UNIVERSITY COLLEGE LONDON

Faculty of Mathematics and Physical Sciences

Department of Physics & Astronomy

NOVEL ALGORITHMS FOR THE UNDERSTANDING OF THE CHEMICAL COSMOS

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

by

Antonios Makrymallis

Supervisors:	Examiners:
Prof. Serena Viti	Prof. Martin McCoustra
Dr. Jeremy Yates	Prof. Sarah Bridle

April 15, 2015

To my family. Obviously.

I, Antonios Makrymallis, 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

Molecular data from the interstellar medium (ISM) contain information that holds the key to understanding our chemically controlled cosmos and to unlocking the secrets of our universe. Observational data, as well as synthetic data from chemical codes, provide a cornucopia of digital information that conceals knowledge of the ISM. Astrochemistry studies the chemical interactions in the ISM and translates this information into knowledge of the physical characteristics of the ISM. As larger datasets and more complex models are being employed in astrochemistry, the need for intelligent data mining algorithms will increase. Machine learning algorithms provide novel methods for human-driven analysis of astrochemical data by augmenting scientific intelligence. The aim of this thesis is to introduce machine learning methods for solving typical astrochemical problems. The main application focus will be the physical parameter profile of dark molecular clouds.

Time-dependent chemical codes are typically used as a tool to interpret observations, but their potential to explore a large physical and chemical parameter space is often neglected due to the computational complexity or the complexity of the parameter space. We will present clustering analysis methods, using traditional and probabilistic hierarchical clustering, for the efficient discovery of structure and patterns in vast parameter spaces generated solely from an astrochemical code. Moreover, we will demonstrate how Bayesian methods in conjunction with Markov Chain Monte Carlo sampling algorithms can efficiently solve nonlinear inverse problems for the probabilistic estimation of chemical and physical parameters of dark molecular clouds. The computational cost of sampling algorithms can be preventive for a full Bayesian approach in some cases, hence we will also present how artificial neural networks can accelerate the inference process without much loss of accuracy. Finally, we will demonstrate how the Bayesian approach and smart

sampling techniques can tackle uncertainty about surface reactions and rate coefficients, even with vague and not very informative observational constraints, and assist laboratory astrochemists by guiding experimental techniques probabilistically .

Acknowledgements

First and foremost I would like to especially thank Serena Viti. I am really grateful for the knowledge, the inspiration, the support and the mindset. I would also like to express my thanks to all my friends and collaborators from UCL for the science, motivation, discussions and great times. Thank you to my second supervisor Jeremy Yates for all his help and opportunities. I also acknowledge IMPACT studenship and Columba Systems, the funding sources that made my PhD work possible. Obviously, I would like to thank my family for all their love and encouragement. Finally, I am especially grateful to all the people that helped me break out of my routine and without whom I would have finished this PhD and thesis earlier.

... throw off the bowlines. Sail away from the safe harbor. Catch the trade winds in your sails. Explore. Dream. Discover.

Mark Twain

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When one thinks of our Universe, wondering about all the planets, stars, galaxies, intergalactic and interstellar space or all the matter, energy and the smallest subatomic particles, some scientific, ethical or religious questions might spring to mind. The Universe usually answers back with chemistry. More than anything, our cosmos is chemically controlled and even though only about 0.5% of the total mass of the universe is composed of molecules, astronomers still insist on the title 'Molecular Universe' (Fraser et al. 2002; Cernicharo & Bachiller 2012; Tielens 2013). Understanding the origin and evolution of interstellar molecules through chemical interactions has become a fundamental goal of modern astrophysics and holds the key to understanding the universe and our place in it. Astrochemistry, the study of the chemical interactions in the interstellar medium (ISM), is a fascinating topic that can give answers to some of the most exciting questions concerning astrophysics.

It is the molecules that regulate the interstellar gas temperature and function as our thermometers and barometers to investigate local physical conditions. It is them that interact and form larger prebiotic molecules, the building blocks of life for complex species such as the human species. And, among other processes, it is the molecular gas that functions as a reservoir of matter that is to be processed into galaxies, stars and planets. There is still so much to learn about the behavior of molecules in a variety of physical situations.

This knowledge enables the understanding of core processes such as the formation of stars within interstellar clouds, the death of many stars as supernovae explosions, galaxy structure and evolution and plenty more processes of scientific interest. Astrochemical research is an essential tool, that becomes functional only with a plethora of data and information on atomic, molecular, surface and solid-state physics and chemistry. This information is the main building block of the chemical cosmos' understanding and comes mainly in two forms: Observational data taken at the telescope and synthetic data produced using theoretical astrochemical codes. The exponential growth of new data (Becla et al. 2006; Brunner et al. 2001; Szalay et al. 2002), which is often described as a data-driven revolution in astrophysical research, empowers effective and fast research results, but also challenges, requiring new 'Big Data' approaches. The call for novel methods entails inspiration or solutions that might already be available in other data-driven research areas. Specifically, areas such as data mining, machine learning or statistical learning provide a structured way to tackle common tasks in astrophysics such as data organization, data description, astronomical classification taxonomies, astronomical concept ontologies, probabilistic inference, visualization and pattern recognition. That brings us to the subject, scope and aim of this thesis.

The ISM and understanding of its processes is still hampered by incomplete knowledge about the dense, cold and dark molecular clouds. Theoretical chemical models, in conjunction with observations, provide essential data and information to probe molecular clouds and if properly interpreted can lead to useful insights about the molecular conditions and processes. However, the vast parameter space to explore and the overabundance of observational information and synthetic data aggravate the knowledge retrieval process. In this thesis, we concentrate on molecular clouds when both gas and solid (dust) phase chemistry occur and we present probabilistic and machine learning methods to address common data mining and inverse astrochemical problems. On one hand, we demonstrate algorithms that enable qualitative and quantitative data modelling for efficient, fast and structured inference from large molecular abundance data sets. On the other hand, we employ Bayesian inference to address inverse problems of estimating the structure, dynamics and processes of molecular clouds from observational data. The aim of this thesis is to provide a prototype of human-driven analysis of astrochemical data by augmenting scientific intelligence using novel techniques for astrochemistry.

The first three sections of the introduction provide a theoretical background on the

Figure 1.1: The Chemically Controlled Cosmos (Image credit: Satoshi Kambayashi).

pertinent astrophysical processes. Section 1.1 reviews the ISM processes. The physics and chemistry of molecular clouds are discussed in Section 1.2, while Section 1.3 focuses on the interstellar ices and the chemistry that takes place on the grain surfaces. Section 1.4 introduces chemical models and specifically the chemical code used in this thesis. In Section 1.5 we discuss how machine learning can be a novel approach to common knowledge discovery and inverse problems in astrochemistry. Finally, in Section 1.6 we talk about the work in this thesis. This introduction is intended to establish fundamental theoretical principles that guided the development of the chapters in this thesis.

1.1 The Interstellar Medium

The ISM in our galaxy is filled with gas (99%) and silicate and carbonaceous dust grains (1%). It consists of roughly 89% hydrogen, 9% helium and 2% heavier atoms (Dyson &

Table 1.1: Types of interstellar medium and their physical characteristics. (Adapted from Williams & Viti (2014))

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

Williams 1997). This interstellar matter can be found in neutral, ionized, atomic and molecular form and in the gas phase or in the solid phase and can be in several physical phases. Table 1.1 summarizes the physical characteristics of these phases. Temperature is a dominant and critical parameter for the ISM processes. Heating of the interstellar gas and dust is achieved through energy from interstellar radiation, cosmic rays and X-rays, which is transfered to the gas through molecular collisions. Cooling is achieved through several molecular microscopic processes that depend on the local physical characteristics of the gas and result in emitted radiation (Dyson & Williams 1997).

It is impressive how the ISM gas and dust evolution follows a cyclic process. This evolutionary cycle inside the ISM is schematically depicted in Figure 1.2. The ashes from present or past stars are injected in the ISM, and it is these ashes that drive the interstellar gas and dust to gravitational collapse that eventually forms new stars. Hence, the ISM is the birthplace of stars, but stars are the ones that regulate the structure and processes of the gas and therefore the star formation process. Both low and high mass stars control the mass balance of interstellar gas and inject dust and aromatic hydrocarbon molecules (PAHs) into the ISM through stellar winds. Dust is an important source of opacity, while PAHs are important heating agents of the ISM gas. Furthermore, stellar winds and supernova explosions from high-mass stars contribute towards star formation and the support of clouds against self gravity by controlling the mechanical energy injection into the ISM and therefore, the turbulent pressure. High-mass stars also control and regulate the cosmic-ray radiation and the FUV photon energy and hence to some extent, the heating ionization and dissociation of the interstellar gas. As will be discussed later, dust opacity and cosmic rays have a catalytic role in the formation of molecular gases. Then, molecular processes trigger the gravitational instability of the molecular clouds and

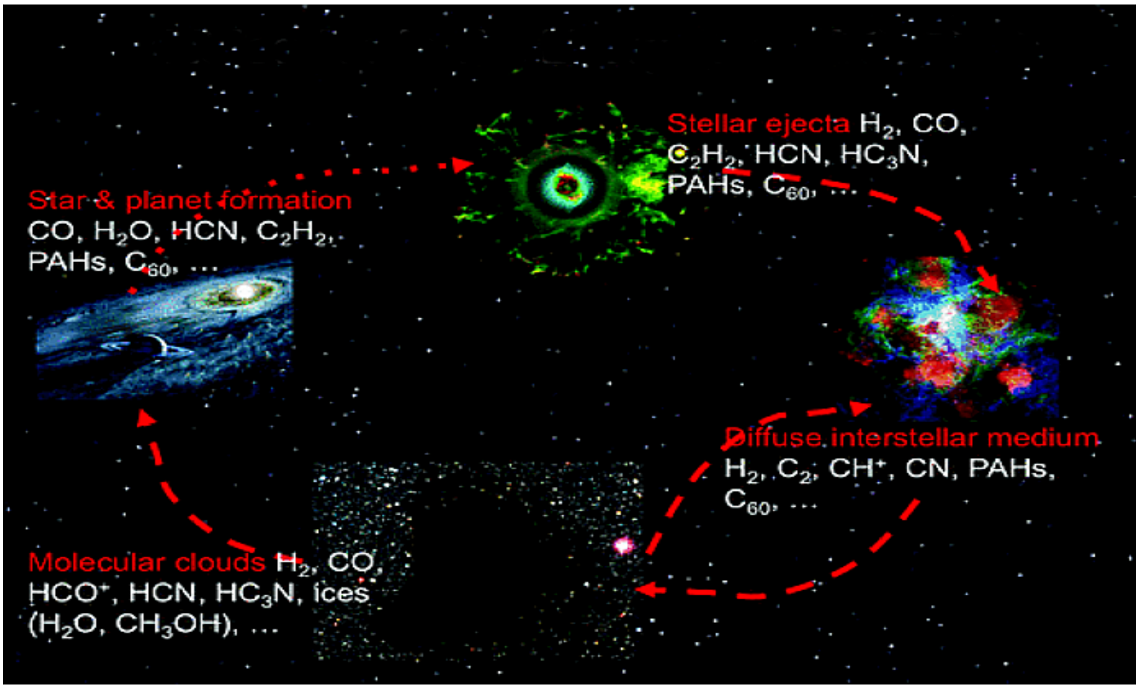

Stellar ejecta H_2, CO, C_2H_2, HCN, HC_3N, PAHs, C_{60}, ...
Star & planet formation CO, H_2O, HCN, C_2H_2, PAHs, C_{60}, ...
Diffuse interstellar medium H_2, C_2, CH^+, CN, PAHs, C_{60}, ...
Molecular clouds H_2, CO, HCO^+, HCN, HC_3N, ices (H_2O, CH_3OH), ...

Figure 1.2: The ISM cyclic process (Image credit: Tielens (2013))

consequently the star formation processes.

This clear but complex interconnection between star formation and the ISM regulates the composition, chemical evolution, structure and observational properties of the ISM in any galaxy all the way back to when the first stars and galaxies formed. To comprehend this interaction, we need to unravel the chemical and physical processes that interconnect interstellar gas to the thermal, mechanical and photon energy inputs from stars. In this thesis we will focus on novel methods for understanding chemical processes in molecular cold regions of the ISM, commonly known as molecular clouds. A complex of molecular clouds is depicted in Figure 1.3 and a more detailed description of molecular clouds and their processes follows.

1.2 Dark Molecular Clouds

Molecular hydrogen is the key to interstellar chemistry. Star formation processes start in cold clouds of molecular hydrogen which are simply called molecular clouds. Molecular clouds are irregularly shaped regions where extinction by dust is high ($A_V \gtrsim$5 mag), temperatures are low ($\sim$ 10 K), densities are inhomogeneous and high ($n_{\rm H} \geq 10^3$ cm^{-3}) and where most of the gas is molecular. They contain high density clumps or cores (Myers & Benson 1983), some of which may become gravitationally unstable and initiate the early

Figure 1.3: Orion Molecular Cloud Complex: an enormous cloud of interstellar gas and dust within the Milky Way Galaxy. (Image credit: Josh Knutson)

stages of star formation. Understanding the life cycle of dark molecular clouds is very important for comprehending star formation and for getting insight into the processes of the interstellar medium and to that extent galaxy formation. Molecules provide a paramount tool for the analysis of the chemical and physical conditions of star forming regions. Every stellar or planetary evolutionary stage is characterized by a chemical composition, which represents the physical processes of its phase.

1.2.1 The Chemistry of Molecular Clouds

The initial gas-phase chemistry and the efficiency of the gas to form molecules depends highly on how much of the hydrogen is molecular, on the abundance of heavier elements of the ISM gas and on cosmic rays. The relative elemental abundance of these heavier elements differs from galaxy to galaxy and defines an important galaxy parameter, the metallicity. The main chemical processes of molecular clouds start when most of the hydrogen is molecular and the rest of the reactive elements are initially atomic. When the density is greater than $10^3\,cm^{-3}$, the UV radiation field will not affect the gas chemistry due to the high extinction by the dust particles ($A_V \gtrsim$5 mag), while only cosmic ray radiation can penetrate through the interior of the cloud. Cosmic rays are particles that ionize the gas and form predominantly H_2^+, but also H^+. The role of cosmic rays is very important, since the low gas temperature prevents oxygen or other species reacting with molecular hydrogen. However, the ion H_2^+ has a high probability for reacting with molecular hydrogen forming H_3^+, which then provides the necessary route to more complex molecules. H_3^+ is very reactive and can donate a proton to many species such as oxygen and carbon. Further reactions with molecular hydrogen can occur until a saturated molecule is reached such as H_2O. An example sequence of reactions for what is called the backbone chemistry of the H_3^+ and oxygen as well as H_3^+ and carbon is shown in Figure 1.4. The products of these backbone chemistry routes react to further enrich the interstellar molecular canvas with important molecules such as CO. However, the backbone chemistry concept is not applicable to all atomic and neutral elements. For example, nitrogen-bearing species such as NO and CN require the reaction of nitrogen with OH and CH. At that point, it is important to emphasize the critical role of cosmic rays for the ionization of the gas and dust for shielding the cloud from the interstellar radiation field. The role of dust though, is versatile and will be discussed in more detail in the next paragraph and Section 1.3.

The dust particles constitute about 1% of the molecular clouds, are irregularly shaped and are composed of silicates, carbon, ice, and/or iron compounds. The various and paramount roles of dust in the ISM entitle dust as the interstellar catalytic converter (Hartquist & Williams 2008). The first step towards chemistry in the ISM is the formation of molecular hydrogen from atomic hydrogen. Molecular hydrogen forms in dense regions within molecular clouds, on the surfaces of dust grains. The high density and low temperature profile of star-forming regions create the perfect environment for atoms and molecules to collide with dust grains frequently. These interactions and reactions on dust grain surfaces are considered as important and maybe even more important than gas-phase chemistry since they form the molecular hydrogen. Another important role of dust that has already been mentioned is to shield gas from UV and visual radiation that would dissociate the molecular gas back to its atomic state. Furthermore, some of the absorbed radiation releases photoelectrons from the dust grains to the gases, which constitute an important energy source for the molecular gas. Finally, in the center of dark clouds, atomic and molecular gas, other than H, freeze out on to icy mantles accumulated on the dust grain surfaces. Atoms such as oxygen, carbon and nitrogen will freeze into the mantles and by hydrogen addition will be converted to water, methane and ammonia (Tielens 2010). These frozen molecules acquire mobility and are believed to form new, more complex species. Figure 1.5 depicts the core structure of a dust grain and the main routes of interstellar ice processing in a dark cloud.

1.2.2 The Physics of Molecular Clouds

The virial theorem states that in order to maintain equilibrium, the internal thermal energy must equal half the gravitational potential energy. For a timescale over $\sim 3 \times 10^7$ years and depending on the mass of the cloud, turbulent and magnetic pressure and gravity will be in balance and keep the molecular cloud in a stable state. However, when part of the magnetic or turbulent support is lost, the core will start to collapse gravitationally. A molecular cloud can and may collapse and fragment into smaller cores. Each of the cold cores collapses in an isothermal manner since the gas (atoms and molecules) releases energy in the form of radiation (Bergin & Tafalla 2007). During the collapse the density will increase and after a point ($10^5\,cm^{-3} - 10^7\,cm^{-3}$) the fragmented cores will be optically opaque. This opacity will make the energy release through radiation less efficient and the central cores, called protostars, will gradually warm up. The increase in temperature makes solid species still frozen on the grains more mobile and, therefore, drives a rich grain-surface chemistry. When the temperature increases further (20 K $<$ T $<$ 100 K), H atoms no longer reside long enough on the grain surfaces to be dominant reactants. Some species, especially the most volatile ones, such as CO, O_2 and N_2 will start to sublimate.

When the molecular cloud reaches the hot core phase (i.e. dense, compact cores with temperatures of 100-300 K, hosting the birth of a massive star), the mantle molecules are injected back to the gas phase, where they react and form even more complex molecules for up to 10^5 years (Herbst & van Dishoeck 2009). Simultaneously to the collapse, a fraction of matter is violently ejected outward in the form of highly supersonic collimated jets and molecular outflows. When the outflowing material encounters the quiescent gas of the envelope and molecular cloud, it creates shocks, where the grain mantles are (partially) sputtered and the refractory grains are shattered.

The next stages of the star evolution are not covered in this thesis, but are briefly introduced for completeness. The objects resulting from the cloud collapse are called Young Stellar Objects (YSOs) and are simply rotating spheres of gas with a central protostar. Different mechanisms are believed to form stars of different masses. Due to insightful observations, for single low-mass and intermediate-mass star formation these mechanisms are roughly understood. The conservation of the angular momentum leads the collapse of a rotating sphere of gas and dust to the formation of an accretion disk through which matter is channeled onto a central protostar. However, for high mass stars these mechanisms are not fully understood (Bonnell et al. 1998; Yorke & Sonnhalter 2002). It is believed though that in general the mechanisms are similar to the ones for low-mass star formation. At the final stages of the formation, protostars with masses less than 0.08 $M_{\odot}$, known as brown dwarfs, will not reach temperatures high enough for hydrogen nuclear fusion. Protostars with masses between 0.08 $M_{\odot}$ and 8 $M_{\odot}$ will stay $10^7 - 10^{10}$ years on the main-sequence phase and through nuclear fusion elements up to C, O and N will be formed. After that phase, if the mass of the star is relatively low ($<$ 0.23 $M_{\odot}$) it will become a white dwarf, while stars with higher masses will move into the Red Giant and Asymptotic Giant Branch (RGB and AGB) phases and will evolve to a planetary nebula with a white dwarf core. Red giant winds and planetary nebulae enrich the ISM with gas and dust and complete the cyclic evolution process. High-mass stars burn elements up to Fe until no more energetically favorable nuclear reactions can occur and the core collapses. The core can become either a neutron star, a pulsar or a black hole depending on the initial mass of the star. With supernova explosion, the outer shells of the star explode in a violent event, perturb the surrounding ISM and potentially trigger star formation. Hence, completing the cyclic evolution process.

1.3 Interstellar Ices

Icy mantles on top of dust grains were first detected by Gillett & Forrest (1973), even though Eddington (1937) had first postulated interstellar ice. Interstellar ice chemistry is controlled by the accretion rate of the gas phase species onto the grains, the desorption rate and the surface reactions network and rates. We have already remarked that during the gas-phase chemistry some atoms and molecules, called the adsorbates, freeze onto

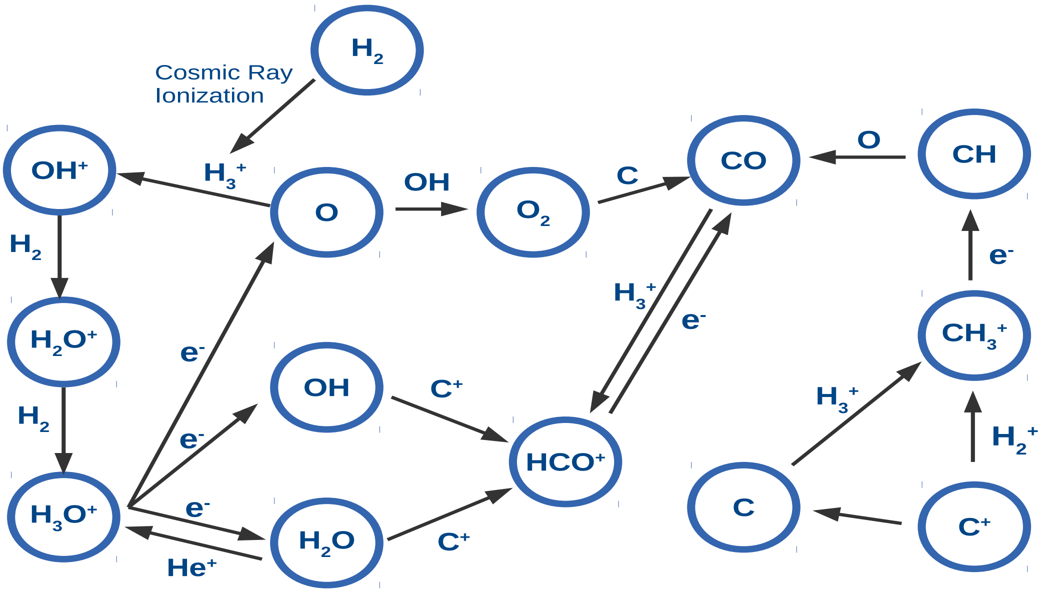


Figure 1.4: Backbone chemistry (Adapted from Bergin (2011))

the dust grains, forming an icy mantle. The rate at which species freeze-out depends on the density of the grains, the grain radius, the temperature of the dust, the mass of the species and the sticking coefficient of each species. This coefficient can be considered as the efficiency of the freeze-out and usually is treated as a free parameter. For weakly bound species though, such as CO, experiments indicate that the efficiency of the free-out is close to 100% (Bisschop et al. 2006). The accretion can occur through either weak van der Waals forces (physisorption) or chemical valence forces (chemisorption). This process is very important not only because species are removed from the gas, but also because it allows surface reactions to occur and complex molecules to form on dust grains. An understanding of freeze-out is crucial because it has a great impact on the cooling rate of the molecular gas and is also necessary to interpret observations of the emission from molecules such as gas CO. Since the timescale of the freeze-out process is much less than the expected lifetime of a typical molecular cloud, we would expect no evidence of heavy gas-phase species. However, significant observed abundances of gas phase species (Smith et al. 2004; Wakelam et al. 2006) suggest that desorption mechanisms must be in place. When the temperature is low ($< 20K$) desorption can occur either by sublimation (thermal desorption) for very light species or by non-thermal desorption mechanisms. The main non-thermal desorption mechanisms are desorption resulting from H_2 formation,

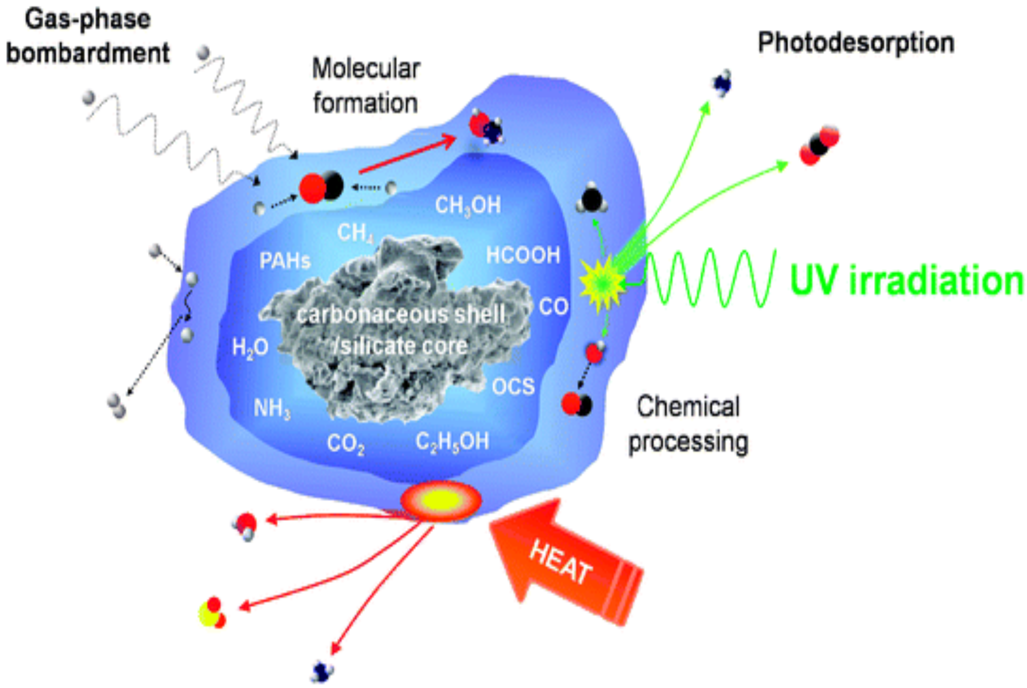


Figure 1.5: Suggested core structure and main routes of interstellar ice processing. Image credit to Burke & Brown (2010).

desorption by direct cosmic ray heating and cosmic ray induced photodesorption (Roberts et al. 2007).

As long as the molecules have frozen onto the grains, surface reactions occur through 3 main mechanisms: the Langmuir-Hinshelwood, the Eley-Rideal and the hot atom or Harris-Kasemo (Herbst & van Dishoeck 2009 and references within). These mechanisms are depicted in Figure 1.6. In the Langmuir-Hinshelwood mechanism both reactants lie in adjacent sites on the grain surface and diffusion happens through either tunneling or thermal hopping over an energy barrier between one site to an adjacent one. In both the Eley-Rideal and the hot atom mechanism, the surface reaction involves a gas phase species and an adsorbate. In the first case, the gas phase lands on the adsorbate, while in the second case the gas phase species lands and moves significantly before thermalization, so that it is able to collide with the adsorbate. The main type of surface chemical reactions that occur are hydrogenation reactions and the surface species produced are saturated ones. That is because atomic hydrogen is very mobile and a very efficient reactant on the grain surface. The most dominant of the ice species is water ice and is produced either by two sequential hydrogenations of O atoms landing on a grain:

$$O \rightarrow OH \rightarrow H_2O$$

or via a more complex hydrogenation of O_2 and O_3 (Tielens & Hagen 1982). Similarly, NH_3 and CH_4 are formed from N and C atoms respectively. Complex molecules such as methanol can be formed through hydrogenation surface reactions as well. In specific, after CO is produced in the gas-phase and accreted on the grains we can have:

$$\mathrm{CO} \rightarrow \mathrm{HCO} \rightarrow \mathrm{H_2CO} \rightarrow \mathrm{H_2COH} \rightarrow \mathrm{CH_3OH}.$$

The above process has been studied and confirmed in the laboratory by two different groups (Ioppolo et al. 2007; Watanabe & Kouchi 2002). Formation routes to species even more complex than methanol are being explored. The question is whether heavier species can be reactive enough since they diffuse much slower than atomic hydrogen. However, the efficiency or reactions that lead to the formation of ethanol and acetaldehyde from CO is found to be satisfactory.

Initially, it was only chemical intuition and gas phase chemistry analogues that was driving the surface reaction network knowledge. It took many decades for laboratory astrochemists to initiate the use of experimental techniques to test and evaluate the surface reaction inventory. Through laboratory experimentations efficiency of reaction routes are explored and even new reaction routes are revealed. However, the experimentation process is neither simple nor fast. The truth is that little experimental information is yet available for the interstellar ices. Many questions need to be answered regarding the surface reaction efficiencies, the ice composition and the energetics that have an impact on the processed ices. To disentangle the chemistry of ISM ices, laboratory work combined with chemical models constitute an invaluable tool.

1.4 Chemical Models

In recent years the molecular complexity of star forming regions has evolved and led to the development of complex, multi-point time-dependent, gas-grain chemical and photon-dominated models which accurately simulate the physics and the chemistry of the observed interstellar material (e.g. Allen & Robinson 1977; Tielens & Hagen 1982; Viti & Williams 1999; Vasyunin et al. 2009). This thesis utilizes chemical models that belong to the category of time-dependent single-point models or the time-dependent depth-dependent models that provide astrochemists with time series of molecular abundances as a function of the physical conditions of the molecular cloud and the chemical parameters of the

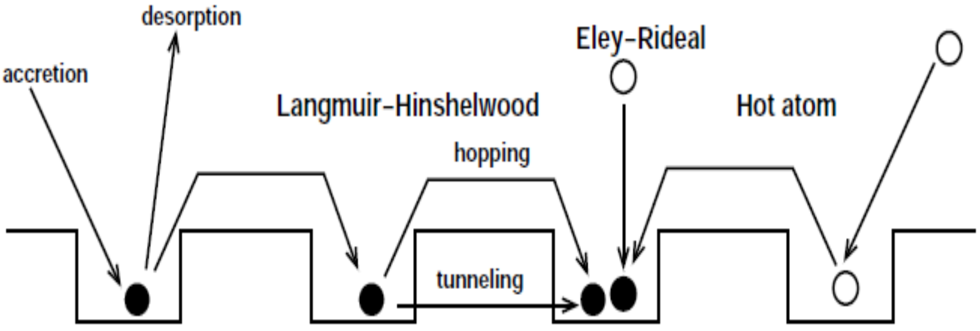


Figure 1.6: Three mechanisms for surface reactions. S is the sticking coefficient, E$_D$ the binding energy of the species to the surface and E$_b$ is the barrier from one site to the an adjacent one. Image credit to Ioppolo (2010).

defined chemical network. In particular, we consider molecular clouds as continuous time dynamical systems, where the abundance of K species $x(t) = [x_1(t), x_2(t), ..., x_K(t)]^T$ are represented by a set of K ODEs:

$$\dot{\boldsymbol{x}}(t) = \frac{d}{dt}\boldsymbol{x}(t) = f(\boldsymbol{x}(t), \boldsymbol{\theta}) = \sum production - \sum destruction$$

where $\boldsymbol{\theta}$ is a vector of physical and chemical parameters. The production and destruction terms refer to all chemical and physical processes that produce and destroy atomic and molecular species (Wakelam et al. 2013). In this thesis, our chemical modeling work will use or be based on the UCL_CHEM chemical code. This code was first implemented by Viti & Williams in 1999 and subsequently developed further by Viti et al. (2004). UCL_CHEM is a time and depth dependent gas-grain chemical model that can be used to estimate the fractional abundances (with respect to hydrogen) of gas and surface species in every environment where molecules are present. The model includes both gas and surface reactions and determines molecular abundances in environments where not only the chemistry changes with time but also local variations in physical conditions lead to variations in chemistry. Regardless of the object that is modeled, the code will always start from the most diffuse state where all the gas is in atomic form and evolve the gas to its final density. Depending on the temperature, atoms and molecules from the gas freeze on to the grains and they hydrogenate where possible. The advantage of this approach is that the ice composition is not assumed but it is derived by a time-dependent computation of the chemical evolution of the gas-dust interaction process. The main categories for the physical and chemical input parameters are the initial elemental abundances, cosmic ray ionization rate (ζ) in s^{-1}, radiation field strength ($G_{\circ}$) in $Habing$, gas density (n_H) in cm^{-3}, dust grain characteristics, freeze-out (species depletion rate), desorption processes and reaction database. In this thesis, the initial fractional elemental abundances, compared to the total number of hydrogen nuclei, were taken to be 0.14, 4.0×10^{-4} , 1.0×10^{-4}, 7.0×10^{-5}, 1.3×10^{-7}, 1.0×10^{-7} for helium, oxygen, carbon, nitrogen, sulphur and magnesium (Sofia & Meyer 2001). The gas phase network used by UCL_CHEM is based on the UMIST database (Millar et al. 2000). The chemical network also includes surface reactions as in Viti et al. (2004). In total we have 208 species and 2391 gas and surface reactions included in our network. As an output, the code will compute the fractional abundances of all atomic and molecular species included in the network as a function of time.

