## INTRODUCTION

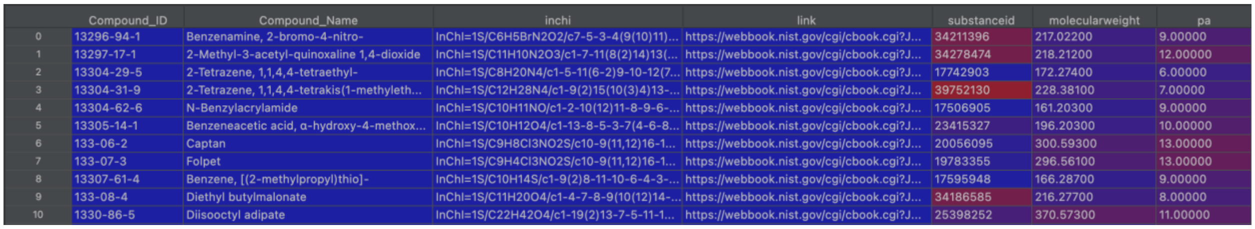
To date, there is no definition of life that is universally accepted among scientists. There have been conflicting opinions on the overall importance of various characteristics (e.g. metabolism, Darwinian evolution, self-reproduction, etc.) common among living systems. Using a strict definition tends to place complex biomolecules with many ‘living’ characteristics (e.g. viruses) into an ambiguous classification between abiotic and biotic [1]. Many have suggested treating using a non-binary classification system of life, with different clusters of organisms having different ‘living’ characteristics, however this complicates the ability to test for signatures of living systems [1]. Therefore, developing methods of testing for living systems that are standardized and non-arbitrary remains a desirable goal for scientists aiming on discovering extraterrestrial biotic systems.

## RELATED WORK

Marshall S.M., *et al.* [2] develop a framework, Molecular Assembly (MA), that proposes extremely complex molecules that are driven to a high copy-number (i.e. out-of-equilibrium) are unique to living systems and, thus, make strong candidates as biosignatures for detecting biotic samples. Molecular assembly derives from assembly pathways, which are joining operations that start with basic building blocks and end with a final product (i.e. a molecule), during which subunits generated within the assembly pathway can combine with other subunits or basic components later within the pathway to generate complex structures [2]. Their framework uses tandem mass spectrometry (MS/MS) to differentiate various ions and calculate their overall complexity based on the number of associated spectral peaks, and the ion with the highest MA being representative of the MA of the sample [2]. However, due to the complexity of MS/MS data, I will be focusing on a theoretical model they generated and published in their supplemental material which uses a Convolutional Neural Network (CNN) to correlate single-MS spectra with the molecule's MA.

## DATASET AND FEATURES

The dataset for this project relies on a .pickle file (data frame) published in the supplemental information for Marshall S.M., *et al.*, which has the calculated MA for 2.5 million compounds. Additionally, we collected the total number compounds identified in the NIST chemistry webbook [3] based on their "species list" [4], which were then filtered for molecules containing a webpage, and further filtered for molecules with associated MS data. Once all possible elements were identified, we combined the two dataframes based on their InChi key (Figure 1). Once a full list of possible datapoints was extracted from NIST, we ended up collecting ~1500 total spectral JDX-CAMP files to re-create images in a black and white figure (Figure 2).



