Investigating the use of Atomic Property Weighted Radial Distribution Functions to predict the Water Stability of Metal-Organic Frameworks using Machine Learning

George Irving (gi238@cam.ac.uk)

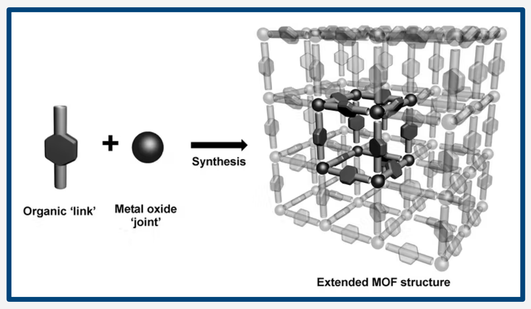
Supervised by: Nakul Rampal & Prof. David Fairen-Jimenez

Department of Chemical Engineering and Biotechnology, University of Cambridge

Background

What are Metal-Organic Frameworks?

- ✔ Exciting class of materials formed by the composition of organic ligands with inorganic metals
- ✔ Highly open structures with extraordinary internal surface areas (sometimes over 10,000 $m^2\ g^{-1}$) making them attractive for use in many fields such as adsorption, catalysis, gas storage, and even drug delivery



C. Sumby (2012)

Atomic Property Weight Radial Distribution Function (AP-RDF)

$$AP\text{-}RDF(R) = \sum_{i,j} P_i P_j e^{-B(r_{ij}-R)^2}$$

- ✔ Can be interpreted as: the weighted probability distribution of finding an atom pair separated by a distance R in the unit cell
- ✔ Weighting by e.g. electronegativity or Van der Waals radius
- ✔ Previously successfully used to predict MOF gas uptake capacity (M. Fernandez et al., 2013) but has not been used to predict chemical stability

Water stability? Why...?

- ✔ Water molecules can attack the metal-organic coordination bonds and lead to structural collapse
- ✔ Water vapour often found in industrial process streams
- ✔ Significant energy required to separate out water before an adsorption bed
- ✔ Therefore useful to be able to predict stability of new MOFs
- ✔ Previously predicted by R. Batra et al. (2020) using various chemical features but RDFs have not been used in the literature

AP-RDF → **Machine learning?** → **Water stability**

Methods

1. Collect dataset → 2. Create AP-RDFs → 3. Develop and train ML models → 4. Evaluate performance

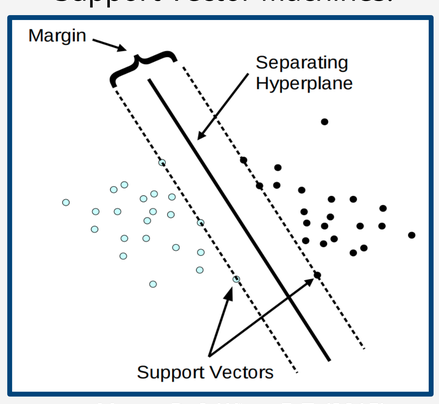
1a. Dataset – precursor: N. Burtch et al. (2014) review of water stability gives classifications

1b. Dataset – preparation: Find .cif (structural information file) for each MOF using Cambridge Structural Database and reference to original papers

2. AP-RDFs: Implement AP-RDF calculations and evaluate across the dataset

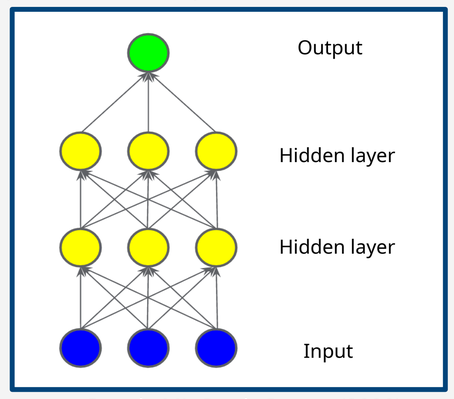
3. Machine Learning method:

Support Vector Machines:



Meyer, D. & Wien, F. T. (2015)

(Convolutional) Neural Networks:



Google ML Crash Course (2020)

4. Evaluate performance: k-fold cross validation was used to allow the majority of the dataset to be used for training. The most useful metric representing the ability of an algorithm to classify data is the **Matthews Correlation Coefficient (MCC)** which can be calculated from the **confusion matrix.**

Results

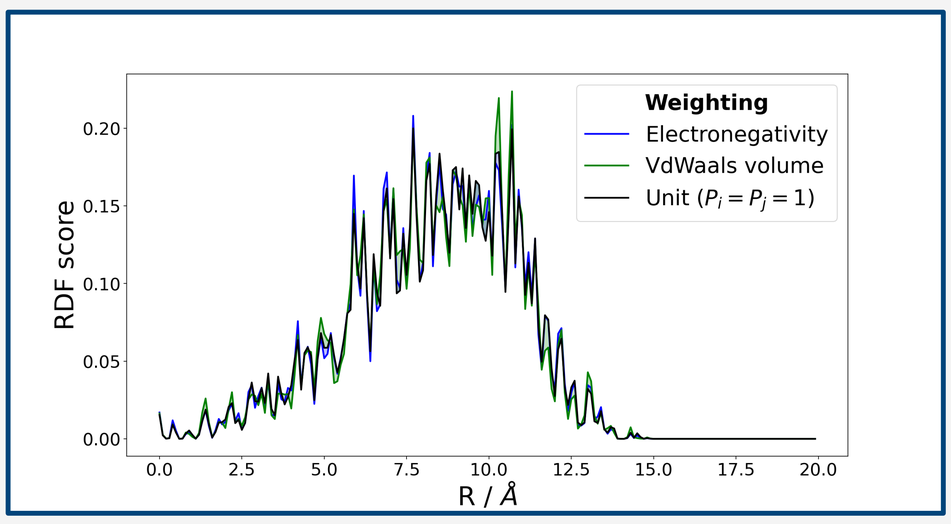


Figure 1: An example of AP-RDFs - $Zn_2(BDC)_2$(2,2′-dimethyl-BPY)

Each of the MOFs in the dataset have had three AP-RDFs calculated using unit, electronegativity, and Van der Waals volume weighting.

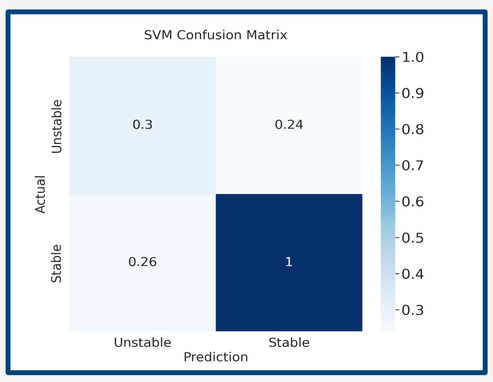


Figure 2: SVM learning results for all weightings combined.

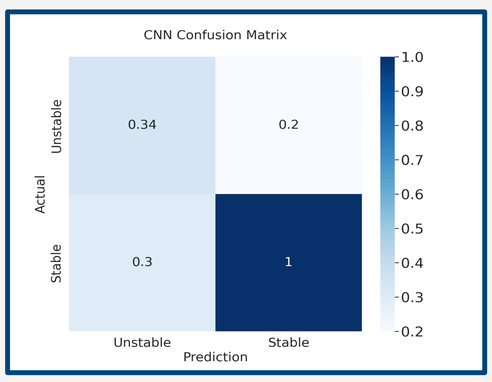
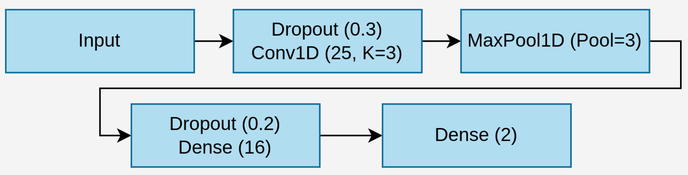


Figure 3: CNN learning results for all weightings combined.