1.5 Machine Learning and Astrochemistry

1.5.1 Astrochemical Inference Problems

In astrochemistry, as in astrophysics and other fields, scientists contribute to the growth of knowledge by formulating and solving problems that depend on the understanding of some observed phenomenon or some instances of it. These problems can be classified as either forward or inverse problems. For the forward problems, a theoretical model is formulated that relates or maps the model parameters with observed or experimental data. Predicting the molecular abundances of chemical species in molecular clouds given the physical parameters of the cloud is a forward problem. This problem is solved by the development of a theoretical chemical model, such as UCL_CHEM, that simulates the cloud processes and relates the molecular cloud parameters with molecular abundances. The forward approach is usually employed to explore molecular abundances of a large number of species, under a large number of physical and chemical conditions. On the other hand, inferring the physical parameters of a molecular cloud, given molecular abundance observations, is an inverse problem. Such an inverse problem usually requires numerous runs of chemical codes until a satisfactory solution to the problem is reached. It is apparent that large data collections are or can be produced by both forward and inverse approaches. Even though gathering and maintaining these large collection of data is a problem that can be tackled with chemical codes and data management solutions, extracting useful information from the data is a very challenging task. Apart from the size of the data, the complicated nature of astrochemical models, and to that extent the complexity of the astrochemical data, amplifies the challenge.

Molecular cloud dynamics depend on a complicated, time-dependent, non-linear chemistry that strongly depends on the physical environment. The interaction of gas and dust, and hence the gas composition varies within very short timescales and the effects of chemistry and dynamics are interlocked in a complex non-linear problem. The potential complex interconnection of all the parameters with each other or with extra unknown parameters augments our difficulty to determine, specify or explore the parameter space in a straightforward way. On top of that, the large parameter space and the complexity of the physical system makes the task of parameter estimation highly perplexing. Traditionally, astrochemistry and molecular cloud physics have 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 and/or parameter space are large, complex or heterogeneous. But even if we take efficiency and tractability out of the equation, the core principles of traditional error treatment in astrochemistry can be fundamentally wrong. Traditional practices might account for observational error, but they keep treating modeling error in a deterministic way. The uncertainty of the arbitrary selected reaction network, the degree to which grain processes are incorporated in the model and the general uncertainty on the rate coefficients of numerous reactions make a deterministic treatment fallacious.

On the contrary, machine learning algorithms represents a rigorous, automated framework that intelligently discovers pertinent scientific information in a scalable and efficient way even for large datasets, with impressive results on stochastic treatment, pattern recognition, extrapolation and probabilistic inference.

1.5.2 Machine Learning

Machine learning is a novel and rapidly expanding research domain that combines artificial intelligence, statistics and computer science. Fundamentally, machine learning is about the construction of intelligent enough systems to learn and perform various tasks such as pattern discovery, extrapolation and data mining, without being explicitly programmed. Machine learning is finding its way to many scientific and industrial domains at the moment, because its fully automatic and generic methods simplify and sophisticate most of the typical data scientist tasks. For detailed and advanced information on machine learning algorithms, we refer the reader to Bishop (2006) and references therein. Here, we present a broad list of machine learning approaches which can be categorized into supervised and unsupervised methods:

> **Supervised Learning**: In supervised learning, machine learning algorithms learn to infer a function or relation between input and output data from labeled training examples. After the algorithm is trained, it can generalize the function and infer outputs from any given new input. Supervised learning is further divided to regression and classification problems. Classification is the problem of learning to group a new observation to a predefined set of classes or subpopulations, given a training set of already labeled observations. Learning to automatically classify objects detected in deep surveys to either galaxies or stars, using only the infrared information and a set of already labeled objects is a classification problem (Kovács & Szapudi 2014). On the other hand, regression is the problem of learning to infer the relationship among variables. The prediction of photometric redshifts using training samples of galaxies from the Sloan Digital Sky Survey would be a regression problem (Hoyle et al. 2014). Both regression and classification approaches can be utilized for anomaly or outlier discovery as well. Popular supervised learning algorithms include Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Gaussian processes (GPs), k-nearest neighbors (k-NN) etc.

> **Unsupervised Learning**: In unsupervised learning, machine learning algorithms learn and uncover the structures and patterns within data when the output is not known and there are no labeled or training data. Unsupervised learning is associated with a main class of machine learning problems, clustering. Clustering algorithms identify the inherent structures in data through common properties and groups data points in such a way that data in the same group are more similar to each other compared to those that belong to different groups. Partitioning galaxies in dissimilar groups of similar galaxies based on their morphology type would be a clustering problem (Peth et al. 2014). Popular unsupervised learning algorithms include k-means, mixture models, hierarchical clustering, principal component analysis, independent component analysis etc.

If we extract the high level essence of most, if not all, scientific tasks astronomers are called to accomplish, it is apparent that they coincide perfectly with the machine learning framework: characterize the known (unsupervised learning), assign the new or unknown (supervised learning, classification) and discover or predict the unknown (supervised learning, regression, outlier detection). The advantages of machine learning and data mining methods over traditional methods in astronomy and astrophysics are reflected through the numerous scientific publications that employ such methods, as well as reviews that report how machine learning fully exploits the exponentially increasing amount of available data, promising great scientific advance in astronomy (Borne 2009; Ball & Brunner 2010). However, astrochemistry is a field that still has not efficiently benefited from 'the perks' of machine learning. We hope that the benefits of machine learning in astrochemical research will be clarified and confirmed in the following chapters of this thesis. In the next sections, we review some machine learning concepts that are fundamental to the methods used in this thesis and that will appear in the upcoming chapters. We discuss Bayesian methods, Gaussian Processes and Neural Networks in general, providing an introduction to the algorithms developed and presented in this thesis

1.5.3 Bayesian Methods

Bayesian methods provide maybe the only way to make consistent and sound decisions in the face of uncertainty. Bayesian inference uses Bayes' rule to update probability of events based upon the model parameters, observed data and the evidence known already about the modeled situation. This mathematical handling of uncertainty has risen to be the basis of many machine learning systems. Bayes rule states that:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}, \tag{1.1}$$

where x is a data point and θ some model parameters. The probability of θ before any observations are made is referred to as the *prior* and denoted as $P(\theta)$. $P(x|\theta)$ is the probability of observing x given θ and is usually known as the *likelihood* and denoted as $\mathcal{L}(\theta|x)$. $P(\theta|x)$ is the *posterior probability* of the parameters θ, after we have observed x. Finally $P(x)$ is just a normalization factor, called the *evidence*. The Bayes' rule can be rewritten as:

$$P(\theta|x) = \frac{\mathcal{L}(\theta|x)P(\theta)}{P(x)} \propto \mathcal{L}(\theta|x)P(\theta) \tag{1.2}$$

The decision-making power in Bayesian methods lies with the evidence. If $P(x, \theta)$ is the joint probability of x and θ, in order to get the probability of x, $P(x)$, we need to marginalize over θ:

$$P(x) = \int P(x, \theta) d\theta = \int P(x|\theta) P(\theta) d\theta. \tag{1.3}$$

This quantity is called *marginal likelihood* and is fundamental in many machine learning algorithms.

1.5.4 Gaussian Processes Training

Gaussian process training involves the learning of the parameters θ of a Gaussian process (GP) for the estimation of a non-linear function in light of observed data. However, instead of assuming a specific model for the function (e.g. a quadratic, cubic, polynomial function etc.), GP represents the function by letting the data 'speak' for themselves, without making any assumptions about the form of the function in advance. In this section, the terms of observed data or observations are not to be confused with astronomical observations. We simply refer to statistical observations. Let us start by describing a GP and what parameters define it.

Imagine that we have a distribution of functions. Each function generated by this particular distribution has a characteristic form that is uniquely defined by the parameters of the distribution. We want to be able to decide in a probabilistic way if an unknown observed function is likely to have been generated by our distribution of functions. We would also like to find out the parameters of the distribution that have generated one or more observed functions. All these goals define the essence of Gaussian processes. A typical example would be to estimate the dependency of an observed variable y on an input $x \in X$, given by a function $f : X \to \mathbb{R}$. The data comes in the form of $D = \{(y_i, x_i), i = 1, ..., n\}$, where n is the number of observations. The inputs are given by $\boldsymbol{x} = [x_1, ..., x_n]$ and the outputs by $\boldsymbol{y} = [y_1, ..., y_n]$. A parametric approach would assume a model for f (e.g. a polynomial), and express f as a prior distribution on the weights/parameters of the model. However, in cases when we can not make any assumptions about the model, the parametric approach is too restrictive. Therefore, instead of assuming a model and trying to fit the data to the model, we would like to let the data define their dependency through the way they covary. GP offers this exact type of representation, by viewing any finite number of points, i.e. a subset of a function, as generated by a multivariate Gaussian distribution with a particular covariance matrix.

A GP is a stochastic process where the joint distribution of any finite subset of its random variables $\boldsymbol{f} = [f(x_1), ..., f(x_{n'})]$ associated with inputs $\boldsymbol{x} = [x_1, ..., x_{n'}]$ is a multivariate Gaussian distribution. GP is a generalization of a Gaussian distribution and is fully specified by its mean function $m(x)$ and covariance function $k(x, x')$. This is denoted as $f(x) \sim \mathcal{GP}(m, k)$. If we think of GP as a distribution over functions, the latter formula means that function f is distributed as a GP with mean function m and covariance function k. Without loss of generality, many authors assume that the mean function is zero, hence the properties of the process are entirely determined by the covariance function k. Any positive definite function can be used as covariance function. A popular choice though is the 'squared exponential' kernel:

$$k(x, x') = \sigma_f^2 \, exp(\frac{-(x - x')^2}{2l^2}). \tag{1.4}$$

The parameter σ_f is the maximum covariance and the parameter l is the characteristic length-scale parameter which defines how much effect distant observations have on each other. These two parameters specify fully the covariance matrix and are denoted as $\theta_k = \{\sigma_f, l\}$. We can now write that:

$$P(f|\theta_k) = \mathcal{N}(0, k(x, x')). \tag{1.5}$$

Usually, $f(x)$ can not be observed directly, but only through noisy samples, so that $y = f(x) + \varepsilon$. We assume that ε is independent and identically distributed and follows a normal distribution $\mathcal{N}(\varepsilon|0, \sigma_\varepsilon^2)$. Therefore, it follows that $y \sim \mathcal{GP}(f, k + \sigma_\varepsilon^2 \delta_{ii'})$, where $\delta_{ii'}$ is the Kronecker delta function or written differently :

$$P(y|f, \sigma_\varepsilon^2) = \mathcal{N}(f, \sigma_\varepsilon^2 \boldsymbol{I}), \tag{1.6}$$

where I is the $n \times n$ identity matrix. We can now formulate the likelihood of the data:

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta_k}, \sigma_\varepsilon^2) = \mathcal{N}(0, k + \sigma_\varepsilon^2 \boldsymbol{I}). \tag{1.7}$$

The likelihood function of y is called *marginal likelihood* and quantifies the probability that a group of measurements was generated by the same underlying stochastic process. Therefore, if we find the optimal θ_k that maximizes the marginal likelihood, we have defined both the GP that generated the data and the likelihood that these data were generated by this particular process. In order to optimize θ_k, we can maximize the log marginal likelihood [1] :

$$log P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta_k}, \sigma_\varepsilon^2) = -\frac{1}{2} log|\boldsymbol{K}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}^{-1}\boldsymbol{y} - \frac{n}{2} log(2\pi), \tag{1.8}$$

where $K = k + \sigma_\varepsilon^2 I$. This space is smooth, so any numerical optimization routine such as conjugate gradients can be used to approximate a good parameter setting. This approximation is known as type II maximum likelihood (ML-II).

By maximizing the marginal likelihood, we get a better understanding of the data and the underlying function. In reality, by learning the optimal parameters θ_k, we specify a distribution over functions that not only best describes the dependency between our data, but also quantifies in a probabilistic way the degree of belief that new data points belong to the same function or that new points were produced by the same process. As an illustrative example, Figure 1.7 depicts our belief for the distribution that generated a set of observed data.

Even though not covered in this thesis, Gaussian process regression is naturally a common extension of GP training. For completeness, we briefly describe a simple regression approach with GP. Consider trying to estimate the value y_* at a new data point x_*. Our data can be thought as a sample from a Gaussian distribution and their joint distribution will be:

$$\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K_*}^T \\ \boldsymbol{K}_* & \boldsymbol{K}_{**} \end{bmatrix}\right),$$

where we use K for training set covariances, K_* for training-test set covariances and K_** for test set covariances. Using the formula for conditioning a joint Gaussian distribution we have:

[1] A multivariate Gaussian distribution with mean vector $\boldsymbol{m}$ of length D and a symmetric positive definite covariance matrix Σ of size $D \times D$ has a joint probability density given by:
$P(\boldsymbol{x}|\boldsymbol{m}, \Sigma) = (2\pi)^{-D/2}|\Sigma|^{-1/2} exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{m}))$

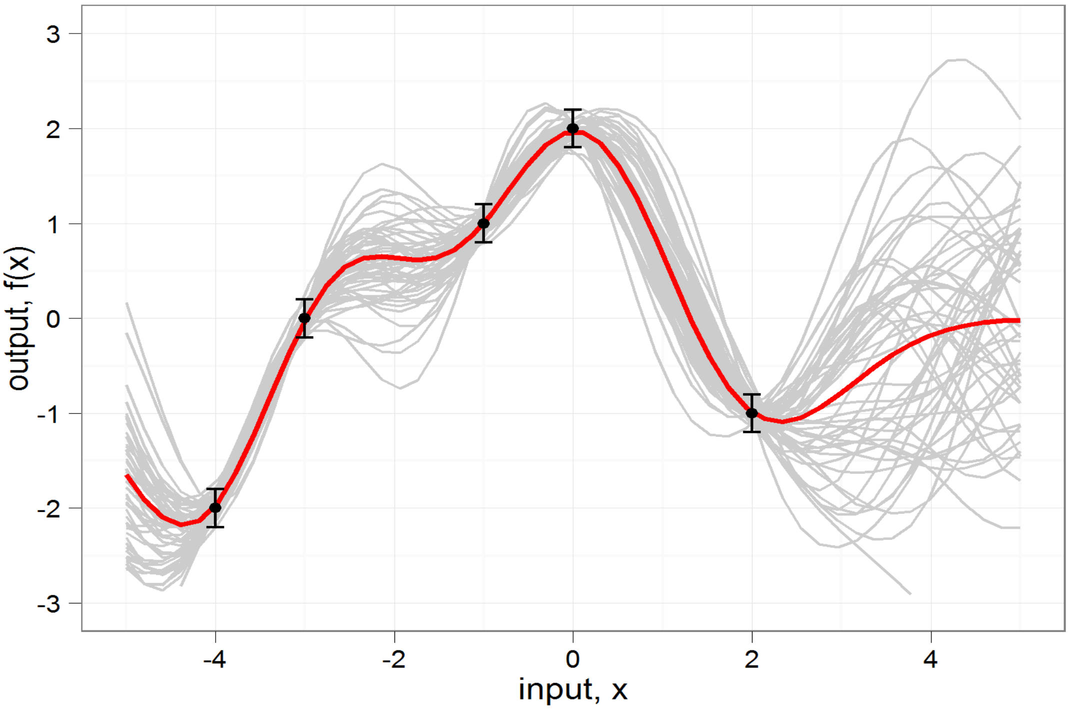


Figure 1.7: Example of Gaussian process trained on noisy data. The black dots indicate observations, the red line the mean function $m(x)$ and the gray lines samples from the Gaussian process within two standard deviations (a 95% confidence interval). Image credit: 'R-bloggers'.

$$P(y_*|\boldsymbol{y}) \sim \mathcal{N}(K_*K^{-1}\boldsymbol{y}, K_{**} - K_*K^{-1}K_*^T). \tag{1.9}$$

The best estimate is the mean of this distribution, $\bar{y}_* = K_*K^{-1}\boldsymbol{y}$ and the uncertainty about the estimate is given by the variance $var(y_*) = K_{**} - K_*K^{-1}K_*^T$. We refer readers seeking more details about Gaussian process to Rasmussen & Williams (2005).

1.5.5 Artificial Neural Networks

Artificial Neural Networks are machine learning computational models, capable of learning by example and configured for specific applications through a training process. The structure of ANNs and the way they process information is inspired from biological nervous systems such as the human brain. ANNs associate and map inputs with outputs through a group of interconnected nodes or artificial neurons. The nodes in ANNs are arranged into layers. The first layer is known as the input layer, the last one as the output layer, while the in between layers as the hidden layers. A simple ANN is shown in Figure 1.8. The more nodes per hidden layer and the more hidden layers, the higher the complexity of an ANN and hence its computational capacity, but also learning difficulty. Information is processed, transformed and passed along to other nodes and layers through weighted connections and transformation functions, known as activation functions, until an output node is reached. The objective of the training period is to learn the values of these weights, so that given a new input, the NN will predict the correct output.

There are many computational models able for machine learning. ANNs are preferred by many researchers because their attributes make them extremely suitable for various learning tasks. The universal approximation theorem (Hornik 1991) states that an ANN with one or more hidden layers, containing a finite number of nodes can construct complex input-output mappings and approximate any continuous function as long as the activation functions are locally bounded, piecewise continuous, and not a polynomial. Furthermore, even though the tuning and training of ANNs can be a very challenging task, it is one of the most studied problems in the fields of machine learning and artificial intelligence, with very well established learning techniques and continuous research and progress in training methods of even complex deep neural networks. We refer readers seeking more details about ANNs and feed forward networks to MacKay (2002).

1.6 This Thesis

This thesis reflects established and new developments in the field of machine learning and pattern recognition for astrochemical problems. We briefly present an outline of the thesis:

> **Chapter 2**: *Machine Learning and Data Mining in Time Series of Molecular Abundances*
>
> Chapter 2 provides statistical procedures to detect and highlight structure within synthetic time series of molecular abundances. The aim is to identify groups of cloud parameters that appear to regulate and control the chemical mechanism in a similar way by clustering together sets of models and parameters that exhibit similar dynamics. Hence, the goal from an astrochemical perspective is to introduce clustering techniques that will aid the understanding of the underlying parameter system and the possible pathways that lead to specific molecular abundance behavior.

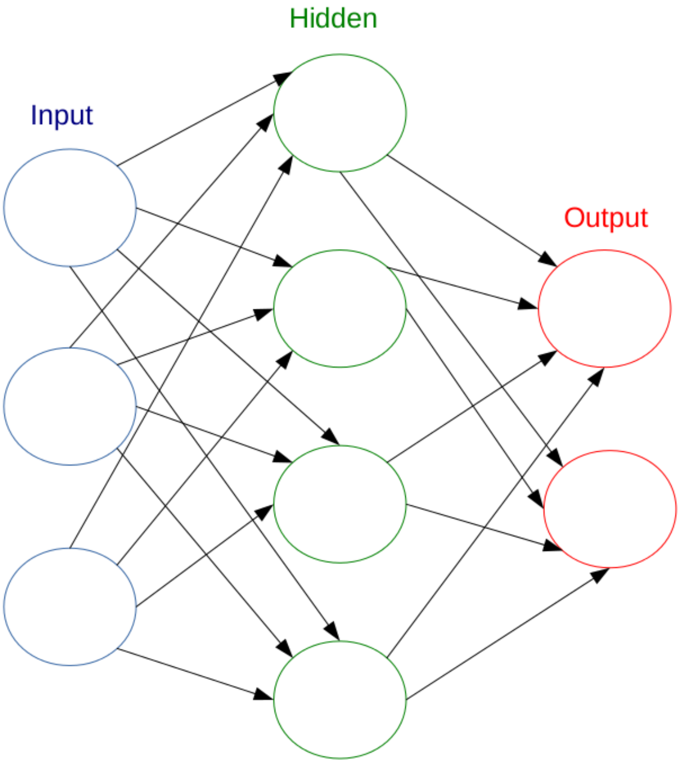


Figure 1.8: A 3-layer neural network with 3 inputs, 4 hidden nodes, and 2 outputs.

Chapter 3: *Understanding the Formation and Evolution of Interstellar Ices: A Bayesian Approach*

Understanding the physical conditions of dark molecular clouds and star forming regions is an inverse problem subject to complicated chemistry that varies non-linearly with time and the physical environment. In this chapter we apply a Bayesian approach based on a Markov Chain Monte Carlo (MCMC) method for solving the non-linear inverse problems encountered in astrochemical modelling. We use observations for ice and gas species in dark molecular clouds and a time dependent, gas grain chemical model to infer the values of the physical and chemical parameters that characterize quiescent regions of molecular clouds. We show evidence that in high dimensional problems, MCMC algorithms provide a more efficient and complete solution than more classical strategies. The results of our MCMC method enable us to derive statistical estimates and uncertainties for the physical parameters of interest as a result of the Bayesian treatment.

Chapter 4: *Fast Astrochemical Parameter Estimation with Neural Networks*

Estimating the physical and chemical conditions in dark molecular clouds is a common inverse problem in astrochemistry. Bayesian inference and Monte Carlo sampling algorithms provide a systematic and consistent approach to tackle these kind of problems. However, a fast evaluation of the likelihood and the speed of the analysis remains still a challenge. In this chapter, we present an algorithm that incorporates ANN to learn the likelihood function and substitute it with a much more rapid evaluation. We demonstrate the performance of the algorithm against an already studied inverse problem and we show evidence that ANN can be efficiently used to any astrochemical inverse problem with computationally expensive likelihood function.

Chapter 5: *Bayesian Uncertainty Analysis of Surface Reactions*

There is still too much uncertainty about surface reactions and rate coefficients. Laboratory experiments can shed some light on the solid phase reactions, but the large parameter space and the vast number of possible reactions make the task highly challenging. Chapter 5 demonstrates whether and how we can use Bayesian inference methods to explore the solid phase chemical network parameter with vague and abstract constraints. We developed a simple grain chemical code and with the help of MCMC sampling algorithms we exploited the Bayesian inference principles in order to get information about the reaction rate constants of a simplified chemical network. We show evidence that Bayesian methods provide an efficient approach to get insight on chemical parameters even with vague and not very informative constraints.

Chapter 6: *Conclusions*

Chapter 6 presents the concluding remarks of this thesis and discusses future work.

Chapter 2

Clustering Time Series of Molecular Abundances

Let us consider for a moment what inhibits our full understanding of a physical system such as molecular clouds. Clearly, we have no direct experience of the systemic processes in place and our knowledge reserve is solely based on observations. The main drawback on observations though, is our lack of control over the observational information we obtain. We can not plan, replicate or control in any way neither the kind nor the evolution of the data we observe, but only settle with collecting as much observational data as possible. The scientific significance of chemical models lies exactly upon that absence of enough, relevant and controllable information that would allow us to connect the dots between observations and molecular clouds. All chemical models, similarly to all models in general, are incomplete and inaccurate, but, without doubt, extremely useful. And they are useful not only as a tool to interpret observations, but also as a tool to reproduce, replicate and control synthetic data as substitute of real observations. In other words, our ability to understand molecular cloud processes scales with the amount of available information and chemical models are our only reliable source for 'bespoke' information about molecular clouds.

One might wonder that since there is a plethora of available chemical models in the astrochemical academic community, why we are still struggling to comprehend the ISM processes. It is certainly not that simple. The difficulty lies as much in the incompleteness of the models as in the complicated processes of the modeled system. Uncertainty or how to model uncertainty and what our models do not account for, is a problem discussed in Chapter 3 and Chapter 5. The subject of the present chapter is how to get insight out of complicated molecular cloud synthetic data. By complicated data, we mean large data sets of non-linearly interconnected data points and parameters, that are difficult to process by eye or traditional data processing applications. The complex processes of dark clouds are reflected on the complex relations between model data attributes and parameters and at the same time, the size of the data has to scale up in order to reach enough expressive power for such complex processes. To get insight out of complex synthetic data, this chapter presents an exploratory data mining method based on cluster analysis. Clustering is an unsupervised machine learning method that provides a straightforward, but sophisticated way to uncover hidden data structure, patterns and provide scientific information retrieval. In simple words, by creating natural groupings of molecular information across a large parameter space, clustering can tell us what is missing, tell us what should not be there, tell us what we can ignore and especially what we need to pursue. To our knowledge, this is the first time a systematic way to statistically explore large synthetic astrochemical data sets with 'intelligent' machine learning methods is suggested in the field of astrochemistry.

The goal of this chapter is to assist the astrochemist by providing statistical procedures to detect and highlight latent structure within synthetic time series of molecular abundances, by grouping together sets of models and parameters that exhibit similar dynamics. The aim is to identify groups of cloud parameters that appear to regulate and control the chemical mechanism in a similar way. Hence, the astrochemist' goal is to understand the underlying parameter system and the possible pathways that lead to specific molecular abundance behavior and maybe also identify possible chemical code deficiencies. Section 2.1 provides a thorough description of the astrochemical data the analysis of this chapter is based on. Hierarchical clustering, a traditional agglomerative clustering method, is presented in Section 2.2. A probabilistic approach to hierarchical clustering is described in Section 2.3 and finally, Section 2.4 concludes this chapter.

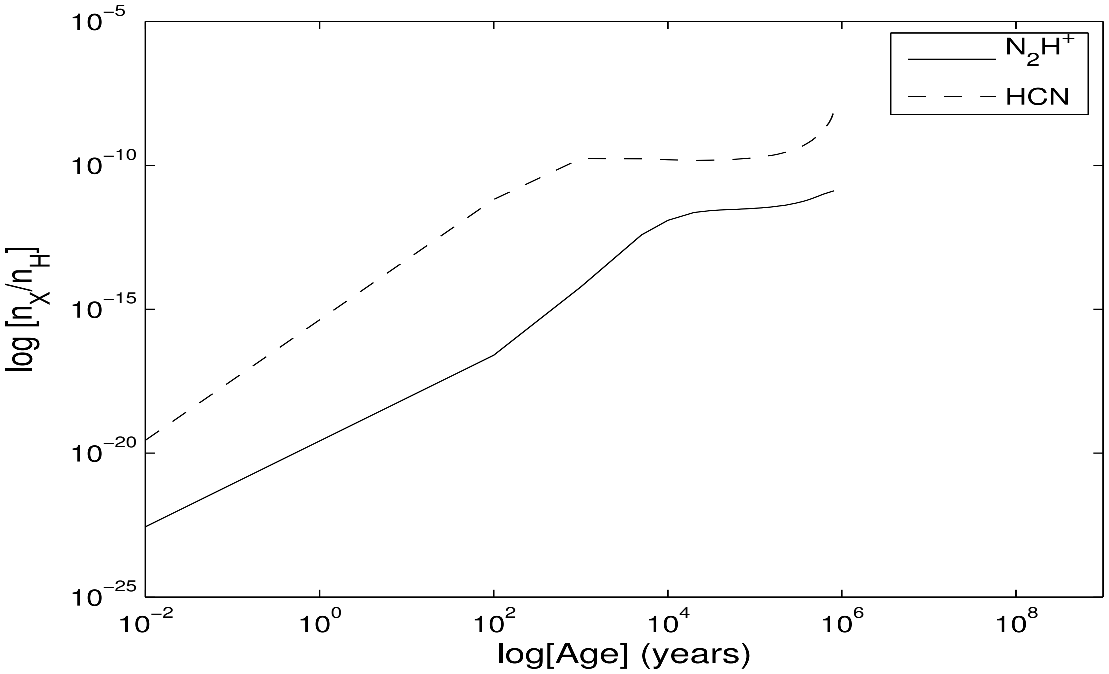


Figure 2.1: Example UCL_CHEM time series output for two random chemical species, N_2H^+ and HCN. The notation n_x indicates the abundance of species x, where x is any of our random species.

2.1 Data Understanding

For the cluster analysis we produced a large time series database using UCL_CHEM, described in detail in Section 1.4. For the database, all the models included gas-phase reactions, freeze-out, surface reactions, thermal and non-thermal desorption, with 155 gas-phase and 53 mantle species. Since we are focusing on cold molecular clouds and the evolution of ices, only the first phase of the chemical code was run, which corresponds to the period before any star is born. In total, the database consists of more than 10^5 chemical models that extend over a large parameter space, covering broad astronomical conditions. Each model consists of 208 gas-phase and mantle species time series outputs, produced according to the model specific physical and chemical parameters. For each species the database stores a sequence of molecular abundance data points, measured at successive points in time and spaced at uniform time intervals for each specific model. Example time series for two random species from two random models are shown in Figure 2.1.

Our time series data set consist of models that explore the physical conditions of molecular clouds by altering 5 basic, but critical cloud parameters: the final cloud density n_H, the cosmic ray ionization rate ζ, the radiation field rate $G_\circ$, the freeze-out parameter

Table 2.1: Explored Parameter Domain

Parameters	Unit	Explored Domain
ζ	$10^{-17} \cdot s^{-1}$	1-12
$G_\circ$	$Habing$	1-12
n_H	cm^{-3}	$10^4 - 10^8$
fr	-	$0 - 100\%$
C_f	-	$0.5 - 3$

fr and the cloud collapse rate C_f. Table 2.1 summarizes the explored parameters and the explored domain space for each one of them. To ease result interpretation we note that the Milky Way average cosmic ray ionization rate is $10^{-17} \cdot s^{-1}$. Similarly, the mean interstellar radiation field is $1.7\,Habing$ where one $Habing$ expresses the strength of a field that is equal to $10^8\,photons \cdot cm^{-2} \cdot s^{-1}$. By selecting a subgroup of the total number of 10^5 models in the database, we basically adjust the granularity of the explored parameter grid. The freeze-out parameter in our code is effectively the sticking coefficient, a number in the range of $0 - 100\%$ that adjusts the rate per unit volume at which species deplete on the grains. For the collapse to a particular n_H we used the modified formula of Rawlings et al. (1992), where parameter C_f is considered to be a retardation factor (to the free-fall) with a value less than one, to roughly mimic the magnetic and/or rotational support, or an acceleration factor with a value greater than one to simulate a collapse faster than a free-fall (e.g. due to external pressure).

