Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2023, **2**, 952

# Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Received 23rd February 2023
Accepted 8th May 2023

DOI: 10.1039/d3dd00020f

rsc.li/digitaldiscovery

Machine learning techniques have been previously used to model and predict column densities in the TMC-1 dark molecular cloud. In interstellar sources further along the path of star formation, such as those where a protostar itself has been formed, the chemistry is known to be drastically different from that of largely quiescent dark clouds. To that end, we have tested the ability of various machine learning models to fit the column densities of the molecules detected in source B of the Class 0 protostellar system IRAS 16293-2422. By including a simple encoding of isotopic composition in our molecular feature vectors, we also examine for the first time how well these models can replicate the isotopic ratios. Finally, we report the predicted column densities of the chemically relevant molecules that may be excellent targets for radioastronomical detection in IRAS 16293-2422B.



## 1 Introduction

The observation of interstellar molecules is a central component of astrochemical studies. Molecular species have shaped our understanding of star[1] and planet formation,[2] can trace stellar outflows,[3] interstellar shocks,[4] and protoplanetary disks,[5] and can serve as probes of the physical conditions of interstellar sources such as the temperature.[6] However, until recently, in order to model interstellar abundances and predict new molecules for detection, observations have relied on complex chemical models based on a vast network of interconnected reactions (*e.g.* Ruaud *et al.*,[7] Wakelam *et al.*[8]). While these astrochemical models can be excellent tools to explore specific chemical processes that occur in space, their predictive ability can also be quite limited for several reasons (*e.g.* McGuire *et al.*[9]). Firstly, these models are by definition incomplete representations of the true chemical complexity of the interstellar medium because network expansions rely on human input. Additionally, the networks are oftentimes dependent on uncertain extrapolated rate constants.[10]

In an attempt to predict molecular abundances without the need for complete networks, Lee *et al.*[11] introduced a novel methodology involving machine learning. A major benefit of their approach contrasts traditional astrochemical modeling, as it requires no prior knowledge of the conditions of an interstellar source or any reaction pathways involving the previously detected molecules. Instead, abundances are expressed purely in terms of a chemical vector space. Simple regression algorithms were shown to significantly outperform traditional astrochemical models in reproducing the abundances of molecules already observed, and provided a straightforward way to extrapolate to yet undetected molecules.

An interstellar source for which this machine learning technique could be effectively applied is the Class 0 protostar IRAS 16293-2422B (hereafter referred to as IRAS 16293B). IRAS 16293B is one component of the protostellar system IRAS 16293, which is located in the L1689 region of the ρ Ophiuchus cloud complex. Interferometric observations initially revealed two protostellar sources in IRAS 16293 (source A and source B), separated by around 5.1″.[12–14] Further high-resolution studies then confirmed that source A is in fact composed of two compact sources (source A1 and A2), making IRAS 16293 a triple protostellar system.[15] Extensive observations have been made of this source with the Atacama Large Millimeter/submillimeter Array (ALMA) as part of the Protostellar Interferometric Line Survey (PILS) program.[16] The submillimeter spectrum toward IRAS 16293B is especially rich with more than 10 000 features detected.[16] The line widths of the spectral peaks are also extremely narrow for a star forming region (~1 km $s^{-1}$ FWHM), which significantly reduces line confusion and makes this an excellent source for molecular detections.

[a]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: zfried@mit.edu*

[b]*Accelerated Computing Systems and Graphics Group, Intel Corporation, 2111 NE 25th Ave., Hillsboro, OR 97124, USA*

[c]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

[d]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: brettmc@mit.edu*

[e]*National Radio Astronomy Observatory, Charlottesville, VA 22903, USA*

The predictive power of the machine learning method introduced by Lee *et al.*[11] may be especially useful for IRAS 16293B since a large portion of the molecular lines in the interstellar line survey remain unassigned. In fact, as of 2018, Taquet *et al.*[17] noted that approximately 70% of the 5$\sigma$ transitions identified in the ALMA Band 6 dataset were unidentified. If successful, this method might be able to provide an unbiased list of astrochemical targets not yet detected but which might be abundant enough to be contributing to the unidentified molecular lines. If subsequently detected, these molecules and their abundances could then be used to further constrain both the machine learning model and traditional network-based astrochemical models of low mass protostars. These models provide invaluable insight into the chemical processes and conditions relevant to the formation of stars similar to our Sun.

One aspect of interstellar chemistry that was not treated in the work of Lee *et al.*[11] was the incorporation of isotopically substituted species. While certainly such rare isotopologues are present and detectable in TMC-1,[18] detections of these species are more common toward IRAS 16293B and therefore provide substantially more insight into the chemical and physical history of this source. Therefore, it is desirable to update the machine learning model to include isotopologues. Molecules in IRAS 16293B consistently display isotopic ratios that are enhanced compared to the mean solar value and other interstellar sources, especially deuterium (D) and $^{13}$C substituted species. In fact, the deuterated isotopologues of ethanol, ketene, acetaldehyde, formic acid, formamide, and isocyanic acid were all first detected toward this source.[19,20] Various doubly and triply deuterated species have been detected as well (*e.g.* Ilyushin *et al.*,[21] Persson *et al.*[22]). Additionally, the $^{12}$C/$^{13}$C ratios of dimethyl ether, methyl formate, ethanol, and glycolaldehyde toward IRAS 16293B are all much lower than the $^{12}$C/$^{13}$C ratio of the local ISM.[16,19] By convention, the deuterium ratios are reported as D/H while the $^{13}$C ratios are reported in the inverse manner. Therefore, a high D/H ratio and low $^{12}$C/$^{13}$C ratio both denote isotopic enhancement.

A large portion of the remaining unassigned spectral peaks are predicted to arise from isotopically substituted species. In fact, Jørgensen *et al.*[16] note that only 25% of the transitions correspond to the most common organic molecules detected in hot cores, including formaldehyde, methanol, methyl cyanide, isocyanic acid, ethanol, acetaldehyde, methyl formate, dimethyl ether, and ketene. Following this, they predict that the majority of the remaining transitions are likely related to various isotopically substituted molecules as well as more complex organic species. Thus, it is also vital to accurately model the column densities of isotopically substituted molecules so that the high abundance isotopologues in this source can be predicted, measured as needed in the laboratory, and their signals in the PILS survey identified and assigned.

Additionally, machine learning predictions of isotopic ratios are also useful since these ratios can act as tracers of the evolutionary history of an interstellar source along with the conditions, timescales, and pathways of molecular formation. For example, deuterium fractionation relies on gas-phase isotope exchange reactions that are strongly dependent on the temperature. Consequently, the deuterium fraction is a tracer of the conditions of the interstellar environment during molecular formation, with a high D/H ratio (*i.e.* high deuterium fraction) indicating cold formation temperatures.[23,24] Therefore, accurate prediction of isotopic ratios would allow us to gain insight into the details of molecular formation and source history without requiring a dedicated search for these isotopically-substituted species that are often present in fairly low abundance.

In this work we apply the machine learning technique introduced by Lee *et al.*[11] to IRAS 16293B. The machine learning pipeline used for this project along with the isotopic encoding is described in Section 3. Section 4.1 then presents the ability of the supervised machine learning regressors to model the molecular column densities in this source. Using these trained regression models, we obtain an unbiased list of predicted high-abundance targets for astronomical observation. Analysis of these molecular targets is presented in Section 4.2. Next, in Section 4.3 we test the ability of the regressors to model the isotopic ratios in this source. Finally, a list of high predicted column density isotopologues is provided in Section 4.4.

## 2 Dataset

The molecules included in the dataset for this work were mostly detected through observations from the PILS survey.[16] IRAS 16293A is an edge-on disk system while IRAS 16293B has a face-on orientation. This results in the line widths of the spectral peaks toward source A being much wider.[16] Consequently, there is much more overlap of the spectral peaks, making the identification of individual signals more challenging and resulting in fewer definitive molecular detections toward source A. Our analysis therefore focused solely on IRAS 16293B. Most molecules were detected at a one-beam (0.5″) offset position from the continuum peak of source B in the south-west direction. The coordinates of this one-beam offset position are $\alpha_{J2000} = 16^h32^m22^s.58$, $\delta_{J2000} = -24°28'32.8''$. A few of the species, however, were detected at a half-beam offset. For these molecules, the column densities have been reduced by a factor of 2.136 to account for this different pointing position.[25] In total, our dataset includes 98 molecules. Of these, 43 are main isotopologues and 55 are isotopically substituted species. All molecules in the dataset are listed in Table 5 in the Appendix. Our dataset contains 27 deuterium substituted species, 15 $^{13}$C substituted species, two $^{15}$N substituted species, four $^{34}$S substituted species, two $^{33}$S substituted species, two $^{17}$O substituted species, and three $^{18}$O substituted species. Additionally, several doubly deuterated molecules along with one triply deuterated molecule is included.

## 3 Model description

The machine learning pipeline has three main components: (1) molecular featurization, (2) modeling the column densities in IRAS 16293B using supervised regressors, and (3) prediction of high abundance astrochemical targets. An outline of this entire process is depicted in Fig. 8 in the Appendix. In this work, the process of molecular featurization and regression is quite similar to the methods introduced by Lee *et al.*[11] In summary,

a Mol2vec[26] model is first trained on a dataset of 3 634 046 molecules collected from various online databases like Pubchem,[27] ZINC,[28] and the NASA PAH database.[29–31] This trained embedding model then creates 70 dimensional feature vectors for all molecules in the dataset, in addition to the detected interstellar species.

In order to include isotopologues in the training set and investigate isotopic ratios, it was necessary to encode isotopic composition in the feature vectors. In its current form, Mol2vec is not able to fully capture isotopic information. For example, it creates unique vectors for deuterium-substituted molecules but not for molecules that are substituted with $^{13}$C. This is because the molecular substructures are first encoded using Morgan fingerprints, which do not by-default capture differences in $^{13}$C-substituted isotopologues (*e.g.* the default RDKit-constructed[32] Morgan fingerprint[33] of $H_2CO$ and $H_2{}^{13}CO$ are identical). To differentiate between each of the isotopologues in our dataset, we ensure that the Mol2vec-generated vectors are identical for all isotopologues of the same species and then add 19 extra dimensions that encode isotopic information as well as the chemical environment of the isotopic substitution (Fig. 1). The isotopic encoding is designed as follows:

- Dimensions 1–9: number of D, $^{34}$S, $^{33}$S, $^{36}$S, $^{13}$C, $^{17}$O, $^{18}$O, $^{15}$N, and $^{37}$Cl atoms in the molecule.
- Dimensions 10–12: whether the $^{13}$C atoms are sp, sp$^2$ or sp$^3$ hybridized.
- Dimension 13: whether the substituted $^{13}$C atom is bonded to oxygen.
- Dimension 14–15: whether a deuterium atom is bonded to carbon or oxygen.
- Dimension 16: number of non-hydrogen atoms in the molecule.
- Dimensions 17–18: whether there is an oxygen or carbon atom two bonds away from the substituted deuterium.
- Dimension 19: number of deuterium atoms bonded to carbonyl carbons.

