Astronomy & Astrophysics manuscript no. aanda
July 11, 2022

Predicting binding energies of astrochemically relevant molecules via machine learning

T. Villadsen[1], N.F.W. Ligterink[2] and M. Andersen[1,3]

[1] Department of Physics and Astronomy - Center for Interstellar Catalysis, Aarhus University, Aarhus C, DK-8000 Denmark
e-mail: mie@phys.au.dk
[2] Physics Institute, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland
e-mail: niels.ligterink@unibe.ch
[3] Aarhus Institute of Advanced Studies, Aarhus University, Aarhus C, DK-8000 Denmark

Received May 23, 2022; accepted

ABSTRACT

Context. The behaviour of molecules in space is to a large extent governed by where they freeze out or sublimate. The molecular binding energy is thus an important parameter for many astrochemical studies. This parameter is usually determined with time-consuming experiments, computationally expensive quantum chemical calculations, or the inexpensive, but inaccurate, linear addition method.
Aims. In this work we propose a new method based on machine learning for predicting binding energies that is accurate, yet computationally inexpensive.
Methods. A machine learning model based on Gaussian Process Regression is created and trained on a database of binding energies of molecules collected from laboratory experiments presented in the literature. The molecules in the database are categorized by their features, such as mono- or multilayer coverage, binding surface, functional groups, valence electrons, and H-bond acceptors and donors.
Results. The performance of the model is assessed with five-fold and leave-one-molecule-out cross validation. Predictions are generally accurate, with differences between predicted and literature binding energies values of less than ±20%. The validated model is used to predict the binding energies of twenty one molecules that have recently been detected in the interstellar medium, but for which binding energy values are not known. A simplified model is used to visualize where the snowlines of these molecules would be located in a protoplanetary disk.
Conclusions. This work demonstrates that machine learning can be employed to accurately and rapidly predict binding energies of molecules. Machine learning complements current laboratory experiments and quantum chemical computational studies. The predicted binding energies will find use in the modelling of astrochemical and planet-forming environments.

Key words. Astrochemistry – ISM: molecules – molecular processes – molecular data

1. Introduction

One of the objectives of the field of astrochemistry is to understand the formation, destruction, and survival of molecules in astrophysical environments, such as star-forming regions or planet-forming disks (Jørgensen et al. 2020; Öberg & Bergin 2021). This concerns, for instance, the molecular reactions on dust-grain surfaces, where the surface acts as a catalyst (Cuppen et al. 2017) and evaporation of ices in hot cores near young massive stars (Viti et al. 2004). To model such phenomena certain molecule-specific parameters are needed. Both examples require a measure of how strongly the molecule binds to the surface, that is the binding energy (BE). In astrophysical environments physisorption is the primary source of the BE. The main contributors of physisorption are van der Waals forces, which arises from dipole-dipole interactions, and hydrogen bonding between the adsorbed molecule and surface. These forces act at a distance between 2-4 Å. If no dynamical barrier is present, the BE is equal to the activation energy of desorption ($E_{\rm des}$) as the system reference energy (i.e., $E = 0$) corresponds to the adsorbed molecule and surface at infinite separation Minissale et al. (2022).

Experimentally, a prominent technique to determine the BE is by temperature programmed desorption (TPD). This applies to studies of catalysis (Luo et al. 1997), surface science (Zhou et al. 2007), as well as astrochemistry (Muñoz Caro et al. 2010). The TPD process consists of three steps; first, the molecule is adsorbed to the surface at cold temperatures. The coverage may be below or above a single monolayer depending on various factors such as the molecular flux and deposition time. If more molecules than available surface adsorption sites are deposited, the coverage is also termed multilayer. Second, the temperature is linearly increased, resulting in desorption of molecules at specific temperatures. Third, the desorbed molecules are detected, often with a mass spectrometer. This produces spectra of desorption rate as a function of temperature. The process of thermal desorption is generally described by the Polanyi-Wigner equation, a modified Arrhenius law:

$$-\frac{dN}{dt} = k_{\rm des}(T) \cdot N^{\rm n}, \tag{1}$$

$$-\frac{dN}{dt} = \nu_{\rm n} \cdot N^{\rm n} \cdot \exp\left(-\frac{E_{\rm des}}{k_{\rm B} \cdot T}\right), \tag{2}$$

where $k_{\rm des}(T)$ is the desorption rate constant in $\rm s^{-1}$ at temperature T, N the number of adsorbed molecules on a surface, n the

order of desorption (usually 1 for monolayer desorption and 0 for multilayer desorption), ν_n the pre-exponential frequency factor with value molecules^{1-n} s^{-1} (also often denoted as A), E_{des} the desorption energy and k_B the Boltzmann constant. For a more thorough discussion of the technique and analysis, we refer to De Jong & Niemantsverdriet (1990), Burke & Brown (2010), and Minissale et al. (2022) for contextual reviews.

While TPD has been successful in determining the BEs for many molecules that are of astrochemical relevance (e.g., Brown & Bolina 2007; Burke et al. 2015a; Smith & Kay 2018; Behmard et al. 2019; Salter et al. 2019), there are limitations to experimental BE investigations with this and other techniques. Experiments are time-consuming and the focus usually is on the scientifically most impactful systems. With the vast number of known interstellar molecules, this inevitably means that some of them are not yet studied. Furthermore, certain molecular species are difficult to work with, either because they are unstable or highly reactive (e.g., vinyl alcohol, cyanopolyynes), highly toxic (e.g., methyl isocyanate, propyl cyanide), or simply challenging to produce an ice film with (e.g, carbamide).

Alternatively, Bayesian inference (Heyl et al. 2022) or quantum chemical computational methods can be used to determine BEs (e.g., Das et al. 2018; Rimola et al. 2018; Balbisi et al. 2022). Quantum chemical calculation can also take into account BE distributions on amorphous and highly anisotropic surfaces (e.g., Tinacci et al. 2022; Ferrero et al. 2020; Duflot et al. 2021). However, since these methods are computationally expensive, many astrochemical studies rely on the so-called linear addition method. With this method, the BE of a molecule is determined by splitting its components in atoms and molecular fragments for which the BEs are known and subsequently adding them together (e.g., Garrod & Herbst 2006; Shingledecker et al. 2020). This method is computationally inexpensive, but it is also inaccurate. Novel methods that are computationally inexpensive, but more accurate are required.

Machine learning (ML) has become one of the most prominent scientific tools of the 21st century as it provides high accuracy at low computational cost. It has the ability to handle and interpret data in ways impossible to humans, which allows for the discovery of unprecedented patterns (Jordan & Mitchell 2015). These properties make ML an interesting alternative to the above-mentioned theory-based approaches. With immense versatility ML has applications ranging from self-driving cars and social media to banking and image recognition. In recent years, ML has also made its entrance as a powerful tool in astrochemistry and astrophysics. Notable deployments include Lee et al. (2021) for reproducing and predicting chemical abundances in interstellar inventories and Shallue & Vanderburg (2018) for exoplanet identification. Another type of ML models are ML interatomic potentials, which can be directly employed as low-cost alternatives to quantum chemical calculations for investigating for example molecular reactivity, adsorption and diffusion on dust grains (e.g., Mazo-Sevillano et al. 2021; Molpeceres, G. et al. 2021; Zaverkin et al. 2021). In surface science and catalysis, ML has also been used extensively to identify predictive models for BEs (e.g., Gu et al. 2020; Fung et al. 2021; Andersen & Reuter 2021).

In this work, we apply supervised ML to a data set of BEs obtained from literature TPD data and thereby develop a model to predict BEs between new molecules and surfaces relevant to astrochemical environments. The methods used are discussed in Sect. 2. The results and discussion of the analysis are presented in Sect. 3. Astrophysical implications are covered in Sect. 4 and the conclusions of this work are given in Sect. 5.

2. Methods and data

Supervised learning algorithms are constructed to make a model that can recognise particular patterns within the data when given training examples by the user (supervisor). A strong limitation of such algorithms is that they can only recognise data that are related to the training data, therefore, any anomaly or unseen data structures would be difficult for the model to grasp. The training data given to the model in our work is a data set of BEs obtained from TPD experiments collected from the literature as well as relevant features of each system such as the surface category and atoms and functional groups present in the adsorbed molecule. Hence, the trained model can be expected to predict BEs for new examples of molecules and surfaces that are not too different from those seen in the training data. We quantify the predictive accuracy of our model by carrying out two types of cross validation analysis. The workflow of the process is shown in Fig. 1. In the following sections, the essential components of this workflow are described.

2.1. Gaussian process regression

BEs are here predicted using the supervised ML technique Gaussian Process Regression (GPR). It is a probabilistic, non-parametric supervised learning method frequently used for regression and classification problems in the ML community. Being based on Bayesian probability theory, it learns a posterior probability distribution over all admissible target functions. Here, these are functions describing the relationship between surface/molecular features (the input x) and the TPD BE (the output y). A Gaussian Process prior is assumed, which means that both the prior and posterior probability distributions are Gaussian distributed (normal) and can be specified using a mean function, $\mu(x)$, and a covariance function, $k(x, x')$, also called the kernel function. The posterior distribution is calculated by conditioning the prior distribution on the training data set. Model predictions on new test data points with input x_* are obtained from the mean of the posterior distribution, $\bar{\mathbf{f}}_*$, given by

$$\bar{\mathbf{f}}_* = \boldsymbol{\mu_*} + k(x_*, x)[k(x, x) + \sigma_n^2 I]^{-1}(\mathbf{y} - \boldsymbol{\mu}) \quad , \tag{3}$$

and variances are obtained from the diagonal of the covariance matrix, cov($\mathbf{f}_*$), given by

$$\mathrm{cov}(\mathbf{f}_*) = k(x_*, x_*) - k(x_*, x)[k(x, x) + \sigma_n^2 I]^{-1}k(x, x_*) \quad . \tag{4}$$

Here $\boldsymbol{\mu}$ and $\boldsymbol{\mu_*}$ are the mean vectors, $k(x, x_*)$ denotes the covariance matrix evaluated at all pairs of training and test points, and similarly for the other entries $k(x, x)$, $k(x_*, x_*)$ and $k(x_*, x)$. The target function is assumed to be noisy, which is accounted for by the incorporation of independently, identically distributed Gaussian noise, $\sigma_n^2 I$. In practise, a small or vanishing noise level will cause the fitted model to follow the training data points closely, whereas a higher noise level will result in a smoother model. The latter can be useful for extrapolating to unseen data since more emphasis is put on trends rather than the individual training examples seen. The variance provides an uncertainty estimate, that is how confident we can be about the model predictions, and it is also affected by the noise level. The direct access to an uncertainty estimate is a great advantage of GPR compared to other types of ML methods such as neural networks (Scalia et al. 2020).

GPR belongs to the class of kernelized ML methods that employ internally the "kernel trick". A kernel is a function that corresponds to an inner product in some high-dimensional feature

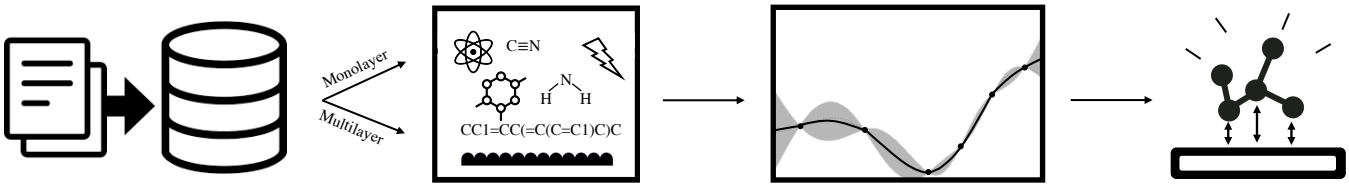


Fig. 1: Schematic overview of the workflow. First the data is collected from the literature and divided into monolayer and multilayer coverage. Secondly, specific features are designated to the data, including atomic composition, functional groups and valence electrons. Thirdly, the Gaussian Process Regression model is constructed and trained on the data to be able to predict BEs for new molecules.

space, whose values can be interpreted as a similarity measure of input data points. In this way, the kernel function implicitly maps the inputs into a higher dimensional space and applies the linear algorithm there, which is the basic idea how kernelized ML methods tackle the computational complexity of dealing with high-dimensional feature spaces. We refer to Rasmussen & Williams (2006) for an extensive textbook coverage of GPR and to Gibson et al. (2012)) and Aigrain et al. (2012) for examples of GPR applied to exoplanet data sets.

For the actual GPR implementation, we rely in this work on the ML library Scikit-learn (Pedregosa et al. 2011), which provides a number of built-in kernels. By testing several different combinations of kernels using five-fold cross validation (cf. Sect. 3.1), we found that the best performance is achieved by using the sum of the radial basis function (RBF) kernel

$$k_{\mathrm{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\frac{1}{2\ell_1^2}(\mathbf{x}_i - \mathbf{x}_j)^2\right] \tag{5}$$

and the rational quadratic (RQ) kernel

$$k_{\mathrm{RQ}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[1 + \frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2\alpha\ell_2^2}\right]^{-\alpha} . \tag{6}$$

The length-scale parameters, ℓ_1 and ℓ_2, indicate how quickly the correlation between two points drops as their distance increases. A higher ℓ gives a smoother function and a smaller ℓ gives a wigglier function (Rasmussen & Williams 2006). α determines the relative weighting of large-scale and small-scale variations in the RQ kernel (Duvenaud 2014). ℓ_1, ℓ_2 and α, along with the noise level from Eq. 3 and 4, are hyperparameters. Here we determine these parameters during the model training by maximising the marginal likelihood function using the standard Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm for bound-constrained optimization (L-BFGS-B) Byrd et al. (1995); Zhu et al. (1997).

2.2. Data set and features

We have compiled the training data by analysing laboratory studies presented in the literature and extracting the relevant information. From these publications, BEs determined with TPD experiments and information about the binding surface are retrieved, which resulted in a data set that initially contained 354 (167) entries for the monolayer (multilayer) case and 117 different molecules, many of which have been detected in the interstellar medium. They range from simple diatomics like N_2 and CO, organic molecules like ethanol (CH_3CH_2OH) and glycolaldehyde ($HOCH_2CHO$), long carbon chains such as octane (C_8H_{18}), and biomolecules like the nucleobase adenine.

Besides the BEs, the training data set also contains input features, which are descriptive attributes of the surfaces and molecules. It is essential to choose the best possible features as they govern how well the model predicts; too few features and the model will not be able to differentiate between different training data points (e.g. molecules), and too many features could invoke the curse of dimensionality, which is a term expressing how increasing the volume of feature space dilutes the data (Bellman 1966). Therefore, the best approach is to minimise the number of features while maximising the amount of information they contain.

In this work, molecular features, such as atomic compositions (C, H, O, N, Cl) and functional groups (alcohol, –OH; carbonyl, –C(O)–; carboxyl, –COOH; ester, –C(O)O–; ether, –O–; amine, –NH_2; cyanide, –CN; amide, –NC(O)–), are used. Several features are obtained from the python module RDkit (Landrum 2020). These are calculated by converting the molecules to SMILES[1] strings, which are then fed into RDkit. The considered features are the number of valence electrons, hydrogen bond donors, hydrogen bond acceptors, and topological polar surface area (TPSA). The latter is a property defined as the molecule's sum of surface area of polar atoms and is measured in Å^2. The motivation for applying the features from RDkit is to try to encapsulate the origin of the dominating forces that govern the binding of molecules. This includes in particular hydrogen bonding and van der Waals interactions. Finally, we also included the molecular mass and the number of atoms in the molecule as features. An overview of all molecular features used is given in Table 1 and a selection of the feature values of each molecule are presented in Table B.1.

For the consideration of features related to the surface, we first divide our data set into two categories; monolayer and multilayer coverage. Entries recorded at monolayer coverage (that is, an adsorbate layer of one molecule thickness) or less are assumed to have their BE dominated by the surface-molecule interaction. A wide variety of different surfaces are present in the analysed literature, the number of which greatly exceeds what is reasonable for the model to handle. To reduce this, the surfaces are placed in four different sub-categories based on their com-

[1] Simplified Molecular-Input Line-Entry System (SMILES) is a formalism widely used in the chemistry community to describe molecular structures as a string of ASCII characters. The SMILES string also encodes which chemical functional groups are present in the molecule.

Table 1: Overview of the features used to describe the molecules and surfaces.

