Using statistical emulation and knowledge of grain-surface diffusion for bayesian inference of reaction rate parameters: an application to a glycine network

Heyl, J.; Holdship, J.R.; Viti, S.

The Astrophysical Journal, 931:26 (16pp), 2022 May 20 https://doi.org/10.3847/1538-4357/ac6606

Using Statistical Emulation and Knowledge of Grain-surface Diffusion for Bayesian Inference of Reaction Rate Parameters: An Application to a Glycine Network

Johannes Heyl[1], Jonathan Holdship[2,1], and Serena Viti[2,1]
[1] Department of Physics and Astronomy, University College London, Gower Street, WC1E 6BT, London, UK; johannes.heyl.19@ucl.ac.uk
[2] Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands
Received 2021 June 20; revised 2022 April 3; accepted 2022 April 7; published 2022 May 20

Abstract

There exists much uncertainty surrounding interstellar grain-surface chemistry. One of the major reaction mechanisms is grain-surface diffusion for which the binding energy parameter for each species needs to be known. However, these values vary significantly across the literature which can lead to debate as to whether or not a particular reaction takes place via diffusion. In this work we employ Bayesian inference to use available ice abundances to estimate the reaction rates of the reactions in a chemical network that produces glycine. Using this we estimate the binding energy of a variety of important species in the network, by assuming that the reactions take place via diffusion. We use our understanding of the diffusion mechanism to reduce the dimensionality of the inference problem from 49 to 14, by demonstrating that reactions can be separated into classes. This dimensionality reduction makes the problem computationally feasible. A neural network statistical emulator is used to also help accelerate the Bayesian inference process substantially. The binding energies of most of the diffusive species of interest are found to match some of the disparate literature values, with the exceptions of atomic and diatomic hydrogen. The discrepancies between these two species are related to the limitations of the physical and chemical models. However, the use of a dummy reaction of the form $H + X \longrightarrow HX$ is found to somewhat reduce the discrepancy with the binding energy of atomic hydrogen. Using the inferred binding energies in the full gas–grain version of UCLCHEM results in almost all the molecular abundances being recovered.

Unified Astronomy Thesaurus concepts: Astrostatistics strategies (1885); Bayesian statistics (1900); Reaction rates (2081); Astrochemistry (75); Dark interstellar clouds (352); Interstellar abundances (832)

1. Introduction

Interstellar dust plays a very significant role in the rich chemistry that is produced in the interstellar medium. In fact, it is widely believed that complex organic molecules (COMs) form on interstellar dust (Herbst & van Dishoeck 2009; Caselli & Ceccarelli 2012). In certain cases, grain-surface reactions are more efficient than gas-phase reactions, due to the dust grains acting as energy sinks. However, there exists much debate about the stage of star formation during which these molecules are produced. While modeling has shown that molecules such as glycine can be formed during the warm-up phase of star formation (Garrod 2013), there is also evidence that suggests that dark interstellar cloud conditions would suffice (Ioppolo et al. 2020).

Bayesian inference can be used to estimate reaction rate parameters using observations. While this tool has become a staple in many areas of astrophysics, it is only recently that it has found use cases in astrochemistry (Makrymallis & Viti 2014; Holdship et al. 2018; de Mijolla et al. 2019; Heyl et al. 2020). In Holdship et al. (2018), reaction rates were inferred using a toy network. In Heyl et al. (2020), the topology of this network was also considered, specifically the placement of constraints within the network. Both of these works considered the rates of the reactions, without considering the actual, underlying reaction mechanisms. However, it was noted in both works that the paucity of grain-surface species abundances means that many of the reaction rates will remain undetermined, due to the high levels of degeneracy. This work seeks to circumvent this issue by using the physics of the grain-surface diffusion mechanism to reduce the number of free parameters and therefore break this degeneracy.

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

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

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

We begin by first explaining the chemical code and network that will be used in this work in Section 2. Additionally, we describe the grain-surface diffusion mechanism that lies at the heart of our investigation. We will explain how we can make approximations regarding a species' mobility to estimate the binding energy of said species. In Section 3, we will discuss how statistical emulation can be leveraged to accelerate the running of the forward model, before describing how Bayesian inference will tie all of this together in Section 4. Following this, we will present the resulting binding energies estimates using this method in Section 5. In Section 6, we then look to see how well we are able to recover abundances when we run a full gas–grain chemical code using the estimated values.

2. The Chemical Code and Network

2.1. The Chemical Model and Code

The code that was used was based on the gas–grain chemical code UCLCHEM (Holdship et al. 2017).

The surface chemistry is modeled through the rate equation approach. The code has to solve a system of coupled ordinary differential equations of the form

$$\frac{dn_i}{dt} = \sum_{l,m} k_{lm}^{i} n_l n_m - n_i \sum_{i \neq r} k_r n_r - k_i^{\mathrm{des}} n_i + k_i^{\mathrm{ads}} n_{i,\mathrm{gas}}, \quad (1)$$

where $k_{lm}^{i}$is the reaction rate of the reactions between species l and m to produce species i, n_i is the abundance of species i, k_r represents the reaction rates of all reactions where species i is consumed as a reactant, and k_i^{des} and k_i^{ads} represent the desorption and adsorption rates of the species. The coupled differential equations represent the formation and destruction mechanisms for all the relevant species.

However, in order to reduce the run time of the inference process, some changes had to be made to reduce the time taken for UCLCHEM to run. These are described in detail in Holdship et al. (2018) but are outlined briefly here.

The code that was used considered only grain-surface chemistry to reduce the complexity of the system of coupled ordinary differential equations. However, it was important to still include the key processes that couple the gas and grain chemistry. It should be noted that the final two terms in Equation (1) represent the net flux of gas-phase molecules adsorbing into the grain surface. As such, if one only wishes to consider grain-surface chemistry, then one just needs to parameterize this net "freeze-out." The net freeze-out was found by running a single point model of the full gas–grain version UCLCHEM. The net movement of each species between the gas and grain phases as a function of time was then extracted. Only the species which were deposited in abundances relative to n_H greater than 10^{-7} on the grains were included. These species were: H, O, OH, C, CO, N, CH, and CH_3. These were all species that would form in the gas phase and were involved in the reactions listed in Table 1. The freeze-out rates were inserted as source terms into the grain-surface models. The freeze-out of the more complex species was not considered, as these species were unlikely to form in the gas phase at 10 K. The advantage of doing this is that one avoided

Table 1
Reactions Used in this Work

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

Note. Taken from Ioppolo et al. (2020) and Linnartz et al. (2015). The values of $\frac{E_b}{E_D}$ used for the more mobile species for each diffusion-based reaction are given. Reactions 34 and 35 are not assumed to be diffusive reactions.

needing to consider the system of ODEs for gas-phase reactions, thereby significantly reducing the computational complexity.

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

2.2. The Chemical Network

Our network is composed of radicals that react to form glycine, the simplest amino acid. The reactions that make up this chemical network are shown in Table 1. The grain-surface network used in this work is based on the one used in Ioppolo et al. (2020) with the final two reactions being taken from Linnartz et al. (2015). In Ioppolo et al. (2020), laboratory and chemical modeling found that the first 47 reactions were able to produce glycine in dark interstellar conditions, long before the warm-up phase of star formation, without requiring any energetic input (Ioppolo et al. 2020). This is in contrast to previous work that assumed that the formation of glycine required an increased temperature as well as energetic processing (Garrod 2013). Based on Ioppolo et al. (2020), it was expected that this network would be sufficient to learn about COMs in the prestellar phase with the help of observed abundances. Reactions 48 and 49 were included, as they involved species already present in the network. Furthermore, one of the end products, NH_4^+, had a constraint on its abundance that could be used for the Bayesian inference to further constrain the parameters.

2.3. Grain-surface Chemistry

2.3.1. Grain-surface Diffusion

An understanding of the actual grain-surface mechanisms will prove crucial in this work. The diffusion mechanism described in Hasegawa et al. (1992) was implemented in UCLCHEM in Quénard et al. (2018).