Architecture of CNN that gave best results:



The key results:

- ✔ Both models successfully learn and are able to give reasonable averaged MCC values (Table 1)
- ✔ Differences in learning between weightings is negligible. However CNNs were consistently able to learn greater information when the RDFs were combined
- ✔ Learning curves suggest that more data would significantly aid learning

Table 1: The Matthews Correlation Coefficients achieved for both ML methods for different atomic property weightings and the combination of all three.

Weighting	SVM MCC score	CNN MCC score
Unit	0.38 ± 0.02	0.34 ± 0.05
Electronegativity	0.36 ± 0.02	0.34 ± 0.05
Van der Waals Vol.	0.38 ± 0.02	0.32 ± 0.04
Combined	**0.39** ± 0.02	**0.42** ± 0.04

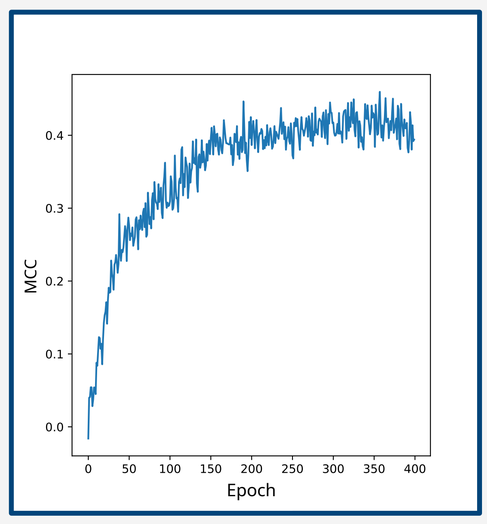


Figure 4: CNN MCC score by epoch shows computation time is not the limiting factor.

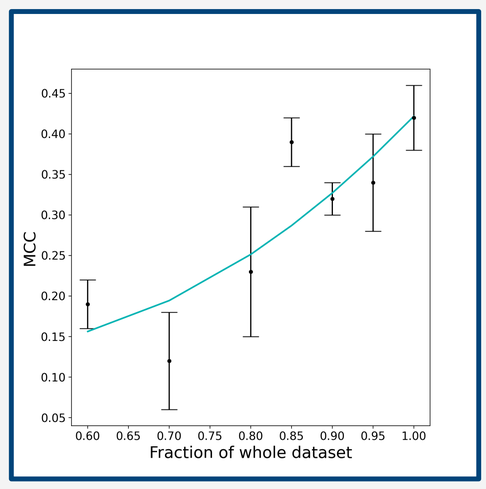


Figure 5: CNN MCC score by fraction of dataset suggests more data would aid learning.

Future work

Short term:

- ✔ Continue investigations into **optimal CNN architecture**
- ✔ Create RDFs based on **further atomic properties**
- ✔ Assess if particular MOFs are frequently miscategorised and **gain insights** into the chemical/structural reasons for this

Longer term:

- ✔ **Larger stability datasets** are likely to significantly improve learning performance
- ✔ Incorporate AP-RDFs into **other feature sets** to evaluate if learning is **orthogonal**
- ✔ Extend into predictions of stability in **other chemical environments**

References

Batra, C, et al. (2020). Prediction of water stability of metal–organic frameworks using machine learning. Nature Machine Intelligence, 2(11), 704-710

Burtch, N. et al. (2014). Water Stability and Adsorption in Metal–Organic Frameworks. Chemical Reviews, 114(20), 10575–10612.

Fernandez, M. et al. (2013). Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity. The Journal of Physical Chemistry C, 117(27), 14095–14105.

Meyer, D., & Wien, F. T. (2015). Support vector machines. The Interface to libsvm in package e1071, 28.

Google ML Crash Course (2020). Neural Networks: Structure

Monday 28th February 2022 : Part IIB Research Project Poster Presentation