Atoms	Functional Groups	RDKit & Misc.	Surface	Examples of surface
Carbon	Alcohol (–OH)	Number of H-bond acceptors	Carbon	Graphene, graphite and
Chlorine	Amide (–NC(O)–)	Number of H-bond donors		highly orientated pyrolytic graphite
Hydrogen	Amine ($–NH_2$)	Number of valence electrons	Metal	Gold and nickel
Nitrogen	Carbonyl (–C(O)–)	Topological polar surface area		
Oxygen	Carboxyl (–COOH)		Silicate	Amorphous silicate and forsterite
	Cyanide (–CN)	Mass		
	Ester (–C(O)O–)	Number of atoms	Water	Amorphous solid water and
	Ether (–O–)			crystalline water

mon traits. The categories and their main contributors annotated with the percentage they comprise of the total entries in the category are the following: Carbon (graphene, graphite and highly orientated pyrolytic graphite, 94 % of data entries), metal (gold, 56 % of data entries), silicate (amorphous silicate and forsterite, 93 % of data entries) and water (amorphous solid water and crystalline water, 98 % of data entries). We note that this approximation could be a source of significant noise in the training data, as different types of surfaces here placed in the same category (e.g. nickel and gold for metals) may in reality bind the studied molecules with different strength, however, they cannot be distinguished from the used input features. Finally, in order to input the surface category feature to the ML algorithm it is converted to numerical values using one hot encoding.

For multilayer entries, the coverage is greater than one molecule in thickness and the BE is assumed to be dominated by intermolecular interactions of the adsorbed species with itself. For this reason, we have chosen to neglect surface features for the multilayer data set.

Finally, we note that, while for further analysis only the BE is used, it is important to be aware of the influence of the pre-exponential factor. This value can be experimentally determined, but often (including in many of the studies that contribute to the data set in this work) an assumed value is used in the analysis to retrieve the BE. Compared to an experimentally determined pre-exponential factor, the assumed value may result in significantly deviating BEs and their inclusion in the training data will affect the results of the ML model. It is also worth noting that adsorbates on amorphous and highly anisotropic surfaces usually have a distribution of BEs rather than a single value (Shimonishi et al. 2018). However, since the available TPD data mainly consists of a single binding energy per surface/molecule combination, this distribution is presently not possible to include in our model.

2.3. Data preparation

Since ML algorithms generally do not perform well on data points that are rare or very different from the remaining of the data set, we combed our data set for outliers. This was done by applying the 'isolation forest algorithm' from Scikit-learn (Pedregosa et al. 2011). It identifies anomalies that are both few in numbers and different in feature space. The isolation forest has been applied in both the mono- and multilayer case. It resulted in removal of three outliers for both cases: the C_{60} fullerene and the two polycyclic aromatic hydrocarbons (PAHs) coronene and ovalene. For the multilayer case we also removed two other molecules (dotriacontane and guanine) since these molecules, together with the fullerene and the PAHs, are outliers in the sense that they have BEs that are 3-4 times as large as the other molecules in the data set.

Lastly, we addressed the issue that data points with the same feature representation can have different labels (BEs). This can arise when several different experimental measurements of the same surface/molecule combination are present in the data set, which differ only in experimental parameters that in our model have been assumed not to influence the BE, such as temperature ramp, starting temperature and pre-exponential factor. In this case an average over the measured BEs are used. If, on the other hand, this issue arises due to differences in parameters such as the initial coverage and the specific surface (for the monolayer data set), a more thoughtful approach is fruitful. If data points at both sub-monolayer and monolayer coverage are present, we use the monolayer data point since the sub-monolayer case may be dominated by specific surface sites exhibiting a more favourable (stronger) binding than the others. If data points for several specific surfaces or facets are present within the same surface category, we use the surface that occurs most frequently within the category in order to get a more coherent data set. After this data preparation step, the final data set contains 143 (46), entries for the monolayer (multilayer) case and 114 individual molecules. After the data preparation step the features were normalised to have zero mean and unit variance, with the exception of the one-hot-encoded surface features. This is a standard procedure in the ML community to make the learning task easier. All data points used in this work, their actual surface and assigned surface category are given in Table B.2 and B.3 for monolayer and multilayer coverage species, respectively[2].

3. Results and discussion

In the following sections we validate the performance of the model using two different types of assessment; five-fold cross validation and leave-one-molecule-out cross validation.

3.1. Five-fold cross validation

Five-fold cross validation is a standard approach in the ML community for comparing models and for assessing how well they can predict new data points. As illustrated in Fig. A.1 in Appendix A, it is carried out by splitting the data set into five disjoint equally sized sets. The model is then trained on four of the sets while the fifth set is used to validate the model (i.e. by comparing model predicted BEs to actual BEs). This is repeated five

[2] Electronic versions of the data files, along with Python scripts for producing the results presented below, can be found in the Github repository here.

times until every data point has been used as validation data exactly once. In Fig. 2 we show parity plots of ML predictions on the combined validation data set from the five folds versus actual literature BEs for the monolayer and multilayer data sets. The closer the points are to the diagonal dotted line, the better the model performs. The performance is quantified using the root-mean-square error (RMSE) and more absolutely by the coefficient of determination, R^2.

It is found that the model has the highest R^2 value for the monolayer case shown in Fig. 2a. One contributor to this could be the fact that there are more than three times as many entries for the monolayer set than for the multilayer set. However, a much more important difference between the two data sets is that for the monolayer case the model is exceptionally accurate for data points with a BE above 17000 K. The reason for this behaviour is that this part of the data set is immensely uniform as it consists only of carbohydrate chains of varying lengths adsorbed on graphene surfaces. Since the BE increases proportionally with the length of the carbon chain, the model is able to learn these BEs with a very high accuracy. In fact, if only the entries with BE above 17000 K are included in the analysis, the model achieves a RMSE of 25.7 K and an R^2 value of approximately one. This RMSE is much lower than the overall RMSE of 879 K, especially when taken into consideration that the scale of the data set is higher. However, it should be kept in mind that the model is only as nuanced as the data it has been provided. This implies that the model has a very narrow knowledge of molecules in the high BE regime, and thus the predictive capability would most likely be limited for molecules different from the simple carbohydrates. If only the low energy regime is considered, the R^2 value decreases to 0.946. The corresponding parity plot, Fig. A.2, can be found in the appendix.

3.2. Leave-one-molecule-out cross validation

As a second type of assessment, we next evaluate how well the model can predict individual new molecules. This is done by first removing a chosen molecule from the data set (including all surface categories for the monolayer case), then training the model on the reduced data set, and finally comparing the ML-predicted BE of the chosen molecule with the literature value. For this assessment eight different molecules have been chosen, namely acetone ($CH_3C(O)CH_3$), acetonitrile (CH_3CN), allyl alcohol (C_3H_5OH), ammonia (NH_3), methane (CH_4), methylformate (CH_3OCHO), the alkane nonacosane ($C_{29}H_{60}$) and the nucleobase thymine ($C_5H_6N_2O_2$), which represent diverse chemical compositions and molecular sizes.

The comparison between ML-predicted and actual BEs are shown in table 2. The model uncertainty estimates are obtained from the standard deviation of the posterior distribution (i.e. the square root of its variance), as described in Sect. 2.1. As seen, the overall predictive capability of the model is reasonably good. Furthermore, there is a direct correlation between how well different types of molecules are represented in the training data and how well the model predicts them. For example, it is found that nonacosane is exactly predicted, which is presumably a consequence of the many similar molecules in the data set. It is further noticeable that ammonia is predicted quite reasonably, even though the model is mostly trained on organic molecules. We can identify that the good prediction accuracy mainly comes from the inclusion of the following molecular features; number of valence electrons, H-bond acceptors, H-bond donors and TPSA, since the predicted BE for a monolayer of ammonia on a carbon surface is only 2070 ± 1740 K (a deviation of -31 % compared to the literature value) if these features are excluded. For the multilayer case the deviation would be even more pronounced at -36 %. The model struggles the most with allyl alcohol, acetontrile and methyl formate, although the deviations are still relatively small at $\pm$ ~20 %. This might be because the training data set contains few molecules like these three or because another factor such as the specific surface has an influence on the BE, which the model cannot account for.

We also recall here that some types of molecules are not well represented in the training data, meaning that the developed model is not suitable for predicting these. This concerns foremost the types of molecules that were removed as outliers in the data preparation step (fullerenes, PAHs, and large aromatic molecules in general). Also phenols are notoriously difficult to perform TPD experiments with, and therefore not much data is available for the model to train on.

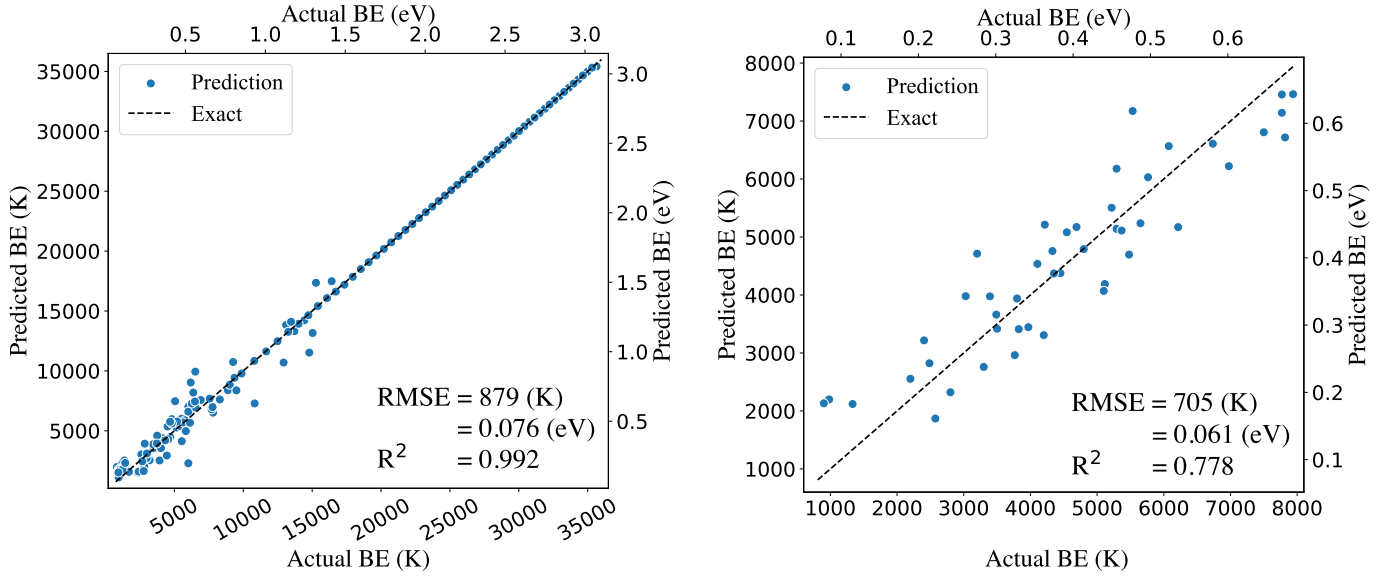
In general, a further limitation of the developed model is that it cannot distinguish between isomers of the same molecule, in the case these have the same feature representation, which will always be the case for structural isomers. However, given the other uncertainties - both in the experimental TPD procedure and in the developed model - it is very likely that the difference in BE between two isomers would anyway be much smaller than what can be resolved with the current approach.

Another limitation of the model is that it struggles with molecules consisting of atoms other than C, H, O, N and Cl. This is because the data set includes very few molecules that contain other than said atoms and because we consequently do not include features representing any other atoms. This means that predicting the BEs of for example sulphurous- and phosphorous-containing molecules will be less accurate. To increase the performance and predictive capabilities of the model, the most essential future step would be to increase and diversify the number of entries in the training data set.

4. Astrophysical implications

In recent years, the number of newly detected molecules in the interstellar medium has skyrocketed, including many complex organic molecules and prebiotic species such as carbamide ($NH_2C(O)NH_2$, Belloche et al. 2019), propargylimine (HC_3HNH, Bizzocchi et al. 2020), propargyl cyanide (HCCCH2CN, McGuire et al. 2020), ethanolamine ($HOCH_2CH_2NH_2$, Rivilla et al. 2021), and allenyl acetylene ($H_2CCCHCCH$, Cernicharo et al. 2021). While these detections underline the molecular complexity that is present in star-forming regions, a limited knowledge of fundamental physico-chemical parameters such as reaction rate constants, photo destruction cross sections, and BEs prevents us from fully understanding how these species form, react, respond to physical conditions, and, ultimately, what their place is in the interstellar chemical factory.

In this section we first employ the ML model to predict the BEs of a number of molecules that have been detected in the interstellar medium, but for which no or limited information about their BEs can be found in the literature, see Table 3. The features of the molecules are encoded in the same way as the training data for the ML model and presented in Table C.1. Predictions of BEs for monolayer coverage are limited to two surfaces, carbon and water, for which the model shows the highest performance. For a number of species, BE estimates based on the linear addition method are available in the literature and have been used in modelling studies. These BEs are generic, meaning that they are not



(a) Monolayer

(b) Multilayer

Fig. 2: Parity plots for a) monolayer and b) multilayer coverage comparing ML-predicted BEs against actual BEs for the combined validation set from five-fold cross validation.

Table 2: Comparison between ML-predicted and literature BE values (rounded to the nearest ten) for leave-one-molecule-out cross validation.

Name	Molecule	Surface or Coverage	Prediction (K)	Literature (K)	Deviation
Acetone	$CH_3C(O)CH_3$	Water	4520 ± 450	4690 ± 240	−3.6 %
Acetonitrile	CH_3CN	Metal	6730 ± 880	5530 ± 360	21 %
Allyl alcohol	C_3H_5OH	Metal	7110 ± 1740	6010[†]	18 %
Ammonia	NH_3	Carbon	2870 ± 1700	2990 ± 240	−4.0 %
Methane	CH_4	Carbon	1870 ± 470	1700 ± 120	10 %
Methyl formate	CH_3OCHO	Water	5390 ± 540	4510 ± 530	19 %
Nonacosane	$C_{29}H_{60}$	Carbon	23720 ± 220	23720 ± 930	0.0 %
Thymine	$C_5H_6N_2O_2$	Metal	11680 ± 2600	12930 ± 240	−9.6 %
Acetonitrile	CH_3CN	Multilayer	4030 ± 470	4800 ± 190	−16 %
Ammonia	NH_3	Multilayer	3330 ± 960	3030 ± 240	9.9 %
Methyl formate	CH_3OCHO	Multilayer	4520 ± 470	4110 ± 200	10 %

Notes. [†]Literature study does not present uncertainty of allyl alcohol BE measurement. A version of this table with the BEs in eV is shown in Table 4

specific to any surface. These literature estimates are included in Table 3 for comparison to our ML predictions.

Several observations are made for the predicted BEs. The uncertainties on the predictions vary from a few % to up to 60% for hydroxyacetone. This is a reflection of the training data and feature representation of the model, with lower uncertainties emerging when features of a certain molecule are better represented in the training data set. In particular for molecules with cyanide groups, the uncertainties on the predicted BEs are low, because a comparatively large number of cyanide molecules are included in the training data. For about half of the molecules their predicted BE is substantially larger on a water surface than on a carbonaceous surface. All species that show this behaviour contain a cyanide group and therefore this trend is explained by the stronger polar interaction between –CN and the H_2O ice surface. For a couple of molecules, all of them isomers, the model cannot differentiate the BEs, because they have the same feature representation.

A comparison between ML predictions and literature estimates shows some discrepancies. The predictions and estimates of methylcyanodiacetylene and the HC_xN species on carbonaceous surfaces, as well as N-methylformamide, agree fairly well, especially within the uncertainties of the ML prediction. However, the linear addition method seems to underestimate the BE of cyanamide, albeit just within the uncertainty of the ML prediction. For most species, the linear addition method severely overestimates the BE of the molecule. In the case of propargylimine the estimated BE is more than three times larger than the ML BE prediction.

Finally, for a number of species we note that the training data does not contain many molecules that represent their features, such as the imine (-N=CH-) group for ethanimine

and propargylimine. Predictions for these species will have a higher degree of uncertainty. Improvements on the predictions can be expected by increasing the size of the training data set, in particular with species containing relevant features, and by expanding the feature representation.

