

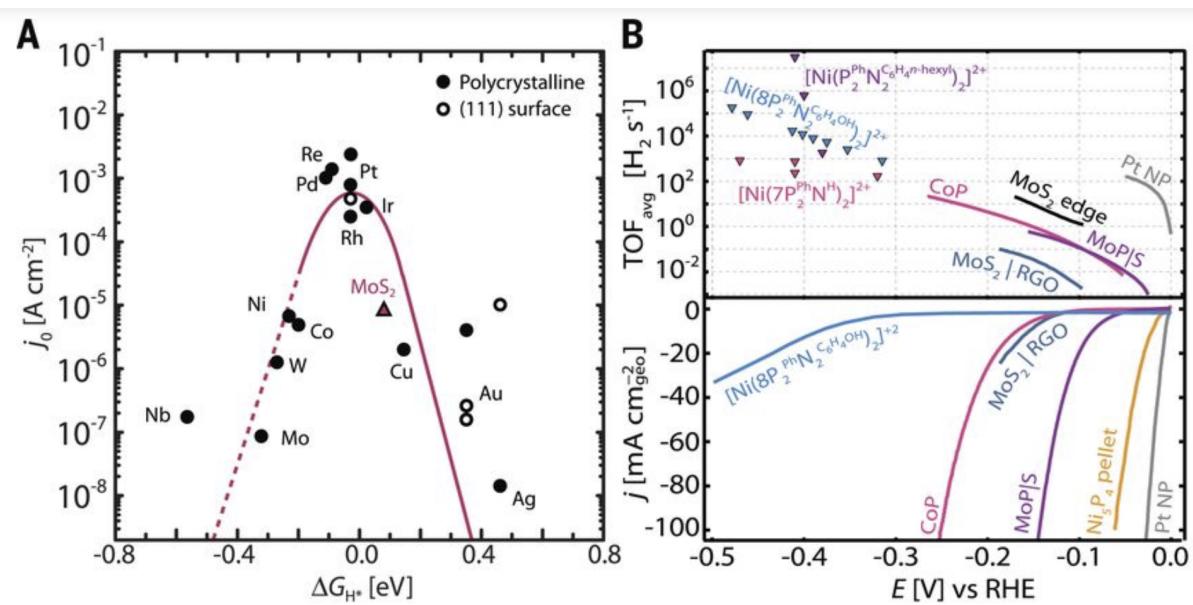


Hands-on 7. ML methods

[CO2RR_data.csv](#)

[handson7_ML_regression&feature_engineering.ipynb](#)

How can we use machine learning methodologies for designing catalyst?



<https://www.science.org/doi/full/10.1126/science.aad4998>

Basic concept: Catalytic activity as a multivariable regression problem.

As Catalytic activity is determined by catalyst character, it could be seen as an optimization problem of multivariables

It could be linear regression

$$y = ax + b$$

catalytic activity = $a^*(\text{catalyst property}) + b$

$\approx a^*f(\text{adsorption energy}) + b$

$\approx a^*f(\text{catalyst element, structure, etc}) + b$

It could be linear regression

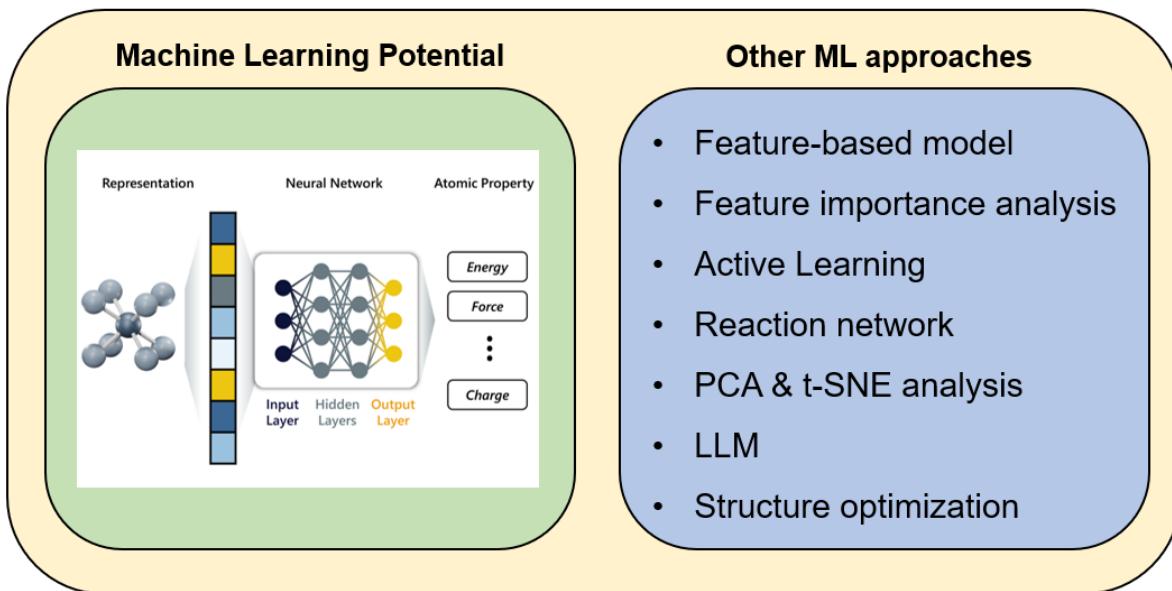
$$y = a_1x_1 + a_2x_2 + a_3x_3 + \dots$$

