Predictive Structure-Property modelling of diverse covalent organic frameworks using ensemble-based Random Forest Regressor

Niraj Bhatt,[1] Kushas Khadka,[2] and Andrew Medeiros[2]

[1]*Department of Mechanical, Industrial and Systems Engineering, University of Rhode Island, Kingston, RI 02881, USA*
[2]*Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881, USA*

(Dated: December 19, 2023)

We developed a machine learning multiple output regression model based on a powerful random forest algorithm that maps the structural and chemical features of the 5000 COFs used in training to their uptake values, heat desorption and deliverable capacities. The choice of random forest model is due to its superior generalizability when compared to other conventional ML algorithms, which is consistent with prior works in material informatics. We then validated the model by predicting the properties of our interest by selecting random 1000 COFs as the test dataset from the database which haven't been used in model training. We further explored other available experimental curated COFs database to predict the properties of our interest eliminating the need of resource intensive Grand Canonical Monte Carlo calculation. Thousands of diverse COFs would then be selected from the database and featurized using material libraries in python. The features are used to uniquely represent each COF which is mapped to their respective target properties of our interest during training of the ML model. The feature space includes structural features like radial distribution functions and density-based structural features. Various dimensionality reduction techniques like feature selection using correlation analysis and feature extraction using principal component analysis are experimented to make an accurate and efficient ML model.

I. INTRODUCTION

Natural gas is an important high-quality clean energy. The main component of the natural gas is alkanes. Among alkanes methane accounts for vast majority. The low carbon-hydrogen ratio in alkanes makes them less prone to pollution hazards. The advantages of CH_4's high calorific value, large reserves, and wide distribution make it a gradual alternative to oily fuels, gradually leading to a global energy transformation.[1,2] However, the biggest challenge in using CH_4 in natural gas applications is complications in its storage. In order to meet the storage requirements for CH_4,it is imperative to develop and design high-performance solid adsorbents. Covalent organic frameworks (COFs) have emerged as promising candidates for methane storage applications because of thier high surface area, flexibility in design, wide density range, porosity and high thermal stability. Covalent organic frameworks (COFs) are a major class of nanoporous materials as compared with MOFs and zeolites. COFs are defined by their composition: they include solely light elements, such as hydrogen, carbon, nitrogen, and oxygen. They have been widely used in the fields of gas storage, separation, interface chemistry, catalysis, and energy storage.[3,4] However, many types and huge amounts of COFs cannot be synthesized. An efficient way is to make a hypothetical database of COFs utilising the flexibility in their design and calculate their deliverable, thermal and working capacities. COFs with high deliverable capacities, high heat of desorption and high uptake (working capacity) are highly desirable. Grand canonical Monte carlo simulations (GCMC) have been used to calculate the working capacities of COFs.[5] This method is accurate but very time-intensive. Since it is not feasible to rely on traditional calculation methods to find suitable storage materials from the massive COFs, it is necessary to use advanced machine learning (ML) technology to solve the material calculation problems.

ML is a technology by which computers use collected data to build a prediction model and then use such a model to analyze unseen data. It is an important subfield of artificial intelligence (AI). Different ML algorithms are suitable for different problems. Multiple linear regression (MLR) can get the formula intuitively. A support vector machine (SVM) can avoid the "curse of dimensionality" to some extent. A decision tree (DT) needs a relatively small amount of calculation and can be easily converted into classification rules. Similarly, a random forest (RF) performs well for many data sets and has relatively high accuracy.[5] ML algorithms such as MLR, SVM, DT, and RF can solve many classification and regression problems, there are still many shortcomings. With the increase in the types of algorithms and problem complexity, researchers need to carefully choose the appropriate model architecture, training process, regularization method, and hyperparameters, all of which have great impact on the final performance of the algorithm. The process of building an accurate and powerful ML model requires advanced data processing and analysis skills. Choosing an appropriate method to solve the problem and configure optimal parameter values for a specific model is quite difficult. We have used random forest regressor in this project becauses of its insensitivity to training data, superior generalizability and accuracy and capacity to handle large number of features, which is consistent with prior works in the field of material informatics.