We next discuss the astrophysical implications of the ML predictions. For this, we construct a simple model for the behaviour of the predicted species during thermal heating in the interstellar medium. At the basis of this model is Eq. 2, which is used to determine the loss of material from a surface at a specified temperature. The BEs presented in Table 3 are used and the pre-exponential factor is assumed to be $1{\times}10^{18}$ s^{-1} for all species. With transition state theory calculations, Minissale et al. (2022) showed for a selection of interstellar molecules that their prefactors are substantially larger than the canonically assumed $1{\times}10^{12}$ or $1{\times}10^{13}$ s^{-1} and that these values increase for molecules of larger size. This motivates the choice of $\nu = 1{\times}10^{18}$ s^{-1}, as the listed molecules are relatively large in size, but we emphasize that it is a rough assumption and values will differ from molecule to molecule. We note that when a prefactor of $1{\times}10^{12}$ s^{-1} is used, the peak desorption temperatures of molecules increase by about 20 – 30 K, with respect to the $1{\times}10^{18}$ s^{-1} value. For each molecule, a monolayer column density of $N = 1{\times}10^{15}$ molecules cm^{-2} is used.

Two model results are presented. Figure 3 shows the desorption profiles of the molecules versus temperature for a linear heating rate of 1 K century^{-1}. Figure 4 shows the same desorption profiles, but plotted against the distance from a protostar based on the following equation:

$$T(r) = 200 \times r^{-0.62}, \tag{7}$$

with r being the radius from the protostar in au. Equation 7 is derived by Andrews & Williams (2007) by averaging the observed disk temperature profiles of a sample of protoplanetary disks in the Taurus-Auriga and Ophiuchus-Scorpius star forming regions. From Fig. 3 and 4 an indication can be given of the location of the snowlines of the molecules presented in Table 3, that is, the radius from a central protostar where volatile molecules sublimate or freeze out (Öberg & Bergin 2021). In both figures the peak desorption temperature (97 K) or location (3.2 au) of water are indicated, which are determined for a monolayer ($1{\times}10^{15}$ molecules cm^{-2}) H_2O coverage on HOPG with $E_{\rm bin} = 5792$ K and $\nu = 4.96{\times}10^{15}$ s^{-1} (Minissale et al. 2022).

The plots display a large variation in peak desorption temperatures of the predicted molecules. For some species, such as CH_3CHNH, CH_3NCO, and $H_2CCCHCCH$, peak desorption coincides with or is lower than that of water. While in this work the binding of molecules to a water surface is considered, it is reasonable to assume that many of these species will in fact be mixed in water ice due to the large H_2O abundance in the ISM (Boogert et al. 2015). Consequently, it is to be expected that these molecule will mostly co-desorb (Burke & Brown 2010) with water when this species sublimates. Snowlines of molecules with $E_{\rm bin,x} \leq E_{\rm bin,water}$ therefore more likely coincide with that of water. Co-desorption with other bulk ice components, such as CO is in principle possible, but laboratory evidence generally indicates that molecules with a significantly higher BE than the bulk medium do not co-desorb (e.g. Ligterink et al. 2018). The species considered in this work, which all have $E_{\rm bin,x} \geq E_{\rm bin,CO}$, are therefore unlikely to co-desorb with CO.

Many species, like CH_2CCHCN, HCCCHCHCN, and CH_3C_5N have a high BE to water surfaces and show desorption traces at much higher temperatures than the peak desorption temperature of water itself. Since at these temperatures water ice has sublimated from grain surfaces and is no longer a binding medium, desorption should instead occur from a surface that is probably made out of silicates, carbonaceous species, or organic residue. Taking this into account, it seems that many of these molecules will in fact desorb quite close to the water snow line, based on their carbon surface BEs. Only a handfull of molecules desorb at temperatures considerably above that of the water snow line, namely NH_2CN, $NH_2C(O)NH_2$, the HC_xN species, and indene. From this visualization it becomes clear that various subgroups of these molecules will end up in distinctly different regions of the planet-forming disk as either gas or ice.

5. Conclusions

In this work, an ML model based on Gaussian Process Regression is created and trained to predict BEs of molecules, specifically those of astrochemical relevance. The BEs determined from laboratory experiments are collected, categorized by their features (e.g., mono- or multilayer coverage, binding surface, functional groups, valence electrons, H-bond acceptors and donors), and used as training data for the model. The performance of the model is assessed with five-fold and leave-one-molecule-out cross validation. A root mean square error of 892 K and 580 K are found, for the mono- and multilayer model, respectively. For individual molecules the deviation between model predicted and literature BEs is found to be within ±20%. We note that sufficient training data and accurate feature representation are essential to predict BEs. Molecules for which features are not well described or insufficient training data points are available will generally have larger uncertainties on their predictions.

The validated model is used to predict the BEs on a water and carbonaceous surface of twenty one molecules that have been detected in recent years in the interstellar medium, but for which no or limited experimental information about their BEs is available. The lowest BE of 2990 K is predicted for methyl isocyanate (CH_3NCO) on a water surface, while the highest BE of 12820 K is predicted for cyanopentacetylene ($HC_{11}N$) bound to a water surface. Uncertainties on the predictions range from just a few percent to about 60% for hydroxyacetone ($CH_3C(O)CH_2OH$), which is presumably a reflection of the lack of training data and feature representation for these molecules. The surface can have a pronounced effect on the predicted BE, showing differences of several 1000's K for some molecules. Finally, a comparison between the ML model predictions and the in the field of astrochemistry widely used linear addition method to predict BEs is presented. We find that the linear addition methods generally overpredicts BEs, in some case by more than a factor of two.

The newly predicted BEs are put into context of interstellar environments with a simple model that shows their desorption profile with respect to a 1 K century^{-1} temperature ramp and a protoplanetary disk temperature profile. From this simple model, the locations of the snowlines of these molecules are determined. Most of them will roughly coincide with the water snowline, but those of cyanamide (NH_2CN), urea/carbamide ($NH_2C(O)NH_2$), and the cyanoacetylenes (HC_xN) are located at much higher temperatures or closer to the protostar.

This work demonstrates that ML can be employed to accurately and rapidly predict BEs of molecules. The approach taken here is based on experimental training data, but we note that ML models can also be trained on BEs obtained from quantum chemical calculations, as already pursued intensively in the hetero-

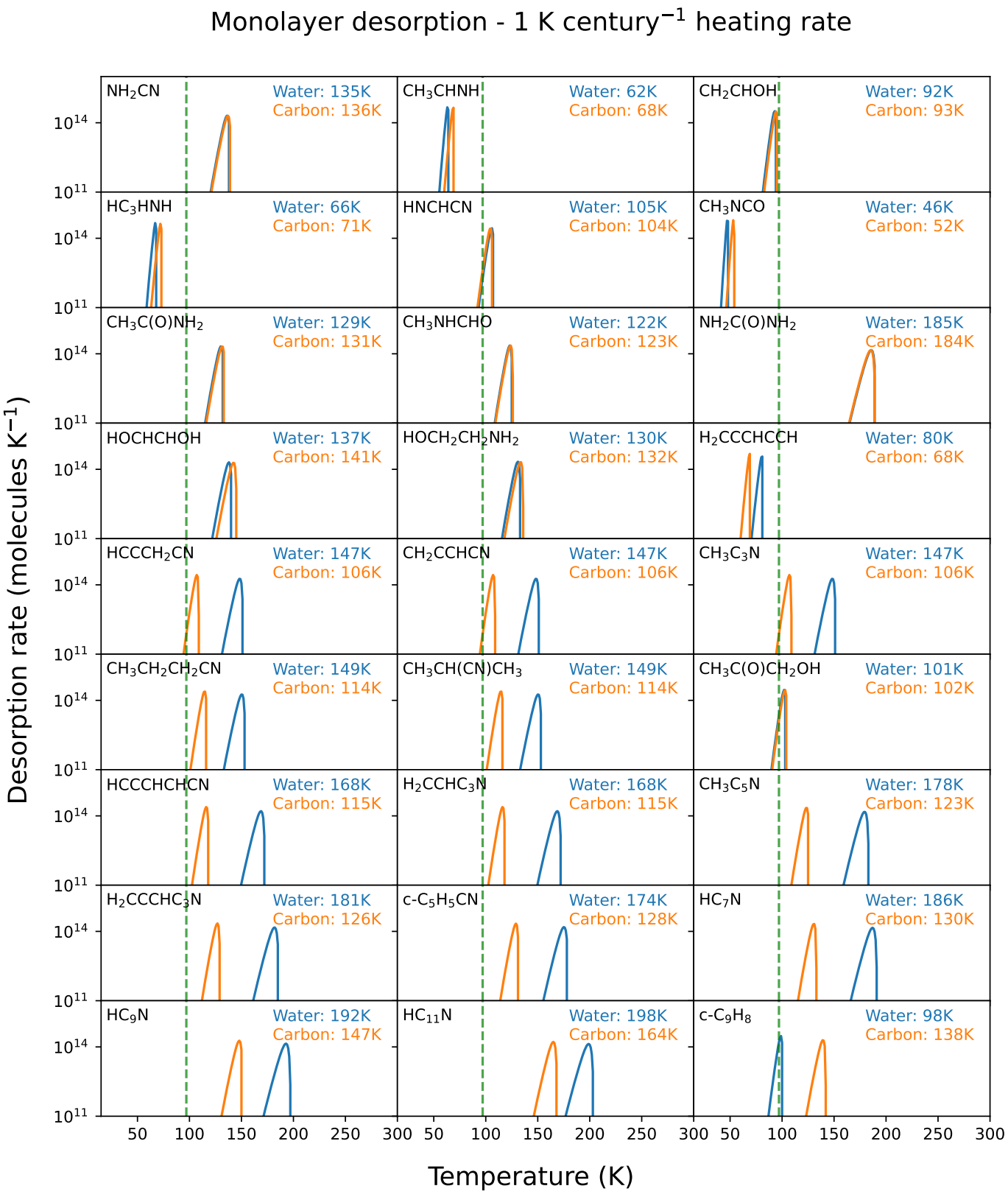


Fig. 3: Desorption traces of molecules for which the BE is determined in this work. The first order desorption profiles are plotted for monolayer (1×10^{15} molecules cm^{-2}) coverage on a water ice (blue) and carbonaceous (orange) surface. The peak desorption temperatures are indicated in the top right corners. A linear heating rate of 1 K century^{-1} is applied. Prefactors are assumed and set at A = 1×10^{18} s^{-1} for all molecules. The peak desorption for water is indicated with a green dashed lines at 97 K.

geneous catalysis community (e.g., Gu et al. 2020; Fung et al. 2021; Andersen & Reuter 2021). In that connection, a natural extension of this work could be to also take into account BE distributions on amorphous and highly anisotropic surfaces, as this distribution is often readily available from quantum chemical calculations (e.g., Tinacci et al. 2022; Ferrero et al. 2020; Duflot et al. 2021). Overall, we believe that the work presented here could pave the way for a stronger collaboration between the communities working on quantum chemical calculations of BEs, laboratory experiments and ML, as the various approaches complement each other. The here predicted BEs will find general use in the modelling of astrochemical and planet-forming environments, while more detailed BE distributions would be critical to more specific modelling such as the reactivity of molecules at dust grains at low temperatures.

Acknowledgements. The authors thank C.N. Shingledecker for providing BE estimates of molecules used in several chemical modelling studies. N.F.W.L. acknowledges funding by the Swiss National Science Foundation (SNSF) under Ambizione grant 193453. M.A. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754513, the Aarhus University Research Foundation, the Danish National Research Foundation through the Center of Excellence 'InterCat' (Grant agreement no.: DNRF150) and VILLUM FONDEN (grant no. 37381).

References

Abdulgalil, A. G. M., Marchione, D., Thrower, J., et al. 2013, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371, 20110586
Acharyya, K., Fuchs, G., Fraser, H., Van Dishoeck, E., & Linnartz, H. 2007, Astronomy & Astrophysics, 466, 1005
Aigrain, S., Pont, F., & Zucker, S. 2012, Monthly notices of the royal astronomical society, 419, 3147
Allouche, A., Verlaque, P., & Pourcin, J. 1998, The Journal of Physical Chemistry B, 102, 89
Andersen, M. & Reuter, K. 2021, Acc. Chem. Res., 54, 2741
Andrews, S. M. & Williams, J. P. 2007, The Astrophysical Journal, 659, 705
Bahr, S. & Kempter, V. 2007, The Journal of chemical physics, 127, 074707
Bahr, S., Toubin, C., & Kempter, V. 2008, The Journal of chemical physics, 128, 134712
Balbisi, M., Horvath, R. A., Szőri, M., & Jedlovszky, P. 2022, The Journal of Chemical Physics, 156, 184703
Behmard, A., Fayolle, E. C., Graninger, D. M., et al. 2019, The Astrophysical Journal, 875, 73
Bellman, R. 1966, Science, 153, 34
Belloche, A., Garrod, R., Müller, H., et al. 2019, Astronomy & Astrophysics, 628, A10
Bertin, M., Doronin, M., Fillion, J.-H., et al. 2017, Astronomy & Astrophysics, 598, A18
Bisschop, S., Fraser, H., Öberg, K., Van Dishoeck, E., & Schlemmer, S. 2006, Astronomy & Astrophysics, 449, 1297
Bizzocchi, L., Prudenzano, D., Rivilla, V. M., et al. 2020, Astronomy & Astrophysics, 640, A98
Bolina, A., Wolff, A., & Brown, W. 2005a, The Journal of chemical physics, 122, 044713
Bolina, A. S., Wolff, A. J., & Brown, W. A. 2005b, The Journal of Physical Chemistry B, 109, 16836
Boogert, A. A., Gerakines, P. A., & Whittet, D. C. 2015, Annual Review of Astronomy and Astrophysics, 53, 541
Borget, F., Chiavassa, T., Allouche, A., Marinelli, F., & Aycard, J.-P. 2001, Journal of the American Chemical Society, 123, 10668
Brown, W. A. & Bolina, A. S. 2007, Monthly Notices of the Royal Astronomical Society, 374, 1006
Burke, D. J. & Brown, W. A. 2010, Physical Chemistry Chemical Physics, 12, 5947
Burke, D. J., Puletti, F., Brown, W. A., et al. 2015a, Monthly Notices of the Royal Astronomical Society, 447, 1444
Burke, D. J., Puletti, F., Woods, P. M., et al. 2015b, The Journal of chemical physics, 143, 164704
Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. 1995, SIAM J. Sci. Comput., 16, 1190
Cernicharo, J., Cabezas, C., Agúndez, M., et al. 2021, Astronomy & Astrophysics, 647, L3
Chaabouni, H., Diana, S., Nguyen, T., & Dulieu, F. 2018, Astronomy & Astrophysics, 612, A47
Collings, M., Dever, J. W., Fraser, H. J., & McCoustra, M. R. 2003, Astrophysics and Space Science, 285, 633
Collings, M. P., Frankland, V., Lasne, J., et al. 2015, Monthly Notices of the Royal Astronomical Society, 449, 1826
Congiu, E., Chaabouni, H., Laffon, C., et al. 2012, The Journal of chemical physics, 137, 054713
Corazzi, M. A., Brucato, J. R., Poggiali, G., et al. 2021, The Astrophysical Journal, 913, 128
Couturier-Tamburelli, I., Toumi, A., Piétri, N., & Chiavassa, T. 2018, Icarus, 300, 477
Cuppen, H., Walsh, C., Lamberts, T., et al. 2017, Space Science Reviews, 212, 1
Danger, G., Duvernay, F., Theulé, P., Borget, F., & Chiavassa, T. 2012, The Astrophysical Journal, 756, 11
Das, A., Sil, M., Gorai, P., Chakrabarti, S. K., & Loison, J.-C. 2018, The Astrophysical Journal Supplement Series, 237, 9
De Jong, A. & Niemantsverdriet, J. 1990, Surface Science, 233, 355
Demers, L. M., Östblom, M., Zhang, H., et al. 2002, Journal of the American Chemical Society, 124, 11248
Dostert, K.-H., O'Brien, C. P., Mirabella, F., Ivars-Barceló, F., & Schauermann, S. 2016, Physical Chemistry Chemical Physics, 18, 13960
Duflot, D., Toubin, C., & Monnerville, M. 2021, Frontiers in Astronomy and Space Sciences, 8
Duvenaud, D. 2014, PhD thesis, Ph. D. dissertation, University of Cambridge, Cambridge, England, United Kingdom
Edridge, J. L., Freimann, K., Burke, D. J., & Brown, W. A. 2013, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371, 20110578
Fayolle, E. C., Balfe, J., Loomis, R., et al. 2016, The Astrophysical Journal Letters, 816, L28
Ferrero, S., Zamirri, L., Ceccarelli, C., et al. 2020, The Astrophysical Journal, 904, 11
Fraser, H. J., Collings, M. P., McCoustra, M. R., & Williams, D. A. 2001, Monthly Notices of the Royal Astronomical Society, 327, 1165
Fuchs, G. W., Acharyya, K., Bisschop, S. E., et al. 2006, Faraday Discussions, 133, 331
Fung, V., Hu, G., Ganesh, P., & Sumpter, B. G. 2021, Nat. Comm., 12, 88
Galvez, O., Ortega, I. K., Maté, B., et al. 2007, Astronomy & Astrophysics, 472, 691
Garrod, R. & Herbst, E. 2006, Astronomy & Astrophysics, 457, 927
Gellman, A. J. & Paserba, K. R. 2002, The Journal of Physical Chemistry B, 106, 13231
Gibson, N., Aigrain, S., Roberts, S., et al. 2012, Monthly notices of the royal astronomical society, 419, 2683
Gu, G. H., Noh, J., Kim, S., et al. 2020, J. Phys. Chem. Lett., 11, 3185
Guennoun, Z., Couturier-Tamburelli, I., Piétri, N., & Aycard, J.-P. 2005, The Journal of Physical Chemistry B, 109, 3437
Haynes, D., Tro, N., & George, S. 1992, The Journal of Physical Chemistry, 96, 8502
He, J., Acharyya, K., & Vidali, G. 2016, The Astrophysical Journal, 825, 89
He, J., Emtiaz, S. M., & Vidali, G. 2017, The Astrophysical Journal, 837, 65
Heyl, J., Holdship, J., & Viti, S. 2022, The Astrophysical Journal, 931, 26
Jordan, M. I. & Mitchell, T. M. 2015, Science, 349, 255
Jørgensen, J. K., Belloche, A., & Garrod, R. T. 2020, Annual Review of Astronomy and Astrophysics, 58, 727
Kruczkiewicz, F., Vitorino, J., Congiu, E., Theulé, P., & Dulieu, F. 2021, arXiv preprint arXiv:2104.10464
Landrum, G. 2020, RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org/
Lasne, J., Laffon, C., & Parent, P. 2012, Physical Chemistry Chemical Physics, 14, 697
Lattelais, M., Bertin, M., Mokrane, H., et al. 2011, Astronomy & Astrophysics, 532, A12
Lee, K. L. K., Patterson, J., Burkhardt, A., et al. 2021, The Astrophysical Journal Letters
Ligterink, N., Walsh, C., Bhuin, R., et al. 2018, Astronomy & Astrophysics, 612, A88
Luo, M.-F., Zhong, Y.-J., Yuan, X.-X., & Zheng, X.-M. 1997, Applied Catalysis A: General, 162, 121
Maté, B., Jiménez-Redondo, M., Peláez, R. J., Tanarro, I., & Herrero, V. J. 2019, Monthly Notices of the Royal Astronomical Society, 490, 2936
Mazo-Sevillano, P. d., Aguado, A., & Roncero, O. 2021, The Journal of Chemical Physics, 154, 094305
McGuire, B. A., Burkhardt, A. M., Loomis, R. A., et al. 2020, The Astrophysical Journal Letters, 900, L10

