Machine Learning Atmospheric Reaction Rates and its Astrobiological Implications

Havishk Tripathi

This report discusses the application of machine learning to predict the relationship between atmospheric chemical reactions and reaction rates. Atmospheric chemistry is an important topic in astrobiology, as it helps in understanding the chemical composition of exoplanets and the early atmosphere, which has implications for prebiotic chemistry and habitability. There are several atmospheric reaction rate datasets that have been analyzed using various modelling strategies, but much remains to be explored. Such modeling, is to some extent, limited by the lack of comprehensive chemical network models. Due to the large data set size, and the abundance of repetitive chemical reaction motifs, machine learning methods may be able to help fill in some of the gaps in atmospheric chemistry models, potentially revealing novel and important chemical phenomena. This report explains the databases explored, the strategies used to extract machine readable data from these databases and use it to produce expanded datasets using machine learning, and suggests where further work is still needed.

Introduction

What is the relevance of Atmospheric Chemistry to Astrobiology?

Understanding atmospheric chemical network kinetics affect the overall composition of planetary atmospheres as well as the surface rainout rate of important prebiotic species is key to creating accurate exoplanetary models. Physical experimentation in atmospheric chemistry is the most important source of data but is also cumbersome and time consuming. “At the molecular level, as computational methods allow for increasingly complex chemistry to be studied computationally, collaboration between laboratory and computational chemists are expected to become more common.”(Burkholder, 1) This paper notes the challenges associated with atmospheric chemistry lab experiments and the importance of exploring computational solutions to atmospheric chemistry questions.

Creating a cohesive, complete chemical network model is also challenging. Atmospheric chemical reaction networks may be intricately sensitive to overall network topologies and the kinetics of seemingly minor reactions. Nevertheless, understanding the types of chemical species atmospheric chemistry can deposit on planetary surfaces, and the rate with which it does so, can allow for better understanding of the types of chemistry which can occur in primitive surface waters on early Earth and exoplanets.

Recent attempts have been made to model exoplanetary atmospheres. Hobbs also uses the same reaction network to generate a chemical network, but do not have a similar approach. The key difference is, we have conducted a novel exploratory application of machine learning techniques to atmospheric chemical network generation. We have specifically tried to understand if the kinetics of reactions not present in reaction databases can be predicted based on similarity to those reactions which are in databases.

We hope to apply this knowledge to test model datasets for various planetary bodies, such as Titan, to either validate chemical reactions that we may not be able to conduct physically or simply explore novel chemistries. The penultimate goal is the development of a practical system for benchmarking, validation and exploratory analysis, to help improve understanding of exoplanetary atmospheric compositions and prebiotic chemistry.

Sources of Data

There are a few standard atmospheric chemistry Databases that contain kinetic data, for example the KIDA

Database, the JPL Data Evaluation, STAND2019 and the NIST Kinetic Database.

Data processing is the most important step doing machine learning and creating neural networks. The issue with most of these databases is that they are not in a format readily useful for machine learning.

The machine learning application discussed in this paper is based on the STAND2019 network provided in Rimmer's "Hydrogen Cyanide in Nitrogen-Rich Atmospheres of Rocky Exoplanets." STAND2019 has hand selected reactions relevant to atmospheric chemistry of astrobiological relevance. Each reaction has 3 kinetic parameters that, in conjunction with a selected temperature, can deliver a reaction rate for nearly 6000 gas-phase reactions. This is preliminarily fed into the machine learning code for further analysis.

Cleaning and parsing the data accurately for the first large dataset proved somewhat easy and provided a dataset for further method development. After creating a method for reading and formatting such databases, new datasets or databases can also be quickly fed into the machine learning algorithms for processing.

