An Application of Machine Learning on Predicting CO_2 Reduction Electrocatalysts

Chen Liang, Yuhang Zhang

Abstract

It has proved that single-atom catalysts (SACs) are able to accelerate CO_2 reduction reaction (CO_2RR) as electrocatalysts, but it is still difficult to design SACs rationally for this reaction. Based on data derived from theoretical computation, the machine learning method can be a solution to the problem. However, how to construct an appropriate representation of materials for machine learning is still an active topic. In this work, we applied machine learning to predict electrocatalysts for CO_2RR using multiple representations. We not only obtained models with good performance in this problem, but also gave interpretation on the reason why different structure representations lead to different performances.

1 Introduction

Carbon dioxide (CO_2) reduction, especially reduction to carbon oxide (CO), is one of the most promising ways to mitigate the greenhouse effect as well as obtain valuable carbon products [1, 2], which has a relatively high selectivity and brings widely-used resultant (CO) [3]. Electroreduction of CO_2 (ECR) is an approach of realizing it, and electricity derived from green energy resources is enough to motivate it [4]. The most important thing we need for the reaction is a high-performance electrocatalyst, and the development of machine learning makes the screening and predicting process much more efficient [5]. Among various kinds of catalysts, single-atom catalysts (SACs), which are constructed using isolated metal atoms (doping atoms) dispersed on substrates, show distinct advantages on accelerating the reduction process [6, 7]. Therefore, we are trying to apply machine learning method to find appropriate SACs for ECR.

One of the most important problems here is to find representations of SACs suitable for machine learning. There are multiple perspectives to characterize materials, leading to different ways to represent them. Descriptors or features of materials can be classified into structure features, which exhibit structure information (for instance, atomic positions and coordinate numbers of atoms) of materials, and element features, which are related to chemical and physical properties (for instance, ionization energy) of elements contained in the materials. Some researchers used both of the kinds of featrues to represent SACs and applied machine learning to screen catalysts from them [5]. Nevertheless, a procedure of feature selection is always necessary, for there are too many element features related to one element. But the selection is related to chemical insights of researchers into the problem, inevitably leading to arbitrariness. Therefore, it may be a better way to describe them only using the coordinates of atoms [8] with one or at most two descriptors to differentiate element types of the doping atoms. In this work, we applied machine learning method to predict high-performance catalysts for CO_2 reduction reaction (CO_2RR) using dataset based on coordinates of atoms in SACs with multiple structure representations. We hope to get a model which can predict whether a SAC is a good catalyst for CO_2RR only using its structure information and some data we can get easily from the internet. After that, we tried to interpret the reason why some structure representation can perform better than others using manifold learning.

2 Dataset and Features

Our dataset is based on computation results of adsorption energy of CO on different SAC surfaces using VASP [9], a quantum chemistry computation software. SACs are constructed on five crystal surfaces of copper (Cu) with around 40 elements replacing one Cu atom on the surfaces each time on different sites, an example of which is shown in Figure 1. After deleting some abnormal structures, we got around 3000 structures with adsorption energy of CO for each of them. The adsorption energy of the resultant of a good catalyst should not be so small or so large, to ensure that the molecule is activated on the surface because of the electron transferring as well as easy to escape from the surface after the reaction. Therefore, the adsorption energy can be a good measurement of catalytic capacity of a material on a chemical reaction, which is set as the label for the regression problem.

In this work, element features describe properties of doping atoms, including atomic mass, radius, electron affinity, Pauling electronegativity, ionization energy and number of valence electrons. Structure features are lattice parameters and coordinates of atoms in SACs, which directly reflect positions of atoms in the materials, or symmetry functions (SFs) [8, 10] of them to extract structure information more precisely. Besides, the transfer charge (Q_t), which measures the movement of charge between the doping atom and the Cu substrate due to the difference of their chemical properties, is related to both of the factors mentioned above.