Minissale, M., Aikawa, Y., Bergin, E., et al. 2022, arXiv preprint arXiv:2201.07512
Molpeceres, G., Zaverkin, V., Watanabe, N., & Kästner, J. 2021, A&A, 648, A84
Muñoz Caro, G., Jiménez-Escobar, A., Martín-Gago, J. Á., et al. 2010, A&A, 522, A108
Noble, J., Congiu, E., Dulieu, F., & Fraser, H. 2012, Monthly Notices of the Royal Astronomical Society, 421, 768
Noble, J., Diana, S., & Dulieu, F. 2015, Monthly Notices of the Royal Astronomical Society, 454, 2636
Noble, J. A., Theule, P., Borget, F., et al. 2013, Monthly Notices of the Royal Astronomical Society, 428, 3262
Öberg, K., Van Broekhuizen, F., Fraser, H., et al. 2005, The Astrophysical Journal Letters, 621, L33
Öberg, K. I. & Bergin, E. A. 2021, Physics Reports, 893, 1
Öberg, K. I., Garrod, R. T., Van Dishoeck, E. F., & Linnartz, H. 2009, Astronomy & Astrophysics, 504, 891
Östblom, M., Liedberg, B., Demers, L. M., & Mirkin, C. A. 2005, The Journal of Physical Chemistry B, 109, 15150
Parmeter, J., Schwalke, U., & Weinberg, W. 1988, Journal of the American Chemical Society, 110, 53
Paserba, K. R. & Gellman, A. J. 2001a, The Journal of Chemical Physics, 115, 6737
Paserba, K. R. & Gellman, A. J. 2001b, Physical review letters, 86, 4338
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, the Journal of machine Learning research, 12, 2825
Quan, D., Herbst, E., Corby, J. F., Durr, A., & Hassel, G. 2016, The Astrophysical Journal, 824, 129
Rasmussen, C. E. & Williams, C. K. 2006, MA: MIT Press [Google Scholar]
Rimola, A., Skouteris, D., Balucani, N., et al. 2018, ACS Earth and Space Chemistry, 2, 720
Rivilla, V. M., Jiménez-Serra, I., Martín-Pintado, J., et al. 2021, Proceedings of the National Academy of Sciences, 118
Salter, T. L., Stubbing, J. W., Brigham, L., & Brown, W. A. 2018, The Journal of chemical physics, 149, 164705
Salter, T. L., Wootton, L., & Brown, W. A. 2019, ACS Earth and Space Chemistry, 3, 1524
Sandford, S. A. & Allamandola, L. J. 1988, Icarus, 76, 201
Sandford, S. A. & Allamandola, L. J. 1990, Icarus, 87, 188
Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P., & Green, W. H. 2020, Journal of Chemical Information and Modeling, 60, 2697, pMID: 32243154
Schriver, A., Coanga, J., Schriver-Mazzuoli, L., & Ehrenfreund, P. 2004, Chemical physics, 303, 13
Shallue, C. J. & Vanderburg, A. 2018, Astronomcal Journal, 155
Shimonishi, T., Nakatani, N., Furuya, K., & Hama, T. 2018, The Astrophysical Journal, 855, 27
Shingledecker, C. N., Molpeceres, G., Rivilla, V. M., Majumdar, L., & Kästner, J. 2020, The Astrophysical Journal, 897, 158
Smith, R. S. & Kay, B. D. 2018, The Journal of Physical Chemistry B, 122, 587
Smith, R. S. & Kay, B. D. 2019, The Journal of Physical Chemistry A, 123, 3248
Smith, R. S., Matthiesen, J., & Kay, B. D. 2014, The Journal of Physical Chemistry A, 118, 8242
Smith, R. S., May, R. A., & Kay, B. D. 2016, The Journal of Physical Chemistry B, 120, 1979
Solomun, T., Christmann, K., & Baumgärtel, H. 1989, The Journal of Physical Chemistry, 93, 7199
Suhasaria, T., Thrower, J., & Zacharias, H. 2015, Monthly Notices of the Royal Astronomical Society, 454, 3317
Suhasaria, T., Thrower, J., & Zacharias, H. 2017, Monthly Notices of the Royal Astronomical Society, 472, 389
Tait, S. L., Dohnálek, Z., Campbell, C. T., & Kay, B. D. 2005, The Journal of chemical physics, 122, 164707
Takeuchi, K., Yamamoto, S., Hamamoto, Y., et al. 2017, The Journal of Physical Chemistry C, 121, 2807
Theulé, P., Borget, F., Mispelaer, F., et al. 2011, Astronomy & Astrophysics, 534, A64
Tinacci, L., Germain, A., Pantaleone, S., et al. 2022, ACS Earth and Space Chemistry, 6
Toumi, A., Piétri, N., Chiavassa, T., & Couturier-Tamburelli, I. 2016, Icarus, 270, 435
Tylinski, M., Smith, R. S., & Kay, B. D. 2020, The Journal of Physical Chemistry A, 124, 6237
Ulbricht, H., Zacharia, R., Cindir, N., & Hertel, T. 2006, Carbon, 44, 2931
Viti, S., Collings, M. P., Dever, J. W., McCoustra, M. R., & Williams, D. A. 2004, Monthly Notices of the Royal Astronomical Society, 354, 1141
Wakelam, V., Herbst, E., Loison, J.-C., et al. 2012, The Astrophysical Journal Supplement Series, 199, 21
Zaverkin, V., Molpeceres, G., & Kästner, J. 2021, Monthly Notices of the Royal Astronomical Society, 510, 3063
Zhou, J.-H., Sui, Z.-J., Zhu, J., et al. 2007, Carbon, 45, 785
Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. 1997, ACM Trans. Math. Softw., 23, 550
Zubkov, T., Smith, R. S., Engstrom, T. R., & Kay, B. D. 2007, The Journal of chemical physics, 127, 184707

Table 3: Predictions of BEs for molecules with astrophysical relevance measured in K and rounded to nearest 10.

Name	Molecule	Surface	Prediction (K)	Prediction (eV)	Estimate (K)
Cyanamide	NH_2CN	Carbon	8780 ± 3730	0.76 ± 0.32	5556[a]
Cyanamide	NH_2CN	Water	8730 ± 3710	0.75 ± 0.32	
Ethanimine	CH_3CHNH	Carbon	4350 ± 1100	0.37 ± 0.09	5580[b]
Ethanimine	CH_3CHNH	Water	4010 ± 1620	0.35 ± 0.14	
Vinylalcohol	CH_2CHOH	Carbon	6000 ± 260	0.52 ± 0.02	–
Vinylalcohol	CH_2CHOH	Water	5910 ± 400	0.51 ± 0.03	–
Propargylimine	HC_3HNH	Carbon	4580 ± 1210	0.40 ± 0.10	14750[c]
Propargylimine	HC_3HNH	Water	4260 ± 1630	0.37 ± 0.14	
Cyanomethanimine	HNCHCN	Carbon	6670 ± 2110	0.57 ± 0.18	10900[c]
Cyanomethanimine	HNCHCN	Water	6750 ± 2100	0.58 ± 0.18	
Methyl isocyanate	CH_3NCO	Carbon	3360 ± 1440	0.29 ± 0.12	6486[d]
Methyl isocyanate	CH_3NCO	Water	2990 ± 1490	0.26 ± 0.13	
Acetamide	$CH_3C(O)NH_2$	Carbon	8420 ± 520	0.73 ± 0.04	–
Acetamide	$CH_3C(O)NH_2$	Water	8350 ± 530	0.72 ± 0.04	–
N-Methylformamide	CH_3NHCHO	Carbon	7920 ± 1290	0.68 ± 0.11	7386[d]
N-Methylformamide	CH_3NHCHO	Water	7880 ± 1340	0.68 ± 0.11	
Carbamide / Urea	$NH_2C(O)NH_2$	Carbon	11930 ± 4350	1.02 ± 0.37	–
Carbamide / Urea	$NH_2C(O)NH_2$	Water	11960 ± 4350	1.03 ± 0.37	–
Ethenediol	HOCHCHOH	Carbon	9130 ± 3230	0.79 ± 0.28	–
Ethenediol	HOCHCHOH	Water	8840 ± 3240	0.76 ± 0.28	–
Ethanolamine	$HOCH_2CH_2NH_2$	Carbon	8550 ± 2250	0.74 ± 0.19	–
Ethanolamine	$HOCH_2CH_2NH_2$	Water	8380 ± 2260	0.72 ± 0.19	–
Allenyl acetylene	$H_2CCCHCCH$	Carbon	4360 ± 370	0.37 ± 0.03	–
Allenyl acetylene	$H_2CCCHCCH$	Water	5120 ± 320	0.44 ± 0.03	–
Propargyl cyanide†	$HCCCH_2CN$	Carbon	6840 ± 650	0.59 ± 0.05	18750[c]
Propargyl cyanide†	$HCCCH_2CN$	Water	9520 ± 310	0.82 ± 0.03	
Cyanoallene†	CH_2CCHCN	Carbon	6840 ± 650	0.59 ± 0.06	–
Cyanoallene†	CH_2CCHCN	Water	9520 ± 310	0.82 ± 0.03	–
Cyanopropyne†	CH_3C_3N	Carbon	6840 ± 650	0.59 ± 0.06	–
Cyanopropyne†	CH_3C_3N	Water	9520 ± 310	0.82 ± 0.03	–
n-Propylcyanide‡	$CH_3CH_2CH_2CN$	Carbon	7320 ± 850	0.63 ± 0.07	21350[c]
n-Propylcyanide‡	$CH_3CH_2CH_2CN$	Water	9650 ± 680	0.83 ± 0.06	
i-Propylcyanide‡	$CH_3CH(CN)CH_3$	Carbon	7320 ± 850	0.63 ± 0.08	–
i-Propylcyanide‡	$CH_3CH(CN)CH_3$	Water	9650 ± 680	0.83 ± 0.07	–
Hydroxyacetone	$CH_3C(O)CH_2OH$	Carbon	6560 ± 3800	0.57 ± 0.33	–
Hydroxyacetone	$CH_3C(O)CH_2OH$	Water	6510 ± 3660	0.56 ± 0.31	–
Cyanovinylacetylene*	HCCCHCHCN	Carbon	7430 ± 860	0.63 ± 0.07	22600[c]
Cyanovinylacetylene*	HCCCHCHCN	Water	10850 ± 340	0.93 ± 0.03	
Vinylcyanoacetylene*	H_2CCHC_3N	Carbon	7430 ± 860	0.63 ± 0.07	–
Vinylcyanoacetylene*	H_2CCHC_3N	Water	10850 ± 340	0.93 ± 0.03	–
Methylcyanodiacetylene	CH_3C_5N	Carbon	7900 ± 1000	0.68 ± 0.09	7880[c]
Methylcyanodiacetylene	CH_3C_5N	Water	11540 ± 480	0.99 ± 0.04	
Cyanoacetyleneallene	H_2CCCHC_3N	Carbon	8120 ± 1050	0.70 ± 0.09	26750[c]
Cyanoacetyleneallene	H_2CCCHC_3N	Water	11700 ± 570	1.01 ± 0.05	
1-cyano-1,3-cyclopentadiene	c-C_5H_5CN	Carbon	8260 ± 1100	0.71 ± 0.09	–
1-cyano-1,3-cyclopentadiene	c-C_5H_5CN	Water	11270 ± 790	0.97 ± 0.07	–
Cyanotriacetylene	HC_7N	Carbon	8360 ± 1130	0.72 ± 0.10	7780[c]
Cyanotriacetylene	HC_7N	Water	12040 ± 630	1.04 ± 0.05	
Cyanotetraacetylene	HC_9N	Carbon	9490 ± 1320	0.82 ± 0.11	9380[c]
Cyanotetraacetylene	HC_9N	Water	12420 ± 1020	1.07 ± 0.09	
Cyanopentaacetylene	$HC_{11}N$	Carbon	10600 ± 1460	0.91 ± 0.12	10980[c]
Cyanopentaacetylene	$HC_{11}N$	Water	12820 ± 1260	1.10 ± 0.11	
Indene	c-C_9H_8	Carbon	8930 ± 230	0.77 ± 0.02	–
Indene	c-C_9H_8	Water	6300 ± 890	0.54 ± 0.08	–

Notes. †,‡,*Isomers with identical feature descriptions have the same BEs. Literature BE estimates are taken from [a]KIDA (Wakelam et al. 2012, http://kida.astrophy.u-bordeaux.fr) ; [b]Quan et al. (2016); [c](Shingledecker et al. 2020), the GOTHAM collaboration, and C. Shingledecker (private communication); [d]Belloche et al. (2019).

