Prediction of Gas Adsorption Isotherm on Zeolite Materials by Using Machine Learning

LU, Chen
九州大学大学院総合理工学府総合理工学専攻機械・システム理工学メジャー

Master Thesis

Prediction of Gas Adsorption Isotherm on Zeolite Materials by Using Machine Learning

Thermal Energy Conversion Systems Laboratory,

Mechanical and Systems Engineering Major,

Interdisciplinary Graduate School of Engineering Sciences,

Kyushu University

LU CHEN

Supervisor(s) Miyazaki Takahiko

Date: 1st, February, 2024

Prediction of Gas Adsorption Isotherm on Zeolite Materials by Using Machine Learning

ABSTRACT

Zeolites find extensive application in adsorption heat pumps and exhaust gas separation treatment because their ion exchange and adsorption properties. Given the diverse range of zeolites, their adsorption performance for different gases varies significantly. When dealing with practical adsorption equipment and operational situations, the pivotal challenge lies in selecting adsorption materials with both high gas adsorption capacity and a rapid adsorption rate to enhance equipment efficiency. Identifying suitable gas adsorption materials through experimental trial and error is a lengthy, expensive, and inefficient process. In recent years, the rapid development of computational simulation and machine learning technology has opened promising prospects for computer-aided material design.

The CIF file is a kind of information file describing the new standard crystallography, which records the species and three-dimensional coordinates of each atom in the zeolite cell in detail. This 3D structural information contains various physicochemical properties of the zeolite itself, therefore, in this paper, we consider the use of deep learning techniques to explore the constitutive relationship between the 3D structural information and the gas adsorption performance.

Using machine learning methods, we can establish non-linear relationships between multiple physical and chemical features and gas adsorption performance. This allows us to identify the physical and chemical features of zeolites that significantly influence gas adsorption performance and subsequently make more accurate predictions for the specified gases.

In this paper, deep learning is applied for the first time to predict the adsorption values of zeolites for multiple gases. While the traditional GBDT model has been considered a good choice in previous machine learning for zeolite adsorption prediction, this study incorporates the GBDT model, the XGBoost model modified by GBDT, and deep learning ANN and DNN models for predicting the adsorption isotherms of zeolites, with a subsequent comparison of results.

The experimental results show that the DNN prediction model proposed in this paper predicts the adsorption values of zeolites on gases with an error of 10% from the true values, with an average absolute

error percentage (MAE) of 15% and proves that the prediction model designed by using a deep neural network is better than the traditional machine learning method in the model comparison experiments.

KEY WORDS: Zeolite, Adsorption isotherm, CIF file, Machine learning, Deeper neural network

CONTENT

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Research background

The continuous development of human society and the increasing level of industrialization have led to a series of pollution issues, with air pollution emerging as a significant global environmental problem. Common primary pollutants like carbon (CO_2) and ammonia (NH_3) are widely prevalent in the atmosphere. The threat posed by air pollution extends to both human health and the ecological environment, presenting a substantial challenge to our well-being and the balance of ecosystems.

In the field of energy and environmental science, porous materials, particularly zeolite-type porous materials, have become ideal candidates for gas adsorption, separation, and storage due to their unique microporous structure and chemical properties. Zeolite porous materials offer a highly controllable pore structure, offering excellent gas adsorption performance. Therefore, they hold broad application prospects in areas such as natural gas storage, carbon dioxide capture, and gas separation [1].

However, the comprehensive and effective evaluation of the adsorption properties of zeolite porous materials through experimental means is a time-consuming and labor-intensive task. Traditional experimental methods are resource-intensive and constrained by experimental conditions. To better understand and optimize the gas adsorption capacity of zeolite porous materials, researchers have increasingly turned to machine learning-based prediction and optimization methods in recent years.

Machine learning technologies, especially supervised learning, and regression analysis, provide an innovative approach to addressing the gas adsorption challenges of porous materials. By collecting a large amount of experimental data and combining it with the physical and chemical characteristics of zeolite materials, machine learning models can learn and predict the adsorption process of various gases in zeolites [2]. This data-driven approach provides researchers with a more comprehensive understanding and has the potential to accelerate the design and development of novel porous materials.

This study aims to explore in-depth the application of machine learning in the prediction and optimization of gas adsorption capacity in zeolite porous materials. By constructing accurate prediction models, we aim to achieve a precise grasp of the adsorption performance of zeolite porous materials and propose targeted optimization strategies to meet diverse gas adsorption requirements. This research

direction is expected to provide robust support for the design and synthesis of novel porous materials, advancing their applications in the field of energy and the environment.

1.2 Theory of adsorption

Adsorption is a surface phenomenon, involves the accumulation of liquid or gas on the surface of a liquid or solid [3]. The adsorption process is illustrated in Fig.1-1. The strength of adsorption forces is determined based on the interaction between the adsorbent and the refrigerant. Additionally, adsorption phenomena are generally categorized into chemical adsorption and physical adsorption based on the forces and interactions between the refrigerant and the adsorbent. Adsorbed molecules do not undergo chemical reactions, but they easily lower their energy when immobilized. This implies that the adsorption process is a phase transition process from fluid to adsorbent, known as an exothermic process. Furthermore, this process should be reversible. Physical adsorption occurs whenever the adsorbable gas contacts with the surface of the adsorbent.

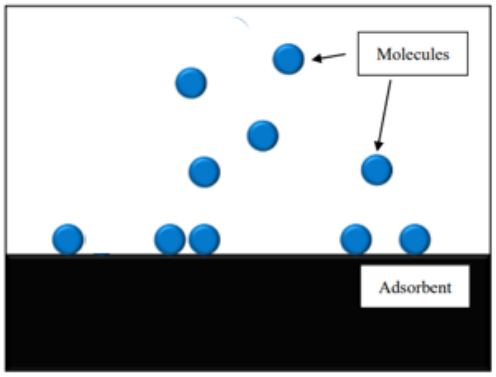


Fig.1-1 Adsorption/desorption process

The interface layer of the adsorbent can be divided into two regions: the adsorbent surface layer and adsorption sites, where adsorption enhancement can occur at the adsorption sites. The adsorbable gas (refrigerant) is commonly referred to as the adsorbate. In physical adsorption, the adsorption events can be caused by intermolecular forces, and these forces are related to van der Waals forces, which are considered equivalent to intramolecular cohesive forces.

Chemical adsorption is a bonding phenomenon between the surface particles of the material and the contacting substance such as gas or liquid. The heat released during chemical adsorption is

approximately 50 to 100 kcal/mol. Since the effective distance of van der Waals forces is greater than the effective distance of chemical reactions, physical adsorption occurs before chemical adsorption. Therefore, when refrigerant molecules approach the adsorbent, physical adsorption will occur first, transitioning into chemical adsorption as the distance decreases [4].

1.3 Zeolite materials overview

Zeolite is a porous crystalline aluminosilicate characterized by a complex crystal structure that gives rise to specific-sized molecular pores [5]. It can be synthesized by hydrothermal treatment of a caustic alkali synthesis solution containing a silicon dioxide source and a suitable structure-directing agent. In practical applications, some widely used zeolite materials closely resemble a pure silicon structure, as depicted in Fig.1-2. Zeolites can exhibit a rich variety of topological structures formed by silicon-oxygen tetrahedra through different connecting methods [6].

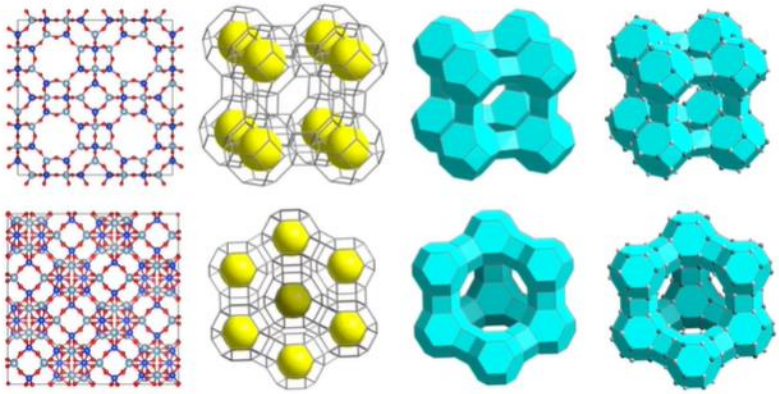


Fig.1-2 Zeolite Topological structure

The skeletal structure of zeolite resembles a honeycomb, facilitating the selective adsorption of molecules with different shapes and sizes. Molecules with diameters smaller than or approximately equal to the pore size of the molecular sieve can diffuse into the cavities of the molecular sieve, while molecules with kinetic diameters much larger than the pore size are blocked outside the crystal pores of the molecular sieve [7]. This adsorption selectivity of zeolites for different molecules is particularly valuable for gas separation purposes.

Zeolites not only possess a relatively high specific surface area within the range of 200-1000 m^2/g and a wide micropore volume [8], but they also exhibit chemical stability and high hydrothermal stability.

Even at elevated temperatures and lower adsorbate partial pressures, zeolites maintain high adsorption capacities. Additionally, zeolites demonstrate the ability to preferentially adsorb polar and unsaturated molecules [9].

1.4 Adsorption isotherm determination methods

Adsorption isotherm describes the equilibrium adsorption of a substance on a specific surface at a constant temperature [10]. When assessing the performance of adsorption materials comprehensively, the single-component adsorption isotherm of the target gas often represents the only available information. Therefore, the following introduces three methods for determining adsorption isotherms: experimental methods, molecular simulation methods, and machine learning methods.

1.4.1 Experimental methods

Experimental methods for determining the adsorption isotherms of single components can be broadly classified into two main types: static methods and dynamic methods. Among these, static methods offer higher precision and can be further divided into classical techniques such as the volumetric method and the gravimetric method [11], as well as the electromagnetic method and the oscillation method developed by Keller et al. [12]. Although the latter two exhibit high measurement precision, their higher costs have limited their widespread adoption. Volumetric and gravimetric methods, on the other hand, are more mature, and fully automated experimental setups are commercially available.

The volumetric method measures adsorption based on the relationship between gas volume and pressure, while the gravimetric method measures adsorption by monitoring the change in sample weight before and after adsorption. The gravimetric method tends to have larger measurement errors for adsorbates with low molecular weights, whereas the volumetric method exhibits the opposite trend. Additionally, dynamic methods have shorter measurement times compared to static methods, but they come with lower precision and are more susceptible to interference from equipment and other factors. Dynamic methods include the permeation method and the chromatographic method, where the permeation method calculates adsorption by simultaneously measuring the outlet gas concentration under certain inlet concentration conditions, and the chromatographic method measures retention time

by injecting a pulse of gas and calculates the equilibrium between the gas phase and the adsorption phase [13].

In general, traditional experimental methods for measuring adsorption isotherms offer a straightforward and intuitive approach in material research. However, these methods require a high level of expertise in equipment, experimental environments, and researchers, making the process labor-intensive and time-consuming.

1.4.2 Molecular simulation methods

Methods guided by computer simulations for measuring adsorption isotherms offer significant reductions in the time and cost required for experimental processes. Compared to experimental methods, theoretical calculations in molecular simulations facilitate the provision of realistic experimental environments, such as high temperature and pressure, since they allow complete control over relevant conditional variables. Moreover, molecular simulations enable the observation of phenomena and details that are not directly observable in experiments.

With the rapid development of computers, molecular simulation has paved a new path for quantitatively measuring adsorption equilibrium data for gases. Molecular simulation methods allow the study of molecular-level properties (such as surface area, pore size, or pore shape) and their impact on adsorption process [14]. Molecular simulations primarily include classical mechanics simulations and quantum mechanics simulations. Classical mechanics simulations rely on molecular mechanics, molecular dynamics, Monte Carlo simulations, and Brownian dynamics. Quantum mechanics simulations are based on ab initio methods, semi-empirical methods, and density functional theory. Among these, quantum mechanics, molecular dynamics, and Monte Carlo simulations are commonly used to simulate the adsorption processes in mesoporous/microporous materials. However, quantum mechanics methods are only suitable for handling simple molecules or systems with a small number of electrons. Even when simulations for larger systems are possible, most of the information needs to be discarded, and the computational time is excessively long. Molecular dynamics mainly simulates the process of individual atoms or molecules passing through the adsorbent material, and the simulation time for molecular diffusion processes is relatively long. Monte Carlo, on the other hand, can simulate

the adsorption positions of molecules without being affected by the diffusion process of guest molecules in the pores, and it has been widely used in the study of adsorption performance in zeolite materials.

