Vermariën.

7.1 Introduction

In light of the recently released JWST ice observational data McClure et al. (2023), we are going to reconsider the analysis from Chapters 4 and 5. We will look to apply the machine learning techniques from Chapter 6 to consider the relationship between binding energies and abundances.

Giant Molecular Clouds in our Milky Way as well as in other galaxies host gas which is almost entirely molecular, with densities above ~ 100 cm^{-3} and temperatures below ~ 100 K. These denser, cooler regions contain a significant fraction of the non-stellar baryonic matter in a galaxy and they are usually much more massive than large tenuous ones. The importance of these regions lies in the fact that they are key for our understanding of how galaxies form and evolve because this denser, cooler gas is the reservoir of matter that forms stars and planets, as well as the gas that fuels the centres of galaxies.

From an astrochemical point of view, due to their high densities and low temperatures, these regions are great laboratories to study the interactions of gas and dust, with species

from the gas phase 'freezing' onto the dust grains present, and forming icy mantles rich in hydrogenated as well as complex organic molecules (COMs), due to the many fast surface reactions that take place. As stars form in these clouds (or if any other energetic process takes place) then the dust temperature may reach the mantle sublimation temperature ($\sim$ 100 K), and the molecules in the mantles are injected into the gas, where they react and form new, more complex, molecules. Associated with star formation, as well as with AGN activity, are highly supersonic collimated jets and molecular outflows. When the outflowing material encounters the quiescent gas of a molecular cloud, it creates shocks, where the grain mantles are (partially) sputtered and the refractory grains are shattered. Again, here, the interaction of gas and dust varies within very short timescales and the effects of chemistry and dynamics are interlocked in a complex non-linear fashion. In summary, the gas and dust surface compositions exhibit a complicated time dependent, non-linear chemistry that strongly depends on the physical environment. There are many open questions - still - about such interaction: what is the unprocessed ice composition? What are the efficiencies of the viable surface reactions? And how do the energetics of the ISM (cosmic rays, UV radiation, shocks) influence the processed ices? In order to determine accurate estimates of the abundances of molecular species as a function of all the parameters that influence their chemistry we need to be able to answer such questions. In other words, we need to understand the chemical pathways towards each molecule and its dependencies on the density, temperature and energetics of the gas and dust before molecules can be truly considered powerful tools.

In recent years coupling chemical and radiative transfer models for the interpretation of molecular emission has been routinely done and the success of such techniques has varied to different degrees, depending on whether one wants to model the physical and chemical structure, or the hydrodynamical history of the gas (Bisbas et al. 2014; Viti et al. 2014; Kazandjian et al. 2016; Huang et al. 2022). However the shortcomings of such methods are two-fold: (i) understanding the physical conditions in molecular gas via a systematic and applicable to many galaxies methodology is an inverse problem subject to complicated chemistry that varies non-linearly with both time and the physical environment (Makrymallis and Viti 2014); hence it may not have a solution, solutions might not be unique and/or might not depend continuously on the observational data. Traditionally astrochemistry has always been dominated by trial and error grid-based analysis combined with simple statistics (Lefèvre et al. 2014), an approach that becomes

impossible or ineffective when datasets (e.g from ALMA) and/or parameter space are large, complex, or heterogeneous; (ii) the knowledge of the micro-physics and chemistry of what occurs on the dust is well behind what is known for the gas-phase. While surface reactions and dynamics (including desorption and diffusion) can be experimentally investigated (but always within a constrained range of laboratory conditions), experimental data for interstellar ices are still limited. In order to make the best use of experimental resources, the chemical data that models require need to be prioritized according to what will have the most impact.

In recent years progress based on the use of Bayesian as well as Machine Learning (ML) techniques to deal with both the issues above has been made, from the creation of neural network based statistical emulators (de Mijolla et al. 2019; Holdship et al. 2021; Grassi et al. 2022) in order to optimize the integration of chemical, radiative transfer and hydrodynamical models to the use of ML techniques to disentangle multiple gas components in unresolved beams (de Mijolla et al. 2023).

In this Chapter we will focus our attention to Bayesian and ML techniques applied to the study of chemical networks and the key parameters that govern their interactions. In recent years there has been a substantial body of work concentrating on reducing the cost of solving chemical networks computations using various techniques from Monte Carlo approaches to constrain important reactions (Holdship et al. 2018), to automated reduction schemes (Grassi et al. 2012; Xu et al. 2019), to topological methods as in Grassi et al. (2013), finally, ML algorithms (Grassi et al. 2022; Tang and Turk 2022). In parallel several studies have concentrated on the estimation of poorly known reaction rates, with particular emphasis on surface chemical networks: an initial approach considered a simple grain-surface network and applied a Bayesian inference method coupled with Markov Chain Monte Carlo sampling in order to infer reaction rates (Holdship et al. 2018). This was followed up with an approach that considered the topological structure of the network in 3, while 4 exploited the characteristics of the chemical reaction mechanism to significantly reduce the dimensionality of the problem under consideration by simply considering the binding energies and the role they play in the determination of grain-surface chemistry. Chapter 5 using the 'Massive Optimised Parameter Estimation and Data compression' (MOPED) algorithm, helped make predictions about which ice species needed to be detected to reduce the variance of binding energy estimates.

Due to the significant role that binding energies play in grain-surface chemistry, we

shall concentrate on the estimation of binding energies as well as on priorization of the ice species that should be observed with instruments such as the JWST to better improve our understanding of their values. We will then use machine learning interpretability to consider the forward relationship between binding energies and the abundances of species of interest. Our methods are described in Section 7.2. The results are presented in Section 7.3 and a brief conclusion is given in Section 7.4.

7.2 Methodology

In this section we first describe the chemical code we use and the chemical assumptions we make followed by a description of the analytical approach we employ.

7.2.1 The Chemical Code and Network

All modelling in this Chapter is done with the open-source astrochemical code UCLCHEM (Holdship et al. 2017)[1]. The chemistry of a collapsing dark cloud was modelled. The dark cloud collapsed isothermally at 10 K from 10^2 cm^{-3} to 10^6 cm^{-3} over 5 million years. The composition of the ices as a result of the ensuing chemistry was then compared to the recent ice observations with the James Webb Space Telescope (JWST) (McClure et al. 2023).

This Chapter focuses solely on grain-surface chemistry. UCLCHEM employs the grain-surface diffusion mechanism to model chemistry on the grains, as has been described in Section 1.3.1. If we wish to better understand grain-surface diffusion-based chemistry, we must have accurate values of the binding energies of species. For most cases, at 10 K, the reactant with the lower binding energy will dominate the total hopping rate, due to the exponential dependence of the hopping rate on the diffusion energy. Across the literature, there is often significant disagreement when it comes to the values of binding energies (McElroy et al. 2013; Wakelam et al. 2017; Quénard et al. 2018). While there exist many different methods of estimating these values (He et al. 2016; Ferrero et al. 2020; Villadsen et al. 2022), we utilise a Bayesian inference approach.

The chemical network used consists of a gas-phase and ice-phase network. The gas-phase network is the UMIST network (McElroy et al. 2013). The ice network used is the same as in Chapter 4, but augmented with a sulphur network based on work done

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

Species	Abundances relative to H
H_2O	$(8.8 \pm 1.1) \times 10^{-5}$
CO	$(2.2 \pm 0.3) \times 10^{-5}$
CO_2	$(1.1 \pm 0.2) \times 10^{-5}$
CH_3OH	$(3.1 \pm 0.7) \times 10^{-6}$
NH_3	$(8.8 \pm 1.6) \times 10^{-6}$
CH_4	$(1.8 \pm 0.1) \times 10^{-6}$
OCN	$\sim 2.0 \times 10^{-7}$
SO_2	$\sim 6.6 \times 10^{-8}$
OCS	$\sim 1.3 \times 10^{-7}$

Table 7.1: *The abundances and uncertainties taken from McClure et al. (2023). These abundances were taken from sources with an A_v of 95.*

to explain the sulphur depletion problem (Laas and Caselli 2019). The inclusion of the sulphur network is important, since recently sulphur-bearing species have been confirmed in the ices (McClure et al. 2023).

7.2.2 Analytical Approach

Bayesian Inference

One of the goals of this Chapter is to estimate the binding energies of the most diffusive species in the network. These species were chosen based on a literature search that suggested they were amongst the species with the lowest values for their binding energies. The binding energy parameters are represented as a vector, $\boldsymbol{\theta} = (E_{b,H}, E_{b,H_2}, E_{b,C}, E_{b,CH}, E_{b,N}, E_{b,CH_3}, E_{b,NH}, E_{b,CH_4}, E_{b,O})$. UCLCHEM was rewritten so that it would take the vector as an input and output the abundances of species of interest. The mapping between the input and output can be summarised as $\boldsymbol{Y} = f(\boldsymbol{\theta})$, where f represents UCLCHEM. We are looking to estimate the binding energies that give us abundances that match our measurements best. This is an inverse problem, as we are trying to determine the best-fitting inputs that give an output of interest. This was done through the use of Bayesian inference as described in Section 1.5.

The prior for all binding energies was selected as a uniform distribution between 400 K and 2000 K. The abundance measurements, given in Table 4.2, were assumed to be Gaussian. The species without associated uncertainty, OCN, SO_2 and OCS, were given a relative uncertainty of 50%. Assuming a Gaussian distribution, the likelihood function can be specified:

$$P(\boldsymbol{d}|\boldsymbol{E}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(d_i - Y_i)^2}{2\sigma_i^2}\right), \tag{7.1}$$

where n_d is the number of observations and σ_i is the uncertainty of the ith observation. Only the species for which there are abundances measurements are indexed over.

The inference was implemented using the UltraNest Python package (Buchner 2021). The package implements efficient methods to construct a neighbourhood to sample from, allowing for better convergence of the sampling of the likelihood (Buchner 2016, 2019). The package conveniently also outputs the maximum likelihood-estimator, $\boldsymbol{\theta_{ML}}$, which will be utilised later.

The MOPED Algorithm

While our knowledge of the molecular inventory in the gas-phase is quite complete, we are still far from being confident about the ice composition as well as the ice chemistry. To this end, we employ the "Massive Optimised Parameter Estimation and Data compression" (MOPED) algorithm (Heavens et al. 2000, 2017; Heavens et al. 2020), the details of which are provided in Section 5.3.3.

The aim of the algorithm is to determine which of the M species in our chemical network would best constrain our knowledge for our p binding energy parameters. In this Chapter, $p = 9$ and $M = 119$. Some binding energies will have a greater influence on certain species than others. The key is to determine the species that are most sensitive to the binding energies of interest. In doing so, we can then make recommendations for future ice observations as was done in Chapter 4.

Machine Learning Interpretability

The previous methods explore the influence of the abundances on the values of the binding energies. This is an inverse problem. In order to tackle the forward problem of assessing the impact of the binding energies on the abundances instead, one needs to use a different set of methods.

As UCLCHEM solves a system of coupled ordinary differential equations, it stands to reason that the relationship between the input parameters (the binding energies) and the output parameters (the abundances of species) is non-linear. As such, the relationship

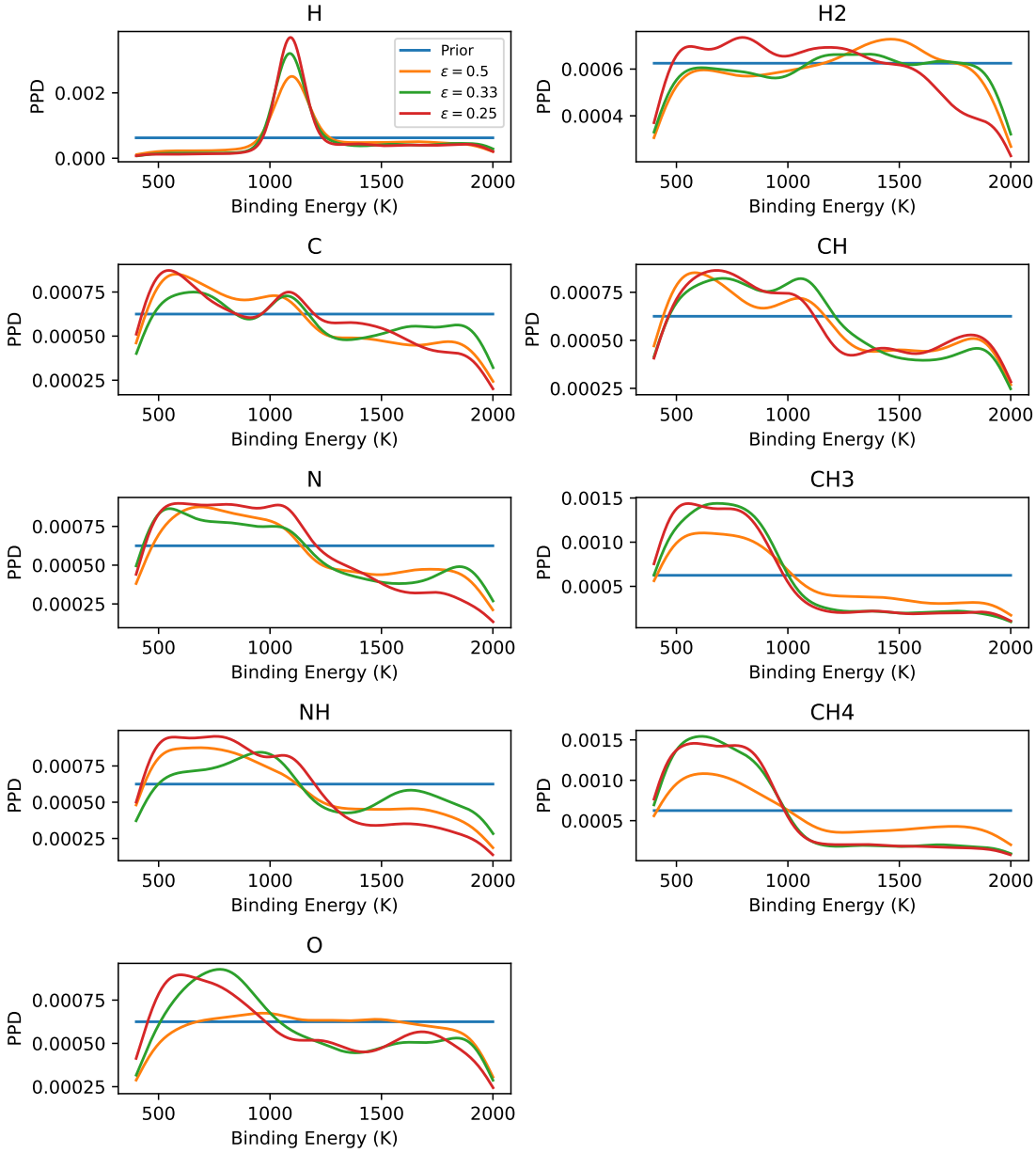

Figure 7.1: *Marginalised posterior distributions of the binding energies of the diffusive species we consider of interest in this Chapter. We also plot the uniform prior distribution. Only H's binding energy marginalised posterior distribution differs significantly from the prior distribution. For the other binding energies, there is less difference. This is due to the lack of enough sufficiently constraining data. We also observe that decreasing the value of ϵ in general decreases the variance of the distribution. Both of these points motivate the need for further ice observations to reduce the variance of the distributions.*