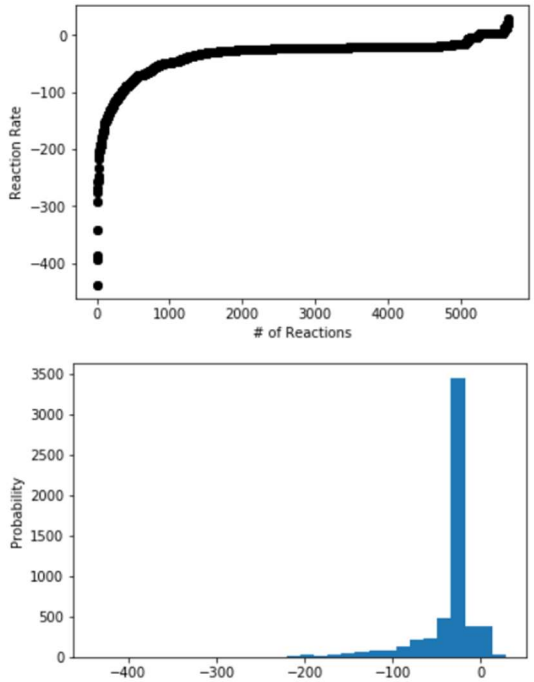


Figure 1: STAND 2019 Reaction Rate Histogram.

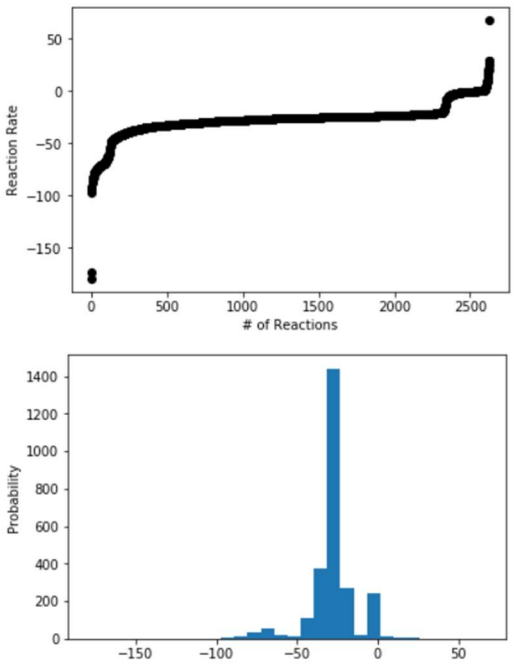


Figure 2: NIST Database Reaction Rate

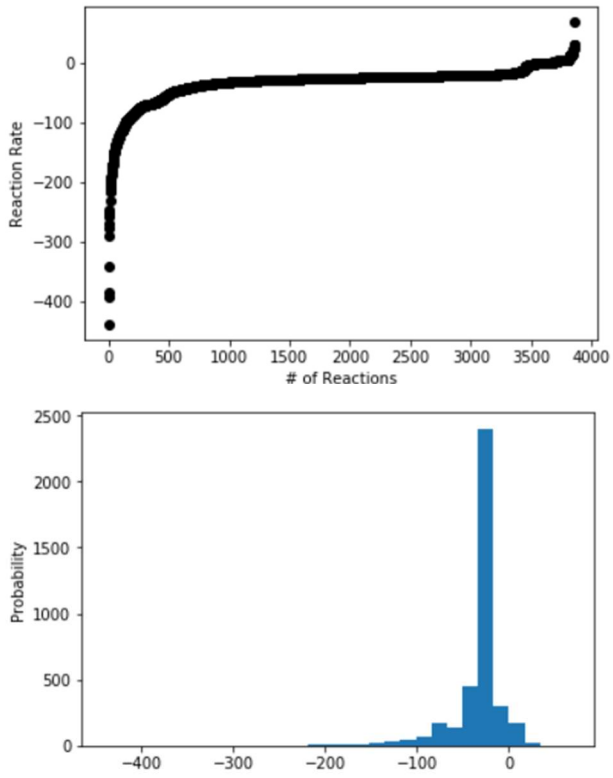


Figure 3: This is a combination of both datasets with only unique reactions.

Figures 1,2 and 3 describe important differences between the evaluated datasets. STAND2019 has more datapoints and has a wider range of reactions. The NIST database, doesn't provide as much data or such a high variance in evaluated data. As we train the neural network process, it's important to remain within a valid basis for the prediction dataset. The combined dataset is the maximal amount of unique information based on curve fitting. Avoiding extrapolation or under-fitting, is important based on how the training data.

Understanding how the Code Works

Machine learning is the application of statistical analysis through inputting preprocessed data.