The RDKit module was used to obtain hybridization and bonding information.[32] Because of the limited number of unique isotopologues in our dataset, the selection of these hand-picked features was largely dependent on which chemical substructures are present in enough molecules to constitute a reasonably sized training set. More specifically, each of the isotopic features denoted by dimensions 10–19 are present in at least three molecules in the dataset. Each of the selected features also has a notable impact on the average isotopic ratio.

Additionally, Mol2vec was unable to differentiate between several conformers of the detected molecules. Examples include ethylene glycol (for which both the aGg′ and gGg′ conformers were detected) as well as monodeuterated $CH_2DCH_2OH$ and $CH_2DOCH_3$. For consistency, we inputted the column density of the most stable or abundant conformer in each case.

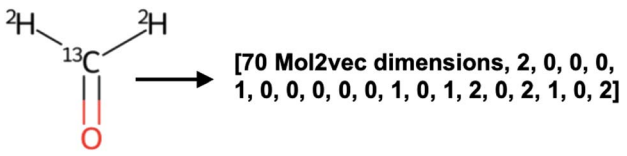


Fig. 1 Depiction of how isotopic composition is encoded in the molecular feature vectors.

Using the resulting feature vectors as inputs and the $\log_{10}$ column densities as outputs, the data was split 80/20 into training and testing sets. In order to mitigate data leakage, all isotopologues of the same molecule were assigned to either the training or testing set. The datapoints were then bootstrapped with Gaussian noise in order to increase the effective dataset size to 800 and control overfitting.

These resulting training and testing sets were then fed into two separate supervised machine learning regressors: Gaussian process regression (GPR)[34] and Bayesian ridge regression (BR). These models learn relationships between the vector components to map the molecular features to the column density data. Each of the models were implemented with the SCIKIT-LEARN Python module.[35] We determined the optimal hyperparameters for each model by first splitting the data into training and testing sets and then running a 5-fold grid search on the training data.

GPR is a nonparametric model that defines a probability distribution over all functions that can map the molecular descriptors to the column densities. It is therefore able to handle nonlinear relationships in the data. A kernel provides the model with prior knowledge regarding the shape and smoothness of the functions. Similarly to Lee *et al.*,[11] the kernel we used was a linear combination of the rational quadratic, dot product, and white noise kernels. Along with the kernel function, additional hyperparameters include a noise value added to the kernel matrix diagonal that denotes the inherent Gaussian noise of the training observations.

BR is a linear regressor that takes a probabilistic approach to optimize the ridge regression model coefficients. It does this by using a gamma distribution prior for the regularization coefficients. These parameters are then optimized through maximization of the log marginal likelihood. For this regressor, the hyperparameters define the shape and inverse scale of the Gamma distribution priors over the various model parameters.

Similarly to Lee *et al.*,[11] a linear model is included in order to provide a baseline performance using an extremely simple model with a limited number of parameters. Bayesian ridge was specifically chosen due to its ability to report prediction uncertainties, which allows us to gauge the confidence level of the predicted values. A GPR model then displays the ability of a more complex model to improve upon this baseline regressor. GPR was also chosen due to its probabilistic and nonlinear nature. The reported uncertainties are fairly informative and interpretable since they can be linked directly to the designed covariance matrix.

Following the training of the regression models, the column densities of the molecules that are most chemically similar to those detected in IRAS 16293B were predicted using the trained models. K-means clustering with $k = 10$ was used to cluster the entire dataset of 3 634 046 feature vectors. Each of the molecules detected in IRAS 16293B was assigned to a single cluster. Thus, we only considered the molecules assigned to this cluster when searching for detectable new species.

When analyzing the ability of the regressors to model the molecules in the training set and subsequently predict the

column densities of the species in the testing set, we were limited to the molecules that have been previously detected toward IRAS 16293B. We were therefore only able to gauge the performance of these models on molecules that are relevant to this fairly small and homogeneous dataset. Consequently, we proceeded to remove the molecules that had no or few chemically similar examples in the dataset since the models did not have sufficient training examples to learn the required relationships for these species and we had no ability to gauge the accuracy of the model predictions. For this additional filtering, we removed molecules that contained atoms other than hydrogen, carbon, sulfur, nitrogen, and oxygen because all of the previously detected molecules were predominantly composed of these atoms. Additionally, we also removed the remaining free radical species because nitrous oxide is the only free radical for which a column density has been derived toward IRAS 16293B. This minuscule training set of free radicals resulted in the models not being able to sufficiently handle these molecules. For example, the free radical counter-parts of various molecules had much higher predicted column densities than the observed values of the parent species. This is very unlikely due to the instability of free radicals and the general under abundance of radicals in protostellar sources. In fact, while 36.5% of interstellar molecules were first detected in star-forming regions, only 13.5% of radical species were first detected in these protostellar sources.[36] This under abundance may be due to the larger gas-phase chemical inventory and warmer kinetic temperatures leading to a greater number of destruction partners for these highly reactive species. Following these filtering steps, the column densities of the remaining 84 863 molecules were then predicted using the previously trained regression models.

# 4 Results and discussion

## 4.1 Regression analysis

Fig. 2 shows the training and testing results of each regression model. The train/test splits were consistent in each case. Just as shown in Lee *et al.*[11] 's study of TMC-1, each of the regressors were able to accurately model the column densities in this source. The strong performance on the test set also provides confidence that the models can generalize well to relevant molecules that were not included in the training set.

Additionally, although the BR model showed an ability to precisely model the column densities with lower uncertainties than the GPR regressor, it was mainly limited by its linear mapping. With our current isotope encoding, a linear model will be unable to fully capture the relevant isotopic fractionation. For example, the difference between a singly and doubly deuterated molecule is in-part denoted with a 2 instead of a 1 in a single vector dimension. That said, the difference in the column densities of singly and doubly deuterated species is typically not simply 2/1 and can differ significantly between molecules. Thus, for the remainder of the analysis, a GPR model was used since a nonlinear mapping was required.

The large error bars on the GPR predictions are in part due to the small size of the dataset. Additional molecular detections (especially of main isotopologues) toward this source will allow for further constrained predictions. Moreover, despite the overall strong performance of the GPR model, this regressor overpredicts ethylene glycol ($OHCH_2CH_2OH$) and dimethyl ether ($CH_3OCH_3$) by over one order of magnitude. This prediction inaccuracy is likely because these molecules have few nearby neighbors in the training set. In fact, these molecules are the $10^{th}$ and $12^{th}$ furthest species from any neighbor in the dataset, respectively. Ethylene glycol is especially unique in that it is the only molecule in the dataset containing two hydroxyl groups. Additionally, ethylene glycol's nearest neighbors are methoxymethanol and ethanol, each of which are more abundant.

Another notable prediction error is the slight overprediction of chloromethane and underprediction of methanol since it highlights the shortcomings of our molecular featurization. Mol2vec creates molecular feature vectors by combining vector representations of chemical substructures. Therefore, small molecules with some shared substructures have extremely

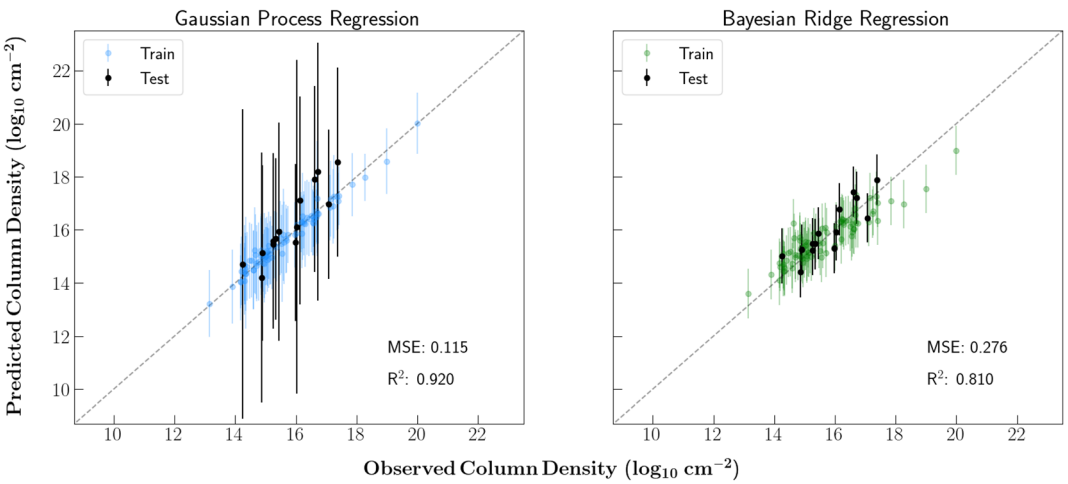


Fig. 2 Training and testing results of each supervised regression model. An 80/20 train/test split was used. The reported MSE and $R^2$ are for the combined testing and training data. 1$\sigma$ uncertainties are shown.

similar feature vectors. In this case, Mol2vec generates similar feature vectors for all molecules that contain a methyl group bonded to a single heteroatom. Despite obvious chemical differences, the resulting vector representations of chloromethane and methanol are therefore very similar. Since methanol is one of the most abundant molecules in the dataset and chloromethane is one of the least abundant, the resulting prediction errors are fairly unsurprising.

In order to test the efficacy of the chosen kernel for the GPR model, we predicted the column density of cimetidine $C_{10}H_{16}N_6S$. This molecule is far more complex than any other species in our dataset and would certainly have an extremely low abundance in the interstellar medium. Therefore, an effective kernel would also produce a low column density prediction for this species. Ultimately, the trained GPR model predicted this molecule to have a column density of $6.94 \times 10^6$ cm$^{-2}$, which is nearly eight orders of magnitude lower than any main isotopologue in the dataset. This provides confidence that the model is not over-fitting to the dataset and simply learning to predict each column density to be in the range of the detected species.

