Prediction of MXene 2D Material's Electronic Property Using Machine Learning Tools.

Jiwoo Seo
School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University

Assoc. Prof Tej Choksi, School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University

Pranav Roy, School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University

Abstract **-** MXene, a family of 2D materials, was recently shown to offer electrochemical qualities that have excellent applicability in catalysis and electronics. MXene offers high tunability by varying different properties such as constituent elements and lattice structure. Understanding those properties is a crucial step in tuning a more efficient and versatile catalyst. MXenes' electric and physical properties, such as workfunction, orbital radii, and electronegativity, are collected from Computational 2D Materials Database (C2DB). Machine learning is used as an alternative data analysis method to extract unseen insights and construct an accurate result model. Fundamental statistical analyses are conducted, followed by diverse machine learning tools such as sure independence screening and sparsifying operator (SISSO), AUTOFEAT feature engineering & selection, and different ensemble methods like stacking and boosting. The high accuracy of models indicates the applicability of these models in future predictions. These models are significantly less time-consuming and expensive to produce than widely used analysis methods, such as quantum mechanical modelling methods like density functional theory (DFT) or direct experimentation. Further research, such as using data of MXene across different environments, should be explored for further industrial application.

Keywords – MXene, workfunction, compressed sensing, feature generation & selection, ensemble learning, machine learning

1 INTRODUCTION

MXene is a recently discovered 2D material with formula of $Mn+1XnTx$ (n=1-3), where M is early transition metal, X is carbon or nitrogen, and T are surface terminations, most commonly OH*, O* or F*. Its unique characteristics enable multiple applications, such as heterogeneous catalysis as it has high thermal stability and ability to stabilize metal atoms. MXene's thermodynamic and electronics properties can be easily tuned with different geometries and/or constituent elements (mainly M & X). Workfunction is the amount of energy required to remove electron from the Fermi level to vacuum level, is the key characteristics when evaluating the performance of a catalyst. As of now, little is known about methods to accurately adjust these electronic properties for direct application such as electronic shielding. Literature shows that workfunction is dependent on the type of termination; Workfunction of bare, F/OH/O functionalized MXenes ranges in 3.3–4.8, 3.1–5.8, 1.6–2.8, and 3.3– 6.7eV, respectively. It is found that change in workfunction is also linearly correlated with change in surface dipole moment, which is affected by the charge transfer of substrate and terminal group. It is also dependent on the location of the termination groups, which is arranged in an energy-stable way to ensure that M-X bonds achieve the highest occupancy.

Currently, computational simulation using Density Functional Theory (DFT) is used to supplement experiments. While DFT is a calculation based on first principles and exhibit high accuracy, it is very expensive and limited for large system calculation. One of the accurate alternatives is machine learning, which can learn from data and rapidly identify complex patterns. Such machine learning models can be a first-line method for property analysis which can guide the direction of DFT calculations and unravel unseen patters. Here, we utilize three main machine learning methods: 1) sure independence screening and sparsifying operator (SISSO), 2) AUTOFEAT feature engineering & selection 3) ensemble methods like stacking and boosting.

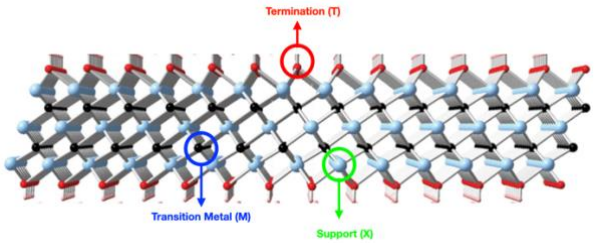


Figure 1: Side-view of MXene 2D material.

