Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **517,** 38–46 (2022) https://doi.org/10.1093/mnras/stac2652
Advance Access publication 2022 September 17

Identifying the most constraining ice observations to infer molecular binding energies

Johannes Heyl [1]★ Elena Sellentin,[2,3] Jonathan Holdship [1,2] and Serena Viti[1,2]★

[1] *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[2] *Leiden Observatory, Leiden University, Huygens Laboratory, Niels Bohrweg 2, NL-2333 CA Leiden, the Netherlands*
[3] *Mathematical Institute, Leiden University, Snellius Building, Niels Bohrweg 1, NL-2333 CA Leiden, the Netherlands*

Accepted 2022 September 14. Received 2022 August 26; in original form 2022 July 9

ABSTRACT

In order to understand grain-surface chemistry, one must have a good understanding of the reaction rate parameters. For diffusion-based reactions, these parameters are binding energies of the reacting species. However, attempts to estimate these values from grain-surface abundances using Bayesian inference are inhibited by a lack of enough sufficiently constraining data. In this work, we use the Massive Optimised Parameter Estimation and Data compression algorithm to determine which species should be prioritized for future ice observations to better constrain molecular binding energies. Using the results from this algorithm, we make recommendations for which species future observations should focus on.

Key words: astrochemistry – methods: data analysis – methods: statistical – ISM: abundances.

1 INTRODUCTION

Interstellar dust grains are a crucial component of interstellar chemistry. Many gas-phase complex organic molecules (COMs) have been detected in our galaxy in cold and hot cores (Boogert, Gerakines & Whittet 2015). There is evidence to suggest that much of the observed chemistry takes place on the grain surfaces as opposed to the gas phase and that these observed gas-phase molecules simply evaporate from the grains some time after formation. As such, if one wishes to understand how such COMs are formed, one must have a thorough understanding of grain-surface chemistry (Herbst & van Dishoeck 2009; Caselli & Ceccarelli 2012).

In order to better understand how grain-surface chemistry proceeds, it is important to know the reaction rate parameters. For grain-surface reactions, these parameters may not necessarily be the rates themselves, but rather parameters that are more specific to the reaction rate mechanism. For diffusion-based reactions, which are typically taken to be the dominant grain-surface reaction mechanism, the reaction rate parameters of relevance are the binding energies of the reacting species and reaction activation energy barriers (Hasegawa, Herbst & Leung 1992). Much experimental work has been done to determine these, but there are often significant disagreements, due to differing laboratory conditions [see Penteado, Walsh & Cuppen (2017) for a survey of binding energy values].

There exist a variety of methods to estimate the binding energies, ranging from experimental approaches (He, Acharyya & Vidali 2016) to density functional theory (Ferrero et al. 2020) to machine learning (Villadsen et al. 2022). However, in our work to estimate these reaction rate parameters given observed abundances, Bayesian inference is typically employed. Bayesian inference has become a ubiquitous tool in astrophysics and has recently found more use within the field of astrochemistry. Previous work has considered the rate-parameter estimation problem (Holdship et al. 2018; Heyl et al. 2020) and has shown that the paucity of available grain-surface species abundances inhibits precise estimates of these rate parameters. The problem due to the lack of sufficiently constraining data has been somewhat ameliorated by considering the network structure (Heyl et al. 2020) or the underlying chemical mechanisms to reduce the dimensionality of the problem (Heyl, Holdship & Viti 2022). However, it remains the case that many binding energies cannot be constrained to the point that they would be useful in chemical codes. This is clear from a survey of the literature that shows quite significant disagreements for some binding energy values (McElroy et al. 2013; Wakelam et al. 2017; Quénard et al. 2018).

Observations of the ices have typically considered the molecular vibration transitions in the infrared (IR) region (Boogert et al. 2015). A number of space telescopes such as the *Infrared Space Observatory* (*ISO*) and *Spitzer* have provided observations of ice band profiles that have been used to determine molecular abundances. However, until now there has been insufficient resolution of the absorption band profiles. The *James Webb Space Telescope* (*JWST*) observes in the IR wavelength range of 0.6–28 μm. It provides higher spectral resolution observations of up 2 mag, especially in the 5–8 μm range that potentially contains the vibrational modes of several molecules of interest (Boogert et al. 2015; Boogert 2016). This is particularly important as IR spectroscopy reveals the features of various functional groups that differ by species but can have similar values (Boogert 2016). As such, having greater resolution will ensure that the various absorption band profiles can be disentangled.

In this work, we wish to provide recommendations of which species should be prioritized for future ice observations in order to reduce the uncertainties on the binding energy values. To achieve

★ E-mail: johannes.heyl.19@ucl.ac.uk (JH); viti@strw.leidenuniv.nl (SV) this, we make use of the 'Massive Optimised Parameter Estimation and Data compression' (MOPED) algorithm (Heavens, Jimenez & Lahav 2000; Heavens et al. 2017; Heavens, Sellentin & Jaffe 2020). A key output of the MOPED algorithm is a measure of how strongly knowledge of a species' ice-phase abundance would constrain the binding energies.

We start by explaining the chemical code and network we will use throughout this work in Section 2. Section 3 will be dedicated to explaining the approach we take in this work, specifically our use of Bayesian inference and the MOPED algorithm. We follow this up in Section 4 by showing the results of the Bayesian inference and the MOPED algorithm as well as by discussing the observational implications of our findings. We briefly conclude in Section 5.

2 THE CHEMICAL CODE AND NETWORK

2.1 The chemical code

In this work, the gas–grain astrochemical code UCLCHEM (Holdship et al. 2017) was used to model the chemistry of a collapsing dark cloud. The cloud was taken to collapse isothermally at 10 K from 10^2 to 10^6 cm^{-3} over a period of 5 Myr. By the end of this collapse, we expect the ice-phase abundances to be representative of a dark cloud.