To appreciate the challenges imposed by the nature of our data, we need to get a better understanding of the time series we want to cluster. For the following data understanding task we are going to assume that the UCL_CHEM user and the person that performs the data analysis are not necessarily the same person. If we assume uniformly spaced time intervals and uniform time length for all the models, the distribution of the time points for all the models should be expected to be uniform. Similarly, if we assume uniformly spaced time intervals, but variable time series' length, the distribution of the time points for all the models should be expected to be step-function shaped. Figure 2.2 uses a histogram to represent graphically the distribution of discrete time points from all the models. With a closer look at the histogram, it is easy to conclude without further investigation that none of the two hypotheses hold. Regarding the length of the time series, it is normal to assume that parameter C_f should be a causal factor. Figure 2.3(a) shows CO time series for a random number of models, color coded by the value of C_f. The segregation of

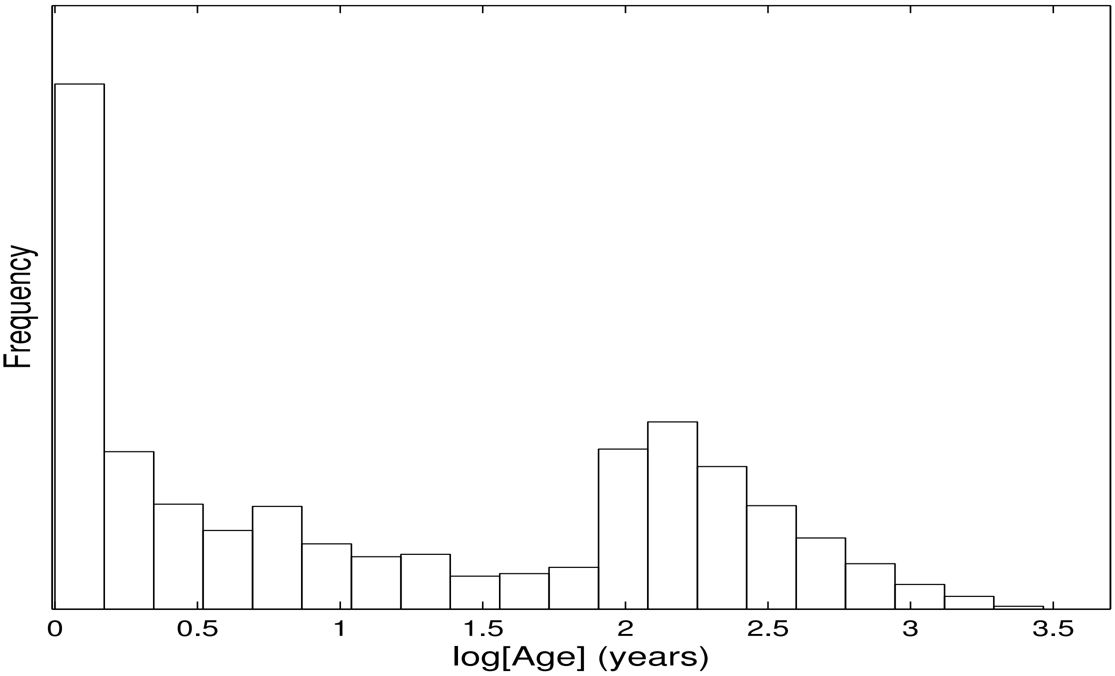
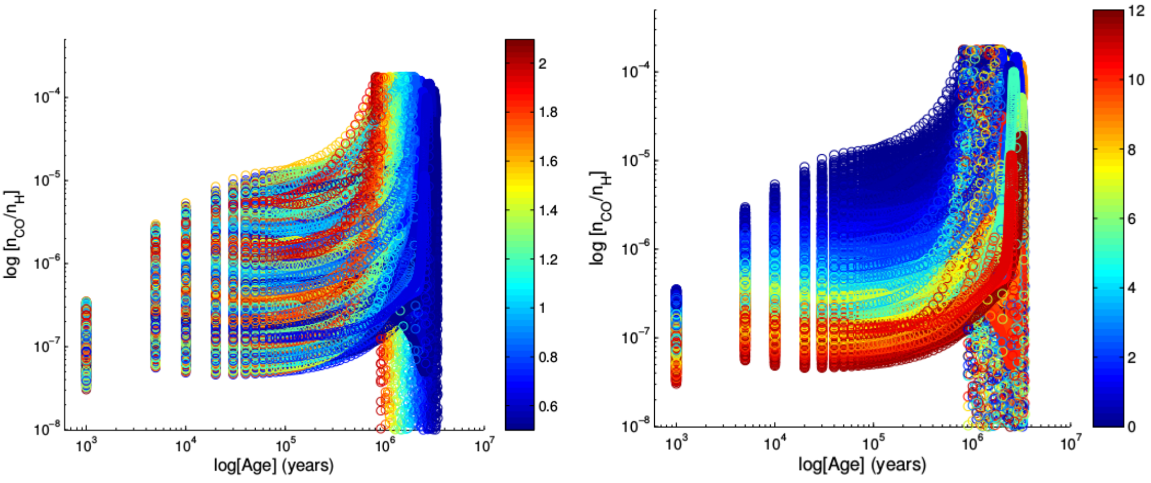


Figure 2.2: Distribution of time series sampling points for all the models of our database.

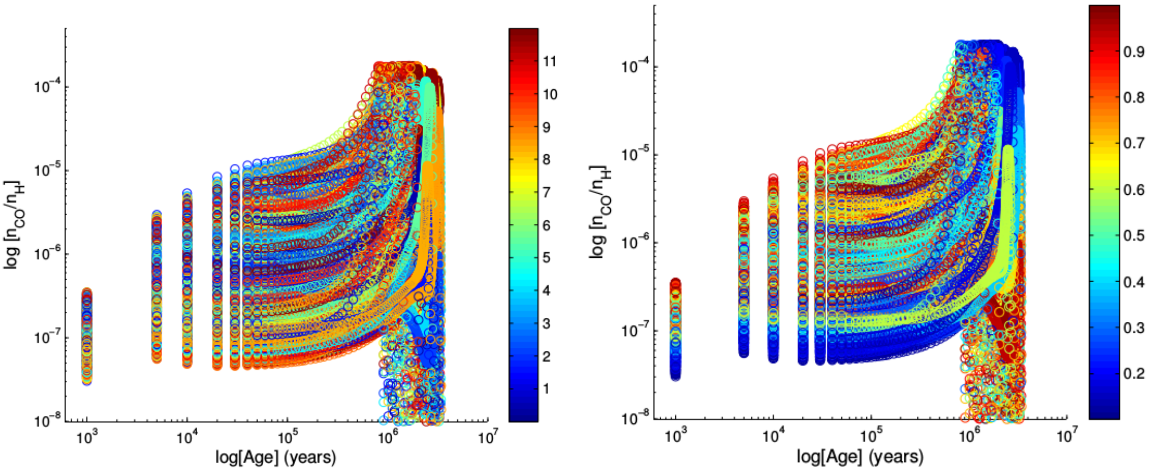
time series groups by color is clear and confirms that the collapse rate parameter has an impact on the length of the time series for each model. Note that similar plots with the same outcome can be reproduced for all the species. The list of simple and code induced parameter correlations stops somewhere here though. With the exception of a correlation between $G_\circ$ and fractional abundances for some of the species (see Figure 2.3(b)), there is no other trivial relationship that can be automatically identified by visual aid. Figure 2.4 depicts exactly that, by reproducing the plots of Figure 2.3, but color coded by ζ and fr this time. We can observe no correlation between the time series and the parameters.

To summarize, our working data set consists of time series that can have both different sampling intervals and different length if produced by a different model. The value of a species' fractional abundance at any given time point depends on the physical and chemical parameters of the model, as well as on the values of the model species at the previous time point. The parameters and species dependence might be trivial to identify, but in most cases the correlations are complex and non linear and hence difficult to uncover in a thoroughgoing way. On top of that, the number of the species and the size of the parameter space make a human driven exploratory analysis very challenging. Cluster analysis can achieve impressive results in identifying relations and uncovering patterns without any explicit definition of what we are looking for.



(a) (b)

Figure 2.3: CO time series scatter plots, color coded by one model parameter: (a) cloud collapse rate C_f, (b) radiation field rate $G_{\circ}$. Both of the parameters create a natural grouping among the CO time series.



(a) (b)

Figure 2.4: CO time series scatter plots, color coded by one model parameter: (a) cosmic ray ionization rate ζ, (b) freeze-out parameter fr. None of the parameters appear to be relevant to CO time series

Algorithm 2.1 Hierarchical Clustering

```
input: Time series data $D = \boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}$ and distance metric $E$
initialize: number of clusters $c = n$, and $D_i = \{\boldsymbol{y}^{(i)}\}$
while $c > 1$ do
Find the pair $D_i$ and $D_j$ that minimize: $dist = E(D_i, D_j)$
Merge $D_k \leftarrow D_i \cup D_j$, Delete $D_i$ and $D_j$, $c \leftarrow c - 1$
end while
output: A list of consecutive cluster merges, and a dendrogram tree visualizing the hierarchy of the cluster merges
```

2.2 Hierarchical Clustering

There are various different methods for clustering data, including hierarchical clustering (Duda & Hart 1973), spectral clustering (Ng et al. 2001), k-means clustering (Hartigan & Wong 1979) and mixture modeling (McLachlan & Peel 2000). Even though all of these methods have been proven useful to a wide range of applications, they all suffer from serious limitations. One important limitation for many of them is the necessity to prespecify the number of clusters. Hierarchical clustering not only does not suffer from that limitation, but also outputs a tree structure that provides more information and insight than the unstructured output returned by typical 'flat' clustering methods.

Given a set of time series as data points, the output of a hierarchical clustering algorithm is a dendrogram (binary tree). The leaves of the tree represent the data points, while the internal nodes of the tree represent nested hierarchies of various sizes. The length of the branches represent the dissimilarity between two time series or two groups of time series. An example dendrogram output of hierarchical clustering is shown in Figure 2.5. The advantages of hierarchical clustering come at the cost of its high complexity order. If n is the number of data points, the complexity order of the hierarchical clustering algorithm is $O(n^2 logn)$ (Jain et al. 1999), which means that it does not scale up nicely to large data sets. However, there are optimal efficient methods of complexity $O(n^2)$ that can speed up the clustering process (Sibson 1973).

For our data, the most typical hierarchical clustering algorithm was employed. That is a bottom-up agglomerative algorithm as described by Duba and Hart (1973) and presented in Algorithm 2.1. According to this algorithm, each data point will be initially assigned to its own cluster. Then, iteratively, the two closest clusters will be merged, until all the data points belong to a single cluster. The choice of the closest clusters is made based on a user defined distance measure. The most popular distance measure, which was also used

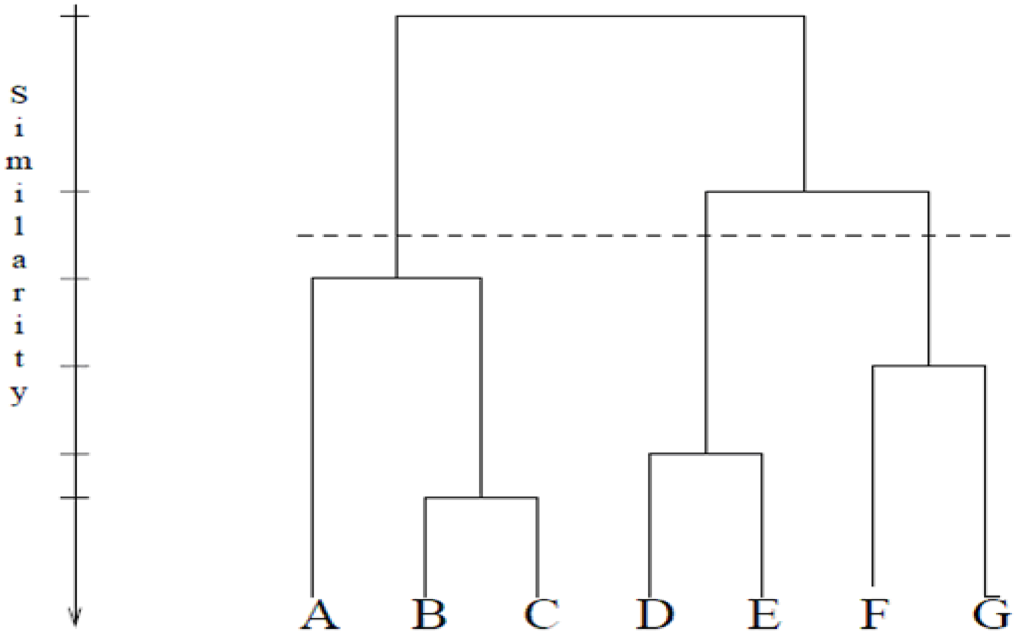


Figure 2.5: A dendrogram obtained using a hierarchical clustering algorithm. The dashed line represents a user defined level to break the tree and yield a desired clustering. Source: Jain et al. (1999)

in our case, is the Euclidean distance. Given two time series $Y^{(1)}$ and $Y^{(2)}$ of the same length N, their Euclidean distance is defined as follows:

$$E(\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}) = \sqrt{\sum_{i=1}^{N} (Y_i^{(1)} - Y_i^{(2)})^2}. \tag{2.1}$$

Both the nature of our data and the large number of time series impose certain challenges to hierarchical clustering of the total number of our time series. The different length and time intervals of time series produced by different models make the Euclidean distance measure impracticable. On top of that, the benefits of hierarchical clustering depend highly on visual inspection of the dendrogram structure by the user. The total number of our time series is massive, and obviously, a large number of time series can be inhibitive for visual assessment of the dendrogram. To overcome the latter problem, clustering was performed to models, hence sets of time series, instead of single time series. In that case, the distance metric was altered to be the sum of Euclidean distances between same species, produced by the models compared. Given two models $M^{(1)}$ and $M^{(2)}$, producing abundances for a number of S species each, their altered Euclidean distance is

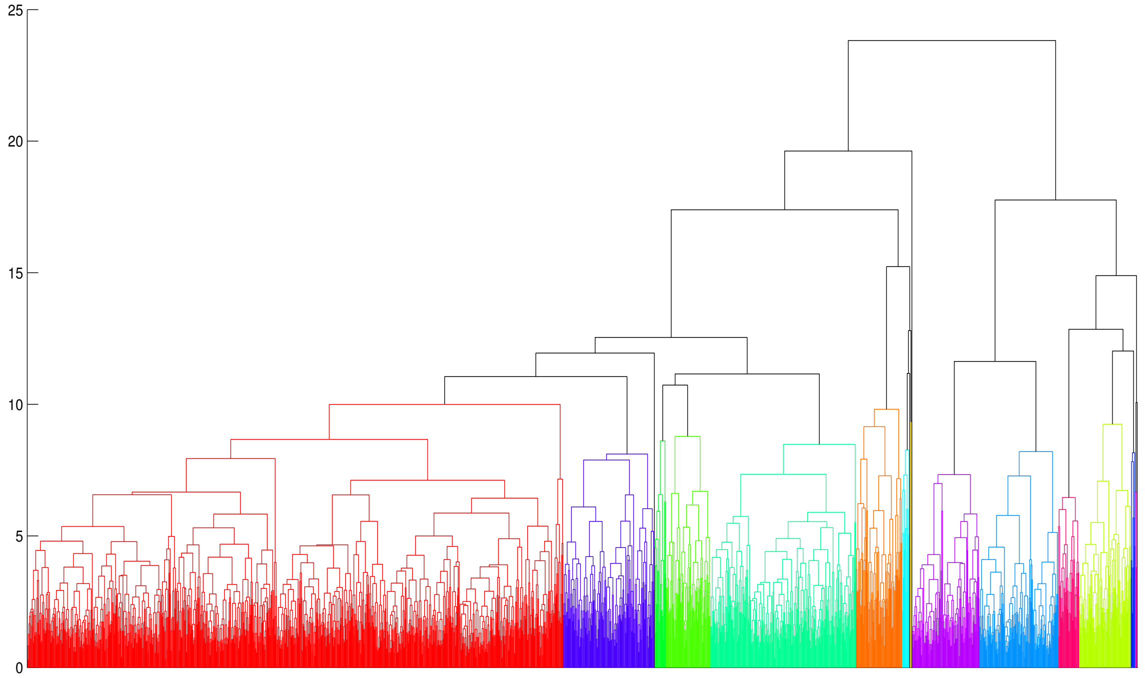


Figure 2.6: Full dendrogram for the bottom-up agglomerative hierarchical clustering on our data. Different clusters are indicated by different color according to a user defined level of pruning the tree. Each of the leaves in the dendrogram represents a model. The x-axis labeling though has been deactivated due to the large number of leaves.

defined as follows:

$$dist(M^{(1)}, M^{(2)}) = \sum_{i=1}^{S} E(M_i^{(1)}, M_i^{(2)}), \tag{2.2}$$

where $M_i^{(j)}$ is the time series of species i for model j. The problem of the different time intervals was solved easily by adapting the most granular set of time intervals as the time interval for all the time series and then filling in the missing abundance values using simple interpolation methods. Measuring similarity between time sequences of different length is a very common problem in literature without any explicit solution. Methods such as dynamic time warping (Sakoe & Chiba 1990), address successfully this problem by determining a measure of similarity that is independent of certain variations in the time dimension. However, in our case the time variations are significant of specific physical processes of the molecular cloud and should be retained and potentially highlighted. Therefore, we resorted into a heuristic way to address this problem, by penalizing uneven sets of time series. In case of time series with the same length, the Euclidean distance between two models is computed normally without any alteration. In case of different lengths

K and N, where $K > N$, the Euclidean distance is computed normally till the time point N and for the rest of the $K - N$ points a user defined penalty is introduced for each extra time point. The value of the penalty was derived by a trial and error method until the distance correctly represented the desired similarity or dissimilarity level between test time series of various different lengths. It has to be noted that all the abundance data were log-normalized and all the time series distances were divided by their length to become length invariant.

We performed hierarchical clustering to 8000 models. The output dendrogram of our cluster analysis can be seen in Figure 2.6. There is no systematic way to decide the level at which to prune the tree in order to get a specific number of clusters. This level is application specific and most of the times subjective. Trial and error methods can be adapted or simply an empirical decision based on examination of the tree and the dissimilarity axis (y-axis). In our case, we decided to prune the tree at a level that yields 18 different clusters. Each cluster can be examined in two different ways: The time series of the cluster members and the distribution of the parameters of a cluster. The time series of the cluster members represent the pattern that was captured by the specific cluster. On the other hand, the parameter distribution reflect the range and type of parameters that reproduce the specific pattern. Time series can be visualized with simple plots, while the distribution of each parameter can be visualized with histograms. This chapter focuses on the methodology, hence lacks any specific application goal. Considering that the astrochemical focus of the present thesis is on interstellar ices and that it is impractical to visualize the time series for all the species, we decided to present several noteworthy clusters using as a basis the ice H_2O:

> Cluster 1: The results of cluster 1 are shown in Figure 2.7. It is obvious that some of the parameter distributions are bimodal, which suggests that different combinations of parameters produce similar time series patterns. The ζ is constrained to low values ($< 4 \cdot 10^{-17}\, s^{-1}$), while the $G_\circ$ seems to have one peak around 2 and one around 9 $Habing$. The n_H has a peak around $8 \times 10^5\, cm^{-3}$, but with considerable probability density throughout the whole explored parameter space. The fr seems better constrained with values around 0.8. Finally, the C_f has one main peak around the expected free fall collapse and one smaller one for a collapse accelerated by a factor of 2. The latter parameter profile seems to slowly increase the abundance

of ice water until the first 1000 years and then rapidly increase its abundance until 10^5 years. At that point, the abundance of ice water appears to plateau. The final fractional abundance of ice water reaches 10^{-4}. High abundance of ice water with a high n_H and fr parameter profile seems to be consistent with the high abundance of ice water reported in literature (Pontoppidan et al. 2005).

Cluster 2: The second cluster represents a parameter profile that fails to produce significant ice water abundances and its result figures are shown in Figure 2.8. Both the ζ and $G_\circ$ are constrained to values higher than $8 \cdot 10^{-17}\, s^{-1}$ and $Habing$ respectively. The n_H has a peak around $2 \times 10^5\, cm^{-3}$, but again with significant probability density throughout the whole explored parameter space. The depletion rate this time though, is far from dominant and is constrained to values less than 0.25. Finally, the C_f has again one main peak around the expected free fall collapse and one smaller one for a collapse accelerated by a factor of 2. The parameter profile of this cluster seems to produce the same trend as cluster 1 for ice water abundance for the first 1000 years. However until 10^5 years the abundance either drops significantly or remains the same and then increases slightly to reach a fractional abundance of about 10^{-11}. The time series profile of cluster 2 is far from any observational data obtained for ice water. The low abundance of ice water can be justified by the high values of ζ and $G_\circ$ in conjunction with the low values of fr. Low freeze-out values obviously restrain adsorption which explains directly the low abundance of ice water. Moreover, the high values of ζ result in high cosmic ray induced desorption. Finally, a strong radiation field could also indirectly have an impact on the abundance of ice water. Consider that the formation of ice water requires the adsorption of O, OH or OH^+, which later hydrogenate to form ice water. High values of $G_\circ$ would dissociate OH to either O and H or OH^+ and an electron. Even though the latter products can still freeze on the grains and produce water, the delay caused by the dissociation can lead to a further decrease in water ice.

Cluster 6: The results of cluster 6 can be seen in Figure 2.9. The probability density of ζ seems to have significant density for all the explored parameter range, peaking around $7 \cdot 10^{-17}\, s^{-1}$. The $G_\circ$ this time is comparable to the interstellar radiation field, with a value around 2 $Habing$. The n_H is lower this time, with a probability peak around $5 \times 10^4\, cm^{-3}$. The depletion rate is by no means dominant with a peak at 0.3, while the C_f is constrained to a free fall collapse. The parameter profile of this cluster seems to produce an ice water abundance time series that resembles the shape of an exponential function. The ice water is increased slowly till 10^4 years and then increases exponentially to reach a fractional abundance higher than 10^{-5}. The parameter profile of this cluster along with the ice water abundance are consistent with ice observations from quiescent molecular clouds that might possibly evolve to low mass stars (Whittet et al. 2011; Makrymallis & Viti 2014).

2.3 Probabilistic Hierarchical Clustering

The traditional hierarchical clustering might present useful insight on our data, but still suffers from several limitations. There is no systematic way to specify or suggest what the 'correct' number of clusters is and the time series must be of the same length and uniformly sampled. The first limitation is usually compensated by the visual benefits of hierarchical clustering, but in cases where the data set is massive, visualization might be intractable. Unfortunately, our data sets are usually very large and the time series are of different lengths and not uniformly sampled. Another limitation is that with traditional hierarchical clustering the outcome does not define a probabilistic model. Therefore, it is usually impossible to evaluate the performance of our clustering, compare the results with a different clustering model or cluster new time series into an existing tree. This section will present a statistical inference approach to perform agglomerative hierarchical clustering that aims to overcome most of these limitations.

Our probabilistic version of hierarchical clustering follows the same principle as the traditional agglomerative hierarchical clustering algorithm, but differs on the merging criteria of potential clusters and how a clustering setting is evaluated. We can assume that similar time series were generated by the same underlying process, so similarity is defined as the probability that the data from two or more time series arose from the same stochastic process. In order to quantify this probability we used the Gaussian process marginal likelihood. The probability of a clustering setting is evaluated by the product of the likelihood of the individual clusters. In other words, the algorithm uses Gaussian process training and marginal likelihood evaluations to quantify whether a pair of time series was produced by the same underlying function and whether a cluster setting is likely to represent homogeneous clusters. We refer the reader back to section 1.5.4 for a reminder of the notation and specifics of Gaussian process training.

Our data comes in the form $D = \{(\boldsymbol{y}_i, \boldsymbol{t}_i), i = 1, \ldots, n\}$, where $\boldsymbol{y}_i$ is a vector of the fractional abundance values of data point i for the time points $\boldsymbol{t}_i$ and n is the number of data points. The set of points represented by the leaves of a subtree T_i are denoted as $D_i \subset D$. The steps of the algorithm are outlined in Algorithm 2.2 and are as follows: At the initial stage of the algorithm, we have n clusters C_i where $i = 1, \ldots, n$, containing a single data point each, so that $D_i = \{y^{(i)}, t^{(i)}\}$ and n sub-trees $\{T_i : i = 1, \ldots, n\}$. Until all clusters are merged into one, each stage of the algorithm evaluates the merging of all possible pairs of existing trees. If the algorithm decides to merge two trees T_i and T_j into T_m, then the new tree would represent a cluster C_m that contains data $D_m = D_i \cup D_j$ (see Figure 2.10). In order to decide about a merge, we use the marginal likelihood of D_m, denoted as $\mathcal{L}(D_m)$, to quantify the probability that D_i and D_j arose from the same underlying stochastic process. We assume a probabilistic model of the form $P(\boldsymbol{y}|\theta_k)$. We recall $\theta_k = \{\sigma_f, l\}$, where σ_f and l are the parameters of the covariance function that specifies fully a Gaussian process. Then, the marginal likelihood of D_m is:

$$\mathcal{L}(D_m) = \prod_{y^{(i)}, t^{(i)} \in D_m} P(y^{(i)}|t^{(i)}) = \int \Big[\prod_{y^{(i)}, t^{(i)} \in D_m} P(y^{(i)}|t^{(i)}, \theta_k) \Big] P(\theta_k) d\theta_k. \tag{2.3}$$

The quantity $\mathcal{L}(D_m)$ can be approximated by type II maximum likelihood approximation as described in Section 1.5.4. After each merge, the probability of the overall cluster setting $\boldsymbol{C}$ is evaluated, i.e. $P(C|D)$. In a fully probabilistic setting, this is proportional to the product of the likelihood $P(D|C)$ times the prior over the clustering $P(C)$. However, for our applications, we rarely have any prior information over the underlying clustering and as such, we may safely ignore the affect of the prior. Therefore, the posterior $P(C|D)$ would be governed by the likelihood $P(D|C)$ and its value is given by:

$$r_c = P(C|D) \propto \prod_{k \in C} \mathcal{L}(D_k). \tag{2.4}$$

When all clusters are merged into one, Algorithm 2.2 evaluates the r_c for each stage and suggests as the best clustering the one before the largest decrease in r_c.

To test and demonstrate the performance of our algorithm we designed the following toy example. From the whole list of species, we selected 6 species with relatively distinct time series' shape. The selected species were CO, N_2H^+, HCO^+, ice H_2O, ice CO and O.

Algorithm 2.2 Probabilistic Hierarchical Clustering

```
input: Time series data $D = \{(\boldsymbol{t}_i, \boldsymbol{y}_i), i = 1, \ldots, n\}$, model $P(\boldsymbol{y}|\theta_k)$
initialize: number of clusters $c = n$, clusters $C_i$ and data $D_i = \{\boldsymbol{t}^{(i)}, \boldsymbol{y}^{(i)}\}$ for $i = 1, \ldots, n$
while $c > 1$ do
Calculate $r_c$
Find the pair $D_i$ and $D_j$ with the highest $\mathcal{L}(D_{i \cup j})$ to merge
Merge $D_m \leftarrow D_i \cup D_j$, $T_m \leftarrow (T_i, T_j)$. Delete $D_i$ and $D_j$, $c \leftarrow c - 1$
end while
output: A list of consecutive cluster merges and the corresponding tree
The tree can be cut where we have the largest decrease in $r_c$
```

For each one of them, we extracted the time series data for 130 random set of parameters following the specifications of Table 2.1. Our final data set consists of 780 time series. We assume that each of the species is represented by a unique process that generates time series that vary according to the parameter settings. Under that hypothesis, we expect that our algorithm will be able to identify the 6 distinct processes which generated 130 time series each. The Figure 2.11 depicts the 780 time series plotted and color coded based on cluster number. As can be seen clusters and species coincide. This toy example demonstrates that our probabilistic hierarchical clustering method can identify clusters of time series that belong to the same 'family' without facing any problem to classify together time series of different sampling rate or varying time length.

In order to discover structure in the parameter space, instead of distinguishing among species, we applied our method to a different data set. Again, considering that the astrochemical focus of the present thesis is on interstellar ices, we created a dataset that includes time series for ice H_2O for more than 10^4 parameter settings. Our algorithm yielded an estimated number of 4 clusters. Considering that it is impractical to visualize all the leaves, a pruned output dendrogram of our cluster analysis can be seen in Figure 2.12. Figure 2.13 depicts sample time series of each cluster, plotted and color coded based on the same cluster color of the output dendrogram. We can notice that clusters 1,2 and 4 of Figure 2.13 are nearly identical to the clusters 1,2 and 6 of our traditional hierarchical clustering analysis. A comparison of their parameter profiles confirms that we have indeed identified the same clusters. The probability distributions for each parameter is exactly the same, hence not presented again. We refer readers back to Figures 2.8 – 2.10. The third cluster is discussed here:

Cluster 3: The characteristic parameter distributions and sample time series for cluster 3 are shown in Figure 2.14. The sample times series of this cluster present a similar profile to the time series of cluster 2 for both the traditional and probabilistic hierarchical clustering analysis. However, the abundance of ice water seems to reach initially lower values, but after 10^5 years, higher values than the abundance profile of cluster 2. Both the ζ and $G_\circ$ are constrained to values higher than $8 \times 10^{-17}\, s^{-1}$ and $Habing$ respectively, even though $G_\circ$ has a second smaller peak around 2 $Habing$. The n_H has a peak around $8 \times 10^5\, cm^{-3}$, but again with significant probability density throughout the whole explored parameter space. The depletion rate, is even lower than cluster 2 and is constrained to values less than 0.2. Finally, the C_f has one main peak that indicates a delayed collapse with a collapse acceleration factor peaking around 0.6. The parameter profile of this cluster initially seems to fail to produce ice water and that is probably because of the high ζ and $G_\circ$ rates, in conjunction with the low depletion rate. However, because of the slow collapse and the high final density, there is enough time for ice water abundance to reach higher values than initially expected.

The reason for not combining our two example applications into a bigger exploratory analysis was solely the computational cost of the method. The probabilistic hierarchical clustering might overcome many traditional limitations, but unfortunately the computational time remains a constraint, mainly because of the type II maximum likelihood approximation.

2.4 Conclusions

In this chapter we have demonstrated how clustering analysis can help astrochemists discover structure in molecular time series data. We have introduced two different clustering methods that are both based on agglomerative hierarchical clustering, but differ on their approach to defining time series similarity. The traditional approach considers two time series to be similar when their Euclidean distance is small, whilst the probabilistic approach considers two time series to be similar if the probability that their joint time series data are generated by the same stochastic process is high. The example applications described in this chapter show that both approaches can provide very useful insight regarding the mapping between the parameter space and the abundance evolution of species, but also present some drawbacks as well. Our main conclusions are the following:

1. Both traditional and probabilistic hierarchical clustering can efficiently discover

structure in molecular abundance time series data and provide scientific insight into the range and type of dark cloud parameters that reproduce specific time series patterns. Each discovered cluster represents a specific time series pattern, while the parameter distribution of the cluster reflects the parameter profile of the pattern.

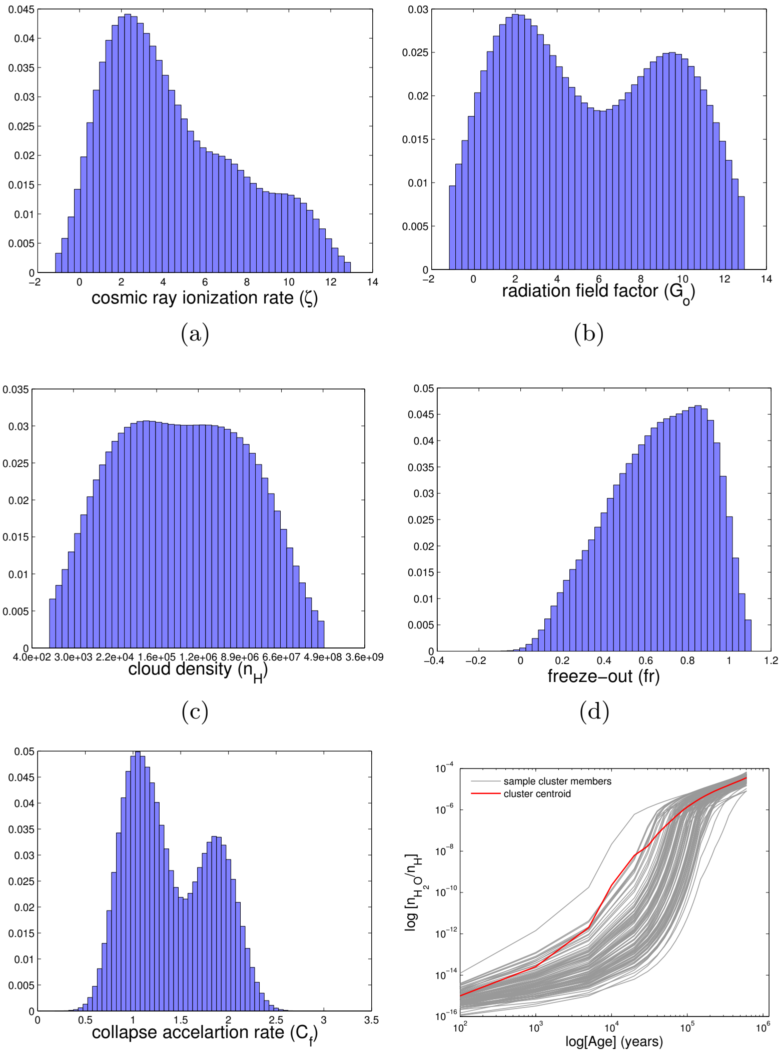
2. The results obtained by both traditional and probabilistic approach indicate that both methods identify the same patterns. Even though the traditional approach was applied considering models as data points and the probabilistic approach considering single time series as data points, results on ice water confirmed that the resulting cluster profiles were the same.

3. Traditional agglomerative clustering shows great performance on clustering efficiently our data, but presents certain limitations. Data and similarity metrics should be heuristically altered in order to accommodate for time series of varying sampling rate and length. The number of clusters is not automatically given by the algorithm and computational time is a significant issue when dealing with very big data sets.

4. The probabilistic hierarchical clustering introduced in this chapter achieves the same efficiency and manages to overcome some of the previous limitations. Time series of varying sampling rate or different length are handled naturally by the algorithm. The number of clusters is also suggested by the algorithm using a statistical model comparison criterion. The computational time though, still remains an issue.

5. The outcome of an exploratory analysis using our clustering methods could be also used in conjunction with observations or experimental data to guide lab experiments, identify chemical code deficiencies or simply constrain cloud physical parameters.