2 DATA PREPARATION

Primary Feature	Description of the Feature	Unit
HoF	Heat of Formation	eV/atom
E_Hull	Energy above convex hull	eV
Length(x)	Lattice Constant	pm
r-(M)	Radius of M atom	pm
r-(X)	Radius of X atom	pm
r-(T)	Radius of T atom	pm
IP(M)	Ionization Potential of M atom	eV
IP(X)	Ionization Potential of X atom	eV
IP(T)	Ionization Potential of T atom	eV
EN(M)	Electronegativity of M atom	(no unit)
EN(X)	Electronegativity of X atom	(no unit)
EN(T)	Electronegativity of T atom	(no unit)
EA(M)	Electron Affinity of M atom	eV
EA(X)	Electron Affinity of X atom	eV
EA(T)	Electron Affinity of T atom	eV

Figure 2: Primary features selected for ML training

Data was obtained using DFT calculation and experiments, at the curtesy of NTU Prof Tej's research team, and Computational 2D Material Database (C2DB), containing work function values of 275 different MXenes with O/F/OH/Bare termination and 15 primary electronic and physical properties. Data was standardized with mean of 0 and standard deviation of 1.

3 SISSO

A compressed sensing method called SISSO (sure independence screening and sparsifying operator) is a feature generation and selection tool recognized in the field of material properties prediction. SISSO first creates different features, by applying mathematical operators (in this case, +, -, ||, ^2, ^3,/) into the selected primary features. The operation is iteratively applied to generate an exponentially growing feature space, with user-selected iteration. Then, SIS (sure independence screening) selects high performance features by inspecting its correlation with not only the target property (work function) nut also the residual error of prior selected features. Finally, SO (sparsifying operator) is used to linearly combine the selected features, to find a highest performing sparse solution. The advantage of using SISSO is that it performs well with complex relationship with limited data.

For this research, no primary features were discarded for some models, as importance of single feature doesn't always relate to importance of that feature as a multi-dimensional descriptor. Out of different SISSO parameters, 5 iterations and 5 non-zero coefficient for the final sparse solution was selected, for its accuracy and rapid computation speed. To check for overfitting, Leave-One-Out Cross-Validation method was selected.

	Model
1D	+ 1.321 |EN(X)-EN(T)| + 3.179
2D	+ 1.534 |EN(X)-EN(T)| + 0.550 (EN(M)+EN(T)) + 0.502
3D	+ 2.125 |EN(X)-EN(T)| + 1.950 (EN(M)+EN(T)) - 1.500 EA(T) - 3.724
4D	+ 2.246 |EN(X)-EN(T)| + 2.366 (EN(M)+EN(T)) - 1.788 EA(T) - 1.008 (EN(M)-EN(X)) - 6.416
5D	+ 1.995 |EN(X)-EN(T)| + 2.683 (EN(M)+EN(T)) - 1.802 EA(T) - 2.445 EN(M) - 0.024 (Length(x)-r-(T)) + 3.743

Figure 3: SISSO performance and model for dimension 1D~5D

Starting with average training RMSE of 1.388 eV in 1D, RMSE rapidly converges to average training RMSE 0.733 eV, and average testing RMSE of ( ) in 5D.

One interesting observation is the presence of feature |EN(X)-EN(T)|, (EN(M)+EN(T)) and EA(T) throughout the sparse solutions of different dimensions, according to the descriptor selection frequency. A new feature, (Length(x)-r-(T)) emerges at 5D.

This is a moderate performance compared to other machine learning methods explored by Pranav Roy et al, and offers an advantage as we can inspect the operators, unlike other black box models such as Neural Network.

Dimension	Feature	Descriptor selection frequency
1D	|EN(X)-EN(T)|	275
2D	|EN(X)-EN(T)|, EN(M)+EN(T)	275
3D	|EN(X)-EN(T)|, EN(M)+EN(T), EA(T)	275
4D	|EN(X)-EN(T)|, (EN(X)+EN(T)), EA(T), (EN(M)-EN(X))	166
	|EN(X)-EN(T)|, (EN(X)+EN(T)), EA(T), |EN(M)-EN(X)|	105