The application of ML in materials science is becoming more and more extensive. In many previous works, material researchers are keen to combine material research with ML. ML methods have become powerful tools for complex data analysis and information mining in materials science. The algorithm model of ML can quickly and accurately predict the performance indicators. Compared with the traditional calculation method of adsorption amount, it can save a lot of calculation time and has been successfully applied in materials science. According to the structural and chemical characteristics of the material, its adsorption performance can be accurately predicted, and impressive results have been achieved. However, getting an accurate prediction model is difficult for researchers. The generation of the model needs to go through complex feature engineering, algorithm selection, hyperparameter optimization, model training, and model evaluation, all of which is demonstrated in this work.

A. Data

Rocio et al., have reported a database of 69840 largely nobel Covalent organic frameworks(COFs) assembled in silico from 666 distinct organic linkers and 4 established synthetic routes to assess the methane storage performance which is the target property of our interest.[6] They have performed high throughput grand-canonical monte carlo (GCMC) simulation to calculate the various properties of each COFs like uptakes at various pressures, heat desorption and deliverable capacities. All the structures and properties are made available on open source materials platform Material Cloud, which we used to collect the input structures in the forms of .cif (crystallography information files). The detailed information regarding the nature of the data can be accessed using the following link: `https://archive.materialscloud.org/record/2018.0003/v2`. Since our database consists of cif files, we have used powerful inbuilt python libraries used in material's informatics namely MATMINER and PYMATGEN to extract the input features. We use PYMATGEN as well as MATMINER to featurize the COFs available in the form of crystallographic information files into a diverse set of chemical as well as structural features. The structural features include a set of density based features and radial distribution function, a distance-based geometric features. Chemical features include 132 distinct features, the density-based features include three distinct features like (Density , Vpa (Volume per atom) , packing fraction) and the distance based features include a set of 200 radial distribution function based features. Our training data consists of featurised-data of 5000 randomly selected COFs from the in-silico database. The overoll property distribution for all three target properties can be visualised from Figure 1. Similarly the property distribution in our training database is vsualised in Figure 2. Clearly, the randomly selected COFs database comprises of COFs that explore a great range of the target properties of our interest across low and high density ranges as explored by the in-silico database as shown in Figure 3.

II. METHODS

This section comprises explanation of the processes employed sequentially from feature engineering to hyperparamter tuning to final regression modelling. To streamline the preprocessing and modeling workflow, a pipeline was constructed using sklearn.pipeline.Pipeline. This pipeline sequentially applied the preprocessing steps, including PCA, median imputation, standard scaling, and polynomial feature transformation, followed by the Random Forest model with the best parameters obtained using GridsearchCV. The pipeline encapsulates the entire process, making it more modular, reproducible, and facilitating efficient deployment. Different features usually have different dimensions, and the values may vary greatly. Failure to process them may affect the results of data analysis. In order to eliminate the influence of the difference between the dimensions and the value range of the indicators, it is necessary to carry out standardization processing, and the data are scaled according to the proportion to make it fall into a specific area, which is convenient for comprehensive analysis.

A. Feature Engineering:

Feature engineering is an important concept in the ML field. It is the work of designing feature sets for ML applications, focusing on how to design features that match the characteristics of the data itself and the application scenario. Feature engineering usually includes feature extraction, feature selection, and feature construction. Feature extraction usually aims to reduce the dimension of features through some functional mapping, while feature construction is used to expand the original feature space. In addition, the purpose of feature selection is to reduce feature redundancy by selecting important features. Feature selection is the process of constructing feature subsets through the reduction of unrelated or redundant features from the original feature set, which is beneficial to simplify the model, thereby avoiding overfitting and improving model performance. Feature construction is the process of constructing new features from the basic feature space or original data to enhance the accuracy and generalization of the model. It's essence is to improve the representation of the original features. raditionally, this process highly depends on human expertise, while the commonlyusedmethodisdata preprocessing. Feature extraction is a dimension reduction process via certain mapping functions. This process extracts highly informative and nonredundant features according to certain specific indicators. Unlike feature selection, feature extraction usually changes the original feature.

1. Principal Component Analysis (PCA):

It was employed to reduce the dimensionality of the feature space. This technique was chosen to address multicollinearity and enhance computational efficiency. We performed PCA using the sklearn.decomposition.PCA module, retaining the top k principal components to capture a significant proportion of the variance.