2.2 Grain-surface chemistry

2.2.1 Grain-surface diffusion

It is important to understand the grain-surface mechanisms, as this is needed to show why this work considers binding energies as the key parameters that govern the reaction rates.

We assume that all grain-surface reactions take place via the Langmuir–Hinshelwood mechanism and use the formalism described in Hasegawa et al. (1992), which was implemented in UCLCHEM in Quénard et al. (2018). We believe this is a reasonable assumption as previous work has shown that including Eley–Rideal reactions does not strongly affect surface abundances (Ruaud et al. 2015). According to the formalism, the rate at which two species A and B react via diffusion is given by

$$k_{\rm AB} = \kappa_{\rm AB} \frac{(k^{\rm A}_{\rm hop} + k^{\rm B}_{\rm hop})}{N_{\rm site} n_{\rm dust}}, \tag{1}$$

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

In equation (1), $k^X_{\rm hop}$ is the thermal hopping rate of species X on the grain surface, which is defined as

$$k^X_{\rm hop} = \nu_0 \exp\left(-\frac{E_{\rm D}}{T_{\rm gr}}\right), \tag{2}$$

where $E_{\rm D}$ 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 a fraction of the binding energy of the species, $E_{\rm b}$. In this work, this fraction is taken to be 0.5, in line with Quénard et al. (2018). While it is known that this value can vary between 0.3 and 0.8, there is significant uncertainty within that range (Garrod & Pauly 2011). Furthermore, the value is not expected to play a significant role at 10 K (Vasyunin et al. 2017).

The characteristic vibration frequency, ν_0, is defined as

$$\nu_0 = \sqrt{\frac{2k_{\rm b} n_{\rm s} E_{\rm b}}{\pi^2 m}}, \tag{3}$$

where $k_{\rm b}$ is the Boltzmann constant, $n_{\rm s}$ is the grain site density, and m is the mass of species. While there exists some debate regarding the validity of this expression [see Minissale et al. (2022) for a more detailed discussion], this equation for the characteristic vibration frequency is what is used in UCLCHEM. While a more accurate equation that takes into account the rotation partition function of the desorbing molecules should be used, this will not affect the ability of Bayesian inference to constrain the binding energies of species of interest, which is the aim of this paper.

The final term, $\kappa_{\rm AB}$, which gives the reaction probability, is

$$\kappa_{\rm AB} = \max\left(\exp\left(-\frac{2a}{\hbar}\sqrt{2\mu k_{\rm b} E_{\rm A}}\right), \exp\left(-\frac{E_{\rm A}}{T_{\rm gr}}\right)\right), \tag{4}$$

where $\hbar$ is the reduced Planck constant, μ is the reduced mass, $E_{\rm A}$ is the reaction activation energy, $k_{\rm b}$ is the Boltzmann constant, and $a = 1.4$ Å is the thickness of a quantum mechanical barrier. While values between 1 and 2 Å have been used (Hasegawa et al. 1992; Garrod & Pauly 2011; Vasyunin et al. 2017), Quénard et al. (2018) found that a value of 1.4 Å matched the ice composition best. The reaction probability represents the competition between the quantum mechanical probability of a tunnelling through a rectangular barrier of thickness a, which is the first term, and the thermal reaction probability, which is the second term.

2.2.2 Reaction-diffusion competition

A modification needs to be made to the $\kappa_{\rm AB}$ term to take into account the possibility that species might diffuse or evaporate before they can react with each other. This is the reaction-diffusion competition (Chang, Cuppen & Herbst 2007; Garrod & Pauly 2011). The reaction probability is now defined as

$$\kappa^{\rm final}_{\rm AB} = \frac{p_{\rm reac}}{p_{\rm reac} + p_{\rm diff} + p_{\rm evap}}, \tag{5}$$

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

$$p_{\rm reac} = \max(\nu^{\rm A}_0, \nu^{\rm B}_0)\kappa_{\rm AB}, \tag{6}$$

$$p_{\rm diff} = k^{\rm A}_{\rm hop} + k^{\rm B}_{\rm hop}, \text{ and} \tag{7}$$

$$p_{\rm evap} = \nu^{\rm A}_0 \exp\left(-\frac{E^{\rm A}_{\rm b}}{T_{\rm gr}}\right) + \nu^{\rm B}_0 \exp\left(-\frac{E^{\rm B}_{\rm b}}{T_{\rm gr}}\right). \tag{8}$$

We replace $\kappa_{\rm AB}$ with $\kappa^{\rm final}_{\rm AB}$ in equation (1).

Overall, we find that equations (1)–(8) show that the key quantities are ν_0, $k^X_{\rm hop}$, $E_{\rm b}$, and $E_{\rm A}$. The first three are all functions of the binding energies of the reacting species, indicating the binding energies are the crucial parameters. We assume that the activation energies in equation (4) are well known. This is reasonable, as these should be independent of the ice composition (unlike the binding energies) and can be determined theoretically or experimentally. Many of the reactions would also be expected to have zero activation energy as they are radical–radical reactions (Quénard et al. 2018).

2.3 The chemical network

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

We believe that the gas-phase network is comprehensive and sufficiently accurate that any deficiencies in the network will not have a great effect on our results. The gas-phase network was benchmarked against observations in McElroy et al. (2013). The abundances of species freezing out from the gas phase are likely to be approximately correct, and we therefore only need to be concerned by the accuracy and completeness of the grain-surface network. We operate under the assumption that the gas-phase network is complete.

