Use of Computational Neural Networks for Rapid Bacterial Strain Identification Using Deinococcus Aquaticus Isolates Obtained from Biofilm Samples

David Ayodele, Stacy Scholz-Ng, Chad Albert and James Tuohy
Glendale Community College, Glendale, AZ. April 2018

ABSTRACT

Convolutional Neural Networks (CNNs) have been applied to a wide-rage of problems in recent years[1,7]. In simple terms, these algorithms can be thought of as curve-fitting schemes, designed to take some input (an unknown curve) and generate an appropriate mathematical model for it. With the model, a computer would then be able to predict the output of future curve inputs. CNNs can be broadly viewed as extending the curve-fitting strategy to many more dimensions[1,7]. MALDI-TOF (Matrix-assisted laser desorption/ionization – Time of Flight) mass spectrometry is a spectroscopic technique commonly used in the analysis of whole cell protein extracts[2,5]. MALDI-TOF has been shown to reliably produce summary spectra of bacterial proteins resulting in a characteristic fingerprint for a given species[2,5]. In this investigation, we use a CNN algorithm written in Python (Computer Language) to model changes in MALDI-TOF spectra with respect to D. Aquaticus strains in an effort to arrive at a model for predicting the identity an unknown strain using its MALDI-TOF spectrum.

INTRODUCTION

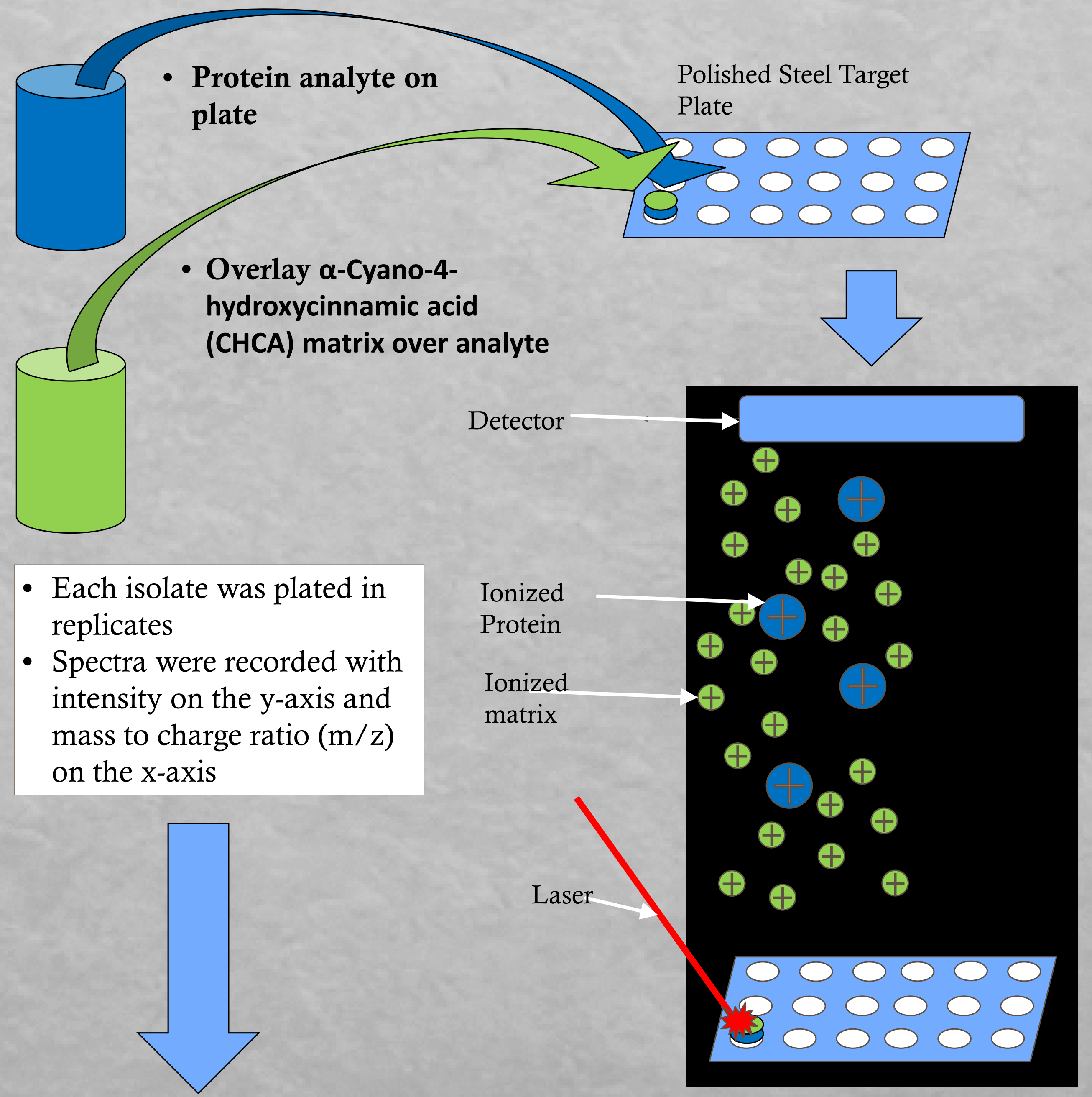
Deinococcus Aquaticus is one of the more studied members of the genus *Deinococcus*. The genus is well-known for being highly resistant to adverse environmental forces such as ionizing radiation, oxidation and desiccation[4]. *Deinococci* have been isolated from environments as dry as deserts, as temperate as hot springs, and as nutrient deficient as steel milling machinery[5]. MALDI-TOF spectrometry has been shown to be an effective method for rapid identification of bacteria when employed with a variety of sample preparation methods[5,6]. We aim to present here a method for much faster identification using machine learning algorithms in conjunction with MALDI-TOF, a technique which may be dubbed MAMALDI-TOF (Machine-learning Algorithm based Matrix-assisted laser desorption/ionization – Time of Flight) mass spectrometry. The algorithms used were written in the Python 3 programming language and make use of machine learning libraries such as Tensor Flow and Keras. These libraries greatly accelerate the processing speed of several computationally intensive steps commonly used in neural network algorithms. Nearly all such algorithms use some variant of gradient descent logic in that they begin with a random approximation of a feature of interest, compute the error in that approximation, compute the gradient of that error (the change in the error with respect to its inputs), and then compute the "direction" of the next approximation. Through use of this generalized gradient descent technique, these algorithms achieve successive reductions in the error of their initial predictions until a desired accuracy level is reached.