Figure 4: SISSO descriptor selection frequency

4 AUTOFEAT

AUTOFEAT is a feature engineering and selection framework inspired from SISSO and other machine learning tools. Firstly, it removes features with high correlation with target property. It generates feature space with mathematical operation (log, root, 1/x, power, absolute, exp, 2^x, sin, cos, tan), while ensuring that it follows the Buckingham pi theorem. Brute force is used to ensure that some important features are not lost in the process. Feature selection is done by multi-step selection approach using L1-regularized logistic regression model and LASSO LARS regression model (scikit-learn). It also utilizes noise filtering approach, where noise features are added and only keep features that has larger coefficient than the largest of noise features. Once high-performing features are selected, instead of simply generating a linear solution, they can go through other complex standard machine learning

operations, such as Neural Network and Random Forest Regression.

The results show rapid decrease in both training and testing RMSEs fir both AUTOFEAT generated solution and Kernel Ridge Regression (KRR) using the AUTOFEAT generated features, with the lowest RMSE of 0.264 eV for KRR at AUTOFEAT step 3.

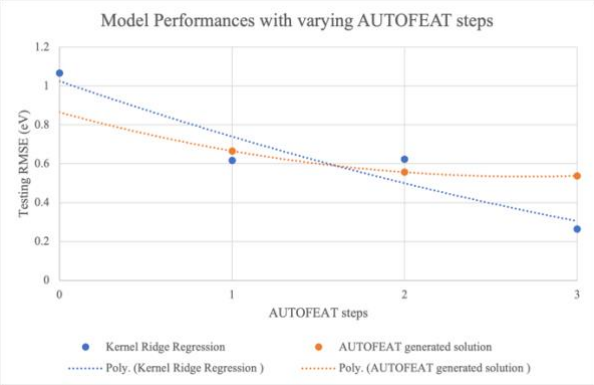


Figure 5: Model performance with varying AUTOFEAT steps

AUTOFEAT generated solution (step 1)
-3.074+4.678EN(X)+4.050cos(EN(X))-2.208EA(X)+1.893324r-(X)/Length(x)+1.847HoF+0.663EN(T)+0.014284r-(X)-0.01*(E_Hull)
AUTOFEAT generated solution (step 2)
8.520-4.902(EA(X))^2*exp(-EN(X))+3.516*sin(EN(X))*cos(EN(X))-1.749*EA(X)*exp(-EN(X))+0.845*sqrt(HoF)*sqrt(EN(X))-0.747*exp(-EN(X))+0.528*cos(EN(X))-0.289*sin(EN(X))-0.0236*IP(X)^3*exp(-EN(X))+0.0125*EA(M)*EA(X)^3-0.00453*E_Hull+0.00294* EN(X)^3*exp(-EN(T))-0.000369*r-(X)^2*cos(EN(X))
AUTOFEAT generated solution (step 3)
6.600 + 0.693 * sin(EN(X)^2*cos(EN(X))) -0.658 * sin(2^EN(X))+0.428*sqrt(EN(X))-0.294*sin(2^EN(X))+0.220* sin(EN(X)^3*exp(EN(X)))-0.149*EA(M)*sin(EN(X))-0.148* cos(EN(X)^2)-0.136*sin(2^EN(X)+Abs(EN(T)))-0.101* sin(EN(X)^2*x012^3)-0.0977*cos(2**EN(X) -2**EN(T))-0.0875 * sin(EN(X) **2*exp(EN(X))) -0.009 * E_Hull

Figure 6: equation of the AUTOFEAT generated solution

5 ENSEMBLES

Ensemble learning is a method that aims to reduce noise, bias and variance by employing different machine learning models together into one. The most popular ensemble methods, bagging, boosting and stacking were used. For bagging, the selected machine learning models are trained independently and later combined in a deterministic averaging method. Boosting learns sequentially and is combined. For stacking, ML models are trained independently and in parallel, then combined by training a ‘meta-model’ based on the predictions of the prior trained ML models.

lowest RMSE of 0.264 eV for KRR at AUTOFEAT step 3.