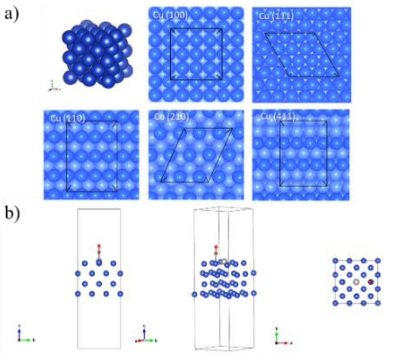

FIG. 1. Structures of a) Cu crystal and five surfaces of it and b) a SAC adsorbing a CO from different perspectives.

3 Methods

Machine learning algorithms we used in this work included Support Vector Machine (SVM, based on polynomial kernel and Gaussian kernel, respectively), Gradient Boosting Regression (GBR) and Neural Network (NN). The structure of NN in this work had two hidden layers, and the numbers of neurons in each layer are 100, 50/150,100 according to different sizes of input. The activation function of this network was ReLU. In order to deal with overfitting, we introduced a dropout layer and regularization techniques. In addition, the learning rate declining method was introduced to improve accuracy.

In the training process, for SVM and GBR, the size of training set to test set was 4:1, and 0.2 of training set were used as the validation part to select proper parameters. For NN, the training set and test set were divided five to one. Every batch had half of training data, and the number of iteration was 8000. In order to determine the values of parameters including learning rate and its declining as well as regularization, sensitivity analysis was implemented to find the optimal value according to MSE and MAE. We chose a large range of these parameters and then found a smaller interval to do sensitivity analysis. The results are shown in Figure 2.

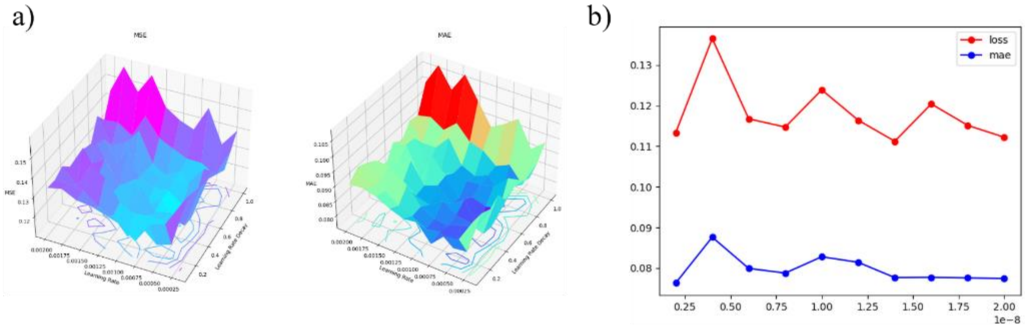

FIG. 2. Parameter setting procedures for NN, including a) learning rate and its declining and b) regularization process.

Manifold learning, a dimension reduction method considering the nonlinear relationship in data, was introduced to interpret the difference of performances based on different representations. We applied Multidimensional Scaling (MDS) [11] and t-distributed Stochastic Neighbor Embedding (t-SNE) [12] for this purpose.

4 Results

4.1 Performances of machine learning algorithms based multiple representations

First of all, we tested performances of machine learning algorithms mentioned above. In this primary test, the element feature was ionization energy, and we used the simplest method to construct structure features without SFs, which is called the simple matrix representation. Table 1 exhibits the results, and the “# of models” column means the number of models with different parameters for each algorithm. We can see clearly that GBR and NN outperform SVM largely, and we only used GBR and NN to do analysis further.

Algorithms	R^2 score	RMSE	MAE	# of models
SVM (poly kernel)	0.701461	0.334459	0.220453	1737
SVM (rbf kernel)	0.778075	0.288367	0.154408	360
GBR	0.924240	0.164814	0.098185	768
NN	0.923467	0.174112	0.096741	125

Table 1. Performances of test set of different algorithms trained on simple matrix representation with ionization energy as the element feature. Q_t feature was included in the test.

