A statistical and machine learning approach to the study of astrochemistry

Heyl, J.; Viti, S.; Vermariën, G.J.P.W.

Faraday Discussions
Cite this: *Faraday Discuss.*, 2023, **245**, 569

PAPER

A statistical and machine learning approach to the study of astrochemistry

Johannes Heyl, †[a] Serena Viti †*[ba] and Gijs Vermariën [b]

Received 15th January 2023, Accepted 8th February 2023
DOI: 10.1039/d3fd00008g

In order to obtain a good understanding of astrochemistry, it is crucial to better understand the key parameters that govern grain-surface chemistry. For many chemical networks, these crucial parameters are the binding energies of the species. However, there exists much disagreement regarding these values in the literature. In this work, a Bayesian inference approach is taken to estimate these values. It is found that this is difficult to do in the absence of enough data. The Massive Optimised Parameter Estimation and Data (MOPED) compression algorithm is then used to help determine which species should be prioritised for future detections in order to better constrain the values of binding energies. Finally, an interpretable machine learning approach is taken in order to better understand the non-linear relationship between binding energies and the final abundances of specific species of interest.

1 Introduction

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

From an astrochemical point of view, due to their high densities and low temperatures, these regions are great laboratories to study the interactions of gas and dust, with species from the gas phase 'freezing' onto the dust grains present, and forming icy mantles rich in hydrogenated as well as complex organic molecules (COMs), due to the many fast surface reactions that take place. As stars form in these clouds (or if any other energetic process takes place) then the dust

[a]*Department of Physics and Astronomy, University College London, Gower Street, WC1E 6BT, London, UK. E-mail: johannes.heyl.19@ucl.ac.uk*
[b]*Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands*
† These authors contributed equally.

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

In recent years, coupling chemical and radiative transfer models for the interpretation of molecular emission has been routinely done and the success of such techniques has varied to different degrees, depending on whether one wants to model the physical and chemical structure, or the hydrodynamical history of the gas.[1–4] However the shortcomings of such methods are two-fold: (i) understanding the physical conditions in molecular gas *via* a systematic and applicable to many galaxies methodology is an inverse problem subject to complicated chemistry that varies non-linearly with both time and the physical environment;[5] hence it may not have a solution, solutions might not be unique and/or might not depend continuously on the observational data. Traditionally astrochemistry has always been dominated by trial and error grid-based analysis combined with simple statistics,[6] an approach that becomes impossible or ineffective when datasets (*e.g.* from ALMA) and/or parameter space are large, complex, or heterogeneous; (ii) the knowledge of the micro-physics and chemistry of what occurs on the dust is well behind what is known for the gas-phase. While surface reactions and dynamics (including desorption and diffusion) can be experimentally investigated (but always within a constrained range of laboratory conditions), experimental data for interstellar ices are still limited. In order to make the best use of experimental resources, the chemical data that models require need to be prioritized according to what will have the most impact.

In recent years progress based on the use of Bayesian as well as Machine Learning (ML) techniques to deal with both the issues above has been made, from the creation of neural network based statistical emulators[7–9] in order to optimize the integration of chemical, radiative transfer and hydrodynamical models to the use of ML techniques to disentangle multiple gas components in unresolved beams.[10]

In this paper we will focus our attention to Bayesian and ML techniques applied to the study of chemical networks and the key parameters that govern their interactions. In recent years there has been a substantial body of work

concentrating on reducing the cost of solving chemical networks' computations using various techniques from Monte Carlo approaches to constrain important reactions,[11] to automated reduction schemes,[12,13] to topological methods[14,15] to, finally, ML algorithms.[9,16] In parallel several studies have concentrated on the estimation of poorly known reaction rates, with particular emphasis on surface chemical networks: an initial approach considered a simple grain-surface network and applied a Bayesian inference method coupled with Markov Chain Monte Carlo sampling in order to infer reaction rates.[11] This was followed up with an approach that considered the topological structure of the network,[15] while another exploited the characteristics of the chemical reaction mechanism to significantly reduce the dimensionality of the problem under consideration by simply considering the binding energies and the role they play in the determination of grain-surface chemistry.[17] Subsequent work using the 'Massive Optimised Parameter Estimation and Data compression' (MOPED) algorithm, helped make predictions about which ice species needed to be detected to reduce the variance of binding energy estimates.[18]

Due to the significant role that binding energies play in grain-surface chemistry, we shall concentrate on the estimation of binding energies as well as on prioritization of the ice species that should be observed with instruments such as the JWST to better improve our understanding of their values. We will then use machine learning interpretability to consider the forward relationship between binding energies and the abundances of species of interest. Our methods are described in Sections 2. The results are presented in Section 3 and a brief conclusion is given in Section 4.

2 Methodology

In this section we first describe the chemical code we use and the chemical assumptions we make in our work (Section 2.1), followed by a description of the analytical approach we employ (Section 2.2).

2.1 The chemical code

All modelling in this work is done with the open-source astrochemical code UCLCHEM.‡[19] The chemistry of a collapsing dark cloud was modelled. The dark cloud collapsed isothermally at 10 K from 10^2 cm^{-3} to 10^6 cm^{-3} over 5 million years. The composition of the ices as a result of the ensuing chemistry was then compared to the recent ice observations with the James Webb Space Telescope (JWST).[20]

As this work focuses solely on grain-surface chemistry, it is pertinent to describe the details of the underlying reaction mechanisms we consider in this work. This will be used as justification to explain why binding energies are of such great importance in the context of this work.