### 4.2 Targets for astrochemical study

Using the trained GPR model, we then predicted the column densities of the aforementioned 84 863 astrochemically relevant molecules assigned to the same cluster as the IRAS 16293B

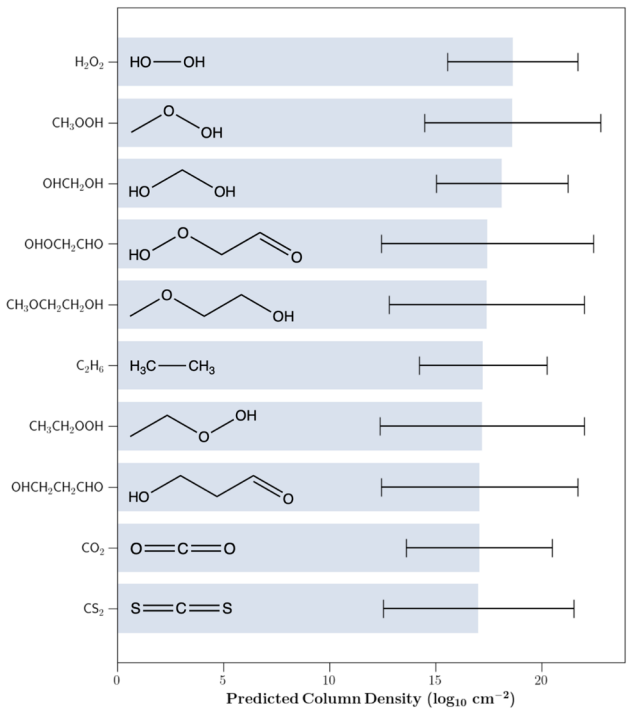


Fig. 3 The 10 undetected molecules with the highest predicted column densities. The molecules were drawn from the collection of species that were assigned to the same cluster as each of the detected molecules through k-means clustering. Predictions were made using the trained Gaussian process regression model. 1$\sigma$ column density uncertainties are shown.

Table 1 Average chemical composition of the 20 highest and lowest predicted abundance astrochemical targets

| | 20 highest predicted abundance molecules | 20 lowest predicted abundance molecules |
|---|---|---|
| Mean # of oxygen atoms | 1.65 | 0.75 |
| Mean # of nitrogen atoms | 0.10 | 1.05 |
| Mean # of sulfur atoms | 0.10 | 0.75 |
| Mean degree of unsaturation | 0.75 | 1.65 |
| Mean # of heavy atoms | 3.85 | 4.25 |
| Mean molecular weight (amu) | 59.66 | 74.35 |

detections. To enhance the confidence in our predicted values and further limit our investigation to chemically relevant molecules, our analysis was solely focused on species for which the 1$\sigma$ prediction uncertainty was less than five orders of magnitude; there were 242 such species. Fig. 3 shows the 10 molecules with the highest predicted column densities. All predictions are provided in the associated GitHub repository.

The chemical composition of the predicted molecules is displayed in Table 1. Oxygenated hydrocarbons were typically predicted to be in high abundance while those containing nitrogen and sulfur were predicted to have lower column densities. Of the 20 highest predicted column density molecules, 15 contain at least one oxygen atom, while 2 contain a nitrogen atom, and 1 contains a sulfur atom. The preference for oxygen-substituted molecules is not surprising since the most abundant detected species in IRAS 16293B are carbon monoxide (CO), methanol ($CH_3OH$), formaldehyde ($H_2CO$), methyl formate ($HCOOCH_3$), dimethyl ether ($CH_3OCH_3$), carbonyl sulfide (OCS), and ethanol ($CH_3CH_2OH$) – each of which contain an oxygen atom.

Highly saturated molecules were also predicted to be very abundant in IRAS 16293B. This is to be expected in a protostellar source since hydrogenation is very efficient on grain surfaces. Therefore, many of the species that are sublimated from grains as the protostar heats the surroundings are highly saturated (*e.g.* Linnartz *et al.*,[37] Fedoseev *et al.*,[38] Woon,[39] Garrod *et al.*[40]).

The predictions also display a preference for lighter molecules that contain less heavy atoms. This also matches the detected chemical inventory in IRAS 16293B, in which the seven highest abundance molecules each contain four or less heavy atoms.

The proceeding subsections highlight the astrochemical relevance of some of the molecules with the highest predicted column densities in Fig. 3. It is important to note that while a high column density is beneficial for interstellar molecular detectability, various additional factors must also be considered including the magnitude of the dipole moment, the spectral pattern, and intrinsic line strengths.

**4.2.1 Hydrogen peroxide ($H_2O_2$).** Hydrogen peroxide was first observed toward the SM1 core in ρ Oph A through several

torsion–rotation transitions.[41] This molecule is proposed to form on grain surfaces *via* successive hydrogen additions to $O_2$.[42] If hydrogen peroxide were to be detected, another possible molecular candidate is $HO_2$. This radical is generated after the first of these successive hydrogen additions to molecular oxygen and was also detected toward ρ Oph A with a similar abundance to hydrogen peroxide.[43]

**4.2.2 Methyl hydroperoxide ($CH_3OOH$).** Methyl hydroperoxide is a very attractive candidate for interstellar detection. This organic peroxide is structurally similar to many of the small chemical species that have been detected in the interstellar medium, such as hydrogen peroxide.[41] Additionally, multiple energetically feasible products of methyl hydroperoxide UV photodissociation have been detected in space, including the OH and $CH_3O$ radicals.[44–46] The microwave and millimeter-wave spectra of this molecule have been previously experimentally studied and assigned, thus allowing for radio-astronomical detection.[47] To the best of our knowledge, no dedicated search for this molecule has been conducted toward any interstellar source.

**4.2.3 Methanediol ($OHCH_2OH$).** Methanediol is the simplest diol molecule and is of astrochemical interest due to its similarity to previously detected species such as methanol. Through modeling efforts, it has been proposed that this molecule is generated *via* grain surface reactions of the OH and $CH_2OH$ radicals.[40] For years, methanediol has been extensively studied in the aqueous phase (*e.g.* Möhlmann,[48] Matsuura *et al.*,[49] Ryabova *et al.*[50]). However, gaseous production and detection of this molecule was only reported in the past year.[51] To our knowledge, there have been no previous high-resolution microwave studies of this species. Thus, without experimental rotational parameters, definitive interstellar detection using radio astronomy is currently impossible.

**4.2.4 Methoxyethanol ($CH_3OCH_2CH_2OH$).** This molecules is in the same chemical family as methoxymethanol ($CH_3OCH_2OH$) and methoxyethane ($CH_3OCH_2CH_3$), which have large column densities of $1.4 \times 10^{17}$ cm$^{-2}$ and $1.8 \times 10^{16}$ cm$^{-2}$ toward IRAS 16293B, respectively.[52] The methoxy radical ($CH_3O$) has been detected in space and is proposed to form through methanol photodissociation, gas-phase reactions between the OH radical and methanol, and hydrogen addition to $H_2CO$.[53] Additionally, methoxymethanol has been shown to form *via* a reaction of the methoxy radical with $CH_2OH$.[40] The methoxy radical could feasibly react with other organic radicals to form the methoxylated versions of various additional organic species. Therefore, the methoxylated counterparts of the high abundance organics in this source could be interesting radio-astronomical targets. However, the microwave spectrum of methoxyethanol has only been experimentally collected and assigned up to 26.5 GHz.[54]

**4.2.5 Ethane ($C_2H_6$), carbon dioxide ($CO_2$), and carbon disulfide ($CS_2$).** For an allowed pure rotational spectrum to be collected, a molecule must have a nonzero dipole moment. Therefore, nonpolar molecules like ethane, carbon dioxide, and carbon disulfide will be undetectable through radio astronomy based on allowed pure rotational transitions regardless of their high predicted column densities. However, ethane was first detected toward the comet C/1996 B2 Hyakutake with high-resolution infrared spectroscopy.[55] Both solid and gaseous $CO_2$ have also been detected toward various interstellar sources using IR techniques.[56,57] $CS_2$ has not been previously detected in the interstellar medium. That said, experimental studies have shown that $CS_2$ can react with oxygen atoms on solid surfaces under astrophysically relevant conditions to form carbonyl sulfide (OCS), which is detected in high abundance toward IRAS 16293B[58,59]

**4.2.6 3-Hydroxypropanal ($OHCH_2CH_2CHO$).** Hydroxyacetone ($CH_3COCH_2OH$), a structural isomer of 3-hydroxypropanal, was detected with a column density of $1.2 \times 10^{16}$ cm$^{-2}$ toward IRAS 16293B.[60] Experiments by Wang *et al.*[61] showed that 3-hydroxypropanal can be formed in methanol–acetaldehyde ices irradiated with energetic electrons at 5 K. They concluded that this molecule can be produced in interstellar ices of star-forming regions that have high abundances of methanol and acetaldehyde (which is the case in IRAS 16293B). This molecule would then be desorbed into the gas phase as the protostar is heated. However, to our knowledge the rotational spectrum of 3-hydroxypropanal has not been experimentally measured.

### 4.3 Isotope ratios

Since isotopic composition was encoded in our molecular feature vectors, we proceeded to test the regressors' ability to predict isotopic ratios. As noted previously, if the machine learning model can accurately predict isotope ratios, information about the evolutionary history of the source and the molecular formation can be deciphered. That said, with such a simple encoding of the isotopic information in our feature vectors as well as the relatively small collection of isotopically substituted molecules, modeling this nuanced chemistry will be a challenge. Our discussion will solely focus on $^{13}$C and D substituted isotopologues because of the extremely small sample size of all other minor isotopes.

Fig. 4 displays the predicted column densities of the D and $^{13}$C substituted isotopologues along with the corresponding isotopic ratios. These predictions stem from 5-fold cross validation on the isotopically substituted data. In this process, the isotopologues are split into five subsets of training and validation data. In each iteration, 20% of the isotopically substituted molecules are left out of the training set. The model is then trained on all molecules in the dataset besides the 20% of rare isotopologues that were assigned to the validation set.

Because the deuterium and $^{13}$C ratios are reported in inverse fashions, the mean squared errors of the two ratio plots differ dramatically. The points within the shaded regions denote the molecules for which the prediction error is less than the mean absolute prediction error. For the D/H ratios, the mean absolute error is 0.032. For the $^{12}$C/$^{13}$C ratios, the mean absolute error is 20.9.