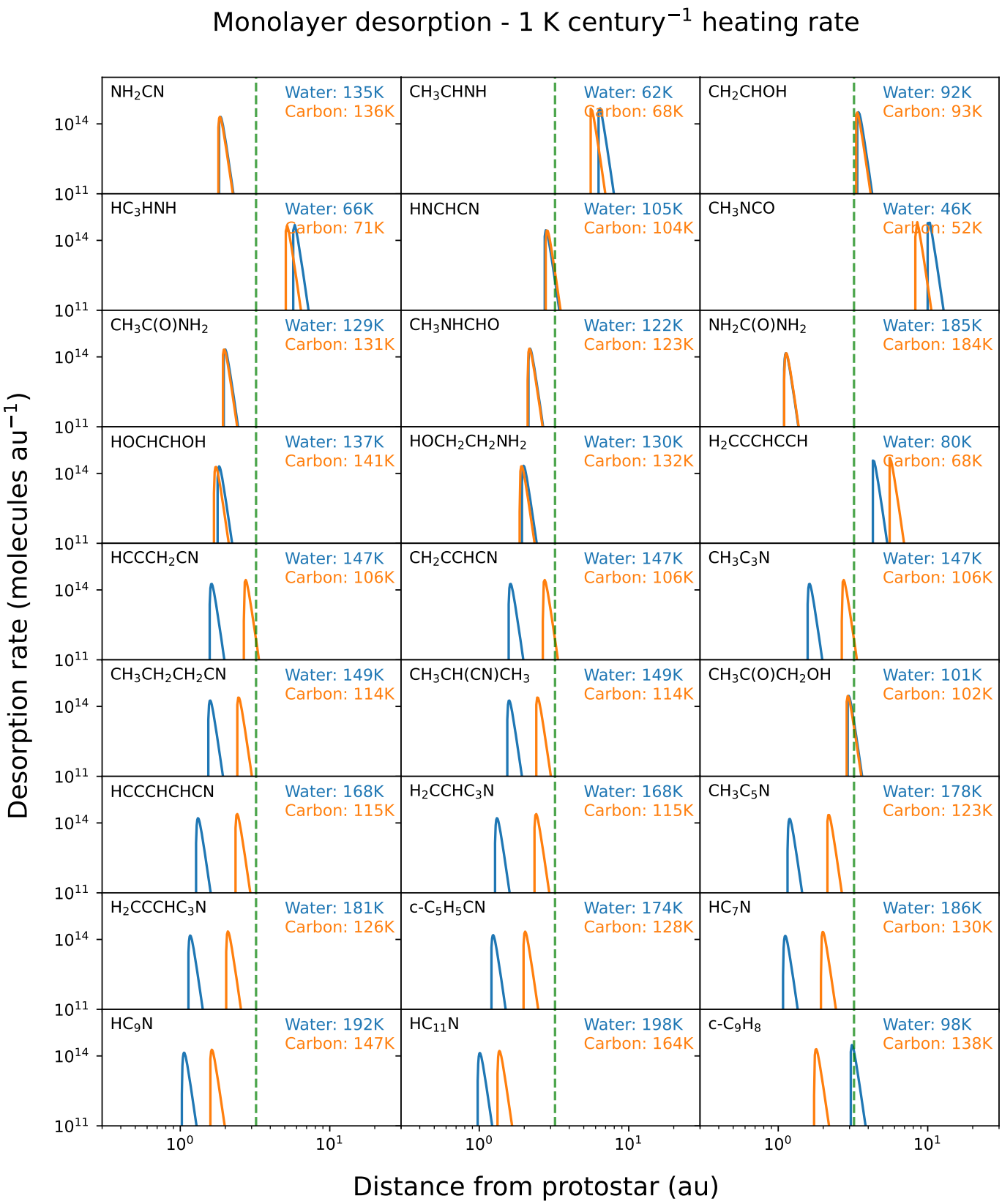


Fig. 4: Same as Fig. 3, except that the desorption trace is plotted against a median disk temperature profile as derived by Andrews & Williams (2007), see main text for more details. Shorter distances are closer to the protostar and thus correspond to higher temperatures. The peak desorption for water is indicated with a green dashed lines at 3.2 au.

Table 4: Comparison between predicted and observed values of BE obtained from the literature measured in eV and rounded to nearest 0.01.

Name	Molecule	Surface or Coverage	Prediction (eV)	Observation (eV)	Deviation
Acetone	$CH_3C(O)CH_3$	Water	0.39 ± 0.04	0.40 ± 0.02	−3.6 %
Acetonitrile	CH_3CN	Metal	0.58 ± 0.08	0.48 ± 0.03	21 %
Allyl alcohol	C_3H_5OH	Metal	0.61 ± 0.15	0.52[†]	18 %
Ammonia	NH_3	Carbon	0.25 ± 0.15	0.26 ± 0.02	−4.0 %
Methane	CH_4	Carbon	0.16 ± 0.04	0.15 ± 0.01	10 %
Methyl formate	CH_3OCHO	Water	0.46 ± 0.04	0.39 ± 0.05	19 %
Nonacosane	$C_{29}H_{60}$	Carbon	2.04 ± 0.02	2.04 ± 0.00	0.0 %
Thymine	$C_5H_6N_2O_2$	Metal	1.01 ± 0.22	1.11 ± 0.02	−9.6 %
Acetonitrile	CH_3CN	Multilayer	0.35 ± 0.04	0.42 ± 0.04	−16 %
Ammonia	NH_3	Multilayer	0.26 ± 0.10	0.26 ± 0.01	9.9 %
Methyl formate	CH_3OCHO	Multilayer	0.43 ± 0.09	0.35 ± 0.02	10 %

Notes. [†]Literature study does not present uncertainty of allyl alcohol BE measurement.

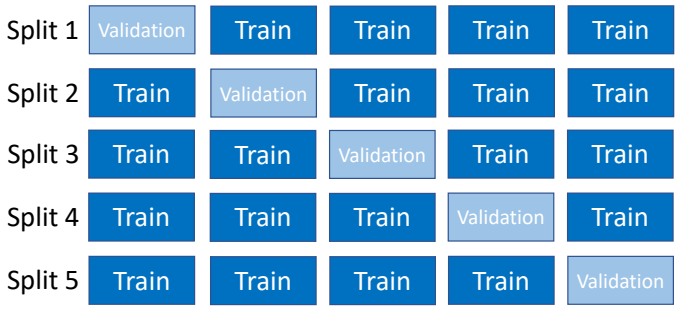


Fig. A.1: Schematic overview of 5-fold cross validation

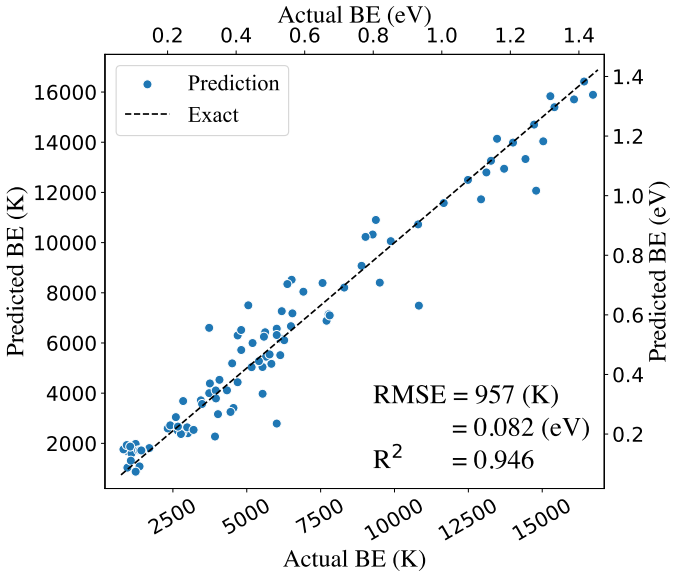


Fig. A.2: Parity plot for the monolayer case and including only BEs up to 17000 K.

Appendix A: Cross validation

Figure A.1 schematically displays how the five-fold cross validation is set up. Figure A.2 depicts the parity plot when only BEs up to 17000 K are included.

Appendix B: Literature data and molecular features

Table B.1 shows the most important features of molecules that are used as training data in this work. Tables B.1 and B.3 present the literature data used to train the ML model on[3].

[3] Electronic versions of the data files, along with Python scripts for producing the results, can be found in the Github repository here.

Table B.1: Features of molecules used for ML

Molecule name	Molecule formula	Mass	atoms	-OH	-C(O)-	-COOH	-C(O)O-	-O-	-NH_2	-CN	-N-C(O)-	Valence electrons	H-bond acceptor	H-bond donor	TPSA
		amu	#	#	#	#	#	#	#	#	#	#	#	#	Å^2
Methane	CH_4	16	5	–	–	–	–	–	–	–	–	8	0	0	0
Ammonia	NH_3	17	4	–	–	–	–	–	–	–	–	8	1	1	35
Water	H_2O	18	3	1	–	–	–	–	–	–	–	8	0	0	32
Acetylene	C_2H_2	26	4	–	–	–	–	–	–	–	–	10	0	0	0
Hydrogen cyanide	HCN	27	3	–	–	–	–	–	–	1	–	10	1	0	24
Carbon monoxide	CO	28	2	–	–	–	–	–	–	–	–	12	1	0	17
Ethylene	C_2H_4	28	6	–	–	–	–	–	–	–	–	12	0	0	0
Ethylene glycole	$(CH_2OH)_2$	62	10	2	–	–	–	–	–	–	–	12	2	2	40
Dinitrogen	N_2	28	2	–	–	–	–	–	–	–	–	10	2	0	48
Ethane	C_2H_6	30	8	–	–	–	–	–	–	–	–	14	0	0	0
Formaldehyde	H_2CO	30	4	–	1	–	–	–	–	–	–	12	1	0	17
Methylamine	CH_3NH_2	31	7	–	–	–	–	–	1	–	–	14	1	1	26
Methanol	CH_3OH	32	6	1	–	–	–	–	–	–	–	14	1	1	20
Oxygen	O_2	32	2	–	–	–	–	–	–	–	–	8	0	0	32
Hydroxylamine	NH_2OH	33	5	1	–	–	–	–	1	–	–	14	2	2	46
Methylacetylene	CH_3CCH	40	7	–	–	–	–	–	–	–	–	16	0	0	0
Acetonitrile	CH_3CN	41	6	–	–	–	–	–	–	1	–	16	1	0	24
methylisocyanide	CH_3NC	41	6	–	–	–	–	–	–	1	–	16	0	0	4
Propene	CH_2CHCH_3	42	9	–	–	–	–	–	–	–	–	18	0	0	0
Isocyanic acid	HNCO	43	4	–	–	–	–	–	–	–	–	16	2	1	41
Acetaldehyde	CH_3CHO	44	7	–	1	–	–	–	–	–	–	18	1	0	17
Carbon dioxide	CO_2	44	3	–	–	–	–	–	–	–	–	16	2	0	34
Ethylene oxide	c-C_2H_4O	44	7	–	–	–	–	1	–	–	–	18	1	0	13
Nitrous oxide	N_2O	44	3	–	–	–	–	–	–	–	–	16	2	0	51
Propane	C_3H_8	44	11	–	–	–	–	–	–	–	–	20	0	0	0
Formamide	NH_2CHO	45	6	–	–	–	–	–	–	–	1	18	1	1	43
Dimethylether	CH_3OCH_3	46	9	–	–	–	–	1	–	–	–	20	1	0	9
Ethanol	CH_3CH_2OH	46	9	1	–	–	–	–	–	–	–	20	1	1	20
Formic acid	HCOOH	46	5	–	–	1	–	–	–	–	–	18	1	1	37
Nitrogen dioxide	NO_2	46	3	–	–	–	–	–	–	–	–	17	3	0	54
Cyanoacetylene	HC_3N	51	5	–	–	–	–	–	–	1	–	18	1	0	24
Acrylonitrile	CH_2CHCN	53	7	–	–	–	–	–	–	1	–	20	1	0	24
Propionitrile	CH_3CH_2CN	55	9	–	–	–	–	–	–	1	–	22	1	0	24
Acetone	CH_3COCH_3	58	10	–	1	–	–	–	–	–	–	24	1	0	17
Allyl alcohol	CH_2CHCH_2OH	58	10	1	–	–	–	–	–	–	–	24	1	1	20
Propionaldehyde	CH_3CH_2CHO	58	10	–	1	–	–	–	–	–	–	24	1	0	17
Acetic acid	CH_3COOH	60	8	–	–	1	–	–	–	–	–	24	1	1	37
Glycolaldehyde	$HOCH_2CHO$	60	8	1	1	–	–	–	–	–	–	24	2	1	37
Glycolonitrile	$HOCH_2CN$	57	7	1	–	–	–	–	–	–	1	24	2	1	44
Methyl formate	CH_3OCHO	60	8	–	–	–	1	–	–	–	–	24	2	0	26
Pentane	C_5H_{12}	72	17	–	–	–	–	–	–	–	–	32	0	0	0

Continued on next page

Table B.1 – *Continued from previous page*

Molecule name	Molecule formula	Mass	atoms	-OH	-C(O)-	-COOH	-C(O)O-	-O-	-NH_2	-CN	-N-C(O)-	Valence electrons	H-bond acceptor	H-bond donor	TPSA
		amu	#	#	#	#	#	#	#	#	#	#	#	#	Å^2
N,N-DMF	$(CH_3)_2NCHO$	73	12	–	–	–	–	–	–	–	1	30	1	0	20
Ethyl formate	CH_3CH_2OCHO	74	11	–	–	–	1	–	–	–	–	30	2	0	26
Dicyanoacetylene	NCCCCN	76	6	–	–	–	–	–	–	2	–	26	2	0	48
Benzene	C_6H_6	78	12	–	–	–	–	–	–	–	–	30	0	0	0
Hexane	C_6H_{14}	86	20	–	–	–	–	–	–	–	–	38	0	0	0
Toluene	$CH_3C_6H_5$	92	15	–	–	–	–	–	–	–	–	36	0	0	0
1,1-Dichloroethane	CH_3CHCl_2	98	8	–	–	–	–	–	–	–	–	26	0	0	0
Heptane	C_7H_{16}	100	23	–	–	–	–	–	–	–	–	44	0	0	0
Benzonitrile	c-C_6H_5CN	103	13	–	–	–	–	–	–	1	–	38	1	0	24
Benzaldehyde	C_6H_5CHO	106	14	–	1	–	–	–	–	–	–	40	1	0	17
Ethylbenzene	$CH_3CH_2C_6H_5$	106	18	–	–	–	–	–	–	–	–	42	0	0	0
o-Xylene	$(CH_3)_2C_6H_4$	106	18	–	–	–	–	–	–	–	–	42	0	0	0
Cytosine	$C_4H_5N_3O$	111	13	–	–	–	–	–	1	–	1	42	3	2	72
Octane	C_8H_{18}	114	26	–	–	–	–	–	–	–	–	50	0	0	0
Trichloromethane	$CHCl_3$	118	5	–	–	–	–	–	–	–	–	26	0	0	0
Thymine	$C_5H_6N_2O_2$	126	15	–	–	–	–	–	–	–	1	48	2	2	66
Naphthalene	$C_{10}H_8$	128	18	–	–	–	–	–	–	–	–	48	0	0	0
Nonane	C_9H_{20}	128	29	–	–	–	–	–	–	–	–	56	0	0	0
Adenine	$C_5H_5N_5$	135	15	–	–	–	–	–	1	–	–	50	4	2	80
Decane	$C_{10}H_{22}$	142	32	–	–	–	–	–	–	–	–	62	0	0	0
1,2-Dichlorobenzene	$C_6H_4Cl_2$	146	12	–	–	–	–	–	–	–	–	42	0	0	0
Guanine	$C_5H_5N_5O$	151	16	–	–	–	–	–	1	–	1	56	4	3	100
Undecane	$C_{11}H_{24}$	156	35	–	–	–	–	–	–	–	–	68	0	0	0
Dodecane	$C_{12}H_{26}$	170	38	–	–	–	–	–	–	–	–	74	0	0	0
Tridecane	$C_{13}H_{28}$	184	41	–	–	–	–	–	–	–	–	80	0	0	0
Tetradecan	$C_{14}H_{30}$	198	44	–	–	–	–	–	–	–	–	86	0	0	0
Pentadecane	$C_{15}H_{32}$	212	47	–	–	–	–	–	–	–	–	92	0	0	0
Hexadecane	$C_{16}H_{34}$	226	50	–	–	–	–	–	–	–	–	98	0	0	0
Heptadecane	$C_{17}H_{36}$	240	53	–	–	–	–	–	–	–	–	104	0	0	0
2-Deoxyadenosine	$C_{10}H_{13}N_5O_3$	251	31	2	–	–	–	1	1	–	–	96	8	3	119
Octadecane	$C_{18}H_{38}$	254	56	–	–	–	–	–	–	–	–	110	0	0	0
2-Deoxyguanosine	$C_{10}H_{13}N_5O_4$	267	32	2	–	–	–	1	1	–	1	102	8	4	139
Nonadecane	$C_{19}H_{40}$	268	59	–	–	–	–	–	–	–	–	116	0	0	0
Icosane	$C_{20}H_{42}$	282	62	–	–	–	–	–	–	–	–	122	0	0	0
Henicosane	$C_{21}H_{44}$	296	65	–	–	–	–	–	–	–	–	128	0	0	0
Coronene	$C_{24}H_{12}$	300	36	–	–	–	–	–	–	–	–	108	0	0	0
Docosane	$C_{22}H_{46}$	310	68	–	–	–	–	–	–	–	–	134	0	0	0
Tricosane	$C_{23}H_{48}$	324	71	–	–	–	–	–	–	–	–	140	0	0	0
Coronene	$C_{24}H_{50}$	338	74	–	–	–	–	–	–	–	–	108	0	0	0
Tetracosane	$C_{25}H_{52}$	352	77	–	–	–	–	–	–	–	–	146	0	0	0
Pentacosane	$C_{26}H_{54}$	366	80	–	–	–	–	–	–	–	–	152	0	0	0
Hexacosane	$C_{27}H_{56}$	380	83	–	–	–	–	–	–	–	–	158	0	0	0