In UCLCHEM, the main grain-surface reaction mechanism is the Langmuir–Hinshelwood mechanism.[21] The rate at which two species A and B react through diffusion is given by:

‡ https://uclchem.github.io/.

$$k_{AB} = \kappa_{AB} \frac{\left(k_{hop}^{A} + k_{hop}^{B}\right)}{N_{site} n_{dust}}, \quad (1)$$

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

In eqn (1), k_{hop}^{X} is the thermal hopping rate of species X on the grain surface which is given as:

$$k_{hop}^{X} = \nu_0 \exp\left(-\frac{E_D}{T_{gr}}\right), \quad (2)$$

where E_D is the diffusion energy of the species, T_{gr} is the grain temperature and ν_0 is the characteristic vibration frequency of species X. The diffusion energy is a fraction of the binding energy of the species, E_b.

The characteristic vibration frequency, ν_0, is defined as:

$$\nu_0 = \sqrt{\frac{2 k_b n_s E_b}{\pi^2 m}}, \quad (3)$$

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

The final term, κ_{AB}, which gives the reaction probability is:

$$\kappa_{AB} = \max\left(\exp\left(-\frac{2a}{\hbar}\sqrt{2\mu k_b E_A}\right), \exp\left(-\frac{E_A}{T_{gr}}\right)\right), \quad (4)$$

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 $a = 1.4$ Angstrom is the thickness of a quantum mechanical barrier that is used as the default in UCLCHEM. The reaction probability encodes 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.

Species do not necessarily need to react with each other on the grains. It is also possible for them to diffuse away from a potential reactant or evaporate. As such, a modification needs to be made to the κ_{AB} term to take this into account. This is the reaction-diffusion competition.[22,23] The reaction probability is now defined as:

$$\kappa_{AB}^{final} = \frac{p_{reac}}{p_{reac} + p_{diff} + p_{evap}}, \quad (5)$$

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

$$p_{reac} = \max(\nu_0^A, \nu_0^B)\kappa_{AB}, \quad (6)$$

$$p_{diff} = k_{hop}^{A} + k_{hop}^{B} \quad (7)$$

and

$$p_{evap} = \nu_0^A \exp\left(-\frac{E_b^A}{T_{gr}}\right) + \nu_0^B \exp\left(-\frac{E_b^B}{T_{gr}}\right). \quad (8)$$

The term κ_{AB} in eqn (1) is replaced with κ_{AB}^{final}.

Eqn (1)–(8) show that the key quantities are ν_0, k_{hop}^X, E_b and E_A. The first two are functions of the binding energies of the reacting species. We assume that the activation energies in eqn (4) are well-known, so do not include these as parameters to be estimated.

If we wish to better understand grain-surface diffusion-based chemistry, we must have accurate values of the binding energies of species. For most cases, at 10 K, the reactant with the lower binding energy will dominate the total hopping rate, due to the exponential dependence of the hopping rate on the diffusion energy. Across the literature, there is often significant disagreement when it comes to the values of binding energies.[24–26] While there exist many different methods of estimating these values,[27–29] we utilise a Bayesian inference approach.

The chemical network used consists of a gas-phase and ice-phase network. The gas-phase network is the UMIST network.[24] The ice network used is the same as in previous work,[18] but augmented with a sulphur network based on work done to explain the sulphur depletion problem.[30] The inclusion of the sulphur network is important, since recently sulphur-bearing species have been confirmed in the ices.[20]

2.2 Analytical approach

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

Bayes' Law was used to solve this inference problem. Given the data, $\boldsymbol{d}$, of abundance measurements of species, the probability distributions of the binding energies of interest are given by:

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

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 normalising factor and represents the marginalised likelihood. The posterior distribution represents the updated probability distribution of reaction rates based on the data, the prior distribution, and the physical model.

The prior for all binding energies was selected as a uniform distribution between 400 K and 2000 K. The abundance measurements, given in Table 1, were

Table 1 The abundances and uncertainties taken from McClure *et al.*[20] These abundances were taken from sources with an visual extinction, A_v, of 95

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

assumed to be Gaussian. The species without associated uncertainty, OCN, SO_2 and OCS, were given a relative uncertainty of 50%. Assuming a Gaussian distribution, the likelihood function can be specified as:

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

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 abundance measurements are indexed over.

The inference was implemented using the UltraNest Python package.[31] The package implements efficient methods to construct a neighbourhood to sample from, allowing for better convergence of the sampling of the likelihood.[32,33] The package conveniently also outputs the maximum likelihood-estimator, $\boldsymbol{E}_{\mathbf{ML}}$, which will be utilised later.

2.2.2 The MOPED algorithm. While our knowledge of the molecular inventory in the gas-phase is quite complete, we are still far from being confident about the ice composition as well as the ice chemistry. To this end, we employ the "Massive Optimised Parameter Estimation and Data compression" (MOPED) algorithm.[34–36]

The aim of the algorithm is to determine which of the M species in our chemical network would best constrain our knowledge for our p binding energy parameters. In this work, $p = 9$ and $M = 119$. Some binding energies will have a greater influence on certain species than others. The key is to determine the species that are most sensitive to the binding energies of interest. In doing so, we can then make recommendations for future ice observations as was done in a proof-of-concept work recently.[18]

There will be uncertainty associated with all of our potential future abundance measurements. It is likely that these uncertainties will vary by species. However, it is difficult to determine these species-dependent uncertainties *a priori*. As such, we assume the uncertainty on each abundance measurement is the same. This is summarised in a covariance matrix: $\boldsymbol{C} = \mathrm{diag}({\sigma_1}^2, {\sigma_2}^2, \ldots {\sigma_M}^2)$.