Monte Carlo simulations typically employ ensembles such as the microcanonical ensemble, canonical ensemble, and grand canonical ensemble. The Grand Canonical Monte Carlo (GCMC) method is often used to simulate the adsorption behavior of gas on adsorbent materials [15]. This is because the grand canonical ensemble requires constant temperature, volume, and chemical potential, making it suitable for studying gas adsorption and separation phenomena. For example, Rahmati et al. [16] found a significant correlation between adsorption capacity and pore size by simulating the isothermal adsorption curves of hydrogen molecules in different zeolite structures at a certain temperature using the grand canonical Monte Carlo method. In the simulation process, adsorbed molecules can be removed from the system or new molecules can be inserted into random positions in the system using random movement methods (translation and rotation, etc.). However, as the number of materials increases, the computational load of the GCMC method also increases. Since computer simulations depend heavily on the microscopic structure of the materials involved, high-performance computing devices are required. Typically, these simulations run on large computing clusters, resulting in high time and cost requirements.

1.4.3 Machine learning methods

In recent years, the integration of Machine Learning (ML) methods has injected new vitality into the efficient and rapid discovery of materials, providing a time-saving approach. Machine Learning is a powerful tool capable of identifying patterns in high-dimensional data. Through using of ML algorithms, computers can simulate linear or nonlinear relationships between material properties and influencing factors, learning from empirical data. Against the backdrop of advanced materials characterization techniques and the ever-growing number of materials, the application of ML techniques and big data methods has successfully addressed the modeling challenges between material properties and complex physical factors [17]. The determination of adsorption isotherms for materials essentially involves extensive data analysis for different materials and combinations of pressure and temperature. Therefore, it is highly suitable for employing machine learning methods. Once appropriate descriptors for material structure and external conditions (such as temperature and pressure) are identified and standardized,

models capable of predicting adsorption isotherms can be trained through machine learning.

1.5 Previous research

Zeolite materials, known for their well-defined microstructure and uniform microporous morphology, showcase exceptional adsorption and separation performance, leading to their widespread applications. A crucial challenge in designing novel high-performance zeolite structures lies in establishing the relationship between pore size, pore structure, and adsorption/separation performance. The research group led by Yu [18] proposed utilizing structural analysis to obtain local atomic distances as descriptors, assessing the differences between various zeolite structures. They further introduced stacking sequence descriptors to characterize the three-dimensional structures formed by stacking different structural units [19]. Lin et al. [20], employing machine learning techniques, investigated the adsorption relationship between zeolite structures and siloxanes. They used lattice parameters, pore diameter, and specific surface area as descriptors, rapidly screening 230 zeolite structures with excellent performance. This approach can be applied for the rapid screening of high-performance adsorbents suitable for other organic or inorganic pollutants. In comparison to screening based on known structures, machine learning techniques can design structures with optimal performance based on specific requirements. Lee et al. [21], addressing the methane storage capacity limit in zeolites, introduced an energy shape descriptor that establishes a mapping relationship with the zeolite structure. By calculating the interaction energy between methane and zeolite, they obtained the energy shape and further trained a Generative Adversarial Network (GAN) to generate more energy shape diagrams. The predicted maximum CH4 capacity (156.74 cm^3/cm^3) closely matches the performance of ACO zeolite (159.41 cm^3/cm^3).

With the rapid advancement of artificial intelligence, materials science has entered a new research paradigm propelled by data-driven machine learning. Machine learning also proven effective in predicting materials adsorption. Jaffari et al. [22], used ten tree-based machine learning (ML) models, including XGBoost, GBDT, and DecisionTree to accurately predict the adsorption capacity of biochar materials toward ECs in aqueous solutions. The rigorous evaluation and comparison of the ML model performances shows that CatBoost model had the highest test coefficient of determination (0.9433) and lowest mean absolute error (4.95 mg/g), clearly outperformed all other models. Zhao et al. [23], used Four

artificial intelligence models to predict the maximum adsorption capacity of hydrochar and identify the key influencing factors. The gradient boosting decision tree (GBDT) showed excellent predictive capability for this study (R^2 = 0.93, RMSE =25.65). Yuan et al. [24], compiled a data set including 527 data points collected from peer-reviewed publications and applied machine learning to systematically map CO2 adsorption as a function of the textural and compositional properties of BWDPCs and adsorption parameters, and the gradient boosting decision tree (GBDT) had the best predictive performance with R2 of 0.98 and 0.84 on the training and test data.

The advantage of machine learning lies in its capacity to extract valuable information relevant to research goals from extensive computational or experimental datasets in high-dimensional data. Presently, due to the complexity of the relationship between material structure and function, as well as limitations in computational resources and experimental costs, the traditional approaches of relying solely on theoretical studies, experimental research, and computational simulations for the discovery of new materials are lagging the increasing demand for the development of high-performance materials driven by numerous emerging technologies. The speed of developing high-performance materials has become a bottleneck in the technological progress of various fields [25]. Therefore, using machine learning-established models for predicting material properties in the design and screening of materials can effectively enhance the efficiency of developing new materials.

1.6 Thesis outline

This thesis includes six chapters, and is organized as follows:

Chapter 1 provides an overview of the background and significance of the research, presenting the current state of research on zeolite performance using machine learning methods. The chapter offers a comprehensive introduction to the main research content of the thesis and outlines the overall structure of the document.

Chapter 2 delves into the types of zeolite adsorption isotherms, detailing the preparation of data for machine learning and the use of CIF files. It covers the collection of adsorption isotherm data, the storage of basic spatial structural information of zeolites in CIF files and emphasizes the advantages of CIF files. The chapter then outlines processing methods for CIF files.

Chapter 3 focuses on feature generation, traditional machine learning Decision Tree, and deep learning Neural Networks. It includes correlation analysis and data preprocessing related to zeolite adsorption features. Additionally, it introduces four machine learning methods used in the research.

Chapter 4 presents three hypothesized zeolite utilization situations, divides the usage data for each situation, and develops machine learning models. The chapter concludes with a comparison of the strengths and weaknesses of the models, determining the most appropriate model for each situation.

Chapter 5 provides a summary of the research content and achievements, offering reasonable recommendations. The chapter concludes with suggestions for future work.

Chapter 2: Data preparation

2.1 Collection adsorption data

2.1.1 Adsorption isotherm

The isotherm study is the first step to determine the adsorption capabilities of adsorbent materials. Adsorption, in this context, refers to the quantity of refrigerant adsorbed on the adsorbent as a function of gas-phase pressure at constant temperature along an isotherm. Majority of gas and/or vapor adsorption isotherms are classified based on the categorization scheme established by the International Union of Pure and Applied Chemistry (IUPAC). The isotherms classification is illustrated in Fig.2-1.

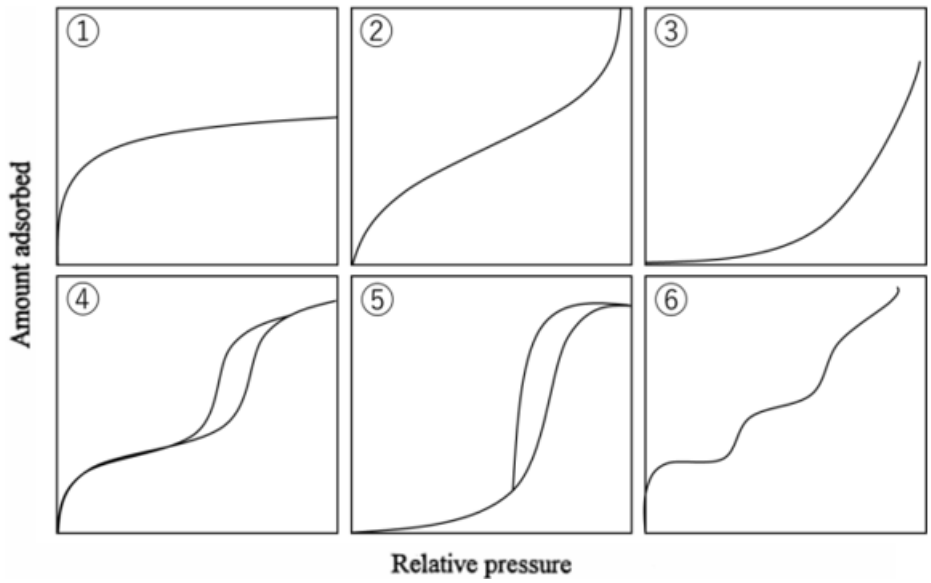


Fig.2-1 The IUPAC classification of adsorption isotherms showing both the adsorption and desorption pathways

For adsorption isotherm type, type I isotherms indicate monolayer adsorption, type II isotherms imply multilayer adsorption, type III isotherms indicate the coexistence of monolayer and multilayer adsorption, type IV isotherms reflect capillary condensation and pore blockage in mesoporous materials, type V isotherms are associated with materials with macropores or large pores, and type VI isotherms show a stepwise increase in the energy of adsorption corresponding to different adsorption energies, indicating the presence of a significant adsorption sites or multiple adsorption processes. These

classifications provide a powerful tool for studying the adsorption capacity of adsorbent materials, enabling scientists to systematically understand and compare adsorption process under different conditions [26].

2.1.2 NIST database

The National Institute of Standards and Technology (NIST) maintains a diverse and comprehensive database that plays a pivotal role in various scientific and technological domains [27]. The NIST database is particularly relevant in the context of machine learning applications, serving as a valuable resource for training and testing machine learning models. Datasets curated by NIST are often utilized in challenges and competitions, fostering innovation and advancements in machine learning algorithms.

The NIST database encompasses a wide range of datasets, standards, and reference materials, addressing the needs of researchers, engineers, and practitioners across different disciplines. Serving as a centralized repository, it facilitates access to high-quality data for benchmarking, validation, and calibration purposes.

The interface of the NIST database is shown in Fig.2-2 and Fig.2-3, which allows you to independently select the adsorbent pairs to obtain the adsorption isotherm data.