It could be regression by neural network

$$y \approx \text{Neuralnet}(a_1, a_2, a_3, \dots, x_1, x_2, x_3)$$

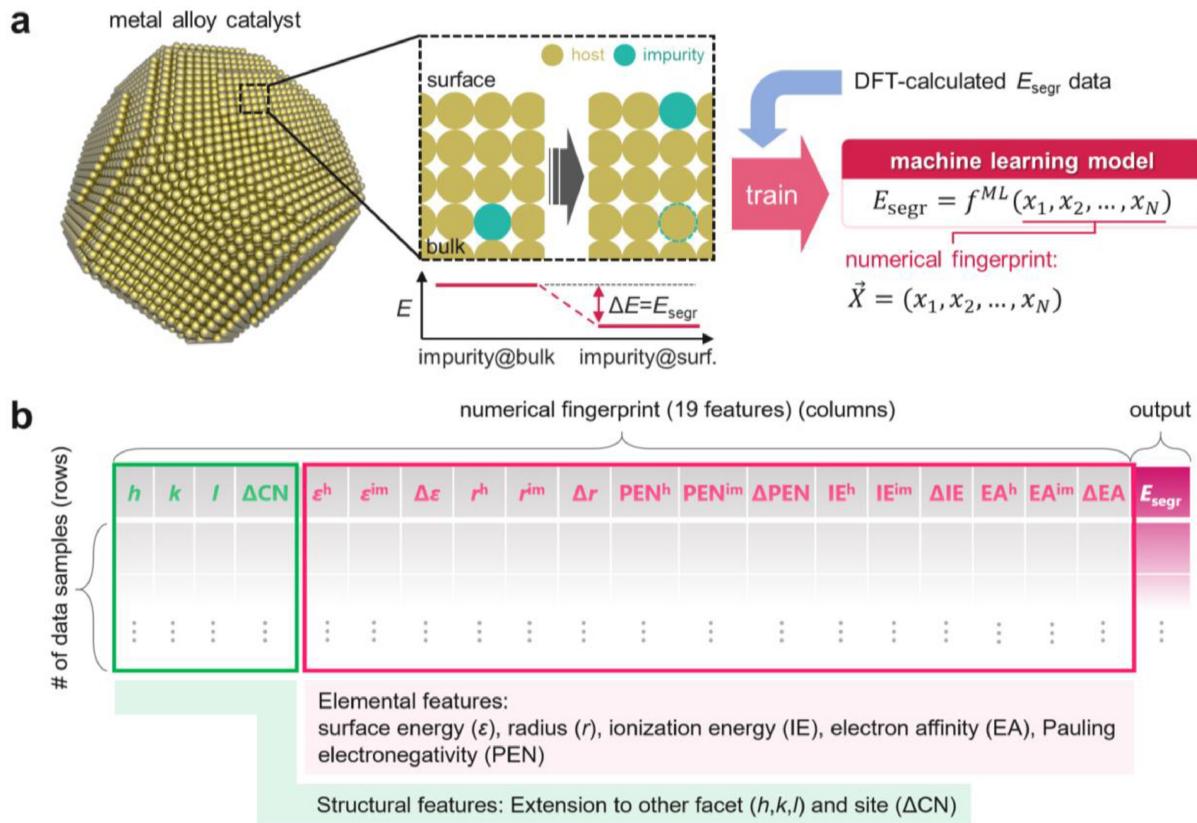
There are several machine learning methods except machine learning potential (MLP)

ML methodologies for catalysis



In this class, we are going to deal various ML approaches in catalysis except MLP