We apply a filtering technique developed by Heavens *et al.*,[34,35,36] who propose using a linear combination of the final abundances of a network, $\boldsymbol{Y}$, to compress data points into numbers. Such a compression takes the form:

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

where α ranges from 1 to p and $\boldsymbol{b}_\alpha$ is a set of orthonormal linear filters. Each filter vector is unique to each parameter and does not contain information contained in any of the other vectors. $\boldsymbol{Y}$ represents a vector containing the final, steady-state abundances for some value of $\boldsymbol{E}$, though we employ $\boldsymbol{E} = \boldsymbol{E}_{\mathbf{ML}}$, which can be obtained from the Bayesian inference, as this has been found to be sufficient as the fiducial model.[34,35] For each c_α, there will be greater dependence on some of the components of $\boldsymbol{b}_\alpha$ than others. As each component represents a different species, this implies that a component with a greater magnitude has more information about that parameter.

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

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

and

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

where $\boldsymbol{Y}_{,\alpha}$ is the partial derivative of $\boldsymbol{Y}$ with respect to the parameter α around the point $\boldsymbol{Y} = f(\boldsymbol{E}_{\mathbf{ML}})$. The equations for $\boldsymbol{b}_\alpha$ were derived *via* a Lagrange multiplier procedure.[34] When it is said that all filters are orthonormal, this means that

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

which is another way of saying that all filter vectors are uncorrelated. Each component of $\boldsymbol{b}_\alpha$ is weighted:

- inversely by the size of the uncertainties associated with each species, as encoded by the covariance matrix.
- by the sensitivity of the species' abundance to the value of the binding energy, which is represented by the $\boldsymbol{Y}_{,\alpha}$.

If one wished to obtain a ranking of species in terms of their importance in helping constrain binding energies, one would need to come up with a 'score' for each species. Recall that as the magnitude of each component of $\boldsymbol{b}_\alpha$ is a weight for that species' influence on the parameter α, one would need to sum over the absolute values of the components of $\boldsymbol{b}_\alpha$ for species across all α. That is, we perform the sum over our linear filters

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

We now have a "filter sum" for each of the M species in our network, which serves as a means of comparing the importance of each species in helping us better constrain binding energy distributions. A species with a larger filter sum will have a larger influence in helping constrain the p binding energies.

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

As UCLCHEM solves a system of coupled ordinary differential equations, it stands to reason that the relationship between the input parameters (the binding energies) and the output parameters (the abundances of species) is non-linear. As such, the relationship between the input and output is not necessarily intuitive and is likely to be different for various 'binding energy regimes'. We make use of machine learning interpretability to help uncover this relationship.

In order to better understand the relationships between the inputs and outputs, we utilise SHapley Additive exPlanations (SHAP).[37] SHAP approximates Shapley values: these are measures of the marginal contribution of a feature to the output value, relative to the mean value of all outputs in the dataset.[38] This is done by considering various coalitions of feature values. A coalition of features represents all subsets of the total set of features. The Shapley value of a feature represents the average change in the prediction when that feature is included in the coalition of features selected. This change is assessed by considering the change in the prediction when the feature is included, averaged over all coalitions.[39] However, this becomes computationally unfeasible as the number of features grows, as the number of subsets grows exponentially with the number of features. SHAP is particularly useful, as it approximates the Shapley values, greatly reducing the time taken to compute them. This is done through the use of the TreeSHAP algorithm.[40]

500 000 data points were created from UCLCHEM by using a Latin Hypercube sampling scheme[41] implemented with the help of the Python surrogate modelling toolbox.[42] We employ the XGBoost Python package§ to build an XGBoost regressor[43] that is made to fit the relationship between the input parameters and the output abundance for each species.

3 Results

3.1 Results of the Bayesian inference

At first, the Bayesian inference was run using the original dataset. However, it was found that despite running the inference in parallel using MPI over 128 cores, that there was no convergence, even after several days. This was attributed to the fact that the model struggled to match the constraints. Many of these constraints have very low relative error, compared to the data used in previous works which typically had relative errors of the order of 50%.[11,15,17] A nested sampler will move from areas of low likelihood to areas of high likelihood. However, if the model struggles to find combinations of parameters that lead to a higher likelihood, then it will inevitably take longer to perform the inference. To properly run the inference, a significantly larger computing cluster would be required. As an alternative, we decided to investigate how the relative error, ε, impacted the obtained posterior probability distributions. We used values of 0.5, 0.33 and 0.25

§ https://xgboost.readthedocs.io/en/stable/index.html.