The column density predictions for isotopically substituted molecules are typically extremely accurate. However, when considering isotopic ratios, the range of realistic values is quite limited; therefore, a small prediction error is very notable. For example, deuterated acetaldehyde in IRAS 16293B is observed to have a D/H ratio of 7.98%. A column density under-prediction

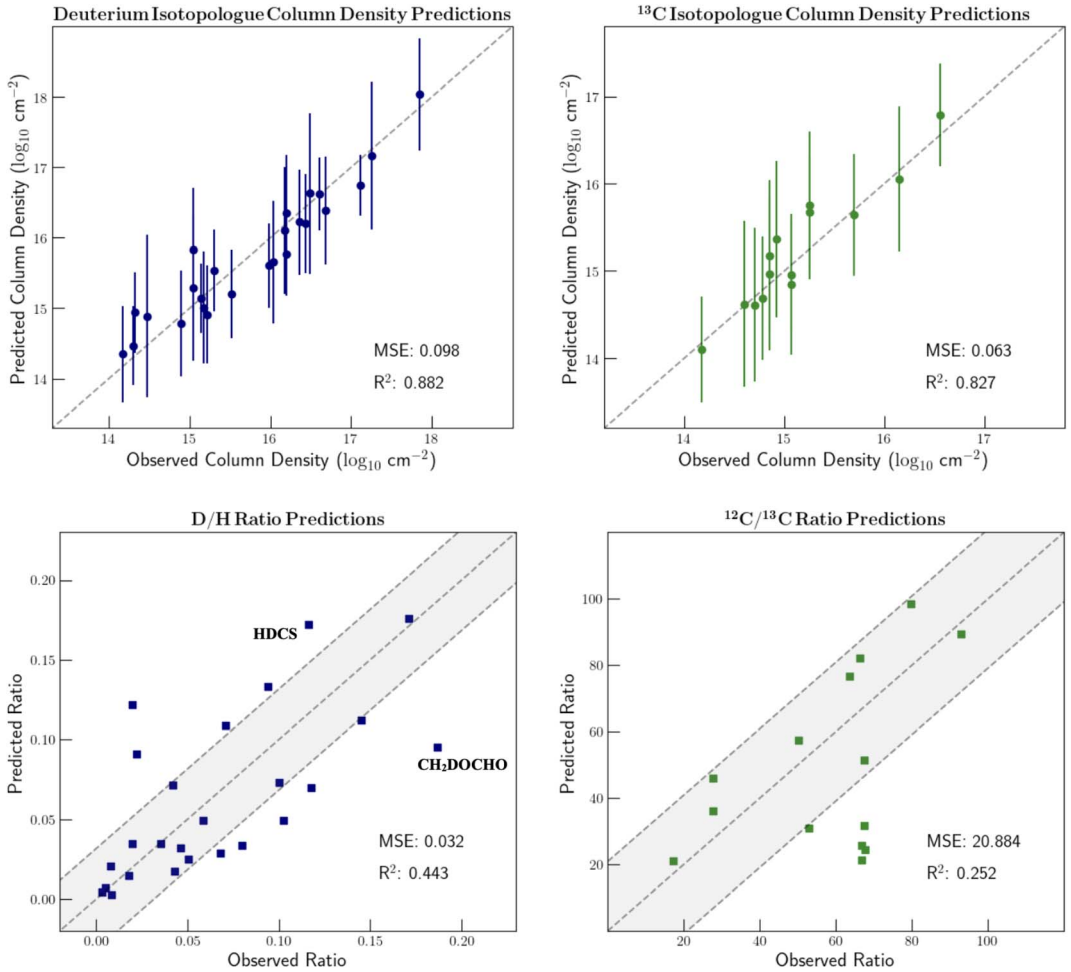


Fig. 4 GPR column density and isotope ratio predictions of deuterium and $^{13}$C substituted isotopologues using hand-picked isotope features. 1$\sigma$ column density uncertainties are shown. Some of the notable prediction errors are labelled and discussed in the text. Points that are within the grey shading denote molecules for which the ratio prediction error is less than the mean absolute error.

by 0.3 orders of magnitude would result in a predicted ratio of approximately 4.00%. This ratio suggests very different temperatures and timescales of formation.

There are a few molecules for which the ratio prediction is especially inaccurate. These species are labelled on Fig. 4. The prediction errors of HDCS and $CH_2DOCHO$ are especially notable since they highlight the shortcomings of using hand-picked descriptors. When encoding the chemical environment of the deuterium substitution, vector dimensions are included that denote whether there is a carbon or oxygen atom two bonds away from the deuterium atom. However, there was no consideration of whether the atom in this position is sulfur since HDCS is the only molecule in the dataset in which this is the case. Therefore, this important chemical environment information is not provided to the model, thus leading to a large prediction error. Additionally, with simple hand-picked features, the model isn't always able to capture the nuances of isotopic fractionation. For example, the isotopic encoding of $CH_2DOCHO$ is very similar to that of $CH_2DOH$. Because $CH_2DOH$ has a D/H ratio of only around 7%, the model inaccurately predicts $CH_2DOCHO$ to have approximately the same ratio.

Preferably we could include a more nuanced encoding of isotopic composition that better captures the local chemical environment instead of simple hand-picked features. However, with only 27 deuterated molecules and 15 $^{13}$C substituted species, the dataset of unique isotopologues is too small to learn the required relationships with a complex featurization. As mentioned previously, Mol2vec is sensitive to some, but not all, isotopic substitutions. It can, however, create unique vectors for deuterated species. Therefore, we tested the ability to learn deuterium ratios from the original Mol2vec-produced vectors that more fully consider chemical context. These results are shown in Fig. 5. The D/H ratio of formic acid is omitted from the graph since a ratio of around 6 is predicted which skews the ability to view the remaining points. As can be seen, these predictions are far less accurate than when hand-picked features were used. This is because the vector representations of many of the deuterated species are quite dissimilar to the main isotopologue in this case. In fact, the vector representation of $CH_2DCH_2OH$ is closer to that of propanal than that of $CH_3CH_2OH$. In order to include more detailed isotopic information in the feature vectors and thus accurately model

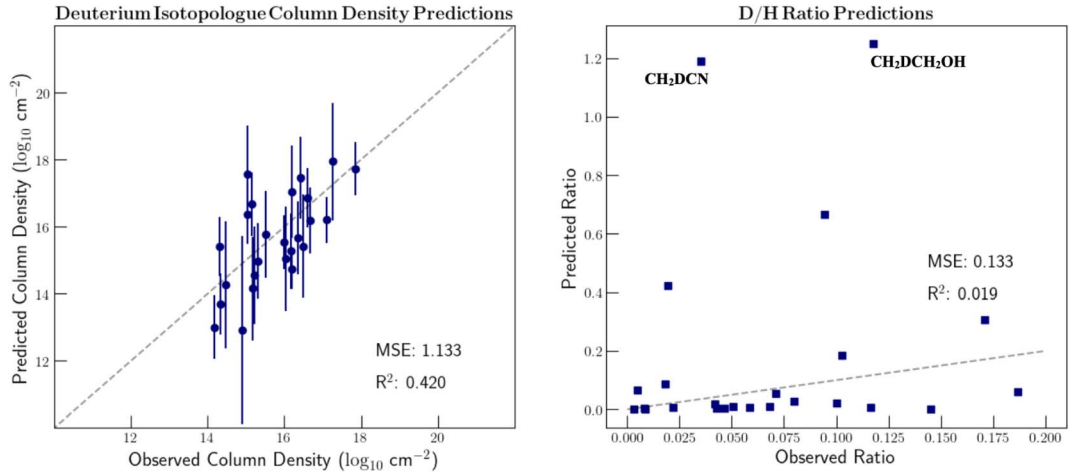


Fig. 5 GPR column density and isotope ratio predictions of deuterium substituted isotopologues. The inputted molecular feature vectors were generated with the Mol2vec algorithm and had no additional isotopic information included. 1$\sigma$ column density uncertainties are shown. The notable prediction errors are labelled and discussed in the text.

Table 2 Measured and predicted $^{12}C/^{13}C$ ratios for CO, CN, and $H_2CO$ toward IRAS 16293B along with the expected ratios that correspond to the $^{13}C$ galactocentric gradient

| Molecule | Galactocentric gradient $^{12}C/^{13}C$ ratio | Predicted $^{12}C/^{13}C$ ratio | Observed $^{12}C/^{13}C$ |
|---|---|---|---|
| CO | 46.04–79.05 | 50.85 | — |
| CN | 41.72–79.52 | 27.34 | — |
| $H_2CO$ | 53.90–104.46 | — | 52.92 |

isotopic fractionation, it is clear that we require additional isotopologue detections.

Finally, in order to test the predictive ability of the model on isotopologues for which no column density has been derived, we proceeded to predict the $^{12}C/^{13}C$ ratio of CO and CN with the trained GPR model from Section 4.1. For these two molecules (along with $H_2CO$), a linear $^{12}C/^{13}C$ trend has been defined as a function of galactocentric distance ($D_{GC}$). The formulae of these galactocentric gradients are displayed in eqn (1)–(3)

$$^{12}CO/^{13}CO = (5.41 \pm 1.07)\text{kpc}^{-1} \times D_{GC} + (19.03 \pm 7.90) \quad (1)$$

$$^{12}CN/^{13}CN = (6.01 \pm 1.19)\text{kpc}^{-1} \times D_{GC} + (12.28 \pm 9.33) \quad (2)$$

$$H_2{}^{12}CO/H_2{}^{13}CO = (7.60 \pm 1.79)\text{kpc}^{-1} \times D_{GC} + (18.05 \pm 10.88) \quad (3)$$

Using a galactocentric distance of 8.043 kpc for IRAS 16293, the range of expected ratios along with the ratios predicted by the GPR model are shown in Table 2. For reference, the observed $^{12}C/^{13}C$ ratio of $H_2CO$ is listed as well.[22] The Galactocentric distance of IRAS 16293 was computed using the ASTROPY Python module.[62] The values used in this calculation were the distances from the Earth to both the Galactic Center and IRAS 16293 (8.178 kpc and 141 pc (ref. 63 and 64)) as well as their respective sky coordinates.

The observed $^{12}C/^{13}C$ ratio of formaldehyde toward IRAS 16293B is very near the lower bound of the galactocentric trend

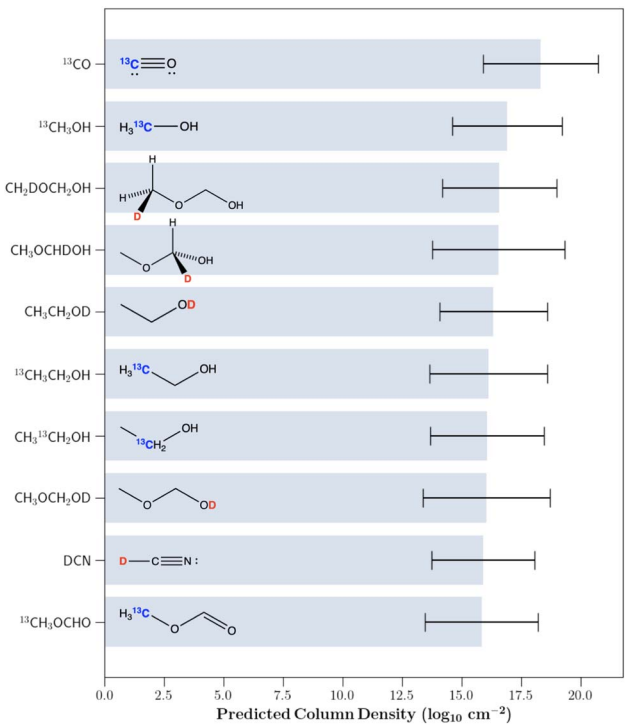


Fig. 6 The 10 undetected rare isotopologues of the molecules identified in IRAS 16293B with the highest predicted column densities. The trained Gaussian process regression model was used for the predictions. 1$\sigma$ column density uncertainties are shown.