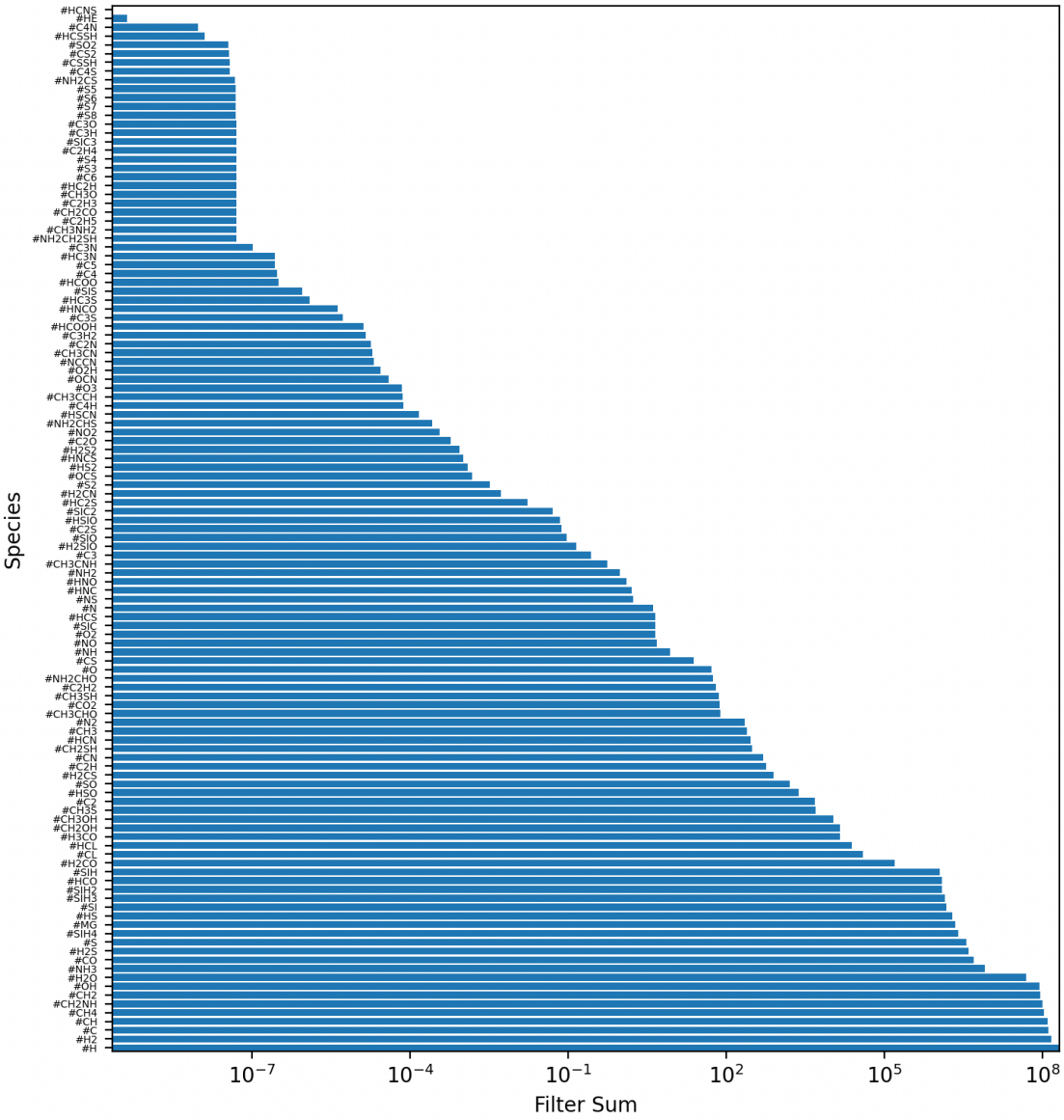

Figure 7.2: *Bar chart displaying the filter sums for all grain-surface species. Species with a larger filter sum are higher priority detection targets, as they are more affected by the binding energies of the species we consider. Some of the highest-ranked species have already been detected, which potentially implies that future observations should aim to improve the level of precision of these abundance measurements.*

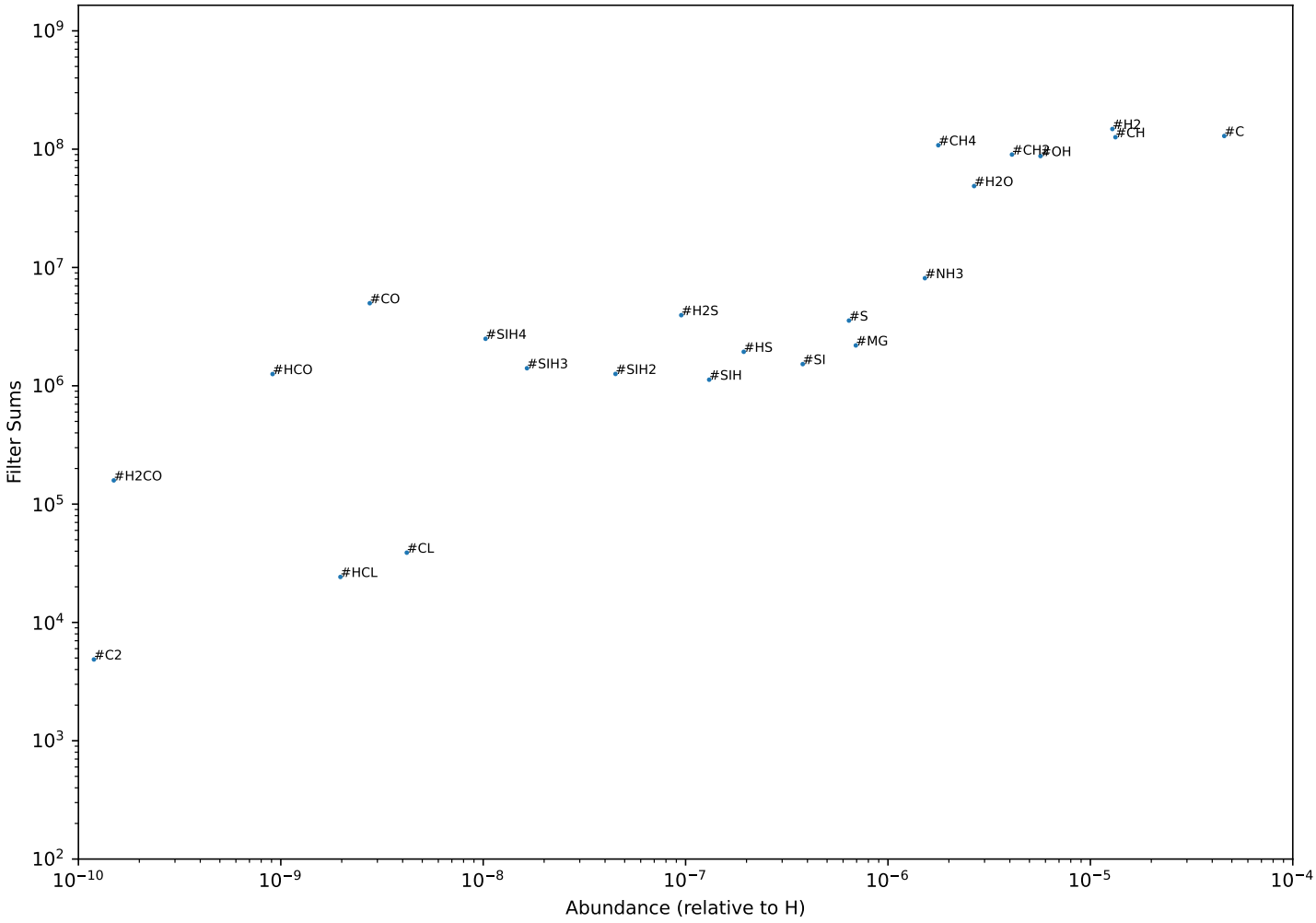

Figure 7.3: *Scatter plot depicting filter sum against the predicted abundances when the maximum-likelihod estimate for the binding energies is input into UCLCHEM. Given constraints on instrumental uncertainties, we should look to prioritise species that are not only important, as determined by their filter sums, but that can also be realistically detected. These include saturated species such as $\#CH_4$, $\#NH_3$, $\#SiH_4$, $\#H_2S$ and $\#H_2O$, as well as their precursors. We find that many of the species we observe are the intermediate species formed during the creation of the saturated species in Table 4.2. This indicates that understanding these intermediate products is essential to better constraining the binding energies of interest.*

between the input and output is not necessarily intuitive and is likely to be different for various 'binding energy regimes'. We make use of machine learning interpretability to help uncover this relationship.

In order to better understand the relationships between the inputs and outputs, we utilise SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017). SHAP approximates Shapley values: these are measures of the marginal contribution of a feature to the output value, relative to the mean value of all output in the dataset (Shapley 2016). This is done by considering various coalitions of feature values. A coalition of features represents all subsets of the total set of features. The Shapley value of a feature represents the average change in the prediction when that feature is included in the coalition of features selected. This change is assessed by considering the change in the prediction when the feature is included, averaged over all coalitions (Molnar 2022). However, this becomes computationally unfeasible as the number of features grows, as the number of subsets grows exponentially with the number of feature. SHAP is particularly useful, as it approximates the Shapley values, greatly reducing the time taken to compute them. This is done through the use of the TreeSHAP algorithm (Lundberg et al. 2018).

500,000 data points were created from UCLCHEM by using a Latin Hypercube sampling scheme (McKay et al. 1979) implemented with the help of the Python surrogate modelling toolbox (Bouhlel et al. 2019). We employ the XGBoost Python package[2] to build an XGBoost regressor (Chen and Guestrin 2016) that is made to fit the relationship between the input parameters and the output abundance for each species.

7.3 Results

7.3.1 Results of the Bayesian Inference

At first, the Bayesian inference was run using the original dataset. However, it was found that despite running the inference in parallel using MPI over 128 cores, that there was no convergence, even after several days. This was attributed to the fact that the model struggled to match the constraints. Many of these constraints have very low relative error, compared to the data used in previous works which typically had relative errors of the order of 50% (Holdship et al. 2018). A nested sampler will move from areas of low likelihood to areas of high likelihood. However, if the model struggles to find combinations

[2]https://xgboost.readthedocs.io/en/stable/index.html

of parameters that lead to a higher likelihood, then it will inevitably take longer to perform the inference. To properly run the inference, a significantly larger computing cluster would be required. As an alternative, we decided to investigate how the relative error, ϵ, impacted the obtained posterior probability distributions. We used values of 0.5, 0.33 and 0.25 and ran the inference each time. Our results are displayed in Figure 7.1. Also plotted are the prior distributions.

We observe that with the exception of hydrogen's binding energy, the binding energy posteriors are prior-dominated. However, it can also be seen that a decrease in the relative error of the data appears to be accompanied by a decrease in the variance of some of the posteriors, such as for CH_3, CH_4, NH and O. This is consistent with lower variance posteriors for H and O binding energy with the artificially reduced uncertainties for H_2O observation in Chapter 4. However, even in this scenario we are finding that our posteriors have a relatively large variance. The best way to address this is to figure out which other species we should observe to further constrain the distributions.

7.3.2 Using the MOPED Algorithm

We now look to analyse the results of the MOPED algorithm. The fiducial model we use is the one with $\epsilon = 0.25$. In Figure 7.2 we plot the filter sums for each species to provide us with an initial ranking. We only consider species formed on the grains. As the UCLCHEM code models both the bulk and the surface abundances, we sum the abundances of each species on the surface as well as in the bulk to provide us with a total abundance on the ices.

However, in order to inform future ice observations, it would be useful to also consider the likely abundances of each species. Ideally, we would wish to observe species that are highly abundant and that have large filter sums. The first requirement means it is easier to observe a species given a particular instrumental uncertainty, whilst the second ensures that we are observing species that are dependent on the binding energies and are therefore relevant to the chemistry we are considering. To do this, we plot the filter sum of each species against the abundance produced when we use binding energies equal to $\boldsymbol{E_{ML}}$. The resulting plot is Figure 7.3. We only consider species with an abundance greater than 10^{-10} relative to H, as anything less abundant is unlikely to be detected in the ices. As in Chapter 4, we observe that the species H_2O, CH_4, NH_3, H_2S, SiH_4, CO and H_2CO are amongst the highest-ranked species with abundances that are predicted to be detectable.

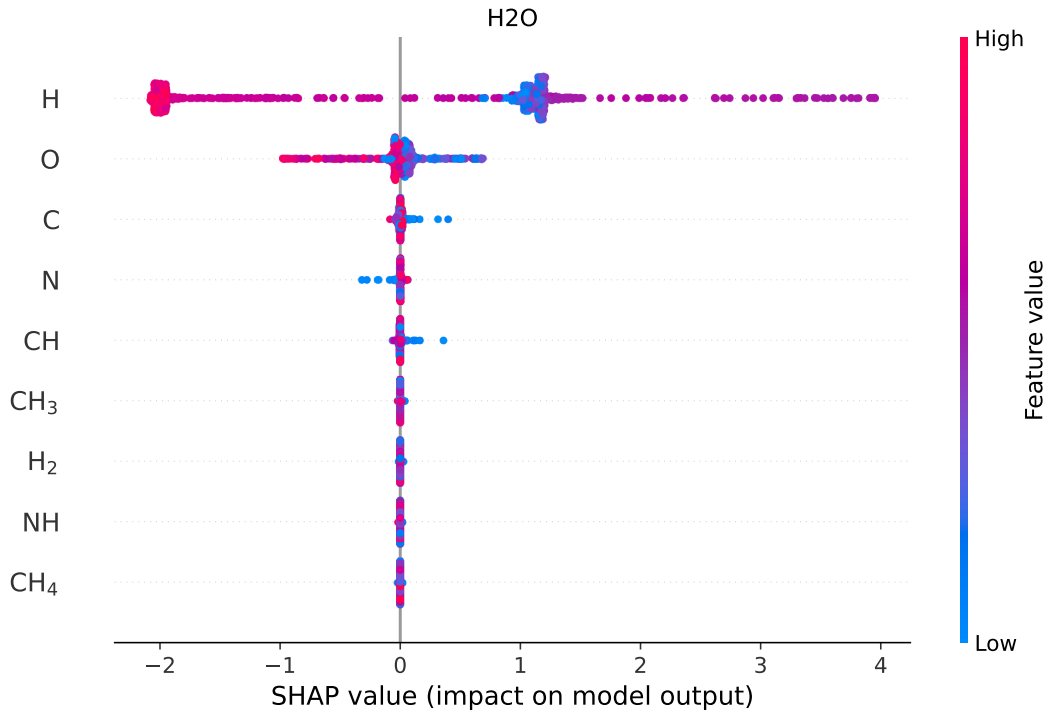

Figure 7.4: *A beeswarm plot for the statistical emulator trained to predict H_2O's abundance. The features are listed from top to bottom in decreasing order of importance to the model output. Along the horizontal axis, individual predictions are plotted in terms of their SHAP value, that is the change to the log-abundance relative to the average value in the dataset. Recall that the SHAP value states the difference in the value of the model output for that prediction relative to the global average. Furthermore, the points are colour-coded in terms of the size of the feature value. We observe that the binding energies of H and O are the most important features. This makes sense, as both species are necessary to form water via successive hydrogenations of an oxygen atom.*

These species all have modes in the range considered by JWST. Unlike in Chapter 5, however, CO_2, CH_3OH and HCN are not amongst the most significant species. This can be attributed to the fiducial model, as we used different constraints, which lead to the maximum-likelihood estimate being different.

7.3.3 Insights from the Machine Learning Interpretability

Previously, we considered the impact of the data, i.e. the species abundances, on the binding energy values and their distributions. We now wish to consider the opposite situation, which is the impact of the binding energy values on the final steady-state abundances of molecules of interest. This is important to consider as the binding energy of a species can be dependent on the ice-composition as well as on the individual sites (Grassi et al. 2020; Das et al. 2021).