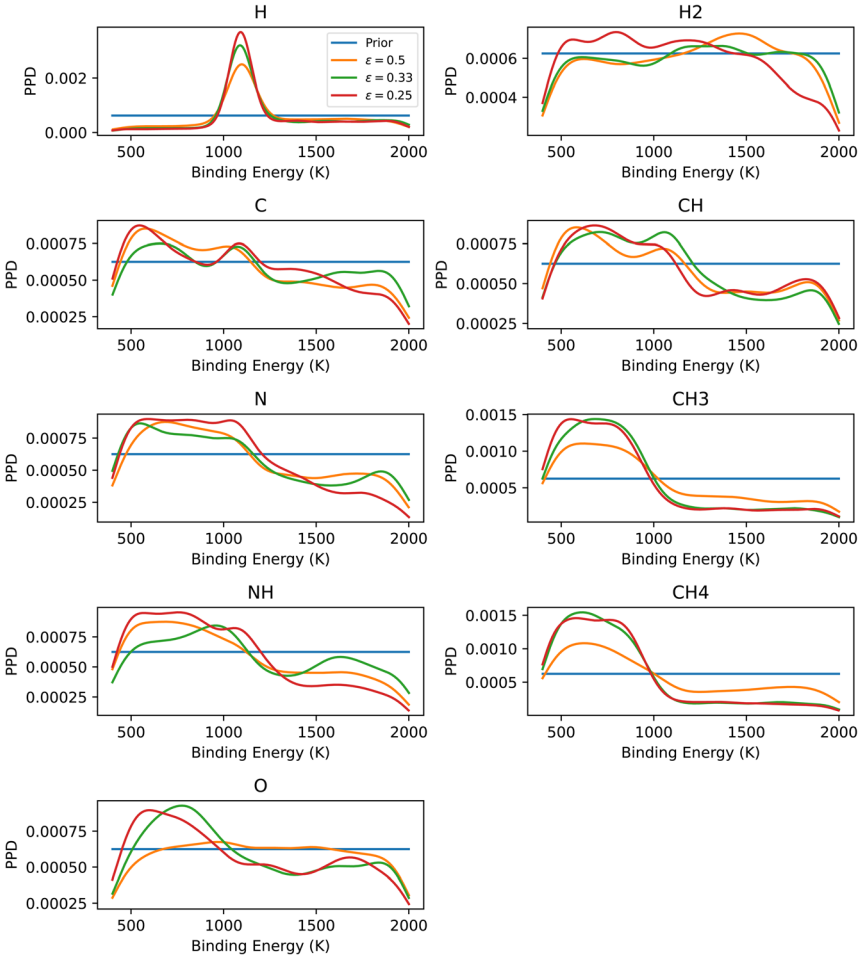

Fig. 1 Marginalised posterior distributions of the binding energies of the diffusive species we consider of interest in this work. We also plot the uniform prior distribution. Only H's binding energy marginalised posterior distribution differs significantly from the prior distribution. For the other binding energies, there is less difference. This is due to the lack of enough sufficiently constraining data. We also observe that decreasing the value of ε in general decreases the variance of the distribution. Both of these points motivate the need for further ice observations to reduce the variance of the distributions.

and ran the inference each time. Our results are displayed in Fig. 1. Also plotted are the prior distributions.

We observe that with the exception of hydrogen's binding energy, the binding energy posteriors are prior-dominated. However, it can also be seen that a decrease in the relative error of the data appears to be accompanied by a decrease in the variance of some of the posteriors, such as for CH_3, CH_4, NH and O. This is consistent with lower variance posteriors for H and O binding energies with the artificially reduced uncertainties for H_2O observation in prior work.[18] However, even in this scenario we are finding that our posteriors have a relatively

large variance. The best way to address this is to figure out which other species we should observe to further constrain the distributions.

3.2 Using the MOPED algorithm

We now look to analyse the results of the MOPED algorithm. The fiducial model we use is the one with $\varepsilon = 0.25$. In Fig. 2 we plot the filter sums for each species to provide us with an initial ranking. We only consider species formed on the grains. As the UCLCHEM code models both the bulk and the surface abundances, we sum the abundances of each species on the surface as well as in the bulk to provide us with a total abundance on the ices.

However, in order to inform future ice observations, it would be useful to also consider the likely abundances of each species. Ideally, we would wish to observe

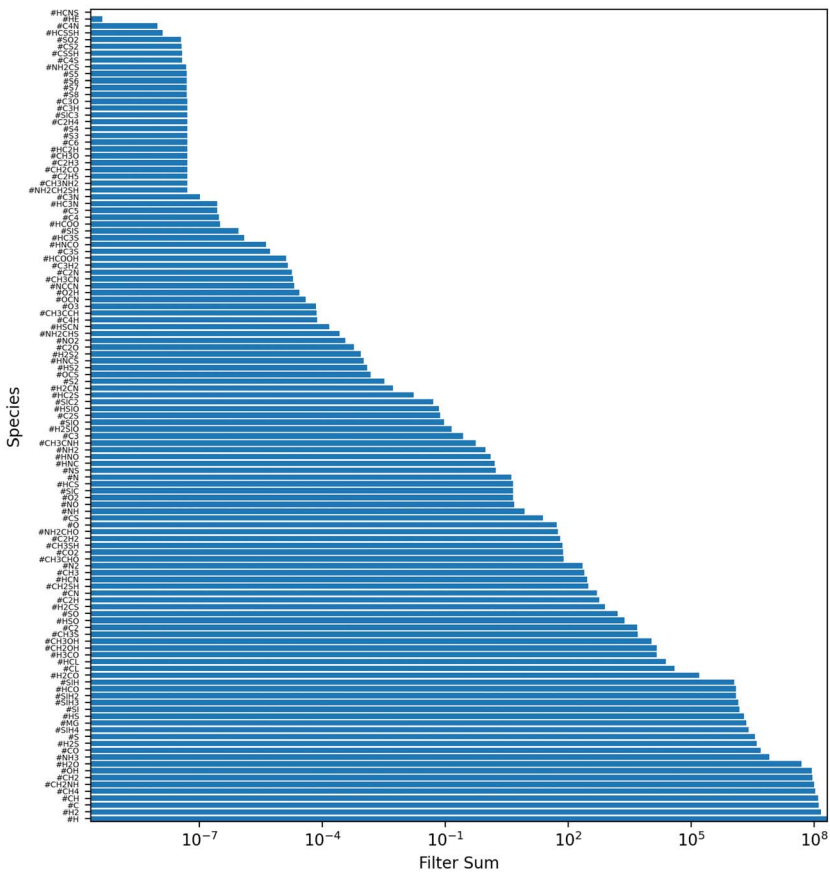

Fig. 2 Bar chart displaying the filter sums for all grain-surface species. Species with a larger filter sum are higher priority detection targets, as they are more affected by the binding energies of the species we consider. Some of the highest-ranked species have already been detected, which potentially implies that future observations should aim to improve the level of precision of these abundance measurements.