According to the mechanism, the rate at which two species A and B react via diffusion is given by

$$k_{\rm AB} = \kappa_{\rm AB}\frac{(k_{\rm hop}^A + k_{\rm hop}^B)}{N_{\rm site} n_{\rm dust}}, \tag{2}$$

where $N_{\rm site}$ is the number of sites on the grain surface and $n_{\rm dust}$ is the number density of dust grains.

In Equation (2), $k_{\rm hop}^X$ is the thermal hopping rate of species X on the grain surface defined as

$$k_{\rm hop}^X = \nu_0 \exp\left(-\frac{E_b}{T_{\rm gr}}\right), \tag{3}$$

where E_b is the diffusion energy of the species, $T_{\rm gr}$ is the grain temperature and ν_0 is the characteristic vibration frequency of species X. The diffusion energy is typically taken to be a fraction of the species binding energy, E_D. There is debate surrounding the value of the fraction $\frac{E_b}{E_D}$, though there is agreement that it should be in the range 0.3–0.8, with lower values in this range being more appropriate for stable molecules (Penteado et al. 2017). However, it has been found that for O and N atoms, a ratio of 0.55 is more suitable (Minissale et al. 2016a). In this work, we follow the convention adopted by Jin & Garrod (2020) where the ratio was set to equal 0.6 for atomic species and 0.35 for stable species. For all other species, a value of 0.5 was used. For each reaction, the value of $\frac{E_b}{E_D}$ for the more mobile species is given in Table 1. The reason we only consider the value of this ratio for the more mobile species is given in Section 4.5. The characteristic vibration frequency, ν_0, is defined as

$$\nu_0 = \sqrt{\frac{2k_b n_s E_D}{\pi^2 m}}, \tag{4}$$

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

Finally, $\kappa_{\rm AB}$, which provides the reaction probability, is taken to be

$$\kappa_{\rm AB} = \max\left(\exp\left(-\frac{2a}{\hbar}\sqrt{2\mu k_b E_A}\right), \exp\left(-\frac{E_A}{T_{\rm gr}}\right)\right), \tag{5}$$

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

2.3.2. Reaction–diffusion Competition

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

$$\kappa_{\rm AB}^{\rm final} = \frac{p_{\rm reac}}{p_{\rm reac} + p_{\rm diff} + p_{\rm evap}}, \tag{6}$$

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

$$p_{\rm reac} = \max(\nu_0^A, \nu_0^B)\kappa_{\rm AB}, \tag{7}$$

$$p_{\rm diff} = k_{\rm hop}^A + k_{\rm hop}^B \tag{8}$$

and

$$p_{\rm evap} = \nu_0^A \exp\left(-\frac{E_D^A}{T_{\rm gr}}\right) + \nu_0^B \exp\left(-\frac{E_D^B}{T_{\rm gr}}\right). \tag{9}$$

In Equation (2), $\kappa_{\rm AB}$ is replaced with $\kappa_{\rm AB}^{\rm final}$.

At 10 K, we observe that the rate of evaporation is far lower than the rates of diffusion and reaction, so will be neglected throughout this work.

As most of the reactions in Table 1 are radical–radical, it was assumed that their activation energies were 0 K. Even for reactions involving a known reaction barrier, such as reaction 5, it was found that $p_{\rm reac} \gg p_{\rm diff}$, which means the activation

energy barrier is lower than the diffusion barrier. As such, $\kappa_{\rm AB}^{\rm final} \simeq 1$. This "diffusion-limited regime" corresponds to the situation where the diffusion process is the rate-limiting step and is due to the fact that the temperature being considered is 10 K.

3. Statistical Emulation

Statistical emulation involves fitting a statistical function to match the inputs and outputs of a forward model (Grow & Hilton 2018). The advantage in doing so is that one replaces the slow-to-evaluate forward model with the fitted emulator in order to save time. This becomes particularly significant when multiple evaluations of the forward model are required, such as in Bayesian inference which typically involves calling the forward model hundreds of thousands of times. Statistical emulators have primarily been used in the past in cosmology (Auld et al. 2007; Schmit & Pritchard 2017; Rogers et al. 2019; Wang et al. 2020), but have also recently found use in astrochemistry (de Mijolla et al. 2019; Holdship et al. 2021). In our case, the forward model requires solving a coupled system of ODEs of the form given in Equation (1). The evaluation of the forward model can be time consuming, especially if this has to be repeated multiple times as would be the case for Bayesian inference. This proves to be particularly important for the analysis we do in Appendices A and B. In this case, a statistical emulator would be particularly useful, as it can interpolate within the range of input values considered. This is computationally faster than making use of the original forward model for evaluation.

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

3.1. Training the Emulator

In order to be able to use the emulator, it must first be trained on some data. It is important that the sampling is done in such a way that the entire parameter space is explored. One cannot simply use random uniform sampling, as each point is drawn independently of the others. This can result in the training points being clustered. This has the consequence of the emulator attempting to match the training data more in these regions, thereby introducing bias in other less well-covered regions of the parameter space. A Latin hypercube sampling scheme was used (McKay et al. 1979) and implemented using the Python surrogate modeling toolbox (Bouhlel et al. 2019). As both the input and output parameters span several orders of magnitude, the emulator was trained to learn the mapping between the logarithm of these two. The training data set spanned the prior range for each parameter. Given that a log-uniform prior between 10^{-15} and 10^{0} was used for the Bayesian inference (see Section 4.2 for details), this ensured that any conceivable input to the emulator from the inference was within the prior range, as outside that range the posterior is zero due to the prior being zero. The parameter ranges defined the range of values over which the emulator could interpolate. The emulator was not needed to extrapolate, as the range of the prior was covered.

Choosing the number of training points is a crucial parameter. It is clear that increasing the amount of training data will improve the emulator's performance. However, this will also result in the time taken for training increasing. As such, a balance needs to be struck. Figure 1 shows the mean-squared error (MSE) on a test set as a function of the number of training points. It was found that using 150,000 training points was sufficient. By evaluating these points on a single Research Capital Investment Fund node with 40 cores, the training time was about 30 minutes.

3.2. The Neural Network

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

4. Bayesian Inference

4.1. Introduction to Bayesian Inference

The aim of this work is to deduce the reaction rates of the reactions in this network, which we represent as a vector, $\boldsymbol{k} = (k_1,\ k_2 ... k_{49})$, and use these inferred reaction rates to determine the binding energies of diffusive species. This is initially a 49-dimensional inference problem. The code used takes this vector as an input and outputs the abundances of all the species in this network, which is represented by the vector $\boldsymbol{Y} = (Y_1,\ Y_2 ... Y_{35})$. There exist measurements for the abundances of a subset of the molecules in this network. These form the data $\boldsymbol{d}$, which are listed in Table 2.

Bayes' Law can be used to determine the probability distribution of the reaction rates given the data

$$P(\boldsymbol{k}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{k})P(\boldsymbol{k})}{P(\boldsymbol{d})}, \tag{10}$$

where $P(\boldsymbol{k}|\boldsymbol{d})$ is the posterior probability distribution, $P(\boldsymbol{k})$ is the prior, $P(\boldsymbol{d}|\boldsymbol{k})$ is the likelihood, and $P(\boldsymbol{d})$ is referred to as the evidence. The prior distribution encodes our initial knowledge of the values of the reaction rates. The likelihood provides the likelihood of the data as a function of the reaction rates. The likelihood provides information about the physical model under consideration. The evidence serves as a normalizing factor, as it