| | Compound_ID | Compound_Name | inchi | link | substanceid | molecularweight | pa |
|---|---|---|---|---|---|---|---|
| 0 | 13296-94-1 | Benzenamine, 2-bromo-4-nitro- | InChI=1S/C6H5BrN2O2/c7-5-3-4(9(10)11)... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 34211396 | 217.02200 | 9.00000 |
| 1 | 13297-17-1 | 2-Methyl-3-acetyl-quinoxaline 1,4-dioxide | InChI=1S/C11H10N2O3/c1-7-11(8(2)14)13(... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 34278474 | 218.21200 | 12.00000 |
| 2 | 13304-29-5 | 2-Tetrazene, 1,1,4,4-tetraethyl- | InChI=1S/C8H20N4/c1-5-11(6-2)9-10-12(7... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 17742903 | 172.27400 | 6.00000 |
| 3 | 13304-31-9 | 2-Tetrazene, 1,1,4,4-tetrakis(1-methyleth... | InChI=1S/C12H28N4/c1-9(2)15(10(3)4)13-... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 39752130 | 228.38100 | 7.00000 |
| 4 | 13304-62-6 | N-Benzylacrylamide | InChI=1S/C10H11NO/c1-2-10(12)11-8-9-6-... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 17506905 | 161.20300 | 9.00000 |
| 5 | 13305-14-1 | Benzeneacetic acid, α-hydroxy-4-methox... | InChI=1S/C10H12O4/c1-13-8-5-3-7(4-6-8... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 23415327 | 196.20300 | 10.00000 |
| 6 | 133-06-2 | Captan | InChI=1S/C9H8Cl3NO2S/c10-9(11,12)16-1... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 20056095 | 300.59300 | 13.00000 |
| 7 | 133-07-3 | Folpet | InChI=1S/C9H4Cl3NO2S/c10-9(11,12)16-1... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 19783355 | 296.56100 | 13.00000 |
| 8 | 13307-61-4 | Benzene, [(2-methylpropyl)thio]- | InChI=1S/C10H14S/c1-9(2)8-11-10-6-4-3-... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 17595948 | 166.28700 | 9.00000 |
| 9 | 133-08-4 | Diethyl butylmalonate | InChI=1S/C11H20O4/c1-4-7-8-9(10(12)14-... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 34186585 | 216.27700 | 8.00000 |
| 10 | 1330-86-5 | Diisooctyl adipate | InChI=1S/C22H42O4/c1-19(2)13-7-5-11-1... | https://webbook.nist.gov/cgi/cbook.cgi?J... | 25398252 | 370.57300 | 11.00000 |

**Figure 1**. First 11 compounds of the labelled dataset. 'Compound_ID' is the unique identifier of the molecule in the NIST webbook, 'Compound_Name' is the common name of the molecule, 'InChi' is a unique formula, 'link' contains information on where to download the MS data for the molecule, 'substanceid' is a unique identifier from the original paper, 'molecularweight' is the weight of the object, and 'pa' is the MA index as calculated by the authors.

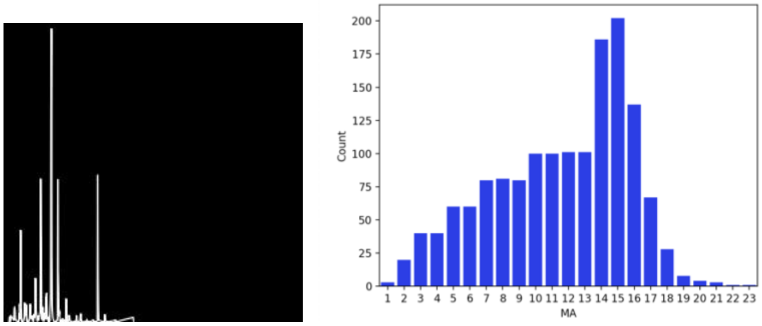


**Figure 2. (Left)** An example MS spectrum which was created using the JDX-CAMP files contained in the 'link' in figure 1. **(Right)** Distribution of downloaded spectral files and their respective MA.

## METHODS

After all data was collected, spectral images were organized into training and validation directories, as well as label directories, such that the name of their host directory was the associated label (i.e. MA index). These datasets were then fed into a CNN model using the VGG16 convolutional layer (without the fully connected layers), with a dense layer using 512 neurons with a .5 dropout rate and 'relu' activation, then 256 neurons with .5 dropout and 'relu' activation, and then a layer with the number of classes with 'sigmoid or 'softmax' activation. The model used the 'adam' or 'SGD' optimizer, while the loss varied from: 'categorical_crossentropy'; 'binary_categorical_crossentropy'; or 'mean_absolute_error'. The variation in activation functions, optimizers, and loss functions stems from testing which methods would result in the highest accuracy of the CNN model in differentiation between the spectral images.

We initially organized it such that only files with a MA >15 were included in order to test the practicality of the model. After initial positive results, we expanded the range of available to data to include MA 1-23. However, due to several failed attempts to create an accurate model, this was simplified to binary categories of MA <15, which can be equated to 'non-living' samples, and samples with MA >=15, which can be equated to 'living' samples. We additionally created a model which classified the groups into 3 categories: low complexity (MA 1-8); medium complexity (9-14); and high complexity (15+).