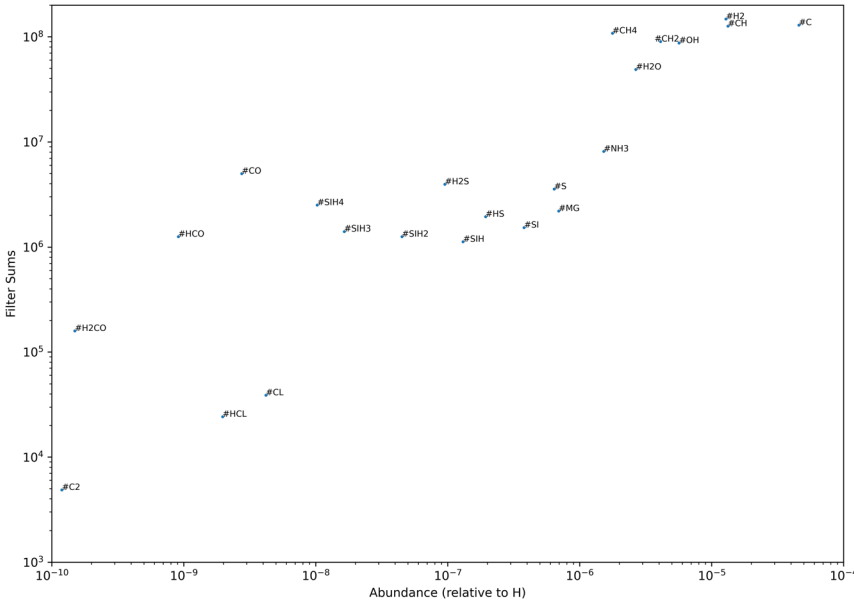

Fig. 3 Scatter plot depicting filter sum against the predicted abundances when the maximum-likelihood estimate for the binding energies is input into UCLCHEM. Given constraints on instrumental uncertainties, we should look to prioritise species that are not only important, as determined by their filter sums, but that can also be realistically detected. These include saturated species such as #CH_4, #NH_3, #SiH_4, #H_2S and #H_2O, as well as their precursors. We find that many of the species we observe are the intermediate species formed during the creation of the saturated species in Table 1. This indicates that understanding these intermediate products is essential to better constraining the binding energies of interest.

species that are highly abundant and that have large filter sums. The first requirement means it is easier to observe a species given a particular instrumental uncertainty, whilst the second ensures that we are observing species that are dependent on the binding energies and are therefore relevant to the chemistry we are considering. To do this, we plot the filter sum of each species against the abundance produced when we use binding energies equal to $\boldsymbol{E}_{\mathbf{ML}}$. The resulting plot is shown in Fig. 3. We only consider species with an abundance greater than 10^{-10} relative to H, as anything less abundant is unlikely to be detected in the ices. As in previous work,[18] we observe that the species H_2O, CH_4, NH_3, H_2S, SiH_4, CO and H_2CO are amongst the highest-ranked species with abundances that are predicted to be detectable. These species all have modes in the range considered by JWST. Unlike in previous work, however, CO_2, CH_3OH and HCN are not amongst the most significant species. This can be attributed to the fiducial model, as we used different constraints, which lead to the maximum-likelihood estimate being different.

3.3 Insights from the machine learning interpretability

Previously, we considered the impact of the data, *i.e.* the species abundances, on the binding energy values and their distributions. We now wish to consider the

opposite situation, which is the impact of the binding energy values on the final steady-state abundances of molecules of interest. This is important to consider as the binding energy of a species can be dependent on the ice-composition as well as on the individual sites.[44,45]

In the interest of brevity, we consider a subset of the molecules so as to demonstrate the effectiveness of this approach as a proof-of-concept. We are interested in better understanding the importance of each of the features in predicting the final abundance of a species of interest, as well as the relative importances of the features. Fig. 4 and 5 are so-called beeswarm plots for H_2O and CO respectively. The features are listed from top to bottom in decreasing order of importance to the model output. Along the horizontal axis, individual predictions are plotted in terms of their SHAP value. Recall that the SHAP value states the difference in the value of the model output for that prediction relative to the global average. Furthermore, the points are colour-coded in terms of the size of the feature value. From this, we can attempt to better understand the directionality of each feature's relationship with the output.