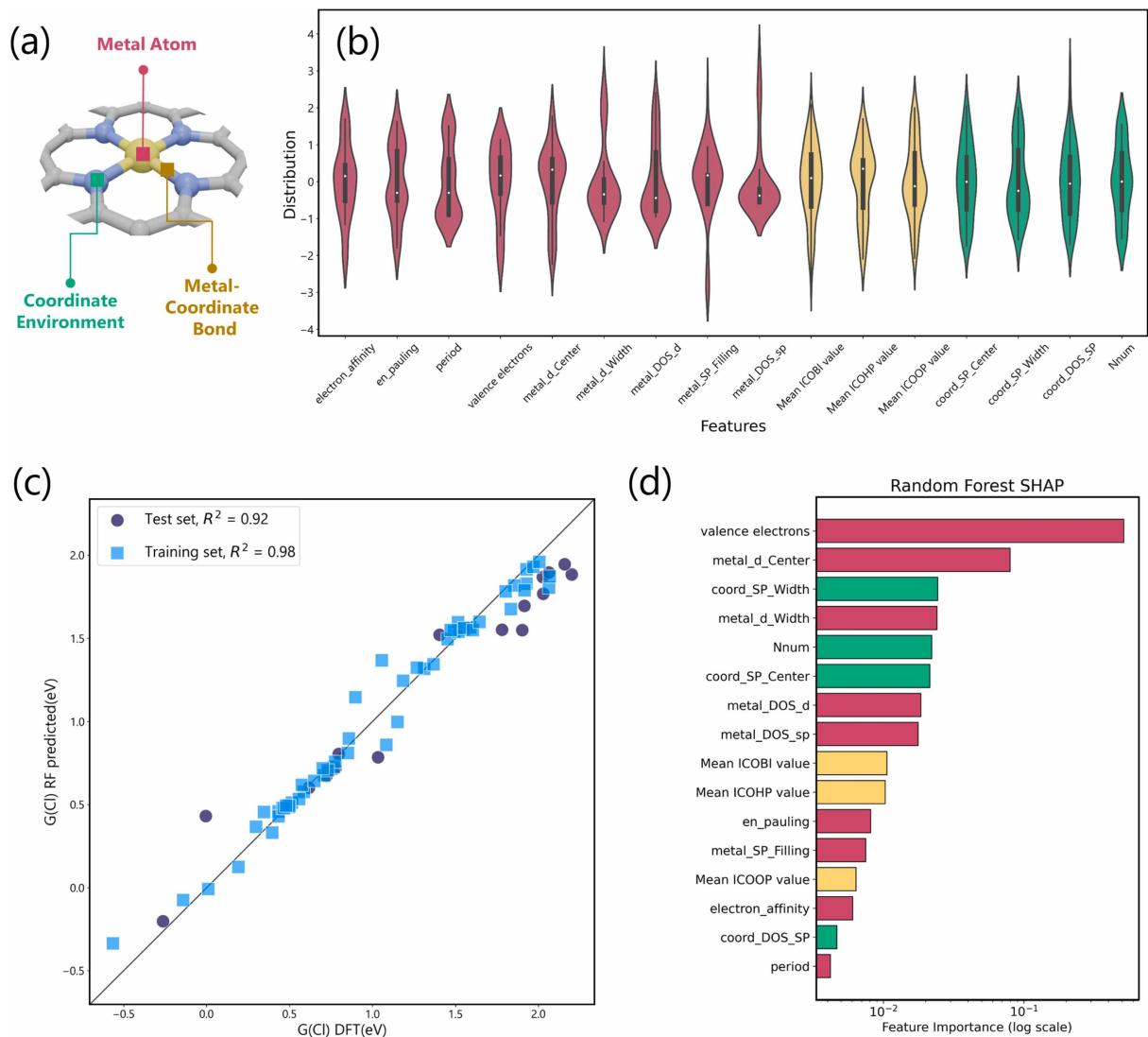
1. Feature-based model



- We can predict surface segregation energy with inexpensive features.