2. Imputer:

To handle missing values within the dataset, a simple imputation strategy was applied using the median. The sklearn.impute.SimpleImputer was utilized, providing a prag-

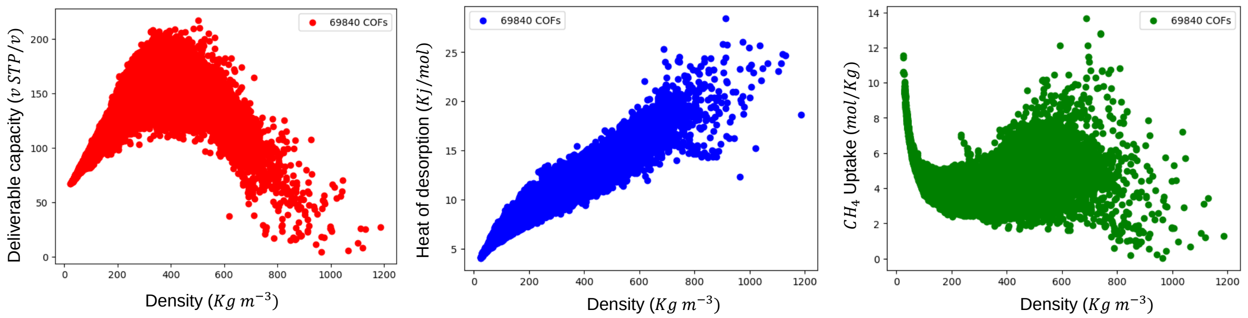


Figure 1. Visualising the three target properties of our interest namely deliverable capacity (figure on the left), heat of desorption(figure in the middle) and CH_4 Uptake (figure in the right) with respect to density of COFs of all the COFs present in the in-silico database. Clearly, the in-silico COF database comprises of COFs that explore a great range of the target properties of our interest across low and high density ranges.

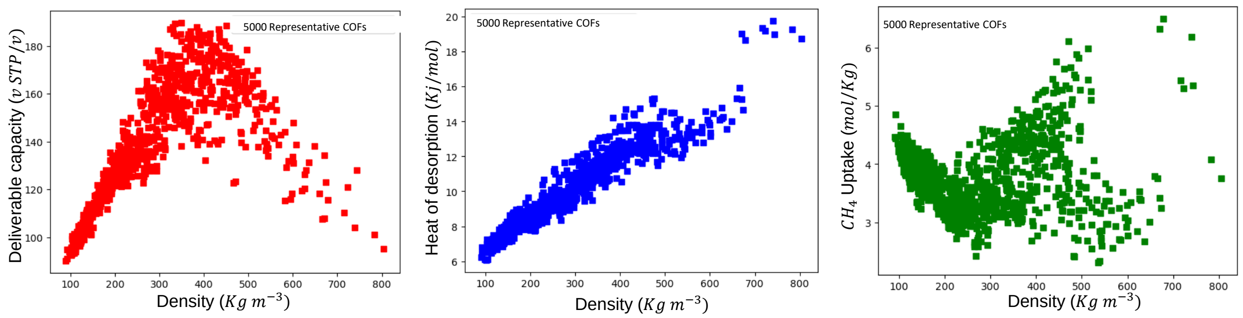


Figure 2. Visualising the three target properties of our interest namely deliverable capacity (figure on the left), heat of desorption(figure in the middle) and CH_4 Uptake (figure in the right) with respect to density of COFs for the COFs chosen for constructing our training as well as test database. We randomly choose 1000 Diverse COFs from the in-silico data base for the purpose of constructing a random feature space that serves as our custom database for training as well as test data. 20% of the COFs are used as test data to assess the accuracy of the model.Clearly, the randomly selected COFs database comprises of COFs that explore a great range of the target properties of our interest across low and high density ranges as explored by the in-silico database.

matic approach to maintain data integrity and facilitate downstream modeling processes.

3. *Standard Scaler:*

Feature scaling is critical for machine learning models, and the standard scaler was applied to normalize feature values. The sklearn.preprocessing.StandardScaler was employed to standardize each feature by removing the mean and scaling to unit variance, ensuring that all features contributed equally to the model.

4. *Polynomial Feature Transformation:*

To capture non-linear relationships within the data, a polynomial feature transformation was conducted. This involved generating polynomial features up to a specified degree using sklearn.preprocessing.PolynomialFeatures. The choice of the polynomial degree was based on empirical analysis to strike a balance between model complexity and performance.