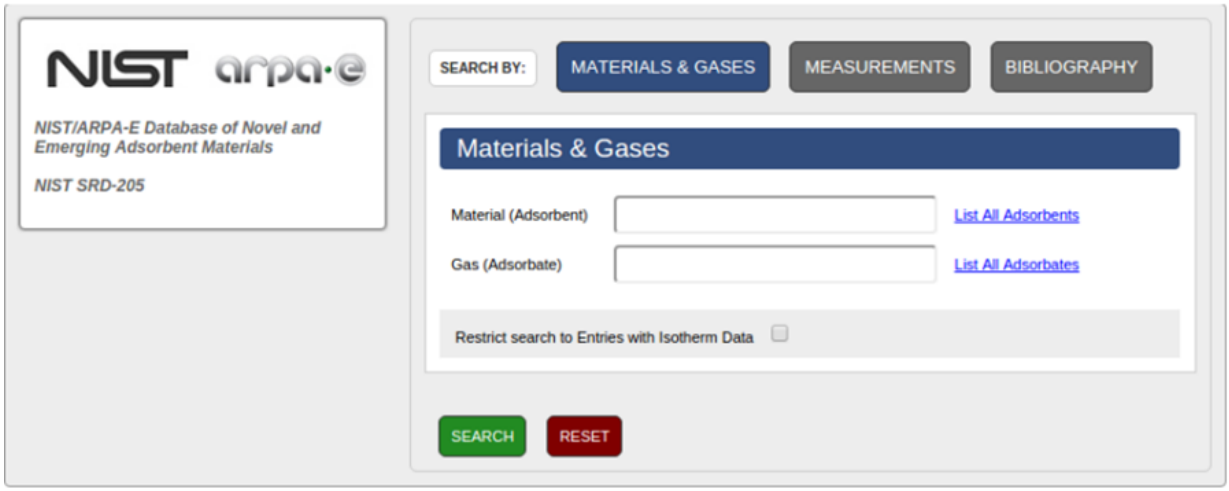


Fig.2-2 NIST database

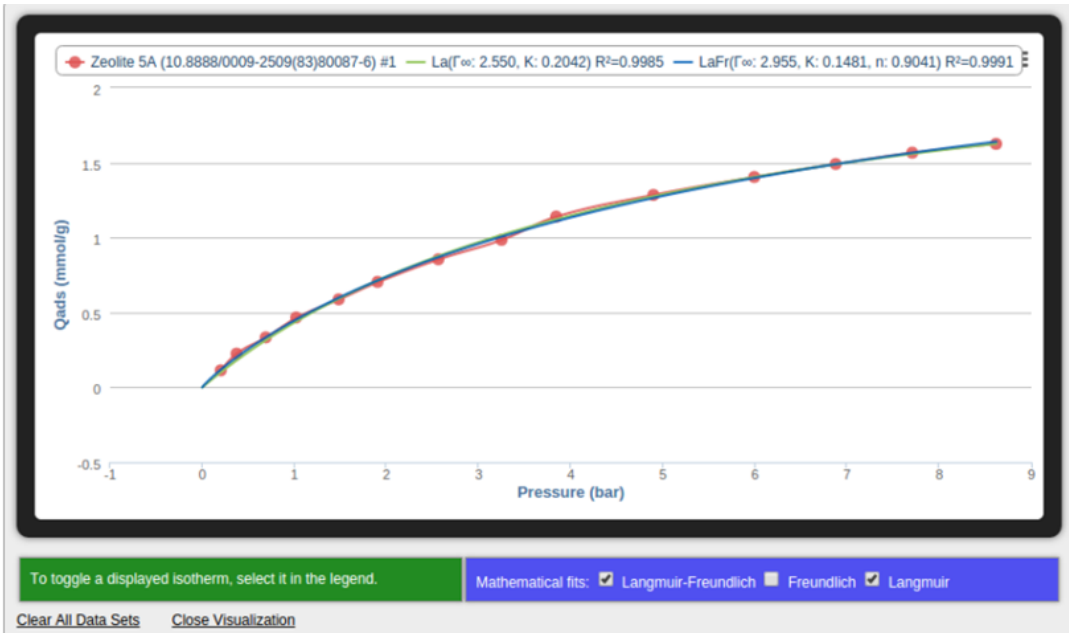


Fig.2-3 Adsorption isotherm

2.2 CIF storage format and application

2.2.1 Introduction of CIF format file

The CIF (Crystallographic Information File) is a mature format widely utilized in crystallography for data exchange and archiving. Since its inception, CI F and its associated applications have supported a broad ontology-based global framework, specifically designed for the exchange and processing of crystallographic data. The initial definition of CIF was presented by Hall et al. in a publication in 1991 [28].

The CIF format encompasses essential information regarding crystal structures, including parameters defining the unit cell, atomic coordinates, and quality indicators for the structural model. As illustrated in Fig.2-4, CIF files are widely used in many fields such as crystal information description, crystal structure mapping and theoretical calculations in material science. They play a pivotal role in the field, serving as a cornerstone for various applications.

Numerous crystal/material/computational processing software tools, including Vesta, Diamond, GSAS, Mercury, Materials Studio, Python, and others, utilize CIF files either as inputs or outputs. These files act as a standardized means of exchanging crystallographic information between different software

platforms, ensuring compatibility and interoperability.

In this context, our study employs CIF files as inputs for Python software to compute characteristic properties of zeolites.

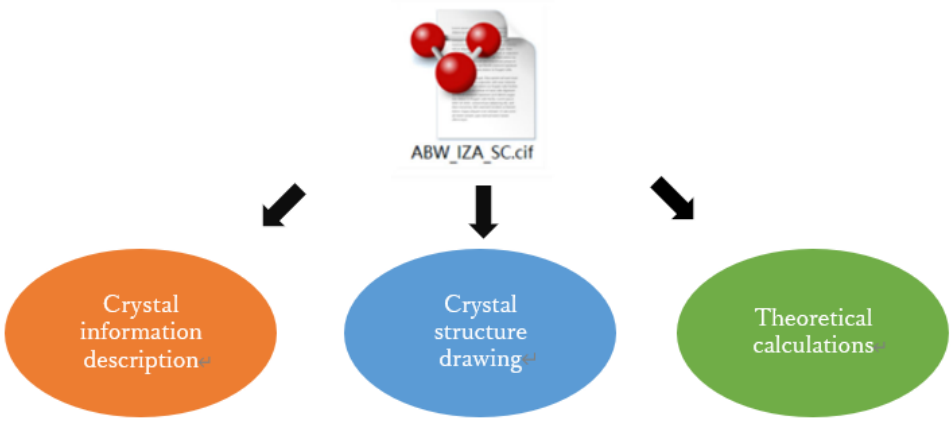


Fig.2-4 The application of CIF file

2.2.2 Application of the CIF file

With the growing popularity of software designed for creating CIFs, these files are now automatically generated in the author's laboratory and submitted electronically. CIFs prove valuable for crystal structure databases, enabling seamless integration of crystallographic information. A CIF file is shown in Fig.2-5 below:

```
data_Li-ABW
_chemical_name_systematic 'Li-ABW'

_cell_length_a 10.313
_cell_length_b 8.194
_cell_length_c 4.993
_cell_angle_alpha 90
_cell_angle_beta 90
_cell_angle_gamma 90
_cell_formula_units_Z 1

_space_group.name_H-M_ref 'P n a 21'
_symmetry_space_group_name_H-M 'P n a 21'
_symmetry_space_group_name_Hall 'P 2c -2n'
_space_group.IT_number 33

loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_B_iso_or_equiv
_atom_site_occupancy
```

Fig.2-5 Example of CIF file information

The CIF file states that the zeolite is named Li-ABW. Lines 3-9 describe the cell structure and other parameters of the zeolite. Lines 10-14 describe the space group of the zeolite. Lines 16-22 describe the atomic labels, the atomic symbols, the A-side scaled coordinates of the cell where the atom is located, the B-side scaled coordinates, and the C-side scaled coordinates.

For those who publish information about the structure of a crystal, CIF provides an easy way to record the crystal structure. The experimenter writes all numerical information and experimental details to the CIF using structure solver software. The information contained in the file can be read easily.

2.2.3 Collection CIF files

The acquisition of CIF files is the most basic and important step. Based on the research purpose of this paper, the download source of the CIF file is the IZA Structure Commission (http://www.iza-structure.org/) [29].

The International Zeolite Association Structure Commission (IZA-SC) is a global organization committed to the study and advancement of zeolite structure science. This commission is devoted to fostering cooperation and research in the field of zeolites, with the overarching goal of promoting the systematic study and understanding of zeolite structure.

The Structure Committee of the International Zeolite Association employs a systematic classification system that categorizes different types of zeolites into specific structural categories, including ABW, CHA, and others. This classification is based on distinctions in structural features and topological configurations of zeolite crystals, encompassing parameters such as zeolite structural characteristics, space groups, and experimental data.

As shown in the Fig.2-6, CIF files for various types of zeolites can be collected directly from the database.

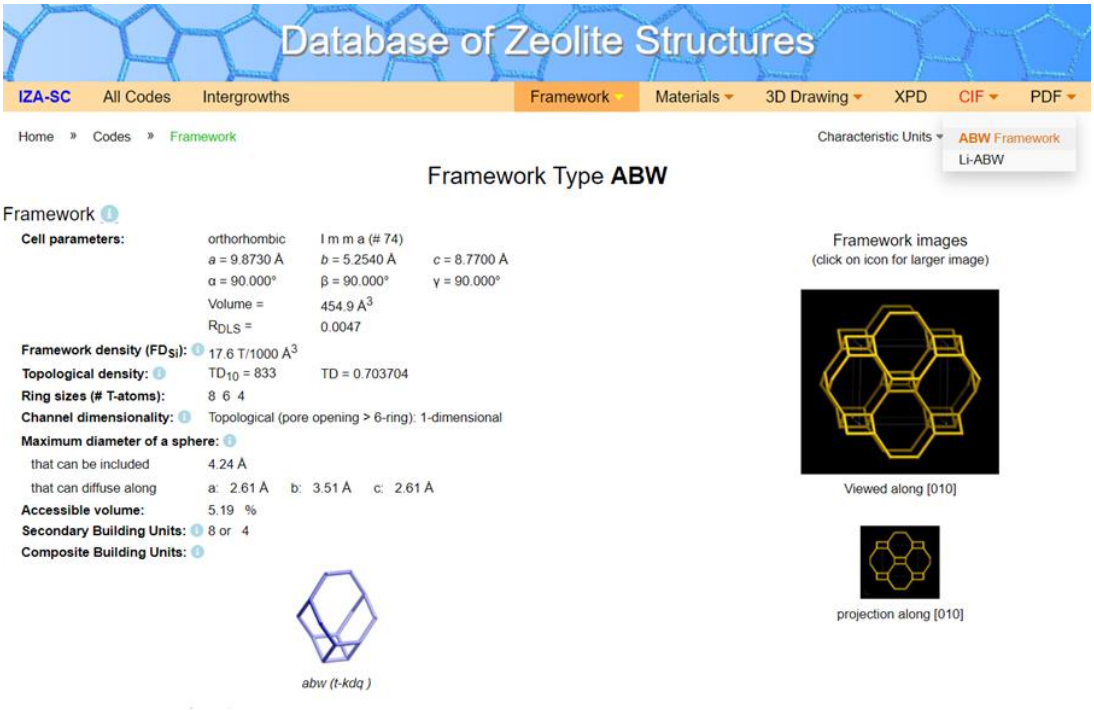


Fig.2-6 CIF file of zeolite ABW

2.3 Processing CIF files

2.3.1 Zeo++

This paper explores the application of the Zeo++ software package within the Python framework for the processing of CIF files.

Zeo++ is a software package designed for the crystal porous materials analysis [30]. Zeo++ provides a comprehensive geometric analysis of internal void structures and topology, as well as the assembly or alternation of structures. The software is versatile, capable of analyzing individual structures or conducting high-throughput analyses on large databases.

The tool calculates geometric parameters to describe voids, utilizing the Voronoi decomposition method. It offers a graphical representation of void spaces within a given atomic arrangement in the periodic domain. The resulting Voronoi network analysis yields geometric parameters, including the diameter of the maximum inscribed sphere and the largest free sphere—commonly used to characterize the geometric shape of voids.

In conclusion, Zeo++ stands out as a powerful tool for the geometric analysis of crystal porous materials. Its capabilities in calculating geometric parameters, conducting network analysis, and

supporting high-throughput processing make it valuable for researchers and practitioners in the field of materials science.