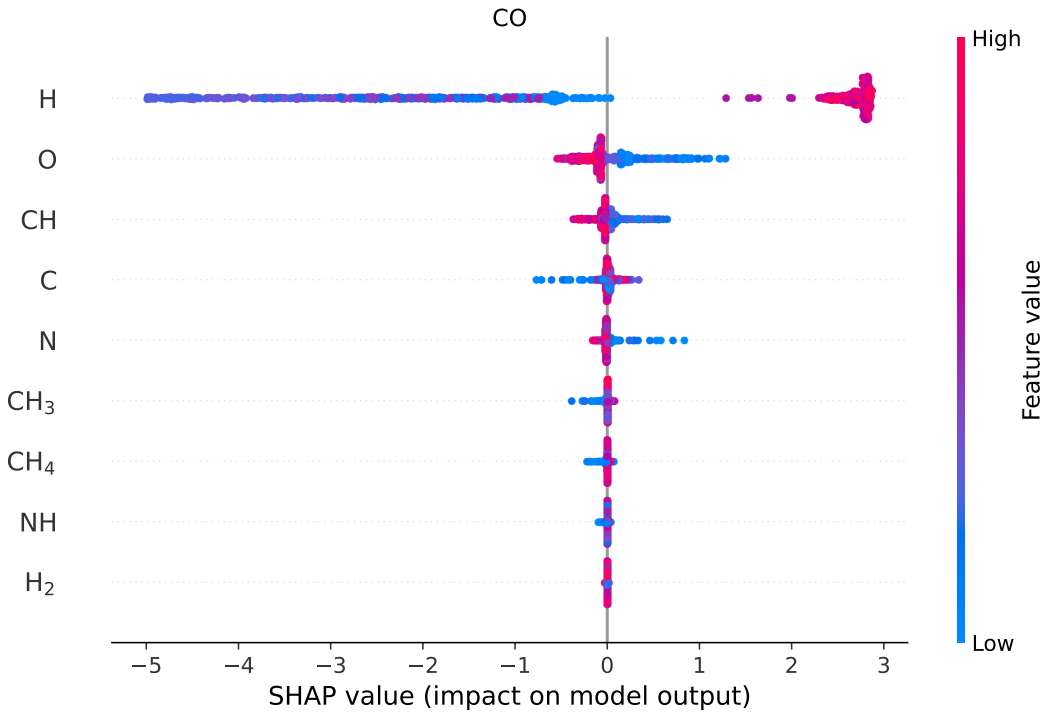

Figure 7.5: *A beeswarm plot for the statistical emulator trained to predict CO's abundance. The features are listed from top to bottom in decreasing order of importance to the model output. Along the horizontal axis, individual predictions are plotted in terms of their SHAP value, that is the change to the log-abundance relative to the average value in the dataset. Recall that the SHAP value states the difference in the value of the model output for that prediction relative to the global average. Furthermore, the points are colour-coded in terms of the size of the feature value. We observe that the binding energies of H and O are the most important features. Increasing H's binding energy appears to increase CO's abundance, which can be attributed to a decrease in the efficiency of the hydrogenation of CO.*

In the interest of brevity, we consider a subset of the molecules so as to demonstrate the effectiveness of this approach as a proof-of-concept. We are interested in better understanding the importance of each of the features in predicting the final abundance of a species of interest, as well as the relative importances of the features. Figures 7.4 and 7.5 are so-called beeswarm plots for H_2O and CO respectively. The features are listed from top to bottom in decreasing order of importance to the model output. Along the horizontal axis, individual predictions are plotted in terms of their SHAP value. Recall that the SHAP value states the difference in the value of the model output for that prediction relative to the global average. Furthermore, the points are colour-coded in terms of the size of the feature value. From this, we can attempt to better understand the directionality of each feature's relationship with the output.

From the beeswarm plots, we can make a number of comments about which binding

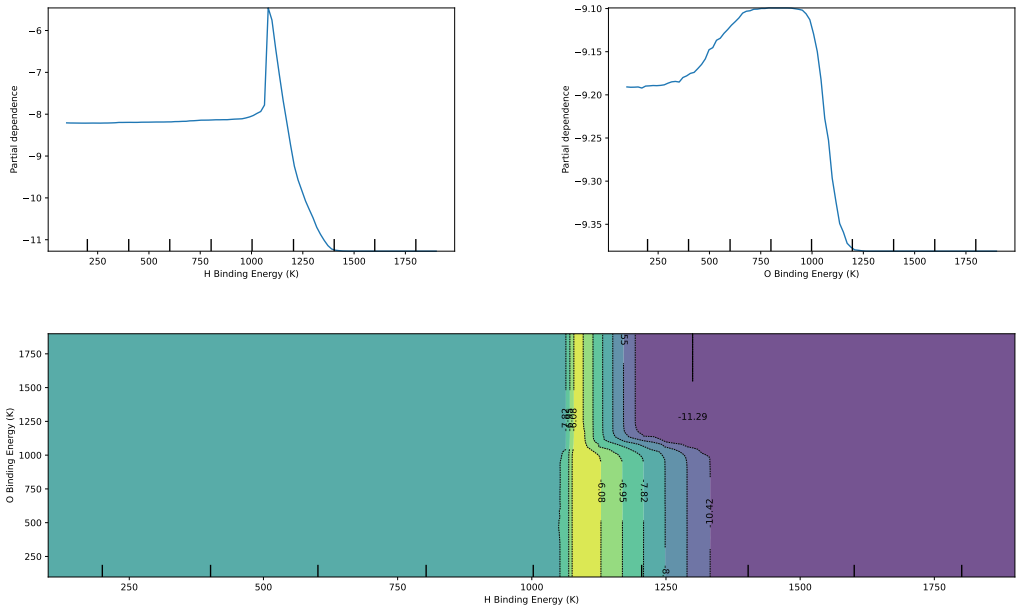

Figure 7.6: *Top: A plot of the 1-D partial dependence plots of the binding energies of H and O for water. The partial dependence represents the expected value of the log-abundance of water as a function of the variable in features, marginalised over all other features. We observe that for a narrow range of atomic hydrogen's binding energies at around 1100 K, there is a sharp increase in the abundance of water. This is roughly the point at which the marginalised posterior distribution for H's binding energy in Figure 7.1 peaks. The dependence for* O *'s binding energy shows a similar consistency with the posteriors, having a clear preference for energies smaller than* $\sim 1000K$*. Bottom: A 2-D partial dependence plot for the binding energies of H and O. Yellow represents the region with the highest abundance of water.*

energies are most relevant for that species. For example, H_2O is unsurprisingly dependent on the H and O. Others seems less intuitive, such as CO's strong dependence on the H binding energy or CO_2's dependence on nitrogen. These can typically be reasoned out by considering the chemical network used.

We can also consider the exact nature of the relationship between the features and the final abundance. To do this, we consider the partial dependence of specific variables relative to the output variable. The partial dependence is defined as the marginal effect of one or several features on the output of a machine learning model (Friedman 2001; Molnar 2022). To demonstrate the utility of the partial dependence, we consider H_2O and CO. Both of these molecules are largely dependent on two binding energies: that of H and O. We plot their 1-D and 2-D partial dependences in Figures 7.6 and 7.7. Note that the y-axis of the 1-D plots are simply the log-abundance of the respective species.

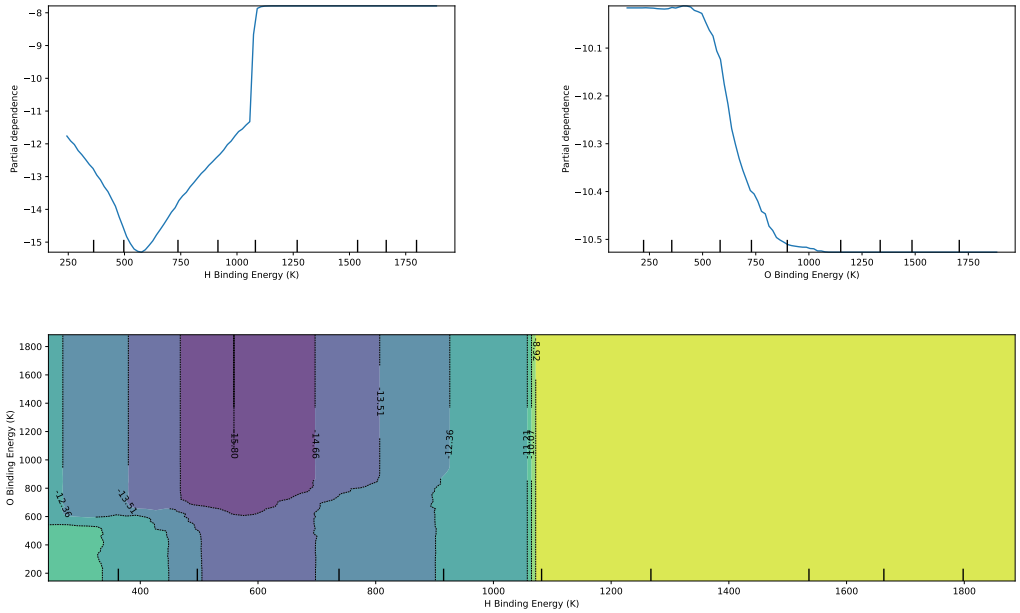

Figure 7.7: *Top: A plot of the 1-D partial dependence plots of the binding energies of H and O for CO. The partial dependence represents the expected value of the log-abundance of CO as a function of the variable in features, marginalised over all other features. Bottom: A 2-D partial dependence plot for the binding energies of H and O. Yellow represents the region with the highest abundance of CO.*

We observe that for water, there is a small area of parameter space in which the abundance peaks. This roughly matches the maximum-posterior hydrogen binding energy value obtained in Figure 7.1. Despite the oxygen's binding energy being the second most important feature, we observe that over the range of binding energies considered, it has far less impact in changing the obtained water abundance. Even so, the parameter favoring binding energies lower than ∼1000 K for oxygen is consistent with the posterior for the inverse problem.

We can make a similar comment about carbon monoxide. The abundance peaks for hydrogen binding energy values greater than 1100 K. This makes sense, as having too low a binding energy for hydrogen would result in CO being hydrogenated efficiently. For binding energies above 600 K for oxygen, we notice a slight decrease in the abundance of carbon monoxide.

7.4 Conclusion

In this Chapter we focus our attention on the estimation of binding energies, key parameters in the interaction among surface reactions in ice. We use three statistical approaches to estimate binding energies, prioritise future ice species to be observed, and to understand better the non-linear relationship between binding energies and abundances of such species. Our conclusions can be summarized as followed:

- As seen in Chapters 3 and 4, we find that Bayesian inference can be a very useful tool to constrain binding energies. However further ice observations are needed in order to reduce the variance of the distributions.

- Indeed, the MOPED algorithm can help towards the prioritization of such observations. As seen in Chapter 5, we find that solid H_2O, CH_4, NH_3, H_2S, SiH_4, CO and H_2CO are the most important species to observe; surprisingly ice observations of CO_2, CH_3OH and HCN are not amongst the most significant species.

- Using SHAP we establish the key relationships between binding energies and the abundances of the ice species. For example, we find that for water and CO the key parameter is the hydrogen binding energy, and to a much lesser extent the oxygen one. Prioritizing which binding energies are keys for the potentially observable species may be of use in prioritizing experiments and calculations of such energies to reduce their errors.

Probabilistic methodologies as well as Machine Learning methods have now started to be used to solve astrochemical problems. As larger chemical reaction networks and more complex models are being employed in astrochemistry, statistical methods and machine learning (ML) techniques will become ever more necessary in order to reduce the uncertainty in such networks.

Chapter 8

Conclusion and future prospects

On a fundamental level, astrochemistry is complex and highly non-linear. This often makes understanding the relationship between various astrochemical parameters and the abundances of various species difficult to interpret. The focus of this thesis has been to develop methodologies to improve our understanding of astrochemistry. This has been done through chemical modelling in Chapter 2, Bayesian statistics in Chapters 3 and 4, the MOPED algorithm in Chapter 5, and machine learning interpretability in Chapters 6 and 7.

In Chapter 2, we considered the impact of adding two H_2-based addition reactions, $C + H_2 \longrightarrow CH_2$ and $CH + H_2 \longrightarrow CH_3$, to a glycine grain-surface chemical network. We found that including molecular hydrogen as an active chemical participant fundamentally changed the hydrogen economy of the system. Previously, grain-surface chemistry has not had many H_2-based reactions, leading to a significant untapped hydrogen reservoir being stored there. With the presence of these reactions, atomic hydrogen is freed up which allows for further hydrogenation reactions to take place elsewhere in the network. As a result, simple hydrogenation productions such CH_4, CH_3OH and NH_3 see increases in their abundances. On the other hand, more complex species such as glycine and its precursors, which rely on the atomic addition of productions of hydrogenation see decreases in their abundances as their reactants are preferentially hydrogenated due to the greater efficiency of the hydrogenation process. This could potentially offer an explanation as to why glycine has not yet been confirmed to be detected. It is possible that the models have

overestimated its abundance. What this demonstrates is that more work needs to be done to focus on establishing a chemically sound glycine network that would build upon the work done in Ioppolo et al. (2020).

Applying Bayesian inference to estimate reaction rate parameters of grain-surface reactions is difficult due to the paucity of available data. There have simply not been enough grain-surface species detections to make inferences. This problem is exacerbated by the large uncertainties present in the available measurements. What this results in is degeneracies in our posterior probability distributions. Furthermore, from a computational point of view, we end up finding that the time taken for the inference process to complete is excessive. We considered a number of approaches to deal with these two problems.

In Chapter 3 we explored how the network topology could be used to address both of these problems. We considered a toy network from Holdship et al. (2018) and looked at how by stripping away certain reactions, we could reduce the variance of our posterior distributions. We also considered how the inclusion of a single "dummy reaction" could help us recover our the distributions obtained from the original network, while reducing the time taken for the inference. We also showed that we could divide up our network and perform inference on each part separately.

Chapter 4 was dedicated to addressing the same problem, but instead of taking a manual approach to reducing the complexity of the problem, we considered a physics-based one. Specifically, we demonstrated how the reaction rates of the grain-surface reactions were ultimately governed by the binding energies of the more mobile reaction. In doing this, we reduced the dimensionality of our inference problem from 49 reactions to 14 binding energies making the inference process computationally feasible. However, even this proved to not be enough to completely eliminate any degeneracies. We attribute this to the fact that while the ratio of parameters to infer to data points was not as large when inferring binding energies, we still did not have enough data.

Chapter 5 focussed on addressing this issue. We employed the MOPED algorithm to identify specific species whose detections would reduce the variance of our posterior distributions. The species we identified were: H_2O, CO_2, NH_3, CH_4, CO, CH_3OH, H_2CO, HCN, SiH_4 and H_2S. Our recommendations to future ice observations for JWST were based on the range of wavelengths that were going to be considered. However, the fact that many of our recommended species are hydrogen-based suggests that our network could stand to be expanded to include a more diverse set of species. Hydrogenation will

be one of the most efficient processes on the grains, by virtue of its low binding energy, and therefore its products will have the greatest abundances. However, grain-surface chemistry extends beyond hydrogen-based species, so a more developed network could go some way towards providing us with a better chemical understanding.

Chapter 6 presented the first application of machine learning interpretability techniques to the study of astrochemistry. We investigated how various physical parameters had an impact on the abundances of species of interest as well as on tracer ratios. A large parameter space was used for this study unlike many of the observational studies we compared our results to. Despite this, we managed to achieve qualitative agreement. In order to achieve quantitative agreement with studies considering specific objects, the parameter space would have to be significantly reduced to match the range of conditions for that specific astronomical setting. However, our work was meant to serve as a proof-of-concept of the utility of SHAP in better understanding the non-linear relationship between physical parameters and output abundances.

In Chapter 7, we utilised the techniques from Chapter 2 to 6. In light of the new JWST ice observations we attempted to update our analysis using Bayesian inference and MOPED. We find that despite the greatly reduced variance in our data points, the lack of a sufficiently high number hinders our ability to run the inference due to the specificity of the small number of constraints. This demonstrated that we need low-variance data points in high numbers to reduce the computational costs. The MOPED analysis suggests the detection of similar molecules would reduce the variance, though due to the change in the likelihood function and the data, not all of the molecules in Chapter 6 are recommended again. This was followed up by applying SHAP to the analysis of the relationship between binding energies and H_2O and CO's abundances, we were able to identify the most important values. These insights could prove to be useful when making recommendations for planning experiments to determine the binding energies of species of interest.

There are many further avenues to pursue following on from this work. One of the recurring issues that was encountered during the inference process was the computational time. This will continue to be a problem for several reasons. For one, the inference process requires the evaluation of the forward model each time the likelihood function is evaluated. A single forward model evaluation of UCLCHEM takes of the order of a minute. With many Bayesian inferences requiring hundreds of thousands of evaluations, this becomes

computationally expensive. To address this, future work might want to consider utilising statistical likelihood emulators in which an emulator is trained to reproduce the output of the likelihood function. This has found widespread use in the field of cosmology through the use of emulators such as the "Blind Accelerated Multi-Modal Bayesian Inference" (BAMBI) algorithm (Graff et al. 2012). There also exist more modern methods such as normalizing-flow enhanced Markov chain Monte Carlo (MCMC) sampling methods that specialise in high-dimensional spaces Gabrié et al. (2022).

While we established in Chapters 3, 4 and 5 that a lack of sufficiently constraining data was hindering out ability to produce low-variance posterior distributions, there is more work that could be done. It is clear from Appendix A that we are in a prior-dominated regime, which means that for many parameters our priors "wash out" the information provided by our data. This is particularly problematic for the binding energies of most of the species besides hydrogen, as hydrogen's binding energy impacts a lot of our observed abundances. To rectify this, a potential approach would be to consider a least-informative prior approach which has found widespread use in the field of Bayesian inference (see Heavens and Sellentin (2018) for an example). Such a prior will have the smallest impact on the posterior distribution, allowing the data to dominate as much as possible. The variance of our posteriors will stem from the variance due to our likelihood, which is encoded in our data and its associated uncertainties, as well as the variance of our prior. Taking a two-pronged approach to reducing the variance by increasing the number of low-variance abundances would alongside making our prior as least-informative as possible would go a long way towards achieving this goal.