In future work, clustering analysis and especially probabilistic clustering can be used to cluster and/or classify the evolution, type and parameter profile of dark clouds or cores. Observational data never come in the form of time series. However, from a particular source we can have a sequence of observational values for a number of species. This sequence of data is called cross-sectional data and can be algorithmically treated exactly as a time series. Cross sectional data can be defined as data collected by observing many attributes (i.e. species in our case) at the same point of time. Given that we have a database of molecular abundance time series that covers a large range of species, parameters and timescales, we can perform cluster analysis on the cross sectional data of the database for each evolutionary stage of a core. Each resulting cluster would contain a group of models that produce similar abundances for a predefined set of species at a particular evolutionary stage. The parameter distribution of this group of models would be the parameter profile of the cluster. Essentially, after the analysis, each cluster will be defined by its evolutionary stage and parameter profile. Therefore, given a set of observed species from a particular source, we could classify/match the source with the corresponding cluster and hence have a parameter profile for possible evolutionary stages of the source. Probabilistic clustering is considered preferable due to its ability to deal better with abundance uncertainties and missing species' data.



(a)

(b)

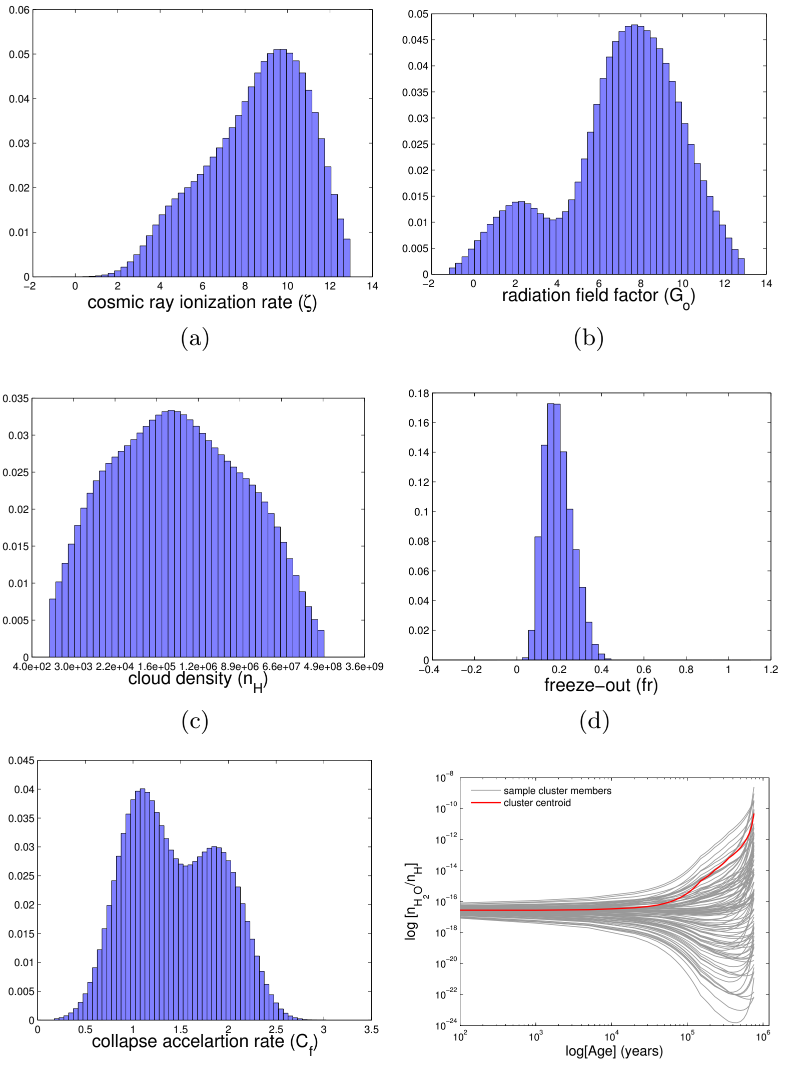
(c)

(d)

(e)

(f)

Figure 2.7: Cluster 1 after traditional hierarchical clustering. Plots (a) - (e) show the histogram of the marginalized probability distribution for each of the five parameters for Cluster 1. These plots show the Gaussian kernel density estimator of each Probability Density Function. Plot (f) shows sample time series of the cluster members for cluster 1 and the centroid of the cluster members.



(a)

(b)

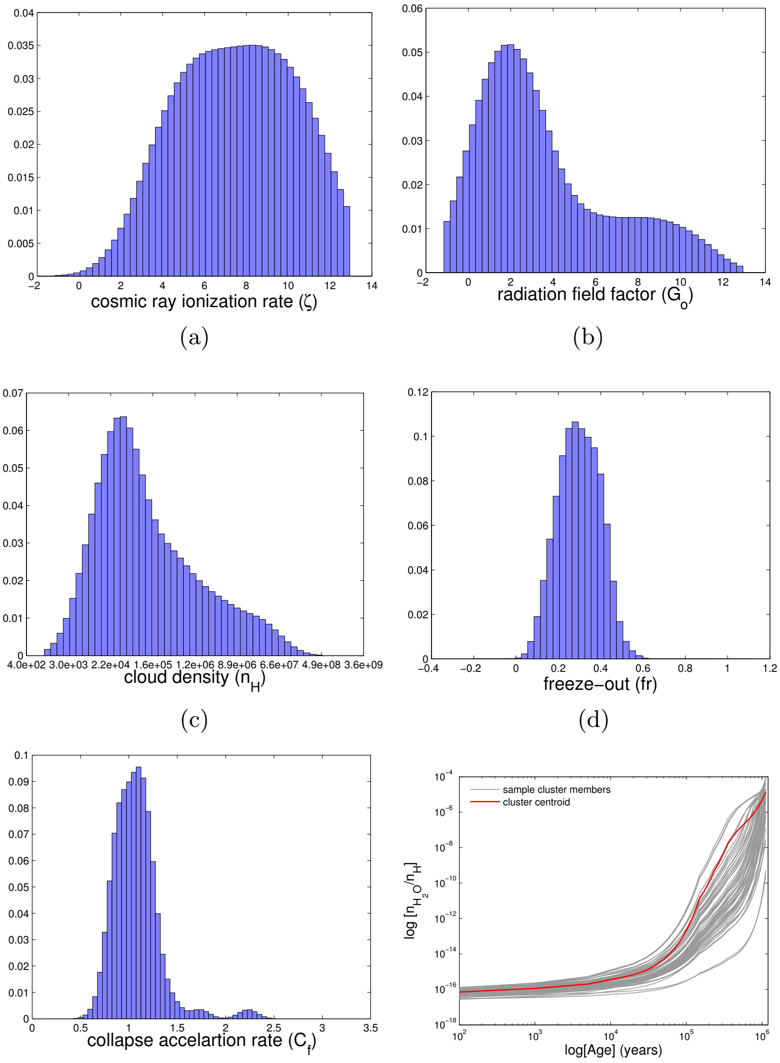
(c)

(d)

(e)

(f)

Figure 2.8: Cluster 2 after traditional hierarchical clustering. Plots (a) - (e) show the histogram of the marginalized probability distribution for each of the five parameters for Cluster 2. These plots show the Gaussian kernel density estimator of each Probability Density Function. Plot (f) shows sample time series of the cluster members for cluster 2 and the centroid of the cluster members.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2.9: Cluster 6 after traditional hierarchical clustering. Plots (a) - (e) show the histogram of the marginalized probability distribution for each of the five parameters for Cluster 6. These plots show the Gaussian kernel density estimator of each Probability Density Function. Plot (f) shows sample time series of the cluster members for cluster 6 and the centroid of the cluster members.

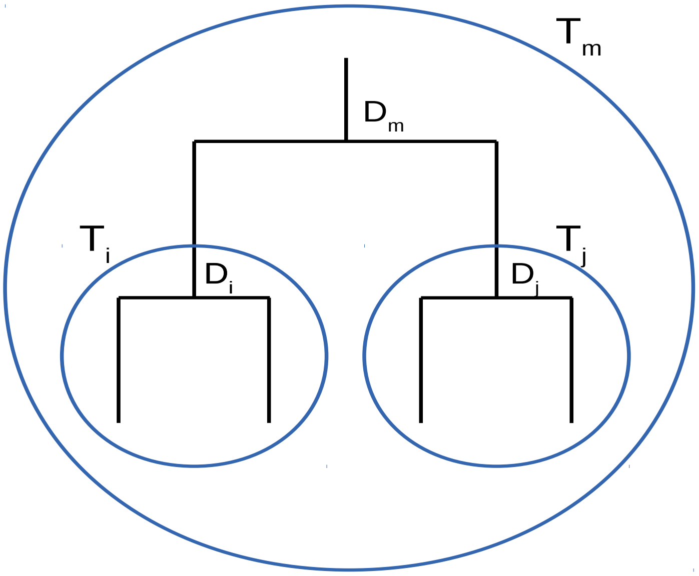


Figure 2.10: Part of a tree where T_i and T_j are merged into T_m, and the associated data sets D_i and D_j are merged into D_k. (Adapted from Heller & Ghahramani (2005)).

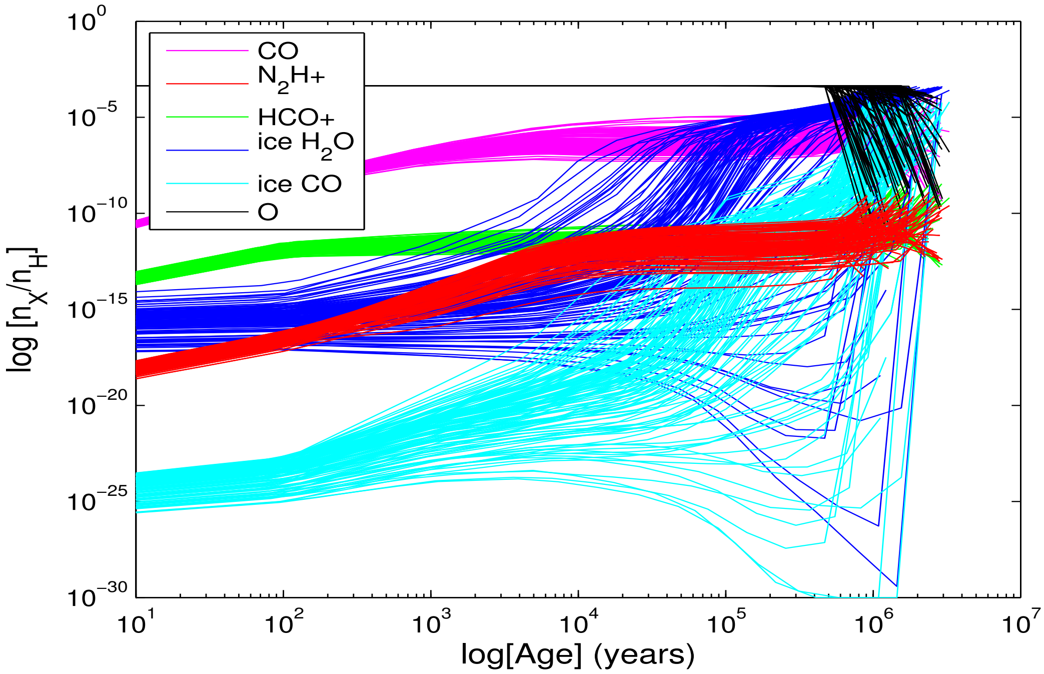


Figure 2.11: Probabilistic hierarchical clustering of our toy example application. Each color represents a distinct cluster and coincides with a distinct species as well.

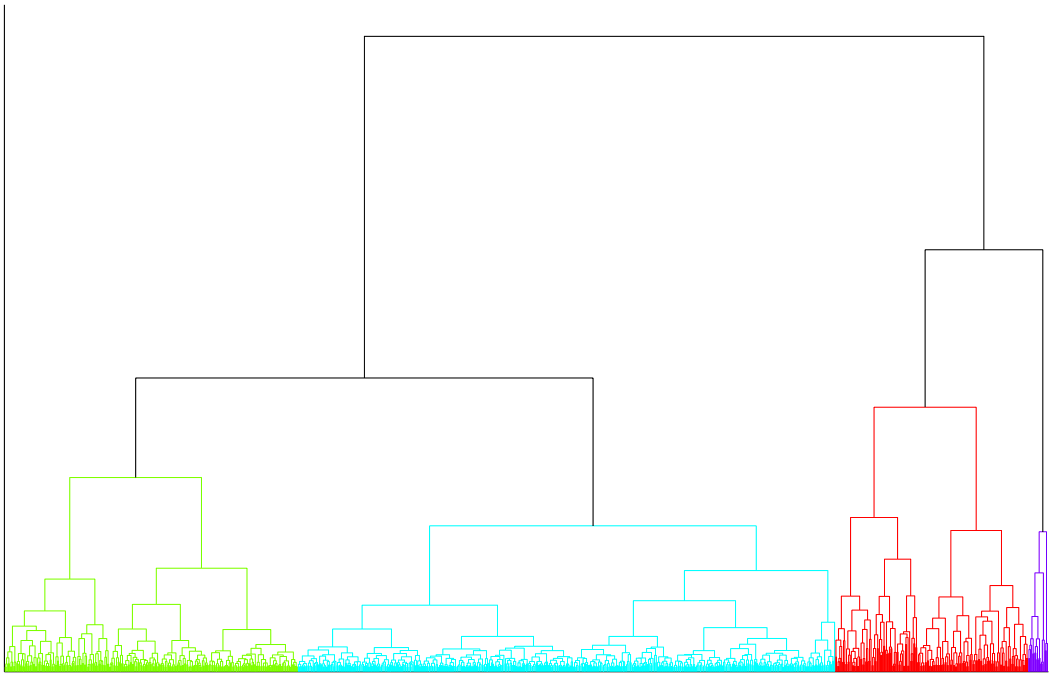


Figure 2.12: Dendrogram for the probabilistic hierarchical clustering on ice water data. Different clusters are indicated by different color according to the algorithm's suggested number of clusters. The tree is pruned for practical visualization reasons.

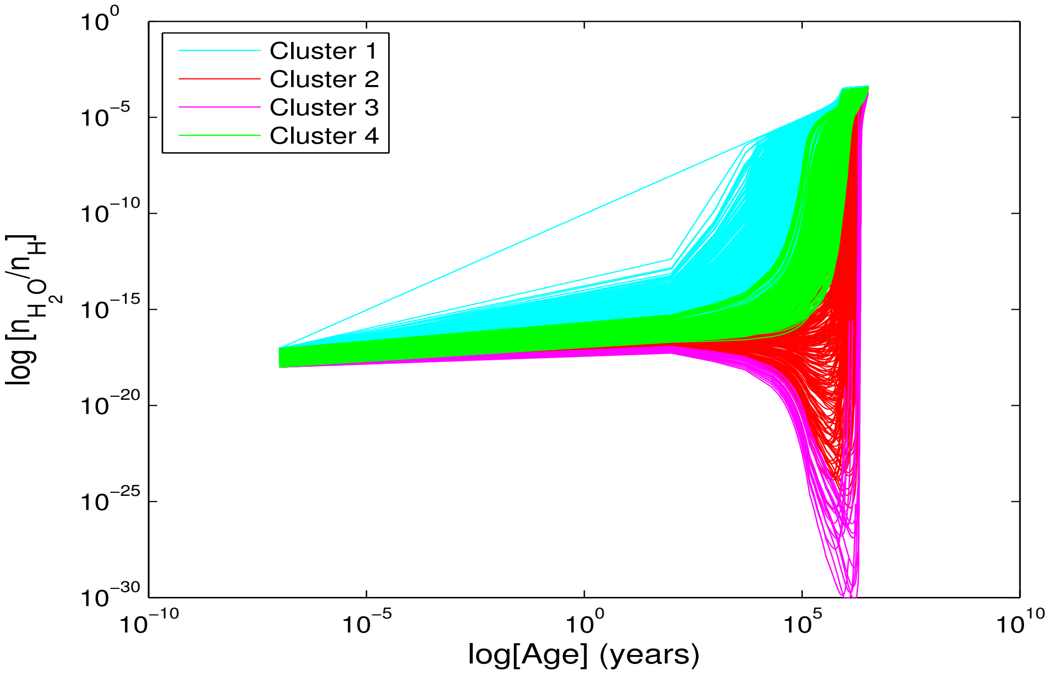
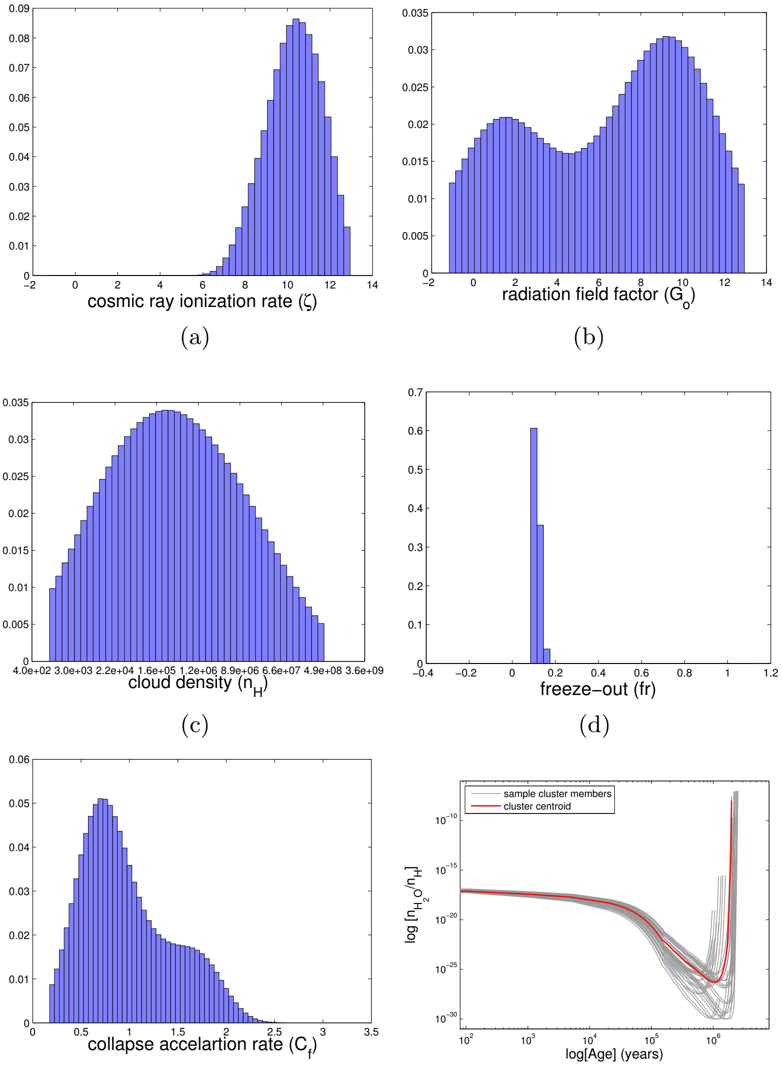


Figure 2.13: Probabilistic hierarchical clustering of different parameter models of ice H_2O. Each color represents a distinct cluster.



(a) (b)

(c) (d)

(e) (f)

Figure 2.14: Cluster 3 after probabilistic hierarchical clustering. Plots (a) - (e) show the histogram of the marginalized probability distribution for each of the five parameters for Cluster 1. These plots show the Gaussian kernel density estimator of each Probability Density Function. Plot (f) shows sample time series of the cluster members for cluster 1 and the centroid of the cluster members.

Chapter 3

Understanding the Formation of Interstellar Ices: A Bayesian Approach

The work presented in this chapter is based on the paper by Makrymallis & Viti (2014)

In the previous chapter, we presented how statistical procedures such as clustering can detect and highlight structure within synthetic time series of molecular abundances. Synthetic data from chemical codes can also function as an excellent tool for interpreting observations. By solely analyzing synthetic data, our knowledge gain towards the understanding of physical and chemical systems such as molecular clouds is constrained. Observational data can both navigate astrochemists towards a good understanding of the corresponding molecular cloud dynamics and point out chemical code deficiencies or gaps. This chapter demonstrates how Bayesian inference methods can alleviate both our uncertainty about physical processes in molecular clouds and our uncertainty about what the chemical codes do not account for.

3.1 Introduction

In the dense cores of molecular clouds, molecules and atoms previously in the gas phase, deplete onto the dust grains. For each atom or molecule, freeze-out (or depletion) depends on a complicated time-dependent, non-linear chemistry that strongly depends on the physical environment (see section 1.3). It is difficult to quantify depletion observationally (e.g. Christie et al. 2012). CO emission can be used to infer the fraction of species that is in the form of icy mantles, by taking the ratio of the observed CO to the expected abundance at a particular density in steady state, if freeze-out did not occur (e.g. Caselli et al. 1999). This however, not only implies that the cores are in steady-state, but also implies a knowledge of the H_2 density, as well as of the efficiency of the non-thermal desorption mechanisms that can return the depleted CO to the gas. Moreover, the CO depletion factor is not necessarily equivalent to the molecular gas depletion factor, because different species freeze and desorb at different rates with different sticking coefficients, which are mostly unknown.

The detection of water ice mantles in cold dark interstellar clouds and star forming regions (Öberg et al. 2011) provides us with direct evidence that surface reactions on dust grains involving oxygen atoms make water molecules, which are then retained on the surface and make water ice. Not all species undergo surface reactions when they stick to dust grains. For example, CO sticks efficiently to surfaces at temperatures below ~ 25 K and is found to be abundant in the ices. Some of this CO can be converted to other species.

The relatively high abundance of CO_2, CH_3OH, and H_2CO in ices (Öberg et al. 2011; Whittet et al. 2011), relative to H_2O, in some clouds indeed suggest that some processing of CO to these products is occurring, due possibly by irradiation, by cosmic rays or by photons generated by cosmic rays inside the cloud. H_2CO and CH_3OH are stages in the surface hydrogenation of CO. Similarly, CO_2 can be the result of oxygenation of CO:

$$\mathrm{CO} + \mathrm{OH} \rightarrow \mathrm{CO_2} + \mathrm{H}.$$

Some ices can be thermally returned to the gas phase when the gas temperature is higher than 20 K. At low gas temperatures non-thermal desorption processes can also return molecules from solid to gas-phase (e.g. Roberts et al. (2007)). However, these mechanisms 'compete' with those of freeze-out. The composition of the icy mantles is clearly a time-dependent process highly dependent on the initial conditions of the gas in any particular cloud. Hence, the ices on dust grain surfaces are of a mixed composition and may reflect the local conditions and evolutionary history. In some dark molecular clouds, the ices are abundant, indicating that non-thermal desorption mechanisms may not be very efficient everywhere. The potential interconnection and linear or non-linear correlation of these parameters with each other or with extra unknown parameters augments our difficulty to determine and specify the parameter network. The large parameter space in combination with the number of parameters and the complexity of the physical system make the task of parameter estimation highly challenging.

The increasingly detailed observations of molecular clouds and star forming regions enable us to identify some of the most important processes at work. Chemical and radiative transfer models can transform molecular observations into powerful diagnostics of the evolution and distribution of the molecular gas. The results of these models though, depend on a number of parameters or group of parameters that are most of the times poorly constrained. Moreover, deriving information about molecular clouds using observational information and, even well established modeling codes, is an inverse problem that usually does not fulfill Hadamard's (Hadamard 1902) postulates of well-posedness. That is, it may not have a solution, solutions might not be unique and/or might not depend continuously on the observational data. The first and second postulates simply state that for a well-posed problem a solution should exist and be unique. The third postulate holds when small changes in the observational data result in small changes in the solution. As shown later in Section 3.3.1, in typical astrochemical problems, only the first postulate holds and we usually have to deal with non linear ill-posed inverse problems.

Employing sampling algorithms is a traditional approach to tackle inverse problems in many scientific fields with large parameter space. Bayesian statistical techniques and Monte Carlo sampling methods such as Markov Chain Monte Carlo (MCMC) algorithms and Nested Sampling have flourished over the past decade in astrophysical data analysis (Christensen & Meyer 2000; Ford 2005; Fitzgerald et al. 2007; Feroz & Hobson 2008; Isella et al. 2009). A summary of a typical MCMC method and an application to quantify uncertainty in stellar parameters using stellar codes is given by Bazot et al. (2012). To our knowledge, MCMC methods have never been applied in the framework of parameter estimation through astrochemical modeling. In this chapter we present a first astrochemical application of gas-grain chemical modeling, molecular abundances and a Bayesian statistical approach based on MCMC methodology.

The motivation of the present chapter is to solve the inverse problem of deriving the physical conditions in interstellar molecular clouds; in particular: the gas density, cosmic ray ionization rate, radiation field, the rate of collapse, the freeze-out rate and non-thermal desorption efficiency. In Section 3.2, we formulate a typical inverse problem for interstellar molecular clouds and describe the Bayesian method and the Metropolis-Hastings (MH) algorithm (an example of a wider class of MCMC techniques). In Section 3.3, we discuss the statistical results and the astrophysical consequences. Finally in Section 3.4, we present our conclusions.

3.2 Parameter Estimation

In this chapter, we are interested in dense, cold, quiescent regions of molecular clouds where atoms and molecules in the gas phase freeze-out on to the dust grains. The observed quantities are molecular abundances for solid and gas phase species. The parameters we want to estimate are the cloud density n_H, the cosmic ray ionization rate ζ, radiation field rate $G_\circ$, the cloud collapse rate C_f and three non-thermal desorption efficiencies ϵ, ϕ, y presented in Section 3.2.3. Due to the nature of the addressed inverse problem, the theoretical and modeled relationship between the parameters and the observed data is highly non-linear. Therefore, we anticipate several degeneracies as well as a multi-modal and non-Gaussian joint parameter distribution. Moreover, the parameters are not uniquely related to the observations. While the forward problem has (in deterministic physics) a unique solution, the inverse problem does not. Different combinations of parameters can produce the same abundances. Furthermore, the possible combinations of parameters are too many to permit an exhaustive search.

Traditional approaches to tackle inverse problems of this nature fail to cope with these kind of issues. Methods based on searching iteratively to minimize an appropriate distance such as the χ^2 error, can be stuck in local minimum and give degenerate solutions. Alternative approaches to aim for a global solution such as simulated annealing would have some benefits, but since we are not just looking for the global optimum of our target distribution, the most comprehensive view is obtained by a Bayesian Monte Carlo sampling method. We selected the Bayesian MCMC approach against other methods that work equally well with complex and multimodal target distributions (e.g. Nested

Sampling), since MCMC constitutes a benchmark algorithm in Monte Carlo sampling and parameter estimation problems.

To overcome the challenges of an ill-posed nonlinear inverse problem we adopted a Bayesian approach based on the use of Metropolis-Hastings (MH) algorithm. The Bayesian framework for inverse problems is based on systematic modeling of all errors and uncertainties from the Bayesian viewpoint. The potential of this approach to solve difficult inverse problems with high noise levels and serious model uncertainties is much higher and also allows for prior information to be incorporated. The Bayesian solution is the whole posterior distribution of the parameters and therefore, there is not only one solution, but a set of possible values. The advantage of MCMC approach is that there is no restriction concerning the non-linearity of the model. Moreover, an appropriate tuning of the MCMC parameters allows the algorithm to explore all modes of the target distribution. Finally, even though it is still not feasible to do an exhaustive search through the parameter space, MCMC methods can effectively explore the parameters joint posterior distribution, since model computations are concentrated around regions of interest in the parameters space.

3.2.1 Bayesian Inverse Problem

Our aim is to obtain information about physical parameters of a molecular cloud $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$, while we measure molecular abundances $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_n)$. These quantities are related to a (forward) function $f(\cdot)$ which represents the physical and chemical processes in the cloud. The main challenge is that there is no closed form function f mapping the parameters to the observations, which could be inverted. However, given a set of parameters, estimated abundance values for the species of interest can be computed with astrochemical models denoted here as $\mathcal{C}(\cdot)$. The addressed problem in our case is how to estimate $\boldsymbol{\theta}$ from

$$\boldsymbol{\mathcal{Y}} = \mathcal{C}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \tag{3.1}$$

and according to Idier (2008) this constitutes an inverse problem. The error term $\boldsymbol{\varepsilon}$, represents both the observational noise and the modeling error between $\mathcal{C}(\cdot)$ and $f(\cdot)$.

We treat $\boldsymbol{\mathcal{Y}}$, $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$ as random variables and define the solution of the inverse problem to be the posterior probability distribution of the parameters given the observations. This allows to model the noise via its statistical properties, even though we do not know the exact instance of the noise entering our data. We can also optionally specify a priori the form of solutions that we believe to be more likely, through a prior distribution. Thereby, we can attach weights to multiple solutions which explain the data. This is the Bayesian approach to inverse problems.

Assume we have K parameters θ_k and N solid phase observable quantities $\mathcal{Y}_n$. The error ε_n on each observation $\mathcal{Y}_n$ is assumed to be normally distributed with variance σ_n^2. In addition, it is assumed that the observational errors are independent. The σ_n^2 is considered to correspond to the uncertainty on $\mathcal{Y}_n$, which is solely dictated by the observation. The probability density function of the errors is given by:

$$p_\varepsilon(\boldsymbol{\varepsilon}) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_n^2} \exp(\frac{\varepsilon_n^2}{2\sigma_n^2}).$$

Using (1), we can define the likelihood function $\mathcal{L}$ of observations given a model parametrized by a set of parameters as

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\mathcal{Y}}) = p_\varepsilon(\boldsymbol{\mathcal{Y}} - \mathcal{C}(\boldsymbol{\theta})) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_n^2} \times \exp(-\frac{1}{2}\sum_{n=1}^{N}[\frac{\mathcal{C}_n(\boldsymbol{\theta}) - \mathcal{Y}_n}{\sigma_n}]).$$

In case any prior information about the unknown parameters is available, the Bayesian approach allows for this information to be taken into account. This information can be integrated through a prior probability distribution on the parameters, say $\pi(\boldsymbol{\theta})$. Then parameter estimation can be performed through the posterior probability distribution (PPD), using Bayes' rule

$$\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}) = \frac{\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\mathcal{Y}})\pi(\boldsymbol{\theta})}{m(\boldsymbol{\mathcal{Y}})}. \tag{3.2}$$

The PPD expresses our uncertainty about the parameters after considering the observations and any prior information. The denominator is simply a normalization factor.

In reality we are not able to access the whole posterior probability distribution. Therefore, computation of parameter estimates or uncertainties is a hard task. MCMC methods are efficient methods that allow to sample from complex probability distributions and approximate complex probability densities.

3.2.2 Markov Chain Monte Carlo

MCMC methods are a powerful class of algorithms that produce random samples distributed according to the distribution of interest. The importance and efficiency of MCMC methods lies in the fact that these samples can be used to approximate the probability density of the distribution by calculating it only for a feasible number of parameter values. The MCMC framework uses a Markov chain to explore the parameter space and approximate the posterior probability distribution. This chain consist of a series of states $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(t)}, ...\boldsymbol{\theta}^{(T)}$, where the probability of $\boldsymbol{\theta}^{(t)}$ depends only on $\boldsymbol{\theta}^{(t-1)}$. MCMC methods require an algorithm for choosing states in the Markov chain in a random way. Among the several implementations of possible algorithms, we employ a MH sampling algorithm (Gilks et al. 1995). The MH algorithm will enable us to explore the parameter space and approximate efficiently the PPD. A theoretical introduction on MCMC and MH is far beyond the scope of this chapter. However, we briefly describe the MH algorithm and how MCMC is employed for parameter estimation in our case. Note that the tuning of the MH algorithm is very crucial when aiming to approximate possibly multi-modal and non-Gaussian distribution, which is the case for this study. The MH is briefly outlined here using the following pseudocode:

1. Select a starting point $\boldsymbol{\theta}^{(1)}$ from the parameter space. Then for $i = 2, 3, ...$ until convergence, repeat the following steps.
2. Propose a random set of parameters according to a proposal distribution q, so that $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^i|\boldsymbol{\theta}^{i-1})$
3. Calculate the posterior probability of the new parameters, $\pi(\boldsymbol{\theta}^*|\boldsymbol{\mathcal{Y}})$, using equation 3.2
4. Accept the new parameters with probability
$$\alpha(\theta^*|\theta^{i-1}) = min\{1, \frac{q(\theta^{i-1}|\theta^*)\pi(\theta^*|\mathcal{Y})}{q(\theta^*|\theta^{i-1})\pi(\theta^{i-1}|\mathcal{Y})}\}$$
5. Calculate $u \sim Uniform(u; 0, 1)$
6. if $u < a$ then accept the proposal, $\theta^i \leftarrow \theta^*$; otherwise, reject the proposal and $\theta^i \leftarrow \theta^{i-1}$

The performance of the MH algorithm is highly dependent on the proposal distribution. The appropriate distribution should account for the complexity of the target distribution but it should still be computationally easy to draw samples from. In non-linear problems such as ours, we expect a multimodal non-Gaussian target joint distribution. Non gaussianity is not a problem for MCMC algorithms. However classical choices for the proposal distribution (i.e. Gaussian distribution) can potentially prevent the MCMC to converge to the target distribution, since the transition of the chain from one mode to another is not very possible. In our specific case, following former similar choices (eg. Bazot et al. (2012)), and taking into account the characteristics of the expected target distribution, we chose the proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^t)$ to be a mixture of two Gaussians distribution centered at $\boldsymbol{\theta}^{(t)}$ and a uniform distribution on $\mathcal{D}_\theta$. Hence, for all parameters θ^*_k for k=1,..,9

$$\boldsymbol{\theta}^* \sim \mathcal{N}_{\mathcal{D}_\theta}(\theta^t_k, \sigma^2_{k,1})\text{with probability } 40\%$$
$$\boldsymbol{\theta}^* \sim \mathcal{N}_{\mathcal{D}_\theta}(\theta^t_k, \sigma^2_{k,2})\text{with probability } 40\%$$
$$\boldsymbol{\theta}^* \sim \mathcal{U}_{\mathcal{D}_\theta}\text{with probability } 20\%$$

The values for $\sigma^2_{k,1}$ and $\sigma^2_{k,2}$ were selected based on test runs. We run $m = 8$ independent Markov chains of length $T = 200000$. By using parallel and independent chains it is easier to understand the dependence of the MH performance on the initial parameter values guesses. Moreover, parallel chains provide insight on whether convergence has been reached. Convergence was also decided based on empirical graphical aid. The length T of the chains was chosen confidently larger than the value of decided convergence. In our case $q(\cdot)$ will be a symmetrical distribution. That means that $q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ and the ratio in the acceptance probability α is simply the PPD ratio computed at $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^t$. In simple words, that parameters that increase the PPD are always accepted, while parameters that decrease the PPD are randomly accepted based on α.

