The Astrophysical Journal Letters, 917:L6 (15pp), 2021 August 10 https://doi.org/10.3847/2041-8213/ac194b

Machine Learning of Interstellar Chemical Inventories

Kin Long Kelvin Lee[1,2], Jacqueline Patterson[2,3], Andrew M. Burkhardt[2], Vivek Vankayalapati[2,4], Michael C. McCarthy[2], and Brett A. McGuire[1,2,5]
[1] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; kelvlee@mit.edu, brettmc@mit.edu
[2] Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA 02138, USA
[3] Indiana University, Bloomington, IN 47405, USA
[4] The University of Utah, Salt Lake City, UT 84112, USA
[5] National Radio Astronomy Observatory, Charlottesville, VA 22903, USA
Received 2021 April 30; revised 2021 July 28; accepted 2021 July 29; published 2021 August 12

Abstract

The characterization of interstellar chemical inventories provides valuable insight into the chemical and physical processes in astrophysical sources. The discovery of new interstellar molecules becomes increasingly difficult as the number of viable species grows combinatorially, even when considering only the most thermodynamically stable. In this work, we present a novel approach for understanding and modeling interstellar chemical inventories by combining methodologies from cheminformatics and machine learning. Using multidimensional vector representations of molecules obtained through unsupervised machine learning, we show that identification of candidates for astrochemical study can be achieved through quantitative measures of chemical similarity in this vector space, highlighting molecules that are most similar to those already known in the interstellar medium. Furthermore, we show that simple, supervised learning regressors are capable of reproducing the abundances of entire chemical inventories, and predict the abundance of not-yet-seen molecules. As a proof-of-concept, we have developed and applied this discovery pipeline to the chemical inventory of a well-known dark molecular cloud, the Taurus Molecular Cloud 1, one of the most chemically rich regions of space known to date. In this paper, we discuss the implications and new insights machine learning explorations of chemical space can provide in astrochemistry.

Unified Astronomy Thesaurus concepts: Astrochemistry (75); Chemical abundances (224); Interdisciplinary astronomy (804)

1. Introduction

In the interstellar medium, molecules act as sensitive probes of their local environment. Their relative abundances can be used to infer myriad physical properties of the target system ranging from the thermal history of the source (Lis et al. 2002) to the kinematic structure of the gas (Pinte et al. 2018; Disk Dynamics Collaboration et al. 2020), the passage of recent hydrodynamic shock (Schilke et al. 1997), or the presence of various radiation fields (Cleeves et al. 2017). The chemical inventories—and abundances—are also deeply tied to the environment itself, from carbon- and silicon-rich evolved stars (Gong et al. 2015) to organic-rich star-forming cores (Belloche et al. 2019) to the salty disks around massive stars (Ginsburg et al. 2019). Thus, the utility of molecules as tracers of the chemical and physical properties and evolutionary history of astrophysical sources increases with the completeness of chemical inventories in these regions: the more knowledge we possess of the inventory, the more astrophysical information we can derive.

Detecting new molecules in space and using these detections to infer astrophysics has underpinned the foundation of astrochemical research for over half a century, although the pace of discovery truly exploded with the advent of molecular radio astronomy in the 1960s (McGuire 2018). Growing alongside laboratory and observational efforts to identify new molecules in space, astrochemical models were developed to attempt to reconstruct the network of chemical reactions occurring in the environments that were being studied (see, e.g., Wakelam et al. 2015 and Garrod et al. 2008). The analysis and refinement of these models, from relatively simple networks focusing on just a few molecules (Herbst & Klemperer 1973; Guzmán et al. 2015) to very large holistic approaches attempting to replicate the observed abundances in a source (van Dishoeck & Black 1986; Garrod 2013), provide a small window into the complex processes occurring toward a diverse set of environments (Schilke et al. 1997). Often, these models are descriptive rather than predictive: while they replicate many of the abundance ratios seen in an observation and provide substantial insight into the associated chemistry and physics, they can also struggle to predict abundances of other species that have not yet been observed (see, e.g., McGuire et al. 2015). For each newly detected molecule, we must contemplate a host of related species and reaction pathways necessary to describe its chemistry. To date, this aspect of astrochemistry has solely depended on chemical intuition: we draw on subjective expertise to determine what *may* be important areas in which to dedicate computational, laboratory, and observational efforts. As the complexity of molecules grows beyond a few carbon atoms, however, the number of possible isomers grows combinatorially and inference based on human intuition becomes neither tractable nor exhaustive.

Interestingly, fields such as drug and materials discovery face a similar problem; open source tools in the cheminformatics and machine-learning space have transformed how novel molecules are discovered and/or designed (e.g., David et al. 2020; Janet et al. 2020; Kulik 2020). By exploiting the scalability of chemically descriptive computer representations of molecules, we can systematically and exhaustively identify attractive candidates for astrochemical study. In this work, we

demonstrate the feasibility and accuracy of such an approach on a well-characterized chemical inventory, the cyanopolyyne peak of the Taurus Molecular Cloud 1 complex (TMC-1). Combining unsupervised machine learning of chemical embeddings with "classical" supervised machine learning regressors, we are not only able to successfully reproduce observed molecular abundances, but recommend and predict the abundances of thousands of chemically related molecules. Perhaps most importantly, our approach does not require prior knowledge of the physical/chemical conditions, contrasting with conventional chemical modeling, which can rely on parameters that are not always known and/or cannot be directly determined.

In this paper, we provide a verbose discussion of the workflow, theory, and implications of applying unsupervised machine learning for astrochemical inference; given the relatively niche intersection of cheminformatics, machine learning, and astrophysics, this paper is written with particular emphasis on the interpretation and reconciliation of machine learning predictions with chemical intuition. We begin by introducing machine representations of molecules, followed by details on the overall workflow and descriptions of various regressors. From there, we discuss and provide visualizations of the learned vector representations, contextualizing them in the broader scopes of chemical inventories and networks, and finally discuss the use of these embeddings for recommendation and regression.

2. Computational Methods

2.1. Molecule Representation Learning

In order for quantitative comparisons to be made with machine-learning methods, molecular features need to first be encoded into vector representations. A common approach is to hand pick features appropriate for the task at hand, for example, the length of hydrocarbons or the number and types of functional groups, and express properties using additivity schemes (Benson & Buss 1958). While these approaches are simple to implement, they are subject to the choice of features and as certain features may only occur in certain groups of molecules, hand-picked features are neither scalable nor balanced in their approach to representations. Alternatively, more systematic (e.g., Coulomb matrices Rupp et al. 2012) and unsupervised approaches such as MOL2VEC (Jaeger et al. 2018) provide means to generate molecule vectors without the need to choose descriptors. For this work, we have chosen the MOL2VEC approach as it does not require molecular structures—instead, MOL2VEC repurposes the WORD2VEC algorithm (Mikolov et al. 2013) developed for natural language processing and operates on linear string representations of molecules, specifically in the simplified molecular-input line-entry system (SMILES) format (Weininger 1988; O'Boyle 2012) commonly enumerated in large data sets.

The MOL2VEC algorithm decomposes the representation task into two aspects; unique atom environments defined by a radius hyperparameter are hashed with the Morgan algorithm (Morgan 1965) to form a dictionary/corpus of substructures, which are subsequently used to train a multilayer perceptron—a continuous bag of words architecture (Mikolov et al. 2013)—to learn a context-aware, unsupervised mapping of substructures (words) to molecule (sentences) vectors. Thus, for every canonical SMILES string that encodes every functional group and connectivity in a molecule, MOL2VEC generates n-dimensional vectors (in our case, 300 dimensions) as a sum of substructure vectors, capturing *every* molecular feature. This model description of chemistry has been successfully used for a number of predictive tasks, for example, drug activity screening (Das et al. 2021), property prediction and rationalization (Zheng et al. 2019), and chemical space exploration and visualization (Shibayama et al. 2020).

Table 1
Composition of the Data Set Used for This Work: Sources and Number of Entries within Each Data Set

Source	Number of Entries	References
ZINC	3,862,980	Sterling & Irwin (2015)
PubChem A	2,444,441	Kim et al. (2021)
PCBA	437,929	Wang et al. (2012)
QM9	133,885	Ramakrishnan et al. (2014)
NASA PAHs	3139	Boersma et al. (2014), Bauschlicher et al. (2018), Mattioda et al. (2020)
KIDA	578	Wakelam et al. (2015)
TMC-1	87	See Table A2

By training the model on a diverse set of SMILES strings—as with many unsupervised approaches—the resulting embeddings become more holistic in their description of chemical properties. Developing a data set for astrochemical purposes, however, requires striking a balance between descriptiveness and utility: cheminformatics data sets typically comprise large, biological molecules whereas those detected in the interstellar medium are smaller and oftentimes transient and unstable. Given that data set bias is currently a well-known problem in word embeddings used for natural language applications (Bolukbasi et al. 2016; Basta et al. 2019), we were mindful not to bias toward terrestrial chemistry albeit only qualitatively so. For this reason, we curated a comparatively small data set of molecule SMILES balanced in its emphasis on molecules relevant to interstellar chemistry (KIDA, NASA PAH database), and small to medium-sized molecules from several generalized data sets (QM9, ZINC, Pubchem, PCBA). For some data sets, SMILES notations for the molecules are not provided, for example, the NASA PAH database and KIDA; the former provides Cartesian coordinates and the latter InCHI notation. In both cases, OPENBABEL (O'Boyle et al. 2011) was used to convert these formats into SMILES strings. The sources of data are summarized in Table 1 and correspond to a total of 6,883,279 entries that are filtered for duplicates, resulting in 3,316,454 unique canonical SMILES strings that constitute the training data for MOL2VEC and for molecule recommendations. While we have not performed a systematic analysis into the influence of each public data set on the resulting embeddings, from Table 1, it is easy to conclude that the number and diversity of molecules contained in astrochemical data sets like KIDA and the NASA PAH database alone would not be sufficient for an adequately descriptive embedding. We note also that this data set comprises radicals and ions, however, we have chosen not to include isotopologues although they can be readily encoded in SMILES strings.