METHODS

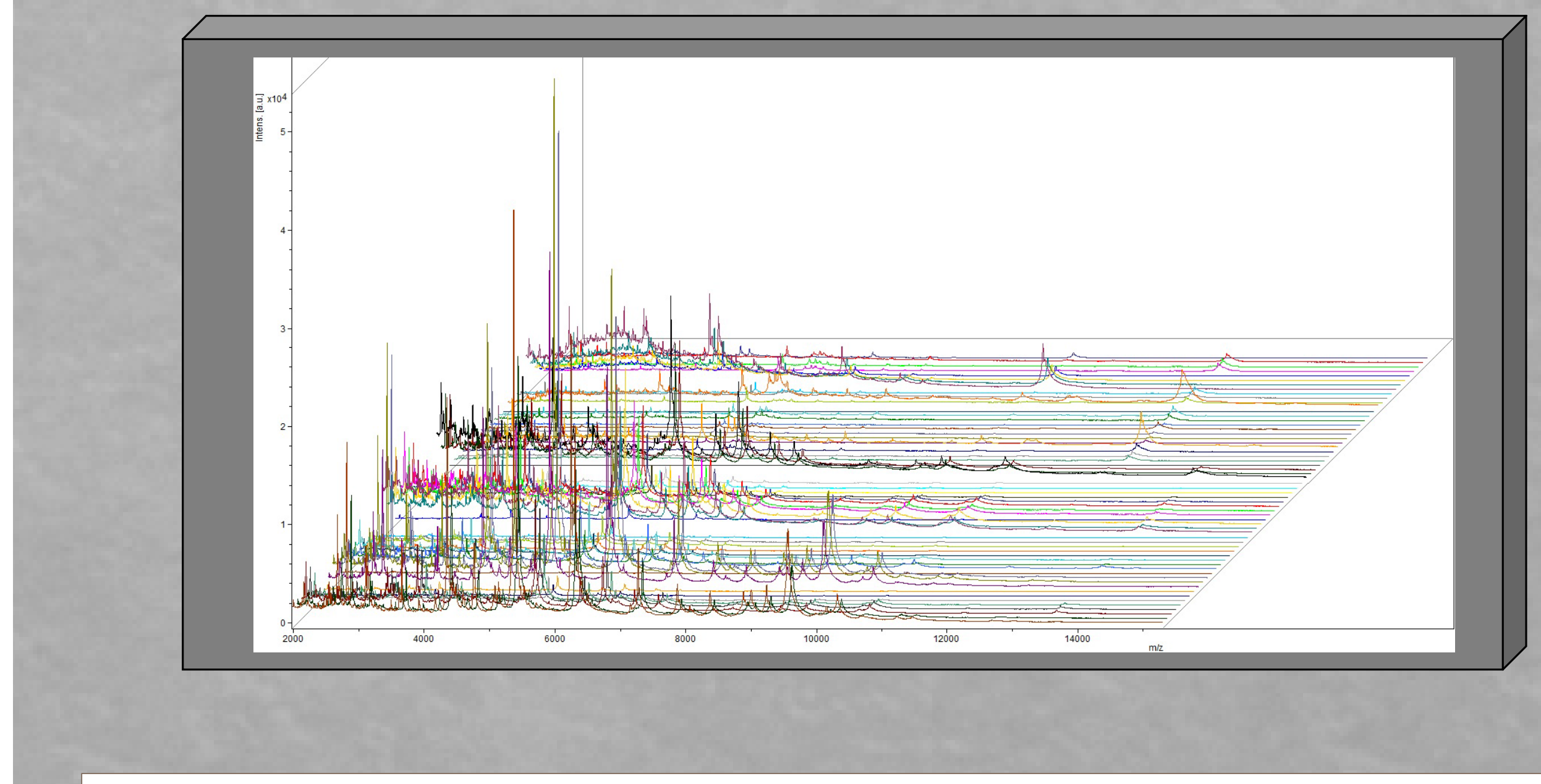
MALDI-TOF Mass Spec

Whole cell protein extraction was conducted for D. Aquaticus

- Spectral data was read into a Jupyter/IPython Notebook Development Environment
- Tensor Flow and Keras libraries were used to build a CNN model for the training data.
- Parameters of the model include ReLU as the activation function, categorical cross-entropy as the loss function, and adam as the optimizer