<https://doi.org/10.1016/j.mcat.2023.113096>

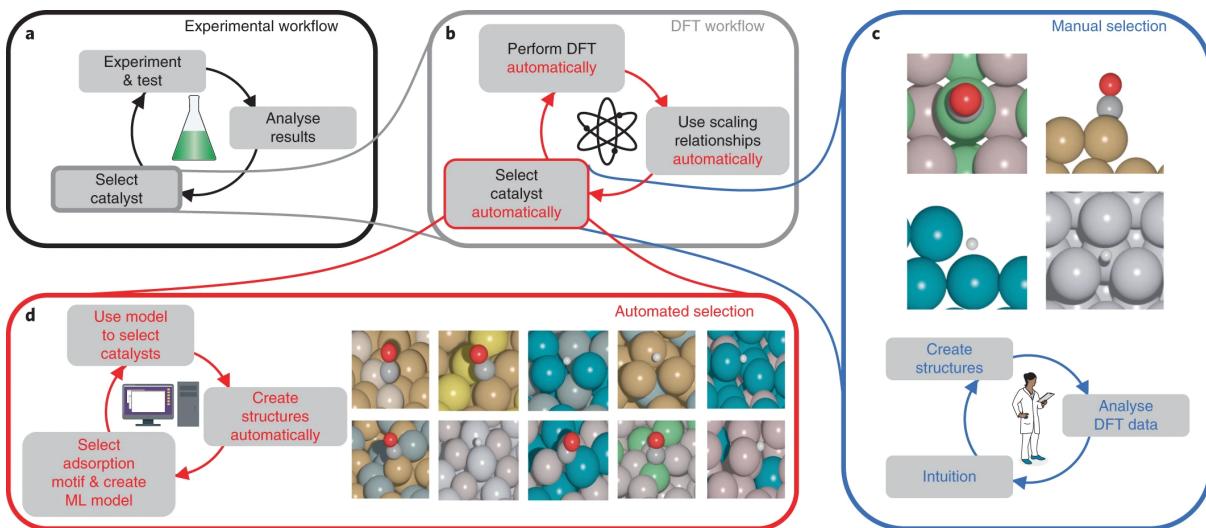
2. Feature importance analysis



- We predicted CER overpotential with inexpensive features and analyzed them with feature importance.
- The number of valence electrons and d-band center of metal can play a significant role for CER overpotential.

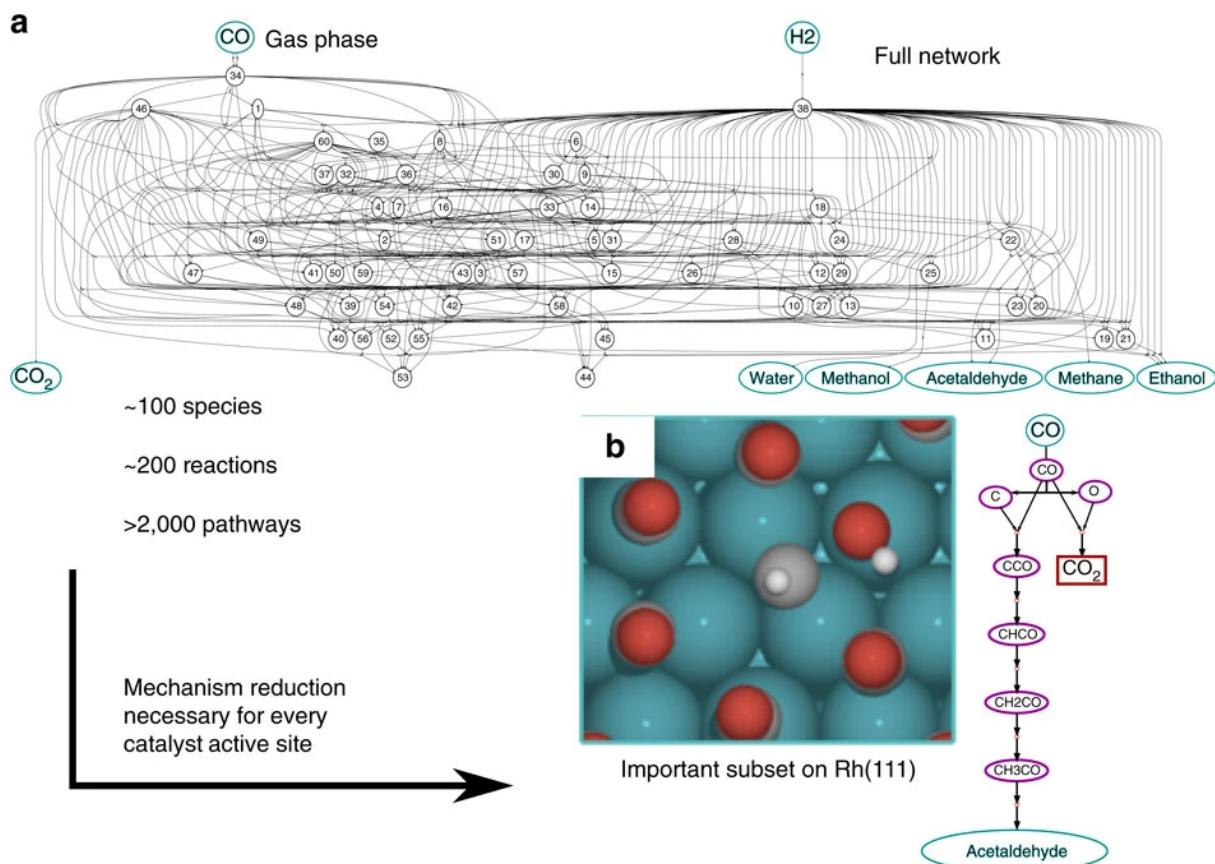
<https://doi.org/10.1016/j.cattod.2023.114358>