3.2.3 Parameter Space

The chemical modeling code used in this chapter and denoted as $\mathcal{C}(\cdot)$ in (3.1) is the UCL_CHEM time-dependent gas-grain chemical code (Viti et al. 2004) which is briefly described in section 1.4. Note that for each set of parameters, $\mathcal{C}(\cdot)$ provide us with time series of chemical abundances. We choose to extract the chemical abundances of interest for the time points when the final density is reached and the cloud collapse has finished. Even though we ignore the previous time points, the time-dependency is still taken into account and investigated through exploration of different final density values.

Parameters $\boldsymbol{\theta}$	Unit	Definition Domain $\mathbb{D}_\theta$
ζ	$10^{-17} \cdot s^{-1}$	1-10
$G_\circ$	$Habing$	1-10
n_H	cm^{-3}	$10^4 - 10^8$
fr	-	$0 - 100\%$
C_f	-	$0.5 - 3$
ϵ	yield per H_2 formed	$0.01 - 1$
ϕ	yield per cosmic ray impact	$10^2 - 10^6$
y	yield per photon	$10^{-3} - 10^2$
r	-	$0 - 100\%$

Table 3.1: Parameter Definition Domain

The parameters for our chemical modeling code create a nine dimensional parameter space (9D) for molecular clouds as used in our MH and described in Table 3.1:

$$\boldsymbol{\theta} = (n_H, \zeta, G_\circ, C_f, fr, \epsilon, \phi, y, r),$$

In a first attempt to employ a Bayesian approach for deriving branching ratios for poorly understood chemical reaction pathways, we also investigated the parameter r, which controls how much of the gas phase oxygen turns into ice H_2O or ice OH. Parameter r reflects the percentage of O that turns into H_2O, so that $1 - r$ reflects the percentage of oxygen that turns into OH. Desorption efficiencies resulting from H_2 formation on grains, direct cosmic ray heating and cosmic ray induced photodesorption are determined by parameters ϵ, ϕ and y, as introduced and studied by Roberts et al. (2007). The freeze-out parameter in our code is effectively the sticking coefficient, a number in the range of $0 - 100\%$ that adjusts the rate per unit volume at which species deplete on the grain. For the free-collapse to a particular n_H we used the modified formula of Rawlings et al. (1992), where parameter C_f is considered to be a retardation factor with a value less than one, to roughly mimic the magnetic and/or rotational support, or an acceleration factor with a value greater than one to simulate a collapse faster than a free-fall (e.g. due to external pressure). Table 3.1 lists the set of physical parameters studied in this chapter along with their definition domain $\mathbb{D}_{\theta_k}$. The joint definition domain $\mathbb{D}_\theta$ represents the parameter space to explore. The selected domain limits refer to the theoretical range of possible values for molecular clouds where atoms and molecules deplete on to the dust, ensuring though that extreme values are included.

3.2.4 Observational Constraints

The observational constraints of our analysis are based on data from the existing literature. Even though in this application we are primarily interested in ices, we include both gas phase and solid phase observations. To avoid confusion, we will denote with $\boldsymbol{\mathcal{Y}}$ a vector containing any observed quantity and if required we will specify whether we refer to solid phase or gas phase observations.

The solid phase observations include column densities and visual extinction data for molecular clouds in front of field stars. Such sources often provide suitable opportunities to observe and study ices in quiescent regions of the clouds (e.g. Boogert et al. 2011). We used 31 observations of H_2O, CH_3OH, CO and CO_2 from 31 different regions of 16 different clouds found in literature and summarized by Whittet et al. (2011). The data suggest some abundance variation, which was attributed to different evolutionary stages for different clouds. The scope of this chapter lies beyond studying the behavior of a specific cloud, but rather on how to get statistical insight into the dynamics of common cloud classes. Therefore, the observational data is transformed into fractional abundances with respect to total H nuclei and then the average value is computed and used for our analysis.

In an attempt to minimize degeneracies we introduce additional gas phase abundances as an optional observational constraint. Due to the ill-posed nature of our problem, it is possible for our chemical model to end up with a solution space that fits perfectly the solid phase observations, but with gas phase abundances far from realistic. Hence, the addition of gas phase observations can be considered as a mathematical regularization by introducing additional prior information. Prior information can be naturally integrated into our Bayesian approach. The gas species observations were collected from more than one study, attempting to match the clouds, regions or evolutionary stage of the observational sources used for the solid phase species. If we were to fit observations of a particular source, then, ideally, every observational gas phase constraint should be able to contribute to the regularization of our methodology. However, as we are here only attempting at exploring a methodology, we found that three gas phase species were adequate to provide insight on the efficiency of gas phase species as a regularization factor. Abundances for NH_3 and N_2H^+ were collected from Johnstone et al. (2010) while HCO^+ from Schöier et al. (2002). The gas phase observations are in the form of fractional abundances with

Solid Phase Species				Gas Phase Species		
H_2O	CH_3OH	CO	CO_2	NH_2	N_2H^+	HCO^+
7.47 ± 1.81	0.23 ± 0.13	1.14 ± 0.84	1.89 ± 0.79	3.10 ± 2.24	0.068 ± 0.049	0.20 ± 0.01

*The fractional abundances are with respect to H nuclei
**Solid phase abundances are in units of 10^{-5}; Gas phase abundances are in units of 10^{-8}

Table 3.2: Observational Constraints (Average Fractional Abundances)

respect to total hydrogen nuclei. Table 3.2 lists the average molecular abundances for all the species along with their uncertainties. We emphasize again that the error on each of the observations $\mathcal{Y}_n$ is assumed to be normally distributed with a variance σ_n^2 that is determined solely by the uncertainty reported in Table 3.2.

3.2.5 Priors

We run two identical sets of 8 MCMC chains that differ on the prior distribution information. For the first set, the prior information is non-informative and in the form of acceptable range of possible values. Therefore, $\pi(\boldsymbol{\theta})$ is just uniformly distributed on $\mathbb{D}_\theta$, as listed in Table 3.1. Note that the observational data $\boldsymbol{\mathcal{Y}}$ refers only to the solid phase molecular abundances and in this case the gas phase species are ignored. In the second case, the prior information includes the observational constraints from the gas phase species as well. Let $\boldsymbol{\mathcal{Y}}$ now include all the observational constraints, $\boldsymbol{\mathcal{Y}}_s$ just the solid phase and $\boldsymbol{\mathcal{Y}}_g$ the gas phase observational constraints. In that case, the PPD is defined as:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}) = \pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}_s, \boldsymbol{\mathcal{Y}}_g) = \frac{\pi(\boldsymbol{\mathcal{Y}}_s|\boldsymbol{\theta}, \boldsymbol{\mathcal{Y}}_g)\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}_g)}{m(\boldsymbol{\mathcal{Y}})}. \tag{3.3}$$

The prior information is simply the likelihood function $\mathcal{L}(\cdot)$ of $\boldsymbol{\mathcal{Y}}_g$ given a model parametrized by $\boldsymbol{\theta}$, since:

$$\begin{aligned} \pi(\boldsymbol{\mathcal{Y}}_s|\boldsymbol{\theta}, \boldsymbol{\mathcal{Y}}_g) &= \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_s) \\ \pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}_g) &\propto \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_g)\pi(\boldsymbol{\theta}). \end{aligned}$$

Including prior information in this way is equivalent to attaching weight to the solutions that explain the gas phase as well as the solid phase chemistry.

Parameters $\boldsymbol{\theta}$	Unit	Test Value
ζ	$10^{-17} \cdot s^{-1}$	2.4
$G_\circ$	$Habing$	2.6
n_H	cm^{-3}	10^5
fr	-	42%
C_f	-	1.3
ϵ	yield per H_2 formed	0.02
ϕ	yield per cosmic ray impact	150
y	yield per photon	0.1
r	-	75%

Table 3.3: Blind Benchmark Test

3.2.6 Blind Benchmark Test

In order to quantitatively investigate the effectiveness of our method to astrochemical problems we performed a benchmark test. This benchmark test is basically our Bayesian analysis applied this time on synthetic observations produced by UCL_CHEM using a pre-defined set of parameters $\boldsymbol{\theta_T}$. Once we have our synthetic observations, we apply our methodology and analyse the results and whether the true parameters are recovered. Knowing the solution to this test a priori, allows us not only to validate the method, but also to critically perceive the non linear and ill-posed nature of our problem. This discussion can be found in section 3.3.1. The reasoning behind the particular selection of parameters was a random choice not far from expected or well accepted values in the literature. The parameter values used in the test can be found in Table 3.3.

3.3 Results

When quoting parameter estimation results and especially multivariate results, it is convenient to decrease the parameter space to posterior intervals about single marginalized parameters. The MH simulations provide us with the joint parameter PPD. However, because of the high dimensionality of the distribution it is impossible to represent graphically the joint probability density. Therefore, we compute the marginal density for each parameter or for a subset of parameters by integrating the PPD over the rest of the parameters except the ones we are interest in. For example, to obtain the joint marginal distribution of $\boldsymbol{\theta}_a = \{d, fr\}$ we integrate over the rest of the parameters $\boldsymbol{\theta}_b = \{\zeta, rad, bc, \epsilon, \phi, y, r\}$,

$$\pi(\boldsymbol{\theta}_a|\boldsymbol{\mathcal{Y}}) = \int \pi(\boldsymbol{\theta}_a, \boldsymbol{\theta}_b|\boldsymbol{\mathcal{Y}})d\boldsymbol{\theta}_b$$

The marginal probability distributions are visualized either with simple histograms for the case of univariate probabilities or with a bivariate histogram with intensity map for the case of bivariate probabilities.

Traditionally, in order to explore the posterior distribution, typical Bayesian estimates, such as the Posterior Mean are used. However, for multi-modal and/or non Gaussian distributions the extraction of any useful estimator is most of the times meaningless. Instead, it is convenient to decrease the parameter space to High Density Regions (HDR) or credible intervals. HDR computation and graphical representation is explained thoroughly by Hyndman (1996). Following his paper we shortly define HDR as follows: Let $f(x)$ be the density function of a random variable X. Then the $100(1-a)\%$ HDR is the subset $R(f_a)$ of the sample space of X such that

$$R(f_a) = x : f(x) \geq f_a$$

where f_a is the largest constant such that $Pr(X \in R(f_a)) \geq 1 - a$. The above definition indicates two very important properties. From all the possible regions, HDR occupy the smallest possible volume and every point in the regions has probability density that is larger or equal than every point that does not belong in the regions. HDR are very useful for analyzing and characterizing multi-modal distributions. In such cases, HDR might consist of several regions that are disjoint due to the number of modes. In the context of ice formation mechanisms these high density regions are very useful statistical outcomes of the Bayesian approach. Such regions provide us with a precise quantitative measure of how the ice and gas observations and their uncertainties impact the cloud parameters.

Figure 3.1 shows the nine 1D marginalized posterior probability distributions of the parameters for the benchmark test using a uniform prior. In Figure 3.2, we present the nine 1D marginalized posterior probability distributions of the parameters and their 68% High Density Regions (HDR), recovered from the uniform prior case. Figure 3.3, presents the same results for the informative prior case. HDR indicate the parameter space where the probability density is higher. In order to compare the 2 prior cases and quantify the level of constraint for each parameter we introduce a measure of parameter constrain, the

Parameters θ	High Density Spread HDS(%)	
	Non-Informative Prior	Informative Prior
ζ	48	36
$G_\circ$	46	30
n_H	16	09
fr	38	28
C_f	45	35
ϵ	50	42
ϕ	33	28
y	38	33
r	43	32

Table 3.4: High Density Spread. The lower the value of HDS the more constraint is a parameter.

High Density Spread (HDS), which is defined as follows: Let $|HDR|$ be the width of a High Density Region of a parameter's k density function with definition domain $\mathbb{D}_{\theta_k}$ and $|\mathbb{D}_{\theta_k}|$ the width of the domain. Width is defined with respect to some simple measure such as the Lebesque measure (Lebesgue 1902). Then the High Density Spread is defined as :

$$HDS = \frac{|HDR|}{|\mathbb{D}_{\theta_k}|}$$

The HDS ratio can be perceived as an index of the level of uncertainty on a predefined definition domain and the higher it is the less constrained is a parameter. Table 3.4 presents HDS for each parameter for both priors used. Figure 3.4 shows the 2 dimensional marginal PPD for parameters that present statistical interest. Finally, Table 3.5 lists the statistical mean and standard deviation for the $\sim 35\%$ HDR of the joint distribution for all the 9 parameter. The general statistical picture we get from Figures 3.2 and 3.3 shows that the distributions of all the parameters are far from Gaussian and most of them have more than one modes. Looking at the models with physical units, we can also notice that most of the density lies away from the limits of our definition domain for both cases, which validates our choice for $\mathbb{D}_\theta$.

3.3.1 Blind Benchmark Test Results

The results of the performed test as shown in Figure 3.1 reveal two important insights. First of all, high probability density regions for all the parameters include and hence recover the true parameters. As we can see in Figure 3.1, all the pre-defined parameter values lie under or very close to the highest density point of the marginal PPD. This result simply validates that both the Bayesian approach makes accurate inference based on the given observations and the MH algorithm samples efficiently the solution space. Secondly, we can observe that in many cases there are additional high probability density regions. These regions prove and highlight the ill-posed nature of our problem by indicating that different parameter sets can produce similar observations. Combining the two insights, we can conclude that the Bayesian method with MCMC sampling is exploring efficiently the parameter space, revealing the solution regions that answer our ill-posed inverse problem. In addition, we can conclude that in order to constrain our solution space we should either introduce numerical regularization factors (e.g. gas phase species) or scientific prior knowledge.

3.3.2 Influence of priors

A visual comparison of Figures 3.2 and 3.3 reveals what we can quantitatively observe in Table 3.4. With non-informative uniform prior the high density regions seem to cover large sections of the distribution, which in some cases reach 50% of the definition domain. This means that most of the parameters are not constrained enough. The most statistically straightforward parameters seem to be clearly the n_H and then the ϕ and fr parameters, presenting distinct modes and relatively low HDS. $G_\circ$ and ζ seem neither constrained nor relevant enough, while fr seems to have a clear mode, followed by a very heavy tail. The rest of the parameters present high HDS, above 40% with several disjoint high density regions and do not allow us to reach credible conclusions about the parameters. Including the prior information from the gas phase species changes the picture significantly as can be seen in both Figure 3.3 and Table 3.4. We can observe that the HDR get smaller and the parameters seem more constrained. The distribution of ζ is now denser around high values (> 6), while $G_\circ$ has to be low (< 4). The n_H remains well constrained with even lower HDS, while the distribution of fr now clearly constraints the parameter to low domain values. The distribution of C_f is also altered significantly: not only the HDS has dropped, but also a large portion of the density has transfered from high accelerated collapse regions to free fall collapse regions. The non-desorption mechanisms still present a multi-modal behavior, but with significantly smaller high density regions. Their distribution clearly highlights the non-linear way these mechanism act together or against each other. For r, the addition of informative prior information seems to reduce the HDS as well, centralizing the density, but still favoring slightly the production of H_2O against OH. Therefore, we conclude that the addition of gas phase species as a regularization factor outperforms the use of just a non-informative uniform prior distribution. The HDS is reduced at an average of $\sim 12\%$, which indicates an equivalent constraint on the parameter space. Section 3.3.3 will discuss the statistical and numerical results of our analysis, while Section 3.3.4 will discuss the astrophysical implications. For both these Sections we shall only concentrate on the results of the informative prior case.

3.3.3 High Density Regions

Before presenting the resulting HDR of our analysis, it is useful to explain what is essentially the meaning of such a region. Our results come in the form of probability distributions. That means that for each parameter, the marginalized posterior probability distributions links every value of the parameter definition domain with its probability of being the value that generated our observational constraints. Now, a HDR indicates a range of values that are significantly more likely than the rest. In practice, this means that if we were to select a value that best describes physically our expectation of a molecular cloud, the value would come from this range.

The n_H is clearly the most constrained physical parameter. The marginal density function reveals that most of the density is between 2.2 and $5 \times 10^4\, cm^{-3}$. The ζ is constrained to values higher than $6 \times 10^{17}\, s^{-1}$, while the $G_\circ$ to values lower than $4\, Habing$. The HDR for the fr stays between 20% and 45%, while the C_f has 1 distinct HDR between 0.5 and 1.55 and one long heavy tail between 2 and 3 times the default free fall rate. The ϵ presents two modes. The first HDR is between 0.4 and 0.8 and the second between 1.2 and 1.4. The marginal distribution of ϕ, also presents two modes. One is centered around 10^5. The second one is centered around 60. The marginal distribution for y, presents 2 disjoint high density regions as well. The first one indicates really low efficiency of about 10^{-6}, while the second one a slightly higher $2 \times 10^{-3} - 8 \times 10^{-2}$. Finally, the distribution for the branching ratio parameter r shows high density between 40% and 70% of oxygen turning into ice water.

In Figure 3.4 we show the marginalized 2D PPD for our parameters. Note that the n_H and the fr are negatively dependent in a nearly linear way. On the other hand $G_\circ$ and n_H seem to have a non-linear positive correlation, hitting a plateau after a certain gas density. Similarly, the fr and the C_f may have a clear peak, but also some evidence of a positive

Parameters $\boldsymbol{\theta}$	Unit	Mean Value
ζ	$10^{-17} \cdot s^{-1}$	$8.39(\pm 2.8)$
$G_\circ$	$Habing$	$1.79(\pm 1.27)$
n_H	cm^{-3}	$4.07(\pm 2.34) \times 10^4$
fr	-	$31(\pm 21)\%$
C_f	-	$1.18(\pm 0.9)$
ϵ	yield per H_2 formed	$0.52(\pm 0.35)$
ϕ	yield per cosmic ray impact	$2.78(\pm 1.12) \times 10^6$
y	yield per photon	$3.35(\pm 2.27) \times 10^{-3}$
r	-	$56(\pm 23)\%$

Table 3.5: Mean and standard deviation for the most probable mode of the Joint Distribution. The mode corresponds to ∼ 35% HDR of the total joint distribution.

correlation. The relation between the cosmic ray desorption efficiency parameters, ϕ and y reveals many distinct peaks throughout the domain space. Note that the marginalized PPD for cosmic ray ionization rate and parameter ϕ shows a clear bimodal structure. However, focusing only on the denser areas of the distribution we can observe a potential non linear correlation between the cosmic rays and the efficiency of the cosmic ray related parameter ϕ. In general though, ζ is evidently a parameter that is not sufficiently constrained. This is already obvious by the 1D marginal distribution of ζ, but the contrast of constrain between ζ and one of the most constrained parameters such as n_H is depicted in Figure 3.2(6).

Due to the non-uniqueness of our solution space, examining the joint probability distribution of the PPD provides a useful insight. The dimensionality of the distribution makes a visualization impossible, so we chose to extract the statistical mean and the standard deviation for each one of the parameters from the most probable mode of the joint distribution. The joint distribution was approximated using a multivariate histogram and the most probable mode was chosen in a heuristic way and corresponds to ∼ 35% HDR of the whole PPD. The values for the mean and standard deviation are given in Table 3.5. As expected, the most probable mode of the joint PPD agrees with the HDR of the marginal parameter distributions. For the unimodal 1D marginalized distributions the most probable mode coincides completely, while for the multi-modal cases the most probable mode coincides with one of the modes. Hence, purely based on the statistical interpretation we conclude that: a molecular cloud that matches the observed abundances should have low n_H, a low fr and a low $G_\circ$. The ζ on the other hand is more likely to have high values, but the high standard deviation leaves room for significant variation. The collapse of the cloud may be insignificantly accelerated, while the branching ratio r favors slightly the branching into water, but with a high standard deviation. In terms of the non-thermal desorption efficiency parameters, we notice increased efficiency for all three of them. As a general result we conclude that the 9D space of the joint distribution has multiple peaks. Both the marginalized distributions and the denser peak of the joint distribution indicate that some of the parameters (n_H, $G_\circ, fr, C_f$) are well constrained, while other parameters ($\zeta, r, \epsilon, \phi, y$) present possible variation that implies further astrophysical or statistical implications.

3.3.4 Astrophysical Consequences

Here, we discuss our results for each of the parameters with regards to their astrophysical implication:

$\boldsymbol{n_H}$: The derived credible intervals for the gas density are in very good agreement with the properties of typical collapsing dark clouds, clumps and cores (Myers & Benson 1983; Benson & Myers 1989; Bacmann et al. 2002; Bergin & Tafalla 2007). Higher cloud densities ($> 10^6\, cm^{-3}$), that are usually expected in hot cores after the cloud has collapsed (Hoare 2004), were explored, but showed nearly zero probability density in our analysis.

$\boldsymbol{fr}$: Our study implies a depletion rate that is not high enough to dominate and is probably lower than 50%. Bacmann et al. (2002) suggest that freeze-out dominates when n_H exceeds $\sim 3 \times 10^4\, cm^{-3}$ which is marginally the case in our study. When the freeze-out dominates and densities exceed $\sim 10^5\, cm^3$, the abundance of CO ice is found to be significantly increased to typical gaseous values ($\sim 10^{-4}$) (Pontoppidan 2006; Bergin & Tafalla 2007). Furthermore, the ice water abundance is typically 5×10^{-5} to 9×10^{-5} and even higher at the highest densities (Pontoppidan et al. 2005). The ice CO and H_2O abundances in our case though, are about $0.5 - 1$ magnitude lower. Therefore, along with the n_H results, the lower freeze-out rate estimated by our analysis can be explained by a different evolutionary stage of the observed clouds. According to Fontani et al. (2012) low depletion values can also imply a cloud that is going to form less massive objects.

$\boldsymbol{G_\circ}$: Our analysis showed that in order to match the observed ice abundances the $G_\circ$ is comparable to the standard interstellar radiation field of 1 *Draine* or $\sim 1.7\, Habing$ (Draine 1978).

$\boldsymbol{\zeta}$: In dense gas ζ is measured to be in the range of 1 to $5 \times 10^{-17}\, s^{-1}$ (Bergin et al. 1999). However, considerable uncertainties have been reported in the literature with derived values as high as $10^{-15}\,s^{-1}$, accounted to x-rays from a central source (Doty et al. 2004). These discrepancies may be due to whether ζ is determined via H_3^+ or HCO^+ measurements. Yet, Dalgarno (2006) claims that given latest evidence that the ζ range is narrow and between 10^{-16} and 10^{-15}, the question should be focused not on why the estimations are different, but on why they are so similar. Our analysis confirms ζ values higher than the typical estimations and is even consistent with the $10^{-16}\,s^{-1}$ estimations through the H_3^+ determination. Most importantly, our study indicates high standard deviation on these values highlighting that such a variation should be expected. Theoretically, this is explained considering the fact that ζ lose energy while ionizing and exciting the gas through which they travel in conjunction with the possible variation in the origin of ζ. Even though our astrochemical model does not account for the latter factors, our probabilistic approach reflects their impact.

$\boldsymbol{C_f}$: Our study shows that the collapse of the cloud should follow the expected free fall collapse. Higher C_f values present moderate probability density, which implies that the observational constraints could potentially also be matched with different but also less likely sets of parameters (e.g. higher values for both C_f and fr).

$\boldsymbol{\epsilon, \phi, y}$: The desorption from H_2 efficiency parameter (ϵ) estimates are significantly higher than the value reported by Roberts et al. (2007) ($\epsilon < 0.1$). The direct cosmic ray desorption efficiency (ϕ), presents two peak values. One of them agrees with Roberts et al. (2007) and is centered around 10^5. The second one is centered around 60, which is lower than the lowest limit studied by Roberts et al. (2007). For the cosmic ray-induced photodesorption efficiency (y), we have two probable estimates as well. The first one indicates really low efficiency. The second one presents a slightly higher efficiency that is still lower than the one estimated by Hartquist & Williams (1990) ($y = 0.1$), but consistent with the results of Öberg et al. (2009) for CO_2. Our analysis in general indicates useful credible intervals for non-thermal desorption efficiencies, highlighting though, that the reported non linearities can be tackled with further regularization factors such as molecule specific analysis and additional grain properties. Note as well that our astrochemical model does non include direct UV photodesorption which has recently be found to be efficient (Zhen & Linnartz 2014).

$\boldsymbol{r}$: The branching ratio proved to be a parameter with high but anticipated variability. Its marginal probability distribution presents the most statistically normal behavior with a mean that implies a shared branching ratio of oxygen freezing into ice H_2O and ice

OH, favoring slightly solid H_2O. The first laboratory experiment to reproduce the ice H_2O formation (Dulieu et al. 2010) implied that the hydrogenation of oxygen is an important route for water formation. Furthermore, Cazaux et al. (2010) state that species such as OH are only transitory and quickly turn into ice water. However, they also state that $\sim 30\%$ of the O coming on the grain is released in the gas phase as OH which can freeze back as H_2O. A pathway that is included in our model and can explain both the high water abundance on the grains and the shared branching ratio r. At last the high branching ratio towards ice OH highlights the importance of ice OH for the production of ice CO_2.

We now look at the correlation between our parameters as presented in Figure 3.4. The relation between the n_H and depletion or freeze-out has been the subject of many studies (Bacmann et al. 2002; Christie et al. 2012; Fontani et al. 2012; Hocuk et al. 2014). They all conclude that the amount of depletion, the ice abundances and the density of the cloud should all scale together, as shown by theoretical studies (Rawlings et al. 1992). Even though our analysis suggests a clear anti-correlation between n_H and fr (Figure 3.4(a)), this result is completely in line with literature, since we are not analyzing the time evolution of the cloud, but instead focus on parameter fitting at specific time points. This negative correlation suggests that the less gas density we have the higher the depletion should be in order to match the observed ice abundances. Our results are also in line with the negative correlation between depletion factor and n_H, derived by Fontani et al. (2012) from CO observations. The positive correlation between n_H and $G_\circ$ depicted in Figure 3.4(b) is confirming that the denser a cloud, the higher $G_\circ$ values are needed to match the observations. The plateau after a density value ($\sim 7 \times 10^4 \, cm^{-3}$) indicates that the explored radiation field domain space is not high enough to penetrate the cloud after a density threshold. When C_f is increased the freeze-out timescale needs to be decreased since the final n_H is reached quicker. This reduced timescale requires higher fr values in order to simulate the observed ice abundances and this relation is depicted in Figure 3.4(c). Even though not very straightforward, the relation between the cosmic ray desorption efficiency parameters, ϕ and y, is very interesting (Figure 3.4(d)). In most cases, the cosmic ray photodesorption efficiency is either low or either high for both direct cosmic ray heating and cosmic ray induced cases. However, there is a significant peak when the direct cosmic ray impact is very efficient, whilst the cosmic ray induced impact is very inefficient.

3.4 Conclusions

In this chapter, we implemented a Bayesian MH parameter estimation analysis to solve a typical ill-posed inverse astrochemical problem. We have employed a chemical modeling code and solid phase observations in order to get a holistic insight into the behavior of physical and chemical parameters that drive ice chemistry in dark molecular clouds. The main conclusions of this work are as follows.

1. The Bayesian method provides a systematic approach to solve nonlinear inverse problems with high noise levels and significant model uncertainties. The MCMC technique allows to sample from complex probability distributions in an efficient way. As highlighted by our Blind Benchmark Test, we can conclude that the latter methods succesfully handle astrochemical ill-posed problems and reveal a more complete set of solution regions. On the contrary, single solution estimates derived from traditional approaches would not have provided a complete picture of the solution space and would have contained a high risk of degeneracy.

2. Our probabilistic approach to physical and chemical parameter estimation used a chemical network with deficiencies (especially for the grain part) and several assumptions. Nevertheless, the results both derived useful credible intervals and highlighted model deficiencies implying even more promising results for tackling physical, chemical and model uncertainties for up to date models with targeted astrophysical goals.

3. We confirm that the joint PPD of the solution space is highly non linear and multimodal and the 1D marginal PPD for each parameter are far from Gaussian highlighting the complexity of the problem.

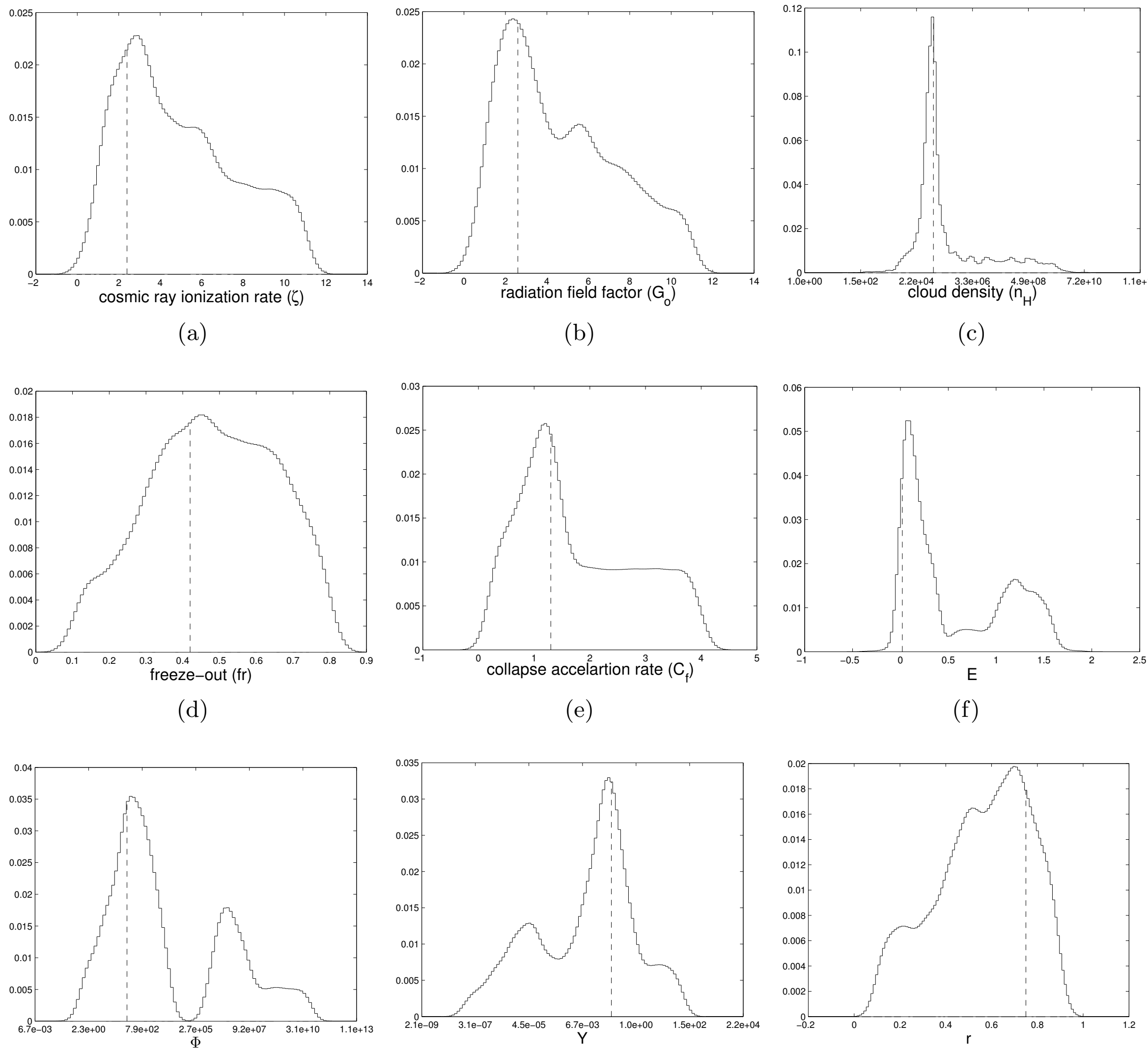
4. Including abundances of gas phase species as a regularization factor and introduced as a Bayesian prior, increases the parameter constrain efficiency by 12%. This result can imply that observational regularization constraints compensate for any chemical code deficiencies. Also, increasing the number of gas phase regularization factors will constrain even more the solution space.