## RESULTS

The initial runs using samples with MA >=15, which had 10 total categories, were only trained shortly (5 epochs), and had moderate success in categorizing the groups with an accuracy if .276 (Figure 3L). We thus expanded this to 23 groups and had initial success with this as well using a small dataset (Figure 3R). However, once we expanded to full scale (~1500 samples) we found that the model did failed to accurately predict different classes resulting in it predicting the same class for all samples resulting in an accuracy of .13 (Figure 4). Simplifying the overall model to have less classes seems to create a stronger classifying model, with the lightly trained 3-class model having an accuracy of .63 (Figure 5L), while the binary class ('living' vs 'non-living') had an accuracy of .79 (Figure 6)

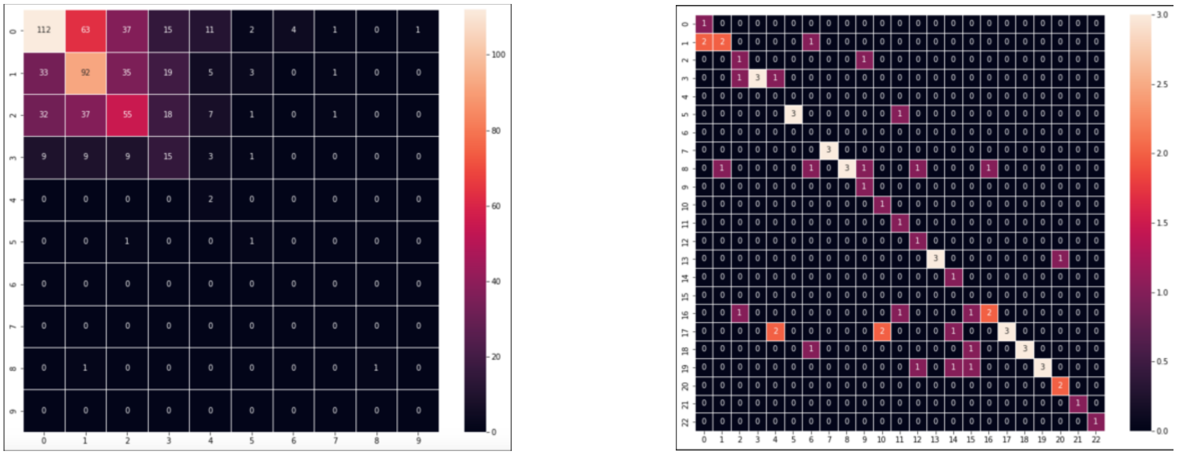
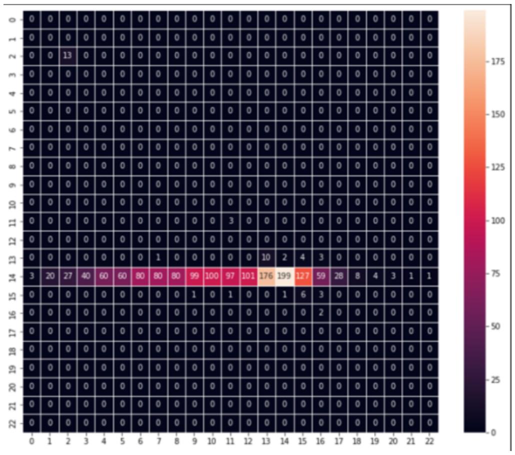
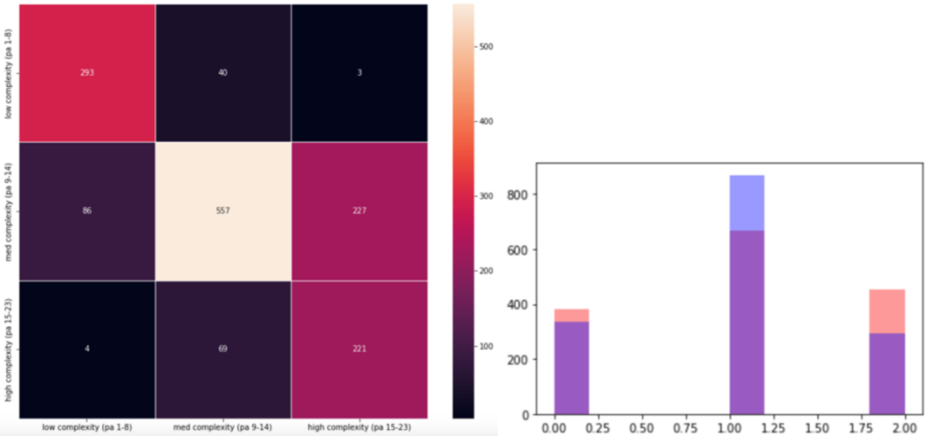