Table 3 Tentative column densities of various ethanol isotopologues derived by Jørgensen *et al.*[19] corresponding to marginal detections along with the column densities predicted by the trained Gaussian process regression model

| Molecule | Tentative column density ($cm^{-2}$) | Predicted column density ($cm^{-2}$) |
|---|---|---|
| $CH_3CH_2OD$ | $1.1 \times 10^{16}$ | $2.15 \times 10^{16}$ |
| $^{13}CH_3CH_2OH$ | $9.1 \times 10^{15}$ | $1.33 \times 10^{16}$ |
| $CH_3{}^{13}CH_2OH$ | $9.1 \times 10^{15}$ | $1.18 \times 10^{16}$ |

Table 4 Analysis of the simulated spectra of the three deuterated isotopologues of methoxymethanol, $^{13}C$ substituted methyl formate, and deuterated methoxyethane. The spectral line catalogs were produced using SPCAT and the spectra were simulated with the MOLSIM Python module. The simulations span ALMA Band 7 (329.147–362.896 GHz). $1\sigma$ transitions have intensities of 7 mJy per beam and $3\sigma$ transitions have intensities of 21 mJy per beam

| Molecule | # of $3\sigma$ transitions | # of $1\sigma$ transitions | Intensity of strongest transition (mJy per beam) |
|---|---|---|---|
| $^{13}CH_3OCHO$ | 81 | 86 | 571.0 |
| $CH_3CH_2OCH_2D$ | 12 | 18 | 64.8 |
| $CH_2DOCH_2OH$ | 0 | 0 | 0.39 |
| $CH_3OCHDOH$ | 0 | 0 | 0.18 |
| $CH_3OCH_2OD$ | 0 | 0 | $3.92 \times 10^{-8}$ |

error bars at 8.043 kpc. Interestingly, the GPR model predicts the CO and CN ratios to also be fairly near the lower bounds of the respective ratio gradients. This matches the observed trends of various other molecules in IRAS 16293B, which typically show high levels of $^{13}C$ substitution. A high abundance of $^{13}CO$ could stem from favorable ion-neutral isotope exchange reactions in the cold interstellar gas before CO freeze-out.[65] Complex organic species that were then formed on grain surfaces from CO following freeze-out would inherit this small $^{12}C/^{13}C$ ratio. The $^{13}C$ enhancement in organic molecules would be even more notable at later timescales since laboratory experiments have shown that $^{12}CO$ desorbs slightly more efficiently than $^{13}CO$.[66] This enables $^{12}CO$ to sublimate from the grain at a lower temperature. Therefore, as the protostar begins to heat the surroundings to around the CO sublimation temperature, the more efficient $^{12}CO$ sublimation would result in grain surfaces that are further enhanced with $^{13}CO$.

Overall, while the machine learning regressor is not precise enough to adequately model the exact isotopic ratios, the $^{13}CO$ and $^{13}CN$ predictions show that it still is able to learn the general overabundance of $^{13}C$ in the organic species of IRAS 16293B.

### 4.4 Isotopologue targets for astrochemical study

As noted in Section 1, because of the extreme isotopic fractionation in this source, multiple groups have predicted that a significant portion of the unidentified spectral peaks in the line survey arise from isotopically substituted species.[16,17] Additionally, Fig. 4 shows that the GPR regressor is able to precisely model column densities of the isotopically substituted molecules. Therefore, using the trained GPR model we proceeded to predict the column densities of the D and $^{13}C$ monosubstituted isotopologues of the previously identified molecules in IRAS 16293B for which no accurate column density has been derived. Due to the inaccurate predictions of the main isotopologues of ethylene glycol, chloromethane and dimethyl ether, the predictions of the rare isotopologues of these species were omitted. The highest 10 predicted column density isotopologues are displayed in Fig. 6. All predictions are provided in the associated GitHub repository.

Of these highest 10 predicted column density species, both $^{13}C$-substituted ethanol isotopologues and OD substituted ethanol were marginally detected toward IRAS 16293B. While we did not include these molecules in our training set, Jørgensen *et al.*[19] derived tentative column densities for these

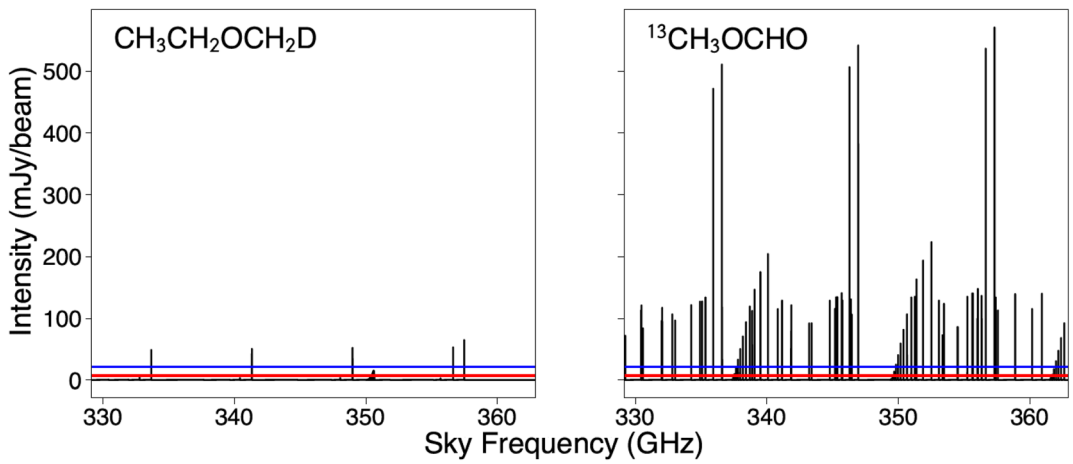


Fig. 7 Simulated spectra of $CH_3CH_2OCH_2D$ and $^{13}CH_3OCHO$ toward IRAS 16293B. The column densities used for the simulations are $3.98 \times 10^{15}$ $cm^{-2}$ for $CH_3CH_2OCH_2D$ and $6.80 \times 10^{15}$ $cm^{-2}$ for $^{13}CH_3OCHO$. These values were predicted by the trained Gaussian process regression model. The red line denotes the approximate RMS noise level of the PILS observations in ALMA Band 7. The blue line denotes the intensity required for a transition to have $3\sigma$ significance. Other simulation parameters are noted in the text.

species. These tentatively derived column densities along with the values predicted by the GPR model are listed in Table 3. As can be seen, all predictions closely match the tentative column densities. Additionally, while transitions corresponding to $^{13}CH_3OH$ and DCN have been identified toward IRAS 16293, no derived column density is listed.[67]

Therefore, the remaining molecules of interest are the three deuterated isotopologues of methoxymethanol along with the $^{13}C$ isotopologue of methyl formate. Beyond these largest 10 predicted column density isotopologues, another high predicted abundance isotopologue is the deuterated isotopologue of methoxyethane ($CH_3CH_2OCH_2D$). As mentioned previously, while large column densities are beneficial for detection, various other factors impact the detectability of a molecule. For all of these molecules, we therefore simulated their spectra toward IRAS 16293B using the predicted column densities in order to assess their true detectability. The microwave and submm spectra of $^{13}CH_3OCHO$ have been experimentally studied and assigned, thus making interstellar detection currently possible. In fact, this rare isotopologue was detected in the Orion molecular cloud.[68] However, this particular isotopologue is not present in the CDMS molecular spectroscopy database.[69] All of the other aforementioned deuterated isotopologues have not been studied experimentally.

In order to simulate the spectra of these previously unstudied isotopologues, the rotational constants must first be calculated. A low-cost method to obtain molecular rotational constants for isotopically substituted species can be achieved by combining experimental data and *ab initio* calculations. In this process, the experimental rotational and distortion constants of the main isotopologue are first collected. The *A*, *B*, and *C* rotational constants of the parent species are then calculated at a given level of theory and basis set. For our work, we ran the calculations with the PSI4 [70] Python package and used the M06-2X functional with the 6-311++G(d,p) basis set. Assuming that the geometry remains constant upon isotopic substitution, the same computational methods are then used to calculate rotational constants of the rare isotopologues. Finally, it is assumed that the scaling factor between the experimental and calculated rotational constants is the same for the main isotopologues and the isotopically substituted molecules. For example, the *B* rotational constant of the isotopically substituted species can be calculated using eqn (4).

$$B_{\text{scaled}} = \frac{B_{\text{exp(parent)}}}{B_{\text{calc.(parent)}}} \times B_{\text{calc.}} \quad (4)$$

The experimental distortion constants and dipole moments of the main isotopologues were used as-is for the isotopically substituted molecules.

Following the rotational constant calculations, a rotational line catalog was generated using Pickett's SPCAT.[71] Only the *A*, *B*, and *C* rotational constants and distortion constants (when available) were used. Internal rotational was not considered during the catalog simulations. The molsim[72] Python package was then used to simulate the spectra of the isotopologues toward IRAS 16293B with the predicted column densities. Molsim assumes that the molecular emission can be described by a single excitation temperature, and accounts for the effects of optical depth. For the simulations, the excitation temperature and $v_{\text{lsr}}$ of the main isotopologue were used. A source size of 0.5″, beam diameter of 0.5″, and line width of 1.0 km $s^{-1}$ were used for each simulation.

Since ALMA Band 7 (329.147–362.896 GHz) is fully covered with the PILS observations, this frequency range is the predominant focus of our spectral simulations. Given the noise level of the PILS observations,[16] any transition with a peak intensity stronger than ~21 mJy per beam should be detectable at a 3$\sigma$ significance. Analysis of the spectral simulations can be seen in Table 4. The simulated spectra of the $^{13}C$-substituted methyl formate isotopologue and the deuterated methoxyethane isotopologue toward IRAS 16293B are presented in Fig. 7.

Despite having higher predicted column densities, the rare isotopologues of methoxymethanol are predicted to have much weaker spectral peaks than the other isotopologues considered due to the limited dipole moment. That said, with several transitions predicted to be stronger than 3$\sigma$, $CH_3CH_2OCH_2D$ is an excellent candidate for experimental study for astrochemical purposes. Additionally, since the spectrum of $^{13}CH_3OCHO$ has already been collected and assigned, we recommend that this molecule be searched for in the PILS data.

# 5 Conclusions

In this work, we applied the machine learning pipeline introduced by Lee *et al.*[11] to source B of the Class 0 protostellar system IRAS 16293-2422. In order to include isotopologues in the dataset, we also concatenated a simple encoding of the isotopic composition to the feature vectors. Gaussian process regression and Bayesian ridge regression were both able to accurately model the column densities of the detected molecules in this source. The trained Gaussian process regression model then provided a list of 242 well-constrained targets for astrochemical study. Small, oxygenated, and fairly saturated hydrocarbons were predicted to be in high abundance in this protostellar source. While the column density predictions of isotopologues were quite precise, the nuances of isotopic ratios were only modeled with moderate accuracy. Additional isotopologue detections will be required to allow for a more complex encoding of isotopic substitution that better captures local chemical environment. Finally, since it has been predicted that many of the unassigned transitions in the PILS survey arise from isotopically substituted molecules, we provided a list of 92 isotopologue column density predictions.