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

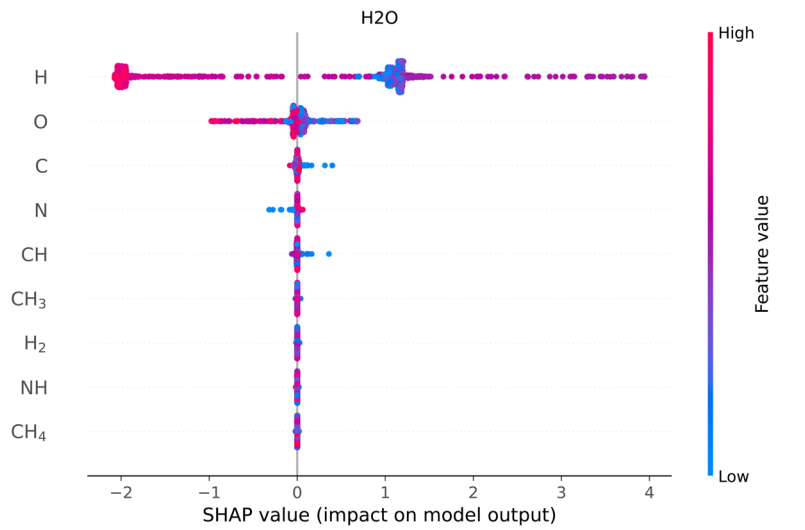

Fig. 4 A beeswarm plot for the statistical emulator trained to predict H_2O's abundance. The features are listed from top to bottom in decreasing order of importance to the model output. Along the horizontal axis, individual predictions are plotted in terms of their SHAP value, that is the change to the log-abundance relative to the average value in the dataset. Recall that the SHAP value states the difference in the value of the model output for that prediction relative to the global average. Furthermore, the points are colour-coded in terms of the size of the feature value. We observe that the binding energies of H and O are the most important features. This makes sense, as both species are necessary to form water *via* successive hydrogenations of an oxygen atom.

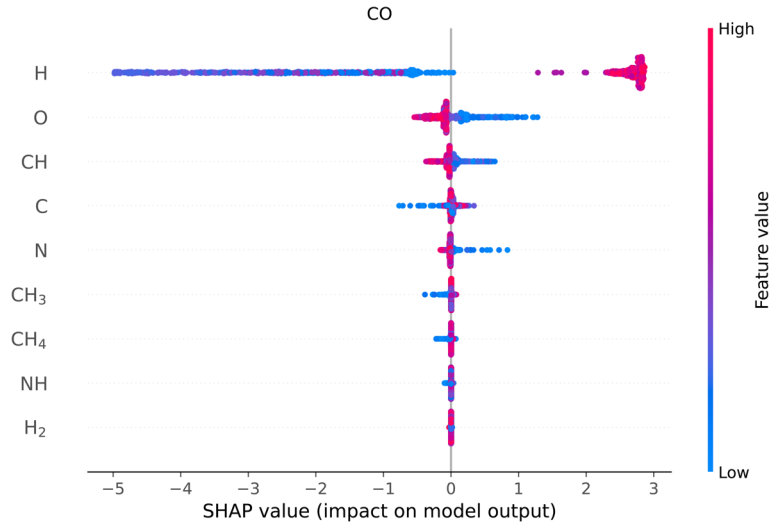

Fig. 5 A beeswarm plot for the statistical emulator trained to predict CO's abundance. The features are listed from top to bottom in decreasing order of importance to the model output. Along the horizontal axis, individual predictions are plotted in terms of their SHAP value, that is the change to the log-abundance relative to the average value in the dataset. Recall that the SHAP value states the difference in the value of the model output for that prediction relative to the global average. Furthermore, the points are colour-coded in terms of the size of the feature value. We observe that the binding energies of H and O are the most important features. Increasing H's binding energy appears to increase CO's abundance, which can be attributed to a decrease in the efficiency of the hydrogenation of CO.

We can also consider the exact nature of the relationship between the features and the final abundance. To do this, we consider the partial dependence of specific variables relative to the output variable. The partial dependence is defined as the marginal effect of one or several features on the output of a machine learning model.[39,46] To demonstrate the utility of the partial dependence, we consider H_2O and CO. Both of these molecules are largely dependent on two binding energies: that of H and O. We plot their 1-D and 2-D partial dependences in Fig. 6 and 7. Note that the y-axis of the 1-D plots are simply the log-abundance of the respective species.

We observe that for water, there is a small area of parameter space in which the abundance peaks. This roughly matches the maximum-posterior hydrogen binding energy value obtained in Fig. 1. Despite the oxygen's binding energy being the second most important feature, we observe that over the range of binding energies considered, it has far less impact in changing the obtained water abundance. Even so, the parameter favoring binding energies lower than $\sim$1000 K for oxygen is consistent with the posterior for the inverse problem.

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

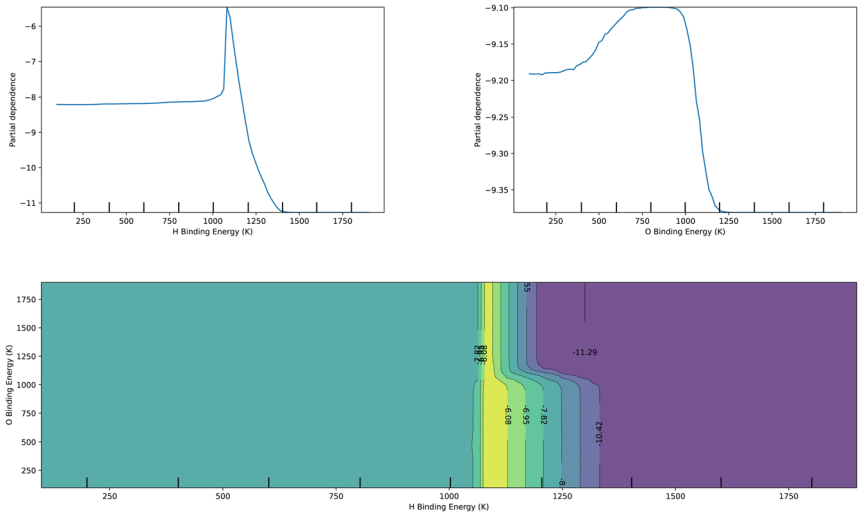

Fig. 6 Top: A plot of the 1-D partial dependence plots of the binding energies of H and O for water. The partial dependence represents the expected value of the log-abundance of water as a function of the variable in features, marginalised over all other features. We observe that for a narrow range of atomic hydrogen's binding energies at around 1100 K, there is a sharp increase in the abundance of water. This is roughly the point at which the marginalised posterior distribution for H's binding energy in Fig. 1 peaks. The dependence for O's binding energy shows a similar consistency with the posteriors, having a clear preference for energies smaller than ~1000 K. Bottom: A 2-D partial dependence plot for the binding energies of H and O. Yellow represents the region with the highest abundance of water.