**Figure 3. (Left)** Confusion matrix including samples with MA >=15 (Axes shifted -15) (Predictions on left axis, true labels on bottom axis) **(Right)** Confusion matrix with 23 classes (Axes shifted -1) using a small sized training batch (~3 samples per class) with a diagonal patterned result that indicates a correctly working model.



**(Left) Figure 4.** Confusion matrix for expanded model with ~1500 samples. Nearly all predictions were of MA==15.

**(Below) Figure 5. (Left)** Confusion matrix of 3 classification system using low (1-8), med (9-14), and high (15-23) complexity. **(Right)** Distribution of predicted values (red) and labels (blue).



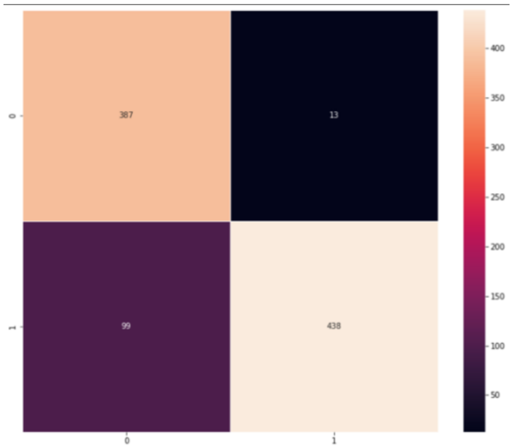


**Figure 6.** Confusion matrix of binary classification model. “0” refers to samples with MA below 15 (range from 10-14) or non-living samples, while “1” refers to samples with MA of 15 or greater.

## DISCUSSION & FUTURE DIRECTIONS

Ultimately, the goal of creating a model that could accurately predict a molecules MA index based on its spectra did not pan out at the full scale. This is likely partly due to high overlap between neighboring values, as well as not cropping images to reduce empty space. Further, higher training time may increase overall accuracy at the full scale, but it remains unclear. Lastly, low samples of higher MA molecules made it difficult to make a generalized model. That said, once the number of classes the model was required to predict was reduced, it was able to become more succesful. The lightly trained 3-class model (Fig. 5) only had 4 low-complexity molecules called as high complexity, which likely could have been reduced with higher training. This is ability to differentiate general patterns is additionally shown in the binary classification model, which uses a MA threshold of 15, as Marshall S.M., *et al.*, found no samples derived from non-living material to have a complexity higher than that. This model had an .8 accuracy in differentiation between the 2 classes, which could be improved upon with further training in addition to more higher complexity molecule data. Ultimately, if I wanted to improve the model, I would start with re-sizing the x-axis of the spectra to better capture the peak ranges, in addition to resizing the line width. An additional step would be to increase the weights of higher MA molecules.

# CITATIONS

[1] C. Mariscal and W. F. Doolittle, “Life and life only: a radical alternative to life definitionism,” *Synthese*, vol. 197, no. 7, pp. 2975–2989, Jul. 2020, doi: 10.1007/s11229-018-1852-2.

[2] S. M. Marshall *et al.*, “Identifying molecules as biosignatures with assembly theory and mass spectrometry,” *Nat Commun*, vol. 12, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41467-021-23258-x.

[3] N. O. of D. and Informatics, “NIST Chemistry WebBook.” https://webbook.nist.gov/chemistry/ (accessed May 02, 2023).

[4] “Species List.” https://webbook.nist.gov/chemistry/download/ (accessed May 02, 2023).