B. Hyper-parameter Tuning:

In order to search for the best parameter configuration as quickly as possible, we set the grid search range for hyper-parameter optimization of ML algorithm. Due to the numerous parameters of the model, we only list some optimization ranges that have a greater impact on the model. To determine the optimal hyperparameters for the Random Forest model, a grid search approach was employed using sklearn.model_selection.GridSearchCV. This systematic exploration involved defining a hyperparameter grid containing various combinations of parameters, such as the number of trees, maximum depth, and minimum samples split. GridSearchCV performed an exhaustive search over the specified parameter grid, identifying the hyperparameter values that resulted in the best model performance.

C. Ensemble-based modelling:

The Random Forest algorithm was selected as the primary machine learning model due to its ensemble nature, robustness, and ability to handle complex relationships within the data. The sklearn.ensemble.RandomForestRegressor imple-

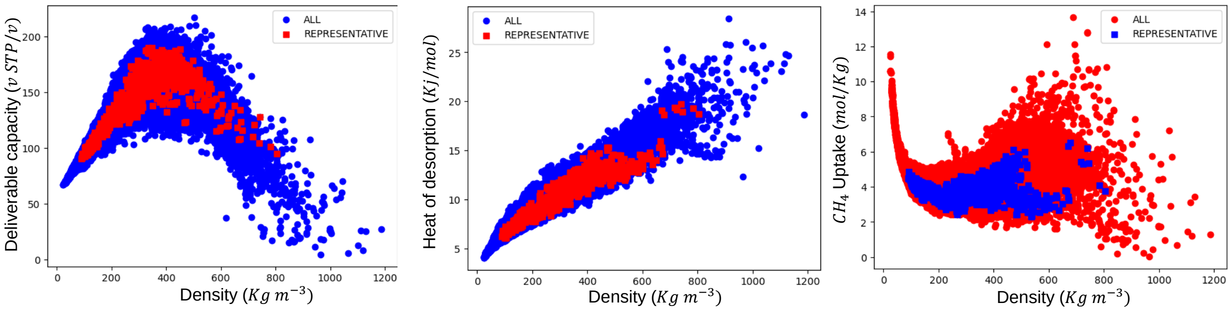


Figure 3. Visualising the scope of target properties of our interest namely deliverable capacity (figure on the left), heat of desorption(figure in the middle) and CH_4 Uptake (figure in the right) with respect to density of COFs in our randomly selected database with respect to the distribution in the total in-silico COFs database. The choice of our COFs significantly explore the in-silico database providing us confidence in using just 1000 COFs for training data.