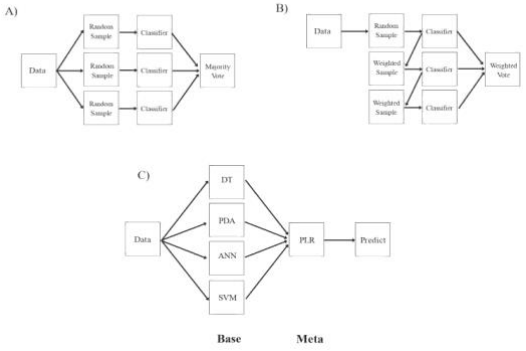
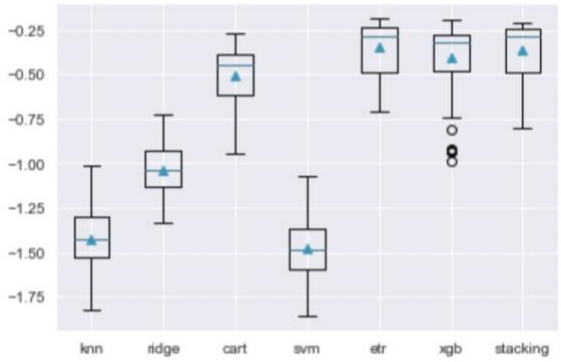


Figure 7: comparison of different ensemble methods A) Bagging B) Boosting C) Stacking. [5]

Before ensemble learning, data were imputed using k-nearest neighbours. Different from the previous tools used, categorical variables (M,X,T) were included into the dataset by one hot encoding. The hyperparameter of these ensemble models were chosen using Bayesian hyperparameter optimization.

5.1 STACKING




```
>knn -1.425737 (0.182463)
>ridge -1.040427 (0.136737)
>cart -0.505438 (0.157488)
>svm -1.480610 (0.173885)
>etr -0.349671 (0.146236)
>xgb -0.404283 (0.195835)
>stacking -0.363133 (0.163268)
```

Figure 8: comparison of different ensemble methods A) Bagging B) Boosting C) Stacking

Figure: (left) negative adjusted RMSE and computational speed of base models and stacking model. (right) box and whisker plot of the negative adjusted RMSE for base models and staking model.

The machine learning models selected were K-Nearest Neighbours, Ridge Regression, Support Vector Machine, XGboost, Decision Tree Regressor, Extra Trees Regressor. These diverse ranges were used with different weaknesses (such as linearity, overfit, simplicity) to supplement each other. The resulting stacking results show RMSE of -0. The result shows that the models with varying RMSE (from 1.4806 eV to 0.3496) combine to give a high performance result in stacking, with RMSE of 0.3631. This shows that, if models are selected and adjusted correctly, stacking creates stronger model from weak learner's predictions.

5.2 BAGGING

For bagging, the highest performing model was the CatBoost model, with RMSE of 0.29. The high performance is likely because it's a leaf-wise (not level-wise) and histogram-based, which often has a high performance and speed for large parameters.