Continued on next page

Table B.1 – *Continued from previous page*

Molecule name	Molecule formula	Mass	atoms	-OH	-C(O)-	-COOH	-C(O)O-	-O-	-NH_2	-CN	-N-C(O)-	Valence electrons	H-bond acceptor	H-bond donor	TPSA
		amu	#	#	#	#	#	#	#	#	#	#	#	#	Å^2
Heptacosane	$C_{28}H_{58}$	394	86	–	–	–	–	–	–	–	–	164	0	0	0
Octacosane	$C_{29}H_{60}$	408	89	–	–	–	–	–	–	–	–	170	0	0	0
Octacosane	$C_{30}H_{62}$	422	92	–	–	–	–	–	–	–	–	170	0	0	0
Nonacosane	$C_{31}H_{64}$	436	95	–	–	–	–	–	–	–	–	176	0	0	0
Dotriacontane	$C_{32}H_{66}$	450	98	–	–	–	–	–	–	–	–	194	0	0	0
Tritriacontane	$C_{33}H_{68}$	464	101	–	–	–	–	–	–	–	–	200	0	0	0
Tetratriacontane	$C_{34}H_{70}$	478	104	–	–	–	–	–	–	–	–	206	0	0	0
Pentatriacontane	$C_{35}H_{72}$	492	107	–	–	–	–	–	–	–	–	212	0	0	0
Hexatriacontane	$C_{36}H_{74}$	506	110	–	–	–	–	–	–	–	–	218	0	0	0
Heptatriacontane	$C_{37}H_{76}$	520	113	–	–	–	–	–	–	–	–	224	0	0	0
Octatriacontane	$C_{38}H_{78}$	534	116	–	–	–	–	–	–	–	–	230	0	0	0
Nonatriacontane	$C_{39}H_{80}$	548	119	–	–	–	–	–	–	–	–	236	0	0	0
Tetracontane	$C_{40}H_{82}$	562	122	–	–	–	–	–	–	–	–	242	0	0	0
Hentetracontane	$C_{41}H_{84}$	576	125	–	–	–	–	–	–	–	–	248	0	0	0
Dotetracontane	$C_{42}H_{86}$	590	128	–	–	–	–	–	–	–	–	254	0	0	0
Tritetracontane	$C_{43}H_{88}$	604	131	–	–	–	–	–	–	–	–	260	0	0	0
Tetratetracontane	$C_{44}H_{90}$	618	134	–	–	–	–	–	–	–	–	266	0	0	0
Pentatetracontane	$C_{45}H_{92}$	632	137	–	–	–	–	–	–	–	–	272	0	0	0
Hexatetracontane	$C_{46}H_{94}$	646	140	–	–	–	–	–	–	–	–	278	0	0	0
Heptatetracontane	$C_{47}H_{96}$	660	143	–	–	–	–	–	–	–	–	284	0	0	0
Octatetracontane	$C_{48}H_{98}$	674	146	–	–	–	–	–	–	–	–	290	0	0	0
Nonatetracontane	$C_{49}H_{100}$	688	149	–	–	–	–	–	–	–	–	296	0	0	0
Pentacontane	$C_{50}H_{102}$	702	152	–	–	–	–	–	–	–	–	302	0	0	0
Henpentacontane	$C_{51}H_{104}$	716	155	–	–	–	–	–	–	–	–	308	0	0	0
Dopentacontane	$C_{52}H_{106}$	730	158	–	–	–	–	–	–	–	–	314	0	0	0
Tripentacontane	$C_{53}H_{108}$	744	161	–	–	–	–	–	–	–	–	320	0	0	0
Tetrapentacontane	$C_{54}H_{110}$	758	164	–	–	–	–	–	–	–	–	326	0	0	0
Pentapentacontane	$C_{55}H_{112}$	772	167	–	–	–	–	–	–	–	–	332	0	0	0
Hexapentacontane	$C_{56}H_{114}$	786	170	–	–	–	–	–	–	–	–	338	0	0	0
Heptapentacontane	$C_{57}H_{116}$	800	173	–	–	–	–	–	–	–	–	344	0	0	0
Octapentacontane	$C_{58}H_{118}$	814	176	–	–	–	–	–	–	–	–	350	0	0	0
Nonapentacontane	$C_{59}H_{120}$	828	179	–	–	–	–	–	–	–	–	356	0	0	0
Hexacontane	$C_{60}H_{122}$	842	182	–	–	–	–	–	–	–	–	362	0	0	0

Notes. We note that this table only lists some of the most significant features of the molecules, but does not provide the full feature list for training.

Table B.2: BEs of molecules at monolayer coverage

Name	Formula	Surface	Simplified Surface	E_{bin} (K)	E_{bin} (eV)	Reference
1,1-Dichloroethane	CH_3CHCl2	HOPG	Carbon	6134 ± 361	0.529 ± 0.031	Ulbricht et al. (2006)
1,2-Dichlorobenzene	$C_6H_4Cl_2$	HOPG	Carbon	8299 ± 722	0.715 ± 0.062	Ulbricht et al. (2006)
2-Deoxyadenosine	$C_{10}H1_3N_5O_3$	Au(100)	Metal	13471 ± 481	1.161 ± 0.041	Demers et al. (2002)
2-Deoxycytidine	$C_9H1_3N_3O_4$	Au(100)	Metal	13711 ± 241	1.182 ± 0.021	Demers et al. (2002)
2-Deoxyguanosine	$C_{10}H1_3N_5O_4$	Au(100)	Metal	14433 ± 241	1.244 ± 0.021	Demers et al. (2002)
Acetaldehyde	CH_3CHO	Ni	Metal	2847 ± 30	0.245 ± 0.003	Corazzi et al. (2021)
Acetaldehyde	CH_3CHO	ASW	Water	4330 ± 217	0.373 ± 0.019	Lasne et al. (2012)
Acetone	CH_3COCH_3	ASW	Water	4691 ± 241	0.404 ± 0.021	Lasne et al. (2012)
Acetonitrile †	CH_3CN	Ni	Metal	3305 ± 13	0.285 ± 0.001	Corazzi et al. (2021)
Acetonitrile †	CH_3CN	Olivine grains	Si	4400 ± 200	0.379 ± 0.017	Corazzi et al. (2021)
Acetonitrile †	CH_3CN	amorphous silica	Si	4595 ± 120	0.396 ± 0.010	Abdulgalil et al. (2013)
Acetonitrile †	CH_3CN	Pt(111)	Metal	4876 ± 361	0.420 ± 0.031	Tylinski et al. (2020)
Acetonitrile †	CH_3CN	alpha-quartz(0001)	Si	5225 ± 696	0.450 ± 0.060	Bertin et al. (2017)
Acetonitrile †	CH_3CN	Graphene	Carbon	5292 ± 361	0.456 ± 0.031	Tylinski et al. (2020)
Acetonitrile †	CH_3CN	Au(100)	Metal	5533	0.477	Solomun et al. (1989)
Acetonitrile †	CH_3CN	Graphene	Carbon	5653 ± 361	0.487 ± 0.031	Tylinski et al. (2020)
Acetonitrile	CH_3CN	ASW	Water	6184 ± 406	0.533 ± 0.035	Bertin et al. (2017)
Acetylene	C_2H_2	compact ASW	Water	3200 ± 220	0.276 ± 0.019	Behmard et al. (2019)
Acrylonitrile	CH_2CHCN	Au	Metal	4215 ± 51	0.363 ± 0.004	Toumi et al. (2016)
Adenine	$C_5H_5N_5$	Au	Metal	14794 ± 241	1.275 ± 0.021	Östblom et al. (2005)
Allyl alcohol	CH_2CHCH_2OH	Pd(111)	Metal	6014	0.518	Dostert et al. (2016)
Ammonia †	NH_3	HOPG	Carbon	2790 ± 144	0.240 ± 0.012	Bolina et al. (2005b)
Ammonia †	NH_3	Au	Metal	3007 ± 120	0.259 ± 0.010	Noble et al. (2013)
Ammonia †	NH_3	HOPG	Carbon	3007 ± 241	0.259 ± 0.021	Ulbricht et al. (2006)
Ammonia †	NH_3	Au	Metal	3067 ± 12	0.264 ± 0.001	Kruczkiewicz et al. (2021)
Ammonia	NH_3	Cut Forsterite (Mg_2SiO_4)	Si	3747 ± 397	0.323 ± 0.034	Suhasaria et al. (2015)
Benzaldehyde	C_6H_5CHO	ASW	Water	4811 ± 241	0.415 ± 0.021	Lasne et al. (2012)
Benzene	C_6H_6	ASW	Water	4691	0.404	Bahr & Kempter (2007)
Benzene	C_6H_6	HOPG	Carbon	5623 ± 962	0.485 ± 0.083	Ulbricht et al. (2006)
Benzonitrile	c-C_6H_5CN	Au(100)	Metal	9020	0.777	Solomun et al. (1989)
$C_{11}H_{24}$	$C_{11}H_{24}$	Graphite	Carbon	13266 ± 530	1.143 ± 0.046	[a]
$C_{12}H_{26}$	$C_{12}H_{26}$	Graphite	Carbon	14011 ± 560	1.207 ± 0.048	[a]
$C_{13}H_{28}$	$C_{13}H_{28}$	Graphite	Carbon	14726 ± 589	1.269 ± 0.051	[a]
$C_{14}H_{30}$	$C_{14}H_{30}$	Graphite	Carbon	15413 ± 616	1.328 ± 0.053	[a]
$C_{15}H_{32}$	$C_{15}H_{32}$	Graphite	Carbon	16077 ± 643	1.385 ± 0.055	[a]
$C_{16}H_{34}$	$C_{16}H_{34}$	Graphite	Carbon	16718 ± 668	1.441 ± 0.058	[a]
$C_{17}H_{36}$	$C_{17}H_{36}$	Graphite	Carbon	17340 ± 693	1.494 ± 0.060	[a]
$C_{18}H_{38}$	$C_{18}H_{38}$	Graphite	Carbon	17944 ± 717	1.546 ± 0.062	[a]
$C_{19}H_{40}$	$C_{19}H_{40}$	Graphite	Carbon	18531 ± 741	1.597 ± 0.064	[a]
$C_{20}H_{42}$	$C_{20}H_{42}$	Graphite	Carbon	19103 ± 764	1.646 ± 0.066	[a]
$C_{21}H_{44}$	$C_{21}H_{44}$	Graphite	Carbon	19661 ± 786	1.694 ± 0.068	[a]

Continued on next page

Table B.2 – *Continued from previous page*

Name	Formula	Surface	Simplified Surface	E_{bin} (K)	E_{bin} (eV)	Reference
$C_{22}H_{46}$	$C_{22}H_{46}$	Graphite	Carbon	20206 ± 808	1.741 ± 0.070	[a]
$C_{23}H_{48}$	$C_{23}H_{48}$	Graphite	Carbon	20738 ± 829	1.787 ± 0.071	[a]
$C_{24}H_{50}$	$C_{24}H_{50}$	Graphite	Carbon	21259 ± 850	1.832 ± 0.073	[a]
$C_{25}H_{52}$	$C_{25}H_{52}$	Graphite	Carbon	21770 ± 870	1.876 ± 0.075	[a]
$C_{26}H_{54}$	$C_{26}H_{54}$	Graphite	Carbon	22270 ± 890	1.919 ± 0.077	[a]
$C_{27}H_{56}$	$C_{27}H_{56}$	Graphite	Carbon	22761 ± 910	1.961 ± 0.078	[a]
$C_{28}H_{58}$	$C_{28}H_{58}$	Graphite	Carbon	23242 ± 930	2.003 ± 0.080	[a]
$C_{29}H_{60}$	$C_{29}H_{60}$	Graphite	Carbon	23715 ± 949	2.044 ± 0.082	[a]
$C_{30}H_{62}$	$C_{30}H_{62}$	Graphite	Carbon	24181 ± 967	2.084 ± 0.083	[a]
$C_{31}H_{64}$	$C_{31}H_{64}$	Graphite	Carbon	24638 ± 986	2.123 ± 0.085	[a]
$C_{33}H_{68}$	$C_{33}H_{68}$	Graphite	Carbon	25531 ± 1021	2.200 ± 0.088	[a]
$C_{34}H_{70}$	$C_{34}H_{70}$	Graphite	Carbon	25967 ± 1039	2.238 ± 0.090	[a]
$C_{35}H_{72}$	$C_{35}H_{72}$	Graphite	Carbon	26397 ± 1056	2.275 ± 0.091	[a]
$C_{36}H_{74}$	$C_{36}H_{74}$	Graphite	Carbon	26821 ± 1073	2.311 ± 0.092	[a]
$C_{37}H_{76}$	$C_{37}H_{76}$	Graphite	Carbon	27239 ± 1090	2.347 ± 0.094	[a]
$C_{38}H_{78}$	$C_{38}H_{78}$	Graphite	Carbon	27652 ± 1106	2.383 ± 0.095	[a]
$C_{39}H_{80}$	$C_{39}H_{80}$	Graphite	Carbon	28059 ± 1122	2.418 ± 0.097	[a]
$C_{40}H_{82}$	$C_{40}H_{82}$	Graphite	Carbon	28461 ± 1138	2.453 ± 0.098	[a]
$C_{41}H_{84}$	$C_{41}H_{84}$	Graphite	Carbon	28858 ± 1154	2.487 ± 0.099	[a]
$C_{42}H_{86}$	$C_{42}H_{86}$	Graphite	Carbon	29250 ± 1170	2.521 ± 0.101	[a]
$C_{43}H_{88}$	$C_{43}H_{88}$	Graphite	Carbon	29637 ± 1185	2.554 ± 0.102	[a]
$C_{44}H_{90}$	$C_{44}H_{90}$	Graphite	Carbon	30020 ± 1201	2.587 ± 0.103	[a]
$C_{45}H_{92}$	$C_{45}H_{92}$	Graphite	Carbon	30399 ± 1216	2.620 ± 0.105	[a]
$C_{46}H_{94}$	$C_{46}H_{94}$	Graphite	Carbon	30773 ± 1231	2.652 ± 0.106	[a]
$C_{47}H_{96}$	$C_{47}H_{96}$	Graphite	Carbon	31144 ± 1246	2.684 ± 0.107	[a]
$C_{48}H_{98}$	$C_{48}H_{98}$	Graphite	Carbon	31510 ± 1260	2.715 ± 0.109	[a]
$C_{49}H_{100}$	$C_{49}H_{100}$	Graphite	Carbon	31873 ± 1275	2.747 ± 0.110	[a]
$C_{50}H_{102}$	$C_{50}H_{102}$	Graphite	Carbon	32232 ± 1289	2.778 ± 0.111	[a]
$C_{51}H_{104}$	$C_{51}H_{104}$	Graphite	Carbon	32587 ± 1303	2.808 ± 0.112	[a]
$C_{52}H_{106}$	$C_{52}H_{106}$	Graphite	Carbon	32939 ± 1318	2.838 ± 0.114	[a]
$C_{53}H_{108}$	$C_{53}H_{108}$	Graphite	Carbon	33288 ± 1332	2.869 ± 0.115	[a]
$C_{54}H_{110}$	$C_{54}H_{110}$	Graphite	Carbon	33633 ± 1345	2.898 ± 0.116	[a]
$C_{55}H_{112}$	$C_{55}H_{112}$	Graphite	Carbon	33975 ± 1359	2.928 ± 0.117	[a]
$C_{56}H_{114}$	$C_{56}H_{114}$	Graphite	Carbon	34314 ± 1373	2.957 ± 0.118	[a]
$C_{57}H_{116}$	$C_{57}H_{116}$	Graphite	Carbon	34650 ± 1386	2.986 ± 0.119	[a]
$C_{58}H_{118}$	$C_{58}H_{118}$	Graphite	Carbon	34983 ± 1399	3.015 ± 0.121	[a]
$C_{59}H_{120}$	$C_{59}H_{120}$	Graphite	Carbon	35314 ± 1413	3.043 ± 0.122	[a]
$C_{60}H_{122}$	$C_{60}H_{122}$	Graphite	Carbon	35641 ± 1426	3.071 ± 0.123	[a]
C_9H_20	C_9H_20	Graphite	Carbon	11667 ± 467	1.005 ± 0.040	[a]
Carbon dioxide †	CO_2	ASW	Water	2105 ± 902	0.181 ± 0.078	Edridge et al. (2013)
Carbon dioxide †	CO_2	non-porous ASW	Water	2236 ± 25	0.193 ± 0.002	Noble et al. (2012)
Carbon dioxide †	CO_2	np-ASW	Water	2250 ± 20	0.194 ± 0.002	He et al. (2017)
Carbon dioxide †	CO_2	Amorphous silicate	Si	2271 ± 25	0.196 ± 0.002	Noble et al. (2012)