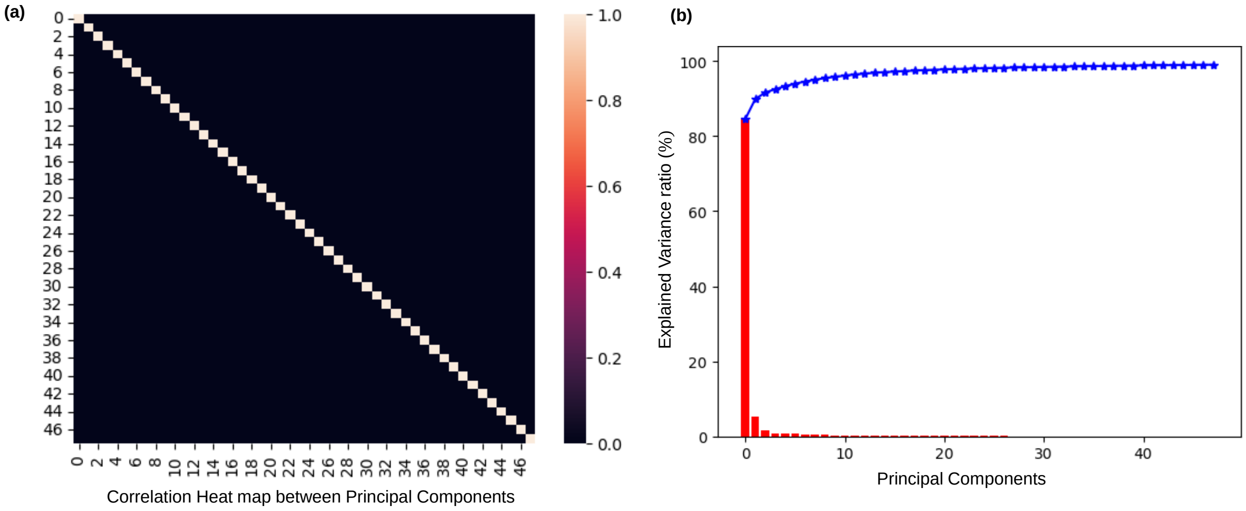


Figure 4. (left) Heat map illustrating the pearson's correlation coefficients between the 48 different principal components capturing 99% variance of the feature space. Clearly, application of principal component analysis (PCA) generates highly uncorrelated set of new features i.e, the principal components. This helps to eliminate the multicollinearity of the problem helping in achieving better accuracy and generalizibilty of the model. (right) Scree plot showing the explained variance ratio captured by principal components. Clearly the first few (10) of the principal components are sufficient to capture variance > 99%.

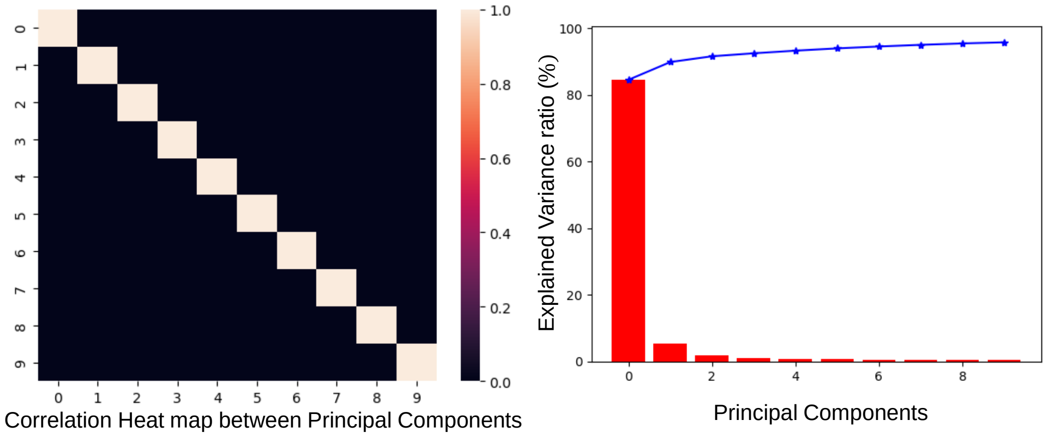


Figure 5. (a) Heat map illustrating the pearson's correlation coefficients between the 10 prime principal component's capturing over 95% variance of the feature space. (b) Scree plot showing the explained variance ratio captured by the selected first 10 principal components. Clearly the first few (10) of the principal components are sufficient to capture variance > 95%.

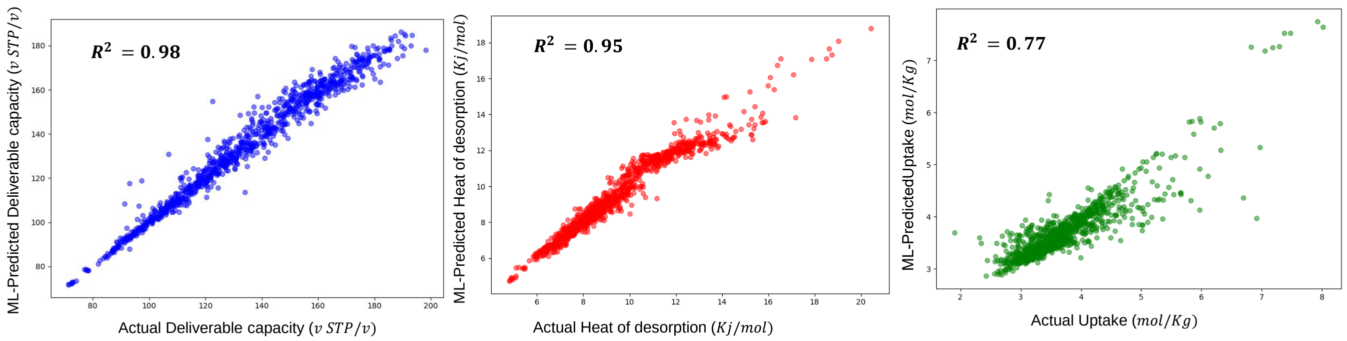
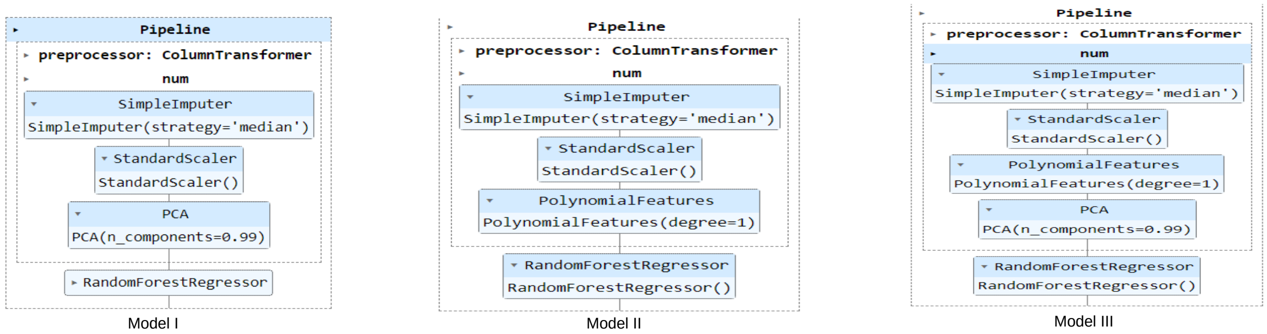