2.2. Model Pipeline

Figure 1 illustrates the computational flow: molecular structures encoded as SMILES strings are passed to the trained

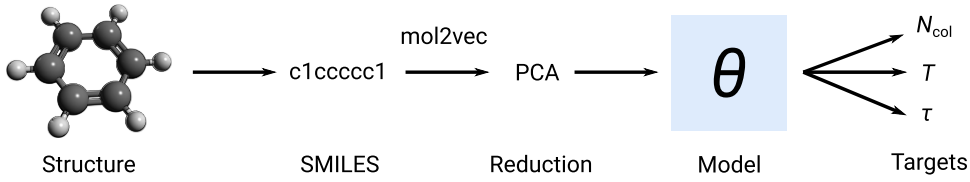

Figure 1. Proposed workflow for unsupervised training and prediction of molecular properties. Molecular structures can be encoded in a number of ways, ranging from atomic Cartesian coordinates to internal coordinate (or Z-) matrices. Structures are standardized with canonical SMILES, ensuring uniqueness in the data set, and transformed into molecule vectors using MOL2VEC. The dimensionality of the vectors is reduced via principal components analysis (PCA) and is used by various models to predict target properties.

Table 2
Summary of Models Used in This Work, Their Types, and Respective Optimized Hyperparameters during Cross-validation

Model	Abbreviation	Category	Hyperparameter Space
Linear regression	LR	P	
Ridge regression	RR	P	α
Bayesian ridge regression	BR	P	
Support vector regression	SVR	P	L_2 and ε regularization, γ
k-nearest neighbors	*k*NN	NP	Num. neighbors, distance metric
Random forest	RFR	NP	Num. trees
Gradient boosting	GBR	NP	Learning rate, num. estimators, subsample fraction, min. samples
Gaussian process	GPR	NP	α, kernel

MOL2VEC embedding model, generating 300-dimension feature vectors. Subsequently, we carry out a dimensionality reduction with principal components analysis (PCA) followed by clustering with *k* means. Of the regression algorithms surveyed here, Gaussian processes (GPs) are the most memory and computationally intensive due to the factorization of large matrices, which, for a naive implementation, scales with $O(n^3)$. Reduction with PCA decreases the memory usage substantially and *k*-means clustering attempts to include only molecules that are of immediate relevance to those found in the astrophysical source under investigation. For details on the PCA dimensionality reduction, see Appendix A. Prior to regression, we also perform feature standardization by scaling the feature values by their mean and variance, determined during training. In conjunction with the regressor-specific regularization methods, standardization attempts to preserve sparsity in the model which in turn contributes toward mitigating overfitting and improving accuracy. The code—including software environment specification—can be found at https://github.com/laserkelvin/umda.

2.3. Model Specification and Selection

The final step of the pipeline shown in Figure 1 is to perform supervised machine learning to predict column densities of molecules detected in TMC-1. To establish a baseline for performance, we tested a variety of commonly used machine-learning algorithms, chosen for their simplicity, and some for their well-documented performance and interpretability. Table 2 organizes the methods used and in Appendix B we provide a short overview of advantages and disadvantages of each. Each method is classified into whether there are learnable parameters or not; as we will discuss in subsequent sections, this will motivate the choice of method to apply.

The primary goal of each regressor is to accurately reproduce the observed column densities in TMC-1; for our purposes, this corresponds to the column densities of 87 molecules spanning from methylidyne CH to cyanonaphthalene (c-$C_{10}H_7CN$), with model accuracy measured by the mean squared error of the $\log_{10}$ column densities. To briefly summarize the data used for this work: 3.3 million molecules constitute the data set for training the embedding and PCA model; 455,461 molecules form the subset which is considered astrochemically relevant to TMC-1 through *k*-means clustering. For column density prediction, the data comprises 87 molecules with observed column densities toward TMC-1, bootstrapped (randomly sampled with replacement) with Gaussian noise ($\sigma = 0.5$) added to the log column densities to yield an effective data set of 800 observations—see Appendix C for more details.

For all models except LR and BR, we perform hyperparameter optimization using grid search combined with ten-fold cross-validation, whereby the bootstrapped data set is split into ten subsets of training and validation data and 20% of the species (17 molecules per split) are not used to fit the regressor. Table 2 summarizes the hyperparameters that are tuned as part of the exhaustive grid search. All models used in this work are implemented in SCIKIT-LEARN (Pedregosa et al. 2011). The hyperparameter and test scores for each model are summarized in Table A1. To assess the degree of overfitting, we perform learning curve analyses, whereby the trained model performance is evaluated as a function of data set size based on ten-fold cross-validation; these details can be found in Appendix C.

3. Results and Discussion

3.1. Vector Representations of Chemistry

The first step in the proposed pipeline is the generation of vector representations of molecules via unsupervised machine learning using the MOL2VEC method, which in turn is adapted from the WORD2VEC algorithm from the natural language processing domain. In this section, we explore a few properties of the learned embedding including the possible manipulations and information compression.

To infer how chemical intuition is encoded in the MOL2VEC vectors, Figure 2 shows how the similarity, or conversely distance, changes over chemical space defined as a spectrum between two extremes: small molecules like methyl cyanide (CH_3CN) and monolithic structures such as buckminsterfullerene (C_{60}). The two metrics, Euclidean distance and cosine

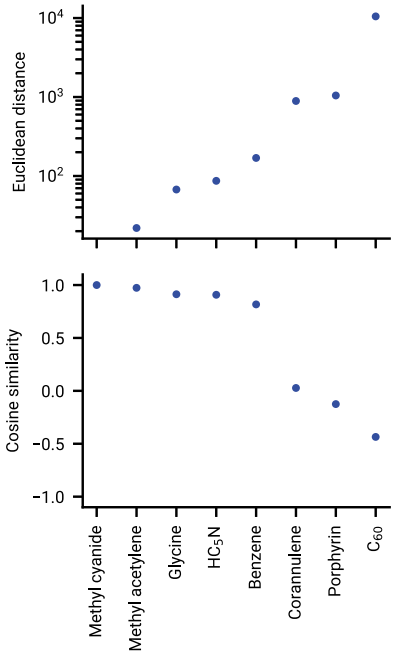

Figure 2. Similarity of arbitrarily chosen molecules with respect to methyl cyanide, given as a function of Euclidean distance (top) in log space and cosine similarity (bottom).

similarity, provide slightly different insight into how the vectors behave—the latter is scale invariant as it simply measures the alignment of two vectors, while the former is not. This is particularly important in differentiating between molecules that are highly similar; for example, the margin between CH_3CN and methyl acetylene (CH_3CCH), and glycine (NH_2CH_2COOH). Intuitively, the two methyl chains should be much more similar/closer in distance than CH_3CN and NH_2CH_2COOH in terms of molecule size and functionalization (i.e., the former have methyl groups).

To better understand the learned representation as well as the chemical inventory of TMC-1, we applied the Uniform Manifold Approximation and Projection (UMAP) method (McInnes et al. 2020) to visualize the embeddings of molecules detected toward TMC-1. This approach attempts to learn an approximation to the manifold of the embedding space, and produces a mapping between the approximate manifold and a lower dimensional representation in an unsupervised fashion. For our purposes, the goal is to visualize the two-dimensional chemical space comprised by molecules in TMC-1, while preserving the topology of the PCA reduced 70-dimensional vectors.

As shown in Figure 3, the UMAP method provides a unique perspective on chemical inventories and validates some assertions of what is contained in the embeddings. For example, chemically similar molecules such as the cyanopolyynes and their methylated variants are located in the same region (left side of Figure 3) and trends within these families are also observable (i.e., chain elongation). Near the center-right of Figure 3 we see clusters of smaller species, which constitutes the other extreme of molecules detected toward TMC-1, contrasting the large carbon chains. This dichotomy illustrates how chemical inventory characterization to date has largely followed a linear progression in chemical space—from right to left, as shown by the dashed line. The recent detections of large aromatic species such as indene (C_9H_8, Burkhardt et al. 2021) and cyanonaphthalenes ($C_{11}H_7N$, McGuire et al. 2021) intuitively correspond to a different type of chemistry, and is indeed identified in the UMAP projection as a cluster of molecules as somewhat orthogonal to the rest of the detections, progressing from c-C_3H_2 through to $C_{11}H_7N$ toward the top.

Taking the inventory of TMC-1 into a broader context, Figure 4 shows a UMAP projection of molecules detected in TMC-1 and species contained in the KIDA network. The main observations here are that by in large, the KIDA network overlaps well with the inventory of TMC-1, corroborating the current intuition of which species are important in the description of chemistry in dark molecular clouds. Where the KIDA points in chemical space are sparse, however, pertain to the recently detected aromatic molecules, which are clustered toward the center of Figure 4. This naturally indicates that, for the proper description of aromatic chemistry, a network such as KIDA would need to increase coverage of aromatic species, to "fill in the gap" in the embedding space.

3.2. Targets for Astrochemical Study

One aspect in astrochemistry that is currently poorly defined is the identification of potential molecules of interest for laboratory, observational, and/or modeling efforts. Using the chemical embeddings, likely targets for study can be readily identified simply by proximity in the latent space: molecules discovered in a given source or survey can be used as cluster centers, and molecules from the full data set within an arbitrary distance threshold can be proposed for study in a systematic fashion. The basis for this is chemical similarity, in a way similar to how isomers and conformers of detected species are viable targets; we have simply vectorized this process.