5. We show that physical parameters such as n_H, $G_\circ$, C_f are highly constrained and their variation has a great impact on the derived ice abundances.

6. The high variation of ζ contradicts the theoretical ζ standard values in dense gas and indicates a larger credible interval instead.

7. Non-thermal desorption efficiencies act and counteract in a non-linear way with each other or ζ. This complex behavior should be analyzed with extra regularization factors.

8. Branching ratio parameters such as r can be successfully estimated through Bayesian MCMC methods. Our results even though with high variability, indicate that the detail or simplicity of the dust grains chemical network can be encapsulated and reflected as certainty or uncertainty respectively.

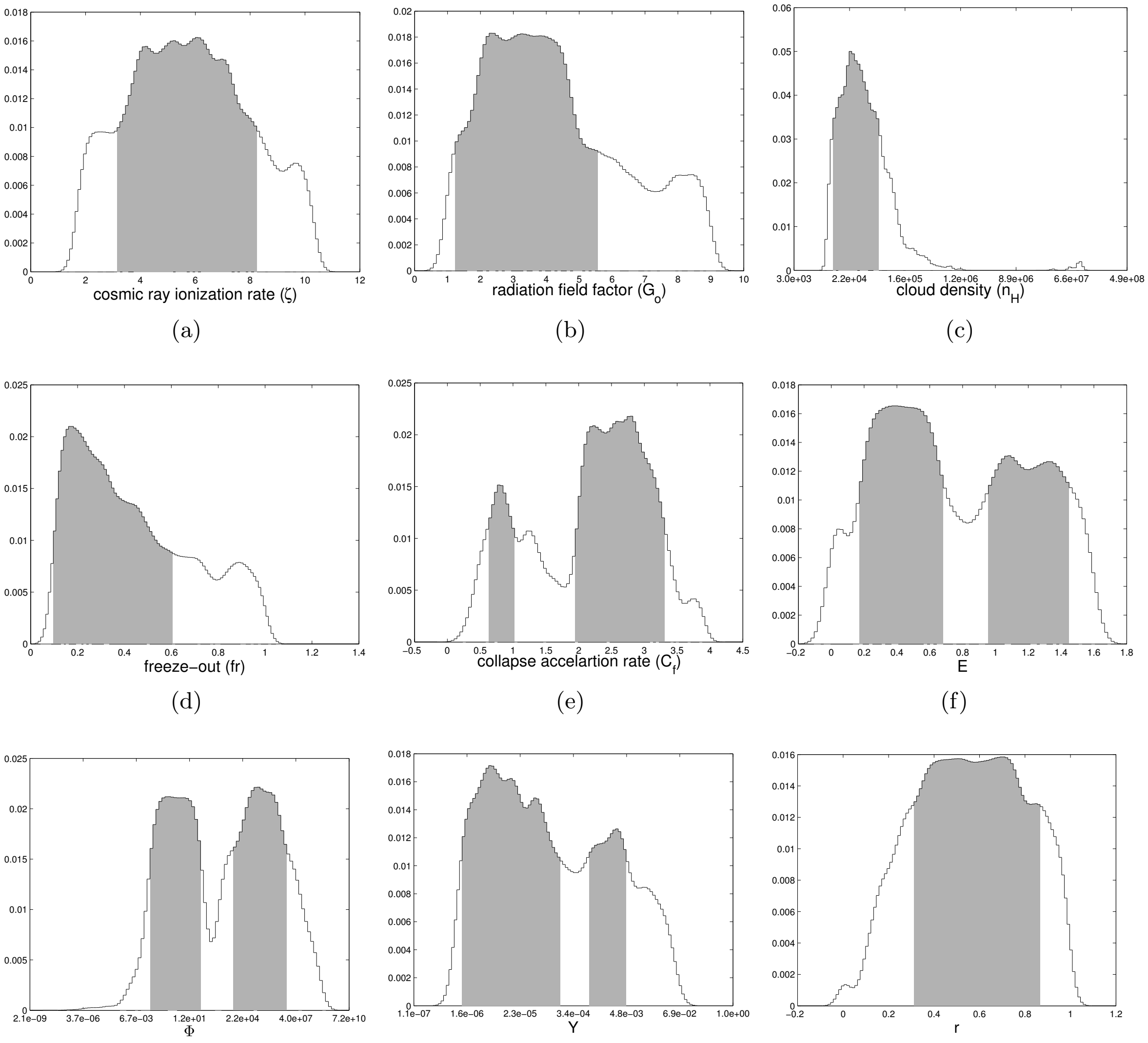


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3.1: 1D Marginalized PPD for each of the nine parameters for the Blind Benchmark Test. The plots show the Gaussian kernel density estimator of each Probability Density Function. Dashed lines indicate the pre-defined parameter values $\boldsymbol{\theta_T}$ we wish to recover.

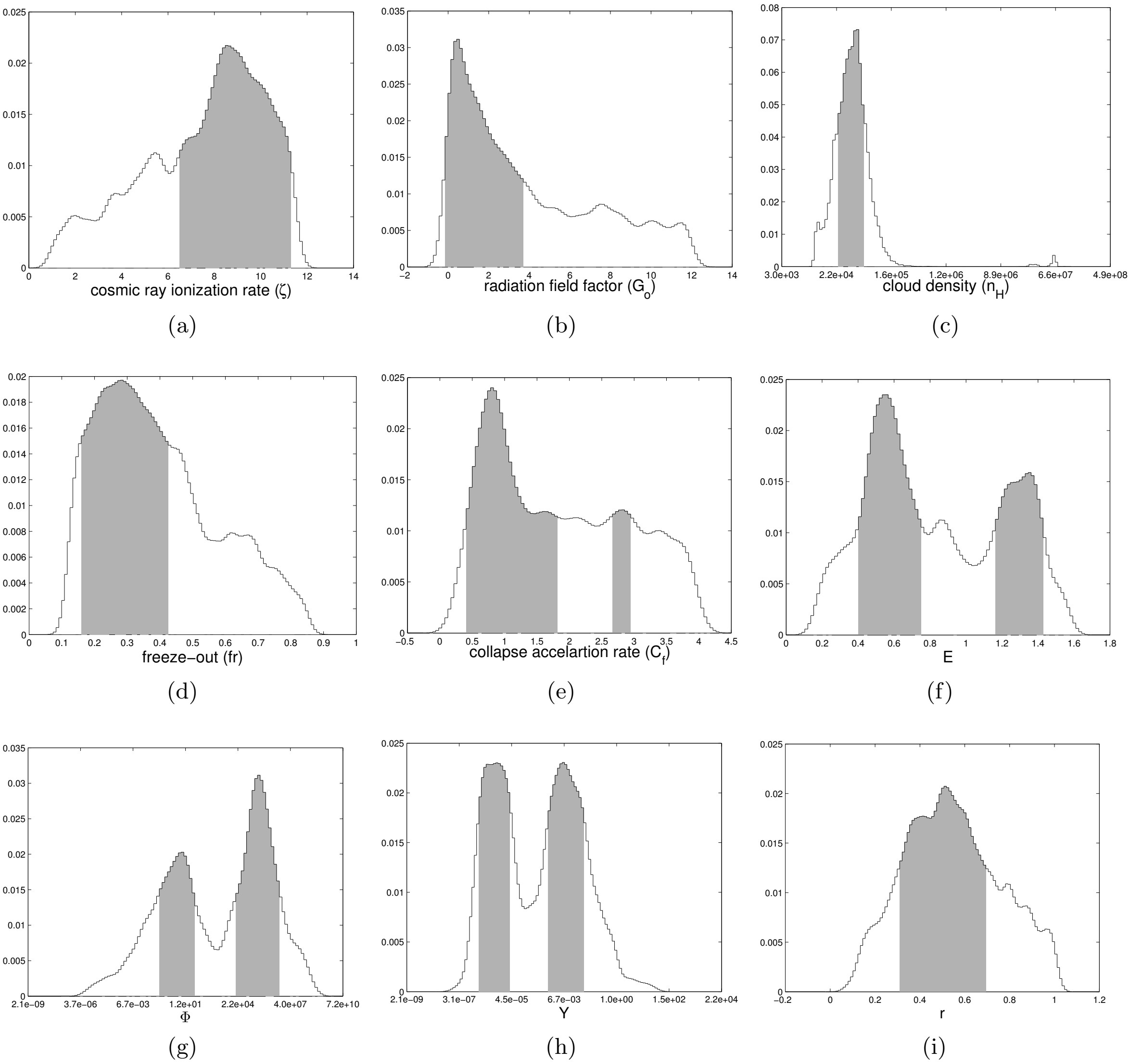


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3.2: 1D Marginalized PPD for each of the nine parameters using uniform non informative prior. The plots show the Gaussian kernel density estimator of each Probability Density Function.Darker regions indicate 68% HDR.

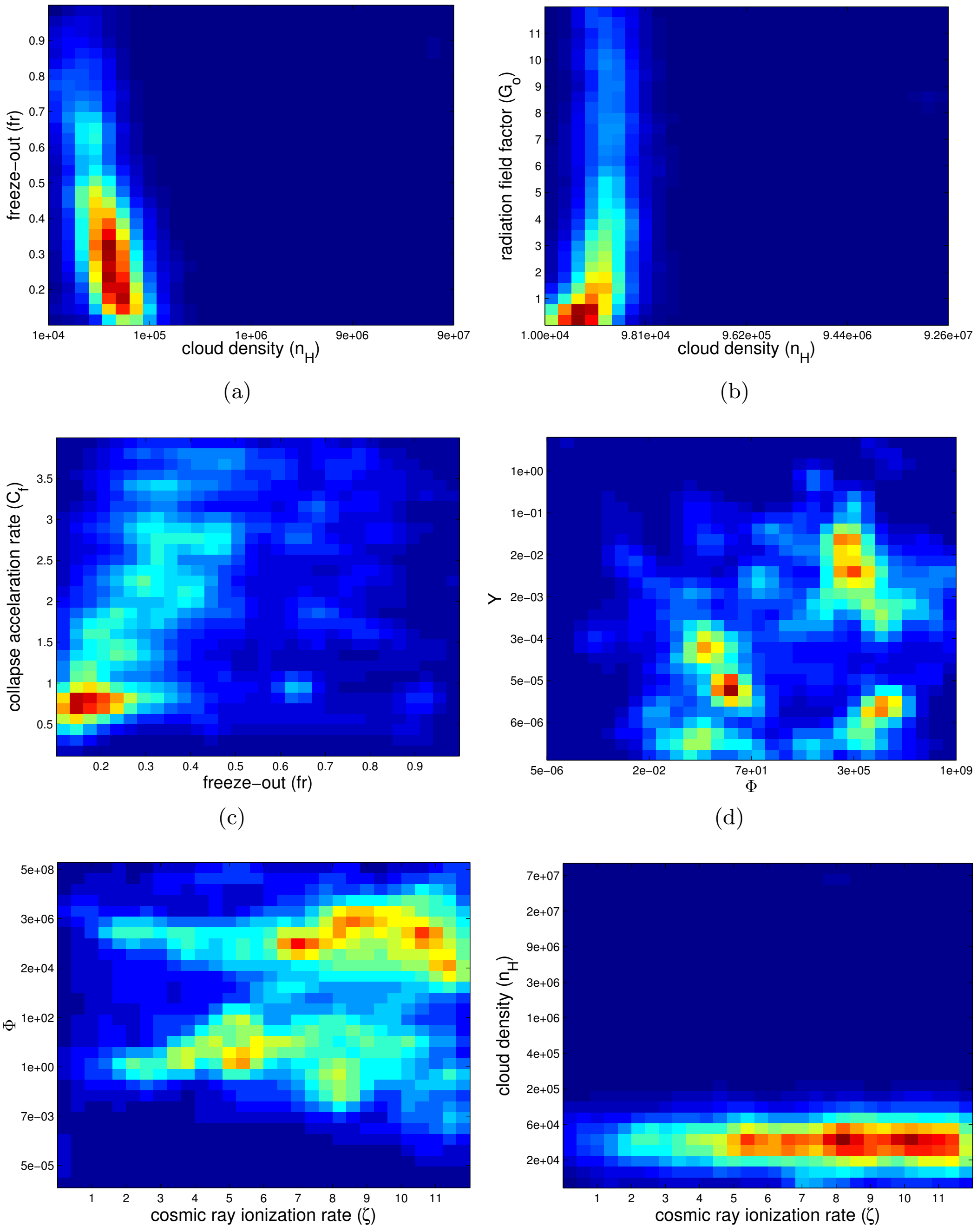


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3.3: 1D Marginalized PPD for each of the nine parameters using informative prior from gas phase chemistry. The plots show the Gaussian kernel density estimator of each Probability Density Function. Darker regions indicate 68% HDR.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.4: 2D marginalized posterior probability density functions.Warmer colors indicate higher probability density.

Chapter 4

Fast Astrochemical Parameter Estimation with Neural Networks

The present chapter extends our work on Bayesian inference methods for astrochemical inverse problems by introducing a machine learning solution to the speed problem of the inference process. With the size of parameter spaces commonly encountered in astronomy, most researchers have to wait for hours or even days for a result. As introduced in section 1.5.5, Artificial Neural Networks (ANNs) are machine learning computational models that can learn and substitute computationally expensive functions and speed up significantly the whole Bayesian inference process. Section 4.1 introduces the problem and how ANNs can contribute towards a solution to astrochemical problems. Multilayer Perceptron is the type of ANNs employed in this chapter and is presented in Section 4.2. The training process of our network is described in Section 4.3 and both an algorithm to speed up Bayesian inference and a simple application example are discussed in Section 4.4. Finally in Section 4.5, we present our conclusions.

4.1 Introduction

Bayesian inference methods are consistently becoming a common practice for constraining parameters in astronomy, cosmology and other fields of astrophysics (Christensen &

Meyer 2000; Ford 2005; Fitzgerald et al. 2007; Feroz & Hobson 2008; Isella et al. 2009; AMI Consortium et al. 2012; Bazot et al. 2012; Makrymallis & Viti 2014). The parameter exploration and estimation is performed with a Monte Carlo sampling algorithm, which usually is a Markov Chain Monte Carlo (MCMC) variant or a nested sampling algorithm. In Chapter 3 and in Makrymallis & Viti (2014), we showed that Bayesian inference techniques based on sampling algorithms can also be successfully applied in astrochemical problems for the estimation of physical and chemical parameters of molecular clouds. The probability distribution of molecular cloud physical properties, as well as chemical reaction coefficients can be accurately approximated, even when the inverse problem is ill-posed or the posterior distribution is multi-modal with inherent degeneracies. Despite the great benefits, this method can present some drawbacks usually related with the computational cost and the speed of the process.

Bayesian inference methods require the evaluation of a likelihood function for each sample point of the explored parameter space. The likelihood function represents the probability of reproducing the observed data for a given set of parameters and in most cases requires cumbersome runs of complex chemical codes. A standard chemical code is usually a time and depth dependent gas-grain chemical model that can be used to estimate the abundances of gas and surface species in every environment where molecules are present. The model can include both gas and surface reactions and determines molecular abundances in environments where not only the chemistry changes with time, but also local variations in physical conditions lead to variations in chemistry. One model run can take from a few seconds up to a few minutes, depending on the complexity of the model and the number of reactions. Monte Carlo sampling algorithms can reduce the number of likelihood evaluations, but only up to a point. Furthermore, when the target distribution is complex and multi-modal, which is usually the case, most sampling algorithms would require more time to adequately explore all the modes. That simply means even more likelihood computations. If we also consider the fact that the time and computational cost required to explore the parameter space increases exponentially with the number of parameters we wish to estimate, it is easy to conclude that great gains can be achieved if we are able to speed up the evaluation of the likelihood function.

In this chapter, we present a real time accelerated astrochemical parameter estimation algorithm based on ANNs. The algorithm follows the successful use of machine learning techniques and specifically ANN for similar tasks in other fields of astrophysics (Auld et al.

2007; Graff et al. 2012) and evaluates the suitability and efficiency of ANN for learning and replacing the likelihood function, speeding up parameter estimation tasks in Astrochemical inverse problems. For the rest of the chapter, we will assume that the problem to tackle is to constrain physical and chemical parameters of dark molecular clouds using Bayesian inference and a Metropolis-Hastings (MH) algorithm as our MCMC method.

4.2 Multilayer Perceptron

In this chapter we will only consider one class of ANNs, the multilayer perceptron (MLP) with one hidden layer as shown in Figure 4.1. The choice of a MLP with one hidden layer was made as the simplest, yet adequately efficient type of ANNs to prove the concept of using ANNs for accelerating astrochemical parameter estimation. The efficiency of a MLP with one hidden layer is reassured by the universal approximation theorem (see Section 1.5.5). Multilayer Perceptron Neural Networks are feed-forward directed networks composed of multiple layers. Each layer consists of perceptron nodes and is fully connected to the next layer. MLP maps input data $\mathbf{x} \in \Re^n$ onto scalar output $y_i(\mathbf{x}; \mathbf{w}, q)$ through linear or non-linear function nodes. The number of the input nodes correspond to the number of physical and chemical parameters we want to constrain, while the output in our case will be just one node corresponding to the likelihood that these parameters describe a system that can reproduce the observed molecular abundances. The number of hidden nodes is a user defined parameter that adjusts the complexity of the network. The outputs of the nodes in the hidden and output layers are as follows:

$$\text{hidden layer: } h_j = g^{(1)}(f_j^{(1)});\ f_j^{(1)} = q_j^{(1)} + \sum_l w_{jl}^{(1)} x_l, \tag{4.1}$$

$$\text{output layer: } y_i = g^{(2)}(f_i^{(2)});\ f_i^{(2)} = q_i^{(2)} + \sum_j w_{ij}^{(2)} h_j, \tag{4.2}$$

where $\boldsymbol{w}$ is the ‘weight’ parameter and q the ’bias’ parameter of the perceptron. Index l runs over input nodes, j runs over hidden nodes, and i runs over output nodes that in our case is just one. An example of a simple MLP and its corresponding outputs is shown in Figure 4.1.

The weights and biases are the values we wish to determine in our network training session. MLP learns a non-linear relationship between input and output nodes by adjusting

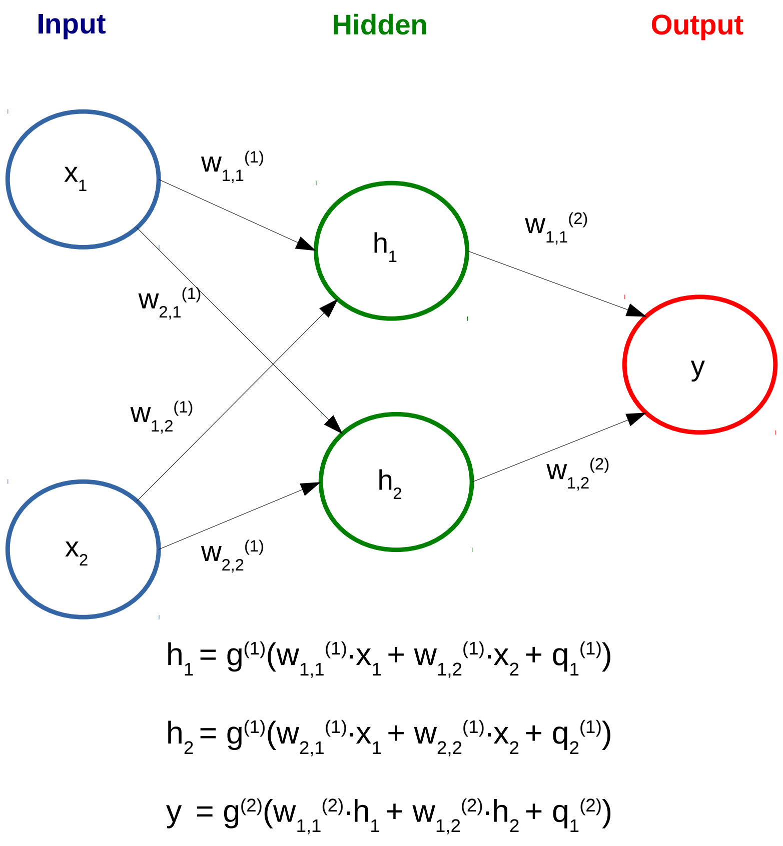

Input
Hidden
Output
x_1
x_2
h_1
h_2
y
$w_{1,1}^{(1)}$
$w_{2,1}^{(1)}$
$w_{1,2}^{(1)}$
$w_{2,2}^{(1)}$
$w_{1,1}^{(2)}$
$w_{1,2}^{(2)}$
$h_1 = g^{(1)}(w_{1,1}^{(1)} \cdot x_1 + w_{1,2}^{(1)} \cdot x_2 + q_1^{(1)})$
$h_2 = g^{(1)}(w_{2,1}^{(1)} \cdot x_1 + w_{2,2}^{(1)} \cdot x_2 + q_2^{(1)})$
$y = g^{(2)}(w_{1,1}^{(2)} \cdot h_1 + w_{1,2}^{(2)} \cdot h_2 + q_1^{(2)})$

Figure 4.1: A MLP with one hidden layer, 2 inputs, 2 hidden nodes, and 1 output, along with the the outputs of the nodes in the hidden and output layers. For brevity, the bias nodes are ommitted.

the weighted connections and the bias given a set of training data and then can make predictions of the output for new input data. The number of hidden nodes is a parameter that has a crucial effect on the performance of the MLP. As a rule of thumb, the number of training points should function as an upper limit for the number of hidden nodes. However, the more hidden nodes we use, the better accuracy we will achieve on our training data. This accuracy though is not representative of how well the NN generalizes to situations not presented during training. This issue is known as overfitting and in Section 4.3 we will discuss ways to deal with it. The activation functions $g^{(1)}$ and $g^{(2)}$ are both selected in accordance to the universal approximation theorem to be bounded, smooth, monotonic and one of them non-linear, allowing the network to model non-linear functions. We chose $g^{(1)}(x) = \tanh(x)$ and $g^{(2)}(x) = x$.

4.3 Network Training

The training data set $\mathcal{D} = \{\mathbf{x}^{(k)}, y^{(k)}\}$, consists simply of the parameter points explored by the MH algorithm and the corresponding evaluated likelihood values respectively. The objective of the training session is to both tune appropriately the values of the weight and bias parameters and optimize the number of hidden nodes, so that maximum ANN performance is achieved, avoiding overfitting. The ANN performance is defined as the mean squared error between the network output and the real likelihood value for each particular set of parameters. The training method used in our case is the backward propagation of errors, known as backpropagation (Rumelhart et al. 1988), and as an optimization method the Levenberg–Marquardt algorithm (Marquardt 1963). In reality, any standard numerical optimization algorithm can be used instead. Backpropagation performs computations backward through the network and computes the gradient of the mean squared error with respect to all the weights and biases in the network. The Levenberg–Marquardt algorithm uses this gradient to update the weights and biases, minimizing gradually the mean squared error.

For the training session, $\mathcal{D}$ is randomly split into two subsets. The first of them is the actual training subset and is used to tune the weight and bias values. The second is used as a validation subset to avoid overfitting the training data. The training error and the validation error are monitored together and both are expected to decrease while the training progresses. However, if and when the ANN starts to overfit the data, the training error will keep decreasing, while the validation error will start increasing. The weight and bias values are chosen to minimize the validation error. The validation subset is also used to compare different ANN models after the training is over. In our case the different models are networks with different number of hidden nodes. For a given training data set $\mathcal{D}$ and N candidate models, the validation error reflects the accuracy of the respective network. If the validation error is less than a user defined threshold, it can be used to decide between the best network structure. The default ratios for training and validation are 0.7 and 0.3 respectively.

4.4 ANN Accelerated Bayesian Inference

Our algorithm combines Bayesian statistical methods and Monte Carlo sampling techniques with ANNs in order to solve astrochemical non-linear inverse problems faster. We adopt the exact problem and notation of Makrymallis & Viti (2014) and Chapter 3. We refer readers back to Chapter 3 for a reminder on the notation and specifics of the Bayesian inverse problem and we simply revisit here the main ingredients of the solution. We wish to constrain a set of physical parameters $\boldsymbol{\theta}$, given molecular observations $\boldsymbol{\mathcal{Y}}$ for mantle species H_2O, CH_3OH, CO and CO_2 and gas phase species NH_3, N_2H^+ and HCO^+. Parameter estimation can be performed through the posterior probability distribution of the parameters, given the observational data:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{Y}}) = \frac{\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\mathcal{Y}})\pi(\boldsymbol{\theta})}{m(\boldsymbol{\mathcal{Y}})} \tag{4.3}$$

The likelihood function $\mathcal{L}$ of observations, given a model parametrized by a set of parameters is defined as:

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\mathcal{Y}}) = p_\varepsilon(\boldsymbol{\mathcal{Y}} - \mathcal{C}(\boldsymbol{\theta})) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_n^2} \times \exp(-\frac{1}{2}\sum_{n=1}^{N}[\frac{\mathcal{C}_n(\boldsymbol{\theta}) - \mathcal{Y}_n}{\sigma_n}]).$$

The likelihood function $\mathcal{L}$ involves runs of the chemical code UCL_CHEM, denoted here as $\mathcal{C}(\cdot)$. UCL_CHEM though, slows down the likelihood evaluation significantly, therefore we wish to train our ANN to learn and replace the likelihood function.

The algorithm used to perform the parameter estimation is shown in Algorithm 4.1. We assume that we have already set up a Bayesian inference approach, based on a MH algorithm and we need to accelerate the inference process. Initially the user defines N ANN models with different number of hidden nodes, a maximum accepted validation error V_{err} and a number M which represents the minimum number of samples required before the ANN training is initiated. The V_{err} is a user defined threshold that represents our belief that a ANN has learned the likelihood function satisfactory. The MH algorithm proceeds as usual. Sets of parameters are generated, and for each set of parameters the likelihood function is evaluated as well as the posterior probability of the specific set. Every time a set of parameters is accepted by the MH, the training and validation data set is incremented by one data point. Each data point consist of one input, the vector of parameters, and one output, the value of the likelihood function computed for these parameters. When the size of our data set reaches M, all N ANNs are trained and then validated. Then, the minimum validation error is compared to V_{err}. If it is smaller than V_{err}, then the ANN with the minimum validation error replaces the likelihood function

Algorithm 4.1 Accelerated Bayesian Inference

```
input: MH algorithm, N $ANN^{(j)}$ models with $j = \{1, ..., N\}$,
maximum accepted validation error $V_{err}$,
number of minimum samples required before ANN training M
initialize: random set of parameters $\boldsymbol{\theta}^{(1)}$, empty training/validation ANN dataset $D = \{\}$
Compute $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}})$
Add $\{\boldsymbol{\theta}^{(1)}, \mathcal{L}\}$ to $D$
while $size(D) < M$ do
    $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^i | \boldsymbol{\theta}^{i-1})$
    MH accepts/rejects $\boldsymbol{\theta}^*$
    if $\boldsymbol{\theta}^*$ is accepted by MH
        Compute $\mathcal{L}(\boldsymbol{\theta}^*; \boldsymbol{\mathcal{Y}})$
        Add $\{\boldsymbol{\theta}^*, \mathcal{L}\}$ to $D$
    end if
end while
Train the N ANNs and compute validation error $Err^{(j)}$ for $j = \{1, ..., N\}$
if $argmin\{Err\} = Err^{(j)}$ and $Err^{(j)} < V_{err}$
    Continue with MH until convergence with $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}) = ANN^{(j)}(\boldsymbol{\theta})$
else Repeat while clause with $M = M + \frac{1}{2}M$
output: MH output
```

and the MH algorithm proceeds until convergence. If it is not smaller than V_{err}, then we need more data points to train our ANNs better. The MH proceeds normally with likelihood function evaluations until the data set reaches $M + \frac{M}{2}$. This is repeated until the minimum validation error gets smaller than V_{err}.

To evaluate and demonstrate the efficiency of ANN to learn complex likelihood functions, such as complex chemical code functions, our algorithm was benchmarked and tested against a simplified and controlled version of the inverse problem studied in Chapter 3. We assume a smaller set of parameters $\boldsymbol{\theta} = \{n_H, \zeta, G_\circ, fr\}$ and we create a controlled test environment such as the blind benchmark test of Chapter 3. The rest of the parameters studied in the previous chapter were kept fixed at the most probable values, as found by our analysis in Chapter 3. We selected a set of parameters $\boldsymbol{\theta} = \{10^5\, cm^{-3}, 2.2\, s^{-1}, 1.2\, Habing, 0.5\}$ and we used UCL_CHEM to produce $\boldsymbol{\mathcal{Y}}$. The particular selection of control parameter values was a random choice not far from expected literature values or the result values of Chapter 3, but different enough to ensure the reliability of the benchmark test. As a control result, we perform Bayesian parameter estimation by running a MH algorithm normally until convergence. We will refer to this as the normal Bayesian inference method to differentiate from the fast Bayesian inference method that employs our algorithm. The final Markov chain was of length 10^5 and the 1D Marginalized posterior probability distribution and the 68% highest density region for each of the four parameters after normal Bayesian inference is shown in Figure 4.2. For our algorithm we set 18 ANN models with different hidden nodes. For the first 16 models the hidden nodes vary between 5 and 20 nodes and in addition we have two more models with 30 and 40 nodes. We run our fast Bayesian inference algorithm for the same observational values and after a Markov chain of length 45000 the ANN with 8 nodes took over the evaluation of the likelihood function. The 45000 samples translate to 18000 unique training data points. In terms of duration, the benefits from the fast Bayesian inference method were great. In normal mode the duration of the whole inference process was 108 hours. The fast Bayesian inference method managed to complete the whole process in under 49 hours. The speed of the algorithm though would be of no actual purpose, if the results were not accurate. Figure 4.3 shows the 1D Marginalized posterior probability distribution and the 68% highest density region for each of the four parameters after the fast Bayesian inference algorithm.

The results from the two inference methods as shown in Figure 4.2 and Figure 4.3 reveal some important insights. First of all, high probability density regions for all the parameters include and hence recover the true parameters. As we can see, all the pre-defined parameter values lie under or very close to the highest density point of the marginal posterior probability distribution. To support numerically this insight we chose to extract the statistical mean and standard deviation for each one of the parameters from the most probable mode of the joint distribution. The results can be seen in Table 4.1 and confirm our visual intuition. This result simply validates that both normal and fast Bayesian approaches make accurate inference based on the given observations and the MH algorithm samples efficiently the solution space with both the normal and the approximated likelihood function. We can also observe that most distributions are not Gaussian and in some cases we can observe two modes (see Figure 4.2(d) and Figure 4.3(d)). Finally, it is obvious that even though the highest density regions of the distributions between both the normal and the fast Bayesian inference methods are similar, the general shape of the distributions and especially across the less dense probability regions are different in some cases.

Parameters $\boldsymbol{\theta}$	Unit	True Parameter Value	Normal Bayesian Inference	Fast Bayesian Inference
			Mean and Standard Deviation	
ζ	$10^{-17} \cdot s^{-1}$	2.2	2.8(±0.74)	2.89(±0.64)
$G_\circ$	$Habing$	1.2	1.28(±0.48)	1.87(±0.89)
n_H	cm^{-3}	10^5	$1.27(\pm1.02)\cdot10^5$	$5.81(\pm1.44)\cdot10^5$
fr	-	50%	54(±0.08)%	52(±0.07)%

Table 4.1: Summary statistics for the most probable mode of the joint distribution.The mode corresponds to the 68% of the highest density of the distribution.