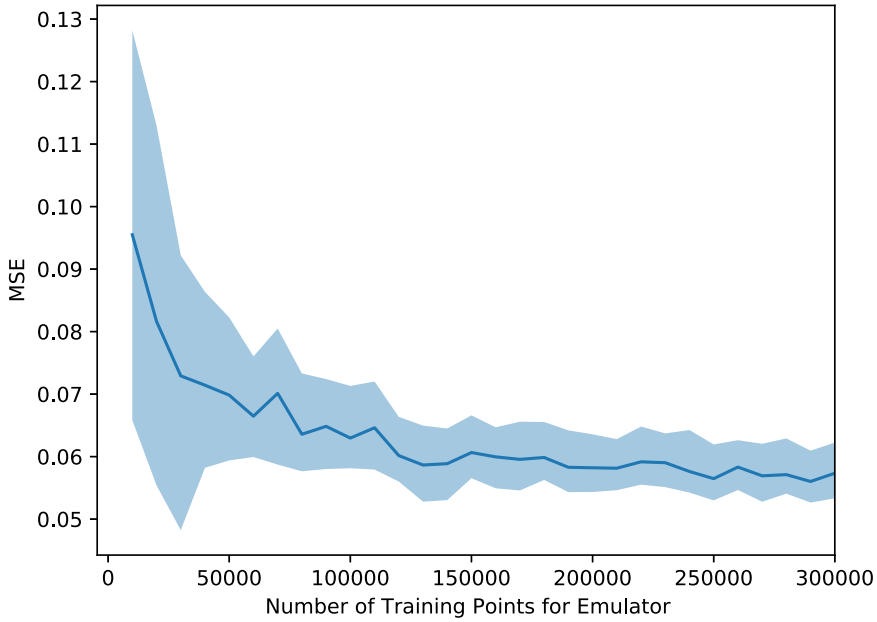

Figure 1. A plot of the mean-squared error of the emulator as a function of the number of training points used to train the emulator. The shaded area represents the 95% confidence interval around the mean-squared error.

Table 2
Abundances and Uncertainties

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

Note. Constraints used adapted from Boogert et al. (2015). There were two distinct values for the upper limit on the abundance of O_2, so the higher one was selected.

represents the marginalized likelihood. The posterior distribution represents the updated probability distribution of reaction rates given the data, information encoded in the prior distribution, and the physical model.

4.2. Implementation

To obtain the posteriors of the reaction rates, a prior must be specified. As has been done previously, a log-uniform prior was chosen, so as to equally weight rates over different orders of magnitude. However, a different range is chosen compared to Holdship et al. (2018) and Heyl et al. (2020) to accommodate the fact that the reaction rates, $\boldsymbol{k}$, are normalized by the cloud density. Additionally, it was found by Holdship et al. (2018) and Heyl et al. (2020) that the probability density is very low in the range 10^{-30}–10^{-15}. As such, a log-uniform prior between 10^{-15} and 10^{0} was used.

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

$$P(\boldsymbol{d}|\boldsymbol{k}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(d_i - Y_i)^2}{2\sigma_i^2}\right), \tag{11}$$

where n_d is the number of observations and σ_i is the uncertainty of the ith observation. Only the species for which there are abundances are multiplied over. Table 2 contains species for which we have abundances with Gaussian uncertainties. Observed abundances will be referred to as constraints in this work as they constrain the prior parameter space of reaction rate posteriors.

Boogert et al. (2015) also contains upper limits for the abundances of some species of interest. The upper limits for O_2, N_2, H_2O_2, and glycine are also included in Table 2. Equation (11) can be rewritten to account for these upper limits, as was done in Holdship et al. (2018).

$$P(\boldsymbol{d}|\boldsymbol{k}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\delta_i \frac{(d_i - Y_i)^2}{2\sigma_i^2}\right)(1 - S(C_i))^{1-\delta_i}, \tag{12}$$

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

$$S(C_i) = 1 - \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{C_i - Y_i}{\sigma_i^{\mathrm{UL}}}\right)\right), \tag{13}$$

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

there might be some level of uncertainty on the value of the upper limit.

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

4.3. Degeneracy Problem

Before performing the inference, it is important to consider the problem in more depth. There are 49 parameters to estimate, but there are only 12 measurements. As was observed in Holdship et al. (2018) and Heyl et al. (2020), having far more parameters than constraints introduces a significant amount of degeneracy into the problem. Some rates of reactions do not influence the abundances of species with constraints. As was observed in Holdship et al. (2018), this will result in the majority of reaction rate posteriors being uniform. Additionally, many posteriors will only deviate weakly from uniformity. This was found to be the case for linear reaction chains with successive hydrogenations, such as the successive hydrogenation of CO to form methanol. The degeneracy stemmed from the fact that the reactions were tightly coupled. Provided one rate took a minimum value and acted as the rate-limiting step, the other reaction rate was free to vary above this minimum rate. The high level of degeneracy inherent to this problem meant that despite running a sampler for several weeks, it never converged.

4.4. Degeneracy Solution

To reduce the degeneracy of the problem, one can exploit information about the underlying grain-surface diffusion mechanism. Ultimately, the reaction rate is strongly dependent on the hopping rates of the reactant species, which, assuming the grain temperature is constant, implies that the reaction rate is set by the binding energies. Given the strong dependence of the hopping rate on the binding energy, it is clear that a small difference in the binding energy between two species will mean that the hopping rate of the more mobile species (the one with the lower binding energy) will dominate the reaction rate. In Equation (2), this corresponds to $k_{\rm hop}^{A} \gg k_{\rm hop}^{B}$ and yields

$$k_{\rm AB} = \kappa_{\rm AB}^{\rm final} \frac{k_{\rm hop}^{A}}{N_{\rm site} n_{\rm dust}}, \tag{14}$$

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

Based on this, one can separate reactions into various classes, depending on which of their reactants is more mobile. Even though the actual values of the binding energies will differ across the literature (see Penteado et al. (2017), for a discussion on this), most works agree on the "hierarchy" of mobility—that is, which of the two species is more mobile. By making an assumption or by considering literature values, one can make a decision on which species should be treated as more mobile. In this work, the more mobile species was assumed to be the one with the lower binding energy in at least two of Penteado et al. (2017), UMIST, and Wakelam et al. (2017). The groupings used are shown in Table 3. For reactions 34 and 35, one does not expect diffusion to be the dominant reaction mechanism. However, since this is the diffusion-limited regime, one can assume that these two reactions will have the same reaction rate.

Table 3
Main Reaction Groupings

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

Note. Groupings are separated by the molecule that the literature suggested was more dominant. Any reaction not included in this table had its reaction rate inferred separately.

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

4.5. Deriving the Binding Energies

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

Our goal is to determine the binding energies of various species. In this work, we will be inferring reaction rates for the various reactions and use these to solve for the binding energies. The quantity $\boldsymbol{k}$ varies as a function of time, as seen in Equation (2) due to the dependence on the total hydrogen number density. However, by multiplying by n_H on both sides, one obtains

$$k'_{\rm AB} = k_{\rm AB} n_H = \kappa_{\rm AB} \frac{(k_{\rm hop}^{A} + k_{\rm hop}^{B})}{N_{\rm site} \frac{n_{\rm dust}}{n_H}}, \tag{15}$$

where $\frac{n_{\rm dust}}{n_H}$ is a constant. Note now that the expression on the right-hand side only consists of constants. This implies that $k'_{\rm AB}$ is constant with respect to the density of the cloud. While $k'_{\rm AB}$ can still be interpreted as a reaction rate, it has units of $\rm s^{-1}$.

Due to the exponential dependence of the hopping rates on the binding energies, one finds that in most cases, one species dominates the reaction rate. If species A has a binding energy of, say, 500 K and species B has a 10% higher binding energy, then A's hopping rate is almost 150 times greater, due to the low grain temperature of 10 K. This difference will only get larger as the binding energies under consideration increase.

Hence, we can state that $k_{\rm hop}^{A} \gg k_{\rm hop}^{B}$ and then determine the binding energy of the species by substituting Equation (3) into Equation (14):

$$k'_{\rm AB}\frac{n_{\rm dust}N_{\rm site}}{n_H\kappa_{\rm AB}}\sqrt{\frac{\pi^2 m}{2k_b n_s}} = \sqrt{\frac{E_b^A}{f}}\exp\left(-\frac{E_b^A}{T_{\rm gr}}\right), \quad (16)$$

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

4.6. Constraints

The final component required to perform Bayesian inference is the data, which in this case would be measured abundances of species. A number of constraints for molecules in this network can be found in Boogert et al. (2015), which provides the median abundance as well as lower and upper quartile. As in Holdship et al. (2018), we assume the measurements are Gaussian distributed, which implies the median is the mean. Additionally, the upper and lower quartiles are 0.68σ from the mean. Using this information, the abundances used in this work are listed in Table 2. We combine measured molecular ice abundances from the dark, quiescent cloud as well as large young stellar objects (LYSOs). We observe that the species CO, CO_2, H_2O, CH_3OH, and NH_4^+ have similar abundances in quiescent clouds and LYSOs. Using this, we assume that other species, which have only been detected in LYSOs, will have broadly similar dark cloud abundances. We argue that while chemistry is expected to happen during the warm-up phase for LYSOs, this will be relatively short lived, and any abundances will likely have been built up during the cold phase of star formation. However, even though the warm-up phase will be shorter, the chemical timescales will decrease due to the reaction rate's dependence on temperature. Overall, while there is justification for using LYSO abundances for dark cloud conditions, it should be noted that we are adding additional uncertainty to our analysis.