Appendix A

Appendices to Chapter 3

A.1 Convergence

Any MCMC chain needs to be checked for convergence. In the limit of an infinitely long chain, the sampled posterior distribution can be said to approximate the true posterior. However, when dealing with a finitely long chain, one needs to check that the posterior is not changing by much. We made use of two diagnostics to check. However, it should be emphasised that these diagnostics do not guarantee that the the chains are converged. Rather, if these checks fail, then we know the chain has not converged. Satisfying the conditions of the diagnostics simply lends credence to the hypothesis that the chains have converged.

A.1.1 Geweke Diagnostic

The Geweke diagnostic calculates a z-score between two sections of a chain, typically the first 10% and the final 50% (Geweke (1991); Roy (2020)). The z-score is calculated by

$$z = \frac{\overline{\theta_a} - \overline{\theta_b}}{\sqrt{\overline{\sigma_a^2} + \overline{\sigma_b^2}}}, \tag{A.1}$$

where the quantities with subscript a refer to the first 10% of the chain and the quantities with subscript b refer to the final 50%. In the limit of the chain length going to infinity, the Geweke diagnostic is expected to follow a normal distribution (Cowles and Carlin

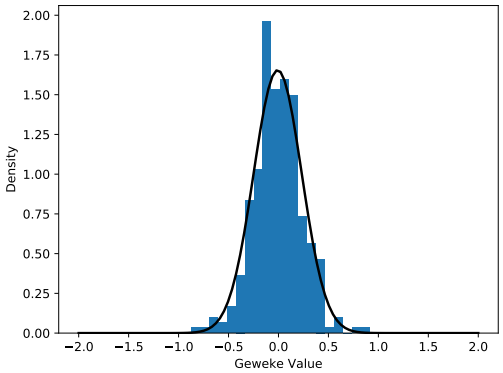

Figure A.1: *A plot of the distribution of the Geweke diagnostic for the samples obtained in Configuration 6. We find that most of the points are within 2 z-scores of the mean. A normal distribution with zero mean is overlaid to show that the Geweke diagnostic is approaching a normal distribution.*

(1996)). In Figure A.1, we plot the distribution of the Geweke diagnostic for the chains of configuration 1 along with a normal distribution overlaid. We observe that the vast majority of points diagnostic stay within one standard deviation of the mean.

A.1.2 Gelman-Rubin

The Gelman-Rubin diagnostic provides a means of comparing the variance across all chains with the variance of the individual chains Gelman and Rubin (1992); Hogg and Foreman-Mackey (2018). There are several quantities of interest at play here. They are calculated for each scalar parameter of interest separately (Gelman and Rubin (1992); Gelman and Shirley (2011)).

For m chains, assumed to be of length n, the between-chain variance is defined as

$$B = \frac{n}{m-1} \sum_{i=1}^{m} \left(\widehat{\theta}_i - \widehat{\theta}\right)^2, \tag{A.2}$$

where $\widehat{\theta}_i$ is the estimator of the mean for chain i and $\widehat{\theta}$ is the estimator of the mean of the sample. The latter is simply the average of all chain mean estimators.

The within-chain variance, W, is defined as the average of the variances of all chains

$$W = \frac{1}{m} \sum_{i=1}^{m} \widehat{\sigma_i^2}, \tag{A.3}$$

where $\widehat{\sigma_i^2}$ is the estimator of the variance for chain i.

The pooled variance estimate is defined as

$$\widehat{V} = \frac{n-1}{n}W + \frac{B}{n}. \tag{A.4}$$

The quantity of interest is

$$\widehat{R} = \frac{\widehat{V}}{W}, \tag{A.5}$$

which is referred to as the potential scale reduction factor (PSRF). In the limit of infinitely long chains, $\widehat{R}$ tends to 1 from above. The closer $\widehat{R}$ is to 1, the better. In practice, a cut-off is used to determine convergence, typically 1.1 though these is some debate surrounding this (Vats and Knudson (2018); Roy (2020)). The PSRF is related to the autocorrelation time that was used in H18, in that a value of $\widehat{R}$ that satisfies the criterion corresponds to a chain length greater than the autocorrelation time. There exists some debate as to whether to discard the first half of the chain when evaluating $\widehat{R}$ (see Roy (2020) and Gelman and Shirley (2011) for two opposing views on the matter). Regardless of whether we do this, we find $\widehat{R} \leq 1.09$ for the reaction rates of the non-constrained reactions and $\widehat{R} \leq 1.03$ for the constrained reactions.

A.2 Bayesian Sensitivity Analysis

Bayesian parameter inference depends on two fundamental quantities: the likelihood and the prior. It is important to understand the relative information content of these two quantities to determine whether one's conclusions are driven by one's initial beliefs or the data at hand. In Chapter 3, we have used a log-uniform prior in the reaction rates, k, encoding our ignorance of their order of magnitude (Jeffreys 1946). One of the key aspects when performing Bayesian inference is to determine if the posterior distributions are driven by the data or by the prior. This was considered in detail by Fischer (2019) in the context of chemical kinetics, but has also been discussed with regards to climate models in Tomassini et al. (2007). In Chapter 2, we motivated our choice of prior using chemical considerations and therefore used a log-uniform prior in the reaction rate parameter vector $\boldsymbol{\theta}$ (Chuang et al. 2016; Ioppolo et al. 2020). Fischer (2019) argues that if we are completely ignorant about $\boldsymbol{\theta}$, then we are also ignorant in $f(\boldsymbol{\theta})$, where f is an arbitrary function in $\boldsymbol{\theta}$. We could equivalently use a prior that is uniform in $f(\boldsymbol{\theta})$. If the posterior distributions

differ depending on the choice of prior, this means that the posteriors are "prior-driven", as opposed to "data-driven".

In this appendix, we repeat a portion of our analysis using two alternative priors in order to demonstrate the impact of different prior assumptions. Figure A.2 shows the posterior distributions for Configuration 1 using three different priors. Alongside the uniform prior in $y = log(k)$ that we have used throughout Chapter 2, we also use uniform priors in $t = \frac{1}{log(k)}$ and $u = k$. We observe that the posteriors can differ significantly depending on the prior. However it is interesting to note that the posteriors for reactions 1 and 2 are relatively similar when uniform priors in y and t are used. As was discussed in section 3.5.1, this is due to the fact that these reactions are related to the formation of H_2O, the abundance of which is known to differ from zero at the 3σ level. This appears to suggest that the marginal probability distributions for Reactions 1 and 2 are more "data-driven" than the posteriors for the other reactions. The other reactions are associated with the weaker abundance constraints. Their associated posterior distributions do not rule out much of the parameter space, so it is unsurprising that the posteriors are affected by the priors. We observe similar trends for the posterior distribution for Configuration 2, shown in Figure A.3. We have cropped the ordinate at 1 in order to make sure the posteriors that came from using uniform priors in y and t are more visible.

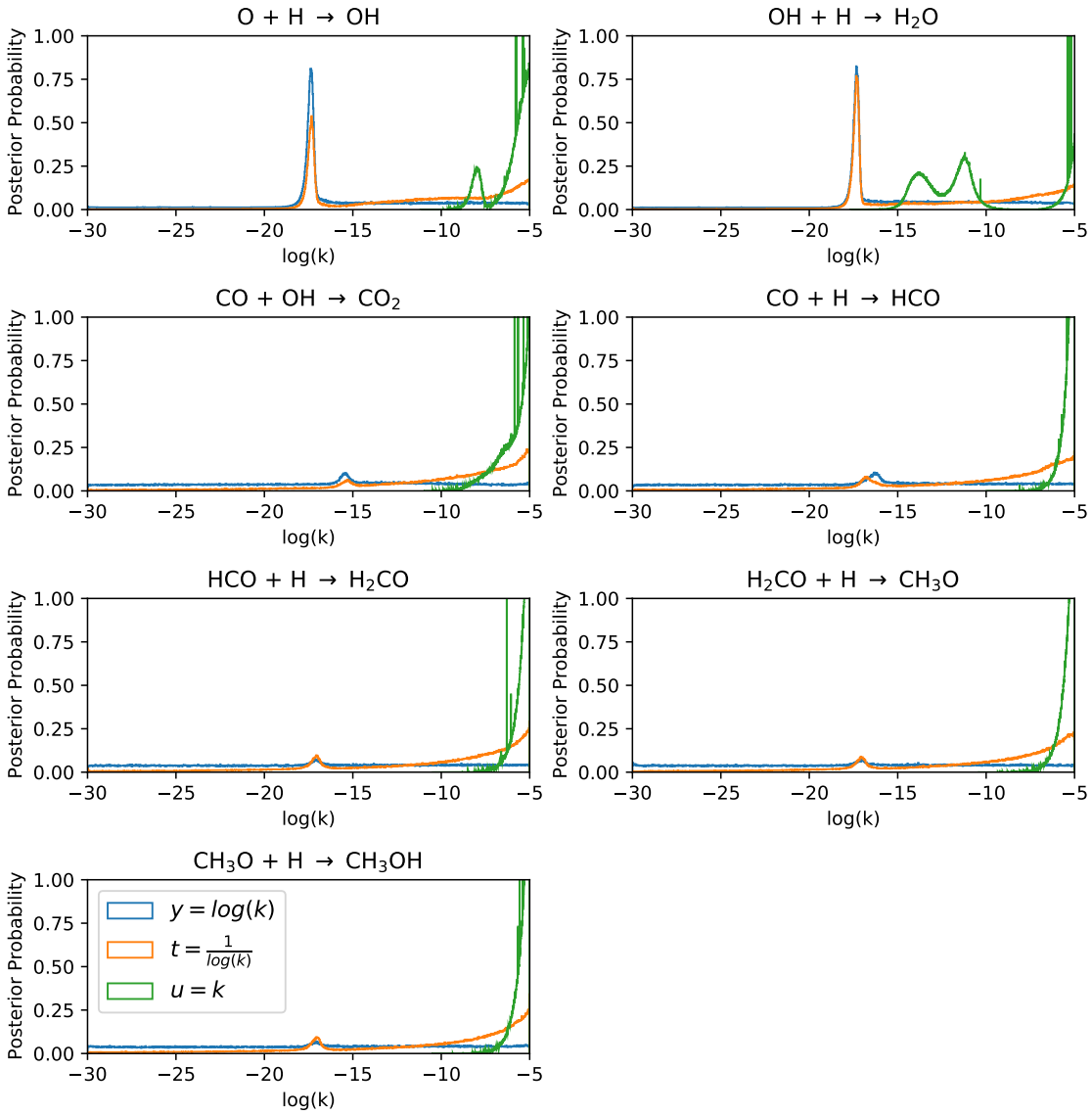

Figure A.2: *Posterior probability distributions for Configuration 1 using three different priors. Alongside the uniform prior in $y = log(k)$, we also consider uniform priors in the variables $t = \frac{1}{log(k)}$ and $u = k$.*

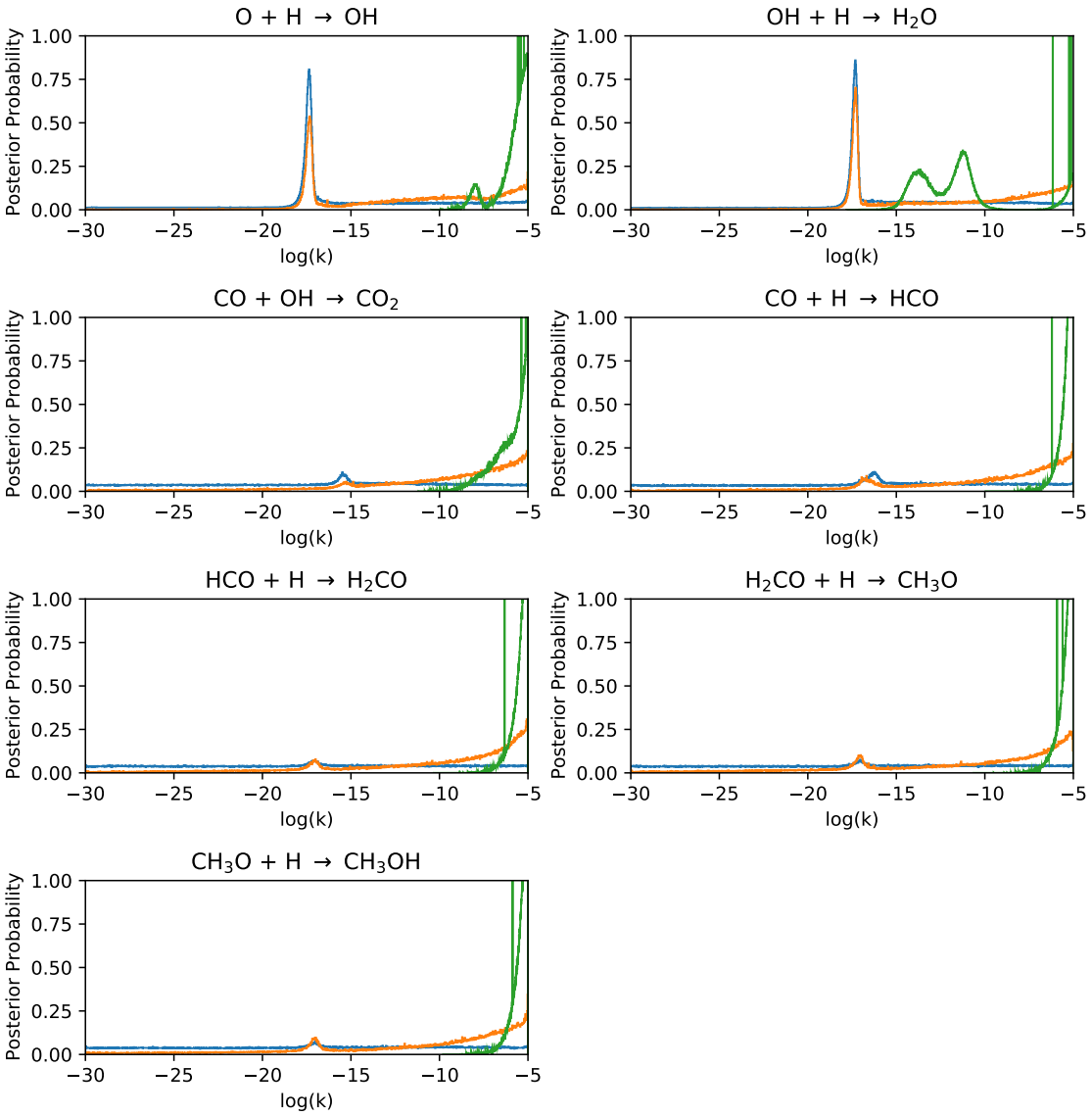

Figure A.3: *Posterior probability distributions for Configuration 2 using three different priors. Alongside the uniform prior in $y = log(k)$, we also consider uniform priors in the variables $t = \frac{1}{log(k)}$ and $u = k$.*

Appendix B

Appendices to Chapter 4

B.1 Evaluating the Frequentist Properties of the Bayesian Estimators

In this section, we seek to determine whether the constraints imposed are significantly influencing the resulting posterior distribution. To determine this, we run the forward model using reaction rates drawn from a Gaussian distribution with a mean value equal to the maximum-posterior reaction rate obtained in Chapter 4 (which represents the "true" reaction rate) and a standard deviation equal to 1. Bayesian inference was then used to recover these reaction rates. This was repeated 20 times using the statistical emulator. The strip plots in Figure B.1 show the values of the reaction rates recovered with the associated uncertainties. This analysis is meant to demonstrate to what extent the constraints imposed by Equation 4.2 are influencing the posteriors obtained.

It becomes clear that the extent to which the constraints affect the posteriors depends on the parameter. For the hydrogenation reaction rate, we see that the 65% highest density regions contain the true reaction rates used in the simulation 65% of the time, that is for 13 of the 20 strips. We find that the high density regions are jittered around the true value, suggesting there is no bias. Overall, what we find is that the constraints are significantly influencing the hydrogenation reaction rate posterior. This is perhaps unsurprising given that most of the constrained species are products of hydrogenation. It is also for this reason that the binding energy for hydrogen has the lowest uncertainty.

However, when the other posteriors are considered, it is clear that there is a greater level of prior domination. While the HDRs are all significantly smaller than the prior range from -15 to 0, we still find that within the HDR the posteriors are not as sharp as for the hydrogenation. This can be confirmed visually by considering the posteriors shown in Figures 4.2 and 4.3. While some of the strips, such as for CO-based reaction, show more jitter around the true value, it is clear that more data is required to counter the influence of the prior distribution. However, we observe that there is no bias in the obtained posteriors.