2.3.2 Running Zeo++

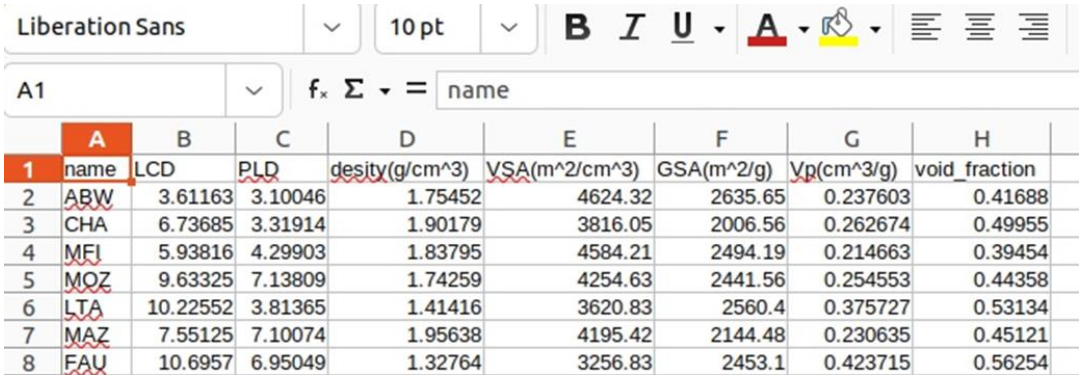
First, change the parameters in the configuration file according to your needs. Fig.2-7 shows the running parameter file of Zeo++.

```
[ZEO_CONFIG]
# zeo++ 的安装目录（The installation directory of zeo++）
zeo++_dir = /home/luxiuyang/zeo++-0.3

# cif文件所在目录，程序将遍历此目录下的所有cif文件并加以计算
# The directory where the cif files are located,
# the program will traverse all the cif files in this directory and calculate
cif_dir = /home/luxiuyang/RASPA_tools/test_cifs

# CPU核心数（Number of CPU cores on your computer）
number_of_threads = 3

# 计算比表面积所用的分子探针半径（Molecular probe radius used to calculate specific surface area）
radius_of_area_probe = 0

# 计算孔隙率、孔体积所用的分子探针半径（Molecular probe radius used to calculate porosity）
radius_of_porosity_probe = 0

# 用于计算比表面积的蒙特卡洛采样次数
#（The number of Monte Carlo samples used to calculate the specific surface area,
# in most cases does not need to be changed）
area_monte_carlo_samples = 2000

# 用于计算孔隙率的蒙特卡洛采样次数
#（The number of Monte Carlo samples used to calculate the porosity,
# in most cases does not need to be changed）
porosity_monte_carlo_samples = 100000

# 输出文件的名称(The name of the output file, in most cases does not need to be changed）
output_file_name = result.csv
```

Fig.2-7 Zeo++ running parameter

After setting the parameters to match your computer's configuration, run the Zeo++ package in Pycharm. Zeo++ calculates CIF files and generates structural representations, handles large CIF databases simultaneously, and generates material property parameters.

The result of the CIF files processing is shown in Fig.2-8. Physical parameters such as pore limit diameter (PLD) and largest cavity diameter (LCD) of zeolite can be calculated.



Liberation Sans | 10 pt | B I U A

A1 | f(x) Σ = name

	A	B	C	D	E	F	G	H
1	name	LCD	PLD	desity(g/cm^3)	VSA(m^2/cm^3)	GSA(m^2/g)	Vp(cm^3/g)	void_fraction
2	ABW	3.61163	3.10046	1.75452	4624.32	2635.65	0.237603	0.41688
3	CHA	6.73685	3.31914	1.90179	3816.05	2006.56	0.262674	0.49955
4	MFI	5.93816	4.29903	1.83795	4584.21	2494.19	0.214663	0.39454
5	MOZ	9.63325	7.13809	1.74259	4254.63	2441.56	0.254553	0.44358
6	LTA	10.22552	3.81365	1.41416	3620.83	2560.4	0.375727	0.53134
7	MAZ	7.55125	7.10074	1.95638	4195.42	2144.48	0.230635	0.45121
8	FAU	10.6957	6.95049	1.32764	3256.83	2453.1	0.423715	0.56254

Fig.2-8 CIF files processing result

2.4 Conclusions

This chapter provides an overview of the types of adsorption isotherms and methods for collecting adsorption data for materials. It introduces CIF files and discusses the Zeo++ toolkit, a valuable tool for processing CIF files. Additionally, the chapter outlines how to use Zeo++ to process CIF files to obtain material physical property data.

Chapter 3: Machine learning

3.1 Feature generation

Achieving high-precision predictive outcomes from machine learning models requires the extraction of pertinent features from the data for prediction. A successful approach in crafting features suitable for adsorption equilibrium predictions involves referencing the physical phenomena of adsorption and the inherent physical properties of materials [31]. This method is considered effective in capturing the intricacies of adsorption processes and enhancing the overall predictive accuracy of the model.

3.1.1 Adsorption temperature

The adsorption temperature of the adsorbent is utilized as a feature, as it constitutes a physical quantity associated with both adsorption capacity and energy. The temperature used in subsequent feature calculations refers to the adsorption temperature of the adsorbent.

3.1.2 Adsorption pressure

The pressure of the adsorbent P is used as a characteristic quantity, as it is a function of the adsorption value.

3.1.3 Adsorption value

Equilibrium adsorption value is the objective to be predicted in this study.

3.1.4 Pore limit diameter

Pore limit diameter (PLD) refers to the minimum bottleneck or opening diameter in the pore structure of a material [32]. This parameter is essential for measuring the size of microscopic pores and is critical for understanding molecular sieving capabilities and diffusion restrictions.

3.1.5 Largest cavity diameter

Largest cavity diameter (LCD) represents the maximum diameter of cavities or pores within the material structure. It serves as a crucial indicator for understanding permeability, adsorption characteristics, and catalytic activity in material science applications.

3.1.6 Density

Density is the ratio of mass to volume in a substance. It provides information about the distribution of material mass. Solids usually have a higher density, while gases have a lower density [33].

3.1.7 Volumetric surface area

Volumetric surface area (VSA) represents the surface area per unit volume [34]. In the context of porous materials, VSA characterizes the internal surface area, offering valuable insights applicable to areas like adsorption.

3.1.8 Gravimetric surface area

Gravimetric surface area (GSA) indicates the surface area per unit mass of the material [35]. This parameter is crucial for characterizing internal surface area in porous materials, with implications for catalysis and adsorption applications.

3.1.9 Pore volume

Pore Volume (PV) signifies the collective internal volume of pores per unit mass of the material. Comprehending PV is imperative for gaining in-depth insights into the pore structure and storage capabilities of materials.

3.1.10 Void fraction

Void Fraction (VF) denotes the ratio of pore volume to the total volume within a material, commonly

expressed as a percentage [36]. This metric provides valuable insights into the internal structure of materials, contributing significant information for analysis.

3.1.11 Adsorption potential

Adsorption value is inherently linked to the adsorption potential, a function influenced by various factors such as temperature (T), pressure of non-adsorbed gas (P), and the pressure of adsorbed layer liquid molecules at the saturation pressure (P_0). The energy required to transfer a unit mass of adsorbate from the gas phase to the adsorbed layer, denoted as E, is expressed by Eq. 3.1.

$$E = RT \ln \frac{P_0}{P} \qquad Eq.3.1$$

This E is referred to as the adsorption potential, representing the change in Gibbs free energy per mole during the adsorption process. It serves as a characteristic quantity providing insights into the energetics of the adsorption phenomenon [37].

3.1.12 Adsorbate type

An adsorbate refers to a substance or molecule that undergoes the process of adsorption on the surface of a solid or liquid material. In the context of this thesis, the adsorbents include water vapor, nitrogen, carbon dioxide, and other gases.

The surface of zeolite can contain various functional groups, enabling interactions with specific types of gas molecules. As a result, the adsorption capacity of zeolite varies for different gases. This inherent variability in the adsorption process emphasizes the importance of considering the specific adsorbate type when evaluating the adsorption properties of zeolite for various applications. [38].

3.1.13 Feature summary

Table 3.1 shows the features created. The total number of training data is 13068 in this paper. The features were mainly based on temperature, pressure, and material property.

Table 3.1 Machine learning features

No.	Feature name	Description	Description Term
1	Temp	Temperature	3.1.1
2	P	Pressure	3.1.2
3	Target	Adsorption value	3.1.3
4	PLD	Pore Limit Diameter	3.1.4
5	LCD	Largest Cavity Diameter	3.1.5
6	Density	Density	3.1.6
7	VSA	Volumetric Surface Area	3.1.7
8	GSA	Gravimetric Surface Area	3.1.8
9	PV	Pore Volume	3.1.9
10	VF	Void Fraction	3.1.10
11	E	Adsorption potential	3.1.11
12	Type	Adsorbate type	3.1.12

3.2 Correlation analysis

In machine learning, Correlation Analysis (CA) is a method used to measure the strength and direction of the relationship between two variables. Specifically, correlation analysis assesses the linear relationship between two variables by calculating the correlation coefficient. The most common correlation coefficient is the Pearson's correlation coefficient, which has a value ranging from -1 to 1 and indicates the strength and direction of the linear relationship between two variables [39].

Correlation analysis serves to unveil the relationship between features and the target variable, aiding in the identification of influential features for the task at hand. Features exhibiting high correlation may encapsulate crucial information, while those with low correlation might contribute minimally to the model's performance. However, when features exhibit high correlation among themselves, it can lead to

multicollinearity issues, impacting the stability and interpretability of the model. Correlation analysis is employed to detect and address such problems, selectively removing or merging highly correlated features. Furthermore, correlation analysis contributes to comprehending the interrelations among various features in the dataset. This is paramount for uncovering potential patterns, trends, or outliers, proving instrumental for subsequent data preprocessing.

Fig.3.1 shows the results of calculating the correlation coefficients between the predicted equilibrium adsorption and the feature values.

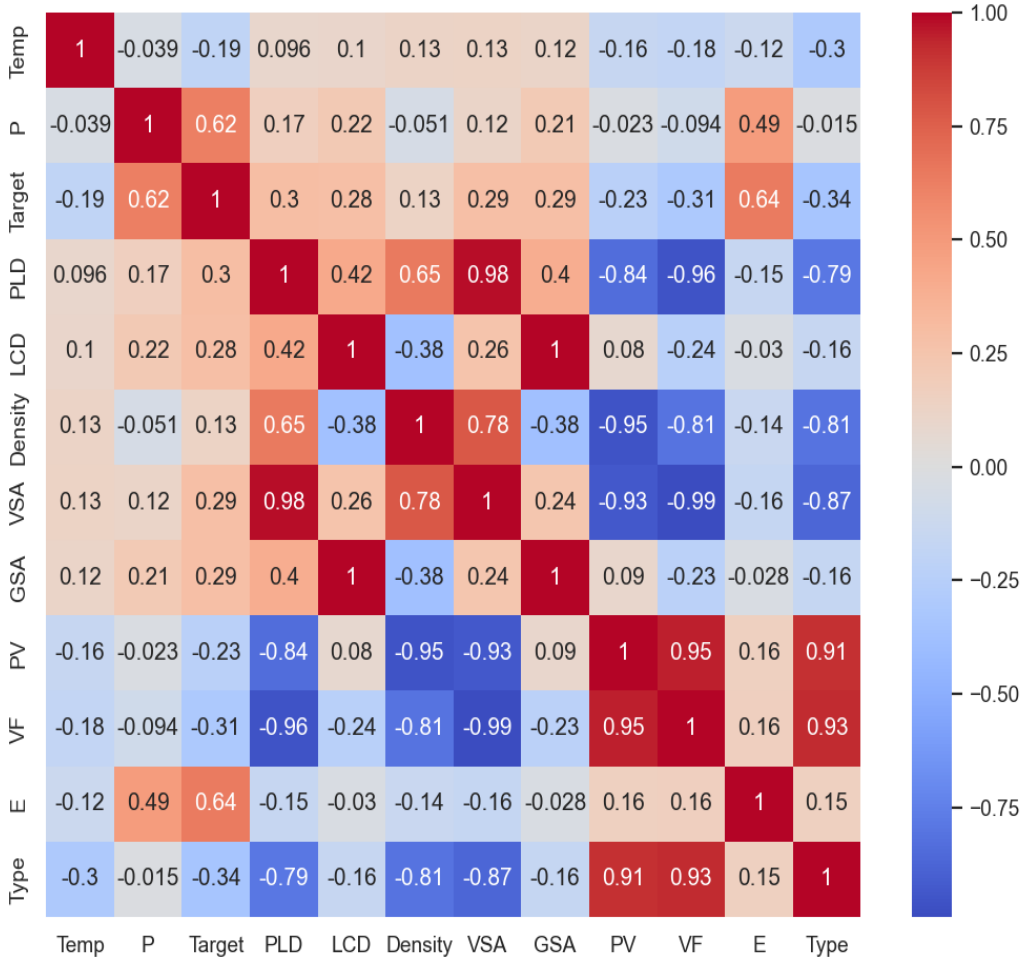


Fig.3-1 Pearson correlation coefficient

The correlation coefficient between equilibrium adsorption value and adsorption potential E is -0.64, confirming a strong negative correlation. Therefore, this may have a significant impact on adsorption predictions.