Protein analyte on plate
Polished Steel Target Plate
Overlay α-Cyano-4-hydroxycinnamic acid (CHCA) matrix over analyte
Detector
Ionized Protein
Ionized matrix
Laser

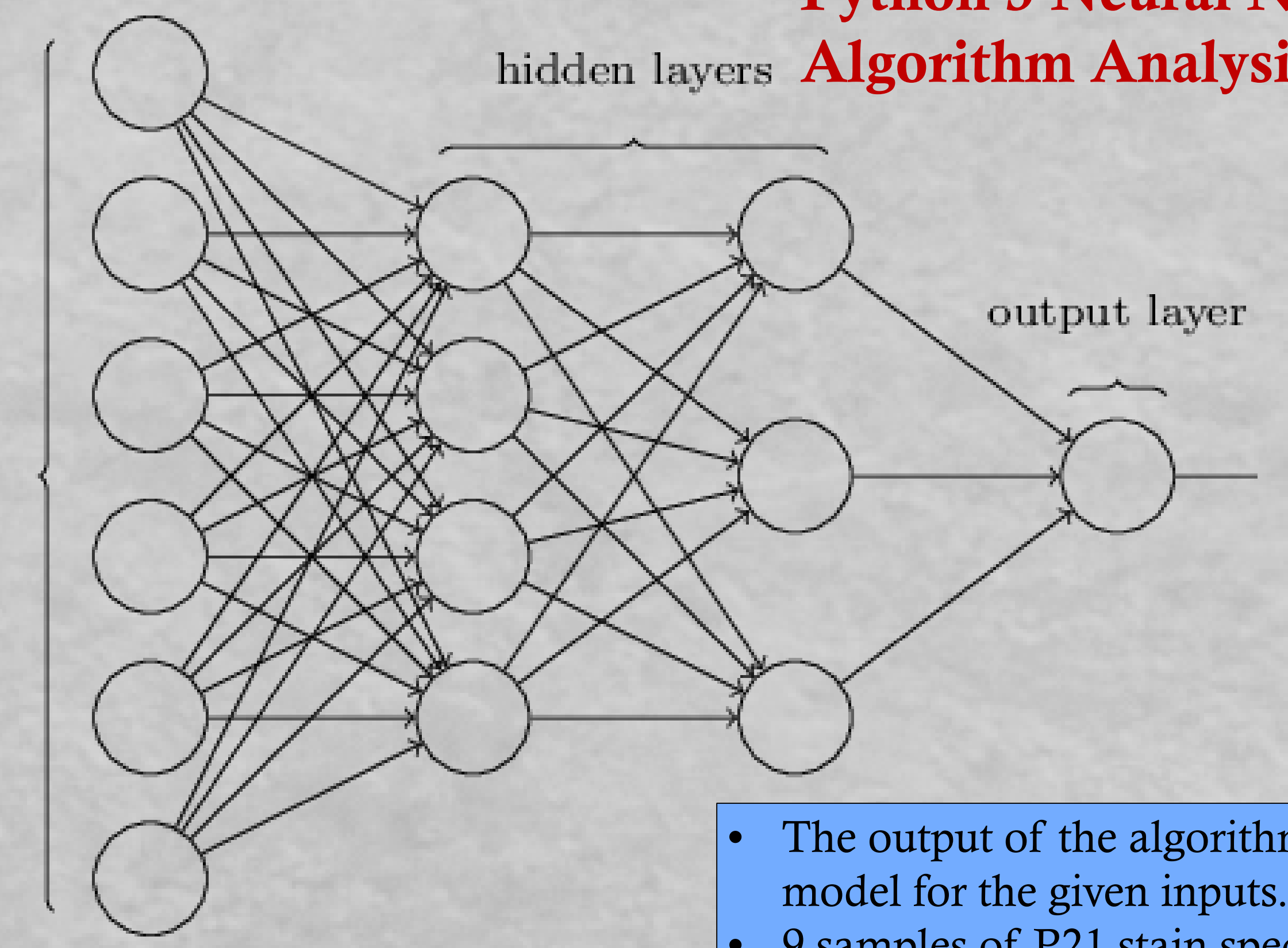
- Each isolate was plated in replicates
- Spectra were recorded with intensity on the y-axis and mass to charge ratio (m/z) on the x-axis