Our grain-surface network is less comprehensive, but we argue it is sufficient to reproduce the abundance of major species, given the results of Makrymallis & Viti (2014), Holdship et al. (2018), and Heyl et al. (2020, 2022), which used smaller networks. The network includes the freeze-out of all species, hydrogenation reactions of all species up to their saturated forms, and radical–radical reactions that have been shown to be efficient in laboratory experiments, as well as other diffusion reactions from the literature (see above). By including all reactions known to be the main routes through which species like H_2O and CH_3OH are formed on the grain surfaces, our network is sufficient to produce accurate ice-phase abundances of these species. Therefore, we can properly predict how important the binding energies of those species are to the surface chemistry.

3 ANALYTICAL APPROACH

3.1 Parameters

The aim of this work is to determine the binding energies of the chemically reactive species. While it would be ideal to determine the binding energies of all species in the network, the reality of the situation is that this is not strictly necessary. In Heyl et al. (2022), it was demonstrated that at 10 K, a moderate difference in binding energies between two species results in a significant difference in reaction rates. As such, one can significantly reduce the dimensionality of the problem one is trying to solve by only considering the most diffusive species. These are those species that will be the more reactive species with the greater hopping frequency for at least one reaction in the network. The more reactive species were determined by considering the literature. Even though there is widespread disagreement about the values of the binding energies, there is less disagreement about the hierarchy of binding energy values. This can be seen by considering the values given in Wakelam et al. (2017), McElroy et al. (2013), and Penteado et al. (2017). For reactions where the literature was not definitive in specifying which species had the lower binding energy, both species' binding energies were included as parameters. The binding energies we considered as parameters were the binding energies of H, H_2, C, CH, N, CH_3, NH, CH_4, and O.

3.2 Bayesian inference

3.2.1 Introduction to Bayesian inference

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

Table 1. The abundances and uncertainties taken for the network adapted from Boogert et al. (2015).

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

In order to solve the inverse problem, we require abundance measurements of grain-surface species, $\boldsymbol{d}$. These are listed in Table 1. These are taken from Boogert et al. (2015).

Bayes' law can be used to determine the posterior distribution of the binding energies given the data:

$$P(\boldsymbol{E}|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\boldsymbol{E})P(\boldsymbol{E})}{P(\boldsymbol{d})}, \qquad (9)$$

where $P(\boldsymbol{E}|\boldsymbol{d})$ is the posterior probability distribution, $P(\boldsymbol{E})$ is the prior, $P(\boldsymbol{d}|\boldsymbol{E})$ is the likelihood, and $P(\boldsymbol{d})$ is referred to as the evidence. The prior distribution encodes the initial understanding of the binding energy distribution. The likelihood gives the data's likelihood as a function of the binding energies. Within the likelihood function, the physical model is encoded. The evidence serves as a normalizing factor and represents the marginalized likelihood. The posterior distribution represents the updated probability distribution of reaction rates based on the data, the prior distribution, and the physical model.

3.2.2 Implementation

The prior for all binding energies was specified as a uniform distribution between 400 and 2000 K. The abundance measurements in Table 1 were assumed to be Gaussian, which allowed for the specification of a Gaussian likelihood function:

$$P(\boldsymbol{d}|\boldsymbol{E}) = \prod_{i=1}^{n_d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(d_i - Y_i)^2}{2\sigma_i^2}\right), \qquad (10)$$

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

The ULTRANEST PYTHON package (Buchner 2021) was used for the Bayesian inference, which is based on the MLFriends algorithm (Buchner 2016, 2019). The package also outputs the maximum-likelihood estimator, $\boldsymbol{E}_{\rm \bf ML}$. We will use this later for the MOPED algorithm.

3.3 The MOPED algorithm

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

Recall that we wish to determine a set of parameters $\boldsymbol{E}$. The species that are found to be important may include the species already listed in Table 1, in which case we would aim to improve the uncertainties surrounding their values. However, it is also possible that we would need to detect species that have not been detected yet.

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

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

$$c_\alpha = \boldsymbol{b}_{\boldsymbol{\alpha}}^{\boldsymbol{T}} \boldsymbol{Y}, \tag{11}$$

where α ranges from 1 to p and $\boldsymbol{b}_{\boldsymbol{\alpha}}$ is a set of orthonormal linear filters, such that each one contains as much information about that parameter that is not contained in any other $\boldsymbol{b}_{\boldsymbol{\alpha}}$. $\boldsymbol{Y}$ represents a vector containing the final abundances for some arbitrary value of $\boldsymbol{E}$. As a fiducial model, we typically take $\boldsymbol{E} = \boldsymbol{E}_{\mathbf{ML}}$, which we can determine using the Bayesian inference discussed in Section 3.2. Using the maximum-likelihood parameters as a fiducial model has been found to be sufficient (Heavens et al. 2000, 2017). The value of each c_α will ultimately be more strongly influenced by the components $\boldsymbol{b}_{\boldsymbol{\alpha}}$ that are larger in magnitude. As there is one species for each component, this means that if a component has a greater magnitude, then it contains more information about that parameter.