The absence of a strong correlation between adsorbate type and adsorption value may be attributed to the complexity of equilibrium adsorption, which is a function of temperature, pressure, and energy. The

relationship between these variables and equilibrium adsorption cannot be accurately captured through a simple correlation analysis. The lack of a strong correlation between parameters such as temperature and physical properties with equilibrium adsorption could be due to the non-linear nature of their relationship.

Given the non-linear nature and complexity of the relationship, feature selection was not performed. Identifying unnecessary feature quantities that could be excluded from the analysis was challenging, as the factors influencing equilibrium adsorption do not exhibit a linear correlation with each other.

3.3 Machine learning model

3.3.1 Decision Tree

The two ensemble algorithms used in this section both rely on tree models as base estimators. Decision Tree is a fundamental and commonly employed machine learning algorithm. As the name suggests, a decision tree makes decisions based on a tree-like structure. Typically, a decision tree starts from a root node and based on the problem and parameter settings, progressively splits into internal nodes and leaf nodes [40].

Decision Tree perform decision splits by assessing the purity of data samples. Information entropy is a commonly used metric to quantify the purity of a dataset. The information entropy $Ent(D)$ for a dataset D is defined as follows:

$$Ent(D) = -\sum_{k=1}^{m} p_k \log_2 p_k \qquad Eq.3.2$$

where m is the number of classes, and p_k is the proportion of samples belonging to class k. The decision tree recursively splits the dataset based on features to maximize the information gain or minimize the information entropy, resulting in a tree structure that represents decision-making rules.

This tree-based approach offers simplicity and interpretability, making decision trees suitable for various machine learning tasks. However, the potential for overfitting, especially with deep trees, has led to the development of ensemble methods, such as random forests and gradient boosting trees, which combine multiple decision trees to enhance performance.

The basic algorithm for decision tree learning is shown in Table 3.2.

Table 3-2 Basic algorithm for Decision Tree learning

Decision Tree code

```
class TreeNode:
    def __init__(self):
        self.attribute = None    # The attribute selected by this node
        self.branches = {}    # Branches, where key is the attribute value and value is a child node
        self.label = None    # Class label for leaf nodes
def generate_decision_tree(D, attributes):
    node = TreeNode()

    # 1. Check if samples all belong to the same class
    if all_same_class(D):
        node.label = D[0].class    # Mark the node as a leaf with the class label
        return node

    # 2. Check if samples have the same values on the attribute set
    if all_same_values(D, attributes):
        node.label = majority_class(D)    # Mark the node as a leaf with the majority class label
        return node

    # 3. Choose the best splitting attribute
    best_attribute = choose_best_attribute(D, attributes)
    node.attribute = best_attribute

    # 4. Generate branches for each attribute value
    for value in get_attribute_values(D, best_attribute):
        subset_D = get_subset(D, best_attribute, value)

        # 5. If the subset is empty, mark the node as a leaf
        if len(subset_D) == 0:
            child_node = TreeNode()
            child_node.label = majority_class(D)
            node.branches[value] = child_node
```

```
        else:
            # 6. Recursively generate child nodes
            child_node = generate_decision_tree(subset_D, attributes)
            node.branches[value] = child_node

    return node
```

3.3.1.1 GBDT model

GBDT (Gradient Boosting Decision Tree) algorithm employs CART (Classification and Regression Trees) regression decision trees and is widely utilized for regression, binary classification, and multiclass classification problems [41]. The fundamental concept of GBDT involves iteratively minimizing the discrepancy between actual and predicted values until this discrepancy falls below a threshold close to zero. The primary workflow combines multiple weak classifiers to yield the result of a strong classifier.

The steps for modeling GBDT algorithm are generally shown below:

Step 1: Build the first decision tree with the predicted value $f_0(x)$ initialized to 0.

Step 2: Use split gain to select the splitting attribute and splitting point, dividing the dataset into left and right child nodes. Compute the mean values for the datasets N_L and N_R in the left and right nodes, respectively, as predicted values y_i.

Step 3: Calculate the residuals as shown in Eq. 3.3.

$$r_{m,i} = f_{m-1}(x_i), \qquad i = 1,2,3 \dots n \qquad Eq.\,3.2$$

Step 4: Iterate until the error is below a specified threshold to compute the loss error.

Step 5: Construct a regression tree where residuals serve as the target values for the next iteration.

Step 6: Update the predicted result $y_{1\sim i}$

$$y_{1\sim i} = y_{1\sim i-1} + step * y_i \qquad Eq.\,3.3$$

where $y_{1\sim i}$ is the combined prediction result of the previous i iteration. y_i is the prediction result of the previous i iterations, step is the learning rate, typically taking values between 0 and 1.

These steps describe the process of iteratively building regression trees in a Gradient Boosting framework, updating predictions based on residuals and enhancing the model's performance over

iterations.

3.3.1.2 XGBoost model

The XGBoost (Extreme Gradient Boosting) algorithm is an enhanced version based on the GBDT (Gradient Boosted Decision Trees) algorithm, offering faster data training speed and more accurate prediction precision [42]. The XGBoost model structure is shown in Fig.3-2. In comparison to traditional GBDT algorithms, XGBoost algorithm employs a second-order Taylor expansion of the loss function, mitigating overfitting and providing higher scalability [43].

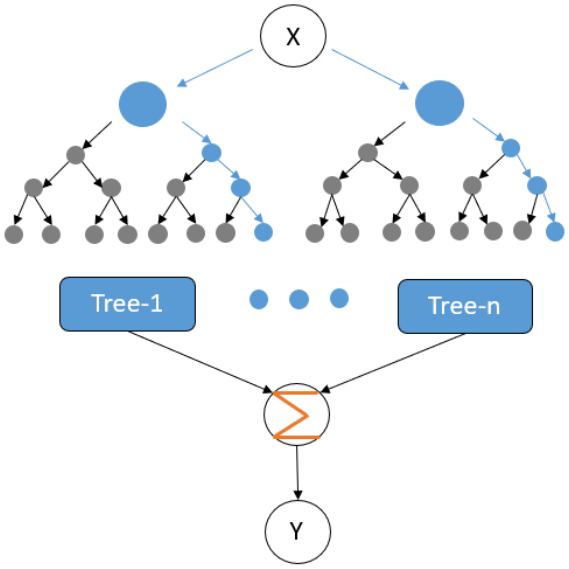


Fig.3-2 XGBoost model structure

While the relationship between trees in XGBoost remains sequential, nodes at the same level can be processed in parallel. The algorithm also enables new trees to automatically learn the splitting direction of missing values based on the residuals from the previous predictions. This capability makes XGBoost well-suited for handling tabular data with high-dimensional sparse features.

In summary, XGBoost incorporates regularization into the objective function, utilizes a second-order Taylor expansion for the loss function, and supports parallel processing of nodes at the same level. These features contribute to its faster training speed, improved accuracy, and enhanced ability to handle high-dimensional sparse feature tabular data compared to traditional gradient boosting decision tree (GBDT) algorithms.

3.3.2 Neural Networks

Neural Networks are a class of machine learning algorithms inspired by the structural organization of the human nervous system. They emulate the functioning principles of the human brain through the arrangement of multiple layers of neurons, also known as nodes. The inception of neural networks can be traced back to Frank Rosenblatt, who introduced a neural network structure known as the perceptron [44]. This perceptron laid the foundation for many subsequent neural network models. Neural networks are widely used for various tasks such as image recognition, speech recognition, natural language processing, recommender systems, etc.

3.3.2.1 ANN model

Artificial Neural Network (ANN) is a computational network inspired by the biological nervous system. Like the structure of the human brain, the neural network model consists of neurons organized in complex nonlinear forms [45].

The main component of the ANN model is shown in Fig.3-3:

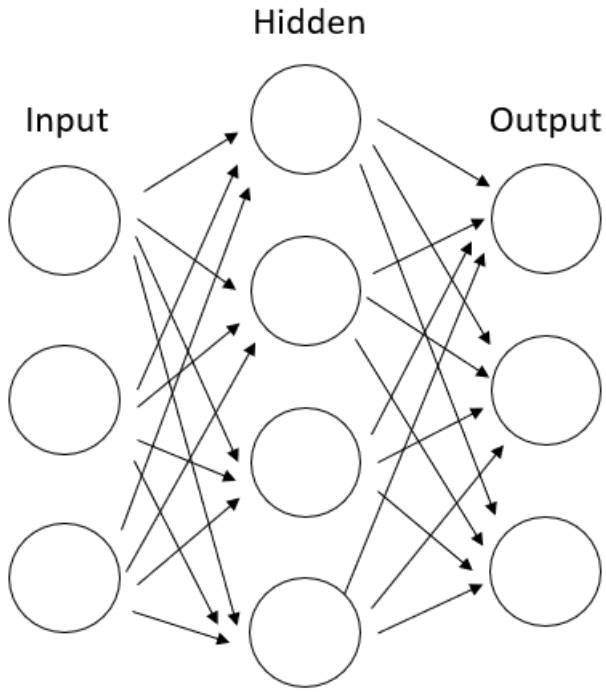


Fig.3-3 Components of the ANN model

Input layer: contains a neuron or node, which is used to receive input data.

Hidden layer: contains neurons or nodes that process the information received from the input layer.

Output layer: contains a neuron or node which is used to produce the final output.

Assuming that the number of layers of the neural network is k (k > 1) and the number of nodes (excluding bias nodes) in each layer from the input layer to the output layer is m respectively, this defines the dimension of the input vector to be m and the dimension of the output vector to be k. The output vectors of each layer of the network are respectively expressed as follows.

Input layer: $Y^{(0)} = [Y_1^{(0)}, Y_2^{(0)}, \dots Y_{m_0}^{(0)}]^T$

Hidden layer I: $Y^{(1)} = [Y_1^{(1)}, Y_2^{(1)}, \dots Y_{m_1}^{(1)}]^T$

Hidden layer II: $Y^{(2)} = [Y_1^{(2)}, Y_2^{(2)}, \dots Y_{m_2}^{(2)}]^T$

…

Output layer: $Y^{(K)} = [Y_1^{(K)}, Y_2^{(K)}, \dots Y_{m_K}^{(K)}]^T$

Next define the weight matrix and bias vector for each layer

$$W^{(1)} \in R^{m_1 \times m_0} \quad b^{(1)} \in R^{m_1 \times 1}$$

$$W^{(2)} \in R^{m_2 \times m_0} \quad b^{(2)} \in R^{m_2 \times 1}$$

…

$$W^{(K)} \in R^{m_K \times m_{K-1}} \quad b^{(K)} \in R^{m_K \times 1}$$

And define the activation function of each layer as $f^{(1)}, f^{(2)}, \dots, f^{(K)}$; choose the corresponding activation function according to different applications, and the type of activation function used in each layer may or may not be uniform. Next, by deriving the expression of the output vector of hidden layerI, the common expression for all layers except the input layer is derived for hidden layerI:

$$net_i^{(1)} = \sum_{j=1}^{m_0} W_{i,j}^{(1)} Y_j^{(0)} + b_i^{(1)}, (1 \le i \le m_i)$$

$$net^{(1)} = W^{(1)} Y^{(0)} + b^{(1)}$$

$$net^{(1)} = [net_1^{(1)}, net_2^{(1)}, \dots, net_{m_1}^{(1)}]^T$$

$$Y^{(1)} = f^{(1)}\left(net^{(1)}\right) = \left[Y_1^{(1)}, Y_2^{(1)}, \dots, Y_{m_1}^{(1)}\right]^T \qquad Eq.3.4$$

Each element in $net^{(1)}$ represents a weighted sum of the input layer vectors as well as the bias

vectors, which can also be referred to as the input vectors of the neurons in hidden layerI.