3. Active Learning



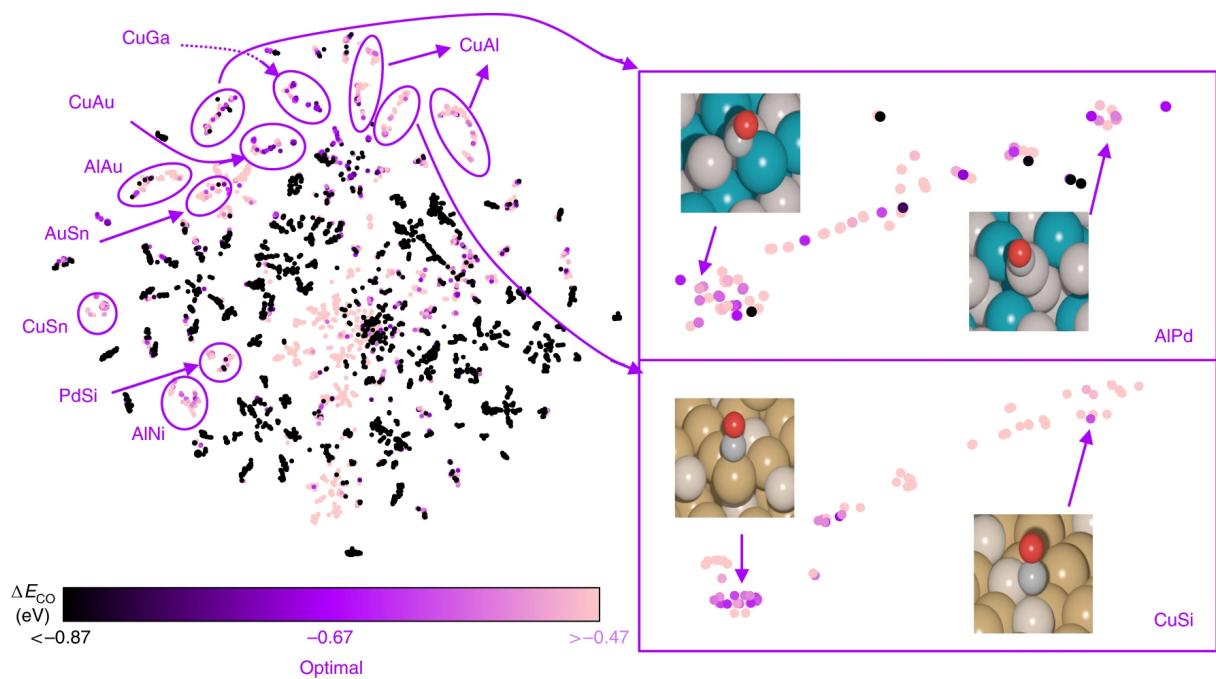
<https://doi.org/10.1038/s41929-018-0142-1>