Here, we provide recommendations for viable candidates of targeted efforts by selecting 100 nearest neighbors for each of the 87 non-isotopologue species detected in TMC-1, and using quantum chemistry to estimate their rotational constants and dipole moments as to assist in their discovery. The list is filtered for: (1) duplicates; (2) heavy elements outside of C, N, O, Si, P, S; and (3) van der Waals complexes, leading to 1510 unique molecules. From this list, Cartesian coordinates for each structure were generated and optimized with the UFF force field Rappe et al. (1992) implemented in OPENBABEL (O'Boyle et al. 2011). The generated 3D structures are then refined at the ωB97X-D/6-31+G(d) level of theory (Rassolov et al. 1998; Chai & Head-Gordon 2008, chosen as a suitable compromise between computational expense and accuracy, in addition to well-known uncertainties and scaling factors (Lee & McCarthy 2020). For open-shell species, an unrestricted reference was used; results for these molecules should be taken conservatively—here we assume that the performance of the electronic structure method and basis provides the same systematic errors in the predicted structure as for the closed-shell case. It is likely that these species will require substantially more sophisticated treatments for desirable accuracy, including estimation of their fine-structure properties. Geometry optimization was performed using the GEOMETRIC package (Wang & Song 2016) with gradients calculated using PSI4 (Parrish et al. 2017). Out of the 1510 molecules, 148 were non-convergent either at the self-consistent field or geometry optimization steps.

The full list of molecules can be found in the Zenodo repository doi:10.5281/zenodo.5146276; for brevity, we highlight and rationalize a few recommendations. Some general observations of the candidates include:

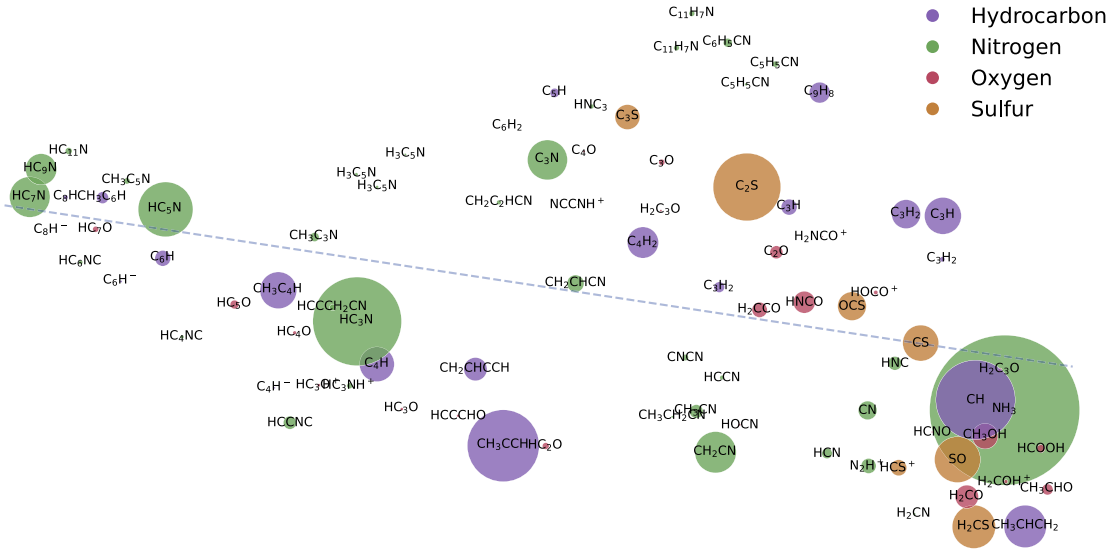

Figure 3. Visualization of the TMC-1 inventory, where the axes represent the UMAP learned two-component projection of the 70 dimensional vector molecule representation. Colors represent arbitrary classification of molecules and were not used for UMAP training. The size of each scatter point corresponds to the molecular column density. The dashed line corresponds to a linear fit to the projection as a visual guide.

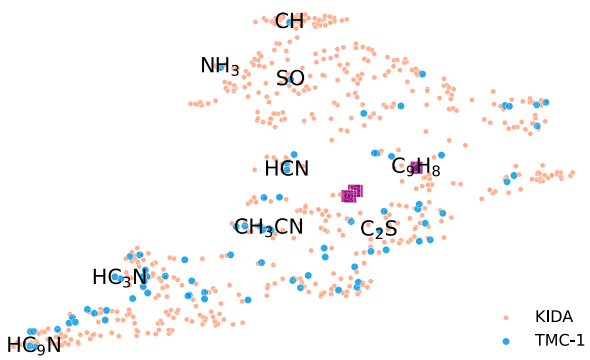

Figure 4. UMAP projections conditioned on the combined TMC-1 and KIDA molecule embeddings. Points with the shaded squares near the center of the image correspond to the aromatic ring molecules detected in TMC-1. Several molecules are annotated to represent the local region of chemical space.

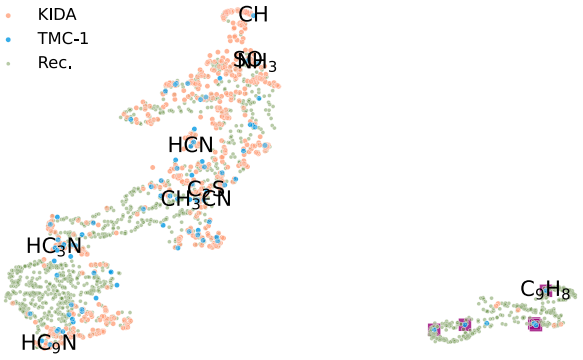

Figure 5. UMAP projection of the combined TMC-1 inventory (blue), KIDA network (peach), and the recommended molecules for study (green). Squares denote aromatic molecules detected in TMC-1.

1. The majority of molecules are unsaturated, containing at least one double or triple bond (1183 molecules, or 78%);
2. The vast majority of candidates are not pure hydrocarbons, with an average of at least 1.3 heteroatoms (338 pure hydrocarbons, 12%);
3. Most contain nitrogen, particularly as a $-C\equiv N$ group (508 cyanides, 33%);
4. Relevant to aromatic/ring species, molecules with up to three rings are recommended. A substantial number of recommended species (436, 29%) contain at least one ring. On a similarity basis, molecules with the same number of rings are typically recommended, although molecules with more or fewer rings are also identified in the search.

These observations reflect the state of the data set and the molecules currently detected in TMC-1: the majority of the molecules detected are indeed highly unsaturated, most are tagged with cyanides, and to date, only six aromatic species have been detected. Given the fact that a nearest-neighbors approach was used to identify potential candidates, it is not surprising that the recommendations closely resemble those already detected.

To understand how the recommended molecules using this nearest-neighbors approach is complimentary to chemical inventories and model networks like KIDA, Figure 5 provides another UMAP learned visualization, trained on the TMC-1 detections, KIDA species, and recommendations. In the case of the former, the group of molecules to the left corresponds to the aromatic molecules detected in TMC-1, with a large number of new recommendations contributing *significantly* toward three subclusters comprising C_9H_8, $C_{11}H_7N$, and C_5H_5CN/C_6H_5CN, respectively. For KIDA, we see that the recommended species act to fill in large gaps in chemical space, particularly in the regions concerning larger species—between HC_3N and HC_9N. In both instances, the main observation is that the recommended molecules add to regions in chemical space that were previously sparse, particularly toward larger molecules. Viewed in this way, there are substantially fewer recommendations for smaller molecules as the inventory is relatively complete, compared to the number of isomers and conformers possible for larger species.

3.3. Machine Learning of Chemical Inventories

From the prior sections, it is clear that our embedding successfully captures aspects of chemical intuition. The aim

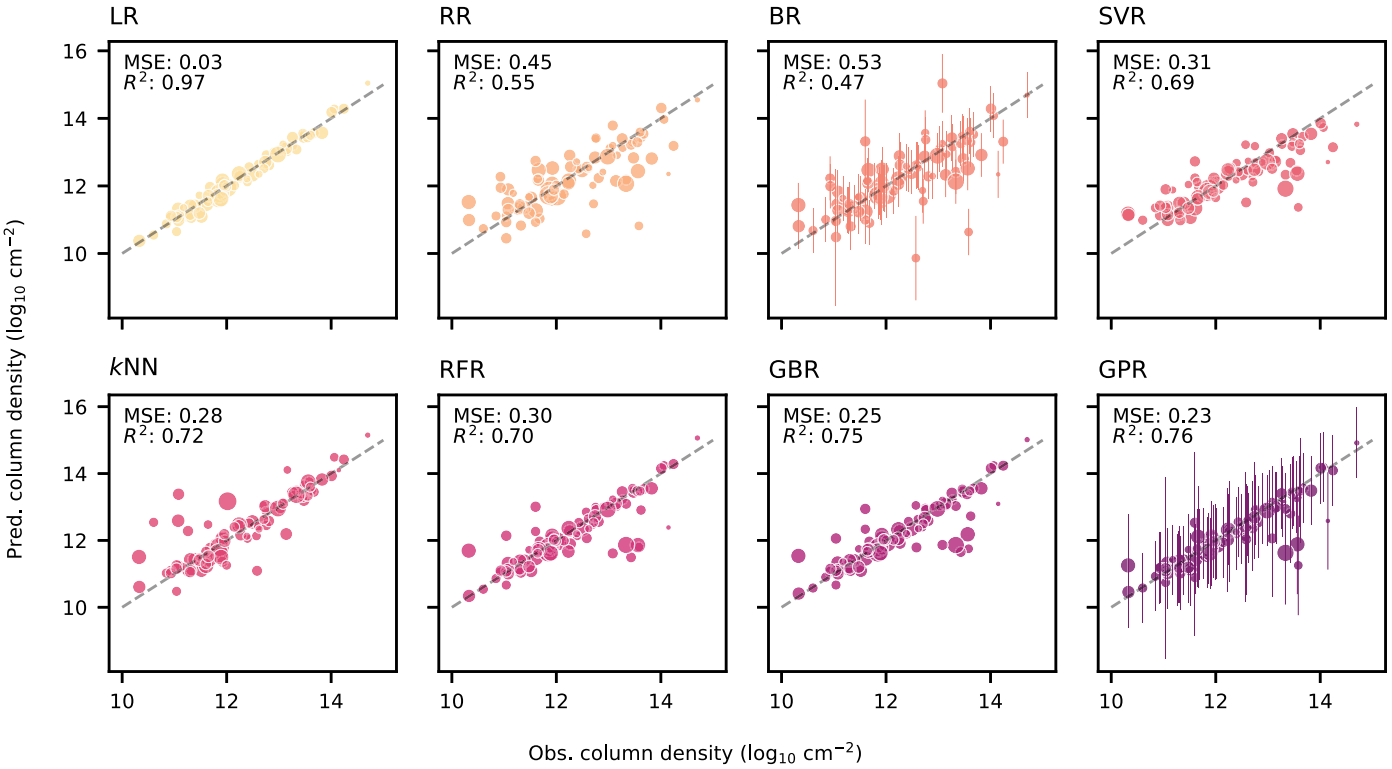

Figure 6. Observed column densities plotted against the corresponding model predictions. The size of each point is proportional to the molecular weight, as an approximate measure for molecular complexity. Three molecules that are underfit by several thousand orders of magnitude by the LR model are excluded from the plot and from the metric calculation (see text). For probablistic models, 1σ uncertainty is given as error bars.

now is to relate the chemical features with physical parameters —the goal of chemical modeling—in a way that links molecules to directly and non-directly observable aspects of the interstellar medium. Here, we demonstrate the approach for column densities.