- Spectral data was exported to txt files for analysis
- Each raw spectrum was exported unprocessed to minimize contributing variables
- Spectra were then also analyzed with Bionumerics® and commonly used software for comparison with our own algorithm's results.

Spectral data of interest

Python 3 Neural Network Algorithm Analysis



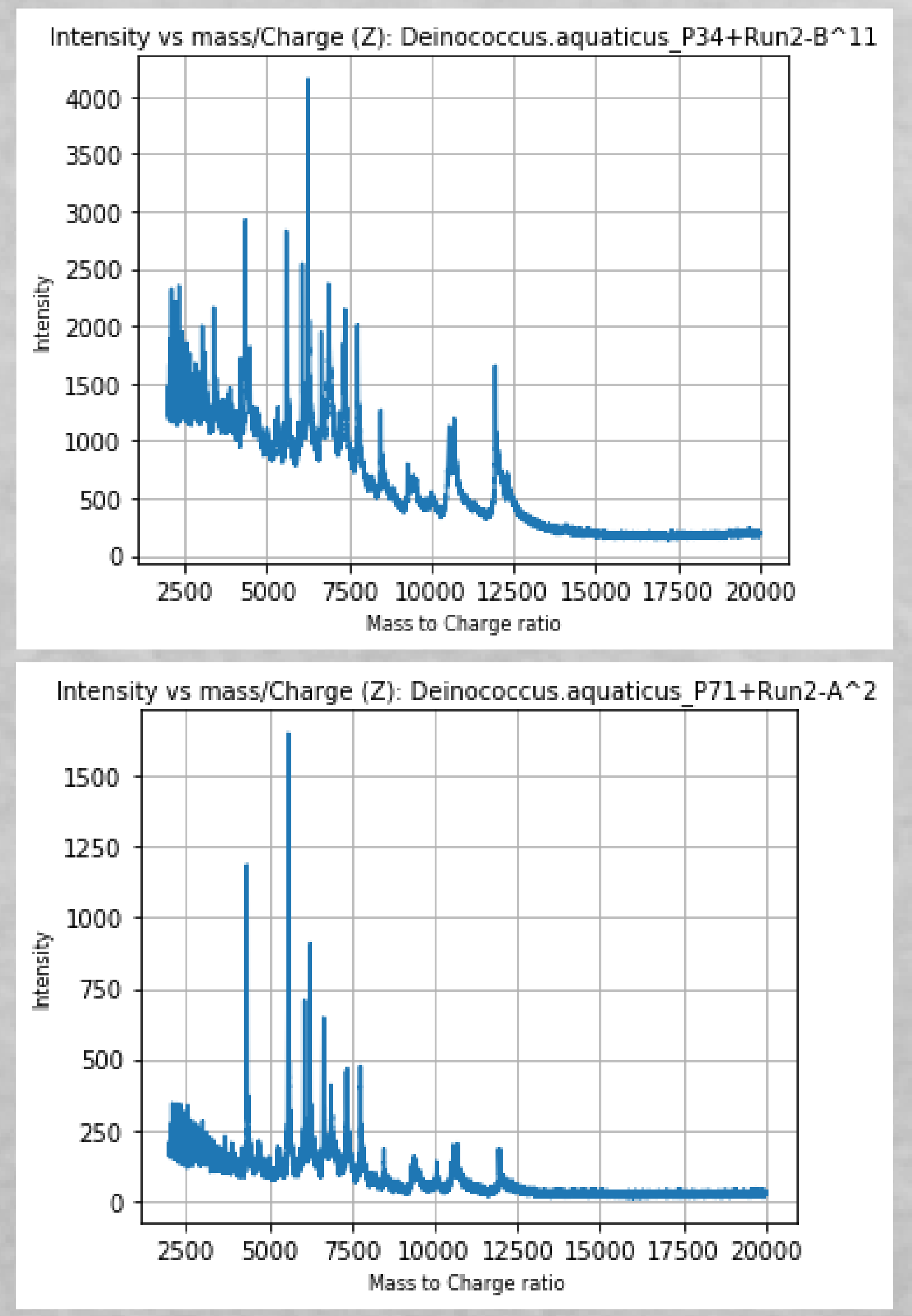
- The output of the algorithm is a mathematical model for the given inputs.
- 9 samples of P21 stain spectra were used to train the algorithm.