5. Results

5.1. Highest Density Regions

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

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

$$R(f_a) = x : f(x) \geqslant f_a, \quad (17)$$

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

5.2. Reaction Rate Marginalized Posteriors

We find that all 14 parameter distributions are nonuniform. As such, this means that we have gained information about the entirety of our 49D reaction network. By exploiting our knowledge of the grain-surface diffusion mechanism and assuming that reaction rates are dominated by the diffusion rates of a subset of molecules, we have been able to significantly reduce the dimensionality of our problem, therefore making it computationally tractable for the sampler. Figures 2 and 3 show the marginalized posterior distributions for the reaction rates with the 65% HDR being the shaded regions when we use the likelihoods expressed in Equations (11) and (12). For all of the posteriors, the 65% HDR lies away from the boundaries of the uniform distribution, implying that our choice of prior was appropriate. We choose not to consider 2D marginalized posterior distributions, due to the fact that the parameters correspond to groups of reactions as opposed to individual reactions. We consider the frequentist properties of the estimators in Appendix A.

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

We observe that the posterior for the reaction rate of hydrogenation is the most constrained in that it rules out more of the prior parameter space than any of the other posteriors do. In Heyl et al. (2020), the lower uncertainty on hydrogen's posterior was related to the size of the constraints on the species formed by hydrogenation, in particular the constraint on water, which is known to have an abundance greater than 0 at the 3.1σ level. It would make sense that this low level of uncertainty on the constraint drives the low uncertainty on the hydrogenation reaction rate posterior, as it penalizes the likelihood function more. In the limit of the uncertainties on the molecular abundances going to zero, one would expect the posterior distribution of the relevant reaction rate to look like a Dirac delta function.

5.3. Binding Energy Posteriors

The advantage of inferring the reaction rates as opposed to directly inferring the species binding energies is that the reaction rate posteriors make no assumption about the exact nature of the reaction mechanism. One can then select specific reactions that one believes occur via diffusion, thereby reducing the dimensionality of the problem. The list of species that were thought to diffuse is listed in Table 4. These are calculated from the reaction rate posteriors by solving Equation (16), with the posteriors shown in Figure 4. In Table 4, these binding energies are compared to the values used in McElroy et al. (2013), Penteado et al. (2017), and Wakelam et al. (2017). We make use of the posteriors obtained using

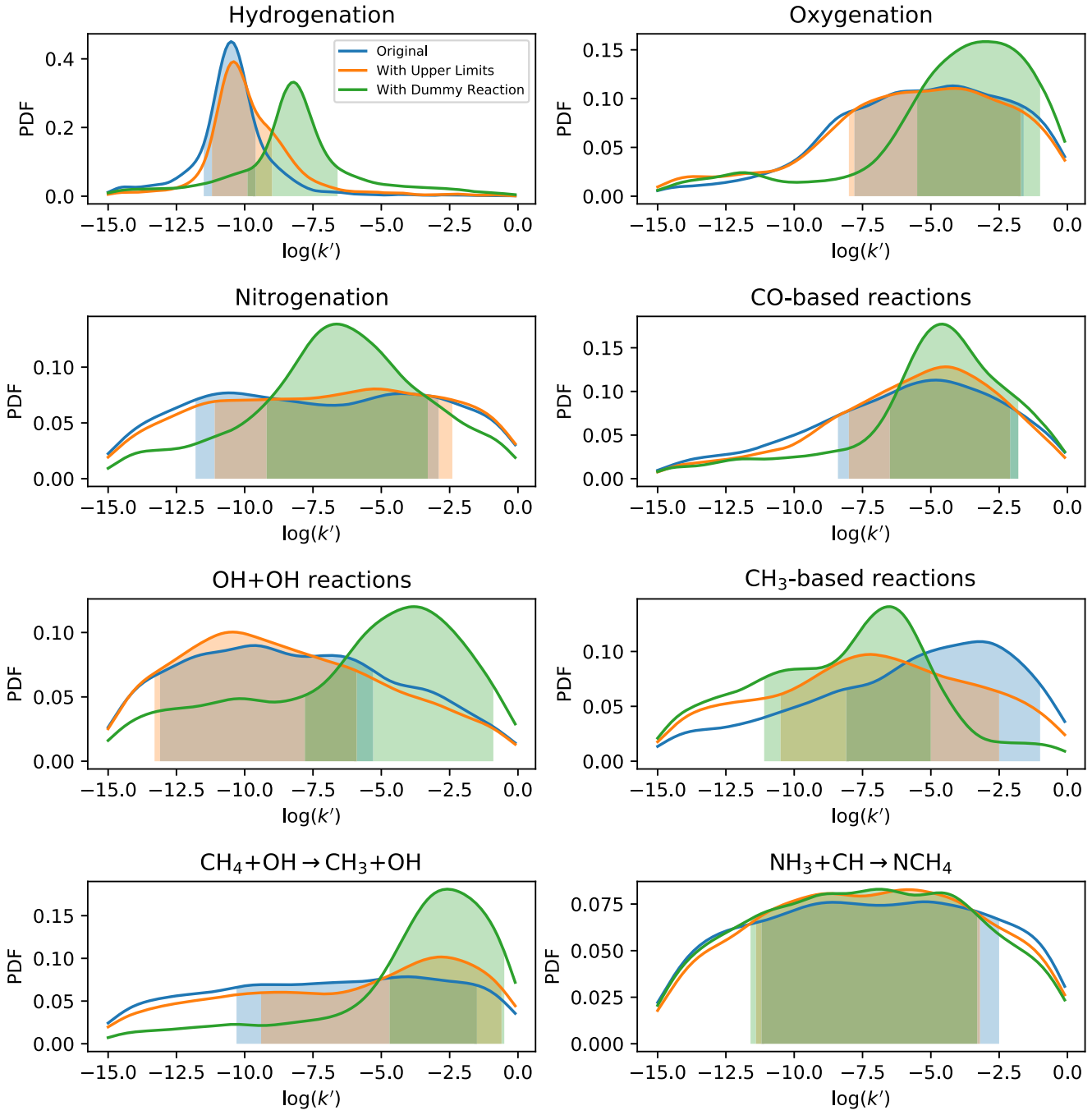

Figure 2. Marginalized posterior distributions for the first eight reaction rate parameters.

Equation (12). This first round of inference is referred to as "Binding Energy 1." We observe that there are no significant differences in the binding energy distributions for most species when we include the upper limits, with the exception of CH_3, for which we see that the inclusion of the upper limits results in a significantly decreased estimated binding energy.

For most of the species for which there are literature binding energies, there is an agreement with at least one literature value and the uncertainty of the values is lower than the spread of literature values. No values for the binding energies of NCH_4 and NH_2CH_3 were found in the literature. The binding energies for O and N were both found to be lower than that of H. This is surprising as the reactions of the species with H were classified as hydrogenations in Table 3.