Figure 6 compares the true observed column densities with model predictions following model selection. We see that most algorithms—with the exception of linear regression, which drastically overfits without regularization—are able to reproduce the observed column densities remarkably well. In the case of linear regression, the coefficients become unrealistically large and severely underfits three molecules (vinyl cyanide, thioformaldhyde, and butatrienylidene) by several thousand orders of magnitude (these are excluded from Figure 6), highlighting a significant need for heavily regularized models owing to the relatively small data sets and chemically simple molecules relevant to astrochemistry.

The ridge regression models (frequentist and Bayesian) provide the necessary regularization to linear models, while performing qualitatively well and provide evidence that, at least locally, the abundance is approximately linear in the chemical space comprised by molecules detected in TMC-1 and that there are no specific *classes* of molecules that demonstrate peculiarities in their abundance. This observation is reinforced by the fact that *k*NN performs extremely well even with only a few neighbors (Table A1). Were the opposite true, there would be systematic effects in the residuals, however, the errors in the linear models appear normally distributed. As each molecule is or can be represented in the same embedding basis—regardless of whether they are ions, radicals, pure hydrocarbon or not—the linear function should be able to readily interpolate between molecules detected in TMC-1 and extrapolate to molecules not yet seen as simple extensions of the vector space they comprise. We note that in some sense, this is an intuitive result given the observed trends for hydrocarbon chains such as the cyanopolyynes and their methylated analogues, however, here we generalize this trend beyond the one-dimensional slices in chemical space (i.e., the length of carbon chains). A short discussion regarding this aspect can be found in Appendix E.

From Figure 6, the bottom row comprises models with the lowest bias, namely *k*NN, RFR, GBR, and GPR. The case of *k*NN is particularly important as it demonstrates that column densities can be expressed as smooth functions of local distance and that learnable parameters are not necessary to describe this behavior. For the ensemble methods RFR and GBR, we see that both regressor types provide approximately the same degree of excellent performance—both methods are able to estimate feature importance, although in the current implementation we are unable to interpret the features directly as they are learned with unsupervised methods. However, future applications could correlate the embedding dimensions with hand-picked features, which in turn could provide a means to translate which aspects of molecules are most critical to the chemistry of an environment.

The two probabilistic models, BR and GPR, warrant some extra discussion. In particular, BR provides a highly attractive approach to modeling chemical inventories as it provides an extremely simple and regularized method to not only obtain abundances but also uncertainties. For the vast majority of molecules, we are able to accurately reproduce the observed abundances within an order of magnitude, and certainly within 2σ uncertainty. In the case of GPR, we substantially increase the modeling flexibility as well as having a significantly more interpretable prediction uncertainty: two outlier points are

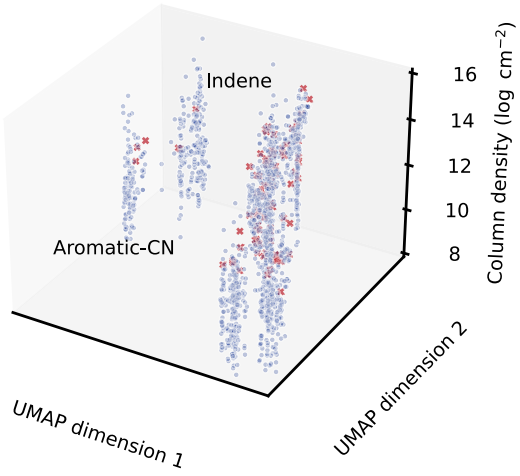

Figure 7. Scatter plot of the UMAP space for the 1510 recommended molecules (blue circles) and molecules detected toward TMC-1 (red crosses), and the corresponding GP predictions of column density. Locations of aromatic molecules are annotated.

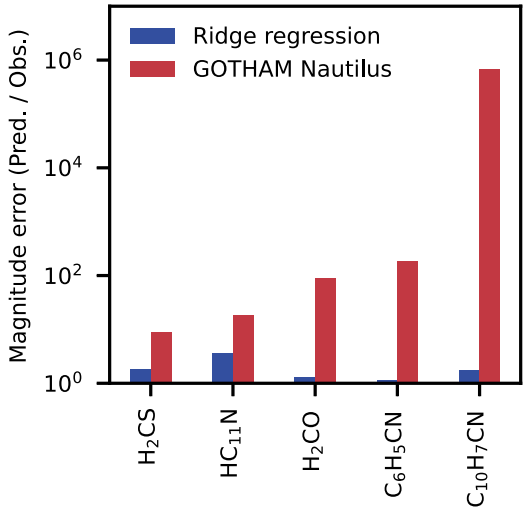

Figure 8. Comparison of molecular abundances predicted by the ridge regression model (blue) and the current state-of-the-art chemical model, NAUTILUS (Ruaud et al. 2016), for TMC-1 (red), given as the magnitude for the predicted and observed ratio.

ascribed much larger uncertainties than others, one of which is consistently overpredicted by each regressor ($NCCNH^+$). This provides a reason to revisit the column density estimate by Agúndez et al. (2015) as to investigate whether there are anomalies in the chemistry or in the embeddings.

Having established the data set performance of each machine-learning method, we can now use them for extrapolation, predicting column densities for unseen molecules. For the sake of brevity, we utilized the trained GPR model to predict column densities of the 1510 recommended molecules from Section 3.2. Figure 7 illustrates the result in three dimensions: the horizontal plane represents the UMAP learned 2D projection of the chemical space spanned by the TMC-1 data set and the recommended molecules and the vertical axis is the GPR predicted column density. In the left cluster of points, which contain the majority of detected molecules, we see that a significant number of recommended species are predicted to have column densities (10^{10-12} cm^{-2}) comparable to those already detected—laboratory and computational efforts would likely enable their detection or at least derivations of upper limits that can be used to refine machine-learning and chemical models.

3.4. Comparisons with Chemical Models

In the preceding sections, we have evaluated the performance of various machine-learning methods for predicting molecular column densities. A natural comparison to be made is with state-of-the-art kinetic chemical models of TMC-1, which have been the main method of choice for predicting abundances of molecules a priori. Here, we utilize the three-phase chemical model NAUTILUS (Ruaud et al. 2016) with the latest chemical network, elemental abundances, and physical conditions used to describe the formation of aromatics by Burkhardt et al. (2021). Figure 8 compares the predicted/observed ratios for five chosen molecules as samples across molecular complexity. Ridge regression is chosen as the baseline algorithm for comparison and we see that it reproduces the abundance of each molecule within an order of magnitude of the observations. For chemical models the molecular abundance is time-dependent and for the purposes of discussion we have chosen the time slice where the cyanopolyynes have their peak abundances. Under these conditions, the relatively simple species H_2CO, H_2CS, and $HC_{11}N$ can be seen to agree with the observed abundances to within one to two orders of magnitude. For the aromatic rings, representing more complex species, the chemical model significantly underpredicts their abundance; for cyanonaphthalene, this discrepancy is nearly six orders of magnitude, as originally discussed in McGuire et al. (2021).

Beyond abundances, however, the current machine-learning approach shares little to no similarities with chemical modeling. While Figure 8 shows quite definitively that ridge regression is able to reproduce observed abundances accurately, chemical models are an avenue for inferring physical information about an astrophysical source by reproducing abundances, albeit a complex and difficult process (Herbst & Klemperer 1973; van Dishoeck & Black 1986; Agúndez & Wakelam 2013). With each new interstellar molecule detection, the chemical network must be updated with new species hypothesized to be important in its formation and destruction, along with reaction rates measured, or more commonly, approximated (Wakelam et al. 2010). In contrast, the allure of the machine-learning approaches is the ease of extension: new molecules can be added simply using SMILES strings, and as we have demonstrated in this work, can readily scale up from hundreds to millions of molecules. As we have alluded to in the previous section, a natural extension of this work is to connect physical parameter inference from chemical models with the generalizability of machine-learning models: the latter informs the former by providing abundances and constraints on unobserved species as well as providing recommendations for new molecules to add to chemical networks. Imputation through machine learning thereby results in a self-consistent and systematic approach to the astrochemical inference—we intend to explore and expand upon these ideas in forthcoming work.

3.5. Generalizations beyond TMC-1

The approach we have detailed here essentially comprises two distinct machine-learning parts: an unsupervised molecule embedding learning and compression task, followed by supervised training of regressors on molecular inventories.