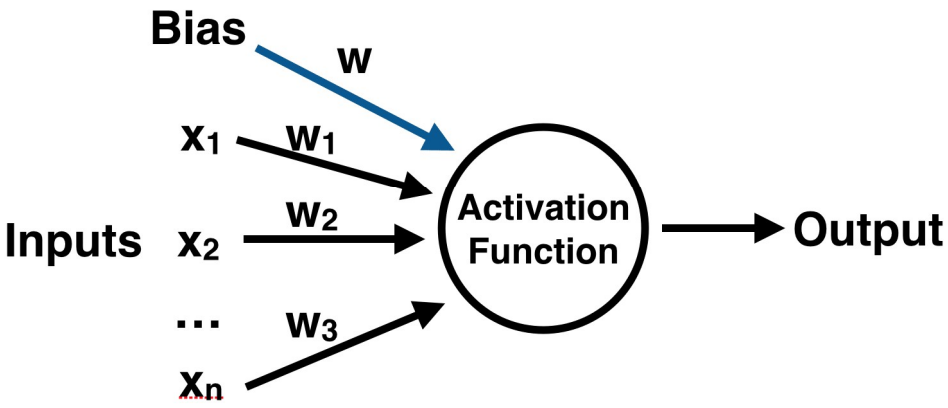


Diagram 1: The mathematical representation behind a neural network.

Figure 3 displays the concept of how a perceptron in a neural network works.
The input layer, in the context of the neural network, has a set of neurons that the network trains a weighted function on. The training is also give a activation function, that is user-defined based on choice. As more points of data are provided to the input layer, the weighted function keeps adding output values into hidden layers, and keep transforming calculating through the hidden layers until you receive a final output function.

The feautures scaling provided to the neural network are very sensitive to changes in the network. It's importtant tot both utlizie and optimiz the features inherent to the mathimatical bias of the neural network, because it has a direct correlatino to the output. , the more nuanced and descriptive of the data the features vector is, the more information the machine learning network can glean from the data.

At the end of the data processing using machine learning, neural networks can be used to test the output against validation data, or to predict new test data.

The first step to any machine learning process is understanding datasets and where they are coming from. Providing the correct features to train the neural network is key to accurate model generation.
The. We import the CSV data into Python, and cleaning the data so that the array features reads [R1.R2.R3.P1.P2.P3.P4.Reaction Rate]. This is standardized between both datasets, befure any mathematical analysis is applied.

We used the features of the STAND2019 database to calculate reaction rates for random reactions selected from the test data. The ultimate goal was to be able to use the predictive model to estimate the reaction rates of reactions not present in the model.

For the most comprehensive analysis, RDKIT was utilized, which is software that allows calculation of chemical features useful for training the chemical network. The data must be provided in a format known as SMILES, which is a standardized language for describing unique chemical compund structure.

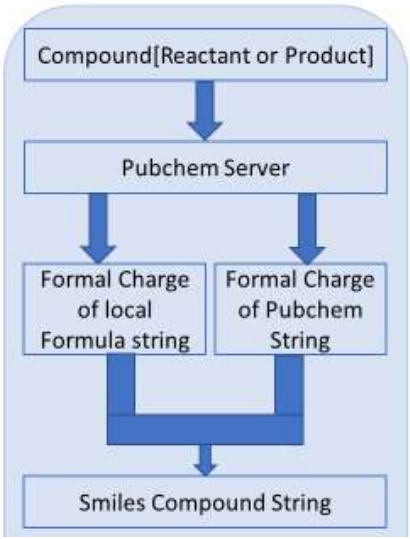


Diagram 2:The process flow diagram for Compound to Smiles String Conversion

We identified all unique compounds in the dataset, assigned them a SMILES string where possible, and remapped this information back to our features vector. If there is a unique circumstance [where either the electronic state is changing, or orbital chemistry takes place], SMILES is sometimes unable

to capture that. Specific compounds, such as XX or YY, do not have a direct SMILES format as they are just an expression of a changing electronic state. Wherever applicable, the closest smiles string was applied that was characteristic of the compound at hand. Certain compounds do not have STRINGS representations because of having a different electronic state, which SMILEs currently doesn't supoprt.