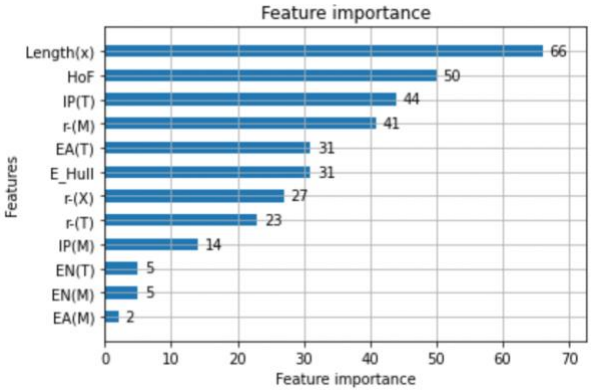


Figure 9: feature importance chart from CatBoost model

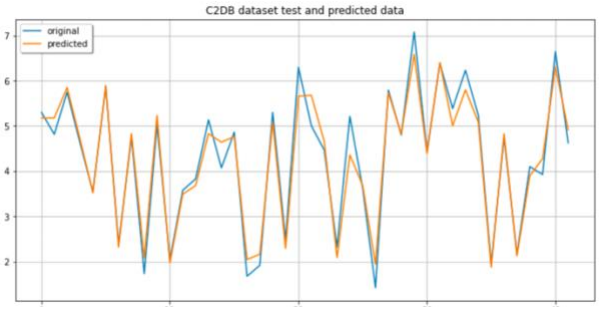


Figure 10: comparison of original and predicted Work function values with CatBoost model

6 CONCLUSION

In this research we presented a range of effective and computationally non-intensive machine learning workflow, with best performance for kernel ridge regression using AUTOFEAT step 3 selected, with RMSE of 0.264. Different models selected different primary features for high feature importance, likely due to the varying levels of model complexity and accuracy. SISSO and AUTOFEAT provides a solution equation that relates these primary properties with work function, which provides an analysis advantage over other black box models. Instead of using DFT computation which is very computationally intensive, we can use the aforementioned model results to predict work function values of new MXenes for further application.

7 LIMITATION & AREAS OF IMPROVEMENT

The dataset is under vacuum and 0K. Therefore this doesn't consider side-reactions such as oxidization or termination desorption that may occur under non-vacuum high temperature environment. More primary properties related to surface diplole moment could have been selected for better insight, such as termination arrangement (atom, bridge, fcc, hcp), transition metal ionic state, and/or structure of MXenes (M2X,M3X2…) Further investigation should take place to ensure the dataset is a sparse dataset suitable for ML tools such as compressed sensing. In our dataset, about 25% of the MXenes are -OH terminated, but according to literature, -OH termination accounts for less than 5% of the available sites. Other machine learning tool should be further investigated, such as genetic programming. Further investigation on properties with high occurrence probabilities should take place.

ACKNOWLEDGEMENT

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project

REFERENCES

[1] Yuri Gogotsi et al, 2D metal Carbides and Nitrides (MXenes), 2019

[2] Runhai Ouyang et al, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Aug 2018

[3] Brownlee, Jason. "Stacking Ensemble Machine Learning with Python." Machine Learning Mastery, 26 Apr. 2021, https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/.

[4] Horn, Franziska, et al. "The Autofeat Python Library for Automated Feature Engineering and Selection." ArXiv.org, 26 Feb. 2020, https://arxiv.org/abs/1901.07329.

[5] Resting Heart Rate Variability Can Predict Track and Field Sprint ... https://www.researchgate.net/publication/331562581_Resting_Heart_Rate_Variability_Can_Predict_Track_and_Field_Sprint_Performance.

[6] Rocca, Joseph. “Ensemble Methods: Bagging, Boosting and Stacking.” Medium, Towards Data Science, 21 Mar. 2021, https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205#:~:text=Very%20roughly%2C%20we%20can%20say,variance%20can%20also%20be%20reduced.

[7] Salim, O., et al. “Introduction to Mxenes: Synthesis and Characteristics.” Materials Today Chemistry, Elsevier, 27 Sept. 2019, https://www.sciencedirect.com/science/article/abs/pii/S2468519419302034.

[8] Pranav Roy, Tej S. Choksi, et al. “Predicting the Work Function of 2D MXenes using Machine-Learning Methods”

[9] Chem. Mater. 2019, 31, 17, 6590–6597. Publication Date:April 4, 2019. https://doi.org/10.1021/acs.chemmater.9b00414

[10] J. Mater. Chem. C, 2017,5, 2488-2503

[11] Electrical4U. “Work Function: Formula & Relation to Threshold Frequency.” Electrical4U, 27 July 2020, https://www.electrical4u.com/work-function/.

[12]Mater. Horiz., 2016, 3, 7-10, 10.1039/C5MH00160A