Continued on next page

Table B.2 – *Continued from previous page*

Name	Formula	Surface	Simplified Surface	E_{bin} (K)	E_{bin} (eV)	Reference
Carbon dioxide †	CO_2	crystalline H_2O	Water	2361 ± 25	0.203 ± 0.002	Noble et al. (2012)
Carbon dioxide †	CO_2	crystalline H_2O	Water	2393 ± 241	0.206 ± 0.021	Galvez et al. (2007)
Carbon dioxide †	CO_2	HOPG	Carbon	2430 ± 277	0.209 ± 0.024	Edridge et al. (2013)
Carbon dioxide †	CO_2	ASW	Water	2490 ± 241	0.215 ± 0.021	Galvez et al. (2007)
Carbon dioxide †	CO_2	Forsterite (Mg_2SiO_4)	Si	2514 ± 36	0.217 ± 0.003	Suhasaria et al. (2017)
Carbon dioxide †	CO_2	HAC1	Carbon	2706 ± 24	0.233 ± 0.002	Maté et al. (2019)
Carbon dioxide †	CO_2	HAC2	Carbon	2814 ± 24	0.242 ± 0.002	Maté et al. (2019)
Carbon dioxide †	CO_2	H_2O	Water	2860 ± 200	0.246 ± 0.017	Sandford & Allamandola (1990)
Carbon dioxide †	CO_2	Graphene	Carbon	3019 ± 180	0.260 ± 0.016	Takeuchi et al. (2017)
Carbon dioxide †	CO_2	Graphene	Carbon	3139 ± 241	0.270 ± 0.021	Smith & Kay (2019)
Carbon monoxide †	CO	Au	Metal	826 ± 24	0.071 ± 0.002	Collings et al. (2003)
Carbon monoxide †	CO	Au	Metal	855 ± 25	0.074 ± 0.002	Bisschop et al. (2006)
Carbon monoxide †	CO	Au	Metal	855 ± 25	0.074 ± 0.002	Öberg et al. (2005)
Carbon monoxide †	CO	Au (polycrystaline)	Metal	855 ± 24	0.074 ± 0.002	Fuchs et al. (2006)
Carbon monoxide †	CO	non-porous ASW	Water	863 ± 25	0.074 ± 0.002	Noble et al. (2012)
Carbon monoxide †	CO	Amorphous silicate	Si	867 ± 25	0.075 ± 0.000	Noble et al. (2012)
Carbon monoxide †	CO	non-porous ASW	Water	870	0.075	He et al. (2016)
Carbon monoxide †	CO	Amorphous silica	Si	878 ± 36	0.076 ± 0.003	Collings et al. (2015)
Carbon monoxide †	CO	Graphene	Carbon	958 ± 52	0.083 ± 0.004	Smith et al. (2016)
Carbon monoxide †	CO	porous ASW	Water	980	0.084	He et al. (2016)
Carbon monoxide †	CO	crystalline H_2O	Water	1009 ± 25	0.087 ± 0.002	Noble et al. (2012)
Carbon monoxide †	CO	ASW	Water	1016 ± 36	0.088 ± 0.003	Allouche et al. (1998)
Carbon monoxide †	CO	Forsterite (Mg_2SiO_4)	Si	1119 ± 12	0.096 ± 0.001	Suhasaria et al. (2017)
Carbon monoxide †	CO	Amorphous silica	Si	1143 ± 157	0.098 ± 0.014	Collings et al. (2015)
Carbon monoxide †	CO	compact ASW	Water	1155 ± 133	0.100 ± 0.011	Fayolle et al. (2016)
Carbon monoxide †	CO	high porous ASW	Water	1179 ± 24	0.102 ± 0.002	Collings et al. (2003)
Carbon monoxide †	CO	compact ASW	Water	1180 ± 131	0.102 ± 0.011	Fayolle et al. (2016)
Carbon monoxide †	CO	HAC1	Carbon	1239 ± 24	0.107 ± 0.002	Maté et al. (2019)
Carbon monoxide †	CO	HAC2	Carbon	1275 ± 24	0.110 ± 0.002	Maté et al. (2019)
Carbon monoxide †	CO	ASW	Water	1419 ± 71	0.122 ± 0.006	Smith et al. (2016)
Carbon monoxide †	CO	HOPG	Carbon	1564 ± 120	0.135 ± 0.010	Ulbricht et al. (2006)
Cytosine	$C_4H_5N_3O$	Au(100)	Metal	15035 ± 481	1.296 ± 0.041	Demers et al. (2002)
Decane	$C_{10}H_22$	MgO(100)	Metal	9369	0.807	Tait et al. (2005)
Decane	$C_{10}H_22$	Graphite	Carbon	12486 ± 499	1.076 ± 0.043	[a]
Dicyanoacetylene	NCCCCN	ASW	Water	5052 ± 601	0.435 ± 0.052	Guennoun et al. (2005)
Dicyanoacetylene	NCCCCN	Au	Metal	6134 ± 601	0.529 ± 0.052	Guennoun et al. (2005)
Dimethylether	CH_3OCH_3	Au (polycrystaline)	Metal	3300 ± 400	0.284 ± 0.034	Öberg et al. (2009)
Dimethylether	CH_3OCH_3	Crystalline H_2O	Water	4076 ± 503	0.351 ± 0.043	Lattelais et al. (2011)
Dotriacontane	$C_{32}H_{66}$	Graphite	Carbon	25089 ± 1004	2.162 ± 0.086	[a]
Ethane	C_2H_6	Au	Metal	2300 ± 300	0.198 ± 0.026	Öberg et al. (2009)
Ethane	C_2H_6	porous ASW	Water	2663	0.229	Behmard et al. (2019)

Continued on next page

Table B.2 – *Continued from previous page*

Name	Formula	Surface	Simplified Surface	E_{bin} (K)	E_{bin} (eV)	Reference
Ethane	C_2H_6	MgO(100)	Metal	2670	0.230	Tait et al. (2005)
Ethane	C_2H_6	Graphene	Carbon	2983 ± 149	0.257 ± 0.013	Smith et al. (2016)
Ethanol	CH_3CH_2OH	HOPG	Carbon	6014 ± 361	0.518 ± 0.031	Ulbricht et al. (2006)
Ethyl formate	CH_3CH_2OCHO	HOPG	Carbon	5196 ± 361	0.448 ± 0.031	Salter et al. (2019)
Ethyl formate	CH_3CH_2OCHO	ASW	Water	5833 ± 241	0.503 ± 0.021	Salter et al. (2019)
Ethylene	C_2H_4	porous ASW	Water	2600	0.224	Behmard et al. (2019)
Formaldehyde	H_2CO	Au	Metal	3765 ± 60	0.324 ± 0.005	Noble et al. (2012)
Formamide	NH_2CHO	Ru(001)	Metal	6545	0.564	Parmeter et al. (1988)
Formamide	NH_2CHO	np-ASW	Water	7700	0.664 ±	Chaabouni et al. (2018)
Formamide	NH_2CHO	HOPG	Carbon	7770	0.670	Chaabouni et al. (2018)
Guanine	$C_5H_5N_5O$	Au(100)	Metal	16418 ± 241	1.415 ± 0.021	Demers et al. (2002)
Heptane	C_7H16	Graphite	Carbon	9874 ± 395	0.851 ± 0.034	[a]
Hexane	C_6H14	MgO(100)	Metal	5581	0.481	Tait et al. (2005)
Hexane	C_6H_{14}	Graphite	Carbon	8886 ± 355	0.766 ± 0.031	[a]
Hydroxylamine	NH_2OH	Amorphous silicate	Si	6519 ± 24	0.562 ± 0.002	Congiu et al. (2012)
Isocyanic acid †	HNCO	HOPG	Carbon	3729 ± 192	0.321 ± 0.017	Noble et al. (2015)
Isocyanic acid †	HNCO	HOPG	Carbon	3957 ± 204	0.341 ± 0.018	Noble et al. (2015)
Isocyanic acid	HNCO	Copper	Metal	4450	0.383	Theulé et al. (2011)
Methane	CH_4	Forsterite (Mg_2SiO_4)	Si	1323 ± 12	0.114 ± 0.001	Suhasaria et al. (2017)
Methane	CH_4	ASW	Water	1371 ± 69	0.118 ± 0.006	Smith et al. (2016)
Methane	CH_4	MgO(100)	Metal	1455	0.125	Tait et al. (2005)
Methane	CH_4	HOPG	Carbon	1702 ± 120	0.147 ± 0.010	Ulbricht et al. (2006)
Methanol	CH_3OH	Au	Metal	4700 ± 500	0.405 ± 0.043	Öberg et al. (2009)
Methanol	CH_3OH	ASW	Water	5412	0.466	Bahr et al. (2008)
Methanol	CH_3OH	HOPG	Carbon	5773 ± 361	0.497 ± 0.031	Ulbricht et al. (2006)
Methyl formate	CH_3OCHO	Au	Metal	4000 ± 400	0.345 ± 0.034	Öberg et al. (2009)
Methyl formate	CH_3OCHO	HOPG	Carbon	4210	0.363	Burke et al. (2015a)
Methyl formate	CH_3OCHO	Crystalline H_2O	Water	4506 ± 529	0.388 ± 0.046	Burke et al. (2015b)
Methylacetylene	CH_3CCH	porous ASW	Water	4550 ± 230	0.392 ± 0.020	Behmard et al. (2019)
Methylamine	CH_3NH_2	np-ASW	Water	4200	0.362	Chaabouni et al. (2018)
Methylamine	CH_3NH_2	HOPG	Carbon	7000	0.603	Chaabouni et al. (2018)
Methylisocyanide	CH_3NC	Au	Metal	4352 ± 116	0.375 ± 0.010	Bertin et al. (2017)
Methylisocyanide	CH_3NC	alpha-quartz(0001)	Si	5165 ± 812	0.445 ± 0.070	Bertin et al. (2017)
Methylisocyanide	CH_3NC	ASW	Water	6267 ± 348	0.540 ± 0.030	Bertin et al. (2017)
N,N-Dimethylformamide	$(CH_3)_2NCHO$	HOPG	Carbon	6374 ± 481	0.549 ± 0.041	Ulbricht et al. (2006)
Naphthalene	$C_{10}H_8$	HOPG	Carbon	9261 ± 1082	0.798 ± 0.093	Ulbricht et al. (2006)
Nitrogen †	N_2	Graphene	Carbon	782 ± 39	0.067 ± 0.003	Smith et al. (2016)
Nitrogen †	N_2	Au	Metal	790 ± 25	0.068 ± 0.002	Öberg et al. (2005)
Nitrogen †	N_2	Au (polycrystaline)	Metal	790 ± 25	0.068 ±	Fuchs et al. (2006)
Nitrogen †	N_2	non-porous ASW	Water	790	0.068	He et al. (2016)
Nitrogen †	N_2	Au (polycrystaline)	Metal	800 ± 25	0.069 ± 0.002	Bisschop et al. (2006)
Nitrogen †	N_2	porous ASW	Water	900	0.078	He et al. (2016)
Nitrogen †	N_2	compact ASW	Water	1034 ± 133	0.089 ± 0.011	Fayolle et al. (2016)

Continued on next page

Table B.2 – *Continued from previous page*

Name	Formula	Surface	Simplified Surface	E_{bin} (K)	E_{bin} (eV)	Reference
Nitrogen †	N_2	nonporous ASW	Water	1082 ± 120	0.093 ± 0.010	Zubkov et al. (2007)
Nitrogen †	N_2	HAC1	Carbon	1119 ± 24	0.096 ± 0.002	Maté et al. (2019)
Nitrogen †	N_2	ASW	Water	1155 ± 58	0.100 ± 0.005	Smith et al. (2016)
Nitrogen	N_2	Amorphous silica	Si	1203 ± 145	0.104 ± 0.012	Collings et al. (2015)
Nitrogen †	N_2	HAC2	Carbon	1203 ± 24	0.104 ± 0.002	Maté et al. (2019)
Nitrogen †	N_2	Graphene	Carbon	1395 ± 70	0.120 ± 0.006	Smith et al. (2016)
Nitrogen dioxide	NO_2	HOPG	Carbon	4450 ± 722	0.383 ± 0.062	Ulbricht et al. (2006)
Nitrous oxide	N_2O	Amorphous silicate	Si	2772 ± 18	0.239 ± 0.002	Congiu et al. (2012)
Octane	C_8H18	MgO(100)	Metal	7565	0.652	Tait et al. (2005)
Octane	C_8H18	Graphite	Carbon	10800 ± 432	0.931 ± 0.037	[a]
Oxygen †	O_2	non-porous ASW	Water	914 ± 25	0.079 ± 0.002	Noble et al. (2012)
Oxygen †	O_2	non-porous ASW	Water	920	0.079	He et al. (2016)
Oxygen	O_2	Au (polycrystaline)	Metal	925 ± 25	0.080 ± 0.002	Fuchs et al. (2006)
Oxygen †	O_2	Amorphous silicate	Si	930 ± 25	0.080 ± 0.002	Noble et al. (2012)
Oxygen †	O_2	crystalline H_2O	Water	969 ± 25	0.084 ± 0.002	Noble et al. (2012)
Oxygen †	O_2	ASW	Water	1107 ± 55	0.095 ± 0.005	Smith et al. (2016)
Oxygen †	O_2	Amorphous silica	Si	1178 ± 218	0.102 ± 0.019	Collings et al. (2015)
Oxygen †	O_2	Forsterite (Mg_2SiO_4)	Si	1179 ± 12	0.102 ± 0.001	Suhasaria et al. (2017)
Oxygen †	O_2	Graphene	Carbon	1419 ± 71	0.122 ± 0.006	Smith et al. (2016)
Oxygen †	O_2	HOPG	Carbon	1443 ± 120	0.124 ± 0.010	Ulbricht et al. (2006)
Pentane	C_5H12	Graphite	Carbon	7808 ± 312	0.673 ± 0.027	[a]
Propane	C_3H_8	compact ASW	Water	3446 ± 280	0.297 ± 0.024	Behmard et al. (2019)
Propane	C_3H_8	MgO(100)	Metal	3488	0.301	Tait et al. (2005)
Propane	C_3H_8	Graphene	Carbon	3752 ± 188	0.323 ± 0.016	Smith et al. (2016)
Propene	CH_2CHCH_3	porous ASW	Water	3950 ± 60	0.340 ± 0.005	Behmard et al. (2019)
Thymine	$C_5H_6N_2O_2$	Au(100)	Metal	12930 ± 241	1.114 ± 0.021	Demers et al. (2002)
Toluene	$CH_3C_6H_5$	HOPG	Carbon	6916 ± 842	0.596 ± 0.073	Ulbricht et al. (2006)
Trichloromethane	$CHCl_3$	HOPG	Carbon	6495 ± 361	0.560 ± 0.031	Ulbricht et al. (2006)
Water †	H_2O	Amorphous silica	Si	4210 ± 120	0.363 ± 0.010	Collings et al. (2015)
Water †	H_2O	Forsterite (Mg_2SiO_4(011))	Si	6014	0.518	Smith et al. (2014)

Notes. †Species with degenerate entries for which the average of the BEs is used in the ML model. [a]Based on Paserba & Gellman (2001a,b); Gellman & Paserba (2002).