To understand the accuracy of the predictions the neural network created, we looked at the R^2 score in order to determine accuracy. The R^2 term is the *Coefficient of Determination*, which describes the variation between the leanred and provided data.

An R^2 value of 1 means the neural network is predicting at a 100% accuracy, while a score of 0 indicates no correlation. The secondary graph shown on each learning set, demonstrates the ability of each neural network to learn per iteration. This information is valuable since it depicts a trend between data to the number of iterations required.

Methods to Develop the Features Vector:

Method 1: Counting Reactants and Products:

The simplest approach involved counting the number of product and reactant species in each reaction and scoring that versus the reaction rate for that reaction. When we just run a features vector with those two values (number of products and number of reactants), the results shown in Figure 3 are obtained. It is important to note that we lose some rows of the NIST dataset because many reactions have the products of misc., which can be 1 or greater than 1. Since we can't know for sure, it's omitted from the training set.

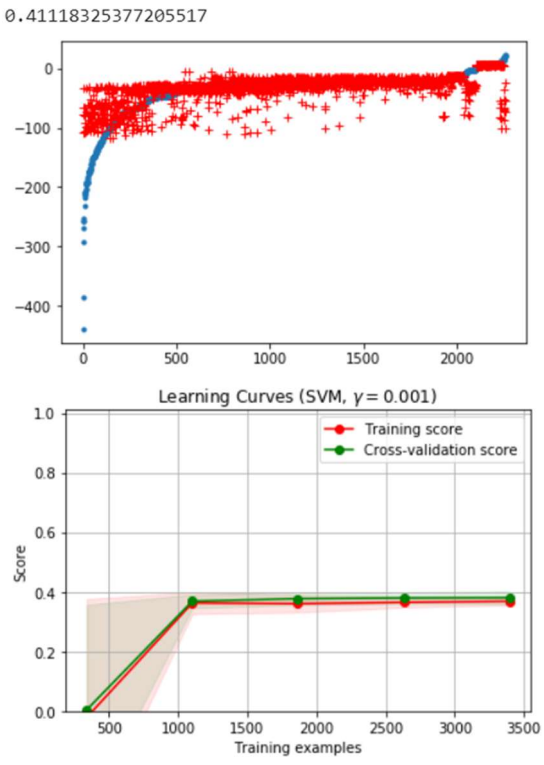


Figure 4: STAND2019 Dataset with neural network trained on reactant and product count.

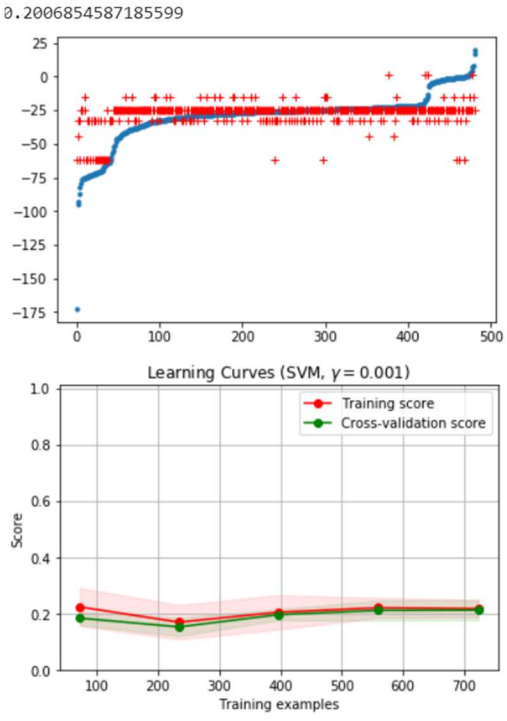


Figure 5: NIST Dataset with neural network trained on reactant and product count.