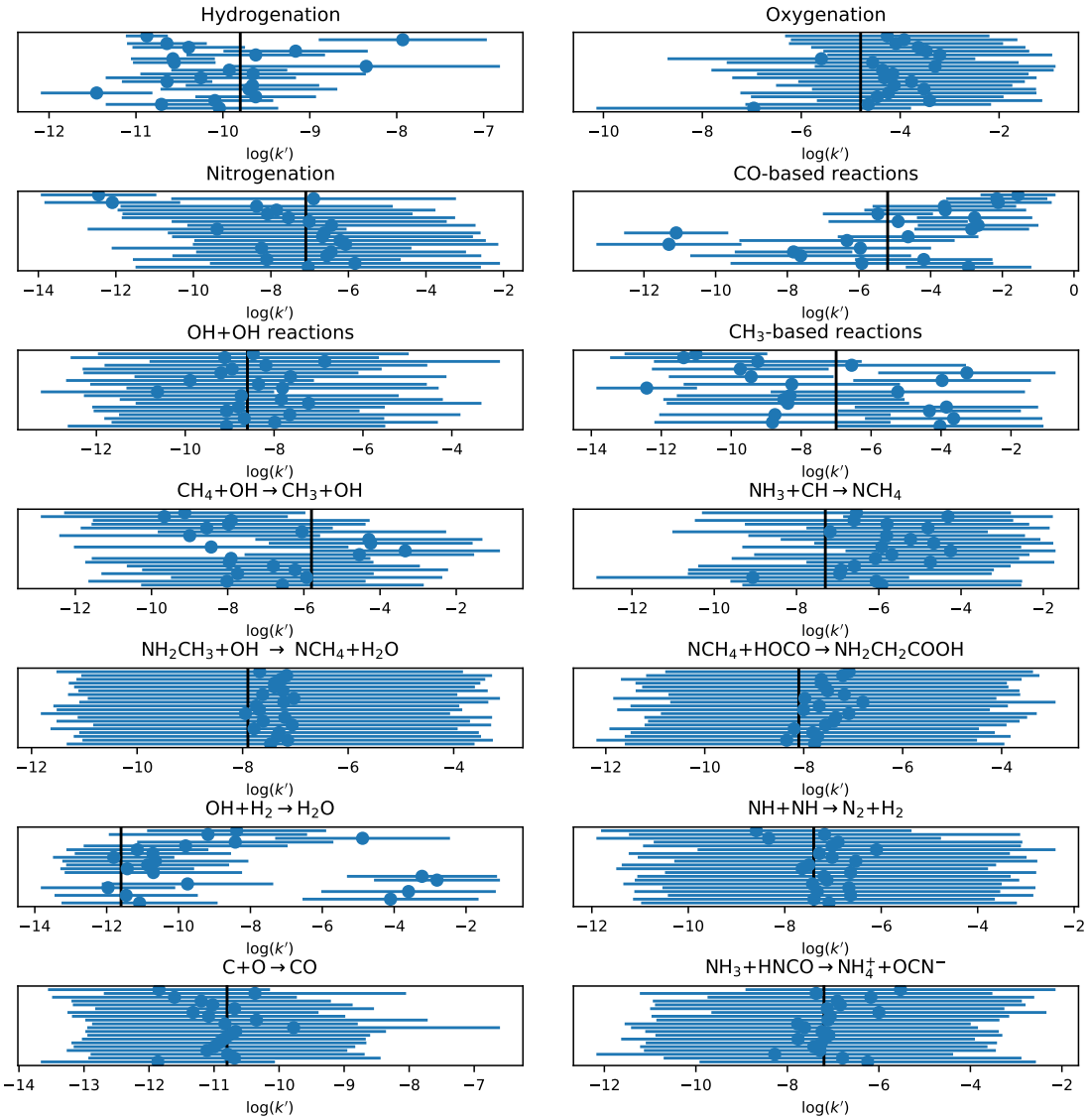

Figure B.1: *A strip plot of the how well the reaction rates are recovered when the forward model is run with some noise on the reaction rates. The vertical black line in each plot represents the "true" reaction rate.*

B.2 Determining the Effect of Constraints on the Inferred Binding Energy Values

We observed that the inclusion of upper limits in the likelihood function given by Equation 4.2 did not have a significant bearing on the binding energy posterior distributions. This suggests that the upper limits listed in Table 4.2 for N_2, O_2, H_2O_2 and glycine may not be sufficiently constraining. In that case it might be more useful to have abundance measurements for these species. To test the effect of these abundance values on the obtained binding energies, we ran the inference 1000 times using the statistical emulator and plotted the distribution of the maximum-posterior binding energy values in Figure B.2. For each inference run, the constraints for each of the species with upper limits in Table 4.2 was taken to be a random value between 0 and the upper limit. These abundance values were sampled uniformly in this range. The relative error on these four measurements was varied to equal 50%, 33.3% and 20%. The relative errors are represented as ϵ.

We observe that the size of the relative errors has some bearing on the maximum-posterior binding energy values obtained as well as the spread of values. For most of the species we observe that there is a significant increase in the spread of inferred values. This demonstrates the importance of detecting further grain-surface species in order to better constrain the binding energy values.

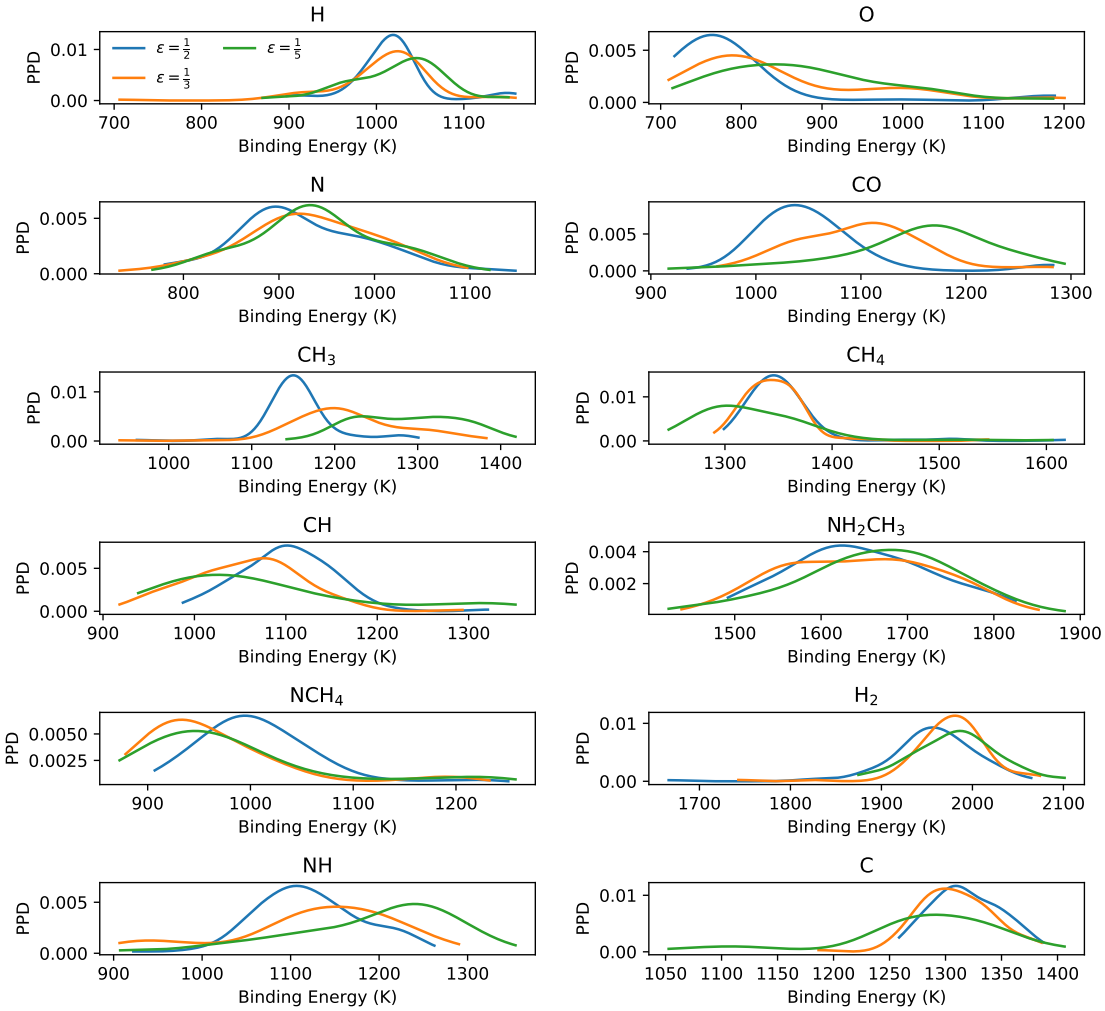

Figure B.2: *Distributions of the maximum-posterior binding energies obtained when the constraints on N_2, O_2, H_2O_2 and glycine are varied.*

Appendix C

Appendices to Chapter 6

Species	Abundance (relative to H nuclei)
He	1.00×10^{-3}
C	1.77×10^{-4}
O	3.34×10^{-4}
N	6.18×10^{-5}
S	3.51×10^{-6}
Mg	2.25×10^{-6}
Si	1.78×10^{-6}
Cl	3.39×10^{-8}
P	7.78×10^{-8}
Fe	2.01×10^{-7}
F	3.60×10^{-8}

Table C.1: *Initial elemental abundances used in UCLCHEM.*

Species/Ratio	$\hat{I}_n$	$\hat{I}_\zeta$	$\hat{I}_T$	$\hat{I}_{m_z}$	$\hat{I}_\psi$
H_2O	0.17	0.15	0.11	0.54	0.02
CO	0.01	0.04	0.04	0.91	0.00
NH_3	0.11	0.14	0.54	0.17	0.03
HCN/HNC	0.03	0.03	0.70	0.24	0.00
HCN/CS	0.16	0.15	0.55	0.13	0.01

Table C.2: *Table summarising the $\hat{I}_i$ for each parameter i.*

Species/Ratio	**Range of Values**
H_2O	$7.7 \times 10^{-12} - 3.8 \times 10^{-6}$ (relative to n_H)
CO	$9.8 \times 10^{-11} - 3.4 \times 10^{-4}$ (relative to n_H)
NH_3	$1.0 \times 10^{-12} - 1.2 \times 10^{-12}$ (relative to n_H)
HCN/HNC	$1.0 - 2418.0$
HCN/CS	$6.0 \times 10^{-3} - 22091.9$

Table C.3: *Table summarising the range of values of the outputs of the abundances and ratios of interest. In the case of NH_3, the lower bound of our values has been clipped at 10^{-12} as dicussed in the text.*

Bibliography

Al-Halabi, A. and van Dishoeck, E. F. (2007). Hydrogen adsorption and diffusion on amorphous solid water ice. *Monthly Notices of the Royal Astronomical Society*, 382(4):1648–1656. 2.1

Ansari, Z., Gall, C., Wesson, R., and Krause, O. (2022). Inferring properties of dust in supernovae with neural networks. *arXiv e-prints*, page arXiv:2207.10104. 6.1

Arasa, C., van Hemert, M. C., van Dishoeck, E. F., and Kroes, G. J. (2013). Molecular dynamics simulations of co2 formation in interstellar ices. *The Journal of Physical Chemistry A*, 117(32):7064–7074. PMID: 23550656. **??**

Auld, T., Bridges, M., Hobson, M. P., and Gull, S. F. (2007). Fast cosmological parameter estimation using neural networks. *Monthly Notices of the Royal Astronomical Society: Letters*, 376(1):L11–L15. 4.3

Awad, Z., Viti, S., Collings, M. P., and Williams, D. A. (2010). Warm cores around regions of low-mass star formation. *MNRAS*, 407(4):2511–2518. 1.2.2

Ayilaran, A., Hanicinec, M., Mohr, S., and Tennyson, J. (2019). Reduced chemistries with the quantemol database (QDB). *Plasma Science and Technology*, 21(6):064006. 3.1

Balucani, N., Bergeat, A., Cartechini, L., Volpi, G. G., Casavecchia, P., Skouteris, D., and Rosi, M. (2009). Combined crossed molecular beam and theoretical studies of the n(2d) + ch4 reaction and implications for atmospheric models of titan. *The Journal of Physical Chemistry A*, 113(42):11138–11152. PMID: 19642633. **??**

Bayet, E., Davis, T. A., Bell, T. A., and Viti, S. (2012). Chemical tracers of high-metallicity environments. *MNRAS*, 424(4):2646–2658. 6.4.2, 6.5

Behrens, E., Mangum, J. G., Holdship, J., Viti, S., Harada, N., Martín, S., Sakamoto, K., Muller, S., Tanaka, K., Nakanishi, K., Herrero-Illana, R., Yoshimura, Y., Aladro, R., Colzi, L., Emig, K. L., Henkel, C., Huang, K.-Y., Humire, P. K., Meier, D. S., Rivilla, V. M., van der Werf, P. P., and Alma Comprehensive High-Resolution Extragalactic Molecular Inventory (Alchemi) Collaboration (2022). Tracing Interstellar Heating: An ALCHEMI Measurement of the HCN Isomers in NGC 253. *ApJ*, 939(2):119. 6.4.2

Belloche, A., Meshcheryakov, A. A., Garrod, R. T., Ilyushin, V. V., Alekseev, E. A., Motiyenko, R. A., Margulès, L., Müller, H. S. P., and Menten, K. M. (2017). Rotational spectroscopy, tentative interstellar detection, and chemical modeling of N-methylformamide. *Astronomy & Astrophysics*, 601:A49. 5.2.2

Benson, P. J. and Myers, P. C. (1989). A Survey for Dense Cores in Dark Clouds. *ApJSS*, 71:89. 6.4.1

Bernstein, M. P., Dworkin, J. P., Sandford, S. A., Cooper, G. W., and Allamandola, L. J. (2002). Racemic amino acids from the ultraviolet photolysis of interstellar ice analogues. *Nature*, 416(6879):401–403. 2.1

Bianchi, E., Ceccarelli, C., Codella, C., Enrique-Romero, J., Favre, C., and Lefloch, B. (2019). Astrochemistry as a tool to follow protostellar evolution: The class i stage. *ACS Earth and Space Chemistry*, 3(12):2659–2674. 6.1

Bisbas, T. G., Bell, T. A., Viti, S., Barlow, M. J., Yates, J., and Vasta, M. (2014). A photodissociation region study of NGC 4038. *Monthly Notices of the Royal Astronomical Society*, 443(1):111–121. 7.1

Black, J. H., Hartquist, T. W., and Dalgarno, A. (1978). Models of interstellar clouds. II. The zeta Persei cloud. *The Astrophysical Journal*, 224:448–452. 2.2.2

Bøgelund, E. G., McGuire, B. A., Hogerheijde, M. R., van Dishoeck, E. F., and Ligterink, N. F. W. (2019). Methylamine and other simple N-bearing species in the hot cores NGC 6334I MM1-3. *A&A*, 624:A82. 2.3.4, **??**

Bonnor, W. B. (1956). Boyle's Law and gravitational instability. *MNRAS*, 116:351. 1.2.2

Boogert, A. A., Gerakines, P. A., and Whittet, D. C. (2015). Observations of the icy universe. *Annual Review of Astronomy and Astrophysics*, 53(1):541–581. (document), 1.3.1, 2.2.3, 2.3.3, 2.3, 3.3.1, 3.3, 3.3.3, 3.6, 3.4, 4.4.2, 4.4.6, 4.2, 5.1, 5.1, 5.3.2, 5.4.3

Boogert, A. C. A. (2016). Telescope Observations of Interstellar and Circumstellar Ices: Successes of and Need for Laboratory Simulations. *IAU Focus Meeting*, 29A:317–318. 1.3.1, 5.1

Bossa, J. B., Duvernay, F., Theulé, P., Borget, F., D'Hendecourt, L., and Chiavassa, T. (2009). Methylammonium methylcarbamate thermal formation in interstellar ice analogs: a glycine salt precursor in protostellar environments. *A&A*, 506(2):601–608. 2.1

Bouhlel, M. A., Hwang, J. T., Bartoli, N., Lafage, R., Morlier, J., and Martins, J. R. R. A. (2019). A python surrogate modeling framework with derivatives. *Advances in Engineering Software*, page 102662. 1.6.2, 4.3.1, 6.3.2, 7.2.2