Figure 6. Comparison of the actual target properties with the ML-predicted target properties in test data. The test accuracy is assessed by value of R-squared as the evaluation metric. The model has highest accuracy in predicting the deliverable capacity followed by heat of desorption and CH_4 uptake respectively. The relatively low R-squared value for CH_4 uptake can be attributed to the fact that the training data consisted of COFs with lesser range of uptake values and thus higher correlation in prediction as compared to other properties. The overall R-squared value of the model is 0.90, indicating high accuracy of our multi-output regression model.


```
Pipeline
preprocessor: ColumnTransformer
num
SimpleImputer
SimpleImputer(strategy='median')
StandardScaler
StandardScaler()
PCA
PCA(n_components=0.99)
RandomForestRegressor
```
Model I

```
Pipeline
preprocessor: ColumnTransformer
num
SimpleImputer
SimpleImputer(strategy='median')
StandardScaler
StandardScaler()
PolynomialFeatures
PolynomialFeatures(degree=1)
RandomForestRegressor
RandomForestRegressor()
```
Model II

```
Pipeline
preprocessor: ColumnTransformer
num
SimpleImputer
SimpleImputer(strategy='median')
StandardScaler
StandardScaler()
PolynomialFeatures
PolynomialFeatures(degree=1)
PCA
PCA(n_components=0.99)
RandomForestRegressor
RandomForestRegressor()
```
Model III

Figure 7. Three different pre-processing experiments.

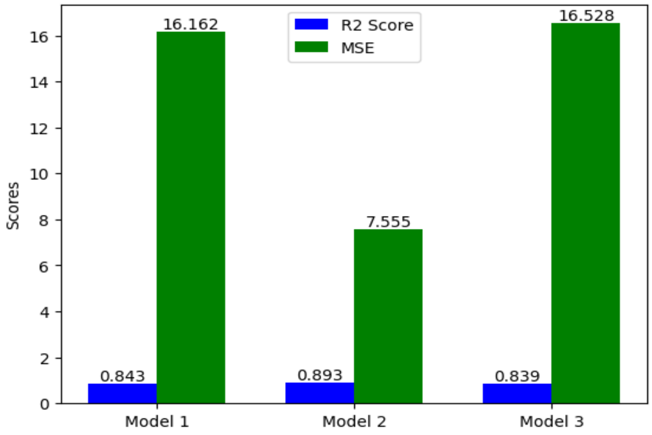


Figure 8. Mean squared error and value of R-squared value for three different model based on different data pre-processing strategies..

mentation was utilized, providing an ensemble of decision trees to enhance predictive accuracy.

D. Experiments

1. *Preprocessing experiments*

Three different model with different pre-processing sequences have been experimented as shown in Figure 8. The performance of each model was assessed through two key metrics: Mean Squared Error (MSE) and R2-squared Score. These metrics provide insights into the accuracy and goodness of fit for the models. The results are visually presented in the accompanying bar graph as shown in Figure 8, offering a comparative view of how each model performs across these evaluation criteria.