For layer k, $k \in \{1,2,\dots,K\}$

$$net_i^{(k)} = \sum_{j=1}^{m_0} W_{i,j}^{(k)} Y_j^{(k-1)} + b_i^{(k)}, (1 \le i \le m_k)$$

$$net^{(k)} = W^{(k)} Y^{(k-1)} + b^{(k)}$$

$$net^{(k)} = [net_1^{(k)}, net_2^{(k)}, \dots, net_{m_k}^{(k)}]^T$$

$$Y^{(k)} = f^{(k)}\left(net^{(k)}\right) = \left[Y_1^{(k)}, Y_2^{(k)}, \dots, Y_{m_k}^{(k)}\right]^T \qquad Eq.\,3.5$$

By forward layer-by-layer computation, the $net^{(k)}$ and $Y^{(k)}$ of each layer in the network, and the input and output vectors of each layer can be obtained, and thus the input and output values of each neuron are also obtained [46].

3.3.2.2 DNN model

The concept of deep neural network proposed by Hinton et al [47] in 2006 has triggered another wave of research and development on neural networks, where deep models have greater potential compared to shallow models.

Deep Neural Network (DNN) is a model evolved from the traditional ANN, which introduces a deep structure, i.e., it contains multiple hidden layers [48]. Compared to ANN, DNN is more complex in its hierarchical structure, in which each hidden layer contains multiple neurons, thus improving the expressive power of the model. By introducing a deep structure, DNN can learn more abstract and complex feature representations, which makes it perform better in handling complex tasks such as image recognition, speech recognition and natural language processing. However, DNN training typically requires more computational resources and data and is relatively more complex.

Due to its deep structure, DNN possess enhanced representational capabilities, making them well-suited for handling large-scale and high-dimensional datasets. They excel in capturing intricate relationships within data. However, the introduction of deep structures poses challenges, such as the issues of vanishing and exploding gradients, particularly when dealing with extensive datasets [49]. Gradient vanishing is a common problem when neural networks have many layers, especially in deep

neural networks. In the backpropagation process, the gradient is the derivative of the loss function with respect to the network parameters. When the network is deep, the gradient needs to be propagated back to the input layer through multiple layers, but during the propagation process, the gradient may become very small or even converge to zero. This results in the deeper layers of the network not being able to learn effective weight updates, which affects the training of the entire network. The opposite of gradient vanishing, gradient explosion is when the gradient becomes very large in backpropagation. This can cause the weight updates to become extreme and the weights of the network to increase rapidly, eventually leading to numerical instability or even overflow. Gradient explosion usually occurs when the network weights are not properly initialized, the learning rate is too large, or the network structure is not appropriate. Mitigating strategies, including appropriate weight initialization, the selection of non-saturating activation functions, batch normalization, and the incorporation of residual connections, have been proposed to address these concerns, enhancing the stability, and learning capacity of DNNs in the face of large-scale datasets.

3.4 Evaluation metrics and performance measures

To evaluate the performance of machine learning models, this paper uses four model evaluation metrics: Coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE) and training time.

3.4.1 Coefficient of determination

Coefficient of determination (R^2) is a metric used to assess the goodness of fit of a regression model. Its values range from 0 to 1, indicating the proportion of the variance in the dependent variable explained by the model. A higher R^2 value, closer to 1, suggests a better fit, while values closer to 0 indicate poorer model performance [50]. The formula for R^2 is shown in Eq. 3.6.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad Eq.\,3.6$$

3.4.2 Mean absolute error

Mean absolute error (MAE) is a commonly used metric in regression evaluation, representing the average absolute difference between predicted and actual values. A lower MAE indicates more accurate predictions. The formula for MAE is shown in Eq. 3.7.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad Eq.\,3.7$$

3.4.3 Mean squared error

Mean squared error (MSE) is another regression evaluation metric, measuring the average squared difference between predicted and actual values. Smaller MSE values imply smaller differences between predicted and actual values [51]. The formula for MSE is shown in Eq. 3.8.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad Eq.\,3.8$$

3.4.4 Training time

Training time is often used to evaluate the performance and efficiency of an algorithm, model training, or other task [52]. In the specific case of training, Time is output to show the time spent on the entire training process of the model. Training time helps to understand how fast the model is being trained and allows for performance tuning if needed.

3.5 Conclusions

In this chapter, key aspects of machine learning are explored in depth, with a focus on generating features and conducting Pearson correlation analyses on them. Follow this, there is a description of the machine learning models and evaluation metrics used in this paper.

Chapter 4: Model calculation and analysis

4.1 Model calculation

4.1.1 GBDT and XGBoost

The code for the construction of GBDT model and XGBoost model is shown in the following Table 4.1 and Table 4.2.

Table 4.1 Initialization GBDT model

Initialization Model
model = GradientBoostingRegressor(n_estimators=200, learning_rate=0.05, max_depth=10, random_state=42)

n_estimators=200: the number of iterations, i.e. the number of weak learners.

learning_rate=0.05: learning rate, which controls how much the weights are updated in each iteration.

max_depth=10: maximum depth of the decision tree.

random_state=42: sets the random seed to ensure reproducibility.

Table 4.2 Initialization XGBoost model

Initialization Model
for i in range(1, 101): model = xgb.XGBRegressor(max_depth=3, learning_rate=0.1, n_estimators=i, objective='reg:squarederror')

max_depth=3: the maximum depth of the decision tree.

learning_rate=0.1: learning rate, which controls how much the weights are updated in each iteration.

n_estimators=i: number of iterations, i.e. number of weak learners. Here range(1, 101) is used, which means the number of iterations is increased one by one from 1 to 100.

objective='reg:squarederror': the objective function for the regression task, using the mean square error (MSE) as the loss function.

Optimizing these parameters, one can fine-tune the gradient boosting model to enhance its fitting capabilities and optimize the performance of the regression model, obtaining the most suitable model.

4.1.2 ANN and DNN

The code for the construction of ANN model and DNN model is shown in the following Table 4.3 and Table 4.4.

Table 4.3 Initialization ANN model

Initialization Model

```
class ANN(nn.Module):
    def __init__(self):
        super(ANN, self).__init__()
        self.layers = nn.Sequential(
            nn.Linear(4, 8),
            nn.ReLU(),
            nn.Linear(8, 8),
            nn.ReLU(),
            nn.Linear(8, 1)

    def forward(self, x):
        outputs = self.layers(x)
        return outputs
model = ANN()

criterion = nn.MSELoss()
```

```
        optimizer = torch.optim.Adam(model.parameters(), lr=0.019,
weight_decay=0.0005)
```

Table 4.4 Initialization DNN model

Initialization Model

```
class DNN(nn.Module):
    def __init__(self):
        super(DNN, self).__init__()
        self.layers = nn.Sequential(
            nn.Linear(4, 8),
            nn.ReLU(),
            nn.Linear(8, 8),
            nn.ReLU(),
            nn.Linear(8, 1)

    def forward(self, x):
        outputs = self.layers(x)
        return outputs
model = DNN()

criterion = nn.MSELoss()
        optimizer = torch.optim.Adam(model.parameters(), lr=0.019,
weight_decay=0.0005)
```

nn.Linear(4, 8): A linear layer with an input dimension of 4 and an output dimension of 8.

nn.ReLU(): ReLU activation function, introducing non-linearity.

nn.Linear(8, 8): A linear layer with an input dimension of 8 and an output dimension of 8.

nn.ReLU(): ReLU activation function.

nn.Linear(8, 1): A linear layer with an input dimension of 8 and an output dimension of 1, suitable for regression tasks.

The model's loss function is Mean Squared Error (MSELoss), and the optimizer is Adam with a learning rate of 0.019 and weight decay of 0.0005.

4.1.3 Residual plot analysis

Residual plot is a visualization tool used to assess the performance of a regression model. In regression analysis, the residuals of a model are the differences between the actual observations and the predicted values of the model [53]. An example residual plot is shown in Fig.4-1. A residual plot visualizes these residual values on the horizontal axis, usually with the model's predicted values as the vertical axis. Residual plots show how well the model fits, and ideally, the residuals should be uniformly distributed around the zero line, with no apparent pattern. By looking at the residual plots, the analyst can better understand the model's performance, identify potential problems, and decide whether the model needs to be improved or adjusted.

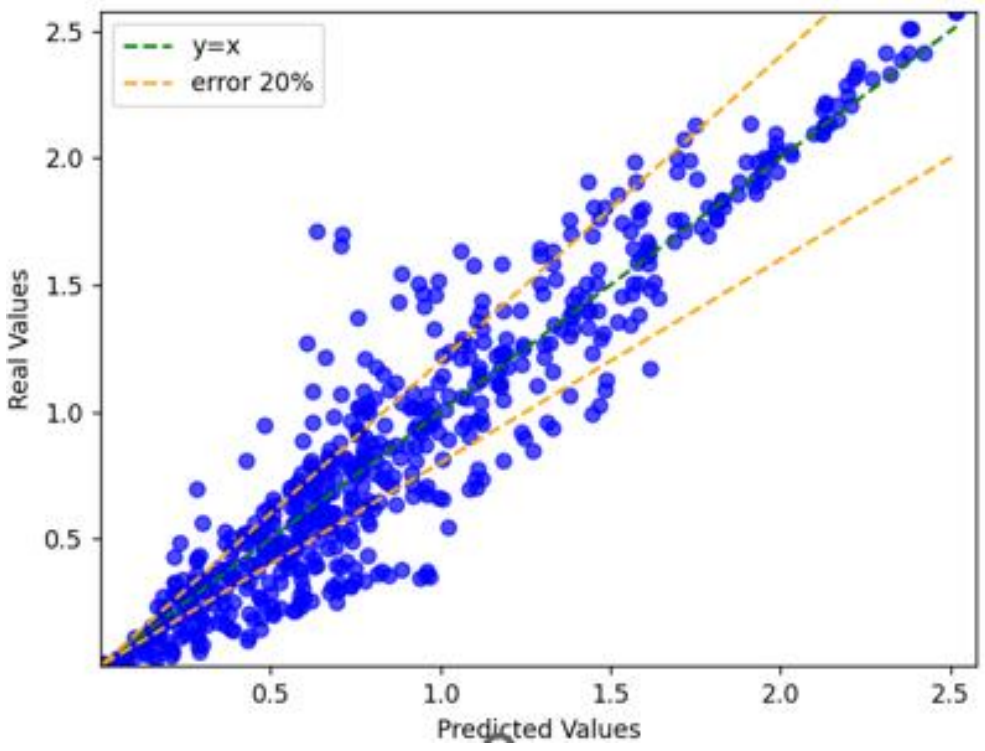


Fig.3-1 Example residual plot

4.2 Adsorption of zeolite at different temperatures

Using zeolite as an adsorption material in industrial production, predicting the adsorption capacity of zeolite at different temperatures can help you to optimize the process conditions and improve production efficiency. For adsorption heat pumps, the temperature of the evaporator decreases as the outside

temperature decreases in cold climates, which leads to a decrease in the saturated vapor pressure of the heat transfer liquid and a decrease in the adsorption capacity. Machine learning can be used to find out the adsorbent-worker pairs suitable for use in cold regions.

This section uses Zeolite 13X data on carbon dioxide adsorption at different temperatures for machine learning.

4.2.1 GBDT prediction model

The GBDT model parameters are set as shown in the following Table 4.5.

Table 4.5 GBDT model parameters

n_estimators	100
learning_rate	0.05
max_depth	3

The change of R2, MSE and MAE when GBDT model running is shown in Fig.4-2 and 4-3. The prediction result is shown in Fig.4-4.