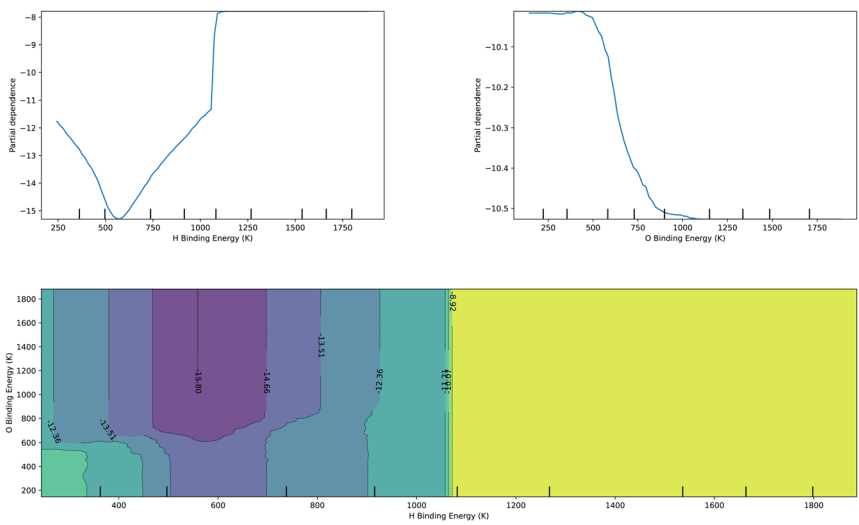

Fig. 7 Top: A plot of the 1-D partial dependence plots of the binding energies of H and O for CO. The partial dependence represents the expected value of the log-abundance of CO as a function of the variable in features, marginalised over all other features. Bottom: A 2-D partial dependence plot for the binding energies of H and O. Yellow represents the region with the highest abundance of CO.

4 Conclusions

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

• As in our previous work, we find that Bayesian inference can be a very useful tool to constrain binding energies. However further ice observations are needed in order to reduce the variance of the distributions.

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

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

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

Author contributions

Johannes Heyl: conceptualization, data curation, formal analysis, validation, writing – original draft, writing – review & editing. Serena Viti: conceptualization, data curation, formal analysis, writing – original draft, writing – review & editing. Gijs Vermariën: writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

J. Heyl 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.

Notes and references

1 T. G. Bisbas, T. A. Bell, S. Viti, M. J. Barlow, J. Yates and M. Vasta, *Mon. Not. R. Astron. Soc.*, 2014, **443**, 111–121.

2 S. Viti, S. García-Burillo, A. Fuente, L. K. Hunt, A. Usero, C. Henkel, A. Eckart, S. Martin, M. Spaans, S. Muller, F. Combes, M. Krips, E. Schinnerer, V. Casasola, F. Costagliola, I. Marquez, P. Planesas, P. P. van der Werf, S. Aalto, A. J. Baker, F. Boone and L. J. Tacconi, *Astron. Astrophys.*, 2014, **570**, A28.
3 M. V. Kazandjian, I. Pelupessy, R. Meijerink, F. P. Israel, C. M. Coppola, M. J. F. Rosenberg and M. Spaans, *Astron. Astrophys.*, 2016, **595**, A124.
4 K. Y. Huang, S. Viti, J. Holdship, S. García-Burillo, K. Kohno, A. Taniguchi, S. Martn, R. Aladro, A. Fuente and M. Sánchez-García, *Astron. Astrophys.*, 2022, **666**, A102.
5 A. Makrymallis and S. Viti, *Astrophys. J.*, 2014, **794**, 45.
6 C. Lefèvre, L. Pagani, M. Juvela, R. Paladini, R. Lallement, D. J. Marshall, M. Andersen, A. Bacmann, P. M. McGehee, L. Montier, A. Noriega-Crespo, V. M. Pelkonen, I. Ristorcelli and J. Steinacker, *Astron. Astrophys.*, 2014, **572**, A20.
7 D. de Mijolla, S. Viti, J. Holdship, I. Manolopoulou and J. Yates, *Astron. Astrophys.*, 2019, **630**, A117.
8 J. Holdship, S. Viti, T. J. Haworth and J. D. Ilee, *Astron. Astrophys.*, 2021, **653**, A76.
9 T. Grassi, F. Nauman, J. P. Ramsey, S. Bovino, G. Picogna and B. Ercolano, *Astron. Astrophys.*, 2022, **668**, A139.
10 D. de Mijolla, J. Holdship, S. Viti and J. Heyl, *Astrophys. J.*, submitted.
11 J. Holdship, N. Jeffrey, A. Makrymallis, S. Viti and J. Yates, *Astrophys. J.*, 2018, **866**, 116.
12 T. Grassi, S. Bovino, F. A. Gianturco, P. Baiocchi and E. Merlin, *Mon. Not. R. Astron. Soc.*, 2012, **425**, 1332–1340.
13 R. Xu, X.-N. Bai, K. Öberg and H. Zhang, *Astrophys. J.*, 2019, **872**, 107.
14 T. Grassi, S. Bovino, D. Schleicher and F. A. Gianturco, *Mon. Not. R. Astron. Soc.*, 2013, **431**, 1659–1668.
15 J. Heyl, S. Viti, J. Holdship and S. M. Feeney, *Astrophys. J.*, 2020, **904**, 197.
16 K. S. Tang and M. Turk, *arXiv e-prints*, 2022, preprint, arXiv:2207.07159, DOI: **10.48550/arXiv.2207.07159**.
17 J. Heyl, J. Holdship and S. Viti, *Astrophys. J.*, 2022, **931**, 26.
18 J. Heyl, E. Sellentin, J. Holdship and S. Viti, *Mon. Not. R. Astron. Soc.*, 2022, **517**, 38–46.
19 J. Holdship, S. Viti, I. Jiménez-Serra, A. Makrymallis and F. Priestley, *Astron. J.*, 2017, **154**, 38.
20 M. K. McClure, W. R. M. Rocha, K. M. Pontoppidan, N. Crouzet, L. E. U. Chu, E. Dartois, T. Lamberts, J. A. Noble, Y. J. Pendleton, G. Perotti, D. Qasim, M. G. Rachid, Z. L. Smith, F. Sun, T. L. Beck, A. C. A. Boogert, W. A. Brown, P. Caselli, S. B. Charnley, H. M. Cuppen, H. Dickinson, M. N. Drozdovskaya, E. Egami, J. Erkal, H. Fraser, R. T. Garrod, D. Harsono, S. Ioppolo, I. Jiménez-Serra, M. Jin, J. K. Jørgensen, L. E. Kristensen, D. C. Lis, M. R. S. McCoustra, B. A. McGuire, G. J. Melnick, K. I. Ã-berg, M. E. Palumbo, T. Shimonishi, J. A. Sturm, E. F. van Dishoeck and H. Linnartz, *Nat. Astron.*, 2023, **7**, 431–443.
21 T. I. Hasegawa, E. Herbst and C. M. Leung, *Astrophys. J., Suppl. Ser.*, 1992, **82**, 167.
22 Q. Chang, H. M. Cuppen and E. Herbst, *Astron. Astrophys.*, 2007, **469**, 973–983.