4.5 Conclusions

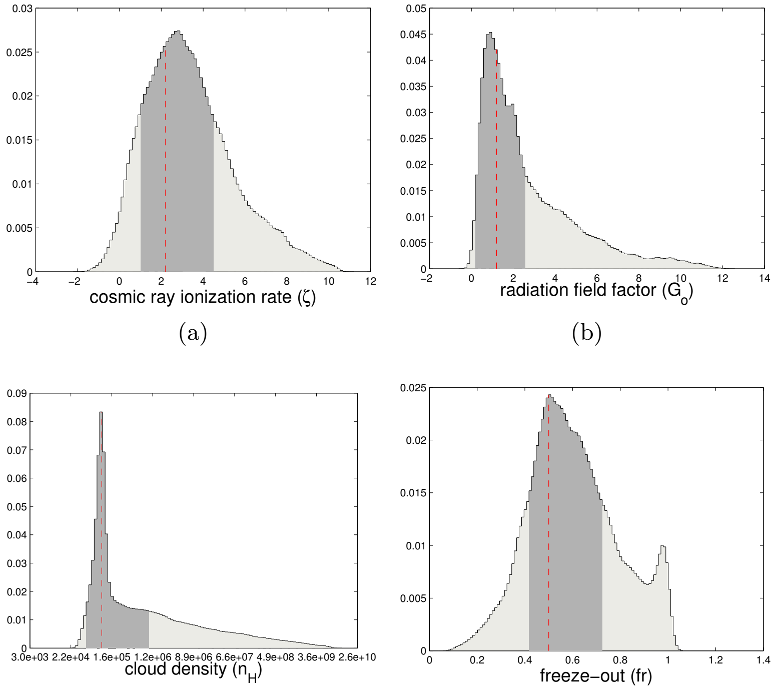
In this chapter we have introduced a fast Bayesian inference algorithm that combines the sampling efficiency of MCMC algorithms such as MH and the approximation efficiency of ANNs. We have demonstrated the performance of our algorithm in both accelerating the Bayesian analysis and approximating efficiently complex and non-linear likelihood function, using a toy example of typical astrochemical inverse problems. Our main conclusions are the following:

1. The fast Bayesian inference method provides an efficient way to speed up typical Bayesian methods for solving inverse astrochemical problems. We succeeded the same parameter estimation results as the normal Bayesian inference method in less than half of the duration.

2. ANNs can approximate successfully complex likelihood functions that include non linear chemical codes. Both the 68% high density regions and the summary statistics for both the normal and fast Bayesian inference method agree. The pre-defined parameter values lie under or very close to the highest density point of the marginal posterior probability distribution .

3. Even in the case of bimodal distribution the fast Bayesian inference method discovers fully all the modes of the posterior distribution (see Figure 4.2(d)), but seems to miss the density proportion between the main and the secondary mode.

4. Even though the high density regions of the posterior distributions are approximated properly by the fast Bayesian inference method, the exact shape of the distribution deviates a bit from the one derived by the normal Bayesian inference method. This



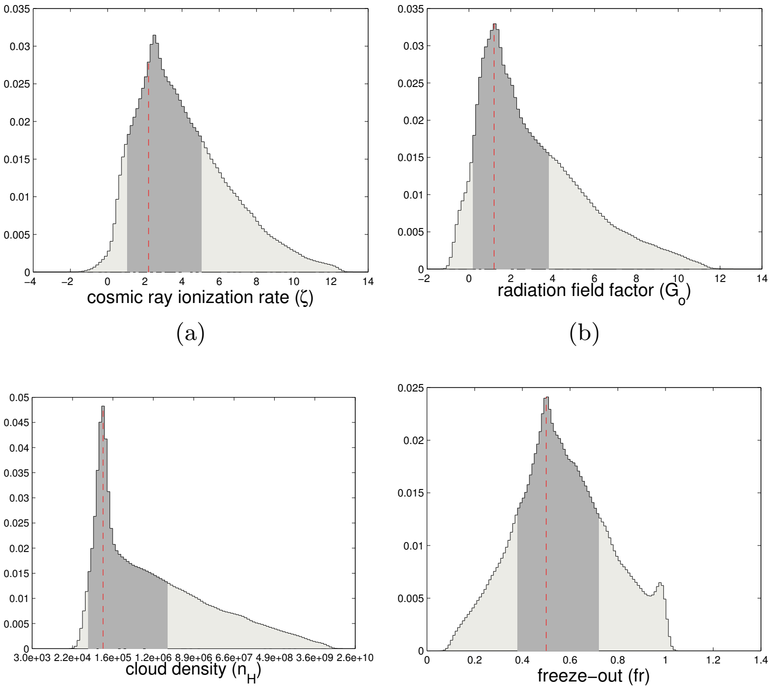
(a) (b)

(c) (d)

Figure 4.2: 1D Marginalized Posterior Probability Distribution for each of the four parameters using normal mode Bayesian inference with MH sampling algorithms. The plots show the Gaussian kernel density estimator of each Probability Density Function. Dark gray area indicate 68% highest density region, while dashed red lines indicate the pre-defined parameter values.

is probably due to the way ANNs generalize for data points that have not been presented during the training period.

Our results and conclusions indicate that ANNs can be especially useful for applications for which the Bayesian inference process would be expected to be tedious. This is usually the case when the set of the parameters and/or the parameter space are too large. In future work, ANNs can be used to estimate a large parameter set that would include both physical and chemical parameters of surface chemistry. The physical parameters can represent the physical conditions of the molecular cloud, the grain properties or the mechanisms that control surface chemistry (e.g. non-thermal desorption mechanisms, freeze-out e.t.c.). The chemical parameters can represent the reaction rates of chemical reactions. Chemical parameters of the latter type are explored in the next Chapter.



(a)

(b)

(c)

(d)

Figure 4.3: 1D Marginalized Posterior Probability Distribution for each of the four parameters using accelerated Bayesian inference with ANNs. The plots show the Gaussian kernel density estimator of each Probability Density Function. Dark gray area indicate 68% highest density region, while dashed red lines indicate the pre-defined parameter values.

Chapter 5

Bayesian Uncertainty Analysis of Surface Reactions

At this point, we can acknowledge that Bayesian inference can and should play an important role in astrochemical problems. The past two chapters have highlighted the benefits of the Bayesian approach towards understanding dark cloud processes and constrain their physical parameters. It is true that our knowledge regarding the molecular gas phase chemistry in the ISM has made impressive progress over the last years. However, we can not claim the same for the solid phase chemistry, where there is still too much uncertainty about surface reactions and rate coefficients. The aim of this chapter is to demonstrate whether and how we can use Bayesian inference methods to explore the solid phase chemical network parameter space and in particular the efficiency of established or new reaction routes. The outcome of such inference processes can be used either directly by theoretical astrochemists or alternatively can guide, in a structured probabilistic way, the laboratory experimentation processes. The latter reasons formed the motivation for the present chapter. Such probabilistic approach against grain chemistry uncertainty has not been attempted in the past. This chapter will function as a proof of concept to demonstrate the feasibility and efficiency of Bayesian methods for inferring grain chemistry parameters and providing helpful insight for laboratory experiments. To prove our concept, a simplified chemical modeling and theoretical problem setting will be utilized as a toy example.

5.1 Introduction

Before laboratory experimentations, all our knowledge about the surface reaction network was based on chemical intuition and gas phase chemistry analogues. Since laboratory astrochemists started using experimental techniques to test and evaluate the surface reaction inventory, efficiency of reaction routes are being properly explored, new reaction routes are constantly discovered and in general important information on how molecules form on grain surfaces is being revealed. The first experimental work on the dust surfaces studied the formation of molecular hydrogen (Pirronello et al. 1997). Several more experiments followed studying either the formation of more complex molecules (e.g. Watanabe et al. 2005; Ioppolo et al. 2009) or the ice morphology and ice mantle mechanisms (e.g. Fraser et al. 2004; Collings & McCoustra 2005). A typical experiment setting consists of a substrate, an Ultrahigh Vacuum (UHV) and Quadrupole Mass Spectrometer (QMS). The substrate represents the dust surface, while the UHV creates an environment that can reach a pressure of $10^{-10} - 10^{-11}$ mbar and a temperature as low as $12 - 15$ K. Ices are monitored by means of the QMS. Surface reactions of simple or more complex ices can be studied and investigated while varying a wide range of laboratory conditions. Typically, these conditions include different atomic fluxes, ice temperatures, ice thicknesses, ice structures, and mixture ratios. The aim of these experiments is to reveal the physics and chemistry of molecule formation on dust surfaces by replicating ice composition in star forming regions. The main focus is to both explore the impact of energetic processing on the interstellar chemistry and characterize ice and dust processes that are relevant to astrochemists. Such processes can include surface molecule formation, thermal desorption, non thermal desorption and diffusion of molecules. However, the truth is that little experimental information is yet available for the interstellar ices. And the main reason is that the experimentation process is neither simple nor fast. There is a huge list of questions that need to be answered regarding the surface reaction efficiencies, the ice composition and the energetics that have an impact on the processed ices. On the other hand, there is an even bigger list of potential experiments that need to be carried out and evaluated before beginning to answer all these questions. It is easy to grasp the complexity of the whole process. By the use of probabilistic inference and chemical models, we can help lab scientists to prioritize specific experiments that are more likely to produce insightful results. At the same time, while our knowledge of the interstellar medium is constantly improving, chemical models have to be employed to simulate the complex chemistry of diverse regions including icy mantles. To ensure the robustness of our models, it is essential to estimate or at least understand the sensitivity of as many chemical parameters as possible. Hence, to disentangle the chemistry of ISM ices, laboratory work combined with chemical models and Bayesian inference can be an invaluable tool.

The addressed problem in this chapter is intended to be as general as possible and is defined as follows: Given a grain or gas-grain chemical network and a set of observational, intuitive or scientifically hypothesized constraints, estimate confidence intervals and/or the sensitivity of the grain reactions or the grain reaction rates to the constraints. The reaction network to explore is defined by the user as an input. The free parameters of our problem are the reaction rates and any prior information on them can be included based on the exploratory level of the project. Therefore, we have an inverse problem similar to the one addressed in Chapter 3, where we want to infer parameter values using some sort of constrain. However, in this case we are focusing mainly on chemical parameters instead of physical parameters. Another crucial difference is that we do not necessarily have actual observational values as constrains for the species of interest, since very few ices have been observed. However, it is important to understand that the priority of the inference process in this case is not to estimate the precise parameter values, but to evaluate the reactions and get insight on the reaction rates. The toy example of this chapter was designed so to abide by the definition of the above problem and at the same time allow us to derive useful conclusions. Specifically, we want to get insight regarding the reactions and reaction rate constants of a grain chemical network. Our constraint is defined through a fractional abundance interval that includes the fractional abundance values that a species needs to reach in order to be observed. We will refer to this interval as observational interval and we can constrain our reaction rates based on whether species lie within that interval or not. The use of observational intervals allows us to make inference about species that we believe/assume that can or can not be observed, but we hold no specific observational information for. Similarly and based on what hypothesis we might wish to form, we can construct different constraints. The choice of the specific constraint was made so to test whether a traditional Bayesian approach can provide useful results from such a general and abstract restriction.

Similarly to Chapter 3, we combine Bayesian inference, a chemical code and a sampling algorithm as our proposed parameter estimation method. The Bayesian inference and the sampling algorithm setting are the same in principle, but adjusted to the specific problem requirements. Specifically for this project we developed a simple grain chemical model that ignores gas chemistry and focuses solely on surface reactions. In Section 5.2, we describe the chemical grain model. The inference process is presented in Section 5.3 and the results of our analysis in Section 5.4. Finally, our conclusions are discussed in Section 5.5

5.2 The Chemical Model

We developed a simple chemical modelling code that accounts solely for the surface chemistry of a dark molecular cloud. Our simplified model is a time-dependent single-point model that generates time series of solid phase molecular abundances as a function of the physical conditions of the molecular cloud and the chemical parameters of the defined chemical network. In total we have a chemical network of 22 species and 23 surface reactions that are listed in Table 5.2 and Table 5.3 respectively.

To model the surface chemistry of a dark cloud the abundance of each solid species is derived by solving rate equations for grain-surface chemistry. The formation and destruction mechanisms for a species i are given by the following kinetic equation:

$$\frac{dn_i}{dt} = \sum_{l,m} k^i_{lm} n_l n_m - n_i \sum_{i \neq r} k_r n_r - k^{des}_i n_i + k^{ads}_i n_i, \tag{5.1}$$

where k^i_{lm} is the reaction rate of all the reactions between species l and m that produce i, n_i is the concentration of species i, k_r represents the reaction rates of all the reactions where species i participates as reactant, while k^{des}_i and k^{ads}_i are the desorption and adsorption rates. For dense cores we expect freeze out to dominate over desorption. Therefore, for simplicity, the desorption rate in our case is assumed to be zero. Such assumptions are made keeping in mind that this project is a proof of concept and we need to focus on the applicability of our method rather than the accuracy of our modeling approach. Gas-phase species can be adsorbed on the grain surfaces and the adsorption rate is assumed zero for all but the following 5 species: CO, CS, O, OH and S. Ideally, we could have defined a gas phase species depletion rate that would be a function of the cross section of the grain, the thermal velocity of the species and the density of the species. However, our model does not include any gas chemistry or gas-grain interaction and it is impossible to retain a species concentration equilibrium. To compensate for the lack of gas phase information

Species i	Q^i_{ads}
CO	10^{-4}
CS	10^{-12}
O	10^{-4}
OH	10^{-8}
S	10^{-7}

Table 5.1: Cumulative gas to grain abundance with respect to H nuclei.

we consider the following. If we knew in advance the maximum cumulative quantity Q^i_{ads} for each species i that would freeze on to the grains after the longest possible collapse period (e.g. $n_H \sim 10^8\,cm^{-3}$), then we could define the adsorption rate as a function of the molecular cloud density. To specify Q^i_{ads} we run UCL_CHEM models for a range of possible physical parameters and set the average value of the sum of the depletion quantities of gas phase species i from all the runs as Q^i_{ads}. The values of Q^i_{ads} can be seen in Table 5.1. Considering that this is a proof of concept project, we chose to use UCL_CHEM instead of setting Q^i_{ads} as a free parameter in order to keep our set of free parameters as small as possible. For the collapse to a particular cloud density n_H we use the modified formula of Rawlings et al. (1992). Finally, all initial fractional abundances are practically zero and their actual value is set to 10^{-30}.

The type of reaction rate coefficients usually included in gas-grain chemical modeling codes, such as UCL_CHEM, for two body reactions is the Kooji-Arrhenius' equation (Côme 2001):

$$k(T) = \alpha \, \frac{T}{300}^{\beta} \, e^{-\gamma/T} \; [cm^3 s^{-1}], \tag{5.2}$$

where T is the gas temperature in K, β is the temperature exponent, γ is the fraction of the activation energy in $J\;mol^{-1}$ to the gas constant ($8.3145\;J\;mol^{-1}K^{-1}$) and α is a constant factor. The above equation accounts for reactions that occur in a three dimensional environment such as gas-phase reactions. Hence, in order to calculate surface rate constants we need to transform the formula or take a different approach. A full description of how the above formula can be transformed to account for surface reactions can be found in Occhiogrosso et al. (2012). In reality, in our case we can simplify the task by ignoring the temperature dependency and assume that both β and γ are equal to zero. Essentially, our parameter estimation is restrained to exploring values for the constant α. In this case, α represents a constant that simply describes the efficiency of a reaction,

Species
CH_3OH, CO, CO_2, CS, CS_2, H, H_2CO, H_2CS, H_2O, H_2S, H_2S_2, HCO, HCS, HOCS, HS, HSO, O, OCS, OH, S, SO, SO_2

Table 5.2: Species

No.	Reactions				
1.	O	+	H	→	OH
2.	OH	+	H	→	H_2O
3.	CO	+	OH	→	CO_2
4.	S	+	H	→	HS
5.	HS	+	H	→	H_2S
6.	H_2S	+	S	→	H_2S_2
7.	CS	+	H	→	HCS
8.	HCS	+	H	→	H_2CS
9.	CO	+	S	→	OCS
10.	OCS	+	H	→	HOCS
11.	H_2S	+	CO	→	OCS
12.	H_2S	+	H_2S	→	H_2S_2
13.	H_2S_2	+	CO	→	CS2 + O
14.	H_2S	+	O	→	SO_2
15.	CS_2	+	O	→	OCS + S
16.	CO	+	HS	→	OCS
17.	S	+	O	→	SO
18.	SO	+	O	→	SO_2
19.	SO	+	H	→	HSO
20.	HSO	+	H	→	SO
21.	CO	+	H	→	HCO
22.	HCO	+	H	→	H_2CO
23.	H_2CO	+	H	→	CH_3OH

Table 5.3: Reaction Network

which at this point is the only aspect we seek to explore. The constant α is normally in units of $mol^{-1}cm^{-3}s^{-1}$. However, in the context of this chapter we will consider it as a unitless index of efficiency.

5.3 Bayesian Inference

Our aim is to obtain information about the set of reaction rates $\boldsymbol{k} = (k_1, k_2, ..., k_{23})$ of our surface chemical network, where k_j is the reaction rate of reaction j, using the simulated molecular abundances $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_{22})$ where $\mathcal{Y}_i$ is the abundance of species i. In our toy example, we will not need to constrain the whole set of $\boldsymbol{\mathcal{Y}}$, but on a limited set of species. Hence, we define the set of simulated molecular abundances $\boldsymbol{S} = \{S_1, S_2\}$, where $\boldsymbol{S} \subset \boldsymbol{\mathcal{Y}}$ and $S_1 = \{H_2O, CO, CO_2, CH_3OH\}$, $S_2 = \{HCO, HOCS, HS, O, S, H_2S_2\}$. Note that S_1 and S_2 refer to the fractional abundances of the noted species. However, for convenience and since there is no risk of confusion, we will use the same notation for the set of the species themselves as well.

These quantities are related through our chemical code $\mathcal{C}(\cdot)$, so that $\boldsymbol{S} = \mathcal{C}(\boldsymbol{k})$. Note that we do not include an error term since we do not attempt to match an actual observed value. Instead, we are trying to generate values that do or do not lie in an observable abundance interval. Let O_{int} be the observable abundance interval, then with respect to H nuclei we define O_{int} as:

$$O_{int} = [10^{-8}, 10^{-4}] = \{x \in \mathbb{R} | 10^{-8} \leq x \leq 10^{-4}\}. \tag{5.3}$$

We want to derive insight for the reaction rate constants that generate fractional abundances so that $S_1 \in O_{int} \wedge S_2 \notin O_{int}$, where the symbol $\wedge$ denotes the logical 'and'. The parameter estimation can be performed through the posterior probability distribution (PPD) of the reaction rates given the species abundances. Using Bayes' rule we have that the PPD is:

$$\pi(\boldsymbol{k}|\boldsymbol{S}) = \frac{\mathcal{L}(\boldsymbol{k};\boldsymbol{S})\pi(\boldsymbol{k})}{m(\boldsymbol{S})} \propto \mathcal{L}(\boldsymbol{k};\boldsymbol{S})\pi(\boldsymbol{k}) \tag{5.4}$$

The PPD expresses our uncertainty about the reaction rates after considering the species abundances and any prior information we might have. The denominator is simply a normalization factor. The prior information on reaction rates is defined as a uniform distribution that is non-zero when the reaction rates are between 10^{-7} and 10^{-18} and zero elsewhere. The limits of the exploration domain (i.e. 10^{-7} and 10^{-18}) were considered to define a reasonable exploration range, however their values are up to the end user. Our specific limit values were chosen so as to represent a more exploratory range than the one we usually meet in gas phase reactions.

The peculiarity of our case is the form of the likelihood function, which is taken to be a Poisson distribution. The likelihood function should express the probability of species in S_1 lying within and species of S_2 lying outside O_{int}, given a set of reaction rates. We require this probability to be maximum when all the species in S lie within their pre-specified interval and minimum when none of the species in S lies within their pre-specified interval. Hence, we require a discrete probability distribution that would express this probability. The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring. An event occurrence in this instance is considered to be the case when $S_1 \in O_{int}$ or $S_2 \notin O_{int}$. We define a random variable X which expresses the number of species in S_1 that lie within the observable abundance interval plus the number of species in S_2 that do not, so that $X \in \{0, 1, ..., 10\}$, where 10 is the cardinality of $\boldsymbol{S}$. The likelihood of a reaction rate set $\boldsymbol{k}$ given that the sum of the species from S_1 that lie within O_{int} and the species from S_2 that do not is n, is equal to the probability mass function of X being equal to n. If X has a Poisson distribution with parameter λ, the probability mass function of X being equal to n is given by:

$$Pr(X = n) = \frac{\lambda^n e^{-\lambda}}{n!}. \tag{5.5}$$

The positive real number λ is equal to the expected value of X and in our case is equal to all possible event occurrences, so that $\lambda = 10$. It is apparent that when a number n of species lie within O_{int} for different sets of $\boldsymbol{k}$, all these sets will be 'scored' equally from our likelihood function, independently of the exact species or the exact abundances. Therefore, we do not aim or expect to estimate an optimal reaction rate setting. Our objective is to estimate probable areas of the reaction rates' that generate observable abundances and to that extend make a comparative study on the efficiency of the reactions.

Again, we are not able to access the whole posterior probability distribution. Therefore, we employ the exact MH sampling algorithm of Chapter 3. We run 5 independent Markov chains of length $500,000$ samples and for each sample we collect the fractional abundance for H_2O, CO, CO_2, CH_3OH at a final density of $10^7\,cm^{-3}$. For more details on MCMC and MH we refer readers to section 3.2.2 and references within.

Before presenting the results of our analysis, it is useful to justify the choice of the sets S_1 and S_2. The species that belong to S_1 are the ones that we require to be abundant enough to be observed. Since this is a toy example, we chose species (i.e. H_2O, CO, CO_2, CH_3OH) that are already observed in icy mantles, so that it would be easier to benchmark our results. Regarding the set S_2, we chose species (e.g. HCO, HOCS) that are not expected to be abundant enough to be observed or species (e.g. O, S) that even though we expect them to be abundant, since they adsorb from the gas phase, they efficiently hydrogenate to form other species.

5.4 Results

Our results are presented in the form of marginalized posterior probability distributions (PPD) for the reaction rate constants. The density of each marginalized PPD reveals the areas where the corresponding reaction rate is more probable based on the imposed constraints. The marginalized PPD can be seen in Figure 5.1 and Figure 5.2. An obvious result is that for 8 of the reaction rates (see Figure 5.1) the distributions present enough variability to reach some constructive conclusions around our constraints. On the contrary, for the remaining 15 of the reaction rates the distributions are nearly uniform. For brevity, the marginalized PPD for 4 of those can be seen in Figure 5.2. Although obvious, it is worth noting what a flat distribution for a reaction rate constant entails. Essentially, the reactions of the reaction rates with uniform distributions do not impact in any way our desired outcome as defined by our imposed constraints. In different words, for the given constraints our chemical network could have been simplified by omitting 15 out of 23 reactions. Such a result was expected and could have been speculated based on the defined constraints. However, we could have not predicted the specific result.

Let us focus now on Figure 5.1 and the reaction rates that present some variability. A general observation would be that for most of the distributions in Figure 5.1 we can notice two or three different probability density levels that are not always smoothly separated. This step-function behavior can be accounted to the fact that our likelihood function is discrete. The PPD of k_1 has a dense peak around 10^{-7}, with a less dense flat region between 10^{-14} and 10^{-9}. The first reaction produces OH and based on the marginal PPD of k_1, it needs to be extra efficient. This result can be explained considering that OH is crucial for the production of H_2O and CO_2, both members of the S_1 set. The PPD of k_2 presents a denser region between 10^{-13} and 10^{-7}, with a slight denser peak close to 10^{-12}. The second reaction produces H_2O, hence needs to be efficient, but also competes for OH with reaction number 3 that produces CO_2. The distribution of the reaction rate constant k_3 shows a nearly uniform behavior, with a small denser region between 10^{-13} and 10^{-9}. The third reaction destroys CO which is a member of S_1, so we wouldn't expect an extremely high reaction rate. At the same time, the third reaction, as already mentioned, competes with the second for OH, which also explains the lack of high dense regions at the right side of the distribution. However, the product of reaction number 3 is CO_2, which is also a member of S_1. That probably explains why the denser bump is somewhere in the middle. The distribution of k_4 shows a somehow smoother behavior and resembles a bimodal distribution. Both modes are quite wide. The denser mode is peaking around 10^{-8}, while the second one around 10^{-17}. The behavior of k_4 shows that intermediate values are way less likely to produce results that align with our constraints. In order to understand or speculate why that might happen we will try to analyze all possible cases. The forth reaction destroys H and S to produce HS. Both S and HS are members of the S_2 set, so our results should point us towards low abundance values for both of them. Let us assume that k_4 is high. That means that S is destroyed, which agrees with our constraint. On the other hand, that would also mean that HS would be abundant. HS though, will only be an intermediate stage and will hydrogenate again. The distribution rate of k_5 that will be discussed later indicates that HS will be destroyed for the production of H_2S. Now, let us assume that reaction number 4 is not efficient. The abundance of HS will be low, however in order to decrease the abundance of S, different reactions should become more efficient and that could potentially increase the abundance of other members of S_2 (see HOCS and reactions 9 and 10). Hence the mode around 10^{-17} is less dense. Lastly, if k_4 was somewhere in the middle, we would probably retain most of the risks discussed in the previous situations without gaining any of the benefits. That explains the deep smooth drop of k_4 marginal PPD around 10^{-13}. Moving to reaction number 5, we have already argued that since HS is a member of S_2, it should only function as an intermediate species. Hence, k_5 is most likely to be high enough to keep HS out of the observational interval. The marginal PPD of k_{21} highlights that the 21st reaction should not be efficient. For the reactions 21, 22 and 23 CO is hydrogenated successively to form HCO and CH_3OH. HCO is a member of S_2, while CH_3OH of S_1. According to the marginal PPD of k_{21}, k_{22} and k_{23}, in order to keep the final abundance of HCO outside the observational interval and at the same time produce enough CH_3OH, the 21st reaction should not be efficient. The k_{21} peaks around 10^{-18}, but a value around 10^{-14} is also very likely. The 22nd reaction on the other hand should be very efficient, with the k_{22} peaking around 10^{-8}. Finally, k_{23} has a wide dense region that peaks around 10^{-10} and indicates a relative efficient reaction.

The marginalized PPD of the reaction rates clearly indicate that there is a clear connection between the species we impose constraints on and the reaction rates that present sensitivity on the results. However, there are species and reactions that are not directly impacted by the constraints, but nevertheless present certain sensitivity. We should be really careful in these cases. By imposing constraints only on certain species from a chemical network we automatically imply that we are either indifferent or ignorant about the rest. Therefore, we must either be certain that we are indeed indifferent or perform some post-algorithm analysis on the behavior of these species for useful insights. Let us make that clearer with an example. Reactions 4,5 and 6 describe the successive hydrogenation of S, to HS and H_2S, which then reacts with S to form H_2S_2. The species S, HS and H_2S_2 belong to the S_2 set, however we have set no constraint on H_2S. Figure 5.3 shows the fractional abundance of the above species as a function of time using a set of reaction rate constants based on our earlier results. The S adsorbs from the gas, but never gets to its Q^i_{ads} value, since it gets destroyed from multiple reactions including reactions number 4 and 6. The abundance of HS seems to increase steadily for about 10^4 years, but then seems to be destroyed more than produced. The high value of k_5 can explain that. At the same time, the abundance of H_2S_2 seems to start increasing, but no matter what is the efficiency of reaction 6, there is not possibly enough time to reach the observational interval levels. On the other hand, the abundance of H_2S increases enough to end up in the observational interval. Therefore, it is safe to assume that by not constraining H_2S, we allowed our algorithm towards a very efficient route to keep S, HS and H_2S_2 outside of the observational interval. Let us assume that this was a real case scenario and not a proof of concept project. In that case, if H_2S should have never reached an observational interval abundance, then by not constraining it we forced our algorithm to a wrong analysis. However, if we assume that we held no prior information at all about H_2S, our analysis would indicate that based on our imposed constraints, H_2S should be abundant enough to be observed. That would be a very useful insight.

While this was clearly only a proof of concept exercise, we succeeded in delivering a small subset of experimentally 'interesting' reactions. From this subset, one could prioritize focusing on the reactions that are extremely sensitive on the value of the reaction rate constant. For example, k_1 presents a very dense mode, which indicates that the first reaction has to be very efficient, since any other alternative reduces the probability density by much. Reactions 21 and 22 present similar behavior, with k_{21} also presenting two very dense modes, that could be worth experimenting with. The rest of the remaining reactions might not present as high dense regions, but still constraint enough the parameter space to either guide experimental scientists or aid theoretical astrochemists with determining possible values for their chemical codes. Finally, by monitoring the evolution of the species' abundances under the most probable reaction rate set, we can identify dependencies or insights on species that we have no prior information at all.

5.5 Conclusions

This chapter presented a novel way to tackle uncertainty about surface reactions and rate coefficients using Bayesian inference. To prove the efficiency of Bayesian techniques to provide insight on the chemical parameters of surface reactions, we tested our algorithm with a proof of concept toy example. For our analysis we developed a simple grain chemical code and with the help of MCMC sampling algorithms we exploited the Bayesian inference principles in order to get information about the reaction rate constants of a simplified chemical network. In order to test for situations where there is no specific observational information, we defined general and vague constraints. The main conclusions of this work are as follows.

1. The Bayesian method provides a systematic approach to get insight on chemical parameters even with vague and not very informative constraints.

2. The results are highly sensitive to the definition of the constraints. The constraints should reflect exactly our knowledge or what we are willing to allow the algorithm to know in order to make inference.

3. Despite the vague constraints, our approach managed to estimate wide but useful intervals for the reaction rates.

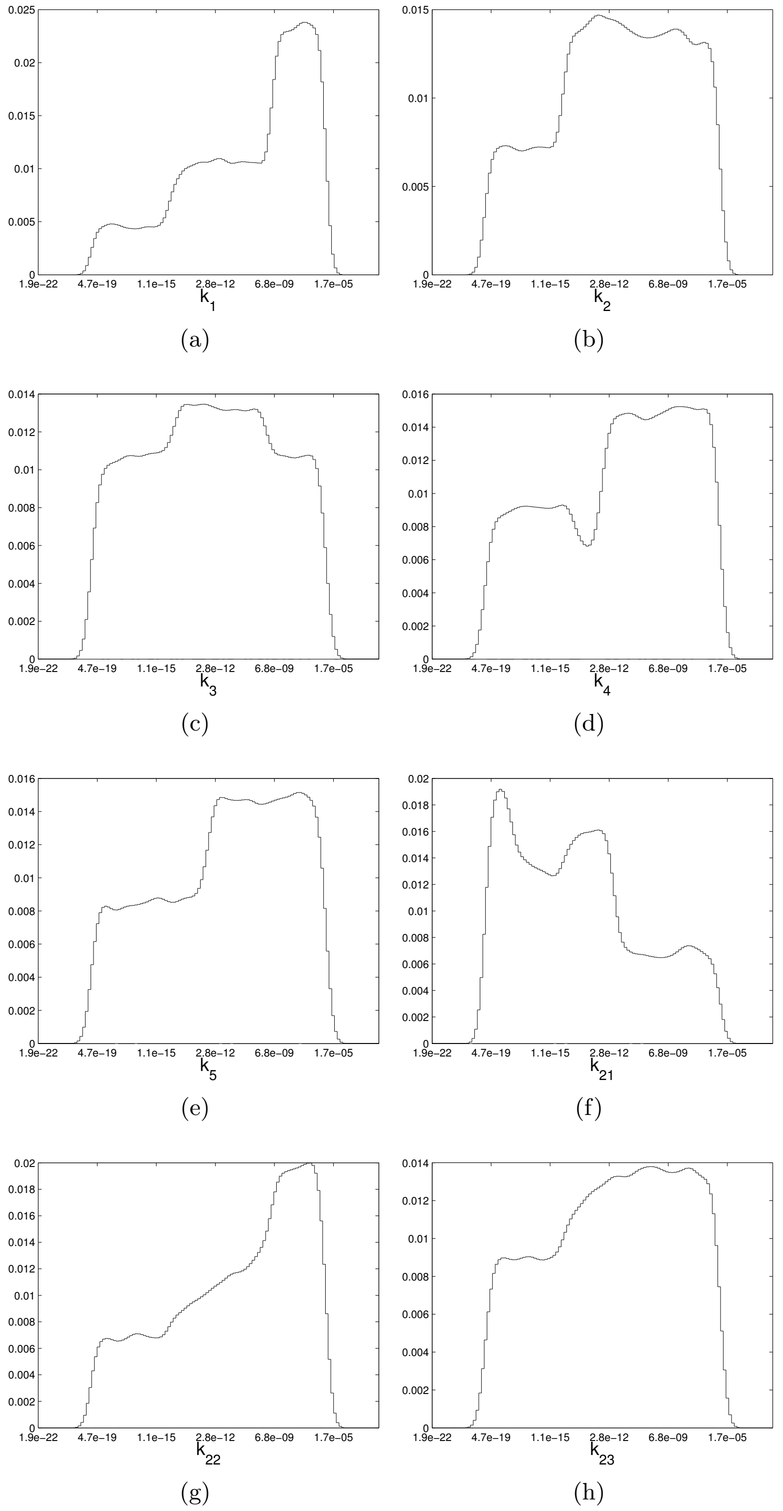
4. Our method managed to identify the list of reactions and species that are important. A simpler chemical network could be designed after our results.