This machine learning method has now been shown to effectively model the molecular column densities in two separate interstellar sources and the resulting trained regression models can be used to predict molecular species that are likely abundant in these various regions of interstellar space. However, these same techniques can be readily applied to terrestrial chemical mixture identifications as well. For example, if a researcher is able to reliably identify a fairly small number of chemical components present in an environmental sample along with their abundances, these supervised regressors could be trained and used to predict other components and contaminants of that mixture.

# 6 Appendix

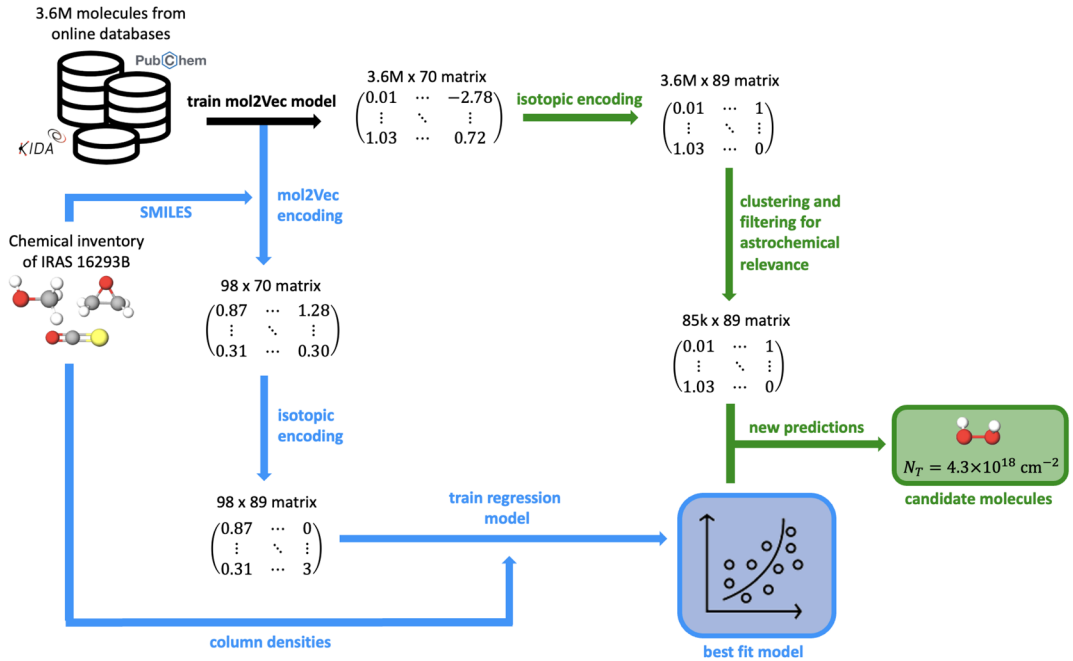


Fig. 8 Schematic that summarizes each of the steps of the machine learning method.

Table 5 Molecules detected at a one-beam or half-beam offset position from the continuum peak of IRAS 16293B in the south-west direction. The units of the observed column densities are $\log_{10}$ cm$^{-2}$

| Formula | SMILES | Observed column density | Reference |
|---|---|---|---|
| CO | [C−]#[O+] | 20.0000 | Drozdovskaya *et al.*[25] |
| $CH_3OH$ | CO | 19.0000 | Jørgensen *et al.*[16,19] |
| $H_2CO$ | C=O | 18.2800 | Persson *et al.*[22] |
| $CH_3CH_2OH$ | CCO | 17.3600 | Jørgensen *et al.*[19] |
| $CH_3OCH_3$ | COC | 17.3800 | Jørgensen *et al.*[19] |
| $HCOOCH_3$ | COC=O | 17.4100 | Jørgensen *et al.*[19] |
| $CH_2OHCHO$ | O=CCO | 16.5100 | Jørgensen *et al.*[16] |
| $CH_3COOH$ | CC(=O)O | 15.4500 | Jørgensen *et al.*[16] |
| $CH_3CHO$ | CC=O | 17.0800 | Jørgensen *et al.*[19] |
| c-$C_2H_4O$ | C1CO1 | 15.7300 | Lykke *et al.*[73] |
| HCOOH | O=CO | 16.7500 | Jørgensen *et al.*[19] |
| aGg′-$(CH_2OH)_2$ | OCCO | 16.7200 | Jørgensen *et al.*[16] |
| $CH_3OCH_2OH$ | COCO | 17.1500 | Manigand *et al.*[52] |
| $C_2H_5CHO$ | CCC=O | 15.3400 | Lykke *et al.*[73] |
| $(CH_3)_2CO$ | CC(C)=O | 16.2300 | Lykke *et al.*[73] |
| $NH_2CHO$ | NC=O | 15.9800 | Coutens *et al.*[20] |
| HCN | C#N | 16.7000 | Drozdovskaya *et al.*[25] |
| $CH_3CN$ | CC#N | 16.6000 | Calcutt *et al.*[74] |
| $CH_3NC$ | [C−]#[N+]C | 14.3000 | Calcutt *et al.*[75] |
| HNCO | N=C=O | 16.5700 | Ligterink *et al.*[76] |
| $HC_3N$ | C#CC#N | 14.2600 | Calcutt *et al.*[74] |
| $H_2S$ | S | 17.2300 | Drozdovskaya *et al.*[59] |
| OCS | O=C=S | 17.4000 | Drozdovskaya *et al.*[59] |
| $CH_3SH$ | CS | 15.6800 | Drozdovskaya *et al.*[59] |
| CS | [C−]#[S+] | 15.5900 | Drozdovskaya *et al.*[59] |
| $H_2CS$ | C=S | 15.1100 | Drozdovskaya *et al.*[59] |
| SO | O=S | 14.6400 | Drozdovskaya *et al.*[59] |

Table 5 (*Contd.*)

| Formula | SMILES | Observed column density | Reference |
|---|---|---|---|
| $CH_3Cl$ | CCl | 14.6600 | Fayolle *et al.*[77] |
| $C_2H_3CHO$ | C=CC=O | 14.5300 | Manigand *et al.*[78] |
| $C_3H_6$ | C=CC | 16.6200 | Manigand *et al.*[78] |
| $CH_3CCH$ | C#CC | 16.0400 | Calcutt *et al.*[79] |
| t-$C_2H_5OCH_3$ | CCOC | 16.2600 | Manigand *et al.*[52] |
| $C_3H_4O_2$ | O=CC=CO | 15.0000 | Coutens *et al.*[80] |
| $CH_3NCO$ | CN=C=O | 15.6000 | Ligterink *et al.*[76] |
| $C_2H_5CN$ | CCC#N | 15.5600 | Calcutt *et al.*[74] |
| $C_2H_3CN$ | C=CC#N | 14.8700 | Calcutt *et al.*[74] |
| $CH_2CO$ | C=C=O | 16.6800 | Jørgensen *et al.*[19] |
| HONO | O=NO | 14.9500 | Coutens *et al.*[81] |
| NO | [N]=O | 16.3000 | Ligterink *et al.*[82] |
| $CH_3C(O)NH_2$ | CC(N)=O | 14.9500 | Ligterink *et al.*[82] |
| $SO_2$ | O=S=O | 15.1100 | Drozdovskaya *et al.*[59] |
| t-HCOOH | O=CO | 16.7500 | Jørgensen *et al.*[19] |
| $CH_2NH$ | C=N | 14.9031 | Ligterink *et al.*[82] |
| $H_2{}^{13}CO$ | [13CH2]=O | 16.5563 | Persson *et al.*[22] |
| $H_2C^{17}O$ | C=[17O] | 14.8573 | Persson *et al.*[22] |
| $H_2C^{18}O$ | C=[18O] | 15.3617 | Persson *et al.*[22] |
| HDCO | [2H]C=O | 17.1139 | Persson *et al.*[22] |
| $D_2CO$ | [2H]C([2H])=O | 16.2041 | Persson *et al.*[22] |
| $D_2{}^{13}CO$ | [2H][13C]([2H])=O | 14.3424 | Persson *et al.*[22] |
| $HC^{15}N$ | C#[15N] | 14.3979 | Drozdovskaya *et al.*[25] |
| $^{13}CH_3CN$ | [13CH3]C#N | 14.7782 | Calcutt *et al.*[74] |
| $CH_3{}^{13}CN$ | C[13C]#N | 14.6990 | Calcutt *et al.*[74] |
| $CH_3C^{15}N$ | CC#[15N] | 14.2041 | Calcutt *et al.*[74] |
| $CH_2DCN$ | [2H]CC#N | 15.1461 | Calcutt *et al.*[74] |
| $CHD_2CN$ | [2H]C([2H])C#N | 14.3010 | Calcutt *et al.*[74] |
| $^{34}SO_2$ | O=[34S=O] | 14.6021 | Drozdovskaya *et al.*[59] |
| $O^{13}CS$ | O=[13C=S] | 15.6990 | Drozdovskaya *et al.*[59] |
| $OC^{34}S$ | O=C=[34S] | 16.0000 | Drozdovskaya *et al.*[59] |
| $OC^{33}S$ | O=C=[33S] | 15.4771 | Drozdovskaya *et al.*[59] |
| $^{18}OCS$ | [18O]=C=S | 14.6990 | Drozdovskaya *et al.*[59] |
| $C^{34}S$ | [C−]#[34S+] | 14.3010 | Drozdovskaya *et al.*[59] |
| $C^{33}S$ | [C−]#[33S+] | 13.9031 | Drozdovskaya *et al.*[59] |
| $C^{36}S$ | [C−]#[36S+] | 13.1461 | Drozdovskaya *et al.*[59] |
| HDCS | [2H]C=S | 14.1761 | Drozdovskaya *et al.*[59] |
| HDS | [2H]S | 16.2041 | Drozdovskaya *et al.*[59] |
| $HD^{34}S$ | [2H][34SH] | 15.0000 | Drozdovskaya *et al.*[59] |
| $CD_3OH$ | [2H]C([2H])([2H)O | 16.4914 | Ilyushin *et al.*[21] |
| $CH_2DOH$ | [2H]CO | 17.8513 | Jørgensen *et al.*[19] |
| $CH_3OD$ | [2H]OC | 17.2553 | Jørgensen *et al.*[19] |
| a-$CH_3CHDOH$ | [2H]C(C)O | 16.3617 | Jørgensen *et al.*[19] |
| $CH_3OCDO$ | [2H]C(=O)OC | 16.1761 | Jørgensen *et al.*[19] |
| $CH_2DOCHO$ | [2H]COC=O | 16.6812 | Jørgensen *et al.*[19] |
| CHDCO | [2H]C=C=O | 15.3010 | Jørgensen *et al.*[19] |
| $^{13}CH_3OCH_3$ | CO[13CH3] | 16.1461 | Jørgensen *et al.*[19] |
| $CH_3CDO$ | [2H]C(C)=O | 15.9823 | Jørgensen *et al.*[19] |
| $H^{13}COOH$ | O=[13CH]O | 14.9191 | Jørgensen *et al.*[19] |
| $CHD_2OCHO$ | [2H]C([2H)OC=O | 16.0414 | Manigand *et al.*[83] |
| $CH_3{}^{37}Cl$ | C[37Cl] | 14.3424 | Fayolle *et al.*[77] |
| $NH_2CDO$ | [2H]C(N)=O | 14.3222 | Coutens *et al.*[20] |
| $NH_2{}^{13}CHO$ | N[13CH]=O | 14.1761 | Coutens *et al.*[20] |
| DNCO | [2H]N=C=O | 14.4771 | Coutens *et al.*[20] |
| $HN^{13}CO$ | N=[13C]=O | 14.6021 | Coutens *et al.*[20] |
| CHDOHCHO | [2H]C(O)C=O | 15.5211 | Jørgensen *et al.*[16] |
| $CH_2ODCHO$ | [2H]OCC=O | 15.1761 | Jørgensen *et al.*[16] |
| $CH_2OHCDO$ | [2H]C(=O)CO | 15.2148 | Jørgensen *et al.*[16] |
| $CH_3{}^{18}OH$ | C[18OH] | 16.2718 | Jørgensen *et al.*[16] |
| $^{13}CH_2CO$ | [13CH2]=C=O | 14.8513 | Jørgensen *et al.*[19] |
| $CH_2{}^{13}CO$ | C=[13C]=O | 14.8513 | Jørgensen *et al.*[19] |
| $13CH_3CHO$ | [13CH3]C=O | 15.2553 | Jørgensen *et al.*[19] |
| $CH_3{}^{13}CHO$ | C[13CH]=O | 15.2553 | Jørgensen *et al.*[19] |