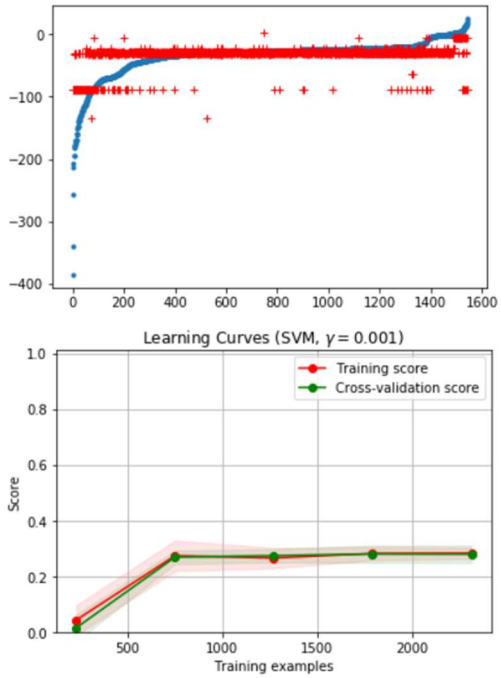


Figure 6: Combined Dataset with neural network trained on reactant and product count.

Method 2: Species Count

This method counts of the number of atoms in each species. There are only a few elements in rotation for each of the compounds. We create a 1x17 vector that has positional intracity based on the element. For example, C2H would be interpreted as [2.1.0.0.0.etc…], where each position represents of carbon, hydrogen, and oxygen, and so on. We consider 17 different elements, based on the distribution of molecules. This conversion is done for all the chemical reactants and is fed as the features vector.

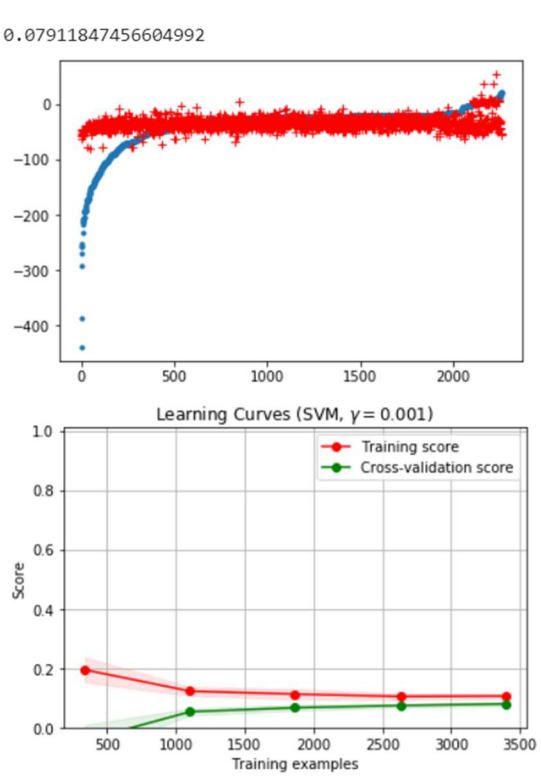


Figure 7: STAND Dataset with neural network trained on Species count.

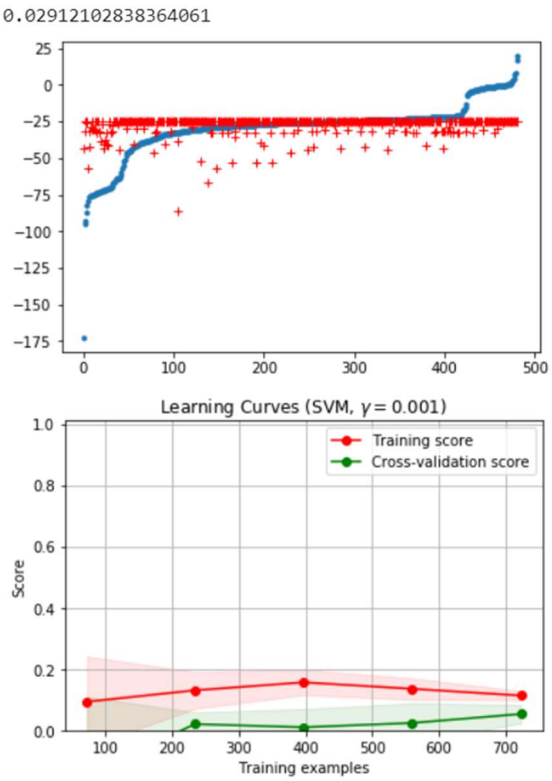


Figure 8: NIST Dataset with neural network trained on Species count.