5. The algorithm can recognize indirect dependencies even among the species that are not directly impacted by the constraints.

6. On the whole, Bayesian inference proved that can be an invaluable knowledge discovery tool against our uncertainty about surface reactions and rate coefficients.

7. Finally, both our type of analysis and potential results can greatly contribute to both experimental and theoretical benefits.

The scope of the present chapter was to demonstrate whether Bayesian analysis techniques can be used by astrochemists to tackle surface reaction uncertainty problems. Since we have established that the Bayesian approach is an invaluable tool and based on our results and conclusions, we now state what can be improved or what further steps should be made. First of all, to increase the validity of our results, better, more complete and accurate chemical modeling should be made. Radiation field and cosmic rays should be included, as well as thermal and non-thermal desorption mechanisms. A more sophisticated way should be found to model the gas-grain interaction by either physical simulation or by including the adsorption parameters as free parameters. In order to get more information, a more complete chemical network can be used. Apart from the adsorption parameters, more parameters can be explored, such as the final density of the cloud. In addition, the benefits of Bayesian analysis could be further exploited in conjunction with a more complicated chemical modeling approach such as the one suggested by Garrod (2008). Finally, the whole approach should be applied to a more realistic and useful project.



(a)

(b)

(c)

(d)

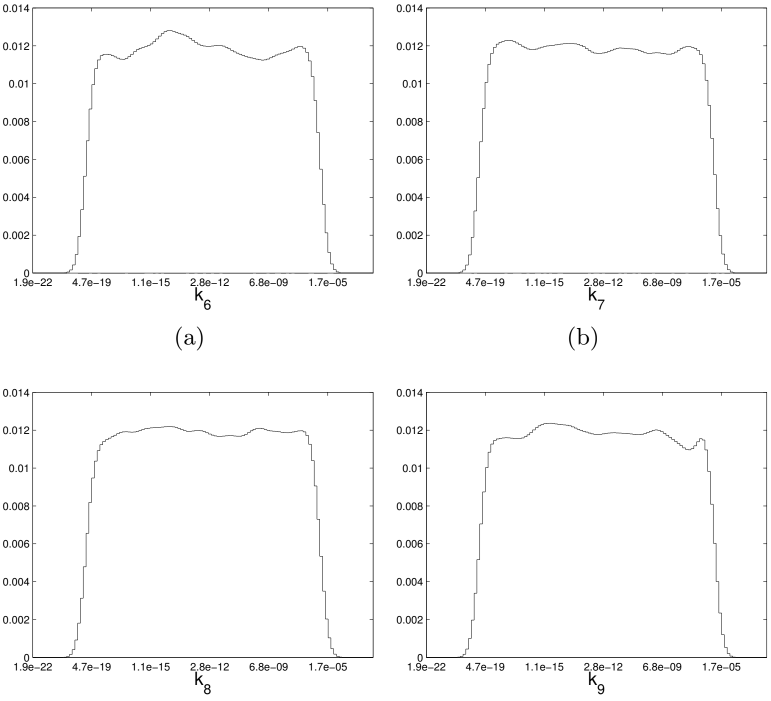
(e)

(f)

(g)

(h)

Figure 5.1: 1D Marginalized Posterior Probability Distribution for 8 reaction rate coefficients using Bayesian inference with MH sampling algorithm. The plots show the Gaussian kernel density estimator of each Probability Density Function.



(a) (b)

(c) (d)

Figure 5.2: 1D Marginalized Posterior Probability Distribution for 4 reaction rate coefficients using Bayesian inference with MH sampling algorithm. The plots show the Gaussian kernel density estimator of each Probability Density Function.

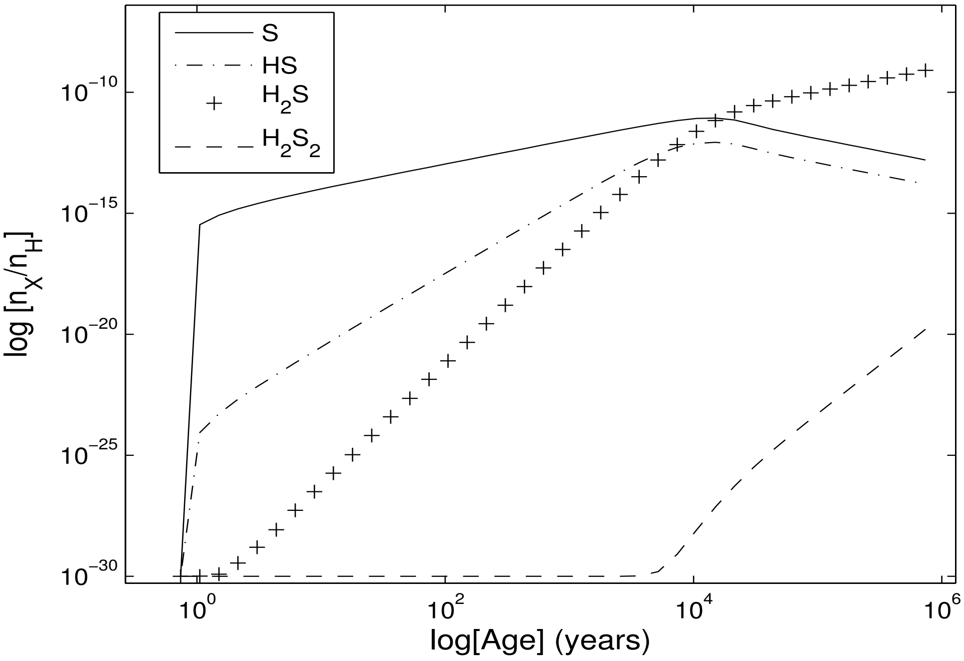


Figure 5.3: Fractional abundance time series for 4 species, using a set of reaction rate coefficients based on our results.

Chapter 6

Conclusions

The aim of this thesis has been to introduce machine learning and probabilistic methods for solving typical astrochemical problems. As larger datasets and more complex models are being employed in astrochemistry, the need for intelligent data mining algorithms will increase. We have explored 3 different machine learning approaches for interstellar knowledge discovery through molecular data and chemical models: Clustering analysis, probabilistic inference and predictive modelling. Each chapter presents a new way to combine, interpret and analyze this cornucopia of data observed in ISM or generated by chemical models. The following points summarize the key conclusions from each chapter of the present thesis:

1. In Chapter 2, we demonstrated how traditional and probabilistic clustering algorithms can provide insight in synthetic time series of molecular abundances. We described the nature of the data produced by chemical codes and the challenges the data present. The analysis of the chapter was based on a traditional clustering algorithm, the hierarchical clustering algorithm. Advantages and disadvantages of the method were discussed, as well as ways to overcome data challenges. A probabilistic version of the algorithm was then suggested as a natural upgrade of the traditional algorithm. By adopting a probabilistic approach we managed to naturally overcome most of the challenges and still discover structure in our data. Both approaches were tested against a database of synthetic time series of molecular abundances for a large

grid of physical and chemical parameters. The results pointed out that clustering methods perform efficiently in finding structure, patterns and insight in big data sets of astrochemical data. Furthermore, our results constituted a perfect showcase on how both traditional and emerging machine learning algorithms can discover and highlight useful astrophysical information in readily available data.

2. Chapter 3 combined a new way of solving astrochemical parameter estimation problems with an application in icy mantles of molecular clouds. Discovering the properties of dark molecular clouds where icy mantles evolve, using observational constraints of solid phase species is a typical inverse problem. Bayesian inference was employed to alleviate our uncertainty about the physical processes in these molecular clouds. We implemented a Bayesian inference algorithm that used Metropolis Hastings (i.e. a MCMC algorithm) to explore the parameter space and a chemical code for the evaluation of the likelihood function. We concluded that the Bayesian method provides a systematic approach to solve non-linear inverse problems with high noise levels and significant model uncertainties. The MCMC technique allows us to sample from complex probability distributions in an efficient way. Our method successfully handled a typical astrochemical ill-posed problem and revealed a more complete set of solution regions compared to traditional approaches.

3. The scope of Chapter 4 was to improve the speed and computational cost of the methods suggested in Chapter 3, without any loss of accuracy. A machine learning supervised algorithm was used and tested against a simplified (compared to Chapter 3) problem. The most computationally and timely expensive component of our Bayesian approach was the evaluation of the likelihood function. This function estimates how likely is for a set of parameters to be valid, based on the chemical abundances it generates and given our observational constraints. For each evaluation of the likelihood function, a single run of a lengthy chemical code is required. We suggested and implemented an ANN that learned the likelihood function after a number of evaluations. Our results demonstrated that ANNs can be used successfully to accelerate Bayesian inference without loss of accuracy. We also concluded that ANNs can efficiently learn complex non-linear function that commonly occur in astrophysics.

4. In Chapter 5, we extended our Bayesian inference algorithm to vaguely constrained

astrochemical problems. There is too much uncertainty regarding surface reactions and rate coefficients. On top of that, there are no structured ways to explore the solid phase chemical network so far, apart from lengthy lab experiments. The complication of the network, the lack of any prior information and the duration of lab experiments makes an actual parameter exploration and significant knowledge discovery unfeasible. In order to contribute towards a faster and smarter parameter exploration, we demonstrated with a proof of concept toy example how Bayesian inference can provide invaluable insight, information and guidance about both surface reactions, reaction rates and specific surface species. This information can be used either directly by theoretical astrochemists or guide experimentation for a more accurate and precise parameter value. We constructed a simplified grain chemistry chemical code over a simple chemical network. We assumed we had no precise observational constraints and tested our method with vague and incomplete observational intervals instead. Based on the results, we concluded that Bayesian inference performs efficiently even under those conditions, discovering and providing useful insights. Bayesian inference proved to be an invaluable knowledge discovery tool against our uncertainty about surface reactions and rate coefficients. We also suggested that further work should be carried out to extend the capabilities of this approach and support real scientific projects.

Before the end of this chapter and thesis, it is worth recapping what this thesis has accomplished at a higher level. In a few words, we managed to build and/or put together the components of an agile and modular framework that can guide and assist an astrochemist through advanced inference techniques and algorithms. Our contribution was not only through the machine learning algorithms that provided and supported the analytical aspect of the framework, but also through the process design and infrastructure of such a framework. The process diagram that visually describes the developed framework can be seen in Figure 6.1. Every rectangular box in the diagram describes a process. Each one of the processes represents a component of the framework that can be redesigned, augmented, altered or replaced by the user with an alternative process that benefits the scope of each project. For example, the chemical code to be used is totally up to the end user, who can choose between UCL_CHEM and other established chemical modeling codes or simply develop his own and plug it in. On top of that, each component may or

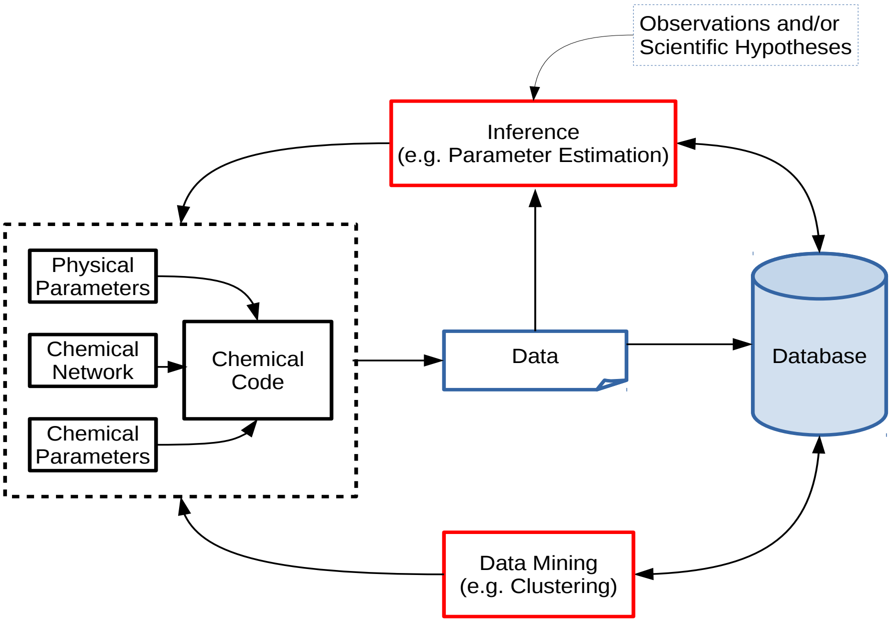


Figure 6.1: Process diagram of the developed framework in this thesis.

may not consist of multiple processes that can either be employed or simply ignored. For example the inference component may consist of a number of modules such as different sampling algorithms to choose from or accelerated likelihood evaluation through ANNs that can be switched on or off. Finally, the database is simply a relational database for data storage, management, retrieval and integration of all the information generated from chemical codes, observations or analytical processes.

Bibliography

Allen, M. & Robinson, G. W., 1977, *The molecular composition of dense interstellar clouds*, *ApJ*, **212**, 396

AMI Consortium, Hurley-Walker, N., Bridle, S., Cypriano, E. S., Davies, M. L., Erben, T., Feroz, F., Franzen, T. M. O., Grainge, K., Hobson, M. P., Lasenby, A., Marshall, P. J., Olamaie, M., Pooley, G., Rodríguez-Gonzálvez, C., Saunders, R. D. E., Scaife, A. M. M., Schammel, M. P., Scott, P. F., Shimwell, T., Titterington, D., Waldram, E. & Zwart, J. T. L., 2012, *Bayesian analysis of weak gravitational lensing and Sunyaev-Zel'dovich data for six galaxy clusters*, *MNRAS*, **419**, 2921

Auld, T., Bridges, M., Hobson, M. P. & Gull, S. F., 2007, *Fast cosmological parameter estimation using neural networks*, *MNRAS*, **376**, L11

Bacmann, A., Lefloch, B., Ceccarelli, C., Castets, A., Steinacker, J. & Loinard, L., 2002, *The degree of CO depletion in pre-stellar cores*, *A&A*, **389**, L6

Ball, N. M. & Brunner, R. J., 2010, *Data Mining and Machine Learning in Astronomy*, *International Journal of Modern Physics D*, **19**, 1049

Bazot, M., Bourguignon, S. & Christensen-Dalsgaard, J., 2012, *A Bayesian approach to the modelling of α Cen A*, *MNRAS*, **427**, 1847

Becla, J., Hanushevsky, A., Nikolaev, S., Abdulla, G., Szalay, A., Nieto-Santisteban, M., Thakar, A. & Gray, J., 2006, in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6270 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*

Benson, P. J. & Myers, P. C., 1989, *A survey for dense cores in dark clouds*, *ApJSS*, **71**, 89

Bergin, E. A., 2011, in M. Röllig, R. Simon, V. Ossenkopf & J. Stutzki (eds.), *EAS Publications Series*, volume 52 of *EAS Publications Series*, pp. 207–216

Bergin, E. A., Plume, R., Williams, J. P. & Myers, P. C., 1999, *The Ionization Fraction in Dense Molecular Gas. II. Massive Cores*, *ApJ*, **512**, 724

Bergin, E. A. & Tafalla, M., 2007, *Cold Dark Clouds: The Initial Conditions for Star Formation*, *Ann. Rev. Astr. Astrophys.*, **45**, 339

Bishop, C. M., 2006, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc., Secaucus, NJ, USA)

Bisschop, S. E., Fraser, H. J., Öberg, K. I., van Dishoeck, E. F. & Schlemmer, S., 2006, *Desorption rates and sticking coefficients for CO and N_2 interstellar ices*, *A&A*, **449**, 1297

Bonnell, I. A., Bate, M. R. & Zinnecker, H., 1998, *On the formation of massive stars*, *MNRAS*, **298**, 93

Boogert, A. C. A., Huard, T. L., Cook, A. M., Chiar, J. E., Knez, C., Decin, L., Blake, G. A., Tielens, A. G. G. M. & van Dishoeck, E. F., 2011, *Ice and Dust in the Quiescent Medium of Isolated Dense Cores*, *ApJ*, **729**, 92

Borne, K., 2009, *Scientific Data Mining in Astronomy*, *ArXiv e-prints*

Brunner, R. J., Djorgovski, S. G., Prince, T. A. & Szalay, A. S., 2001, *Massive Datasets in Astronomy*, *ArXiv Astrophysics e-prints*

Burke, D. J. & 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
URL: *http://dx.doi.org/10.1039/B917005G*

Caselli, P., Walmsley, C. M., Tafalla, M., Dore, L. & Myers, P. C., 1999, *CO Depletion in the Starless Cloud Core L1544*, *ApJL*, **523**, L165

Cazaux, S., Cobut, V., Marseille, M., Spaans, M. & Caselli, P., 2010, *Water formation on bare grains: When the chemistry on dust impacts interstellar gas*, *A&A*, **522**, A74

Cernicharo, J. & Bachiller, R., 2012, *The Molecular Universe (IAU S280)*

Christensen, N. & Meyer, R., 2000, *Bayesian Methods for Cosmological Parameter Estimation from Cosmic Microwave Background Measurements*, *ArXiv Astrophysics e-prints*

Christie, H., Viti, S., Yates, J., Hatchell, J., Fuller, G. A., Duarte-Cabral, A., Sadavoy, S., Buckle, J. V., Graves, S., Roberts, J., Nutter, D., Davis, C., White, G. J., Hogerheijde, M., Ward-Thompson, D., Butner, H., Richer, J. & Di Francesco, J., 2012, *CO depletion in the Gould Belt clouds*, *MNRAS*, **422**, 968

Collings, M. P. & McCoustra, M. R. S., 2005, in D. C. Lis, G. A. Blake & E. Herbst (eds.), *Astrochemistry: Recent Successes and Current Challenges*, volume 231 of *IAU Symposium*, pp. 405–414

Côme, G., 2001, *Gas-Phase Thermal Reactions: Chemical Engineering Kinetics* (Springer) **URL:** *http://books.google.co.uk/books?id=1wflKGgb2rwC*

Dalgarno, A., 2006, *Interstellar Chemistry Special Feature: The galactic cosmic ray ionization rate*, *Proceedings of the National Academy of Science*, **103**, 12269

Doty, S. D., Schöier, F. L. & van Dishoeck, E. F., 2004, *Physical-chemical modeling of the low-mass protostar IRAS 16293-2422*, *A&A*, **418**, 1021

Draine, B. T., 1978, *Photoelectric heating of interstellar gas*, *ApJSS*, **36**, 595

Duda, R. O. & Hart, P. E., 1973, *Pattern Classification and Scene Analysis* (John Willey & Sons, New Yotk)

Dulieu, F., Amiaud, L., Congiu, E., Fillion, J.-H., Matar, E., Momeni, A., Pirronello, V. & Lemaire, J. L., 2010, *Experimental evidence for water formation on interstellar dust grains by hydrogen and oxygen atoms*, *A&A*, **512**, A30

Dyson, J. E. & Williams, D. A., 1997, *The physics of the interstellar medium*

Eddington, A. S., 1937, *Interstellar matter*, *The Observatory*, **60**, 99

Feroz, F. & Hobson, M. P., 2008, *Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses*, *MNRAS*, **384**, 449

Fitzgerald, M. P., Kalas, P. G. & Graham, J. R., 2007, *A Ring of Warm Dust in the HD 32297 Debris Disk*, *ApJ*, **670**, 557

Fontani, F., Giannetti, A., Beltrán, M. T., Dodson, R., Rioja, M., Brand, J., Caselli, P. & Cesaroni, R., 2012, *High CO depletion in southern infrared dark clouds*, *MNRAS*, **423**, 2342

Ford, E. B., 2005, *Quantifying the Uncertainty in the Orbits of Extrasolar Planets*, *AJ*, **129**, 1706

Fraser, H. J., Collings, M. P., Dever, J. W. & McCoustra, M. R. S., 2004, *Using laboratory studies of CO-H_2O ices to understand the non-detection of a 2152cm^{-1} (4.647μm) band in the spectra of interstellar ices*, *MNRAS*, **353**, 59

Fraser, H. J., McCoustra, M. R. S. & Williams, D. A., 2002, *Astrochemistry : The molecular universe*, *Astronomy and Geophysics*, **43**(2), 10

Garrod, R. T., 2008, *A new modified-rate approach for gas-grain chemical simulations*, *A&A*, **491**, 239

Gilks, W., Richardson, S. & Spiegelhalter, D., 1995, *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC Interdisciplinary Statistics (Taylor & Francis)
URL: *http://books.google.co.uk/books?id=TRXrMWY_i2IC*

Gillett, F. C. & Forrest, W. J., 1973, *Spectra of the Becklin-Neugebauer point source and the Kleinmann-Low nebula from 2.8 to 13.5 microns.*, *ApJ*, **179**, 483

Graff, P., Feroz, F., Hobson, M. P. & Lasenby, A., 2012, *BAMBI: blind accelerated multimodal Bayesian inference*, *MNRAS*, **421**, 169

Hadamard, J., 1902, *Sur les problèmes aux dérivés partielles et leur signification physique*, *Princeton University Bulletin*, **13**, 49

Hartigan, J. A. & Wong, M. A., 1979, *A K-Means Clustering Algorithm*, *Applied Statistics*, **28**, 100

Hartquist, T. W. & Williams, D. A., 1990, *Cosmic-Ray Induced Desorption and High Depletions in Dense Cores*, *MNRAS*, **247**, 343

Hartquist, T. W. & Williams, D. A., 2008, *The Chemically Controlled Cosmos*

Heller, K. A. & Ghahramani, Z., 2005, in *Proceedings of the 22Nd International Conference on Machine Learning* (ACM, New York, NY, USA), ICML '05, pp. 297–304
URL: *http://doi.acm.org/10.1145/1102351.1102389*

Herbst, E. & van Dishoeck, E. F., 2009, *Complex Organic Interstellar Molecules*, *Ann. Rev. Astr. Astrophys.*, **47**, 427

Hoare, M. G., 2004, *Star formation at high angular resolution*, *nar*, **48**, 1327

Hocuk, S., Cazaux, S. & Spaans, M., 2014, *The impact of freeze-out on collapsing molecular clouds*, *MNRAS*, **438**, L56

Hornik, K., 1991, *Approximation Capabilities of Multilayer Feedforward Networks*, *Neural Netw.*, **4**(2), 251
URL: *http://dx.doi.org/10.1016/0893-6080(91)90009-T*

Hoyle, B., Rau, M. M., Zitlau, R., Seitz, S. & Weller, J., 2014, *Feature importance for machine learning redshifts applied to SDSS galaxies*, *ArXiv e-prints*

Hyndman, R. J., 1996, *Computing and Graphing Highest Density Regions*, **50**(2), 120
URL: *http://www.jstor.org/stable/2684423*

Idier, J., 2008, *Bayesian Approach to Inverse Problems / Idier, Jérôme. ; Approche Bayésienne Pour Les Problèmes Inverses.; English.*, Digital signal and image processing series; (London : ISTE ; Hoboken, NJ : Wiley)

Ioppolo, S., 2010, *Surface formation routes of interstellar molecules*, Ph.D. thesis, Ph. D. thesis, University of Leiden (2010)

Ioppolo, S., Fuchs, G. W., Bisschop, S. E., van Dishoeck, E. F. & Linnartz, H., 2007, in *Molecules in Space and Laboratory*

Ioppolo, S., Palumbo, M. E., Baratta, G. A. & Mennella, V., 2009, *Formation of interstellar solid CO{_2} after energetic processing of icy grain mantles*, *A&A*, **493**, 1017

Isella, A., Carpenter, J. M. & Sargent, A. I., 2009, *Structure and Evolution of Pre-main-sequence Circumstellar Disks*, *ApJ*, **701**, 260

Jain, A. K., Murty, M. N. & Flynn, P. J., 1999, *Data Clustering: A Review*

Johnstone, D., Rosolowsky, E., Tafalla, M. & Kirk, H., 2010, *Dense Gas Tracers in Perseus: Relating the N_2H^+, NH_3, and Dust Continuum Properties of Pre- and Protostellar Cores*, *ApJ*, **711**, 655

Kovács, A. & Szapudi, I., 2014, *Star-galaxy separation strategies for WISE-2MASS all-sky infrared galaxy catalogs*, *ArXiv e-prints*

Lebesgue, H. L., 1902, *Integrale, Longueur, aire*, *Bernandon de C. Rebeschini*

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. & Steinacker, J., 2014, *Dust properties inside molecular clouds from coreshine modeling and observations*, *A&A*, **572**, A20

MacKay, D. J. C., 2002, *Information Theory, Inference & Learning Algorithms* (Cambridge University Press, New York, NY, USA)

Makrymallis, A. & Viti, S., 2014, *Understanding the Formation and Evolution of Interstellar Ices: A Bayesian Approach*, *ArXiv e-prints*

Marquardt, D. W., 1963, *An algorithm for least-squares estimation of nonlinear parameters*, *SIAM Journal on Applied Mathematics*, **11**(2), 431
URL: *http://dx.doi.org/10.1137/0111030*

McLachlan, G. J. & Peel, D., 2000, *Finite mixture models* (Wiley Series in Probability and Statistics, New York)

Millar, T. J., Farquhar, P. R. A. & Willacy, K., 2000, *UMIST database. 1995 (Millar+, 1997)*, *VizieR Online Data Catalog*, **412**, 10139

Myers, P. C. & Benson, P. J., 1983, *Dense Cores in Dark Clouds - Part Two - NH3 Observations and Star Formation*, *Rev. Mexicana Astron. Astrofis.*, **7**, 238

Ng, A. Y., Jordan, M. I. & Weiss, Y., 2001, in *Advances in Neural Information Processing Systems* (MIT Press), pp. 849–856

Öberg, K. I., Boogert, A. C. A., Pontoppidan, K. M., van den Broek, S., van Dishoeck, E. F., Bottinelli, S., Blake, G. A. & Evans, II, N. J., 2011, *The Spitzer Ice Legacy: Ice Evolution from Cores to Protostars*, *ApJ*, **740**, 109

Öberg, K. I., van Dishoeck, E. F. & Linnartz, H., 2009, *Photodesorption of ices I: CO, N_2, and CO_2*, *A&A*, **496**, 281

Occhiogrosso, A., Viti, S., Ward, M. D. & Price, S. D., 2012, *Modelling of c-C_2H_4O formation on grain surfaces*, *MNRAS*, **427**, 2450

Peth, M., Lotz, J. M., Freeman, P. E., McPartland, C. & CANDELS Collaboration, 2014, in *American Astronomical Society Meeting Abstracts*, volume 223 of *American Astronomical Society Meeting Abstracts*, p. 145.04

Pirronello, V., Liu, C., Shen, L. & Vidali, G., 1997, *Laboratory Synthesis of Molecular Hydrogen on Surfaces of Astrophysical Interest*, *ApJL*, **475**, L69

Pontoppidan, K. M., 2006, *Spatial mapping of ices in the Ophiuchus-F core. A direct measurement of CO depletion and the formation of CO_2*, *A&A*, **453**, L47

Pontoppidan, K. M., van Dishoeck, E. F., Dartois, E., Fraser, H. J., Banhidi, Z., Jørgensen, J. K. & c2d Team, 2005, in D. C. Lis, G. A. Blake & E. Herbst (eds.), *Astrochemistry: Recent Successes and Current Challenges*, volume 231 of *IAU Symposium*, pp. 319–320

Rasmussen, C. E. & Williams, C. K. I., 2005, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (The MIT Press)

Rawlings, J. M. C., Hartquist, T. W., Menten, K. M. & Williams, D. A., 1992, *Direct diagnosis of infall in collapsing protostars. I - The theoretical identification of molecular species with broad velocity distributions*, *MNRAS*, **255**, 471

Roberts, J. F., Rawlings, J. M. C., Viti, S. & Williams, D. A., 2007, *Desorption from interstellar ices*, *MNRAS*, **382**, 733

Rumelhart, D. E., Hinton, G. E. & Williams, R. J., 1988 (MIT Press, Cambridge, MA, USA), chapter Learning Representations by Back-propagating Errors, pp. 696–699
URL: *http://dl.acm.org/citation.cfm?id=65669.104451*

Sakoe, H. & Chiba, S., 1990 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA), chapter Dynamic Programming Algorithm Optimization for Spoken Word Recognition, pp. 159–165
URL: *http://dl.acm.org/citation.cfm?id=108235.108244*

Schöier, F. L., Jørgensen, J. K., van Dishoeck, E. F. & Blake, G. A., 2002, *Does IRAS 16293-2422 have a hot core? Chemical inventory and abundance changes in its protostellar environment*, *A&A*, **390**, 1001

Sibson, R., 1973, *SLINK: An Optimally Efficient Algorithm for the Single-Link Cluster Method*, *The Computer Journal*, **16**, 30

Smith, I. W. M., Herbst, E. & Chang, Q., 2004, *Rapid neutral-neutral reactions at low temperatures: a new network and first results for TMC-1*, *MNRAS*, **350**, 323

Sofia, U. J. & Meyer, D. M., 2001, *Interstellar Abundance Standards Revisited*, *ApJL*, **554**, L221

Szalay, A. S., Gray, J. & Vandenberg, J., 2002, in *American Astronomical Society Meeting Abstracts*, volume 34 of *Bulletin of the American Astronomical Society*, p. 134.06

Tielens, A. G. G. M., 2010, *The Physics and Chemistry of the Interstellar Medium*

Tielens, A. G. G. M., 2013, *The molecular universe*, *Reviews of Modern Physics*, **85**, 1021

Tielens, A. G. G. M. & Hagen, W., 1982, *Model calculations of the molecular composition of interstellar grain mantles*, *A&A*, **114**, 245

Vasyunin, A. I., Semenov, D. A., Wiebe, D. S. & Henning, T., 2009, *A Unified Monte Carlo Treatment of Gas-Grain Chemistry for Large Reaction Networks. I. Testing Validity of Rate Equations in Molecular Clouds*, *ApJ*, **691**, 1459

Viti, S., Collings, M. P., Dever, J. W., McCoustra, M. R. S. & Williams, D. A., 2004, *Evaporation of ices near massive stars: models based on laboratory temperature programmed desorption data*, *MNRAS*, **354**, 1141

Viti, S. & Williams, D. A., 1999, *Time-dependent evaporation of icy mantles in hot cores*, *MNRAS*, **305**, 755

Wakelam, V., Cuppen, H. M. & Herbst, E., 2013, *Astrochemistry: Synthesis and Modelling*, *ArXiv e-prints*

Wakelam, V., Herbst, E. & Selsis, F., 2006, *The effect of uncertainties on chemical models of dark clouds*, *A&A*, **451**, 551

Watanabe, N. & Kouchi, A., 2002, *Efficient Formation of Formaldehyde and Methanol by the Addition of Hydrogen Atoms to CO in H_2O-CO Ice at 10 K*, *ApJL*, **571**, L173

Watanabe, N., Nagaoka, A., Hidaka, H. & Kouchi, A., 2005, in *Protostars and Planets V Posters*, p. 8244

Whittet, D. C. B., Cook, A. M., Herbst, E., Chiar, J. E. & Shenoy, S. S., 2011, *Observational Constraints on Methanol Production in Interstellar and Preplanetary Ices*, *ApJ*, **742**, 28

Williams, D. A. & Viti, S., 2014, *Observational Molecular Astronomy*

Yorke, H. W. & Sonnhalter, C., 2002, *On the Formation of Massive Stars*, *ApJ*, **569**, 846

Zhen, J. & Linnartz, H., 2014, *UV-induced photodesorption and photochemistry of O_2 ice*, *MNRAS*, **437**, 3190

I felt once more how simple and frugal a thing is happiness: a glass of wine, a roast chestnut, a wretched little brazier, the sound of the sea. Nothing else.

Nikos Kazantzakis