4. Reaction network



<https://doi.org/10.1038/ncomms14621>

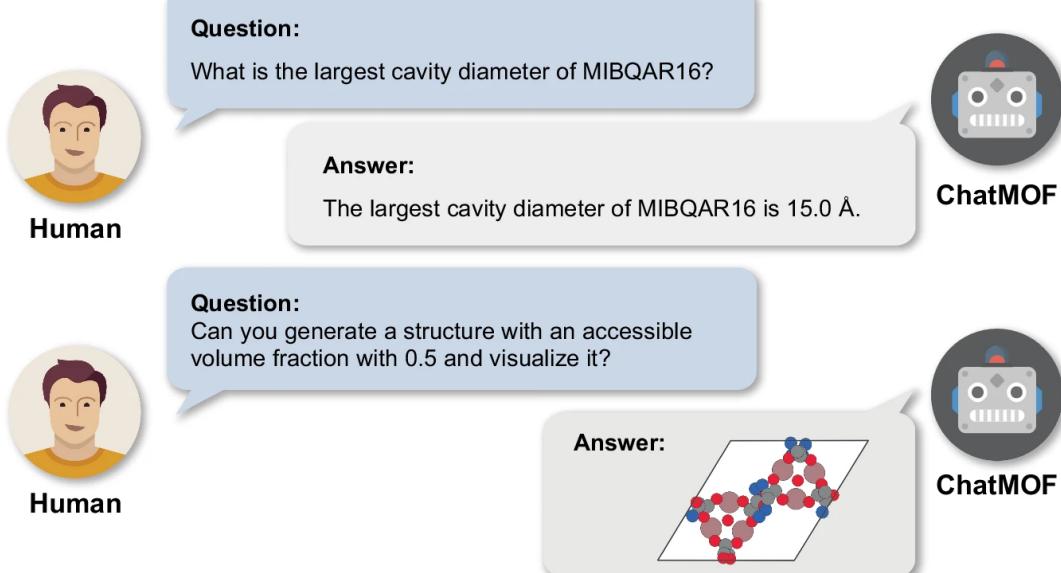
5. PCA & t-SNE analysis



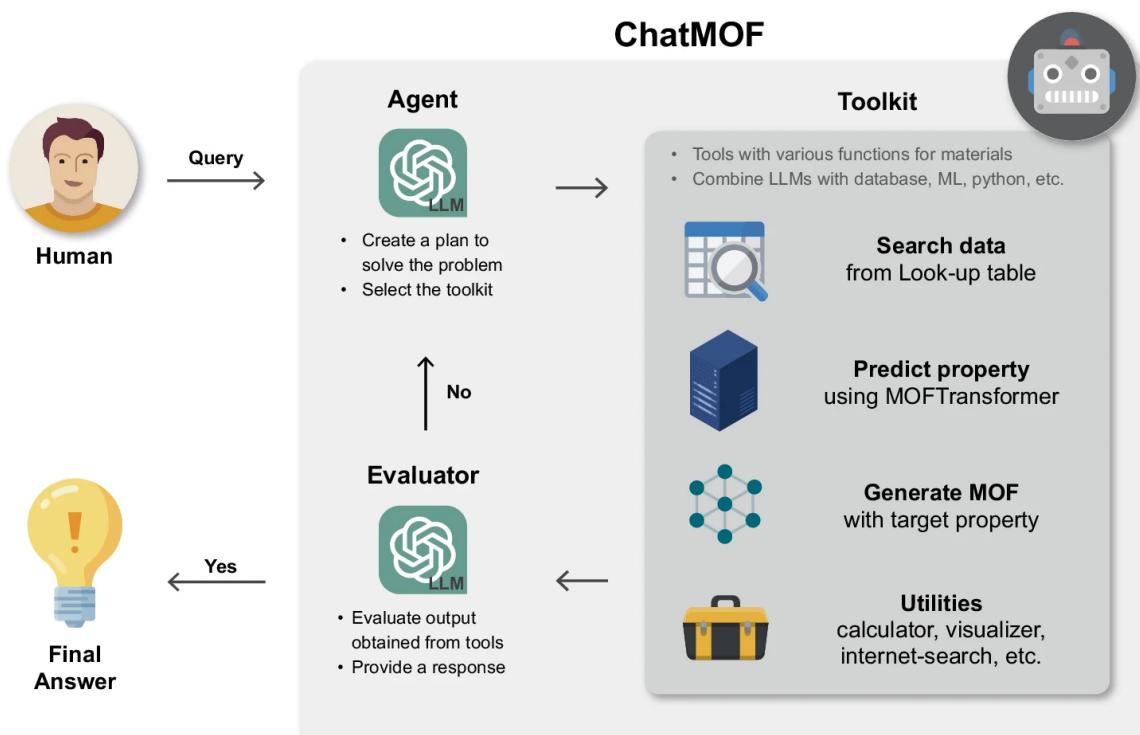
<https://doi.org/10.1038/s41929-018-0142-1>

6. LLM

(a)



(b)



Question: Find the MOF with the highest hydrogen diffusivity at 77 K, 1 bar.

Thought: I need to find the MOF with the highest hydrogen diffusivity at 77 K, 1 bar.

Action: search_csv

...

Thought: The search_csv tool did not provide any information on the hydrogen diffusivity of MOFs at 77 K, 1 bar. I need to find another way to obtain this information.

Action: predictor

Action Input: Predict the hydrogen diffusivity of MOFs at 77 K, 1 bar and find the name of MOF with highest diffusivity

[predictor] **Thought:** I need to gather data on the hydrogen diffusivity of MOFs at 77 K, 1 bar ...

[predictor] **Property:** hydrogen_diffusivity_dilute_77K

[predictor] **Materials:** *

bash

```
load model : hydrogen_diffusivity_dilute_77K/best.ckpt
Predicting DataLoader 0: 100% |██████████| 1246/1246
```

[predictor] **Final Thought:** Based on the result, find the MOF with the highest predicted hydrogen diffusivity at 77 K, 1 bar.

[Table Searcher] **Thought:** To find the MOF with the highest predicted hydrogen diffusivity, we can use the pandas function max() on the 'hydrogen_diffusivity_dilute_77K' column. However, since the values are in **logarithmic form**, we need to apply an exponential to get the original value. We can use the numpy function exp() for this.