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

For some of the species, there is a large variance in the posteriors. This can be attributed to the lack of enough constraints in the network. To demonstrate this, the upper limits

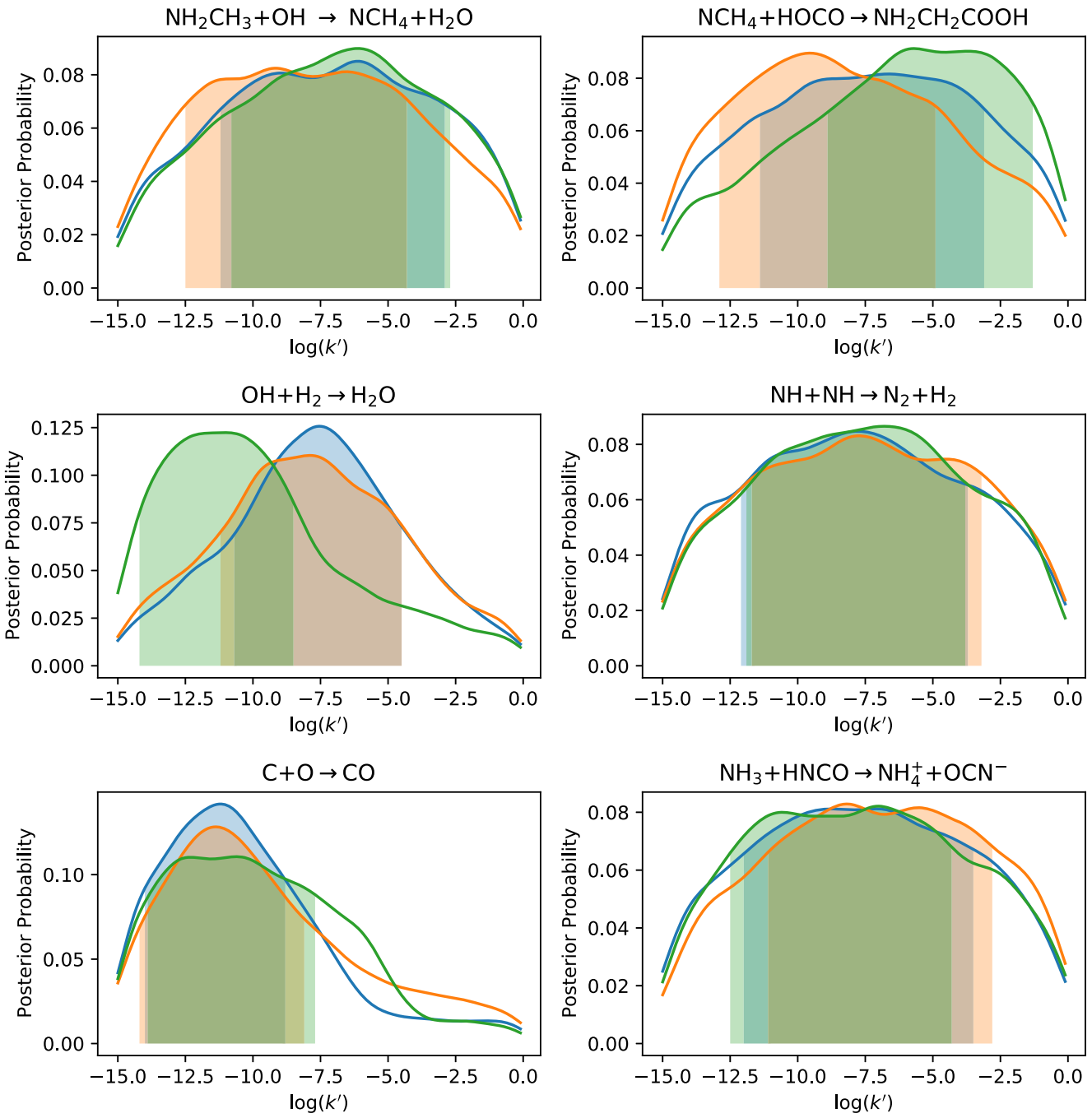

Figure 3. Marginalized posterior distributions for the remaining six reaction rate parameters.

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

It should be noted that even among the literature values, there is not always agreement on the values of the binding energies. The tension in the values can be attributed to varying assumptions made about the grains, such as the ice composition. Additionally, recent work by Bovolenta et al. (2020) and Grassi et al. (2020) suggests that it might be more appropriate to consider binding energy distributions that vary as functions of the individual binding site. In our work, we have assumed that there is a single binding energy value, which implies that the grains are uniform in nature. In reality, this is unlikely to be true and will need to be accounted for in future work.

6. The Binding Energy of Hydrogen

We observe that, despite the high precision of the hydrogenation rate estimate, the binding energy of hydrogen

Table 4
Binding Energies Obtained for Various Species

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

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

is inaccurate and does not match any of the literature values within the error. We now look to address this.

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

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

6.1. Including Chemical Desorption of H_2

The energy released in the reaction of $H + H \longrightarrow H_2$ can cause the product to desorb into the gas phase. An estimate for the fraction of H_2 released was determined by Minissale et al. (2016b) to be

$$\eta_{\rm CD} = \exp\left(-\frac{E_D N_{\rm dof}}{\epsilon_{\rm CD}\Delta H_R}\right), \tag{18}$$

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

$$\epsilon_{\rm CD} = \frac{(m - M)^2}{(m + M)^2}, \tag{19}$$

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

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

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

6.2. Including a Dummy Reaction in the Network

We consider the effect of including the dummy reaction $H + X \longrightarrow HX$ on the entire network. The posteriors for the reaction rates are shown in Figures 2 and 3 with the corresponding binding energy posteriors being shown in Figure 4 and listed in Table 4 as "Binding Energy 2."

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

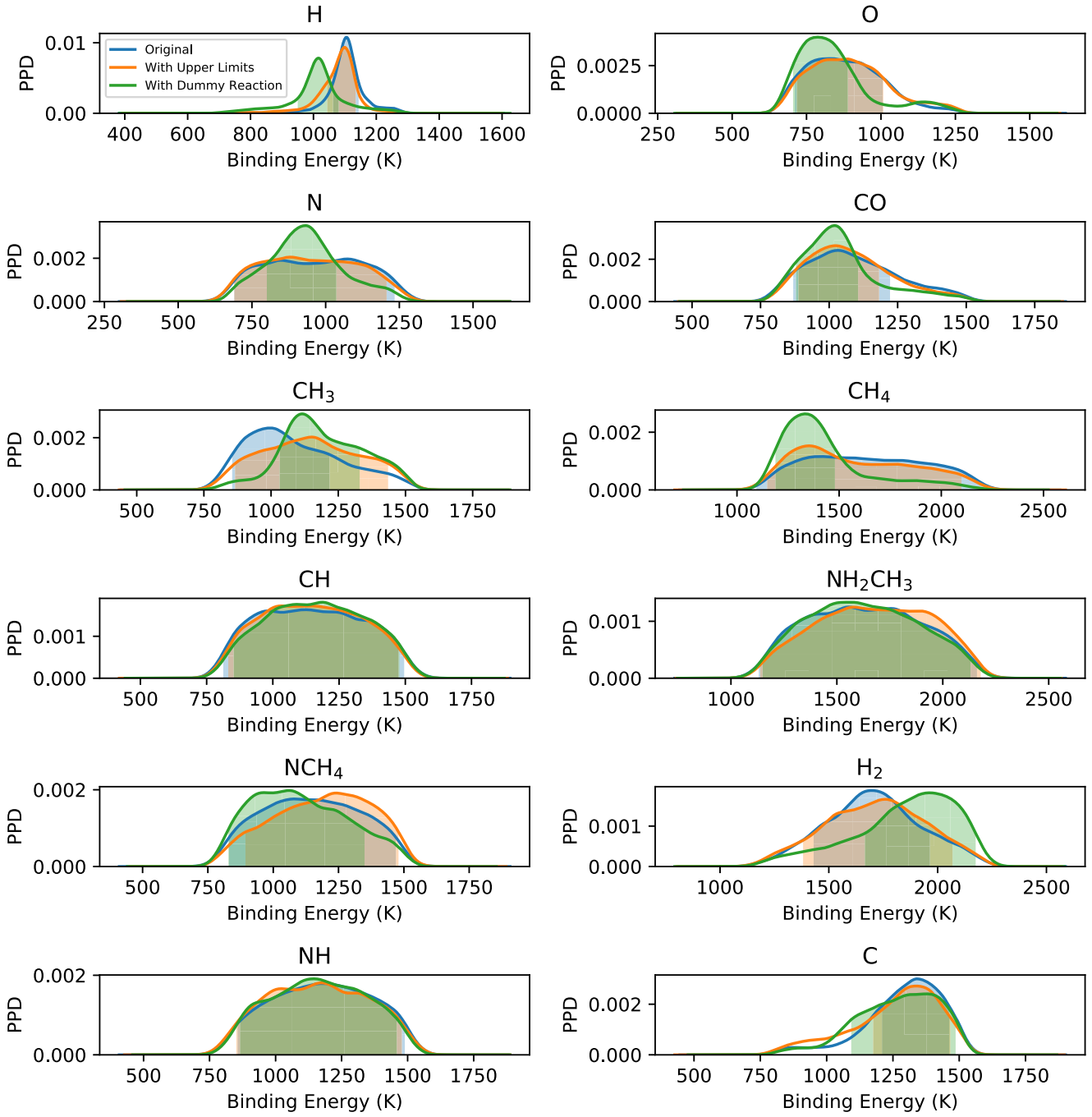

Figure 4. Marginalized posterior probability distributions (PPDs) for the binding energies of the species of interest. The marginalized posterior distributions are also plotted for the case where a dummy reaction for hydrogen is included in the network.