The former creates general purpose vectors in a latent chemical space, and is not limited to the study of chemical inventories or regression—indeed, MOL2VEC vectors have been used for a range of machine-learning tasks such as drug activity prediction. For the latter, chemical inventory regressors can be applied in two ways. First, the pre-trained TMC-1 model can be used to quantitatively assess chemical differences between sources (particularly dark molecular clouds) as a form of "chemical baseline." For example, the TMC-1 regressors can be used to identify systematic offsets (bias) and specific deviations (variance) in chemical abundances without re-training, which could be used to infer dynamics/kinematics unique to a given source, for example core collapse (e.g., L1521E Hirota et al. 2002) and protostar/warm carbon chain chemistry (e.g., L1527 Sakai et al. 2007).

The second application we foresee with machine-learning regressors pertains to other well-characterized chemical inventories outside of TMC-1. Molecule-rich, prototypical sources such as VY Canis Majoris and IRC+10216 make excellent candidates for quantitative inventory analysis, albeit with significantly more complicated dynamics such as shock chemistry and photoinduced processes. For these applications, the simpler regressors reported here are unlikely to model considerations such as radial and angular extent or time-dependence; rather, more sophisticated parametric models such as neural networks will be required.

4. Conclusions

In this work, we have demonstrated the viability for simple machine-learning models to learn and predict entire chemical inventories. Combining the MOL2VEC model embeddings with algorithms as simple as linear regression, we are able to reproduce the column densities of 87 molecules detected toward TMC-1 to well within an order of magnitude without the need for prior knowledge pertaining to the physical conditions of the source. With this, we show that the molecule embeddings can be used to identify new likely candidates for interstellar detection and study, based on quantitative measures of chemical similarity between molecule vectors as in a nearest-neighbors approach, and using the machine-learning models to predict their expected column densities as one way to assess detectability. The attractiveness of our approach is the ability to *systematically* infer the presence of astrophysically important molecules that are not directly observable, for example, those without a rotational spectrum, or where conditions are unfavorable (e.g., partition functions), and to provide a baseline for determining the role of dynamical effects, such as grain-surface chemistry. In this way, predictions from machine-learning models can be used to impute chemical networks used in conventional chemical modeling, from which we can confidently and comprehensively derive astrophysical insight.

The authors thank the reviewers for extremely helpful comments and feedback, in particular regarding model regularization and the suggestion to bootstrap the data set. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The Green Bank Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. J.P. and V.V. acknowledge funding and research support from the SAO REU program; the SAO REU program is funded in part by the National Science Foundation REU and Department of Defense ASSURE programs under NSF grant No. AST-1852268 and by the Smithsonian Institution. A.M.B. acknowledges support from the Smithsonian Institution as a Submillimeter Array (SMA) Fellow. A.M.B. would like to also thank V. Wakelam for use of the nautilus v1.1 code. M.C.M. and K.L.K.L. acknowledge financial support from NSF grant AST-1908576 and NASA grant 80NSSC18K0396.

Appendix A
Dimensionality Reduction

The PCA model is trained on the full 3.3 million molecules to identify an adequate number of dimensions required to explain variation in chemical space, while providing computational benefits. Ultimately, we chose to use 70 principal components, corresponding to a 0.96 explained variance ratio (Figure A1), or just over 2σ of variation accounted for by the components.

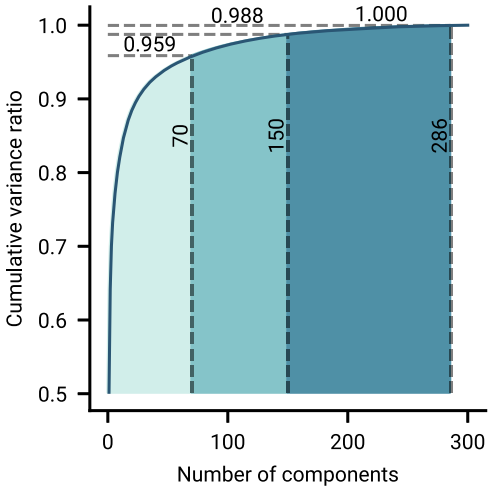

Figure A1. Cumulative explained variance ratio as a function of the number of components in the incremental PCA. The dashed lines represent the number of components that approximately correspond with 2σ, 2.5σ, and 3σ variation.

Appendix B
Supervised Machine Learning Regressors

The simplest parametric and non-parametric models considered in this work are linear regression and *k*-nearest neighbors (*k*NN), respectively, which represent the two types of model abstraction. The former makes the assumption that the column density varies linearly on a *global* scale in chemical space, while the latter expresses the column density as a distance weighted function. Thus, LR parameterizes a function that dictates the abundance of molecules decreases linearly from small to large molecules (i.e., chemical complexity), whereas for *k*NN the focus is *local* chemical similarity.

As we have seen in the results, however, linear regression results in significant overfitting, whereby the linear coefficients become extremely large in order to fit the majority of the data set. To alleviate this, ridge regression includes an L_2 coefficient norm penalty to the loss function, which encourages the linear coefficients to be small. A variation of this is a probabilistic treatment of the model coefficients, defined in Bayesian ridge regression, where predictions are linear coefficients sampled from gamma distributions whose parameters are determined by maximizing the log-marginal likelihood during fitting.

The remaining models are more sophisticated in their use of embedding space. First, ε-support vector regression (SVR; Drucker et al. 1996; Platt 1999) builds on top of linear regression by applying an ε regularization term and a kernel transformation to the features prior to regression: in doing so, the model is capable of capturing nonlinearity in the embeddings while maintaining a simple linear mapping onto the column densities. For this work, we consider a radial basis function kernel with width γ. The ensemble learning methods we employ include random forest (RFR; Liaw & Wiener 2002) and gradient boosting (GBR; Friedman 2002); these often substantially improve upon linear models in both bias (boosting) and variance (forests) performance by aggregating the results of multiple weak models that collectively form an ensemble. The former comprises submodels that are based on randomly selected features with replacement, with the result given as an error-weighted average of all submodel predictions. The latter sequentially trains predictors with weighted data, where the weights are given by the gradient of the error from the prior predictor; a large number of estimators thus tend to yield results with very low bias, and as an ensemble, low variance.

The last method we consider are GPs, which treat *functions* of molecular properties as a stochastic process collectively defined by mean and covariance/kernel functions which—similar to *k*NN—expresses the column density as a function of distance in the embedding space. Among the regression models considered here, GPs are unique in their probabilistic nature and in being the most flexible, given the ability to design kernel functions that optimally suit the embedding space. Given, however, that the embeddings are not directly interpretable, we provide here only a simple mixture kernel comprising three subkernels: the sum of rational quadratic, dot product, and white noise kernels. This covariance function was formulated assuming two components: the rational quadratic kernel describes short ranged, smoothly variability in the latent space, while the dot product kernel explicitly models pairwise, linear contributions between molecules. The former dominates for molecules that are chemically very similar (e.g., HC_5N and HC_7N) and the latter contributes for molecules that are at different scales of chemical complexity or size but share common features such as functional groups (e.g., HC_5N and benzonitrile C_6H_5CN). Finally, the white noise kernel provides modeling flexibility in describing uncertainty/noise in the observed column densities.

For each of the models, we perform hyperparameter tuning using grid search with cross-validation; the optimal hyperparameters are shown in Table A1.

Table A1
Tuned Hyperparameters Based on Grid Search with Ten-fold Cross-validation

Method	Hyperparameter	Value	Validation MSE[a]
LR			$2 \times 10^2$1
RR	α[b]	1.	0.54
BR	α[b]	1.7	0.51
SVR	C	100	0.48
	ε	10^{-1}	0.58
*k*NN	Distance	cosine	0.22
	N	30	
RFR	N	50	0.53
GBR	Learning rate	0.01	0.48
	Samples per leaf	0.3	
	Samples per split	0.1	
	Number of estimators	100	
	Subsampling fraction	0.8	
GPR	α[c]	3×10^{-5}	0.28

Notes. For all other omitted hyperparameter values, the default values in SCIKIT-LEARN were used.
[a] Mean squared error based on the non-bootstrapped observations.
[b] L_2 penalty term.
[c] Noise added to the diagonal of the kernel matrix for numerical stability.

Appendix C
Data Set Bootstrapping

One aspect that became apparent over the course of this work was the susceptibility of supervised regressors overfitting the TMC-1 observations. Given that the PCA vectors constitute 70 dimensions, and that the true data set only constitutes 87 observations, even linear models can be considered as over-parameterized. To alleviate this, we used the bootstrap method to effectively generate "new data," whereby the original data set is resampled and Gaussian noise ($\sigma = 0.5$) is added to the log column densities, yielding an effective data set of 800 points. As seen in Figure A2, more training examples improve the performance of each model, as measured by the mean squared error with respect to the observed TMC-1 column densities.