Table B.3: BEs of molecules at multilayer coverage

Name	Formula	Surface	E_{bin} (K)	E_{bin} (eV)	Reference
1,1-Dichloroethane	CH_3CHCl2	HOPG	5292 ± 361	0.456 ± 0.031	Ulbricht et al. (2006)
1,2-Dichlorobenzene	$C_6H_4Cl_2$	HOPG	6735 ± 601	0.580 ± 0.052	Ulbricht et al. (2006)
Acetic acid	CH_3COOH	Au	5730 ± 24	0.494 ± 0.002	Kruczkiewicz et al. (2021)
Acetonitrile †	CH_3CN	amorphous silica	4595 ± 120	0.396 ± 0.010	Abdulgalil et al. (2013)
Acetonitrile †	CH_3CN	Pt(111)	4876 ± 361	0.420 ± 0.031	Tylinski et al. (2020)
Acetonitrile †	CH_3CN	Graphene	5292 ± 361	0.456 ± 0.031	Tylinski et al. (2020)
Acetylene	C_2H_2	CsI	2800 ± 300	0.241 ± 0.026	Behmard et al. (2019)
Acrylonitrile	CH_2CHCN	Au	4215 ± 51	0.363 ± 0.004	Toumi et al. (2016)
Allyl alcohol	CH_2CHCH_2OH	Pd(111)	5292	0.456	Dostert et al. (2016)
Ammonia †	NH_3	HOPG	2790 ± 144	0.240 ± 0.012	Bolina et al. (2005b)
Ammonia †	NH_3	Au	3007 ± 120	0.259 ± 0.010	Noble et al. (2013)
Ammonia †	NH_3	HOPG	3007 ± 241	0.259 ± 0.021	Ulbricht et al. (2006)
Ammonia †	NH_3	Au	3067 ± 12	0.264 ± 0.001	Kruczkiewicz et al. (2021)
Ammonia †	NH_3	Cleaved Forsterite (Mg_2SiO_4)	3103 ± 108	0.267 ± 0.009	Suhasaria et al. (2015)
Ammonia †	NH_3	Cut Forsterite (Mg_2SiO_4)	3103 ± 84	0.267 ± 0.007	Suhasaria et al. (2015)
Ammonia †	NH_3	Au	3127 ± 120	0.269 ± 0.010	Noble et al. (2013)
Benzene †	C_6H_6	ASW	5052	0.435	Bahr & Kempter (2007)
Benzene †	C_6H_6	HOPG	5383 ± 217	0.464 ± 0.019	Salter et al. (2018)
Carbon dioxide †	CO_2	ASW	2019 ± 168	0.174 ± 0.014	Edridge et al. (2013)
Carbon dioxide †	CO_2	Amorphous silicate	2269 ± 80	0.196 ± 0.007	Noble et al. (2012)
Carbon dioxide †	CO_2	Forsterite (Mg_2SiO_4)	2574 ± 24	0.222 ± 0.002	Suhasaria et al. (2017)
Carbon dioxide †	CO_2	Graphene	2877 ± 241	0.248 ± 0.021	Smith & Kay (2019)
Carbon dioxide †	CO_2	H_2O:CH_3OH	2923 ± 289	0.252 ± 0.025	Edridge et al. (2013)
Carbon monoxide †	CO	Forsterite (Mg_2SiO_4)	806 ± 24	0.069 ± 0.002	Suhasaria et al. (2017)
Carbon monoxide †	CO	Amorphous silicate	831 ± 40	0.072 ± 0.003	Noble et al. (2012)
Carbon monoxide †	CO	crystalline H_2O	849 ± 55	0.073 ± 0.005	Noble et al. (2012)
Carbon monoxide †	CO	Au	855 ± 25	0.074 ± 0.002	Bisschop et al. (2006)
Carbon monoxide †	CO	Au	855 ± 25	0.074 ± 0.002	Öberg et al. (2005)
Carbon monoxide †	CO	Au (polycrystaline)	855 ± 24	0.074 ± 0.002	Fuchs et al. (2006)
Carbon monoxide †	CO	O_2	856 ± 15	0.074 ± 0.001	Acharyya et al. (2007)
Carbon monoxide †	CO	Au	858 ± 15	0.074 ± 0.001	Acharyya et al. (2007)
Carbon monoxide †	CO	Au	858 ± 15	0.074 ± 0.001	Acharyya et al. (2007)
Carbon monoxide †	CO	CsI	866 ± 68	0.075 ± 0.006	Fayolle et al. (2016)
Carbon monoxide †	CO	Amorphous silica	878 ± 36	0.076 ± 0.003	Collings et al. (2015)
Carbon monoxide †	CO	Graphene	958 ± 52	0.083 ± 0.004	Smith et al. (2016)
Carbon monoxide †	CO	ASW	1016 ± 36	0.088 ± 0.003	Allouche et al. (1998)
Carbon monoxide †	CO	CO	1280	0.110	Sandford & Allamandola (1988)
Cyanoacetylene †	HC_3N	ASW	4691 ± 962	0.404 ± 0.083	Borget et al. (2001)
Cyanoacetylene †	HC_3N	crystalline H_2O	4691 ± 962	0.404 ± 0.083	Borget et al. (2001)
Dimethylether	CH_3OCH_3	Au (polycrystaline)	3300 ± 400	0.284 ± 0.034	Öberg et al. (2009)

Continued on next page

Table B.3 – *Continued from previous page*

Name	Formula	Surface	E_{bin} (K)	E_{bin} (eV)	Reference
Ethane †	C_2H_6	Au	2300 ± 300	0.198 ± 0.026	Öberg et al. (2009)
Ethane †	C_2H_6	Graphene	2558 ± 126	0.220 ± 0.011	Smith et al. (2016)
Ethane †	C_2H_6	CsI	2600 ± 300	0.224 ± 0.026	Behmard et al. (2019)
Ethanol	CH_3CH_2OH	HOPG	5367 ± 361	0.462 ± 0.031	Ulbricht et al. (2006)
Ethyl formate	CH_3CH_2OCHO	HOPG	5479 ± 84	0.472 ± 0.007	Salter et al. (2019)
Ethylbenzene	$CH_3CH_2C_6H_5$	HOPG	7818 ± 1203	0.674 ± 0.104	Ulbricht et al. (2006)
Ethylene	C_2H_4	CsI	2200	0.190	Behmard et al. (2019)
Ethylene glycole	$(CH_2OH)_2$	Au	7500 ± 800	0.646 ± 0.069	Öberg et al. (2009)
Ethylene oxide	c-C_2H_4O	Au	2405 ± 241	0.207 ± 0.021	Schriver et al. (2004)
Formaldehyde	H_2CO	Au	3765 ± 60	0.324 ± 0.005	Noble et al. (2012)
Formamide	NH_2CHO	np-ASW	7700	0.664 ±	Chaabouni et al. (2018)
Formic acid †	HCOOH	Au	4607 ± 12	0.397 ± 0.001	Kruczkiewicz et al. (2021)
Formic acid †	HCOOH	Au	5000 ± 500	0.431 ± 0.043	Öberg et al. (2009)
Glycolaldehyde †	$HOCH_2CHO$	HOPG	5629	0.485	Burke et al. (2015a)
Glycolaldehyde †	$HOCH_2CHO$	Au	5900 ± 600	0.508 ± 0.052	Öberg et al. (2009)
Glycolonitrile	$HOCH_2CN$	Au	6976	0.601	Danger et al. (2012)
Hydrogen cyanide †	HCN	Au	3368 ± 120	0.290 ± 0.010	Noble et al. (2013)
Hydrogen cyanide †	HCN	Au	3608 ± 120	0.311 ± 0.010	Noble et al. (2013)
Isocyanic acid	HNCO	Copper	4450	0.383	Theulé et al. (2011)
Methane †	CH_4	HOPG	1266 ± 120	0.109 ± 0.010	Ulbricht et al. (2006)
Methane †	CH_4	Forsterite (Mg_2SiO_4)	1371 ± 36	0.118 ± 0.003	Suhasaria et al. (2017)
Methanol †	CH_3OH	Au	4700 ± 500	0.405 ± 0.043	Öberg et al. (2009)
Methanol †	CH_3OH	HOPG	5533 ± 361	0.477 ± 0.031	Ulbricht et al. (2006)
Methyl formate †	CH_3OCHO	Au	4000 ± 400	0.345 ± 0.034	Öberg et al. (2009)
Methyl formate †	CH_3OCHO	HOPG	4210	0.363	Burke et al. (2015a)
Methylacetylene	CH_3CCH	CsI	4200 ± 300	0.362 ± 0.026	Behmard et al. (2019)
Methylamine	CH_3NH_2	HOPG	3200	0.276	Chaabouni et al. (2018)
methylisocyanide	CH_3NC	Au	4352 ± 116	0.375 ± 0.010	Bertin et al. (2017)
N,N-Dimethylformamide	$(CH_3)_2NCHO$	HOPG	5533 ± 481	0.477 ± 0.041	Ulbricht et al. (2006)
Naphthalene	$C_{10}H_8$	HOPG	9261 ± 1082	0.798 ± 0.093	Ulbricht et al. (2006)
Nitrogen †	N_2	CsI	770 ± 68	0.066 ± 0.006	Fayolle et al. (2016)
Nitrogen †	N_2	Graphene	782 ± 39	0.067 ± 0.003	Smith et al. (2016)
Nitrogen †	N_2	Au	790 ± 25	0.068 ± 0.002	Öberg et al. (2005)
Nitrogen †	N_2	Au (polycrystaline)	790 ± 25	0.068 ± 0.002	Fuchs et al. (2006)
Nitrogen †	N_2	Au (polycrystaline)	800 ± 25	0.069 ± 0.002	Bisschop et al. (2006)
Nitrogen †	N_2	Amorphous silica	830 ± 36	0.072 ± 0.003	Collings et al. (2015)
Nitrogen †	N_2	Graphene	866 ± 43	0.075 ± 0.004	Smith et al. (2016)
Nitrogen †	N_2	Graphene	866 ± 43	0.075 ± 0.004	Smith et al. (2016)
Nitrogen dioxide	NO_2	HOPG	3969 ± 601	0.342 ± 0.052	Ulbricht et al. (2006)
o-Xylene	$(CH_3)_2C_6H_4$	HOPG	6832	0.609	Salter et al. (2018)
Oxygen †	O_2	Amorphous silicate	895 ± 36	0.077 ± 0.003	Noble et al. (2012)
Oxygen †	O_2	non-porous ASW	898 ± 30	0.077 ± 0.003	Noble et al. (2012)

Continued on next page

Table B.3 – *Continued from previous page*

Name	Formula	Surface	E_{bin} (K)	E_{bin} (eV)	Reference
Oxygen †	O_2	Amorphous silica	902 ± 24	0.078 ± 0.002	Collings et al. (2015)
Oxygen †	O_2	Au (polycrystaline)	925 ± 25	0.080 ± 0.002	Fuchs et al. (2006)
Oxygen †	O_2	crystalline H_2O	936 ± 40	0.081 ± 0.003	Noble et al. (2012)
Oxygen †	O_2	Forsterite (Mg_2SiO_4)	998 ± 24	0.086 ± 0.002	Suhasaria et al. (2017)
Oxygen †	O_2	Graphene	1034 ± 52	0.089 ± 0.004	Smith et al. (2016)
Oxygen †	O_2	HOPG	1082 ± 120	0.093 ± 0.010	Ulbricht et al. (2006)
Oxygen †	O_2	Graphene	1119 ± 56	0.096 ± 0.005	Smith et al. (2016)
Propane †	C_3H_8	Graphene	3187 ± 171	0.275 ± 0.015	Smith et al. (2016)
Propane †	C_3H_8	CsI	3600 ± 300	0.310 ± 0.026	Behmard et al. (2019)
Propene	CH_2CHCH_3	CsI	3500 ± 300	0.302 ± 0.026	Behmard et al. (2019)
Propionaldehyde	CH_3CH_2CHO	Pd(111)	4330	0.373	Dostert et al. (2016)
Propionitrile	CH_3CH_2CN	Au	4546 ± 84	0.392 ± 0.007	Couturier-Tamburelli et al. (2018)
Toluene	$CH_3C_6H_5$	HOPG	6074 ± 722	0.523 ± 0.062	Ulbricht et al. (2006)
Trichloromethane	$CHCl_3$	HOPG	5653 ± 361	0.487 ± 0.031	Ulbricht et al. (2006)
Water †	H_2O	HOPG	4799 ± 96	0.414 ± 0.008	Bolina et al. (2005a); Brown & Bolina (2007)
Water †	H_2O	HOPG	5533 ± 361	0.477 ± 0.031	Ulbricht et al. (2006)
Water †	H_2O	Au	5773 ± 60	0.497 ± 0.005	Fraser et al. (2001)
Water †	H_2O	Amorphous silica	5930 ± 240	0.511 ± 0.021	Collings et al. (2015)
Water †	H_2O	$Al_2O_3(11_20)$	5992 ± 101	0.516 ± 0.009	Haynes et al. (1992)

Appendix C: Astrochemically relevant species

Table C.1 presents the features of the astrochemically relevant molecules for which BEs are predicted in this work.

Table C.1: Features of astrochemically relevant molecules used for ML

Molecule name	Molecule formula	Mass	-OH	-C(O)-	-COOH	-C(O)O-	-O-	$-NH_2$	-CN	-N-C(O)-	Valence electrons	H-bond acceptor	H-bond donor	TPSA
		amu	#	#	#	#	#	#	#	#	#	#	#	$Å^2$
Cyanamide	NH_2CN	42	–	–	–	–	–	1	1	–	16	2	1	50
Ethanimine	CH_3CHNH	43	–	–	–	–	–	–	–	–	20	1	1	26
Vinylalcohol	CH_2CHOH	44	1	–	–	–	–	–	–	–	18	1	1	20
Propargylimine	HC_3HNH	53	–	–	–	–	–	–	–	–	20	1	1	24
Cyanomethanimine	HNCHCN	54	–	–	–	–	–	–	1	–	20	2	1	48
Methyl isocyanate	CH_3NCO	57	–	–	–	–	–	–	–	–	22	2	0	29
Acetamide	$CH_3C(O)NH_2$	59	–	–	–	–	–	–	–	1	24	1	1	43
N-Methylformamide	CH_3NHCHO	59	–	–	–	–	–	–	–	1	24	1	1	29
Carbamide / Urea	$NH_2C(O)NH_2$	60	–	–	–	–	–	–	–	2	24	1	2	69
Ethenediol	HOCHCHOH	60	2	–	–	–	–	–	–	–	24	2	2	40
Ethanolamine	$HOCH_2CH_2NH_2$	61	1	–	–	–	–	1	–	–	26	2	2	46
Allenyl acetylene	$H_2CCCHCCH$	64	–	–	–	–	–	–	–	–	24	0	0	0
Propargyl cyanide	$HCCCH_2CN$	65	–	–	–	–	–	–	1	–	24	1	0	24
Cyanoallene	CH_2CCHCN	65	–	–	–	–	–	–	1	–	24	1	0	24
Cyanopropyne	CH_3C_3N	65	–	–	–	–	–	–	1	–	24	1	0	24
n-Propylcyanide	$CH_3CH_2CH_2CN$	69	–	–	–	–	–	–	1	–	28	1	0	24
i-Propylcyanide	$CH_3CH(CN)CH_3$	69	–	–	–	–	–	–	1	–	28	1	0	24
Hydroxyacetone	$CH_3C(O)CH_2OH$	74	1	1	–	–	–	–	–	–	30	2	1	37
Cyanovinylacetylene	HCCCHCHCN	77	–	–	–	–	–	–	1	–	28	1	0	24
Vinylcyanoacetylene	H_2CCHC_3N	77	–	–	–	–	–	–	1	–	28	1	0	24
Methylcyanodiacetylene	CH_3C_5N	89	–	–	–	–	–	–	1	–	28	1	0	24
Cyanoacetyleneallene	H_2CCCHC_3N	89	–	–	–	–	–	–	1	–	36	1	0	24
1-cyano-1,3-cyclopentadiene	c-C_5H_5CN	91	–	–	–	–	–	–	1	–	34	1	0	20
Cyanotriacetylene	HC_7N	99	–	–	–	–	–	–	1	–	34	1	0	24
Indene	c-C_9H_8	116	–	–	–	–	–	–	–	–	44	0	0	0
Cyanotetraacetylene	HC_9N	123	–	–	–	–	–	–	1	–	42	1	0	24
Cyanopentaacetylene	$HC_{11}N$	147	–	–	–	–	–	–	1	–	50	1	0	24

Notes. We note that this table only lists some of the most significant features of the molecules, but does not provide the full feature list for training.