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

7. Application to a Gas–Grain Chemical Code

In this section, we will look to use the binding energies obtained in this work in a full version of the gas–grain chemical code UCLCHEM. This is a form of model checking. To do this, we sample from the binding energy posteriors in Figure 4 and input these into UCLCHEM. We aim to determine how well the abundances of species of interest are recovered when

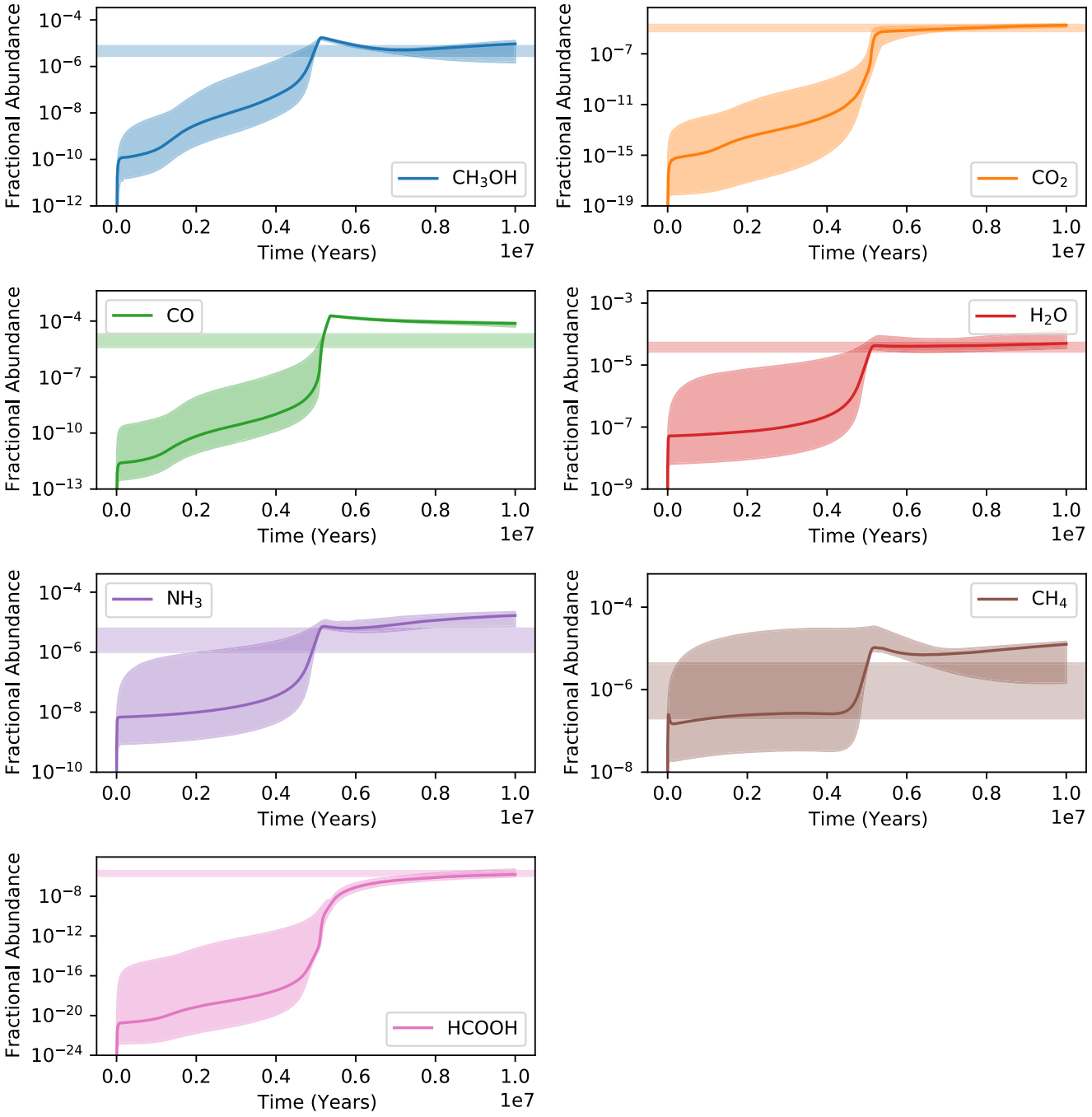

Figure 5. Time series of the fractional abundances for H_2O, CO, CO_2, and CH_3OH. The binding energies for each species were sampled from the marginalized posterior distributions and inputted into the full UCLCHEM code. The horizontal shaded regions are the corresponding measured molecular abundances with their 67% confidence interval. The time series are plotted with their 95% confidence intervals.

the binding energies obtained in this work are inputted into the full gas–grain version of UCLCHEM. Figure 5 shows the time series evolution of the fractional abundances of H_2O, CO, CO_2, CH_3OH, NH_3, CH_4, and HCOOH, with the 95% confidence interval for the time series also shown. These are species from Table 2 that have observed abundances, not simply upper limits that have been converted into weak measurements. The only species with an observed value that has not been included is NH_4^+, but this is not expected to form via diffusion. The purpose of this section is to see how well the inferred binding energy values help in recovering the abundances in a general gas–grain network. We observe that at late times, the final abundances of most of the species become less affected by the binding energies than for earlier times, suggesting that we approach an equilibrium point at a temperature of 10 K. If we were to consider a warmer core, it is likely that our final abundances would be different.

The final abundances of H_2O, CO_2, CH_3OH, CH_4, and HCOOH match the measured values, within the 1σ error. This is not the case for NH_3 and CO. The final abundance for NH_3 is 1.2σ from the mean, which is a relatively small discrepancy. On the other hand, CO's final abundance is 4.6σ from the mean. One possible reason for this is that all the other molecules which are constrained are stable molecules that are

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

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

8. Conclusions

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

This approach yielded binding energy values that were consistent with literature values. The notable exceptions were the binding energies of H and H_2, whose binding energy values were found to be significantly higher than other literature values. This discrepancy was attributed to issues relating to the chemical model used, which decoupled the gas and grain chemistries in the interest of reducing the time taken for the evaluation of the forward model and therefore the time taken for the inference process. While chemical desorption was found to not have a significant effect on the discrepancy, using a dummy reaction of the form $H + X \longrightarrow HX$ to account for all the possible other reactions involving H somewhat reduced the discrepancy, but not enough.

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

In Section 7, we sampled from the obtained binding energy posteriors and input these binding energies into the full gas–grain version of UCLCHEM. We found some agreement between the obtained molecular abundances and the observed values. However, if one wished to infer from UCLCHEM directly, one would need to account for the fact that the inference process would take longer, on account of one evaluation of the full version of UCLCHEM taking of the order of a minute compared to the 0.5 seconds that is typical of the simplified code used in this work. Future work will look to employ statistical emulation to the full version of UCLCHEM to circumvent this problem. Alternative sampling techniques that are adaptive could be utilized to improve emulator performance, even for the emulator in this work (Gramacy & Lee 2008).