Method 3: Counting Molecular Weight [Self Definition vs RDKit]:

The definition begins from using Method 2: that was written outside of package checked each compound entry, and based on the entry format, e.g. C2H2 and C2H4 were read as [2,C,2,H] and [2,C,4,H], respectively. Every compound's molecular weight was determined by summing the products of the atomic weights of atom types and their counts in that molecule.

Using RDKit, the same concept is applied. The SMILES string format is fed into a function defined in the package, and we can easily develop a more accurate molecular weight.

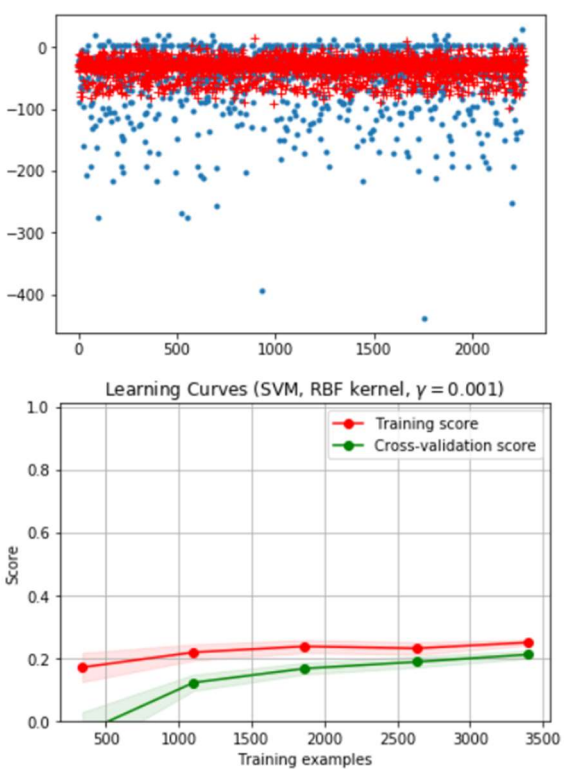


Figure 9: STAND Dataset with neural network trained on Species count.

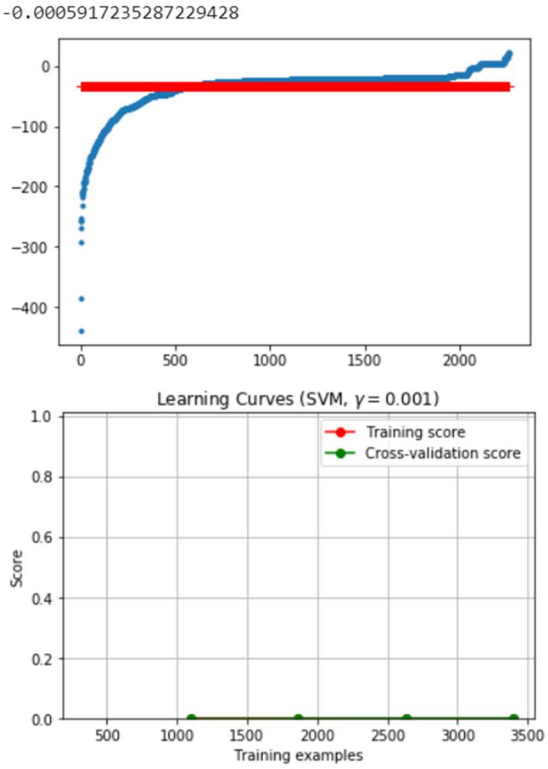
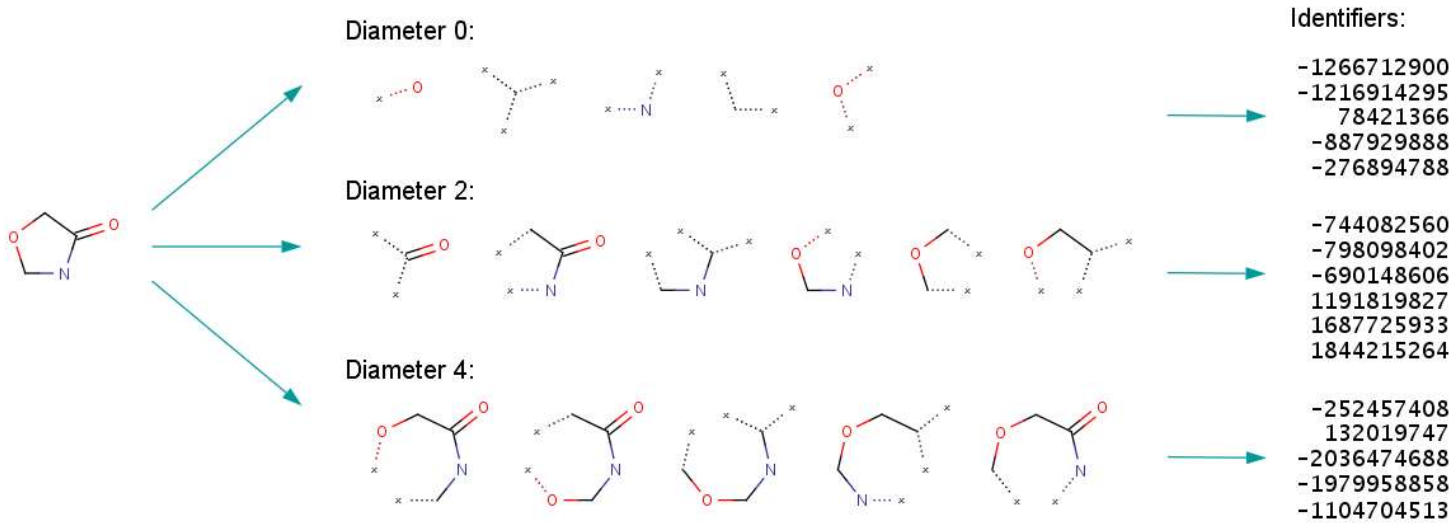