2. *Hyperparameter tuning experiments*

For the Random Forest Regressor, a comprehensive hyperparameter tuning experiment was conducted using Grid Search. The following hyperparameters were explored.

- n_estimators: The number of trees in the forest.
- max_features: The number of features to consider when looking for the best split.

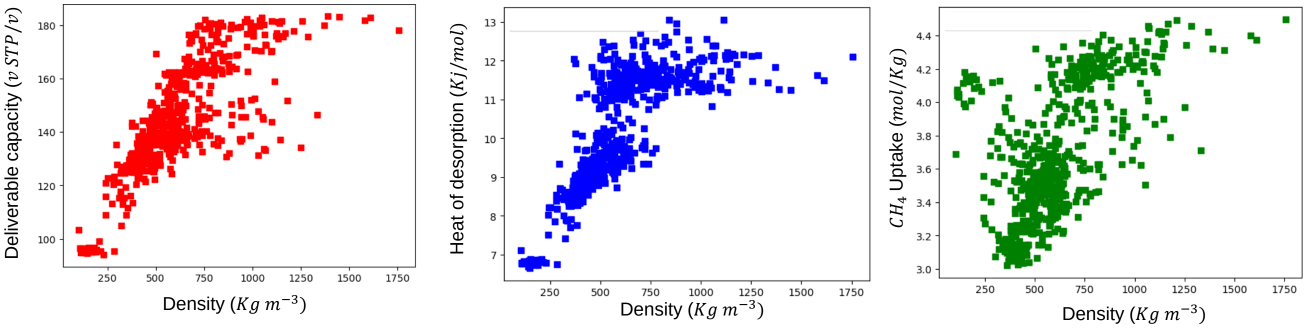


Figure 9. (left) ML-predicted deliverable capacity of experimental(curated) COFs. Clearly the deliverable capacity increases with increase in density upto 1000 Kgm^{-3} and then saturates on further increasing. (middle) The ML-Predicted heat of desorption for the experimental COFs shows monotonic increase of heat of desorption for COFs with increase in density. (Right) Similarly, ML-predicted uptake values for COFs also exhibit monotonic increase with density.

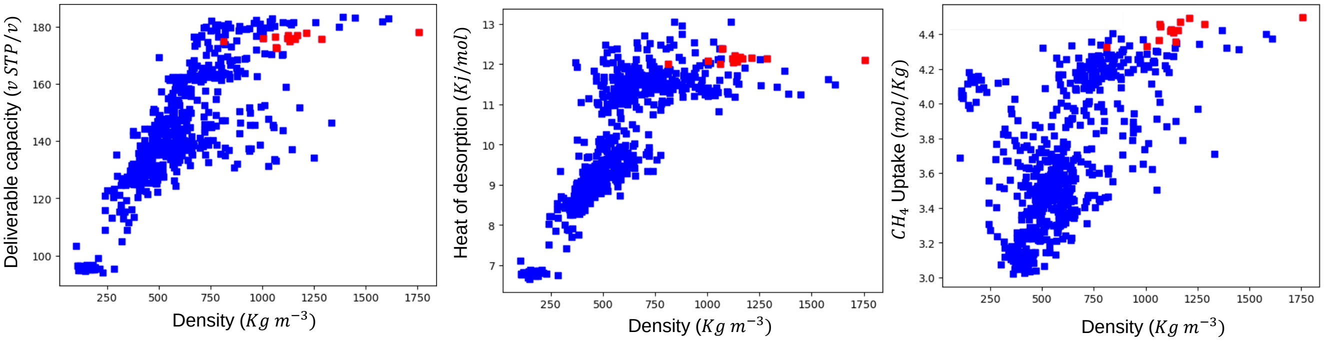


Figure 10. Screening the top performers having highest values for each target property of our interest i.e, deliverable capacity, heat of desorption and CH_4 uptake. The top performers are selected so that they have maximum values for each of the target properties. A total of 14 top performers are screened.

- ❖ max_depth: The maximum depth of the tree.
- ❖ min_samples_split: The minimum number of samples required to split an internal node.
- ❖ min_samples_leaf: The minimum number of samples required to be at a leaf node.
- ❖ Bootstrap: Whether bootstrap samples are used when building trees.

The best set of hyper parameters obtained is:

- ❖ n_estimators: 500
- ❖ max_features: None
- ❖ max_depth: 30
- ❖ min_samples_split: 2
- ❖ min_samples_leaf: 1
- ❖ Bootstrap: True

The best final tuned Random Forest Regressor with the identified hyperparameters were used in the Pipeline along with the preprocessed strategies to obtain the best machine learning model.