To find whether SFs can enhance the performance of machine learning, we applied GBR and NN on four structure representations, respectively, and the results are exhibited in Table 2. Ionization energy is the element feature in this test, and we trained the model with and without Q_t in the dataset. We can see from the results that SF2 performs the best, which introduces cosine function to build up its formulas and considers three-body terms. As for algorithms, the overall performance of NN is better than GBR especially when we dropped the Q_t feature, due to its flexibility. Figure 3 exhibits results of GBR and NN trained on SF2 dataset. In addition, we applied GBR to test which element feature is the most relevant to this task, and Table 3 shows the results. Radii of doping atoms and numbers of valence electrons contained in them have the greatest influence on the adsorption energy of CO.

Models	Representations	With Q_t				Without Q_t			
		R^2 score	RMSE	MAE	# of models	R^2 score	RMSE	MAE	# of models
GBR	Simple matrix	0.924240	0.164814	0.098185	768	0.858092	0.225568	0.139255	768
	SF1 (cos, 1-2b)	0.932955	0.155044	0.088203	768	0.890579	0.198073	0.115673	768
	SF2 (cos, 1-3b)	**0.951977**	**0.13122**	**0.068191**	768	**0.934283**	**0.153501**	**0.082745**	768
	SF3 (tanh, 1-3b)	0.926178	0.168253	0.095829	768	0.880999	0.213621	0.124485	768
NN	Simple matrix	0.923467	0.174112	0.096741	125	0.893924	0.182700	0.114121	125
	SF1 (cos, 1-2b)	0.947504	0.146773	0.103090	125	0.920241	0.176440	0.125913	125
	SF2 (cos, 1-3b)	**0.973240**	**0.100557**	**0.070464**	125	**0.955572**	**0.133868**	**0.093109**	125
	SF3 (tanh, 1-3b)	0.966210	0.113079	0.078114	125	0.947613	0.143289	0.099713	125

Table 2. Performances of test set of GBR and NN trained on dataset of four structure representations with ionization energy as the element feature.

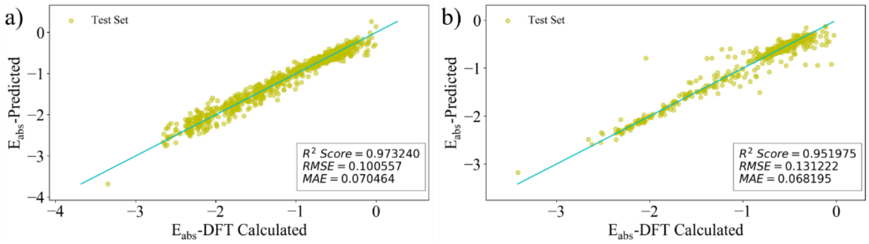

FIG. 3. Test results of a) NN and b) GBR trained on dataset with SF2 as its structure representation and ionization energy as its element feature, including Q_t.

Element features	With Q_t				Without Q_t			
	R2 score	RMSE	MAE	# of models	R2 score	RMSE	MAE	# of models
Electron affinity	0.956053	0.125528	0.065512	768	0.893515	0.195398	0.100101	768
Ionization energy	0.951977	0.13122	0.068191	768	0.934283	0.153501	0.082745	768
Mass	0.952887	0.129971	0.065548	768	0.935897	0.151605	0.079542	768
Electronegativity	0.954798	0.127307	0.065529	768	0.911229	0.178406	0.093158	768
Radius	0.960604	0.11885	0.060189	768	0.937464	0.149741	0.088968	768
Valence elections	0.959973	0.119799	0.059767	768	0.937084	0.150195	0.082222	768

Table 3. Performances of test set of GBR trained on dataset of six element features, respectively, with SF2 being the structure feature.

4.2 Manifold learning analysis of different structure representations

We used manifold learning techniques to find the reason why different representations lead to difference machine learning model performances. In this section, we applied MDS and t-SNE on dataset containing four structure features, respectively, in which the element feature was ionization energy. Besides, dataset with and without Q_t feature was also tested to see why models perform worse without it. Results are shown in Figure 4 for MDS and Figure 5 for t-SNE.