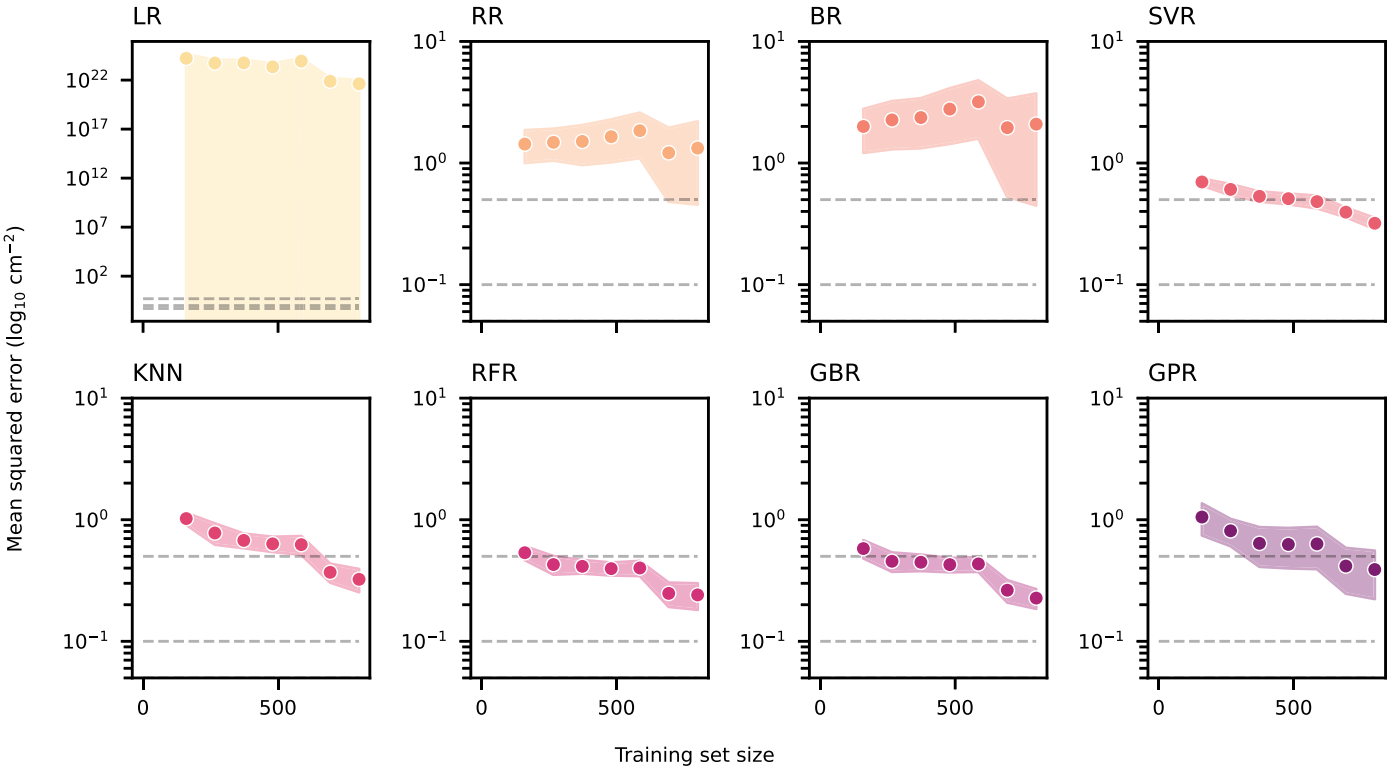

Figure A2. Learning curve analysis using the bootstrapped data set for each supervised regressor. The abscissa corresponds to the number of training examples, while the ordinate represents the mean squared error based on the non-bootstrapped, true column density values. The shaded regions represent 1σ variation in the error, estimated with ten-fold cross-validation.

Appendix D
Comparison of Hand-picked Descriptors and Unsupervised Embeddings

In this work, we utilize molecule embeddings that are learned via unsupervised machine learning. While the main allure of this approach is a scalable method for featurization, it is important to compare with hand-picked features of molecules for supervised regression. To make this comparison, we arbitrarily chose 15 molecule descriptors to constitute the feature vectors to perform regression as was done with the MOL2VEC vectors. The descriptors we chose are implemented in RDKit (Landrum 2020), which include: number of atoms, number of bonds, molecular weight, average bond order, number of aromatic rings, number of valence electrons, *FpDensityMorgan*(1, 2, 3), and the number of atoms for carbon, oxygen, nitrogen, sulfur, and phosphorus. The descriptors should comprise what is necessary to describe the various radicals, charge states, and molecule complexity associated with molecules in TMC-1. Unlike the "production" workflow, we did not perform feature normalization and ridge regression was used to make the comparison; both choices were intended to establish a baseline in the expected performance, and ridge regression was chosen given to its relative model simplicity and its L_2 regularization term to prevent overfitting.

Figure A3 compares the training and test errors for ridge regressors using hand-picked and MOL2VEC vectors, based on a 0.8/0.2 train/test split of the bootstrapped data set. We see that for very small values of α—corresponding to minute regularization—the MOL2VEC vectors tends to overfit and yield a large error with respect to the true column densities. In the range of $\alpha = 1$–30, the MOL2VEC model yields better performance than the hand-picked alternative and at larger values of α both models begin to be over-regularized. The regression performance for hand-picked features

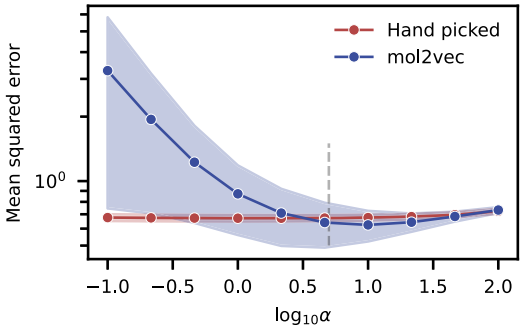

Figure A3. Performance of ridge regressors using hand-picked (red) and MOL2VEC (blue) vectors as a function of regularization term α. The dashed line represents the value of α used in Figure A4. Increasing regularization strength goes from left to right. The shaded region corresponds to 1σ based on twenty-fold cross-validation.

stays relatively constant for the range of α values used in comparison to the results seen for the MOL2VEC features, which shows a large degree of variability.

As an alternative visualization of model performance, Figure A4 (top row) shows the prediction errors and corresponding R^2 values (measured with respect to the true column densities) based on both featurization methods when fit to the full bootstrapped data set, with $\alpha = 5$. The motivation here is to test the performance of features, given the same nominal modeling flexibility and to mitigate variance across cross-validation sets. We see that the hand-picked features display a significant degree of bias (i.e., in the predictions owing to the fact that many molecules are not adequately described). In contrast, the MOL2VEC has significantly less bias

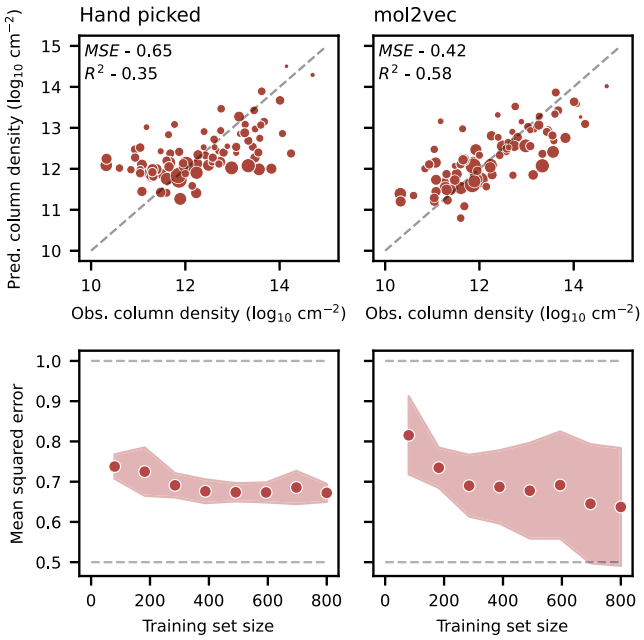

Figure A4. Comparison of R^2 plots (top row) and learning curves (bottom row) between hand-picked features (left column) and MOL2VEC embeddings (right column). Shaded regions in the learning curves represent the $\pm 1\sigma$ in model performance, as estimated with twenty-fold cross-validation.

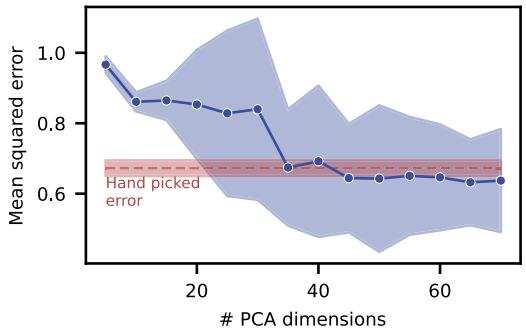

Figure A5. Performance of ridge regression models ($\alpha = 10^{-5}$) using the MOL2VEC embeddings as a function of the number of PCA dimensions used for regression. The red horizontal lines correspond to train (dashed) and test (dotted) errors when using hand-picked features with the same ridge regression model.

and variance, seen both in the scatter as well as the mean squared error and R^2 metrics.

Finally, Figure A5 shows the performance of a ridge regression model with an amount of regularization ($\alpha = 5$) where both models perform well as a function of the number of PCA feature dimensions in the MOL2VEC embeddings, compared to hand-picked features as an effective baseline. We see that for a comparable number of dimensions to the hand-picked features (∼15–20 dimensions), the MOL2VEC features actually result in poorer performance: this is not surprising as the PCA model was trained on the full 3.3 million molecules, which substantially biases the components toward a general description of chemistry not necessarily relevant to that of TMC-1; in particular, highly reactive species do not constitute a significant part of typical public databases. For specific tasks, such as comparisons between dark molecular clouds, one could perform the PCA on the TMC-1 data set specifically to obtain the transformations that best describe dark-cloud chemistry, which would likely then decrease the number of dimensions required for better model performance. Nonetheless, this result makes the distinction that for a small number of dimensions, features chosen by human intuition can potentially provide better performance than embeddings obtained via unsupervised learning.

Appendix E
Abundance and Molecular Complexity

To more quantitatively assess model performance as a function of molecular complexity, we can perform linear interpolation between two arbitrary molecules as one-dimensional slices in chemical space and look for systematic errors in the predicted column densities. Figure A6 interpolates between formaldehyde (H_2CO) and 2-cyanonaphthalene (c-$C_{10}H_7CN$) and identifies the nearest detected molecule to each interpolated point, with the distance from H_2CO providing a naive, relative representation of chemical complexity—the larger the distance, the more "complex" the molecule is. From a modeling perspective, it is important to highlight that there is no individual metric that best describes complexity in a way that reflects the common intuition that abundance is anticorrelated with molecular complexity. Figure A6 highlights this well, as the number of atoms—which is partially encoded in the distance and a common metric used in astrochemistry—barely correlates with column density. However, we see that all methods are able to predict the abundance of molecules within this small subset to well within an order of magnitude, and in particular, the linear models show that the abundance can be expressed as linear functions of 70 PCA dimensions, even if the mapping between individual descriptors and the abundance is not.