III. RESULTS AND DISCUSSIONS

After finalizing the best model pre-processing strategy and choosing the best set of hyperparamters, the model is validated using test data of 1000 COFs which were not used in the training. The model has highest accuracy in predicting the deliverable capacity followed by heat of desorption and CH4 uptake respectively as depicted in Figure 6. The relatively low R-squared value for CH4 uptake can be attributed to the fact that the training data consisted of COFs with lesser range of uptake values and thus higher corelation in prediction as compared to other properties. The overall R-squared value of the model is 0.90, indicating high accuracy of our multi-output regression model. Similarly we use the random forest model to predict the target properties for 613 diverse experimental COFs as shown in Figure 9. Clearly, the deliverable capacity increases with increase in density upto 1000 Kg m^{-3} and then saturates on further increasing. The ML-Predicted heat of desorption for the experimental COFs shows monotonic increase of heat of desorption for COFs with increase in density. Similarly, ML-predicted uptake values for COFs also exhibit monotonic increase with density. We then screened the top performers having highest values for each target property of our interest i.e, deliverable capacity, heat of desorption and CH_4 uptake as shown in Figure 10. The top performers are selected so that they have maximum values for each of the target properties. A total of 14 top performers are screened namely:

☆ linker101_NH_linker91_CO_bor_relaxed.cif

☆ linker108_CO_linker11_NH_sql_relaxed.cif

☆ linker108_CO_linker33_NH_sql_relaxed.cif

☆ linker108_CO_linker37_NH_kgm_relaxed.cif

☆ linker109_CO_linker12_NH_sql_relaxed.cif

☆ linker109_CO_linker16_NH_sql_relaxed.cif

☆ linker111_CO_linker63_NH_sql_relaxed.cif

☆ linker111_CO_linker64_NH_sql_relaxed.cif

☆ linker111_CO_linker78_NH_sql_relaxed.cif

☆ linker111_CO_linker79_NH_sql_relaxed.cif

☆ linker91_CO_linker10_NH_hcb_relaxed.cif

☆ linker91_CO_linker12_NH_hcb_relaxed.cif

☆ linker91_CO_linker3_NH_hcb_relaxed.cif

☆ linker91_CO_linker40_NH_hcb_relaxed.cif

IV. CONCLUSIONS

We explored a large database of over 69000 COFs and randomly selected 5000 COFs out of them. In doing so we make sure we explore the property space or target space as far as practicable. We explored various structural and chemical featurization methods and generated over 300 features to represent each COF uniquely. Variour data pre-processing techniques, dimensionality reduction techniques as well as hyperparameter tuning techniques was explored systematically with the help of pipelines. The model was validated by predicting the target properties of nearly 1000 COFs. We achieved an overoll R-squared value of 0.90 for our multi-output regression model. We explored a totally new experimental COF database by featurizing them exactly the same way we did for training database. The three target properties were predicted for the experimental database and the top performers were screened.

V. ACKNOWLEDGEMENT

We highly thank Dr. Christian Esteves for his encouragement and guidance in the project.

VI. CODE

The code and dataset are uploaded on our github page:`https://github.com/kushasupaya/cof_featurization_ml/tree/main?fbclid=IwAR1FNHcVgpdUeB0W5XAUQ5z8D0USYTnRT6AYiNUsIZMt6NV`

VII. CONFLICT OF INTEREST

The authors declare no conflict of interest.

[1] M. Pardakhti, E. Moharreri, D. Wanik, S. L. Suib, and R. Srivastava, ACS combinatorial science **19**, 640 (2017).

[2] H.-C. Zhou, J. R. Long, and O. M. Yaghi, "Introduction to metal–organic frameworks," (2012).

[3] S.-Y. Ding and W. Wang, Chemical Society Reviews **42**, 548 (2013).

[4] Y. Li and R. T. Yang, AIChE Journal **54**, 269 (2008).

[5] B. Assfour and G. Seifert, Microporous and mesoporous materials **133**, 59 (2010).

[6] R. Mercado, R.-S. Fu, A. V. Yakutovich, L. Talirz, M. Haranczyk, and B. Smit, Chemistry of Materials **30**, 5069 (2018).