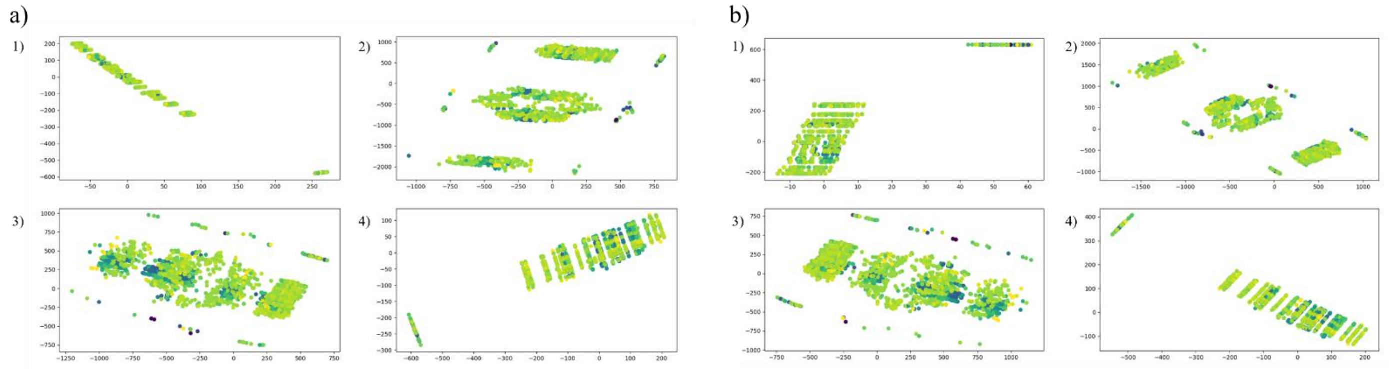

FIG. 4. Visualization of dataset reduced to 2-dimension by MDS with structure representation being 1) simple matrix, 2) SF1, 3) SF2, and 4) SF3 while ionization energy being the element feature. The test was done a) including and b) not including Q_t.

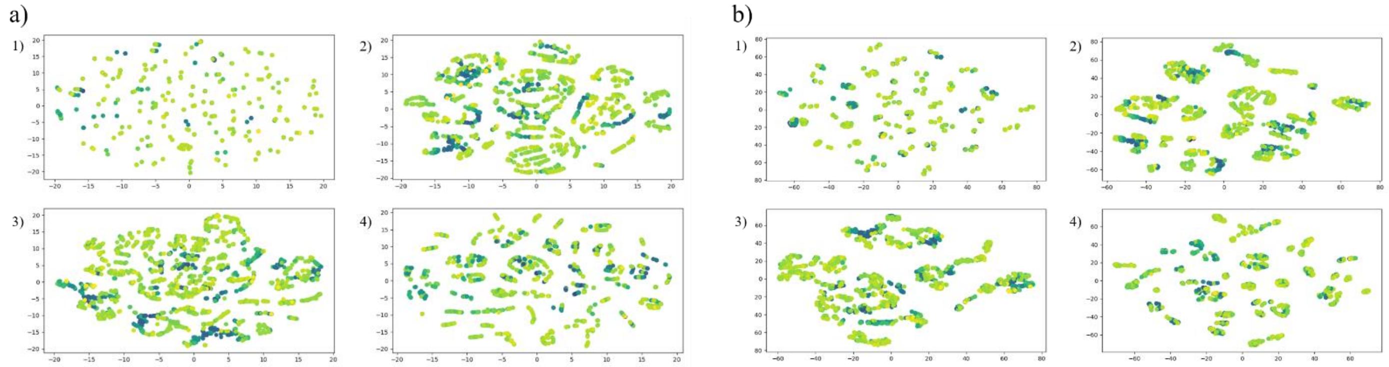

FIG. 5. Visualization of dataset reduced to 2-dimension by t-SNE with structure representation being 1) simple matrix, 2) SF1, 3) SF2, and 4) SF3 while ionization energy being the element feature. The test was done a) including and b) not including Q_t.