The vectors $\boldsymbol{b}_{\boldsymbol{\alpha}}$ are given by

$$\boldsymbol{b}_1 = \frac{\mathbf{C}^{-1}\boldsymbol{Y}_{,1}}{\sqrt{\boldsymbol{Y}_{,1}^{T}\,\mathbf{C}^{-1}\boldsymbol{Y}_{,1}}} \tag{12}$$

and

$$\boldsymbol{b}_\alpha = \frac{\mathbf{C}^{-1}\boldsymbol{Y}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{Y}_{,\alpha}^{T}\boldsymbol{b}_{,\beta})\boldsymbol{b}_{,\beta}}{\sqrt{\boldsymbol{Y}_{,\alpha}^{T}\mathbf{C}^{-1}\boldsymbol{Y}_{,\alpha} - \sum_{\beta=1}^{\alpha-1}(\boldsymbol{Y}_{,\alpha}^{T}\,\boldsymbol{b}_{,\beta})^2}}, \tag{13}$$

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

$$\boldsymbol{b}_{\boldsymbol{\alpha}}^{\boldsymbol{T}}\mathbf{C}\boldsymbol{b}_{\boldsymbol{\beta}} = \delta_{\alpha\beta}, \tag{14}$$

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

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

Recall that the magnitude of each component of $\boldsymbol{b}_{\boldsymbol{\alpha}}$ gives a weighting for that species' influence on the parameter α. To determine the best species to prioritize detection for, we simply add the absolute values of the components of $\boldsymbol{b}_{\boldsymbol{\alpha}}$ for species across all α. That is, we perform the sum over our linear filters

$$\sum_{\alpha=1}^{p}[|b_\alpha^1|, |b_\alpha^2|..., |b_\alpha^M|]. \tag{15}$$

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

4 RESULTS

4.1 Results of the Bayesian inference

Fig. 1 shows the marginalized posterior distributions for the binding energies of interest. The marginalized prior distribution is also plotted for comparison. It is clear that, with the exception of atomic hydrogen's binding energy, the marginalized posterior distributions differ very little from the prior suggesting a lack of sufficiently constraining data. It is for this reason that we now use the MOPED algorithm to identify species we need to detect to better constrain our posterior distributions.

4.2 Using MOPED

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

Fig. 2 shows the sum of the filters for all grain-surface species. The greater the filter sum, the more important it is to detect that molecule. Additionally, one must also consider the likely abundance of each species, as the species will only be observable in the ices if its abundance is above some minimum threshold. We therefore believe that future ice observations should prioritize species that have a high filter sum as well as a high abundance. In order to provide estimates of the abundances, we inserted the maximum-likelihood estimator values for the binding energy, $\boldsymbol{E}_{\mathbf{ML}}$, into UCLCHEM and obtained the fitted abundances for all the species. Fig. 3 is a scatter plot of the filter sum values against the abundances for each species. From this plot, we are able to identify high-importance species that are also likely to be detectable in the ices. However, one needs to also account for which species are realistic targets from a chemical point of view. This is discussed in the next subsection.

4.3 Observational implications

The MOPED analysis has resulted in a clear ranking of which species should be targeted in future ice observations. This ranking is shown in Fig. 2. Of course, we note that many of these species have very low abundances and others are difficult to detect in