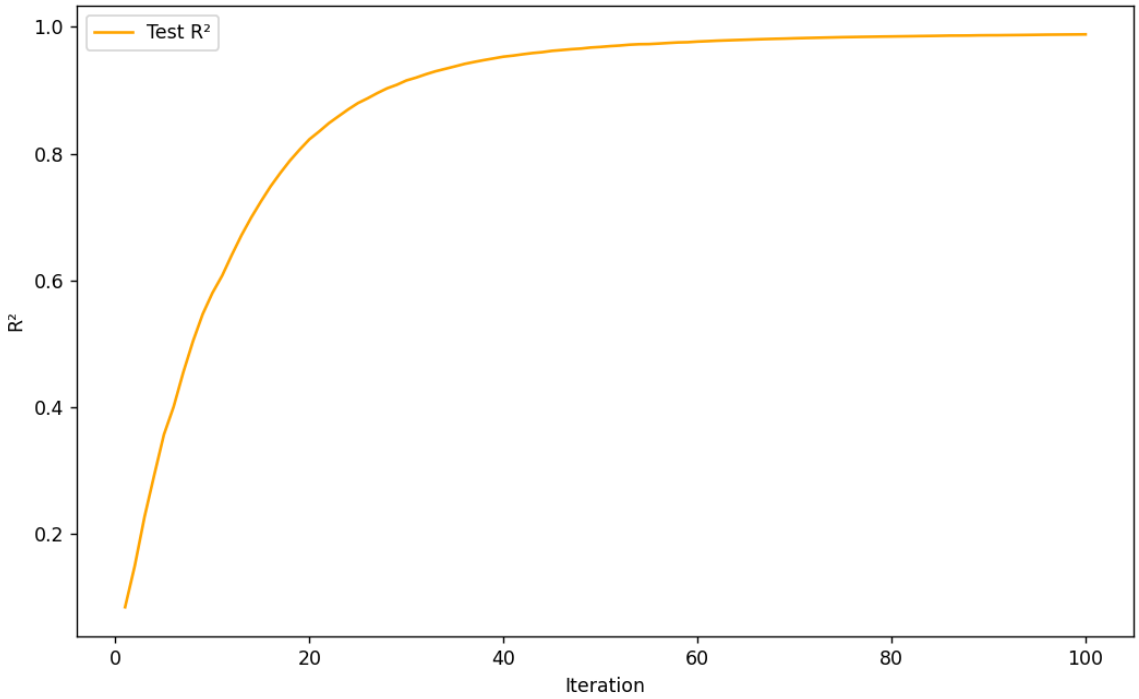


Fig.4-2 Test R^2

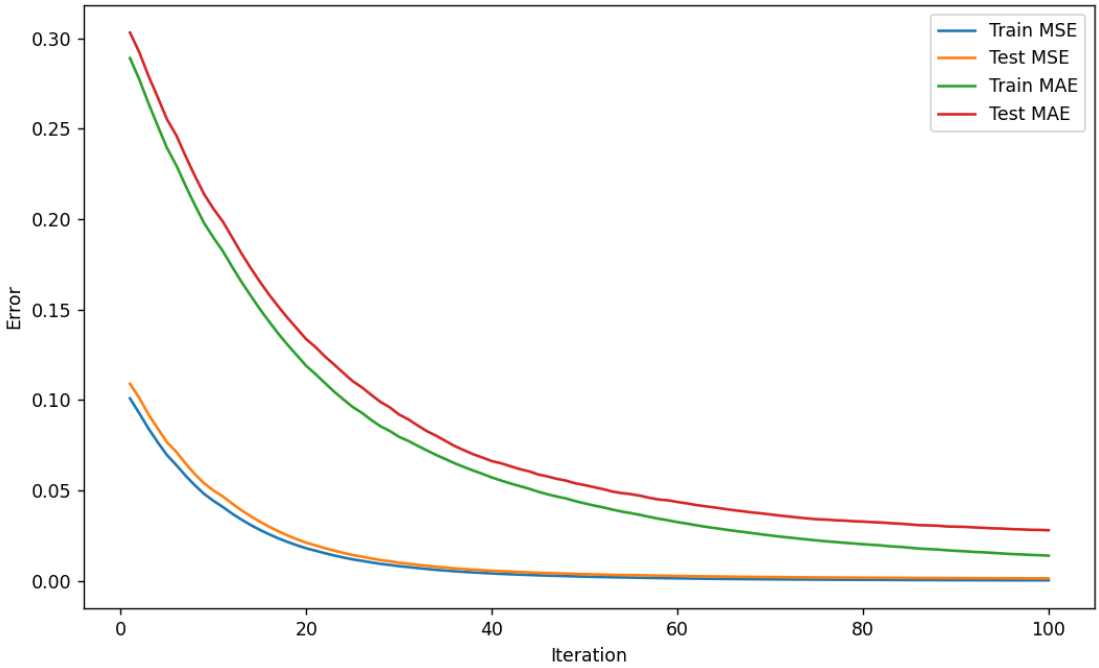


Fig.4-3 Train and Test MSE/MAE

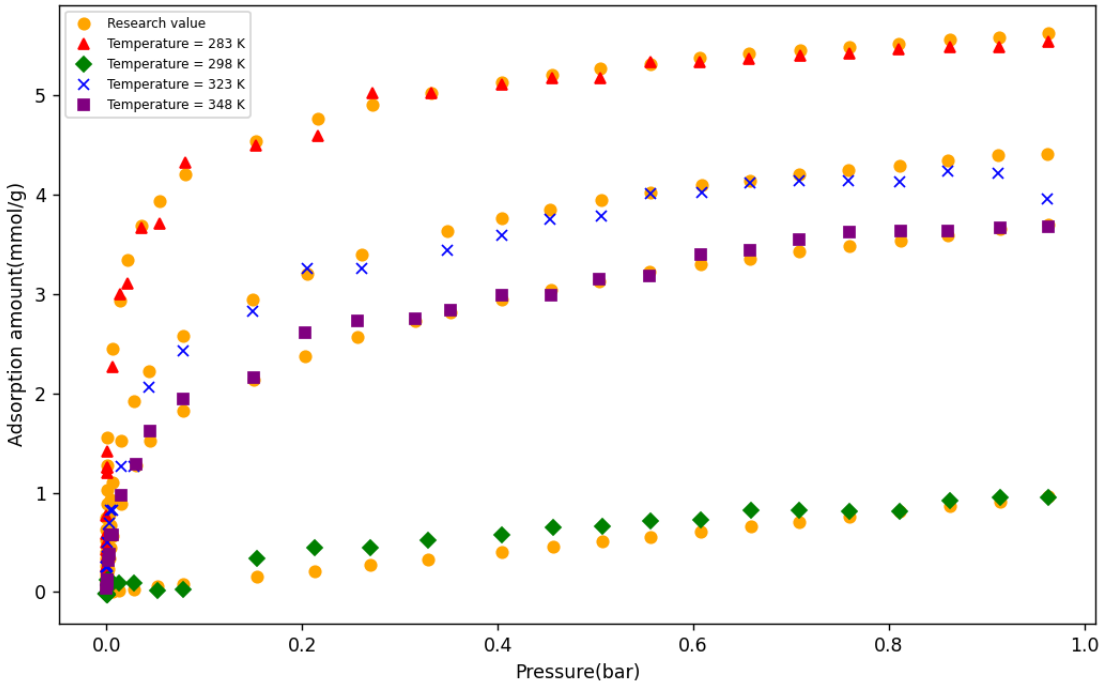


Fig.4-4 Equilibrium adsorption prediction

From error analysis, it can be observed that the model exhibits a satisfactory fitting performance. In the predicted adsorption isotherm plot, noticeable fluctuations are observed around 283K, while predictions at other temperatures demonstrate good accuracy.

4.2.2 XGBoost prediction model

The XGBoost model parameters are set as shown in the following Table 4.6.

Table 4.6 XGBoost model parameters

n_estimators	100
learning_rate	0.05
max_depth	3

The change of R2, MSE and MAE when XGBoost model running is shown in Fig.4-5 and 4-6. The prediction result is shown in Fig.4-7.