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

Further work would also need to address the lack of sufficient abundance data. It is clear that the abundances of more species need to be known in order to better constrain the reaction rate posteriors as well as the binding energy posteriors.

The authors thank the referees for their constructive comments that greatly improved this work. J.H. is funded by an STFC studentship in Data-Intensive Science (grant number ST/P006736/1). This work was also supported by European Research Council (ERC) Advanced Grant MOPPEX 833460. S.V. acknowledges support from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 811312 for the project "Astro-Chemical Origins" (ACO). This work used computing equipment funded by the Research Capital Investment Fund (RCIF) provided by UKRI, and partially funded by the UCL Cosmoparticle Initiative.

Software: UCLCHEM (Holdship et al. 2017), PyMultiNest (Buchner et al. 2014), F2PY (Peterson 2009).

Appendix A
Evaluating the Frequentist Properties of the Bayesian Estimators

In this section, we seek to determine whether the constraints imposed are significantly influencing the resulting posterior distribution. To determine this, we run the forward model using reaction rates drawn from a Gaussian distribution with a mean value equal to the maximum-posterior reaction rate obtained in this work (which represents the "true" reaction rate) and a standard deviation equal to 1. Bayesian inference was then used to recover these reaction rates. This was repeated 20 times using the statistical emulator. The strip plots in Figure 6 show the values of the reaction rates recovered with the associated uncertainties. This analysis is meant to demonstrate to what

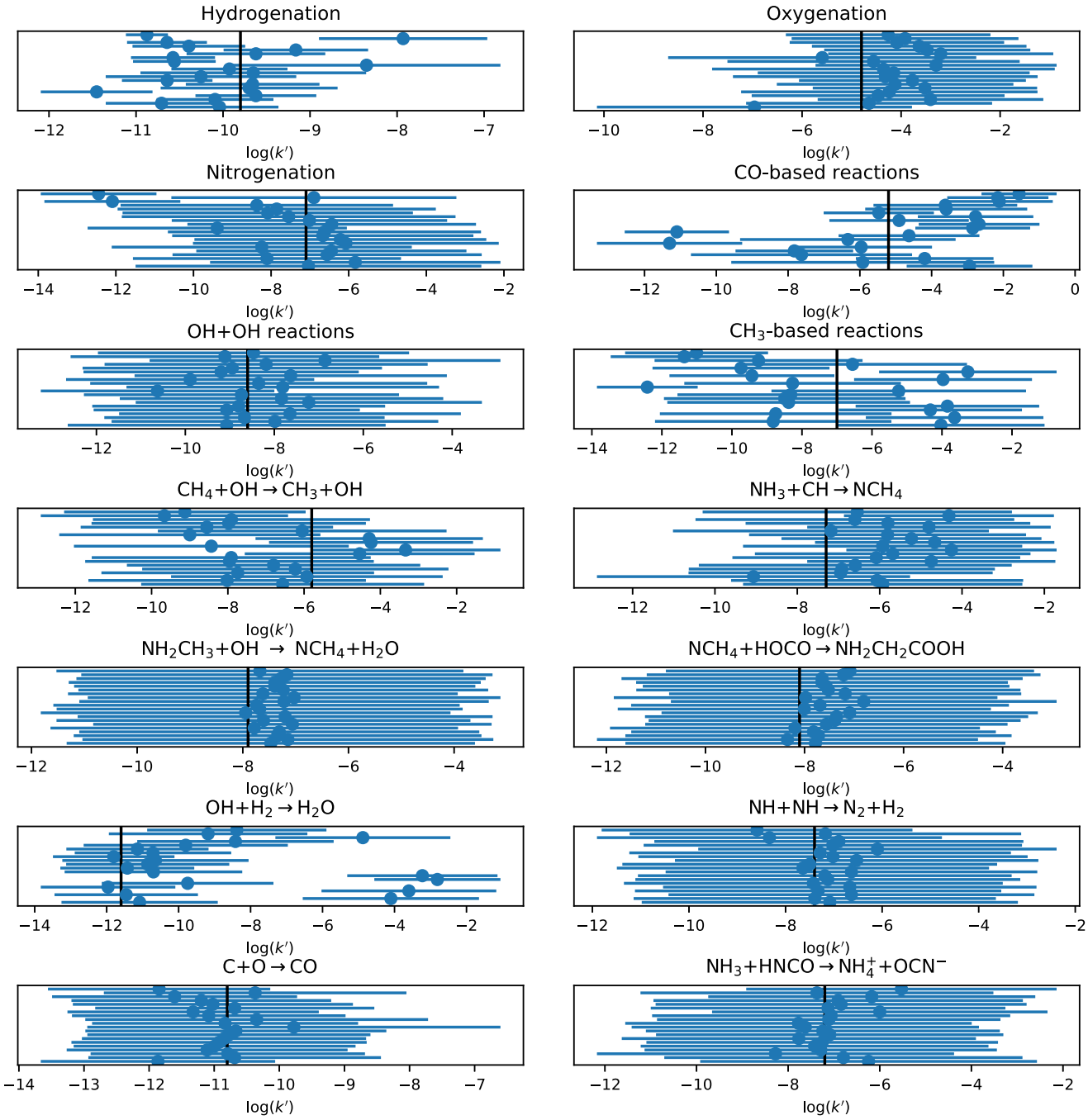

Figure 6. A strip plot of how well the reaction rates are recovered when the forward model is run with some noise on the reaction rates. The vertical black line in each plot represents the "true" reaction rate.

extent the constraints imposed by Equation (12) are influencing the posteriors obtained.

It becomes clear that the extent to which the constraints affect the posteriors depends on the parameter. For the hydrogenation reaction rate, we see that the 65% highest density regions contain the true reaction rates used in the simulation 65% of the time, that is for 13 of the 20 strips. We find that the high-density regions are jittered around the true value, suggesting there is no bias. Overall, what we find is that the constraints are significantly influencing the hydrogenation reaction rate posterior. This is perhaps unsurprising given that most of the constrained species are products of hydrogenation. It is also for this reason that the binding energy for hydrogen has the lowest uncertainty.

However, when the other posteriors are considered, it is clear that there is a greater level of prior domination. While the HDRs are all significantly smaller than the prior range from −15 to 0, we still find that within the HDR the posteriors are not as sharp as for the hydrogenation. This can be confirmed visually by considering the posteriors shown in Figures 2 and 3. While some of the strips, such as for CO-based reaction, show more jitter around the true value, it is clear that more data is required to counter the influence

of the prior distribution. However, we observe that there is no bias in the obtained posteriors.

Appendix B
Determining the Effect of Constraints on the Inferred Binding Energy Values

We observed that the inclusion of upper limits in the likelihood function given by Equation (12) did not have a significant bearing on the binding energy posterior distributions. This suggests that the upper limits listed in Table 2 for N_2, O_2, H_2O_2, and glycine may not be sufficiently constraining. In that case it might be more useful to have abundance measurements for these species. To test the effect of these abundance values on the obtained binding energies, we ran the inference 1000 times using the statistical emulator and plotted the distribution of the maximum-posterior binding energy values in Figure 7. For each inference run, the constraints for each of the species with upper limits in Table 2 were taken to be a random value between 0 and the upper limit. These abundance values were sampled uniformly in this range. The relative error on these four measurements varied to equal 50%, 33.3%, and 20%. The relative errors are represented as ϵ.

We observe that the size of the relative errors has some bearing on the maximum-posterior binding energy values obtained as well as the spread of values. For most of the species, we observe that there is a significant increase in the spread of inferred values. This demonstrates the importance of detecting further grain-surface species in order to better constrain the binding energy values.