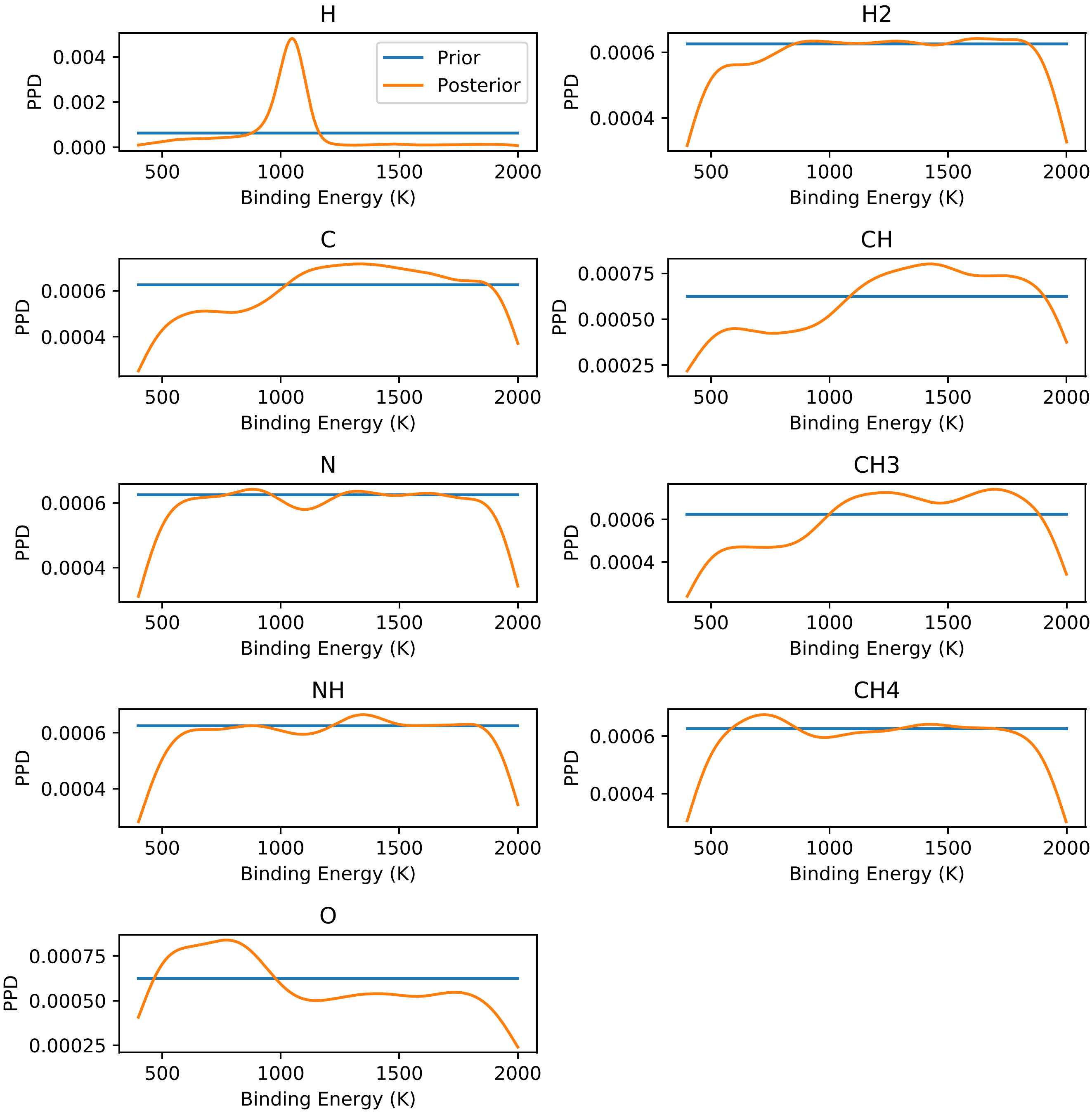

Figure 1. Marginalized posterior distributions of the binding energies of the diffusive species of interest. Also plotted is the prior distribution on the binding energies. With the exception of H, most binding energy distributions differ very little from the prior distribution. This is due to the lack of enough sufficiently constraining data. This motivates the need for further ice observations to reduce the variance of the distributions.

absorption. Diatomic molecules, atomic species, and all radicals except CO will be neglected in our considerations of which species to consider.

We briefly return to the issue of the network's reliability, which was first discussed in Section 2.3. While one can be confident in the abundances of CH_4, H_2CO, CH_3OH, and H_2O as their networks are experimentally derived (Fuchs et al. 2009; Ioppolo et al. 2011; Chuang et al. 2016; Qasim et al. 2020), other species should be viewed more sceptically. This is particularly the case for sulphur. Many works indicate that sulphur may primarily be locked in other forms (Woods et al. 2015; Vidal et al. 2017). It may be that the sulphur reaction network is incomplete. Most concerning is H_2S, which the model suggests is the primary sulphur reservoir on the grains. Observations of ices have never detected H_2S but have instead provided upper limits of $\sim 10^{-6}$ (Boogert et al. 2015). The most likely value of the H_2S abundance derived here is lower than this limit and so it may be correct. However, there are other species in the network such as CS whose surface chemistry is not well understood (Woods et al. 2015). Taking this into consideration, it could be argued that observers should instead target species such as H_2CO or HCN that have similar