Table 5 (*Contd.*)

| Formula | SMILES | Observed column density | Reference |
|---|---|---|---|
| t-DCOOH | [2H]C(=O)O | 15.0414 | Jørgensen *et al.*[19] |
| t-HCOOD | [2H]OC=O | 15.0414 | Jørgensen *et al.*[19] |
| a–a-$CH_2DCH_2OH$ | [2H]CCO | 16.4313 | Jørgensen *et al.*[19] |
| *Asym*-$CH_2DOCH_3$ | [2H]COC | 16.6128 | Jørgensen *et al.*[19] |



## Data availability

The code, analysis scripts, and datasets supporting this article are available on GitHub at **https://github.com/zfried/IRAS_ML_Predictions**.

## Author contributions

Zachary Fried: software, formal analysis, investigation, data curation, writing – original draft, visualization. Kelvin Lee: methodology, supervision, writing – review & editing. Alex Byrne: software, writing – review & editing. Brett McGuire: supervision, writing – review & editing, funding acquisition.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

The authors thank C. W. Coley for providing helpful insight regarding the Mol2vec molecular featurization and the inclusion of isotopic information in Morgan fingerprints. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.

## Notes and references

1 E. F. van Dishoeck and G. A. Blake, *Annu. Rev. Astron. Astrophys.*, 1998, **36**, 317–368.
2 K. I. Öberg, R. Murray-Clay and E. A. Bergin, *Astrophys. J.*, 2011, **743**, L16.
3 R. Bachiller, M. Pérez Gutiérrez, M. S. N. Kumar and M. Tafalla, *Astron. Astrophys.*, 2001, **372**, 899–912.
4 P. Schilke, C. M. Walmsley, G. Pineau des Forets and D. R. Flower, *Astron. Astrophys.*, 1997, **321**, 293–304.
5 A. Ginsburg, B. McGuire, R. Plambeck, J. Bally, C. Goddi and M. Wright, *Astrophys. J.*, 2019, **872**, 54.
6 P. T. P. Ho, A. H. Barrett, P. C. Myers, D. N. Matsakis, A. L. Cheung, M. F. Chui, C. H. Townes and K. S. Yngvesson, *Astrophys. J.*, 1979, **234**, 912–921.
7 M. Ruaud, V. Wakelam and F. Hersant, *Mon. Not. R. Astron. Soc.*, 2016, **459**, 3756–3767.
8 V. Wakelam, J.-C. Loison, E. Herbst, B. Pavone, A. Bergeat, K. Béroff, M. Chabot, A. Faure, D. Galli, W. D. Geppert, D. Gerlich, P. Gratier, N. Harada, K. M. Hickson, P. Honvault, S. J. Klippenstein, S. D. L. Picard, G. Nyman, M. Ruaud, S. Schlemmer, I. R. Sims, D. Talbi, J. Tennyson and R. Wester, *Astrophys. J., Suppl. Ser.*, 2015, **217**, 20.
9 B. A. McGuire, P. B. Carroll, N. M. Dollhopf, N. R. Crockett, J. F. Corby, R. A. Loomis, A. M. Burkhardt, C. Shingledecker, G. A. Blake and A. J. Remijan, *Astrophys. J.*, 2015, **812**, 76.
10 E. Hébrard, M. Dobrijevic, P. Pernot, N. Carrasco, A. Bergeat, K. M. Hickson, A. Canosa, S. D. Le Picard and I. R. Sims, *J. Phys. Chem. A*, 2009, **113**, 11227–11237.
11 K. L. K. Lee, J. Patterson, A. M. Burkhardt, V. Vankayalapati, M. C. McCarthy and B. A. McGuire, *Astrophys. J., Lett.*, 2021, **917**, L6.
12 A. Wootten, *Astrophys. J.*, 1989, **337**, 858.
13 L. G. Mundy, A. Wootten, B. A. Wilking, G. A. Blake and A. I. Sargent, *Astrophys. J.*, 1992, **385**, 306.
14 L. W. Looney, L. G. Mundy and W. J. Welch, *Astrophys. J.*, 2000, **529**, 477–498.
15 M. J. Maureira, J. E. Pineda, D. M. Segura-Cox, P. Caselli, L. Testi, G. Lodato, L. Loinard and A. Hernández-Gómez, *Astrophys. J.*, 2020, **897**, 59.
16 J. K. Jørgensen, M. H. D. v. d. Wiel, A. Coutens, J. M. Lykke, H. S. P. Müller, E. F. v. Dishoeck, H. Calcutt, P. Bjerkeli, T. L. Bourke, M. N. Drozdovskaya, C. Favre, E. C. Fayolle, R. T. Garrod, S. K. Jacobsen, K. I. Öberg, M. V. Persson and S. F. Wampfler, *Astron. Astrophys.*, 2016, **595**, A117.
17 V. Taquet, E. F. v. Dishoeck, M. Swayne, D. Harsono, J. K. Jørgensen, L. Maud, N. F. W. Ligterink, H. S. P. Müller, C. Codella, K. Altwegg, A. Bieler, A. Coutens, M. N. Drozdovskaya, K. Furuya, M. V. Persson, M. L. R. v. Hoff, C. Walsh and S. F. Wampfler, *Astron. Astrophys.*, 2018, **618**, A11.
18 A. M. Burkhardt, E. Herbst, S. V. Kalenskii, M. C. McCarthy, A. J. Remijan and B. A. McGuire, *Mon. Not. R. Astron. Soc.*, 2018, **474**, 5068–5075.
19 J. K. Jørgensen, H. S. P. Müller, H. Calcutt, A. Coutens, M. N. Drozdovskaya, K. I. Öberg, M. V. Persson, V. Taquet, E. F. v. Dishoeck and S. F. Wampfler, *Astron. Astrophys.*, 2018, **620**, A170.
20 A. Coutens, J. K. Jørgensen, M. H. D. v. d. Wiel, H. S. P. Müller, J. M. Lykke, P. Bjerkeli, T. L. Bourke, H. Calcutt, M. N. Drozdovskaya, C. Favre, E. C. Fayolle, R. T. Garrod, S. K. Jacobsen, N. F. W. Ligterink, K. I. Öberg, M. V. Persson, E. F. v. Dishoeck and S. F. Wampfler, *Astron. Astrophys.*, 2016, **590**, L6.