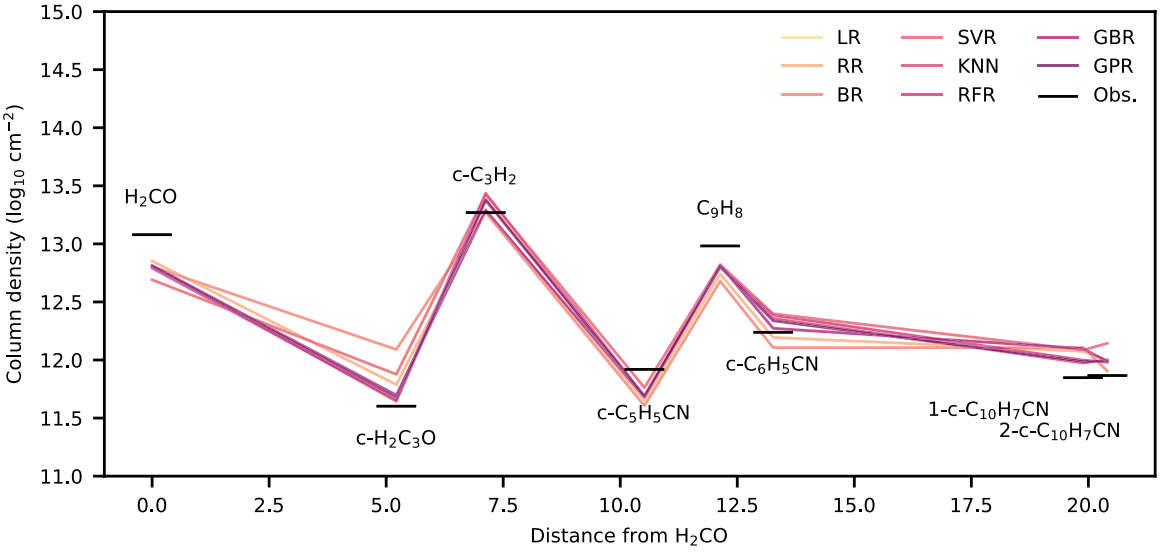

Figure A6. Predicted column densities for select molecules (ordinate) vs. Euclidean distance from H_2CO (abscissa). The black lines represent observed column densities.

Appendix F
Molecule Data Set

Table A2 summarizes the molecules detected toward TMC-1 and their respective references, totaling 87 unique species. Based on our current knowledge of the chemical inventory of TMC-1, the size and complexity of this structure—and on grounds of chemical similarity with the remaining seven—would intuitively lead to a vanishingly small "true" column density.

Table A2
Regression Data Set Comprising Molecules Detected toward TMC-1

Formula	SMILES	Column Density ($\log_{10}$ cm^{-2})	References
CH_3C_6H	CC#CC#CC#C	12.4914	Remijan et al. (2006)
CH_3C_4H	CC#CC#C	13.4771	MacLeod et al. (1984)
CH_3C_5N	CC#CC#CC#N	11.9243	Remijan et al. (2006)
CH_3C_3N	CC#CC#N	12.2553	Remijan et al. (2006)
N_2H^+	N#[NH+]	12.6990	Choi et al. (2017)
NH_3	N	14.6998	Gratier et al. (2016)
CH_3OH	CO	13.1614	Gratier et al. (2016)
C_3H	C1 = C=[C]1	13.4800	Gratier et al. (2016)
C_3H	[CH+]=C=[C-]	12.7497	Gratier et al. (2016)
C_3H_2	C1 = C=C1	11.7701	Gratier et al. (2016)
C_3H_2	C1C#C1	13.2695	Gratier et al. (2016)
C_3H_2	C = C=[C]	12.3979	Cernicharo et al. (1991)
CH_3CCH	CC#C	14.0607	Gratier et al. (2016)
C_2O	[C+]#C[O-]	12.5705	Gratier et al. (2016)
CH_2CN	[CH2]C#N	13.5798	Gratier et al. (2016)
CH_3CN	CC#N	12.6096	Gratier et al. (2016)
HNCO	N = C = O	13.0294	Gratier et al. (2016)
CS	[C-]#[S+]	13.4594	Gratier et al. (2016)
CH_3CHO	CC=O	12.4298	Gratier et al. (2016)
HCS^+	C#[S+]	12.7597	Gratier et al. (2016)
H_2CS	C = S	13.6201	Gratier et al. (2016)
SO	S = O	13.6702	Gratier et al. (2016)
C_4H	[C]#CC#C	13.4298	Gratier et al. (2016)
C_4H_2	C = C = C=[C]	13.3365	Gratier et al. (2016)
C_3N	[C]#CC#N	13.5502	Gratier et al. (2016)
HNC_3	[C-]#C-C#[NH+]	11.6803	Gratier et al. (2016)
C_3O	[C]#C[C]=O	11.9201	Gratier et al. (2016)
HC_3NH^+	C#CC#[NH+]	11.8698	Gratier et al. (2016)
CH_2CHCN	C=CC#N	12.8102	Gratier et al. (2016)
HCCCHO	C#CC=O	11.2601	Gratier et al. (2016)
C_2S	[C+]#C[S-]	14.0086	Gratier et al. (2016)
OCS	O = C = S	13.2601	Gratier et al. (2016)
C_5H	[CH+]=C = C = C=[C-]	12.2695	Gratier et al. (2016)
C_3S	[C-]#CC#[S+]	13.1399	Gratier et al. (2016)
C_6H	[C]#CC#CC#C	12.7404	Gratier et al. (2016)
HC_3N	C#CC#N	14.2430	Xue et al. (2020)
HCCNC	C#C[N+]#[C-]	12.5821	Xue et al. (2020)
HC_5N	C#CC#CC#N	13.8254	Xue et al. (2020)
HC_4NC	C#CC#C[N+]#[C-]	11.5172	Xue et al. (2020)
HC_7N	C#CC#CC#CC#N	13.5623	Xue et al. (2020)
HC_6NC	C#CC#CC#C[N+]#[C-]	11.6064	Xue et al. (2020)
HC_9N	C#CC#CC#CC#CC#N	13.3345	Loomis et al. (2021)
$HC_{11}N$	C#CC#CC#CC#CC#CC#N	12.0170	Loomis et al. (2021)
C_5H_5CN	C1C=CC=C1C#N	11.9191	Lee et al. (2021b)
C_5H_5CN	C1C=CC(=C1)C#N	11.2788	Lee et al. (2021b)
$C_{11}H_7N$	C1=CC=C2C(=C1)C=CC=C2C#N	11.8663	McGuire et al. (2021)
$C_{11}H_7N$	C1=CC=C2C=C(C=CC2=C1)C#N	11.8482	McGuire et al. (2021)
C_6H_5CN	C1=CC=C(C=C1)C#N	12.2380	McGuire et al. (2021)
$HCCCH_2CN$	C#CCC#N	11.9643	McGuire et al. (2020)
H_3C_5N	C#C/C = C/C#N	11.3874	Lee et al. (2021a)
H_3C_5N	C=CC#CC#N	11.0719	Lee et al. (2021a)
H_3C_5N	C#C/C=C\C#N	11.3032	Lee et al. (2021a)
C_8H	[C]#CC#CC#CC#C	11.6628	Brünken et al. (2007)
C_8H-	C#CC#CC#CC#[C-]	10.3222	Brünken et al. (2007)
C_6H-	C#CC#CC#[C-]	11.0792	Brünken et al. (2007)
C_4H-	C#CC#[C-]	10.9294	Brünken et al. (2007)
H_2CCO	C=C=O	12.7118	Soma et al. (2018)
CN	[C]#N	12.8899	Pratap et al. (1997)
HNC	[C-]#[NH+]	12.6201	Pratap et al. (1997)
HC_7O	C#CC#CC#C[C+]=O	11.8921	Cordiner et al. (2017)
HC_5O	C#CC#C[C+]=O	12.2304	McGuire et al. (2017)
H_2CN	C=[N]	11.1761	Ohishi et al. (1994)
H_2CO	C=O	13.0792	Soma et al. (2018)

Table A2
(Continued)

Formula	SMILES	Column Density ($\log_{10}$ cm^{-2})	References
HC_3O^+	C#CC#[O+]	11.3222	Cernicharo et al. (2020)
$HOCO^+$	O=C=[OH+]	11.6021	Cernicharo et al. (2020)
H_2COH^+	C=[OH+]	11.4771	Cernicharo et al. (2020)
H_2NCO^+	NC#[O+]	10.6021	Cernicharo et al. (2020)
HCNO	C#N[O]	10.8451	Cernicharo et al. (2020)
HOCN	OC#N	11.0414	Cernicharo et al. (2020)
C_4O	[C]#CC#[C]=O	11.0792	Cernicharo et al. (2020)
HCOOH	C(=O)O	12.1461	Cernicharo et al. (2020)
HC_2O	C#[C]=O	12.0000	Cernicharo et al. (2020)
HC_3O	C#C[C]=O	11.3010	Cernicharo et al. (2020)
HC_4O	C#CC#[C]=O	11.4771	Cernicharo et al. (2020)
H_2C_3O	C=C=C=O	11.0414	Cernicharo et al. (2020)
H_2C_3O	C1=C(=O)=C1	11.6021	Cernicharo et al. (2020)
CH	[CH]	14.1461	Sakai et al. (2013)
CNCN	[C]#NC#N	11.9542	Agúndez et al. (2018)
$NCCNH^+$	N#CC#[NH+]	10.9345	Agúndez et al. (2015)
C_6H_2	C=C=C=C=C=[C]	10.3284	Langer et al. (1997)
CH_3CHCH_2	CC=C	13.6021	Marcelino et al. (2007)
CH_2C_2HCN	C=C=CC#N	11.6532	Lovas et al. (2006)
HCN	C#N	12.3892	Hirota et al. (1998)
C_9H_8	c1ccc2c(c1)CC=C2	12.9823	Burkhardt et al. 2021
CH_2CHCCH	C=CC#C	13.0792	Cernicharo et al. (2021)
HCCN	N#C[CH+]	11.6435	Cernicharo et al. (2021)
CH_3CH_2CN	CCC#N	11.0414	Cernicharo et al. (2021)

Note. The references provided correspond to the source used for the column density, not initial detection.