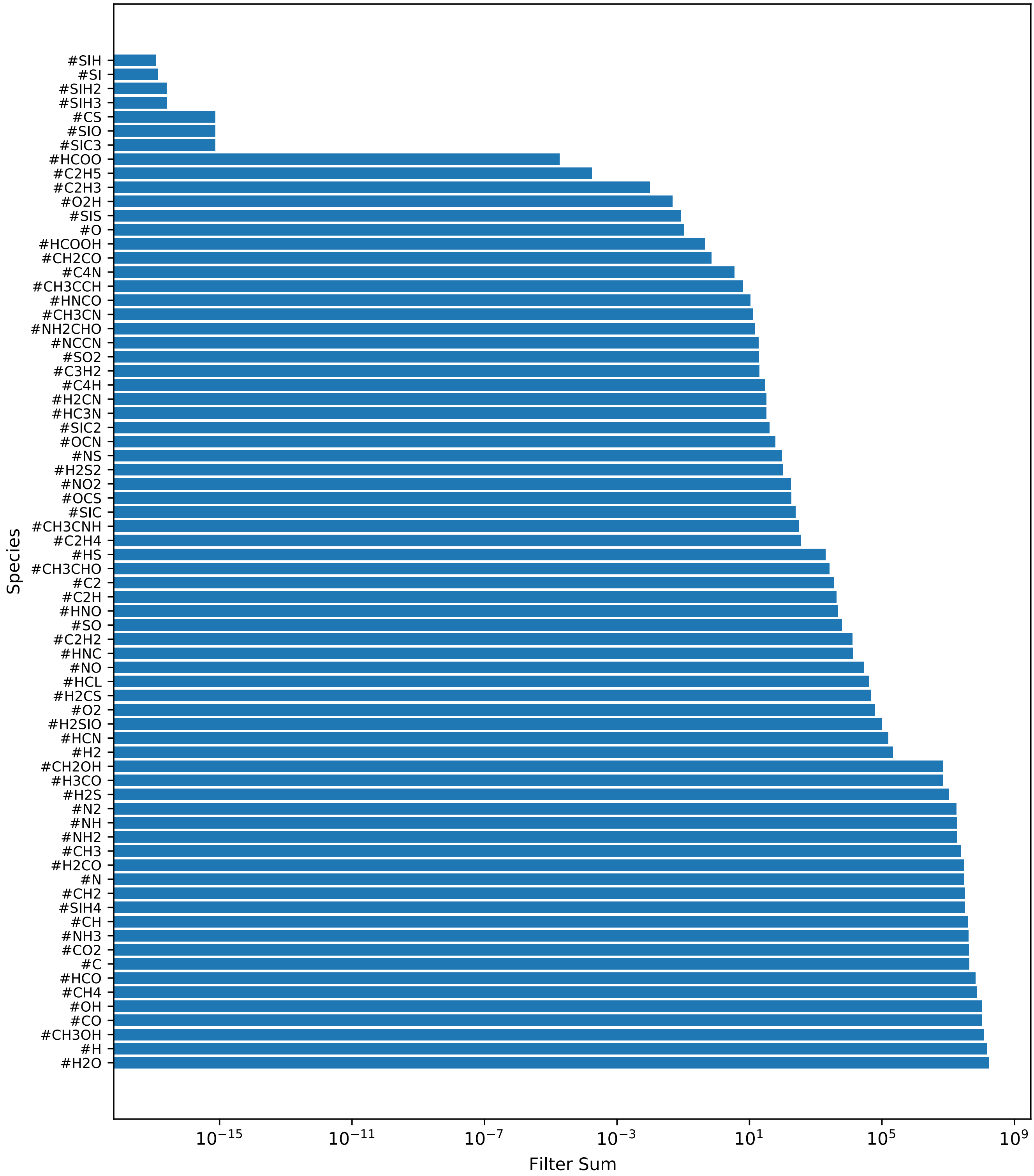

Figure 2. Bar chart showing the filter sums for each species in ascending order. Species with a larger filter sum should be prioritized for detection. Many of the species we observe are the intermediate species formed during the creation of the saturated species in Table 1. This indicates that understanding these intermediate products is essential to better constraining the binding energies of interest. We also note that many of the highest-ranked species have already been detected. This suggests that future observations should aim to improve the level of precision of these abundance measurements.

filter sums and more reliable networks despite their lower predicted abundances.

There is much to be gained from obtaining more precise measurements for the abundances of species listed in Table 1. All of these species except for HCOOH and NH_4^+ have high filter sums and high abundances in the fitted model. However, the uncertainties on the measured abundances are often 50 per cent of the measured value. Our MOPED analysis shows that it would actually be much

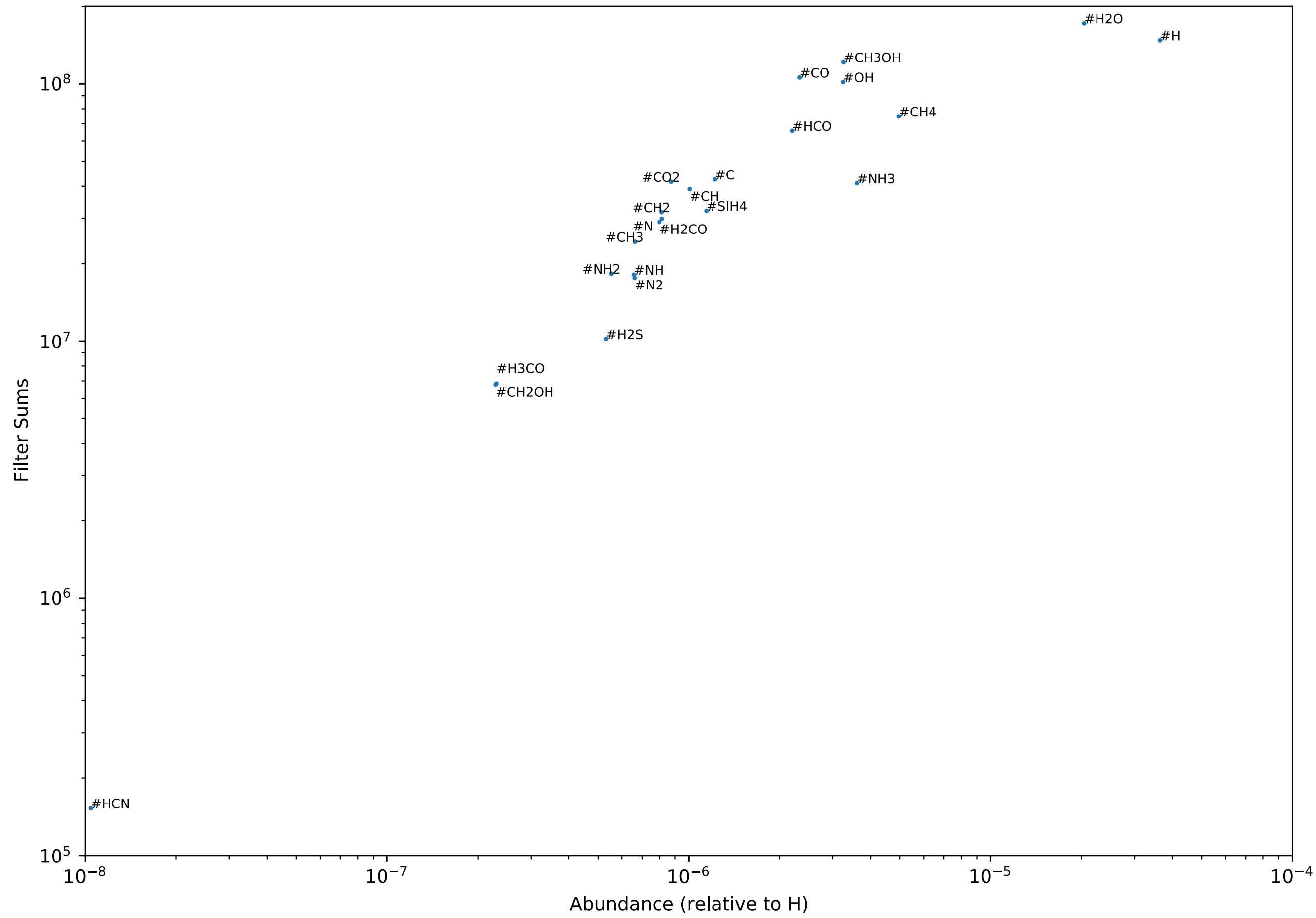

Figure 3. Scatter plot depicting filter sum against the predicted abundances when the MLEs for binding energies are inserted into UCLCHEM. Given constraints on instrumental uncertainties, we should look to prioritize species that not only are important, as determined by their filter sums, but can also be realistically detected. These include saturated species such as #CH4, #NH3, #CO2, and #H2O, but also their precursors.

more valuable to determine these abundances to a smaller degree of uncertainty than it would be to measure the abundance of new species. To demonstrate the effect of reducing the uncertainties on the abundances, we redid the Bayesian analysis, but reduced the uncertainty on water's abundance to 10^{-6}. Fig. 4 shows the resulting binding energy posteriors. We observe significant changes in the posterior distributions for H and O. This suggests that there is much promise in improving the measured ice abundances for those molecules. Many of the absorption band profiles for these species are in the wavelength range of *JWST*, but especially in the 5–8 μm range that will have higher resolution compared to *Spitzer* (Boogert et al. 2015). This is promising as it is certain that H_2O and the other abundant species can be observed and telescope time simply needs to be dedicated to further constraining their abundances.