23 R. T. Garrod and T. Pauly, *Astrophys. J.*, 2011, **735**, 15.
24 D. McElroy, C. Walsh, A. J. Markwick, M. A. Cordiner, K. Smith and T. J. Millar, *Astron. Astrophys.*, 2013, **550**, A36.
25 V. Wakelam, J. C. Loison, R. Mereau and M. Ruaud, *Mol. Astrophys.*, 2017, **6**, 22–35.
26 D. Quénard, I. Jiménez-Serra, S. Viti, J. Holdship and A. Coutens, *Mon. Not. R. Astron. Soc.*, 2018, **474**, 2796–2812.
27 J. He, K. Acharyya and G. Vidali, *Astrophys. J.*, 2016, **825**, 89.
28 S. Ferrero, L. Zamirri, C. Ceccarelli, A. Witzel, A. Rimola and P. Ugliengo, *Astrophys. J.*, 2020, **904**, 11.
29 T. Villadsen, N. F. W. Ligterink and M. Andersen, *Astron. Astrophys.*, 2022, **666**, A45.
30 J. C. Laas and P. Caselli, *Astron. Astrophys.*, 2019, **624**, A108.
31 J. Buchner, *J. Open Source Softw.*, 2021, **6**, 3001.
32 J. Buchner, *Stat. Comput.*, 2016, **26**, 383–392.
33 J. Buchner, *Publ. Astron. Soc. Pac.*, 2019, **131**, 108005.
34 A. F. Heavens, R. Jimenez and O. Lahav, *Mon. Not. R. Astron. Soc.*, 2000, **317**, 965–972.
35 A. F. Heavens, E. Sellentin, D. de Mijolla and A. Vianello, *Mon. Not. R. Astron. Soc.*, 2017, **472**, 4244–4250.
36 A. F. Heavens, E. Sellentin and A. H. Jaffe, *Mon. Not. R. Astron. Soc.*, 2020, **498**, 3440–3451.
37 S. M. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.
38 L. S. Shapley, in *17. A Value for n-Person Games*, ed. H. W. Kuhn and A. W. Tucker, Princeton University Press, Princeton, 2016, pp. 307–318.
39 C. Molnar, *Interpretable Machine Learning*, 2nd edn, 2022.
40 S. M. Lundberg, G. G. Erion and S.-I. Lee, *arXiv e-prints*, 2018, preprint, arXiv:1802.03888, DOI: **10.48550/arXiv.1802.03888**.
41 M. D. McKay, R. J. Beckman and W. J. Conover, *Technometrics*, 1979, **21**, 239–245.
42 M. A. Bouhlel, J. T. Hwang, N. Bartoli, R. Lafage, J. Morlier and J. R. R. A. Martins, *Adv. Eng. Softw.*, 2019, 102662.
43 T. Chen and C. Guestrin, *arXiv e-prints*, 2016, preprint, arXiv:1603.02754, DOI: **10.48550/arXiv.1603.02754**.
44 T. Grassi, S. Bovino, P. Caselli, G. Bovolenta, S. Vogt-Geisse and B. Ercolano, *Astron. Astrophys.*, 2020, **643**, A155.
45 A. Das, M. Sil, R. Ghosh, P. Gorai, S. Adak, S. Samanta and S. K. Chakrabarti, *Front. Astron. Space Sci.*, 2021, **8**, 78.
46 J. H. Friedman, *Ann. Stat.*, 2001, **29**, 1189–1232.