Bovolenta, G., Bovino, S., Vöhringer-Martinez, E., Saez, D. A., Grassi, T., and Vogt-Geisse, S. (2020). High level ab initio binding energy distribution of molecules on interstellar ices: Hydrogen fluoride. *Molecular Astrophysics*, 21:100095. 4.5.3

Branca, L. and Pallottini, A. (2023). Neural networks: solving the chemistry of the interstellar medium. *MNRAS*, 518(4):5718–5733. 6.1

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. 1.6.4

Brewer, B. J. and Foreman-Mackey, D. (2016). DNest4: Diffusive Nested Sampling in C++ and Python. *arXiv e-prints*, page arXiv:1606.03757. 3.3.2

Buchner, J. (2016). A statistical test for Nested Sampling algorithms. *Statistics and Computing*, 26(1-2):383–392. 5.3.2, 7.2.2

Buchner, J. (2019). Collaborative Nested Sampling: Big Data versus Complex Physical Models. *PASP*, 131(1004):108005. 5.3.2, 7.2.2

Buchner, J. (2021). UltraNest - a robust, general purpose Bayesian inference engine. *The Journal of Open Source Software*, 6(60):3001. 5.3.2, 7.2.2

Buchner, J., Georgakakis, A., Nandra, K., Hsu, L., Rangel, C., Brightman, M., Merloni, A., Salvato, M., Donley, J., and Kocevski, D. (2014). X-ray spectral modelling of the agn obscuring region in the cdfs: Bayesian model selection and catalogue. *A&A*, 564:A125. 4.4.2

Burke, D. J. and Brown, W. A. (2010). Ice in space: surface science investigations of the thermal desorption of model interstellar ices on dust grain analogue surfaces. *Phys. Chem. Chem. Phys.*, 12:5947–5969. (document), 1.2

Butterworth, J., Holdship, J., Viti, S., and García-Burillo, S. (2022). Understanding if molecular ratios can be used as diagnostics of AGN and starburst activity: The case of NGC 1068. *A&A*, 667:A131. 6.4.2, 6.4.2

Caselli, P. and Ceccarelli, C. (2012). Our astrochemical heritage. *The Astronomy and Astrophysics Review*, 20:56. 1.3.1, 2.1

Caselli, P., Sipilä, O., and Harju, J. (2019). Deuterated forms of h3+ and their importance in astrochemistry. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377(2154):20180401. 6.4.1

Chang, Q., Cuppen, H. M., and Herbst, E. (2007). Gas-grain chemistry in cold interstellar cloud cores with a microscopic Monte Carlo approach to surface chemistry. *Astronomy & Astrophysics*, 469(3):973–983. 1.3.1, 2.2.1, 6.2

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv e-prints*, page arXiv:1603.02754. 1.6.4, 6.3.2, 7.2.2

Cheng, I. K., Heyl, J., Lad, N., Facini, G., and Grout, Z. (2021). Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. *Scientific Reports*, 11:19009. (document)

Chuang, K. J., Fedoseev, G., Ioppolo, S., van Dishoeck, E. F., and Linnartz, H. (2016). H-atom addition and abstraction reactions in mixed CO, H_2CO and CH_3OH ices - an extended view on complex organic molecule formation. *Monthly Notices of the Royal Astronomical Society*, 455(2):1702–1712. 3.2, 5.4.3, A.2

Ciesla, F. J. and Sandford, S. A. (2012). Organic Synthesis via Irradiation and Warming of Ice Grains in the Solar Nebula. *Science*, 336(6080):452. 2.1

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904. A.1.1

Das, A., Sil, M., Ghosh, R., Gorai, P., Adak, S., Samanta, S., and Chakrabarti, S. K. (2021). Effect of binding energies on the encounter desorption. *Frontiers in Astronomy and Space Sciences*, 8:78. 7.3.3

Davis, T. A., Bayet, E., Crocker, A., Topal, S., and Bureau, M. (2013). ISM chemistry in metal-rich environments: molecular tracers of metallicity. *MNRAS*, 433(2):1659–1674. 6.4.2

de Mijolla, D., Viti, S., Holdship, J., Manolopoulou, I., and Yates, J. (2019). Incorporating astrochemistry into molecular line modelling via emulation. *Astronomy & Astrophysics*, 630:A117. (document), 1.3, 3.1, 4.1, 4.3, 6.1, 6.2, 6.3.2, 7.1

de Mijolla, D., Holdship, J., Viti, S., and Heyl, J. (2023). Disentangling Multiple Emitting Components in Molecular Observations with Non-negative Matrix Factorization. submitted. 7.1

Fathe, K., Holt, J. S., Oxley, S. P., and Pursell, C. J. (2006). Infrared Spectroscopy of Solid Hydrogen Sulfide and Deuterium Sulfide. *Journal of Physical Chemistry A*, 110(37):10793–10798. 5.4.3

Fedoseev, G., Ioppolo, S., Lamberts, T., Zhen, J. F., Cuppen, H. M., and Linnartz, H. (2012). Efficient surface formation route of interstellar hydroxylamine through NO hydrogenation. II. The multilayer regime in interstellar relevant ices. *The Journal of Chemical Physics*, 137(5):054714–054714. ??, ??, ??

Fedoseev, G., Chuang, K. J., van Dishoeck, E. F., Ioppolo, S., and Linnartz, H. (2016). Simultaneous hydrogenation and UV-photolysis experiments of NO in CO-rich interstellar ice analogues; linking HNCO, OCN^-, NH_2CHO, and NH_2OH. *Monthly Notices of the Royal Astronomical Society*, 460(4):4297–4309. 5.2.2

Feroz, F. and Hobson, M. P. (2008). Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463. 4.4.2

Feroz, F., Hobson, M. P., and Bridges, M. (2009). MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614. 4.4.2

Feroz, F., Hobson, M. P., Cameron, E., and Pettitt, A. N. (2019). Importance Nested Sampling and the MultiNest Algorithm. *The Open Journal of Astrophysics*, 2(1):10. 4.4.2

Ferrero, S., Zamirri, L., Ceccarelli, C., Witzel, A., Rimola, A., and Ugliengo, P. (2020). Binding Energies of Interstellar Molecules on Crystalline and Amorphous Models of Water Ice by Ab Initio Calculations. *The Astrophysical Journal*, 904(1):11. 4.1, 5.1, 7.2.1

Fischer, M. (2019). On the usefulness of imprecise Bayesianism in chemical kinetics. In *ISIPTA 2019 - International Symposium on Imprecise Probabilities Theories and Applications*, volume 103, page 203 à 2015, Ghent, Belgium. Foundations Lab for Imprecise Probabilities, a.k.a. FLip and Ghent University, Belgium, PMLR. 1, A.2

Fraser, H., Williams, D., Sims, I., Richards, A., and Yates, J. (2003). Solid-state astrochemistry in star-forming regions. *Astronomy & Geophysics*, 44(4):4.29–4.33. 3.2

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156. 1.6.4

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232. 7.3.3

Fuchs, G. W., Cuppen, H. M., Ioppolo, S., Romanzin, C., Bisschop, S. E., Andersson, S., van Dishoeck, E. F., and Linnartz, H. (2009). Hydrogenation reactions in interstellar CO ice analogues. A combined experimental/theoretical approach. *Astronomy & Astrophysics*, 505(2):629–639. 5.4.3

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. (2022). Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Science*, 119(10):e2109420119. 8

Galagali, N. and Marzouk, Y. M. (2019). Exploiting network topology for large-scale inference of nonlinear reaction models. *Journal of the Royal Society, Interface*, 16(152):20180766. 3.4.1

Garrod, R. T. and Herbst, E. (2006). Formation of methyl formate and other organic species in the warm-up phase of hot molecular cores. *Astronomy & Astrophysics*, 457(3):927–936. 5.2.2

Garrod, R. T., Widicus Weaver, S. L., and Herbst, E. (2008). Complex Chemistry in Star-forming Regions: An Expanded Gas-Grain Warm-up Chemical Model. *The Astrophysical Journal*, 682(1):283–302. 2.3.1, 5.2.2

Garrod, R. T. and Pauly, T. (2011). On the Formation of CO_2 and Other Interstellar Ices. *The Astrophysical Journal*, 735(1):15. 1.3.1, 2.2.1, 2.3.4, 2.4, 6.2

Garrod, R. T. (2013). A Three-phase Chemical Model of Hot Cores: The Formation of Glycine. *The Astrophysical Journal*, 765(1):60. 1.3.1, 2.1, 3.7, 4.2.2

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7(4):457–472. A.1.2

Gelman, A. and Shirley, K. E. (2011). Inference from simulations and monitoring convergence. A.1.2, A.1.2

Gensheimer, P. D., Mauersberger, R., and Wilson, T. L. (1996). Water in galactic hot cores. *A&A*, 314:281–294. 6.4.1

Gerakines, P. A., Yarnall, Y. Y., and Hudson, R. L. (2022). Direct measurements of infrared intensities of HCN and H_2O + HCN ices for laboratory and observational astrochemistry. *Monthly Notices of the Royal Astronomical Society*, 509(3):3515–3522. 5.4.3

Geweke, J. F. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff Report 148, Federal Reserve Bank of Minneapolis. A.1.1

Gibb, E. L., Whittet, D. C. B., Boogert, A. C. A., and Tielens, A. G. G. M. (2004). Interstellar Ice: The Infrared Space Observatory Legacy. *The Astrophysical Journals*, 151(1):35–73. 2.3.4

Goldsmith, P. F. (2001). Molecular Depletion and Thermal Balance in Dark Cloud Cores. *ApJ*, 557(2):736–746. 6.4.1, 6.4.1

Goldsmith, P. F. and Li, D. (2005). HiNarrow self-absorption in dark clouds: Correlations with molecular gas and implications for cloud evolution and star formation. *The Astrophysical Journal*, 622(2):938–958. 2.1

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org. 1.6.3, 1.6.3, 1.6.3

Goumans, T. P. M., Uppal, M. A., and Brown, W. A. (2008). Formation of CO_2 on a carbonaceous surface: a quantum chemical study. *Monthly Notices of the Royal Astronomical Society*, 384(3):1158–1164. ??, ??

Graff, P., Feroz, F., Hobson, M. P., and Lasenby, A. (2012). BAMBI: blind accelerated multimodal Bayesian inference. *MNRAS*, 421(1):169–180. 8

Gramacy, R. B. and Lee, H. K. H. (2008). Adaptive design and analysis of supercomputer experiments. *arXiv e-prints*, page arXiv:0805.4359. 4.8

Graninger, D. M., Herbst, E., Öberg, K. I., and Vasyunin, A. I. (2014). The HNC/HCN Ratio in Star-forming Regions. *ApJ*, 787(1):74. 6.4.2, 6.4.2

Grassi, T., Merlin, E., Piovan, L., Buonomo, U., and Chiosi, C. (2011). MaNN: Multiple Artificial Neural Networks for modelling the Interstellar Medium. *arXiv e-prints*, page arXiv:1103.0509. 6.1

Grassi, T., Bovino, S., Gianturco, F. A., Baiocchi, P., and Merlin, E. (2012). Complexity reduction of astrochemical networks. *Monthly Notices of the Royal Astronomical Society*, 425(2):1332–1340. 3.1, 7.1

Grassi, T., Bovino, S., Schleicher, D., and Gianturco, F. A. (2013). Chemical complexity in astrophysical simulations: optimization and reduction techniques. *Monthly Notices of the Royal Astronomical Society*, 431(2):1659–1668. 7.1

Grassi, T., Bovino, S., Caselli, P., Bovolenta, G., Vogt-Geisse, S., and Ercolano, B. (2020). A novel framework for studying the impact of binding energy distributions on the chemistry of dust grains. *Astronomy & Astrophysics*, 643:A155. 4.5.3, 7.3.3

Grassi, T., Nauman, F., Ramsey, J. P., Bovino, S., Picogna, G., and Ercolano, B. (2022). Reducing the complexity of chemical networks via interpretable autoencoders. *Astronomy & Astrophysics*, 668:A139. 6.1, 7.1

Gray, W. K., Navaratnam, A. V., Day, J., Heyl, J., Hardy, F., Wheeler, A., Eve-Jones, S., and Briggs, T. W. R. (2023). Role of hospital strain in determining outcomes for people hospitalised with covid-19 in england. *Emergency Medicine Journal*. (document)

Grow, A. and Hilton, J. (2018). *Statistical Emulation*, pages 1–8. American Cancer Society. 4.3, 6.1

Hacar, A., Bosman, A. D., and van Dishoeck, E. F. (2020). HCN-to-HNC intensity ratio: a new chemical thermometer for the molecular ISM. *A&A*, 635:A4. (document), 6.4.2, 6.4.2, 6.12, 6.5

Hardy, F., Heyl, J., Tucker, K., Hopper, A., Marchã, M. J., Briggs, T. W. R., Yates, J., Day, J., Wheeler, A., Eve-Jones, S., and Gray, W. K. (2022). Data consistency in the english hospital episodes statistics database. *BMJ Health & Care Informatics*, 29(1). (document)

Hardy, F., Heyl, J., Tucker, K., Hopper, A., Marchã, M. J., Navaratnam, A. V., Briggs, T. W. R., Yates, J., Day, J., Wheeler, A., Eve-Jones, S., and Gray, W. K. (2023). Estimating nosocomial infection and its outcomes in hospital patients in england with a diagnosis of covid-19 using machine learning. *International Journal of Data Science and Analytics*. (document)

Hartquist, T. W., Doyle, H. T., and Dalgarno, A. (1978). The intercloud cosmic ray ionization rate. *A&A*, 68:65–67. 2.2.2

Hasegawa, T. I., Herbst, E., and Leung, C. M. (1992). Models of Gas-Grain Chemistry in Dense Interstellar Clouds with Complex Organic Molecules. *Astrophysical Journal Supplement Series*, 82:167. 1.3.1, 1.4.1, 2.1, 2.3.2, 5.1

He, J., Acharyya, K., and Vidali, G. (2016). Binding Energy of Molecules on Water Ice: Laboratory Measurements and Modeling. *The Astrophysical Journal*, 825(2):89. 4.1, 5.1, 7.2.1

Heavens, A. F., Jimenez, R., and Lahav, O. (2000). Massive lossless data compression and multiple parameter estimation from galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, 317(4):965–972. 5.1, 5.3.3, 5.3.3, 5.3.3, 5.3.3, 7.2.2

Heavens, A. F., Sellentin, E., de Mijolla, D., and Vianello, A. (2017). Massive data compression for parameter-dependent covariance matrices. *Monthly Notices of the Royal Astronomical Society*, 472(4):4244–4250. 5.1, 5.3.3, 5.3.3, 7.2.2

Heavens, A. F. and Sellentin, E. (2018). Objective Bayesian analysis of neutrino masses and hierarchy. *Journal of Cosmology and Astroparticle Physics*, 2018(4):047. 8

Heavens, A. F., Sellentin, E., and Jaffe, A. H. (2020). Extreme data compression while searching for new physics. *Monthly Notices of the Royal Astronomical Society*, 498(3):3440–3451. 5.1, 5.3.3, 7.2.2

Herbst, E., Terzieva, R., and Talbi, D. (2000). Calculations on the rates, mechanisms, and interstellar importance of the reactions between C and NH_2 and between N and CH_2. *MNRAS*, 311(4):869–876. 6.4.2

Herbst, E. and van Dishoeck, E. F. (2009). Complex Organic Interstellar Molecules. *Annual Review of Astronomy and Astrophysics*, 47(1):427–480. 1.1, 1.3.1, 2.1

Heyl, J., Viti, S., Holdship, J., and Feeney, S. M. (2020). Exploiting Network Topology for Accelerated Bayesian Inference of Grain Surface Reaction Networks. *The Astrophysical Journal*, 904(2):197. 3

Heyl, J., Hardy, F., Tucker, K., Hopper, A., Marchã, M. J. M., Navaratnam, A. V., Briggs, T. W. R., Yates, J., Day, J., Wheeler, A., Eve-Jones, S., and Gray, W. K. (2022). Frailty, comorbidity, and associations with in-hospital mortality in older covid-19 patients: Exploratory study of administrative data. *Interact J Med Res*, 11(2):e41520. (document)

Heyl, J., Sellentin, E., Holdship, J., and Viti, S. (2022a). Identifying the most constraining ice observations to infer molecular binding energies. *Monthly Notices of the Royal Astronomical Society*, 517(1):38–46. 5