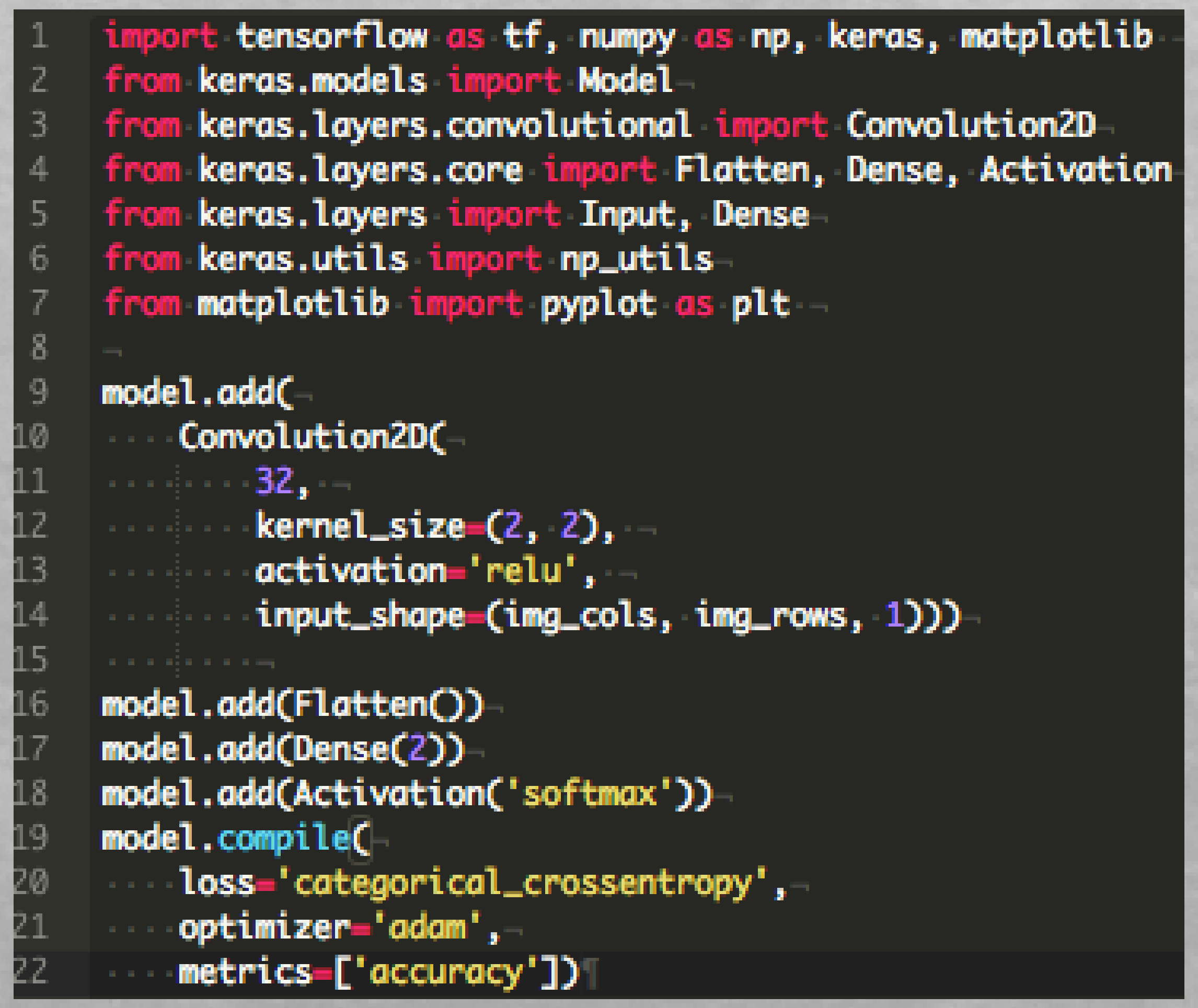
Spectral data for an unknown strain of D. Aquaticus was then read in.