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

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

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

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

5 CONCLUSION

In this work, we have utilized the MOPED algorithm to identify the species that would best constrain binding energies. Bayesian inference was found to result in poorly constrained marginalized posterior distributions for the binding energies. This was due to the lack of enough sufficiently constraining data. The MOPED algorithm allowed us to determine which ice species should be prioritized for

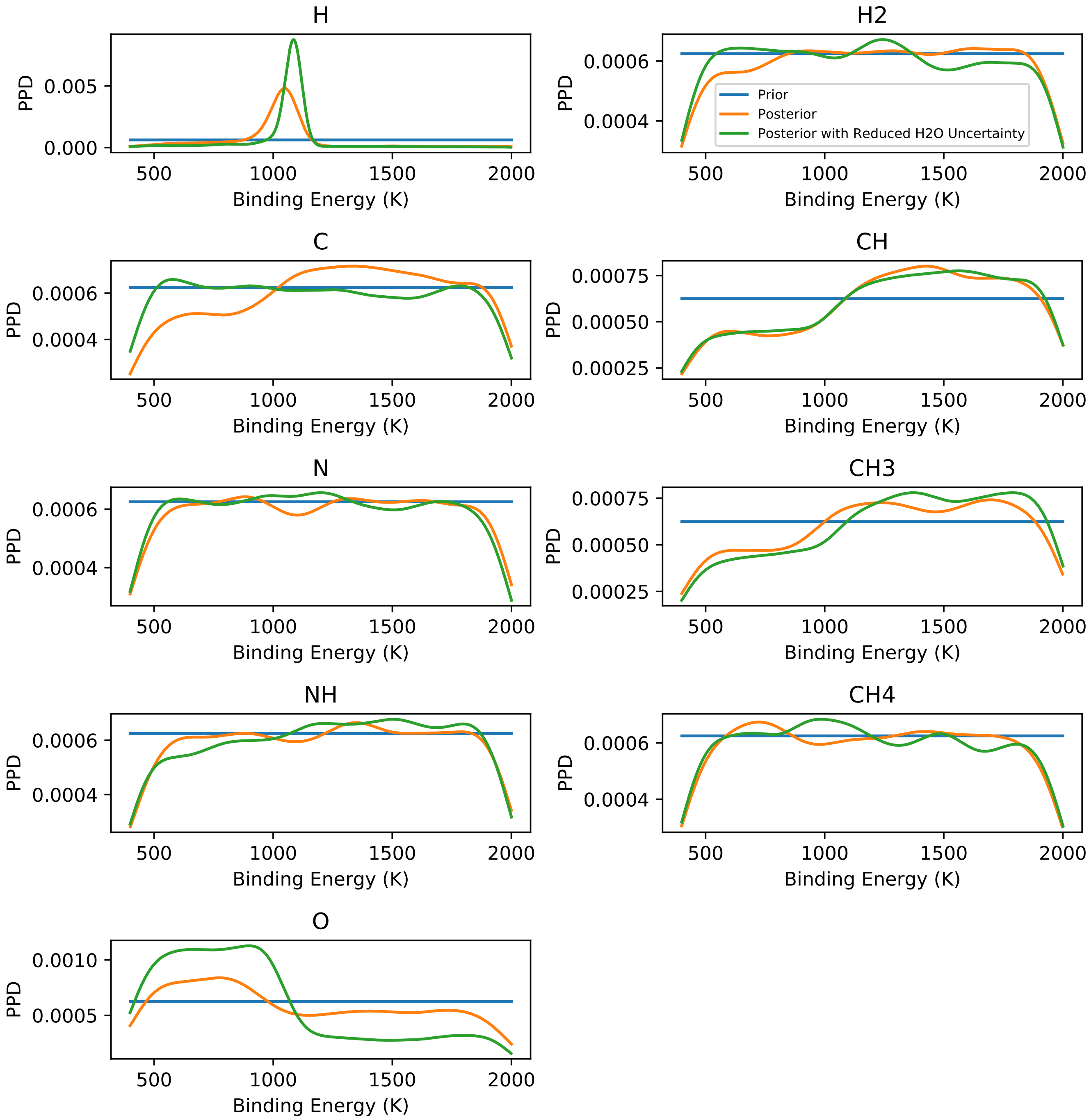

Figure 4. Marginalized posterior distributions of the binding energies of the diffusive species of interest. We also plot the prior distribution and the posterior distributions when the uncertainty on water's abundance is reduced to 10^{-6}. We observe that this has a significant effect on the marginalized posterior distributions of H and O, indicating that there is promise in improving the abundance measurements for species that have already been detected.

future ice observations in such a way that they would further constrain the posteriors. By then considering which species in the fitted model have the highest filter sums as well as the largest abundances, we come up with a list of species that should be targeted. These species are H_2O, CO_2, NH_3, CH_4, CO, CH_3OH, H_2CO, HCN, and H_2S. While some of these species have not been detected, some of them have, which suggests that more precise measurements of these species are necessary. We also comment on which features of each species are likely to appear in the wavelength range considered by *JWST*.

