The prediction of gas isotherms by the structures of Metal-Organic frameworks

Group 6: Wenqin You, Zhe Liu

1 Background

Adsorption is one of the most essential processes in chemical engineering. Adsorption is the adhesion of atoms, ions or molecules from gases, liquid to a surface. Adsorption is usually described through isotherms, which means the amount of adsorbate adsorbed on the adsorbent. Gas isotherms are mainly a function of its pressure and temperatures. Each isotherm data point stands for an adsorption equilibrium at specific temperature and pressure. Adsorption is a consequence of surface energy and isotherms should be related to the structure of adsorbent itself. Metal-organic frameworks (MOFs) are good candidates as the adsorbents for adsorptive separation because their high surface area, porosity, and functionality. The search for the optimal MOF requires aggressive screening of a variety of MOFs. Since the structure can also affect the properties of the adsorbent, it will be interesting for data science to predict the gas isotherms by the analysis of the structures of adsorbents without performing an experiment or running grand canonical Monte Carlo (GCMC) simulation.

This project will first clean the NIST adsorption database, which contains extensive isotherm data. Since the NIST database doesn't include the structures properties of MOFs, we will also link the NIST with other structure databases, for example, CoRE MOF. The structure information contains metal type, density, pore-limiting diameter (PLD), largest-cavity diameter (LCD), volumetric surface area (VSA), gravimetric surface area (GSA), pore volume, and void fraction. We will predict the isotherms by using the structure information and the corresponding adsorption type of the MOFs.

2 Data Description & Challenge

The NIST adsorption database contains around 20,000 adsorption isotherms. It includes 6,000 adsorbents and 300 adsorbates, various temperature and pressure range. The first challenging is that the volume of data is far too large to deal with in an Excel spreadsheet. The isotherms are published by different research groups with either simulation or experimental measurement. The second challenging is that we have to clean the database first, for example, removing some outliers. The third challenging is that we need to find a way of linking NIST database with CoRE MOF database since NIST doesn't contain the structure information. Finally, the adsorbents have different common name or code reference. It will take some efforts to build connection among these databases.

3 Hypotheses/Goals

Goal 1: Combine the NIST adsorption database with CoRE MOF database

The NIST adsorption database doesn't contain structure information about adsorbents. Extraction the structure properties from other databases and adding those features to the NIST adsorption database allows us to explore the correlation between isotherms from various adsorbents and adsorbates.

Since there are 6000 adsorbents in NIST, we will first focus on the metal-organic frameworks porous materials in NIST. As we know, NIST has lots of adsorption isotherms for MOFs. By approximate estimation, we may find that around 200 MOFs have structures information in the CoRE MOF database. We believe we can have enough adsorption isotherms database in NIST because for each MOF people always measured the adsorption of various guest molecules at multiple temperatures. We will know the accurate available number of isotherms with clearly associated structures when we build a connection between NIST, CoRE MOF database.

Goal 2: Isotherm type assignment and PCA for CoRE MOF database

The target of Goal 2 is to identify the isotherm type for each adsorbent. The adsorption isotherm depicts the adsorption process occurring on the surface and reveals the pore structure of the adsorbent. First, we will put the isotherms into a " common T/P frame." To date, 15 different isotherm models were developed and by Brunauer empirical classification, five types of isotherms are observed for solid adsorbents. There is no isotherm type available in NIST, but we need it to predict the isotherm. So the first step is to label each isotherm file with its corresponding adsorption type. Langmuir and Freundlich models are used to fit each isotherm by linear regression. Then assign each isotherm with its model based on the higher R^2 value. Finally, we will only keep the isotherm dataset with R^2 value higher than 0.95.

The feature vectors of structure information contain metal type, density, pore-limiting diameter (PLD), largest-cavity diameter (LCD), volumetric surface area (VSA), gravimetric surface area (GSA), pore volume, and void fraction. The features of the target molecule won't be included in the feature vector since we plan to build independent models for each target adsorbate. The CoRE MOF database only provides the structure information of adsorbents, and there is no information of adsorbates. The atomic structure won't be used since it indeed made the project challenging. We first use PCA to reduce the dimensionality of the MOF structure properties by applying variable selection techniques for each model.

Goal 3: Isotherm classification

From NIST, each isotherm file is labeled with the isotherm model and its corresponding parameters. From the CoRE MOFs database, we have the structure information of the MOFs after PCA. The goal here is to combine these two databases together and predict isotherm type

for each adsorbate based on the structure information of the MOFs by machine learning algorithm. The target adsorbate would be nitrogen. We will find a classifier to predict the isotherm model based on the structure info by various classification method.

Goal 4: Predict isotherm type and adsorption parameters for the adsorbate and adsorbent based on multiple feature variables related to structure information of absorbent

Based on the result from goal 3, if the isotherm is Langmuir model, we can use the structure information to predict this pair of parameters, Q_0 and b (Q_0 is the adsorption capacity and b is the Langmuir constant). If it's Freundlich model, we will relate the structure properties with K and n (K is the Freundlich constant related to the adsorption capacity and n is just a constant). With these pair of parameters and the adsorption model, we can predict the isotherm curve.

In summary, for this regression model, the input will be the feature vector of MOF structures and the adsorption type, and the output is its corresponding adsorption parameters.

4 Definition of Success

Success is defined as a complete workflow that achieves all four of the goals of the project. The lowest level that can be deemed a success is producing a working code to create a new database combined with NIST and CoRE MOFs, to classify the isotherms into two categories, and to build models that can predict the adsorption parameters of each selected adsorbate regardless of accuracy. An expected level of success is that code will utilize a verification and validation strategy, and be able to predict an entire isotherm with R^2 no less than 0.75. A high level of success is defined as a model that can predict saturated loadings with R^2 no less than 0.90.

5 Deliverables

The key deliverable will be a Jupyter notebook that contains:

- Appropriately organized data set
- Code to reproduce all analyses
- Documentation of inputs/outputs of all functions
- Quantitative and visual assessment of classification accuracy
- Quantitative and visual assessment of regression accuracy
- Quantitative and visual assessment of prediction accuracy
- Written critical analysis of success/failures of the model