- A numeric label was assigned to the training set data to distinguish it from non-P21 spectra.
- The dimensionality of the unknown data was augmented to match that of the training set data.
- The prediction function was executed using the unknown spectrum as its input.
- The function produced an output indicating the degree of confidence that the unknown spectrum was derived from D. Aquaticus P21.

RESULTS

Spectra generated with raw data using Python 3 algorithm





```python
import tensorflow as tf, numpy as np, keras, matplotlib
from keras.models import Model
from keras.layers.convolutional import Convolution2D
from keras.layers.core import Flatten, Dense, Activation
from keras.layers import Input, Dense
from keras.utils import np_utils
from matplotlib import pyplot as plt

model.add(
    Convolution2D(
        32,
        kernel_size=(2, 2),
        activation='relu',
        input_shape=(img_cols, img_rows, 1)))

model.add(Flatten())
model.add(Dense(2))
model.add(Activation('softmax'))
model.compile(
    loss='categorical_crossentropy',
    optimizer='adam',
    metrics=['accuracy'])
```

Snippet of Python 3 algorithm showing key parameters

CONCLUSION

- MALDI-TOF Mass Spectrometry provides a rapid and reliable characterization method for *Deinococcus Aquaticus* bacteria
- Our Python 3 algorithm produced unadjusted spectra that appeared to correlate with that of Bionumerics and other commonly used software with MALDI-TOF.
- Our algorithm was unable to provide consistent predictions for the unknown strain. We believe the limited number of training data is a significant factor in the stochastic nature of its outputs. To date, 9 spectra have been acquired and more training set data is expected to be gathered in coming weeks.

ACKNOWLEDGEMENTS

Junior Ayuk Enow, Ideas and encouragement
Biotechnology Program of Glendale Community College
Artemio Chavez, Poster design

Future Work

- More P21 strain data of *Deinococcus* Aquaticus will be gathered and analyzed
- Publish spectra to public databases
- Rigorous statistical analysis of the algorithm's predictive accuracy

REFERENCES

1. Rios A, Kavuluru R. Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles. ACM-BCB . . : the . ACM Conference on Bioinformatics, Computational Biology and Biomedicine ACM Conference on Bioinformatics, Computational Biology and Biomedicine. 2015;2015:258-267. doi:10.1145/2808719.2808746.
2. Brookes B and Murray G. Int. J of Systematic Bacteriology 1981. *Nomenclature for "Micrococcus radiodurans" and Other Radiation-Resistant Cocci: Deinococcaceae fam. nov. and Deinococcus gen. nov., Including Five Species*
3. Eisen JA (1995) The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J Mol Evol 41: 1105–1123.
4. Lane, D.J. 1991. *16S/23S rRNA sequencing. In: Nucleic acid techniques in bacterial systematics. Stackebrandt, E., and Goodfellow, M., eds., John Wiley and Sons, New York, NY, pp. 115-175.*
5. Slade D, Radman M. Oxidative Stress Resistance in Deinococcus radiodurans . Microbiology and Molecular Biology Reviews : MMBR. 2011;75(1):133-191. doi:10.1128/MMBR.00015-10.
6. Sandrin, T. R., Goldstein, J. E. and Schumaker, S. (2013), *MALDI TOF MS profiling of bacteria at the strain level: A review. Mass Spectrom. Rev., 32: 188–217. doi: 10.1002/mas.21359*
7. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular Systems Biology. 2016;12(7):878. doi:10.15252/msb.20156651.