The data dimension is reduced to make them visible to see the distribution. In Figure 4 a), the best performance is derived from dataset of subgraph 3). In this picture, data gathers into several groups with two lines at both sides of them, and the pattern is similar with that of the subgraph 2), whose performance is a little worse. Data points at 1) exhibit in a shape of straight lines with less dispersion, leading to bad performance, and this character can be found in both Figure 4 and 5. The result of t-SNE shows more information of why dataset with Q_t can perform better. In Figure 5, the distribution of data with Q_t is more stretched compared with that of data without Q_t, which is consistent with the conclusion discussed in MDS situation. All information above may provide an interpretation of the better performance of dataset including Q_t and using SF2 as its structure representation.

5 Conclusion

Machine learning is used more and more widely in material science field, while how to represent materials is still a problem remaining to be solved. This work applied machine learning to predict good electrocatalysts of CO_2RR, comparing performances of dataset using different structure and element features, as well as providing an interpretation of it by manifold learning. The application of machine learning on material science is still a topic worth researching on.

References

[1] N. S. Lewis and D. G. Nocera. Powering the planet: Chemical challenges in solar energy utilization. *PNAS* 2006, 103: 15729-15735.

[2] S. Lin, C. S. Diercks, Y. Zhang, N. Kornienko, E. M. Nichols, Y. Zhao, A. R. Paris, D. Kim, P. Yang, O. M. Yaghi and C. J. Chang. Covalent organic frameworks comprising cobalt porphyrins for catalytic CO_2 reduction in water. *Science* 2015, 349: 1208-1213.

[3] T. Zheng, K. Jiang, N. Ta, Y. Hu, J. Zeng, J. Liu and H. Wang. Large-Scale and Highly Selective CO_2 Electrocatalytic Reduction on Nickel Single-Atom Catalyst. *Joule* 2019, 3: 1–14.

[4] Q. Lu, J. Rosen, Y. Zhou, G. S. Hutchings, Y. C. Kimmel, J. G. Chen and F. Jiao. A selective and efficient electrocatalyst for carbon dioxide reduction. *Nature Communications* 2014, 5: 3242.

[5] A. Chen, X. Zhang, L. Chen, S. Yao and Z. Zhou. A Machine Learning Model on Simple Features for CO2 Reduction Electrocatalysts. *J. Phys. Chem. C* 2020, 124: 22471–22478.

[6] Y. Chen, S. Ji, C. Chen, Q. Peng, D. Wang and Y. Li. Single-Atom Catalysts: Synthetic Strategies and Electrochemical Applications. *Joule* 2018, 2: 1242–1264.

[7] M. Jia, Q. Fan, S. Liu, J. Qiu and Z. Sun. Single-Atom Catalysis for Electrochemical CO_2 Reduction. *Current Opinion in Green and Sustainable Chemistry* 2019, 16: 1-6.

[8] Y. Chen, Y. Huang, T. Cheng and W. A. Goddard, III. Identifying Active Sites for CO_2 Reduction on Dealloyed Gold Surfaces by Combining Machine Learning with Multiscale Simulations. *J. Am. Chem. Soc.* 2019, 141: 11651−11657.

[9] G. Kresse and J. Furthmüller. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Computational Materials Science* 1996, 6: 15-50.

[10] J. Behler. Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry* 2015, 115: 1032–1050.

[11] M. A. A. Cox and T. F. Cox. Multidimensional Scaling. In: C. Chen, eds. *Handbook of data visualization*. Berlin: Springer, 2008, 315-347.

[12] L. van der Maaten and G. Hintton, Visualizing data using t-SNE. *Journel of machine learning research* 2008, 9: 2579-2605.