[Table Searcher] **Input:**

python

```
import numpy as np

# Applying the natural logarithm to transform the values
df[“hydrogen_diffusivity_dilute_77K”] = np.exp(df[“hydrogen_diffusivity_dilute_77K”])

# Finding the maximum diffusivity value and its corresponding MOF (cif_id)
max_diffusivity = df[“hydrogen_diffusivity_dilute_77K”].max()
max_diffusivity_mof = df[df[“hydrogen_diffusivity_dilute_77K”] == \
max_diffusivity][“cif_id”].values[0]
```

[Table Searcher] **Observation:** BAZGAM_clean 0.0030176841738998412

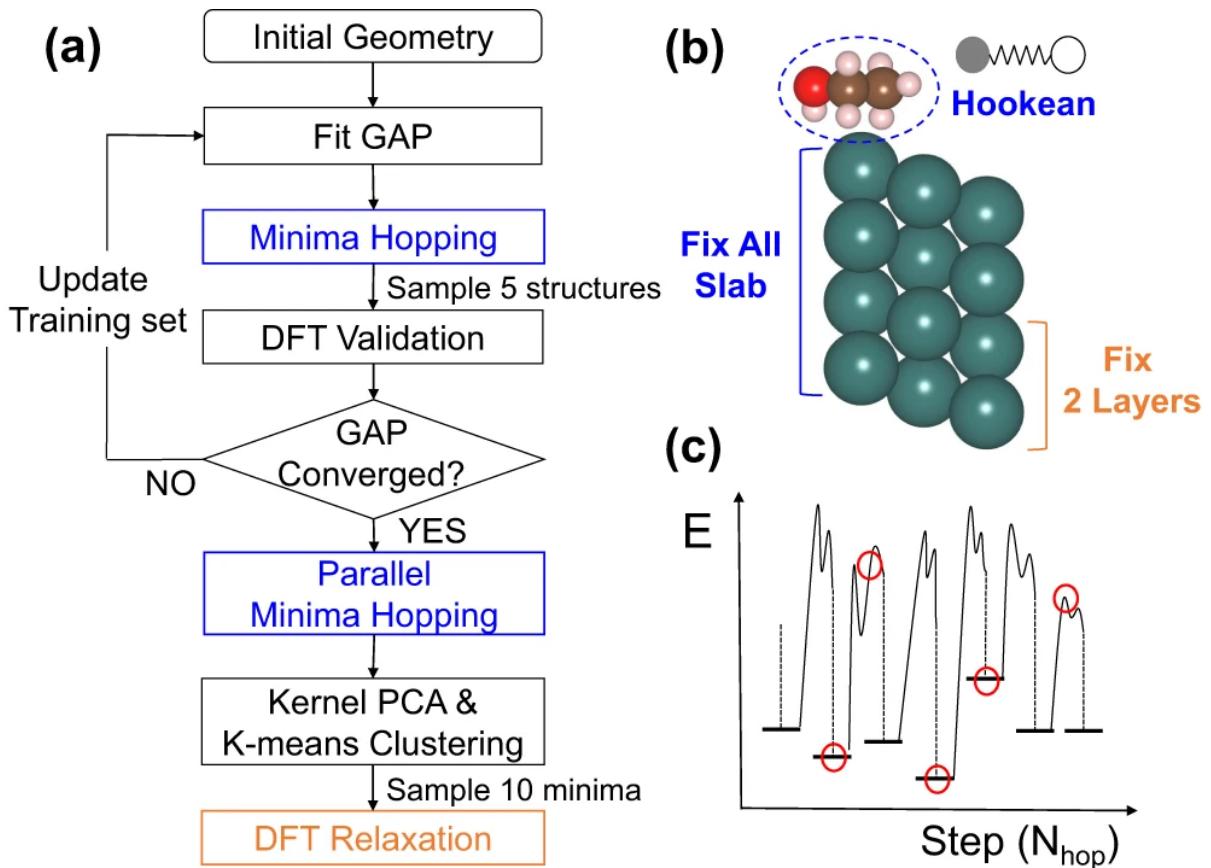
[Table Searcher] **Final Thought:** The MOF with the highest predicted hydrogen diffusivity at 77 K, 1 bar is 'BAZGAM_clean' with a diffusivity of 0.0030176841738998412 cm^2/s.