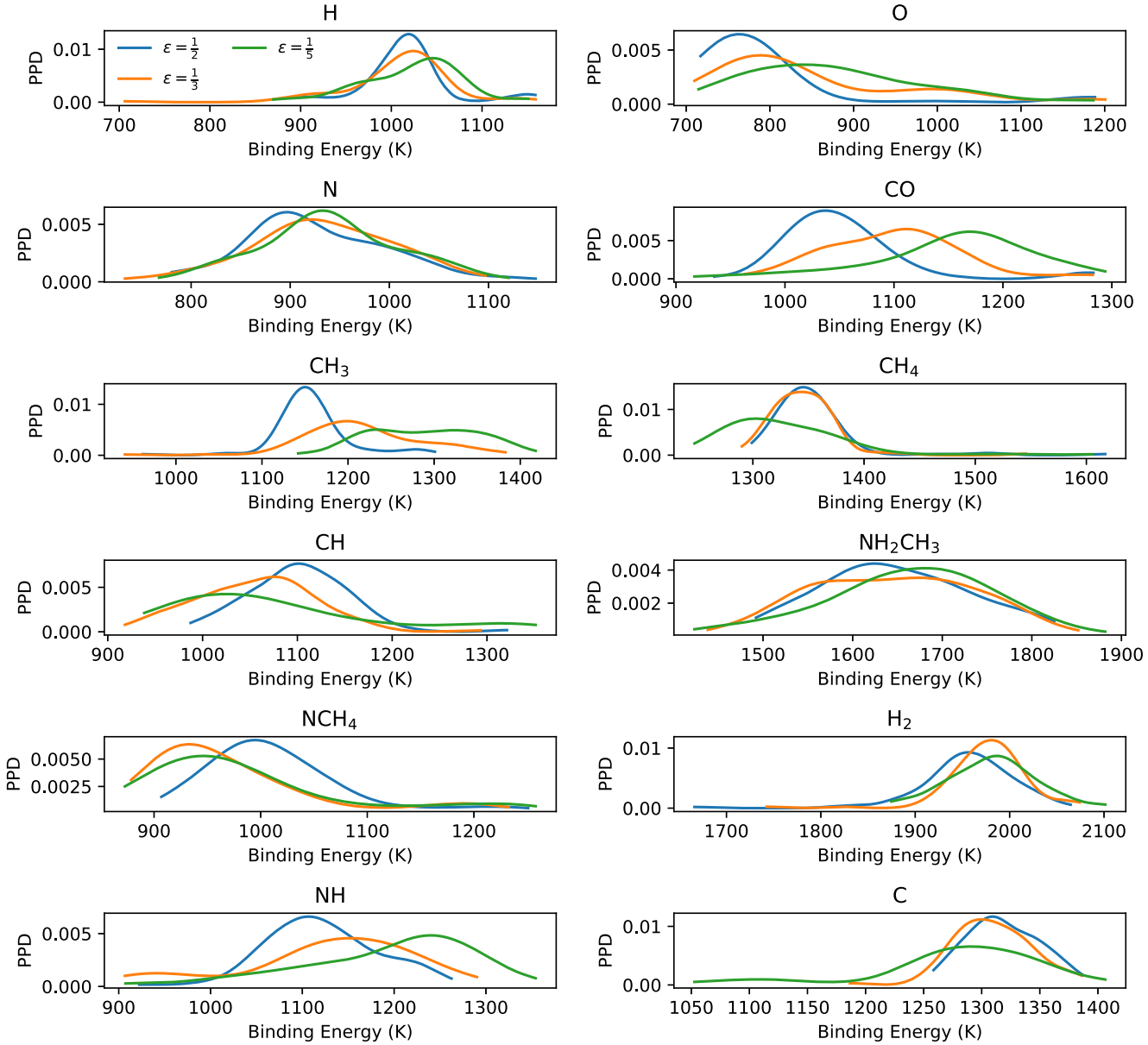

Figure 7. Distributions of the maximum-posterior binding energies obtained when the constraints on N_2, O_2, H_2O_2, and glycine are varied.

ORCID iDs

Johannes Heyl https://orcid.org/0000-0003-0567-8796
Jonathan Holdship https://orcid.org/0000-0003-4025-1552
Serena Viti https://orcid.org/0000-0001-8504-8844

References

Auld, T., Bridges, M., Hobson, M. P., & Gull, S. F. 2007, MNRAS, 376, L11
Boogert, A. A., Gerakines, P. A., & Whittet, D. C. 2015, ARA&A, 53, 541
Bouhlel, M. A., Hwang, J. T., Bartoli, N., et al. 2019, Adv. Eng. Softw., 135, 102662
Bovolenta, G., Bovino, S., Vöhringer-Martinez, E., et al. 2020, MolAs, 21, 100095
Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125
Caselli, P., & Ceccarelli, C. 2012, A&ARv, 20, 56
Chang, Q., Cuppen, H. M., & Herbst, E. 2007, A&A, 469, 973
de Mijolla, D., Viti, S., Holdship, J., Manolopoulou, I., & Yates, J. 2019, A&A, 630, A117
Feroz, F., & Hobson, M. P. 2008, MNRAS, 384, 449
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2019, OJAp, 2, 10
Ferrero, S., Zamirri, L., Ceccarelli, C., et al. 2020, ApJ, 904, 11
Garrod, R. T. 2013, ApJ, 765, 60
Garrod, R. T., & Pauly, T. 2011, ApJ, 735, 15
Gramacy, R. B., & Lee, H. K. H. 2009, Technometrics, 51, 130
Grassi, T., Bovino, S., Caselli, P., et al. 2020, A&A, 643, A155
Grow, A., & Hilton, J. 2018, Statistical Emulation, American Cancer Society, 1, 1
Hasegawa, T. I., Herbst, E., & Leung, C. M. 1992, ApJS, 82, 167
He, J., Acharyya, K., & Vidali, G. 2016, ApJ, 825, 89
Herbst, E., & van Dishoeck, E. F. 2009, ARA&A, 47, 427
Heyl, J., Viti, S., Holdship, J., & Feeney, S. M. 2020, ApJ, 904, 197
Holdship, J., Jeffrey, N., Makrymallis, A., Viti, S., & Yates, J. 2018, ApJ, 866, 116
Holdship, J., Viti, S., Haworth, T. J., & Ilee, J. D. 2021, A&A, 653, A76
Holdship, J., Viti, S., Jiménez-Serra, I., Makrymallis, A., & Priestley, F. 2017, AJ, 154, 38
Hyndman, R. J. 1996, Am. Stat., 50, 120
Ioppolo, S., Fedoseev, G., Chuang, K. J., et al. 2020, NatAs, 5, 197
Jin, M., & Garrod, R. T. 2020, ApJS, 249, 26
Kennedy, M. C., & O'Hagan, A. 2001, StMet, 63, 425
Linnartz, H., Ioppolo, S., & Fedoseev, G. 2015, IRPC, 34, 205
Makrymallis, A., & Viti, S. 2014, ApJ, 794, 45
McElroy, D., Walsh, C., Markwick, A. J., et al. 2013, A&A, 550, A36
McKay, M. D., Beckman, R. J., & Conover, W. J. 1979, Technometrics, 21, 239
Minissale, M., Congiu, E., & Dulieu, F. 2016a, A&A, 585, A146
Minissale, M., Dulieu, F., Cazaux, S., & Hocuk, S. 2016b, A&A, 585, A24
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Pellejero-Ibañez, M., Angulo, R. E., Aricó, G., et al. 2020, MNRAS, 499, 5257
Penteado, E. M., Walsh, C., & Cuppen, H. M. 2017, ApJ, 844, 71
Peterson, P. 2009, Int. J. Comp. Sci. Eng., 4, 296
Quénard, D., Jiménez-Serra, I., Viti, S., Holdship, J., & Coutens, A. 2018, MNRAS, 474, 2796
Rogers, K. K., Peiris, H. V., Pontzen, A., et al. 2019, JCAP, 2019, 031
Ruaud, M., Loison, J. C., Hickson, K. M., et al. 2015, MNRAS, 447, 4004
Schmit, C. J., & Pritchard, J. R. 2017, MNRAS, 475, 1213
Skilling, J. 2006, BayAn, 1, 833
Wakelam, V., Loison, J. C., Mereau, R., & Ruaud, M. 2017, MolAs, 6, 22
Wang, G.-J., Li, S.-Y., & Xia, J.-Q. 2020, ApJS, 249, 25
Whittet, D. C. B., Cook, A. M., Herbst, E., Chiar, J. E., & Shenoy, S. S. 2011, ApJ, 742, 28