Heyl, J., Holdship, J., and Viti, S. (2022b). Using Statistical Emulation and Knowledge of Grain-surface Diffusion for Bayesian Inference of Reaction Rate Parameters: An Application to a Glycine Network. *The Astrophysical Journal*, 931(1):26. 4

Heyl, J., Hardy, F., Tucker, K., Hopper, A., Marchã, M. J., Liew, A., Reep, J., Harwood, K.-A., Roberts, L., Yates, J., Day, J., Wheeler, A., Eve-Jones, S., Briggs, T. W., and Gray, W. K. (2023). Data quality and autism: Issues and potential impacts. *International Journal of Medical Informatics*, 170:104938. (document)

Heyl, J., Lamberts, T., Viti, S., and Holdship, J. (2023a). Investigating the impact of reactions of C and CH with molecular hydrogen on a glycine gas-grain network. *Monthly Notices of the Royal Astronomical Society*, 520(1):503–512. 2

Heyl, J., Viti, S., and Vermariën, G. (2023b). A statistical and machine learning approach to the study of astrochemistry. *Faraday Discussions*, pages 1–17. 7

Heyl, J., Butterworth, J., and Viti, S. (2023c). Understanding Molecular Abundances in Star-Forming Regions Using Interpretable Machine Learning. *MNRAS*. 6

Hocuk, S., Cazaux, S., and Spaans, M. (2014). The impact of freeze-out on collapsing molecular clouds. *Monthly Notices of the Royal Astronomical Society*, 438(1):L56–L60. 3.2

Hoffmann, M., Fröhner, C., and Noé, F. (2019). Reactive SINDy: Discovering governing reactions from concentration data. *The Journal of Chemical Physics*, 150(2). 025101. 6.1

Hogg, D. W. and Foreman-Mackey, D. (2018). Data analysis recipes: Using markov chain monte carlo. *The Astrophysical Journal Supplement Series*, 236(1):11. A.1.2

Holdship, J., Viti, S., Jiménez-Serra, I., Makrymallis, A., and Priestley, F. (2017). UCLCHEM: A Gas-grain Chemical Code for Clouds, Cores, and C-Shocks. *The Astronomical Journal*, 154(1):38. 1.4.1, 2.2.1, 3.2, 4.2.1, 5.2.1, 6.1, 6.2, 7.2.1

Holdship, J., Jeffrey, N., Makrymallis, A., Viti, S., and Yates, J. (2018). Bayesian Inference of the Rates of Surface Reactions in Icy Mantles. *The Astrophysical Journal*, 866(2):116. (document), 3.1, 3.1, 3.2, 4.1, 4.2.1, 4.4.2, 4.4.2, 4.4.3, 4.4.6, 4.7, 5.1, 5.2.2, 7.1, 7.3.1, 8

Holdship, J., Viti, S., Haworth, T. J., and Ilee, J. D. (2021). Chemulator: Fast, accurate thermochemistry for dynamical models through emulation. *Astronomy & Astrophysics*, 653:A76. 4.3, 6.1, 7.1

Holdship, J. and Viti, S. (2022). History-independent tracers. Forgetful molecular probes of the physical conditions of the dense interstellar medium. *A&A*, 658:A103. 6.1

Huang, K. Y., Viti, S., Holdship, J., García-Burillo, S., Kohno, K., Taniguchi, A., Martn, S., Aladro, R., Fuente, A., and Sánchez-García, M. (2022). The chemical footprint of AGN feedback in the outflowing circumnuclear disk of NGC 1068. *Astronomy & Astrophysics*, 666:A102. 7.1

Huijser, D., Goodman, J., and Brewer, B. J. (2015). Properties of the Affine Invariant Ensemble Sampler in high dimensions. *arXiv e-prints*, page arXiv:1509.02230. 3.3.2

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126. 4.5.1

Imanishi, M., Nakanishi, K., and Izumi, T. (2019). ALMA Spatially Resolved Dense Molecular Gas Survey of Nearby Ultraluminous Infrared Galaxies. *ApJSS*, 241(2):19. 6.4.2

Indriolo, N., Geballe, T. R., Oka, T., and McCall, B. J. (2007). H3+ in diffuse interstellar clouds: a tracer for the cosmic-ray ionization rate. *The Astrophysical Journal*, 671(2):1736–1747. 2.2.2

Indriolo, N. and McCall, B. J. (2012). Investigating the cosmic-ray ionization rate in the galactic diffuse interstellar medium through observations of h_3^+. *The Astrophysical Journal*, 745(1):91. 2.2.2

Indriolo, N. and McCall, B. J. (2013). Cosmic-ray astrochemistry. *Chemical Society Reviews*, 42:7763–7773. 2.2.2

Ioppolo, S., van Boheemen, Y., Cuppen, H. M., van Dishoeck, E. F., and Linnartz, H. (2011a). Surface formation of CO_2 ice at low temperatures. *Monthly Notices of the Royal Astronomical Society*, 413(3):2281–2287. 3.2, 5.4.3

Ioppolo, S., Cuppen, H. M., van Dishoeck, E. F., and Linnartz, H. (2011b). Surface formation of HCOOH at low temperature. *Monthly Notices of the Royal Astronomical Society*, 410(2):1089–1095. ??

Ioppolo, S., Fedoseev, G., Chuang, K. J., Cuppen, H. M., Clements, A. R., Jin, M., Garrod, R. T., Qasim, D., Kofman, V., van Dishoeck, E. F., and Linnartz, H. (2020). A non-energetic mechanism for glycine formation in the interstellar medium. *Nature Astronomy*. (document), 1.3.1, 2.1, 2.1, ??, ??, ??, ??, ??, ??, ??, ??, ??, 2.2.1, 2.3, 2.4, 4.2.2, 4.1, 4.4.5, 8, A.2

Izumi, T., Kohno, K., Martín, S., Espada, D., Harada, N., Matsushita, S., Hsieh, P.-Y., Turner, J. L., Meier, D. S., Schinnerer, E., Imanishi, M., Tamura, Y., Curran, M. T., Doi, A., Fathi, K., Krips, M., Lundgren, A. A., Nakai, N., Nakajima, T., Regan, M. W., Sheth, K., Takano, S., Taniguchi, A., Terashima, Y., Tosaki, T., and Wiklind, T. (2013). Submillimeter ALMA Observations of the Dense Gas in the Low-Luminosity

Type-1 Active Nucleus of NGC 1097. *Publications of the Astronomical Society of Japan*, 65:100. 6.4.2

Izumi, T., Kohno, K., Aalto, S., Espada, D., Fathi, K., Harada, N., Hatsukade, B., Hsieh, P.-Y., Imanishi, M., Krips, M., Martín, S., Matsushita, S., Meier, D. S., Nakai, N., Nakanishi, K., Schinnerer, E., Sheth, K., Terashima, Y., and Turner, J. L. (2016). Submillimeter-HCN Diagram for Energy Diagnostics in the Centers of Galaxies. *ApJ*, 818(1):42. 6.4.2

James, T. A., Viti, S., Holdship, J., and Jiménez-Serra, I. (2020). Tracing shock type with chemical diagnostics. An application to L1157. *A&A*, 634:A17. 6.1

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London Series A*, 186(1007):453–461. A.2

Jiménez-Serra, I., testi, L., Caselli, P., and Viti, S. (2014). Detectability of Glycine in Solar-type System Precursors. *The Astrophysical Journall*, 787(2):L33. 2.3.4

Jiménez-Serra, I., Vasyunin, A. I., Caselli, P., Marcelino, N., Billot, N., Viti, S., Testi, L., Vastel, C., Lefloch, B., and Bachiller, R. (2016). The Spatial Distribution of Complex Organic Molecules in the L1544 Pre-stellar Core. *The Astrophysical Journall*, 830(1):L6. 2.2.3

Jiménez-Serra, I., Vasyunin, A. I., Spezzano, S., Caselli, P., Cosentino, G., and Viti, S. (2021). The Complex Organic Molecular Content in the L1498 Starless Core. *The Astrophysical Journal*, 917(1):44. 2.2.3

Jiménez-Escobar, A., Muñoz Caro, G. M., and Chen, Y.-J. (2014). Sulphur depletion in dense clouds and circumstellar regions. Organic products made from UV photoprocessing of realistic ice analogs containing H2S. *Monthly Notices of the Royal Astronomical Society*, 443(1):343–354. 3.6

Jin, M. and Garrod, R. T. (2020). Formation of Complex Organic Molecules in Cold Interstellar Environments through Nondiffusive Grain-surface and Ice-mantle Chemistry. *The Astrophysical Journals*, 249(2):26. 1.3.1, 2.2.1, 2.3.4, 2.4, 4.2.3, 4.6.1, 4.8

Kaiser, R. I. and Osamura, Y. (2005a). Infrared spectroscopic studies of hydrogenated silicon clusters. Guiding the search for Si_2H_x species in the Circumstellar Envelope of IRC+10216. *Astronomy & Astrophysics*, 432(2):559–566. 5.4.3

Kaiser, R. I. and Osamura, Y. (2005b). Laboratory Studies on the Infrared Absorptions of Hydrogenated Carbon-Silicon Clusters: Directing the Identification of Organometallic $SiCH_x$ Species toward IRC +10216. *The Astrophysical Journal*, 630(2):1217–1223. 5.4.3

Kazandjian, M. V., Pelupessy, I., Meijerink, R., Israel, F. P., Coppola, C. M., Rosenberg, M. J. F., and Spaans, M. (2016). Constraining cloud parameters using high density gas tracers in galaxies. *Astronomy & Astrophysics*, 595:A124. 7.1

Keane, J. V., Tielens, A. G. G. M., Boogert, A. C. A., Schutte, W. A., and Whittet, D. C. B. (2001). Ice absorption features in the 5-8 μm region toward embedded protostars. *Astronomy & Astrophysics*, 376:254–270. 5.4.3

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464. 4.3

Krasnokutski, S. A., Kuhn, M., Renzler, M., Jäger, C., Henning, T., and Scheier, P. (2016). Ultra-low-temperature Reactions of Carbon Atoms with Hydrogen Molecules. *The Astrophysical Journall*, 818(2):L31. 2.1

Laas, J. C. and Caselli, P. (2019). Modeling sulfur depletion in interstellar clouds. *Astronomy & Astrophysics*, 624:A108. 3.2, 3.6, 7.2.1

Lamberts, T., Cuppen, H. M., Ioppolo, S., and Linnartz, H. (2013). Water formation at low temperatures by surface o2 hydrogenation iii: Monte carlo simulation. *Phys. Chem. Chem. Phys.*, 15:8287–8302. ??, ??, ??, ??, ??, ??

Lamberts, T., Fedoseev, G., Kästner, J., Ioppolo, S., and Linnartz, H. (2017). Importance of tunneling in H-abstraction reactions by OH radicals. The case of CH_4 + OH studied through isotope-substituted analogs. *A&A*, 599:A132. ??

Lamberts, T. and Kästner, J. (2017). Influence of surface and bulk water ice on the reactivity of a water-forming reaction. *The Astrophysical Journal*, 846(1):43. ??

Lamberts, T., Fedoseev, G., van Hemert, M. C., Qasim, D., Chuang, K.-J., Santos, J. C., and Linnartz, H. (2022). Methane Formation in Cold Regions from Carbon Atoms and Molecular Hydrogen. *The Astrophysical Journal*, 928(1):48. 2.1, 2.1, ??, ??, 2.2.2

Lee, C.-W., Kim, J.-K., Moon, E.-S., Minh, Y. C., and Kang, H. (2009). FORMATION OF GLYCINE ON ULTRAVIOLET-IRRADIATED INTERSTELLAR ICE-ANALOG

FILMS AND IMPLICATIONS FOR INTERSTELLAR AMINO ACIDS. *The Astrophysical Journal*, 697(1):428–435. 2.1

Lefèvre, C., Pagani, L., Juvela, M., Paladini, R., Lallement, R., Marshall, D. J., Andersen, M., Bacmann, A., McGehee, P. M., Montier, L., Noriega-Crespo, A., Pelkonen, V. M., Ristorcelli, I., and Steinacker, J. (2014). Dust properties inside molecular clouds from coreshine modeling and observations. *Astronomy & Astrophysics*, 572:A20. 7.1

Linnartz, H., Ioppolo, S., and Fedoseev, G. (2015). Atom addition reactions in interstellar ice analogues. *International Reviews in Physical Chemistry*, 34(2):205–237. (document), 1.3.1, 3.7, 4.2.2, 4.1

Lundberg, S. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv e-prints*, page arXiv:1705.07874. 6.1, 6.3.1, 6.3.1, 7.2.2

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv e-prints*, page arXiv:1802.03888. 6.3.1, 6.3.1, 7.2.2

Machado Poletti Valle, L. F., Avestruz, C., Barnes, D. J., Farahi, A., Lau, E. T., and Nagai, D. (2021). SHAPing the gas: understanding gas shapes in dark matter haloes with interpretable machine learning. *Monthly Notices of the Royal Astronomical Society*, 507(1):1468–1484. 6.1

Makrymallis, A. and Viti, S. (2014). Understanding the Formation and Evolution of Interstellar Ices: A Bayesian Approach. *ApJ*, 794(1):45. 3.1, 4.1, 5.2.2, 7.1

Maté, B., Tanarro, I., Escribano, R., Moreno, M. A., and Herrero, V. J. (2015). Stability of Extraterrestrial Glycine under Energetic Particle Radiation Estimated from 2 keV Electron Bombardment Experiments. *The Astrophysical Journal*, 806(2):151. 2.1

McClure, M., Bailey, J., Beck, T., Boogert, A. C., Brown, W., Caselli, P., Chiar, J., Egami, E., Fraser, H. J., Garrod, R., Gordon, K. D., Ioppolo, S., Jimenez-Serra, I., Jorgensen, J., Kristensen, L. E., Linnartz, H., McCoustra, M., Murillo, N., Noble, J., Oberg, K., Palumbo, M. E., Pendleton, Y. J., Pontoppidan, K. M., Van Dishoeck, E. F., and Viti, S. (2017). IceAge: Chemical Evolution of Ices during Star Formation. JWST Proposal ID 1309. Cycle 0 Early Release Science. 1.3.1, 2.3.3, 4.8

McClure, M. K., Rocha, W. R. M., Pontoppidan, K. M., Crouzet, N., Chu, L. E. U., Dartois, E., Lamberts, T., Noble, J. A., Pendleton, Y. J., Perotti, G., Qasim, D., Rachid, M. G., Smith, Z. L., Sun, F., Beck, T. L., Boogert, A. C. A., Brown, W. A., Caselli, P., Charnley, S. B., Cuppen, H. M., Dickinson, H., Drozdovskaya, M. N., Egami, E., Erkal, J., Fraser, H., Garrod, R. T., Harsono, D., Ioppolo, S., Jiménez-Serra, I., Jin, M., Jørgensen, J. K., Kristensen, L. E., Lis, D. C., McCoustra, M. R. S., McGuire, B. A., Melnick, G. J., Ã-berg, K. I., Palumbo, M. E., Shimonishi, T., Sturm, J. A., van Dishoeck, E. F., and Linnartz, H. (2023). An Ice Age JWST inventory of dense molecular cloud ices. *Nature Astronomy*, 7:431–443. (document), 1.3.1, 7.1, 7.2.1, 7.1

McElroy, D., Walsh, C., Markwick, A. J., Cordiner, M. A., Smith, K., and Millar, T. J. (2013). The UMIST database for astrochemistry 2012. *Astronomy & Astrophysics*, 550:A36. (document), 1.3.2, 2.2.1, 4.1, 4.4, 4.5.3, 5.1, 5.2.2, 5.3.1, 6.2, 7.2.1

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245. 1.6.2, 4.3.1, 6.3.2, 7.2.2

McKee, C. F. (1989). Photoionization-regulated Star Formation and the Structure of Molecular Clouds. *ApJ*, 345:782. 1.2.2

Meijerink, R., Spaans, M., Loenen, A. F., and van der Werf, P. P. (2011). Star formation in extreme environments: the effects of cosmic rays and mechanical heating. *A&A*, 525:A119. 6.4.2

Meisner, J., Lamberts, T., and Kästner, J. (2017). Atom Tunneling in the Water Formation Reaction H2 + OH → H2O + H on an Ice Surface. *ACS Earth and Space Chemistry*, 1(7):399–410. 2.1, **??**, 2.2.1