21 V. V. Ilyushin, H. S. P. Müller, J. K. Jørgensen, S. Bauerecker, C. Maul, Y. Bakhmat, E. A. Alekseev, O. Dorovskaya, S. Vlasenko, F. Lewen, S. Schlemmer, K. Berezkin and R. M. Lees, *Astron. Astrophys.*, 2022, **658**, A127.
22 M. V. Persson, J. K. Jørgensen, H. S. P. Müller, A. Coutens, E. F. v. Dishoeck, V. Taquet, H. Calcutt, M. H. D. v. d. Wiel, T. L. Bourke and S. F. Wampfler, *Astron. Astrophys.*, 2018, **610**, A54.
23 W. D. Watson, *Rev. Mod. Phys.*, 1976, **48**, 513–552.
24 T. J. Millar, A. Bennett and E. Herbst, *Astrophys. J.*, 1989, **340**, 906.
25 M. N. Drozdovskaya, E. F. van Dishoeck, M. Rubin, J. K. Jørgensen and K. Altwegg, *Mon. Not. R. Astron. Soc.*, 2019, **490**, 50–79.
26 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **58**, 27–35.
27 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2021, **49**, D1388–D1395.
28 T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.
29 C. Boersma, C. W. Bauschlicher, A. Ricca, A. L. Mattioda, J. Cami, E. Peeters, F. S. de Armas, G. P. Saborido, D. M. Hudgins and L. J. Allamandola, *Astrophys. J., Suppl. Ser.*, 2014, **211**, 8.
30 C. W. Bauschlicher, A. Ricca, C. Boersma and L. J. Allamandola, *Astrophys. J., Suppl. Ser.*, 2018, **234**, 32.
31 A. L. Mattioda, D. M. Hudgins, C. Boersma, C. W. Bauschlicher, A. Ricca, J. Cami, E. Peeters, F. S. de Armas, G. P. Saborido and L. J. Allamandola, *Astrophys. J., Suppl. Ser.*, 2020, **251**, 22.
32 *RDKit: Open-Source Cheminformatics*, **http://www.rdkit.org**.
33 H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–113.
34 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, 2005.
35 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
36 B. A. McGuire, *Astrophys. J., Suppl. Ser.*, 2022, **259**, 30.
37 H. Linnartz, S. Ioppolo and G. Fedoseev, *Int. Rev. Phys. Chem.*, 2015, **34**, 205–237.
38 G. Fedoseev, H. M. Cuppen, S. Ioppolo, T. Lamberts and H. Linnartz, *Mon. Not. R. Astron. Soc.*, 2015, **448**, 1288–1297.
39 D. E. Woon, *Astrophys. J.*, 2002, **569**, 541.
40 R. T. Garrod, S. L. Widicus Weaver and E. Herbst, *Astrophys. J.*, 2008, **682**, 283–302.
41 P. Bergman, B. Parise, R. Liseau, B. Larsson, H. Olofsson, K. M. Menten and R. Güsten, *Astron. Astrophys.*, 2011, **531**, L8.
42 F. Du, B. Parise and P. Bergman, *Astron. Astrophys.*, 2012, **538**, A91.
43 B. Parise, P. Bergman and F. Du, *Astron. Astrophys.*, 2012, **541**, L11.
44 M. A. Thelen, P. Felder and J. Robert Huber, *Chem. Phys. Lett.*, 1993, **213**, 275–281.
45 D. McNally, *Sci. Prog.*, 1965, **53**, 83–87.
46 J. Cernicharo, N. Marcelino, E. Roueff, M. Gerin, A. Jiménez-Escobar and G. M. M. Caro, *Astrophys. J., Lett.*, 2012, **759**, L43.
47 M. Tyblewski, T. Ha, R. Meyer, A. Bauder and C. E. Blom, *J. Chem. Phys.*, 1992, **97**, 6168–6180.
48 G. R. Möhlmann, *J. Raman Spectrosc.*, 1987, **18**, 199–203.
49 H. Matsuura, M. Yamamoto and H. Murata, *Spectrochim. Acta, Part A*, 1980, **36**, 321–327.
50 R. S. Ryabova, G. I. Voloshenko, V. D. Maiorov and G. F. Osipova, *Russ. J. Appl. Chem.*, 2002, **75**, 22–24.
51 C. Zhu, N. F. Kleimeier, A. M. Turner, S. K. Singh, R. C. Fortenberry and R. I. Kaiser, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2111938119.
52 S. Manigand, J. K. Jørgensen, H. Calcutt, H. S. P. Müller, N. F. W. Ligterink, A. Coutens, M. N. Drozdovskaya, E. F. v. Dishoeck and S. F. Wampfler, *Astron. Astrophys.*, 2020, **635**, A48.
53 J. Cernicharo, N. Marcelino, E. Roueff, M. Gerin, A. Jiménez-Escobar and G. M. M. Caro, *Astrophys. J., Lett.*, 2012, **759**, L43.
54 P. Buckley and M. Brochu, *Can. J. Chem.*, 1972, **50**, 1149–1156.
55 M. J. Mumma, M. A. DiSanti, N. D. Russo, M. Fomenkova, K. Magee-Sauer, C. D. Kaminski and D. X. Xie, *Science*, 1996, **272**, 1310–1314.
56 L. B. D'Hendecourt and M. Jourdain de Muizon, *Astron. Astrophys.*, 1989, **223**, L5–L8.
57 E. F. van Dishoeck, F. P. Helmich, T. de Graauw, J. H. Black, A. C. A. Boogert, P. Ehrenfreund, P. A. Gerakines, J. H. Lacy, T. J. Millar, W. A. Schutte, A. G. G. M. Tielens, D. C. B. Whittet, D. R. Boxhoorn, D. J. M. Kester, K. Leech, P. R. Roelfsema, A. Salama and B. Vandenbussche, *Astron. Astrophys.*, 1996, **315**, L349–L352.
58 M. D. Ward, I. A. Hogg and S. D. Price, *Mon. Not. R. Astron. Soc.*, 2012, **425**, 1264–1269.
59 M. N. Drozdovskaya, E. F. van Dishoeck, J. K. Jørgensen, U. Calmonte, M. H. D. van der Wiel, A. Coutens, H. Calcutt, H. S. P. Müller, P. Bjerkeli, M. V. Persson, S. F. Wampfler and K. Altwegg, *Mon. Not. R. Astron. Soc.*, 2018, **476**, 4949–4964.
60 Y. Zhou, D.-H. Quan, X. Zhang and S.-L. Qin, *Res. Astron. Astrophys.*, 2020, **20**, 125.
61 J. Wang, J. H. Marks, A. M. Turner, A. A. Nikolayev, V. Azyazov, A. M. Mebel and R. I. Kaiser, *Phys. Chem. Chem. Phys.*, 2023, **25**, 936–953.
62 The Astropy Collaboration, *Astrophys. J.*, 2022, **935**, 167.
63 R. Abuter, A. Amorim, M. Bauböck, J. P. Berger, H. Bonnet, W. Brandner, Y. Clénet, V. C. d. Foresto, P. T. d. Zeeuw, J. Dexter, G. Duvert, A. Eckart, F. Eisenhauer, N. M. F. Schreiber, P. Garcia, F. Gao, E. Gendron, R. Genzel, O. Gerhard, S. Gillessen, M. Habibi, X. Haubois, T. Henning, S. Hippler, M. Horrobin, A. Jiménez-Rosales, L. Jocou, P. Kervella, S. Lacour, V. Lapeyrère, J.-B. L. Bouquin, P. Léna, T. Ott, T. Paumard, K. Perraut,

G. Perrin, O. Pfuhl, S. Rabien, G. R. Coira, G. Rousset, S. Scheithauer, A. Sternberg, O. Straub, C. Straubmeier, E. Sturm, L. J. Tacconi, F. Vincent, S. v. Fellenberg, I. Waisberg, F. Widmann, E. Wieprecht, E. Wiezorrek, J. Woillez and S. Yazici, *Astron. Astrophys.*, 2019, **625**, L10.

64 S. A. Dzib, G. N. Ortiz-León, A. Hernández-Gómez, L. Loinard, A. J. Mioduszewski, M. Claussen, K. M. Menten, E. Caux and A. Sanna, *Astron. Astrophys.*, 2018, **614**, A20.

65 W. D. Langer, T. E. Graedel, M. A. Frerking and P. B. Armentrout, *Astrophys. J.*, 1984, **277**, 581–604.

66 L. R. Smith, M. S. Gudipati, R. L. Smith and R. D. Lewis, *Astron. Astrophys.*, 2021, **656**, A82.

67 J. K. Jørgensen, T. L. Bourke, Q. N. Luong and S. Takakuwa, *Astron. Astrophys.*, 2011, **534**, A100.

68 M. Carvajal, L. Margulès, B. Tercero, K. Demyk, I. Kleiner, J. C. Guillemin, V. Lattanzi, A. Walters, J. Demaison, G. Wlodarczak, T. R. Huet, H. Møllendal, V. V. Ilyushin and J. Cernicharo, *Astron. Astrophys.*, 2009, **500**, 1109–1118.

69 H. S. P. Müller, F. Schlöder, J. Stutzki and G. Winnewisser, *J. Mol. Struct.*, 2005, **742**, 215–227.

70 D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford and C. D. Sherrill, *J. Chem. Phys.*, 2020, **152**, 184108.

71 H. M. Pickett, *J. Mol. Spectrosc.*, 1991, **148**, 371–377.

72 B. A. McGuire and K. Lee, *molsim*, 2020, **https://zenodo.org/record/4122749**.

73 J. M. Lykke, A. Coutens, J. K. Jørgensen, M. H. D. v. d. Wiel, R. T. Garrod, H. S. P. Müller, P. Bjerkeli, T. L. Bourke, H. Calcutt, M. N. Drozdovskaya, C. Favre, E. C. Fayolle, S. K. Jacobsen, K. I. Öberg, M. V. Persson, E. F. v. Dishoeck and S. F. Wampfler, *Astron. Astrophys.*, 2017, **597**, A53.

74 H. Calcutt, J. K. Jørgensen, H. S. P. Müller, L. E. Kristensen, A. Coutens, T. L. Bourke, R. T. Garrod, M. V. Persson, M. H. D. v. d. Wiel, E. F. v. Dishoeck and S. F. Wampfler, *Astron. Astrophys.*, 2018, **616**, A90.

75 H. Calcutt, M. R. Fiechter, E. R. Willis, H. S. P. Müller, R. T. Garrod, J. K. Jørgensen, S. F. Wampfler, T. L. Bourke, A. Coutens, M. N. Drozdovskaya, N. F. W. Ligterink and L. E. Kristensen, *Astron. Astrophys.*, 2018, **617**, A95.

76 N. F. W. Ligterink, A. Coutens, V. Kofman, H. S. P. Müller, R. T. Garrod, H. Calcutt, S. F. Wampfler, J. K. Jørgensen, H. Linnartz and E. F. van Dishoeck, *Mon. Not. R. Astron. Soc.*, 2017, **469**, 2219–2229.

77 E. C. Fayolle, K. I. Öberg, J. K. Jørgensen, K. Altwegg, H. Calcutt, H. S. P. Müller, M. Rubin, M. H. D. van der Wiel, P. Bjerkeli, T. L. Bourke, A. Coutens, E. F. van Dishoeck, M. N. Drozdovskaya, R. T. Garrod, N. F. W. Ligterink, M. V. Persson and S. F. Wampfler, *Nat. Astron.*, 2017, **1**, 703–708.

78 S. Manigand, A. Coutens, J.-C. Loison, V. Wakelam, H. Calcutt, H. S. P. Müller, J. K. Jørgensen, V. Taquet, S. F. Wampfler, T. L. Bourke, B. M. Kulterer, E. F. v. Dishoeck, M. N. Drozdovskaya and N. F. W. Ligterink, *Astron. Astrophys.*, 2021, **645**, A53.

79 H. Calcutt, E. R. Willis, J. K. Jørgensen, P. Bjerkeli, N. F. W. Ligterink, A. Coutens, H. S. P. Müller, R. T. Garrod, S. F. Wampfler and M. N. Drozdovskaya, *Astron. Astrophys.*, 2019, **631**, A137.

80 A. Coutens, J.-C. Loison, A. Boulanger, E. Caux, H. S. P. Müller, V. Wakelam, S. Manigand and J. K. Jørgensen, *Astron. Astrophys.*, 2022, **660**, L6.

81 A. Coutens, N. F. W. Ligterink, J.-C. Loison, V. Wakelam, H. Calcutt, M. N. Drozdovskaya, J. K. Jørgensen, H. S. P. Müller, E. F. v. Dishoeck and S. F. Wampfler, *Astron. Astrophys.*, 2019, **623**, L13.

82 N. F. W. Ligterink, H. Calcutt, A. Coutens, L. E. Kristensen, T. L. Bourke, M. N. Drozdovskaya, H. S. P. Müller, S. F. Wampfler, M. H. D. v. d. Wiel, E. F. v. Dishoeck and J. K. Jørgensen, *Astron. Astrophys.*, 2018, **619**, A28.

83 S. Manigand, H. Calcutt, J. K. Jørgensen, V. Taquet, H. S. P. Müller, A. Coutens, S. F. Wampfler, N. F. W. Ligterink, M. N. Drozdovskaya, L. E. Kristensen, M. H. D. v. d. Wiel and T. L. Bourke, *Astron. Astrophys.*, 2019, **623**, A69.