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

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

ACKNOWLEDGEMENTS

The authors thank the referee for their constructive comments that greatly improved this work. JH was 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. SV acknowledges support from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 811312 for the project 'Astro-Chemical Origins' (ACO).

DATA AVAILABILITY

The data underlying this paper are available in the paper and in its online supplementary material.

REFERENCES

Belloche A. et al., 2017, A&A, 601, A49
Boogert A. A., Gerakines P. A., Whittet D. C., 2015, ARA&A, 53, 541
Boogert A. C. A., 2016, in Benvenuti P., ed., Astronomy in Focus, Vol. 1, Focus Meeting 12 XXIXth IAU General Assembly, Cambridge University Press, Cambridge, p. 317
Buchner J., 2016, Stat. Comput., 26, 383
Buchner J., 2019, PASP, 131, 108005
Buchner J., 2021, J. Open Source Softw., 6, 3001
Caselli P., Ceccarelli C., 2012, A&AR, 20, 56
Chang Q., Cuppen H. M., Herbst E., 2007, A&A, 469, 973
Chuang K. J., Fedoseev G., Ioppolo S., van Dishoeck E. F., Linnartz H., 2016, MNRAS, 455, 1702
Fathe K., Holt J. S., Oxley S. P., Pursell C. J., 2006, J. Phys. Chem. A, 110, 10793
Fedoseev G., Chuang K. J., van Dishoeck E. F., Ioppolo S., Linnartz H., 2016, MNRAS, 460, 4297
Ferrero S., Zamirri L., Ceccarelli C., Witzel A., Rimola A., Ugliengo P., 2020, ApJ, 904, 11
Fuchs G. W., Cuppen H. M., Ioppolo S., Romanzin C., Bisschop S. E., Andersson S., van Dishoeck E. F., Linnartz H., 2009, A&A, 505, 629
Garrod R. T., Herbst E., 2006, A&A, 457, 927
Garrod R. T., Pauly T., 2011, ApJ, 735, 15
Garrod R. T., Widicus Weaver S. L., Herbst E., 2008, ApJ, 682, 283
Gerakines P. A., Yarnall Y. Y., Hudson R. L., 2022, MNRAS, 509, 3515
Hasegawa T. I., Herbst E., Leung C. M., 1992, ApJS, 82, 167
He J., Acharyya K., Vidali G., 2016, ApJ, 825, 89
Heavens A. F., Jimenez R., Lahav O., 2000, MNRAS, 317, 965
Heavens A. F., Sellentin E., de Mijolla D., Vianello A., 2017, MNRAS, 472, 4244
Heavens A. F., Sellentin E., Jaffe A. H., 2020, MNRAS, 498, 3440
Herbst E., van Dishoeck E. F., 2009, ARA&A, 47, 427
Heyl J., Holdship J., Viti S., 2022, ApJ, 931, 26
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., Jiménez-Serra I., Makrymallis A., Priestley F., 2017, AJ, 154, 38
Ioppolo S., van Boheemen Y., Cuppen H. M., van Dishoeck E. F., Linnartz H., 2011, MNRAS, 413, 2281
Kaiser R. I., Osamura Y., 2005a, A&A, 432, 559
Kaiser R. I., Osamura Y., 2005b, ApJ, 630, 1217
Keane J. V., Tielens A. G. G. M., Boogert A. C. A., Schutte W. A., Whittet D. C. B., 2001, A&A, 376, 254
Makrymallis A., Viti S., 2014, ApJ, 794, 45
McElroy D., Walsh C., Markwick A. J., Cordiner M. A., Smith K., Millar T. J., 2013, A&A, 550, A36
Minissale M. et al., 2022, ACS Earth Space Chem., 6, 597
Minissale M., Dulieu F., Cazaux S., Hocuk S., 2016, A&A, 585, A24
Penteado E. M., Walsh C., Cuppen H. M., 2017, ApJ, 844, 71
Qasim D., Fedoseev G., Chuang K. J., He J., Ioppolo S., van Dishoeck E. F., Linnartz H., 2020, Nat. Astron., 4, 781
Quan D., Herbst E., Osamura Y., Roueff E., 2010, ApJ, 725, 2101
Quénard D., Jiménez-Serra I., Viti S., Holdship J., Coutens A., 2018, MNRAS, 474, 2796
Ruaud M., Loison J. C., Hickson K. M., Gratier P., Hersant F., Wakelam V., 2015, MNRAS, 447, 4004
Smith R. G., 1991, MNRAS, 249, 172
Song L., Kästner J., 2016, Phys. Chem. Chem. Phys., 18, 29278
Tegmark M., Taylor A. N., Heavens A. F., 1997, ApJ, 480, 22
Terwisscha van Scheltinga J., Marcandalli G., McClure M. K., Hogerheijde M. R., Linnartz H., 2021, A&A, 651, A95
Vasyunin A. I., Caselli P., Dulieu F., Jiménez-Serra I., 2017, ApJ, 842, 33
Vidal T. H. G., Loison J.-C., Jaziri A. Y., Ruaud M., Gratier P., Wakelam V., 2017, MNRAS, 469, 435
Villadsen, T.,Ligterink N. F. W.,Andersen M.,2022,Predicting binding energies of astrochemically relevant molecules via machine learning, preprint (arXiv:2207.03906)
Wakelam V., Loison J. C., Mereau R., Ruaud M., 2017, Mol. Astrophys., 6, 22
Woods P. M., Occhiogrosso A., Viti S., Kaňuchová Z., Palumbo M. E., Price S. D., 2015, MNRAS, 450, 1256

This paper has been typeset from a TEX/LATEX file prepared by the author.