Figure 10: NIST Dataset with neural network trained on Species count.

Theoretical Method 4: Morgan Fingerprints:

Morgan fingerprints are an impression based on the molecularity of the compound. There is an algorithm that calculates the relationship between the pairing of molecules. As we increase the radius of the Morgan fingerprint, we can achieve a higher resolution of the molecule at hand and their relevant neighbors.



Discussion

For the scope of this paper, we simply need to understand the relationship between reactants, products, and reaction rates, in atmospheric chemistries. We apply neural net technologies to be able to predict reaction rates. We can use this ability

in varying contexts, especially for useful and practical astrobiological applications. A big reason why Titan is explored, and it has interested many people over the years, is simply its ability to form complicated hydrocarbons and its presence of methane, in comparison to other celestial bodies. While many models attempt to expand and derive the nature of Titan's atmosphere, without knowing extensive reaction networks or without overlooking niches reactions that may be occurring, it becomes more about repainting the same image on a different canvas, versus making a clearer picture that dignifies new aspects.

Given error in our results, it is worth further understanding the analysis of the cost of time versus value in pursuing this neural network further. As we can develop a more structured and feasible network, especially for more recent missions like Europa or Titan, is it worth edifying and redefining a neural network where the predictive score can be much higher. However, limited the dataset may be, the value in its interpretation is still abundant. More than that, the cost of understanding or providing a base level search that other can build a more complex neural network, has its own scientific merit. Atmospheric chemistry has its own challenges and caveats that need personalized attention, and the application of recent technologies in this context has been previously unexplored. While it proves difficult to merge these two contexts, the value of using neural net technology, is very practical and applicable. From generating either a reaction rate classifier to apply to a moon like Titan or Europa, or to see and predict atmospheric chemistries specific to those bodies, providing a new scope on underutilized data is a strong selling point.

Bibliography

1. Burkholder, J. The Essential Role for Laboratory Studies in Atmospheric Chemistry. *American Chemical Society* **2017** DOI: https://doi.org/10.1021/acs.est.6b04947.
2. Rimmer, P.; Rugheimer, S. Hydrogen cyanide in nitrogen-rich atmospheres of rocky exoplanets. *Icarus* **2019**, *329*, 124–131 DOI: 10.1016/j.icarus.2019.02.020.

Acknowledgements

This research opportunity was sponsored by the NASA Astrobiology Institute, as well as ELSI, the Earth Life Science Institute