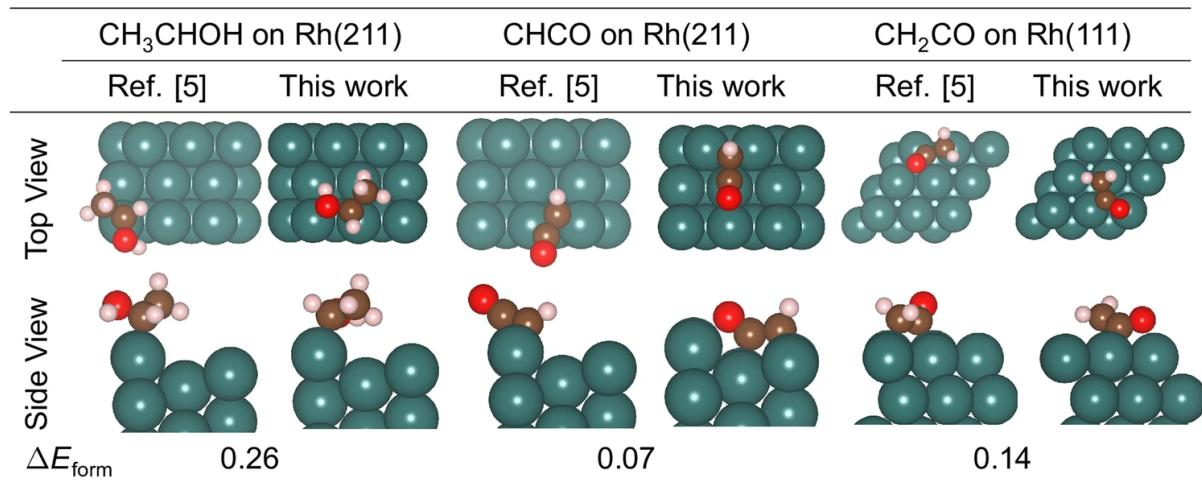
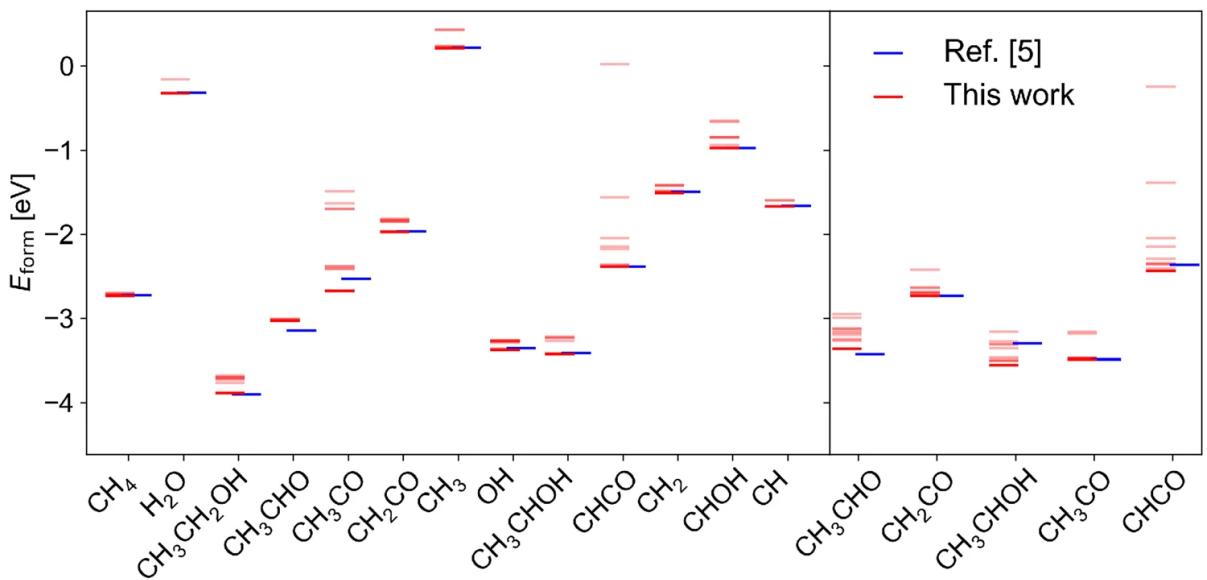
Thought: I now know the final answer

Answer : The MOF with the highest hydrogen diffusivity at 77 K, 1 bar is 'BAZGAM_clean' with a diffusivity of 0.0030176841738998412 cm²/s.

<https://doi.org/10.1038/s41467-024-48998-4>

7. Structure optimization





<https://doi.org/10.1038/s41524-023-01065-w>

Assignment

Please change all of parameters below

1. random_seed
2. n_estimators
3. activation
4. optimizer
5. unit_1 & 2 & 3

Run ML approach code, and attach the parity plots of 3 models and feature importance analaysis.

And you must specify all of the parameter settings you chose.

Assignment

Please change all of parameters below

1. random_seed
2. n_estimators
3. activation
4. optimizer
5. unit_1 & 2 & 3

Run ML approach code, and attach the parity plots of 3 models and feature importance analaysis.

And you must specify all of the parameter settings you chose.