Minissale, M., Congiu, E., and Dulieu, F. (2016a). Direct measurement of desorption and diffusion energies of O and N atoms physisorbed on amorphous surfaces. *Astronomy & Astrophysics*, 585:A146. 4.2.3

Minissale, M., Dulieu, F., Cazaux, S., and Hocuk, S. (2016b). Dust as interstellar catalyst. I. Quantifying the chemical desorption process. *Astronomy & Astrophysics*, 585:A24. 4.6.1, 5.2.2

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition. 1.6.1, 1.6.4, 1.6.5, 6.3.1, 6.3.1, 7.2.2, 7.3.3

Molpeceres, G. and Kästner, J. (2020). Adsorption of h2 on amorphous solid water studied with molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, 22:7552–7563. 2.1

Nogueira, F. (2014). Bayesian Optimization: Open source constrained global optimization tool for Python. 6.3.2

Norton, J. D. (2008). Ignorance and indifference. *Philosophy of Science*, 75(1):45–68. 1

Oba, Y., Chigai, T., Osamura, Y., Watanabe, N., and Kouchi, A. (2014). Hydrogen isotopic substitution of solid methylamine through atomic surface reactions at low temperatures: A potential contribution to the d/h ratio of methylamine in molecular clouds. *Meteoritics & Planetary Science*, 49(1):117–132. **??**

O'Donnell, E. J. and Watson, W. D. (1974). Upper limits to the flux of cosmic rays and X-rays in interstellar clouds. *The Astrophysical Journal*, 191:89–92. 2.2.2

O'Donoghue, R., Viti, S., Padovani, M., and James, T. (2022). The Effects of Cosmic Rays on the Chemistry of Dense Cores. *The Astrophysical Journal*, 934(1):63. 2.2.2

Ohishi, M., Suzuki, T., Hirota, T., Saito, M., and Kaifu, N. (2019). Detection of a new methylamine (CH_3NH_2) source: Candidate for future glycine surveys. *Publications of Astronomical Society of Japan*, 71(4):86. **??**

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. 4.3.2

Pellejero-Ibañez, M., Angulo, R. E., Aricó, G., Zennaro, M., Contreras, S., and Stücker, J. (2020). Cosmological parameter estimation via iterative emulation of likelihoods. *Monthly Notices of the Royal Astronomical Society*, 499(4):5257–5268. 4.3

Penteado, E. M., Walsh, C., and Cuppen, H. M. (2017). Sensitivity Analysis of Grain Surface Chemistry to Binding Energies of Ice Species. *The Astrophysical Journal*, 844(1):71. (document), 4.1, 4.2.3, 4.4.4, 4.4.5, 4.4, 4.5.3, 4.6.2, 5.1, 5.3.1

Pernet, A., Pilmé, J., Pauzat, F., Ellinger, Y., Sirotti, F., Silly, M., Parent, P., and Laffon, C. (2013). Possible survival of simple amino acids to X-ray irradiation in ice: the case of glycine. *A&A*, 552:A100. 2.1

Peterson, P. (2009). F2py: a tool for connecting fortran and python programs. *International Journal of Computational Science and Engineering*, 4(4):296–305. 3.3.2

Pety, J., Guzmán, V. V., Orkisz, J. H., Liszt, H. S., Gerin, M., Bron, E., Bardeau, S., Goicoechea, J. R., Gratier, P., Le Petit, F., Levrier, F., Öberg, K. I., Roueff, E., and Sievers, A. (2017). The anatomy of the Orion B giant molecular cloud: A local template for studies of nearby galaxies. *A&A*, 599:A98. 6.4.2

Plaia, A., Buscemi, S., Fürnkranz, J., and Mencía, E. L. (2022). Comparing boosting and bagging for decision trees of rankings. *Journal of Classification*, 39(1):78–99. 1.6.4

Puzzarini, C. (2022). Gas-phase chemistry in the interstellar medium: The role of laboratory astrochemistry. *Frontiers in Astronomy and Space Sciences*, 8. 1.3.2

Qasim, D., Chuang, K. J., Fedoseev, G., Ioppolo, S., Boogert, A. C. A., and Linnartz, H. (2018). Formation of interstellar methanol ice prior to the heavy CO freeze-out stage. *A&A*, 612:A83. **??**

Qasim, D., Fedoseev, G., Chuang, K. J., He, J., Ioppolo, S., van Dishoeck, E. F., and Linnartz, H. (2020). An experimental study of the surface formation of methane in interstellar molecular clouds. *Nature Astronomy*, 4:781–785. 5.4.3

Quan, D., Herbst, E., Osamura, Y., and Roueff, E. (2010). Gas-grain Modeling of Isocyanic Acid (HNCO), Cyanic Acid (HOCN), Fulminic Acid (HCNO), and Isofulminic Acid (HONC) in Assorted Interstellar Environments. *The Astrophysical Journal*, 725(2):2101–2109. 5.2.2

Quénard, D., Jiménez-Serra, I., Viti, S., Holdship, J., and Coutens, A. (2018). Chemical modelling of complex organic molecules with peptide-like bonds in star-forming regions. *Monthly Notices of the Royal Astronomical Society*, 474(2):2796–2812. 1.4.1, 2.2.1, 5.1, 5.2.1, 5.2.2, 6.2, 7.2.1

Ramesh, J. P. and Yuan-Pern, L. (2022). A chemical link between methylamine and methylene imine and implications for interstellar glycine formation. *Communications chemistry*, 5(1):1–7. 2.3.4

Roberts, J. F., Rawlings, J. M. C., Viti, S., and Williams, D. A. (2007). Desorption from interstellar ices. *MNRAS*, 382(2):733–742. 1.4.1, 6.2

Rogers, K. K., Peiris, H. V., Pontzen, A., Bird, S., Verde, L., and Font-Ribera, A. (2019). Bayesian emulator optimisation for cosmology: application to the lyman-alpha forest. *Journal of Cosmology and Astroparticle Physics*, 2019(02):031–031. 4.3

Roy, V. (2020). Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 7(1):387–412. A.1.1, A.1.2

Ruaud, M., Loison, J. C., Hickson, K. M., Gratier, P., Hersant, F., and Wakelam, V. (2015). Modelling complex organic molecules in dense regions: Eley-Rideal and complex induced reaction. *Monthly Notices of the Royal Astronomical Society*, 447(4):4004–4017. 1.3.1, 4.6.1

Ruaud, M., Wakelam, V., and Hersant, F. (2016). Gas and grain chemical composition in cold cores as predicted by the Nautilus three-phase model. *MNRAS*, 459(4):3756–3767. 6.1

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. 1.6.5

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, page 2:e55. 3.3.2

Sato, A., Kitazawa, Y., Ochi, T., Shoji, M., Komatsu, Y., Kayanuma, M., Aikawa, Y., Umemura, M., and Shigeta, Y. (2018). First-principles study of the formation of glycine-producing radicals from common interstellar species. *Molecular Astrophysics*, 10:11–19. 2.1

Schmit, C. J. and Pritchard, J. R. (2017). Emulation of reionization simulations for Bayesian inference of astrophysics parameters using neural networks. *Monthly Notices of the Royal Astronomical Society*, 475(1):1213–1223. 4.3

Shapley, L. S. (2016). *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton. 6.3.1, 7.2.2

Shi, Y., Wang, J., Zhang, Z.-Y., Gao, Y., Armus, L., Helou, G., Gu, Q., and Stierwalt, S. (2015). The Weak Carbon Monoxide Emission in an Extremely Metal-poor Galaxy, Sextans A. *ApJL*, 804(1):L11. 6.4.1, 6.4.1

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90. 6.3.2

Simončič, M., Semenov, D., Krasnokutski, S., Henning, T., and Jäger, C. (2020). Sensitivity of gas-grain chemical models to surface reaction barriers. Effect from a key carbon-insertion reaction, C + H_2 → CH_2. *A&A*, 637:A72. 2.1, 2.1, **??**, 2.2.2

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833 – 859. 4.4.2

Smith, R. G. (1991). A search for solid H_2S in dense clouds. *Monthly Notices of the Royal Astronomical Society*, 249:172. 5.4.3

Song, L. and Kästner, J. (2016). Formation of the prebiotic molecule NH2CHO on astronomical amorphous solid water surfaces: accurate tunneling rate calculations. *Physical Chemistry Chemical Physics (Incorporating Faraday Transactions)*, 18(42):29278–29285. 5.2.2

Tang, K. S. and Turk, M. (2022). Reduced Order Model for Chemical Kinetics: A case study with Primordial Chemical Network. *arXiv e-prints*, page arXiv:2207.07159. 7.1

Taquet, V., Ceccarelli, C., and Kahane, C. (2012). Multilayer modeling of porous grain surface chemistry. I. The GRAINOBLE model. *A&A*, 538:A42. 6.1

Tegmark, M., Taylor, A. N., and Heavens, A. F. (1997). Karhunen-Loève Eigenvalue Problems in Cosmology: How Should We Tackle Large Data Sets? *The Astrophysical Journal*, 480(1):22–35. 5.3.3

Terwisscha van Scheltinga, J., Marcandalli, G., McClure, M. K., Hogerheijde, M. R., and Linnartz, H. (2021). Infrared spectra of complex organic molecules in astronomically relevant ice matrices. III. Methyl formate and its tentative solid-state detection. *Astronomy & Astrophysics*, 651:A95. 5.4.3

Tielens, A. G. G. M. (2013). The molecular universe. *Reviews of Modern Physics*, 85(3):1021–1081. (document), 1.1

Tomassini, L., Reichert, P., Knutti, R., Stocker, T. F., and Borsuk, M. E. (2007). Robust bayesian uncertainty analysis of climate system properties using markov chain monte carlo methods. *Journal of climate*, 20(7):1239–1254. A.2

Tunnard, R. and Greve, T. R. (2016). How Dense is Your Gas? On the Recoverability of LVG Model Parameters. *ApJ*, 819(2):161. 6.1

van Dishoeck, E. F. and Black, J. H. (1988). The Photodissociation and Chemistry of Interstellar CO. *The Astrophysical Journal*, 334:771. 2.1

van Dishoeck, E. F. (2014). Astrochemistry of dust, ice and gas: introduction and overview. *Faraday Discuss.*, 168:9–47. 1.3.1

van Dishoeck, E. F., Kristensen, L. E., Mottram, J. C., Benz, A. O., Bergin, E. A., Caselli, P., Herpin, F., Hogerheijde, M. R., Johnstone, D., Liseau, R., Nisini, B., Tafalla, M., van der Tak, F. F. S., Wyrowski, F., Baudry, A., Benedettini, M., Bjerkeli, P., Blake, G. A., Braine, J., Bruderer, S., Cabrit, S., Cernicharo, J., Choi, Y., Coutens, A., de Graauw, T., Dominik, C., Fedele, D., Fich, M., Fuente, A., Furuya, K., Goicoechea, J. R., Harsono, D., Helmich, F. P., Herczeg, G. J., Jacq, T., Karska, A., Kaufman, M., Keto, E., Lamberts, T., Larsson, B., Leurini, S., Lis, D. C., Melnick, G., Neufeld, D., Pagani, L., Persson, M., Shipman, R., Taquet, V., van Kempen, T. A., Walsh, C., Wampfler, S. F., Yıldız, U., and WISH Team (2021). Water in star-forming regions: physics and chemistry from clouds to disks as probed by Herschel spectroscopy. *A&A*, 648:A24. 6.4.1

Vats, D. and Knudson, C. (2018). Revisiting the gelman-rubin diagnostic. A.1.2

Vidal, T. H. G., Loison, J.-C., Jaziri, A. Y., Ruaud, M., Gratier, P., and Wakelam, V. (2017). On the reservoir of sulphur in dark clouds: chemistry and elemental abundance reconciled. *Monthly Notices of the Royal Astronomical Society*, 469(1):435–447. 5.4.3

Vidal, T. H. G. and Wakelam, V. (2018). A new look at sulphur chemistry in hot cores and corinos. *Monthly Notices of the Royal Astronomical Society*, 474(4):5575–5587. 3.6

Villadsen, T., Ligterink, N. F. W., and Andersen, M. (2022). Predicting binding energies of astrochemically relevant molecules via machine learning. *arXiv e-prints*, page arXiv:2207.03906. 5.1, 7.2.1

Viti, S., Collings, M. P., Dever, J. W., McCoustra, M. R. S., and Williams, D. A. (2004). Evaporation of ices near massive stars: models based on laboratory temperature programmed desorption data. *Monthly Notices of the Royal Astronomical Society*, 354(4):1141–1145. 1.2.2, 1.4.1, 2.2.2, 6.2

Viti, S., García-Burillo, S., Fuente, A., Hunt, L. K., Usero, A., Henkel, C., Eckart, A., Martin, S., Spaans, M., Muller, S., Combes, F., Krips, M., Schinnerer, E., Casasola, V., Costagliola, F., Marquez, I., Planesas, P., van der Werf, P. P., Aalto, S., Baker, A. J., Boone, F., and Tacconi, L. J. (2014). Molecular line emission in NGC 1068 imaged with ALMA. II. The chemistry of the dense molecular gas. *Astronomy & Astrophysics*, 570:A28. 7.1

Viti, S. (2017). Molecular transitions as probes of the physical conditions of extragalactic environments. *A&A*, 607:A118. 6.1, 6.4.2, 6.4.2

Wakelam, V., Cuppen, H. M., and Herbst, E. (2013). *Astrochemistry: Synthesis and Modelling*, pages 115–143. Springer Berlin Heidelberg, Berlin, Heidelberg. 1.4

Wakelam, V., Loison, J.-C., Herbst, E., Pavone, B., Bergeat, A., Béroff, K., Chabot, M., Faure, A., Galli, D., Geppert, W. D., Gerlich, D., Gratier, P., Harada, N., Hickson, K. M., Honvault, P., Klippenstein, S. J., Picard, S. D. L., Nyman, G., Ruaud, M., Schlemmer, S., Sims, I. R., Talbi, D., Tennyson, J., and Wester, R. (2015). The 2014 kida network for interstellar chemistry. *The Astrophysical Journal Supplement Series*, 217(2):20. 1.3.2

Wakelam, V., Loison, J. C., Mereau, R., and Ruaud, M. (2017). Binding energies: New values and impact on the efficiency of chemical desorption. *Molecular Astrophysics*, 6:22–35. (document), 2.1, 2.2.1, 4.1, 4.4.4, 4.4, 4.5.3, 5.1, 5.3.1, 6.2, 7.2.1

Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2):125 – 148. 1

Wang, G.-J., Li, S.-Y., and Xia, J.-Q. (2020). ECoPANN: A framework for estimating cosmological parameters using artificial neural networks. *The Astrophysical Journal Supplement Series*, 249(2):25. 4.3

Whittet, D. C. B., Cook, A. M., Herbst, E., Chiar, J. E., and Shenoy, S. S. (2011).

Observational Constraints on Methanol Production in Interstellar and Preplanetary Ices. *The Astrophysical Journal*, 742(1):28. 4.4.2

Williams, D. A. and Viti, S. (2013). *Observational Molecular Astronomy: Exploring the Universe Using Molecular Line Emissions*. Cambridge Observing Handbooks for Research Astronomers. Cambridge University Press. (document), 1.1

Woods, P. M., Occhiogrosso, A., Viti, S., Kaňuchová, Z., Palumbo, M. E., and Price, S. D. (2015). A new study of an old sink of sulphur in hot molecular cores: the sulphur residue. *Monthly Notices of the Royal Astronomical Society*, 450(2):1256–1267. 3.2, 3.6, 5.4.3

Woon, D. E. (2002). Pathways to Glycine and Other Amino Acids in Ultraviolet-irradiated Astrophysical Ices Determined via Quantum Chemical Modeling. *The Astrophysical Journall*, 571(2):L177–L180. 2.1, **??**

Xu, R., Bai, X.-N., Öberg, K., and Zhang, H. (2019). Chemical network reduction in protoplanetary disks. *The Astrophysical Journal*, 872(1):107. 3.1, 7.1

Yang, Y.-L., Green, J. D., Pontoppidan, K. M., Bergner, J. B., Cleeves, L. I., Evans, Neal J., I., Garrod, R. T., Jin, M., Kim, C. H., Kim, J., Lee, J.-E., Sakai, N., Shingledecker, C. N., Shope, B., Tobin, J. J., and van Dishoeck, E. (2022). CORINOS I: JWST/MIRI Spectroscopy and Imaging of a Class 0 protostar IRAS 15398-3359. *arXiv e-prints*, page arXiv:2208.10673. 2.3.3