ORCID iDs

Kin Long Kelvin Lee https://orcid.org/0000-0002-1903-9242
Andrew M. Burkhardt https://orcid.org/0000-0003-0799-0927
Michael C. McCarthy https://orcid.org/0000-0001-9142-0008
Brett A. McGuire https://orcid.org/0000-0003-1254-4817

References

Agúndez, M., Cernicharo, J., de Vicente, P., et al. 2015, A&A, 579, L10
Agúndez, M., Marcelino, N., & Cernicharo, J. 2018, ApJL, 861, L22
Agúndez, M., & Wakelam, V. 2013, ChRv, 113, 8710
Basta, C., Costa-jussà, M. R., & Casas, N. 2019, arXiv:1904.08783
Bauschlicher, C. W., Ricca, A., Boersma, C., & Allamandola, L. J. 2018, ApJS, 234, 32
Belloche, A., Garrod, R. T., Müller, H. S. P., et al. 2019, A&A, 628, A10
Benson, S. W., & Buss, J. H. 1958, JChPh, 29, 546
Boersma, C., Bauschlicher, C. W., Ricca, A., et al. 2014, ApJS, 211, 8
Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. 2016, in Advances in Neural Information Processing Systems, Vol. 29, ed. D. Lee et al. (Red Hook, NY: Curran Associates, Inc.), https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf
Brünken, S., Gupta, H., Gottlieb, C. A., McCarthy, M. C., & Thaddeus, P. 2007, ApJL, 664, L43
Burkhardt, A. M., Lee, K. L. K., Changala, P. B., et al. 2021, ApJL, 913, L18
Cernicharo, J., Agúndez, M., Cabezas, C., et al. 2021, A&A, 647, L2
Cernicharo, J., Gottlieb, C. A., Guelin, M., et al. 1991, ApJL, 368, L39
Cernicharo, J., Marcelino, N., Agúndez, M., et al. 2020, A&A, 642, L17
Chai, J.-D., & Head-Gordon, M. 2008, PCCP, 10, 6615
Choi, Y., Lee, J.-E., Bourke, T. L., & Evans, N. J., II 2017, ApJS, 229, 38
Cleeves, L. I., Bergin, E. A., Öberg, K. I., et al. 2017, ApJL, 843, L3
Cordiner, M. A., Charnley, S. B., Kisiel, Z., McGuire, B. A., & Kuan, Y.-J. 2017, ApJ, 850, 187
Das, G., Das, T., Chowdhury, N., et al. 2021, Genomics, 113, 1129
David, L., Thakkar, A., Mercado, R., & Engkvist, O. 2020, J. Cheminf., 12, 56
Disk Dynamics Collaboration, Armitage, P. J., Bae, J., et al. 2020, arXiv:2009.04345
Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. 1996, in Proc. 9th Int. Conf. on Neural Information Processing Systems, NIPS'96 (Cambridge, MA: MIT Press), 155
Friedman, J. H. 2002, Comput. Stat. Data Anal., 38, 367
Garrod, R. T. 2013, ApJ, 765, 60
Garrod, R. T., Weaver, S. L. W., & Herbst, E. 2008, ApJ, 682, 283
Ginsburg, A., McGuire, B. A., Plambeck, R., et al. 2019, ApJ, 872, 54
Gong, Y., Henkel, C., Spezzano, S., et al. 2015, A&A, 574, A56
Gratier, P., Majumdar, L., Ohishi, M., et al. 2016, ApJS, 225, 25
Guzmán, V. V., Pety, J., Goicoechea, J. R., et al. 2015, ApJL, 800, L33
Herbst, E., & Klemperer, W. 1973, ApJ, 185, 505
Hirota, T., Ito, T., & Yamamoto, S. 2002, ApJ, 565, 359
Hirota, T., Yamamoto, S., Mikami, H., & Ohishi, M. 1998, ApJ, 503, 717
Jaeger, S., Fulle, S., & Turk, S. 2018, J. Chem. Inf. Model., 58, 27
Janet, J. P., Kulik, H. J., Morency, Y., & Caucci, M. K. 2020, Machine Learning in Chemistry, ACS In Focus (Washington, DC: American Chemical Society)
Kim, S., Chen, J., Cheng, T., et al. 2021, Nucleic Acids Res., 49, D1388
Kulik, H. J. 2020, WIREs Comput. Mol. Sci., 10, e1439
Landrum, G. A. 2020, RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org/
Langer, W. D., Velusamy, T., Kuiper, T. B. H., et al. 1997, ApJL, 480, L63
Lee, K. L., Loomis, R. A., Burkhardt, A. M., et al. 2021a, ApJL, 908, L11
Lee, K. L. K., Changala, P. B., Loomis, R. A., et al. 2021b, ApJL, 910, L2
Lee, K. L. K., & McCarthy, M. 2020, JPCA, 5, 898
Liaw, A., & Wiener, M. 2002, R News, 2, 18
Lis, D. C., Roueff, E., Gerin, M., et al. 2002, ApJL, 571, L55
Loomis, R. A., Burkhardt, A. M., Shingledecker, C. N., et al. 2021, NatAs, 5, 188
Lovas, F. J., Remijan, A. J., Hollis, J. M., Jewell, P. R., & Snyder, L. E. 2006, ApJL, 637, L37
MacLeod, J. M., Avery, L. W., & Broten, N. W. 1984, ApJL, 282, L89
Marcelino, N., Cernicharo, J., Agúndez, M., et al. 2007, ApJL, 665, L127
Mattioda, A. L., Hudgins, D. M., Boersma, C., et al. 2020, ApJS, 251, 22
McGuire, B. A. 2018, ApJS, 239, 17
McGuire, B. A., Burkhardt, A. M., Loomis, R. A., et al. 2020, ApJL, 900, L10
McGuire, B. A., Burkhardt, A. M., Shingledecker, C. N., et al. 2017, ApJL, 843, L28

McGuire, B. A., Carroll, P. B., Dollhopf, N. M., et al. 2015, ApJ, 812, 76
McGuire, B. A., Loomis, R. A., Burkhardt, A. M., et al. 2021, Sci, 371, 1265
McInnes, L., Healy, J., & Melville, J. 2020, arXiv:1802.03426
Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013, arXiv:1301.3781
Morgan, H. L. 1965, J. Chem. Doc., 5, 107
O'Boyle, N. M. 2012, J. Cheminf., 4, 22
O'Boyle, N. M., Banck, M., James, C. A., et al. 2011, J. Cheminf., 3, 33
Ohishi, M., McGonagle, D., Irvine, W. M., Yamamoto, S., & Saito, S. 1994, ApJL, 427, L51
Parrish, R. M., Burns, L. A., Smith, D. G. A., et al. 2017, J. Chem. Theory Comput., 13, 3185
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Pinte, C., Price, D. J., Ménard, F., et al. 2018, ApJL, 860, L13
Platt, J. C. 1999, Advances in Large Margin Classifiers (Cambridge, MA: MIT Press), 61
Pratap, P., Dickens, J. E., Snell, R. L., et al. 1997, ApJ, 486, 862
Ramakrishnan, R., Dral, P. O., Rupp, M., & Lilienfeld, O. A. v. 2014, Sci. Data, 1, 1
Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., & Skiff, W. M. 1992, JAChS, 114, 10024
Rassolov, V. A., Pople, J. A., Ratner, M. A., & Windus, T. L. 1998, JChPh, 109, 1223
Remijan, A. J., Hollis, J. M., Snyder, L. E., Jewell, P. R., & Lovas, F. J. 2006, ApJL, 643, L37
Ruaud, M., Wakelam, V., & Hersant, F. 2016, MNRAS, 459, 3756
Rupp, M., Tkatchenko, A., Müller, K.-R., & von Lilienfeld, O. A. 2012, PhRvL, 108, 058301
Sakai, N., Sakai, T., Osamura, Y., & Yamamoto, S. 2007, ApJL, 667, L65
Sakai, N., Takano, S., Sakai, T., et al. 2013, JPCA, 117, 9831
Schilke, P., Walmsley, C. M., Pineau des Forêts, G., & Flower, D. R. 1997, A&A, 321, 293
Shibayama, S., Marcou, G., Horvath, D., et al. 2020, Mol. Inf., 39, 1900170
Soma, T., Sakai, N., Watanabe, Y., & Yamamoto, S. 2018, ApJ, 854, 116
Sterling, T., & Irwin, J. J. 2015, J. Chem. Inf. Model., 55, 2324
van Dishoeck, E. F., & Black, J. H. 1986, ApJS, 62, 109
Wakelam, V., Loison, J.-C., Herbst, E., et al. 2015, ApJS, 217, 20
Wakelam, V., Smith, I. W. M., Herbst, E., et al. 2010, SSRv, 156, 13
Wang, L.-P., & Song, C. 2016, JChPh, 144, 214108
Wang, Y., Xiao, J., Suzek, T. O., et al. 2012, Nucleic Acids Res., 40, D400
Weininger, D. 1988, J. Chem. Inf. Comput. Sci., 28, 31
Xue, C., Willis, E. R., Loomis, R. A., et al. 2020, ApJL, 900, L9
Zheng, S., Yan, X., Yang, Y., & Xu, J. 2019, J. Chem. Inf. Model., 59, 914