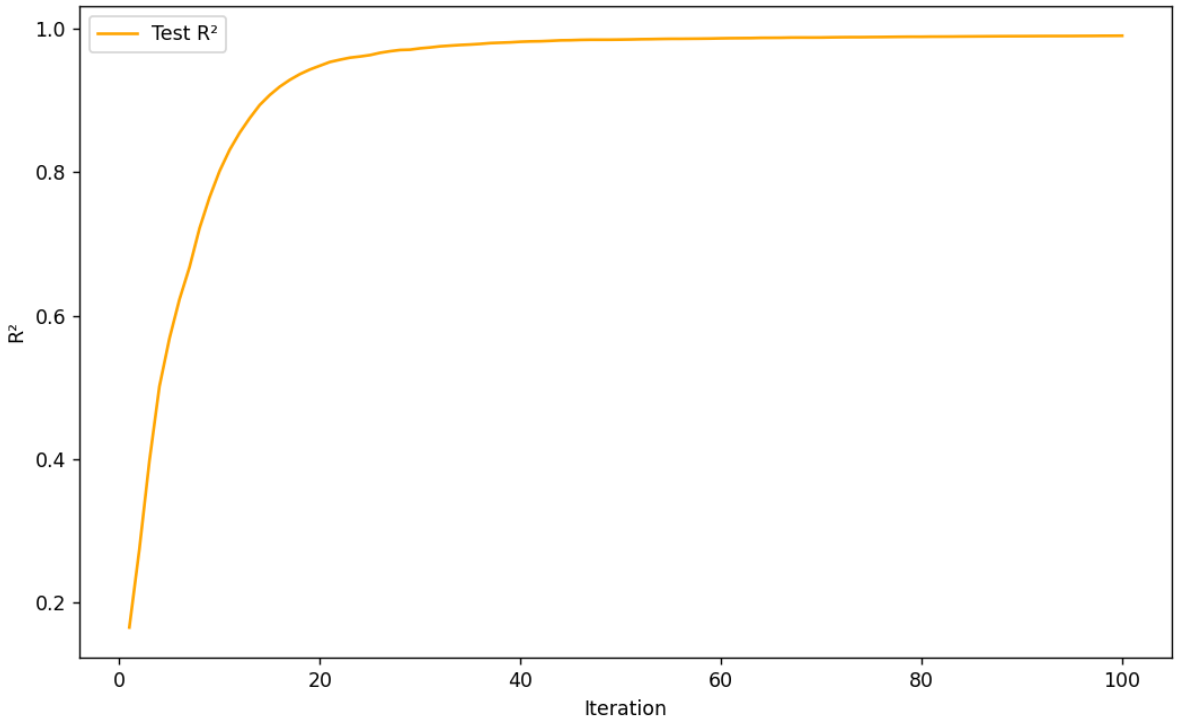


Fig.4-5 Test R^2

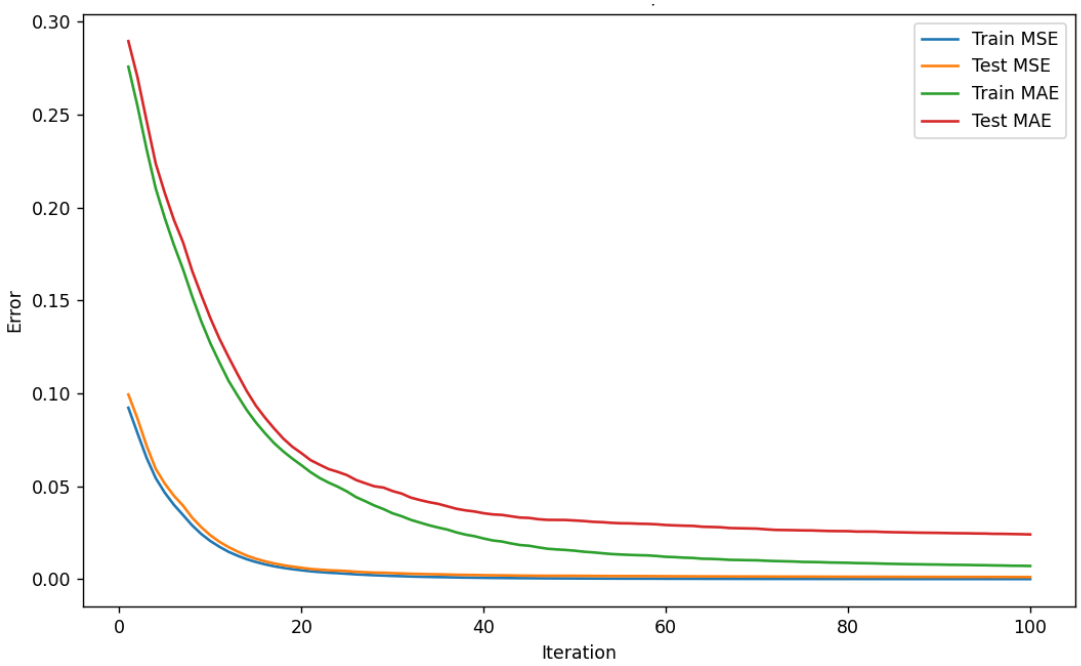


Fig.4-6 Train and Test MSE/MAE

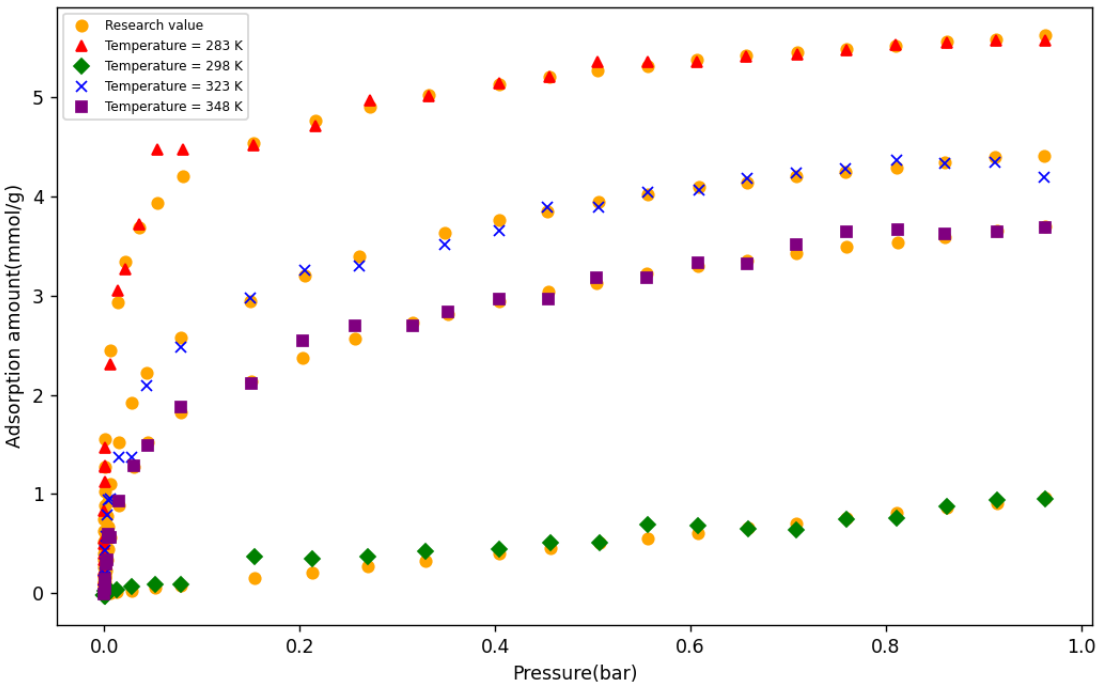


Fig.4-7 Equilibrium adsorption prediction

From the error analysis, it is evident that the model achieves a perfect fit. In comparison to the GBDT model, the XGBoost model demonstrates a quicker convergence to a fitting state. In the predicted adsorption isotherm plot, some individual prediction points show noticeable errors, while predictions for other adsorption points are accurate. This shows that the XGBoost algorithm is very effective in

capturing complex patterns and providing reliable predictions.

4.2.3 ANN prediction model

The ANN model parameters are set as shown in the following Table 4.7.

Table 4.7 ANN model parameters

Input layer	(11, 8)
Hidden layer	(8, 8)
Output layer	(8, 1)
lr	0.019
weight_decay	0.0005
epoch	200

The change of R2, MSE and MAE when ANN model running is shown in Fig.4-8 and 4-9. The prediction result is shown in Fig.4-10.

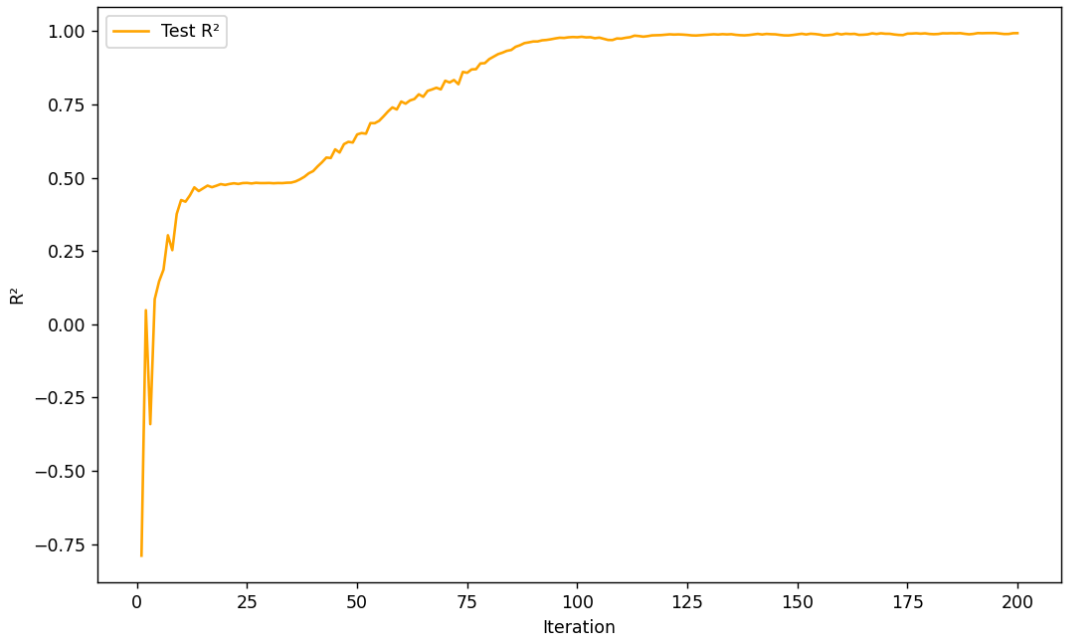


Fig.4-8 Test R^2

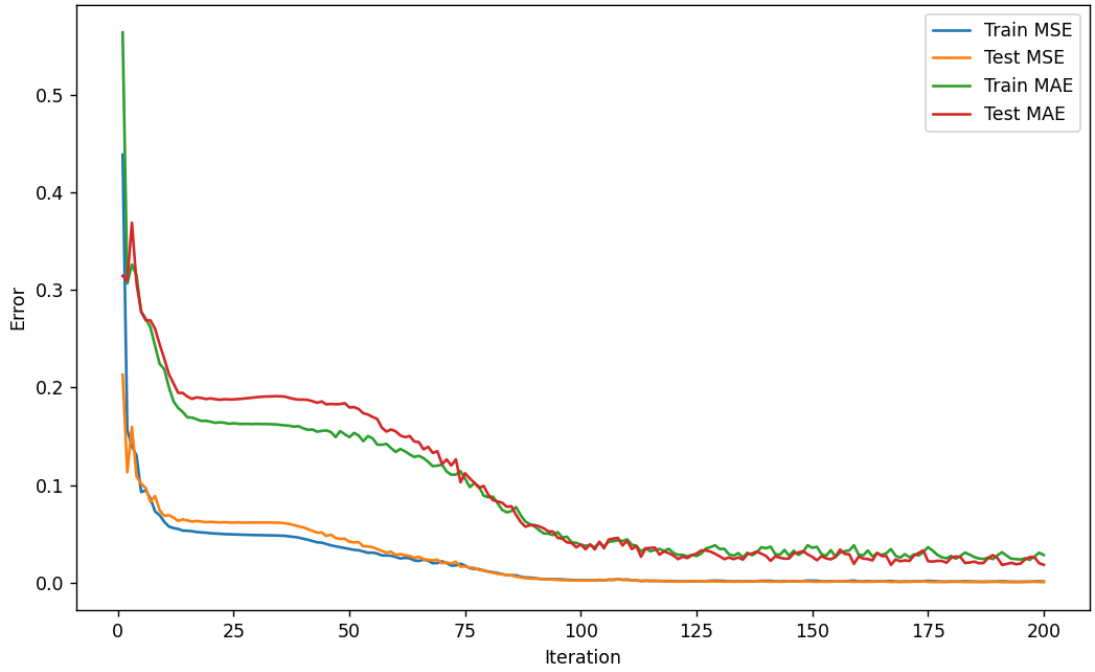


Fig.4-9 Train and Test MSE/MAE

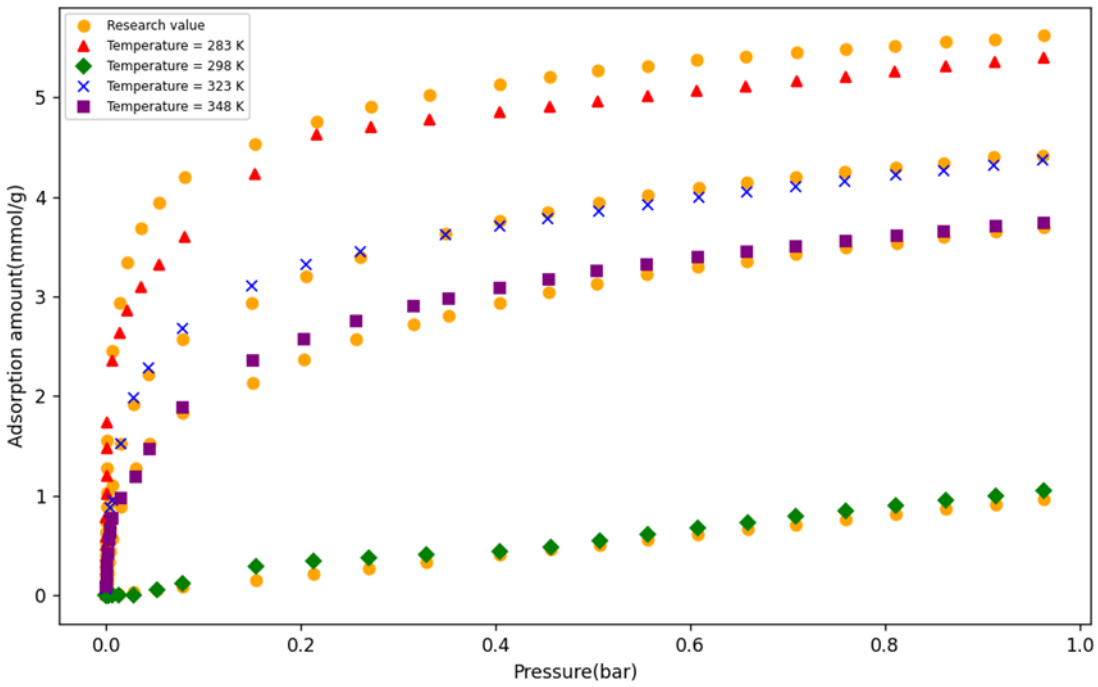


Fig.4-10 Equilibrium adsorption prediction

From the error analysis, the neural network model is more complex to fit and requires a greater number of trainings to reach the model fit. In the predicted adsorption isotherm plots, there is a bias, but it is smoother compared to the adsorption isotherm predicted by decision tree.

4.2.4 DNN prediction model

The DNN model parameters are set as shown in the following Table 4.8.

Table 4.8 DNN model parameters

Input layer	(11, 8)
Hidden layer	(8, 8)
Output layer	(8, 1)
lr	0.016
weight_decay	0.0005
epoch	200

The change of R2, MSE and MAE when ANN model running is shown in Fig.4-11 and 4-12. The prediction result is shown in Fig.4-13.

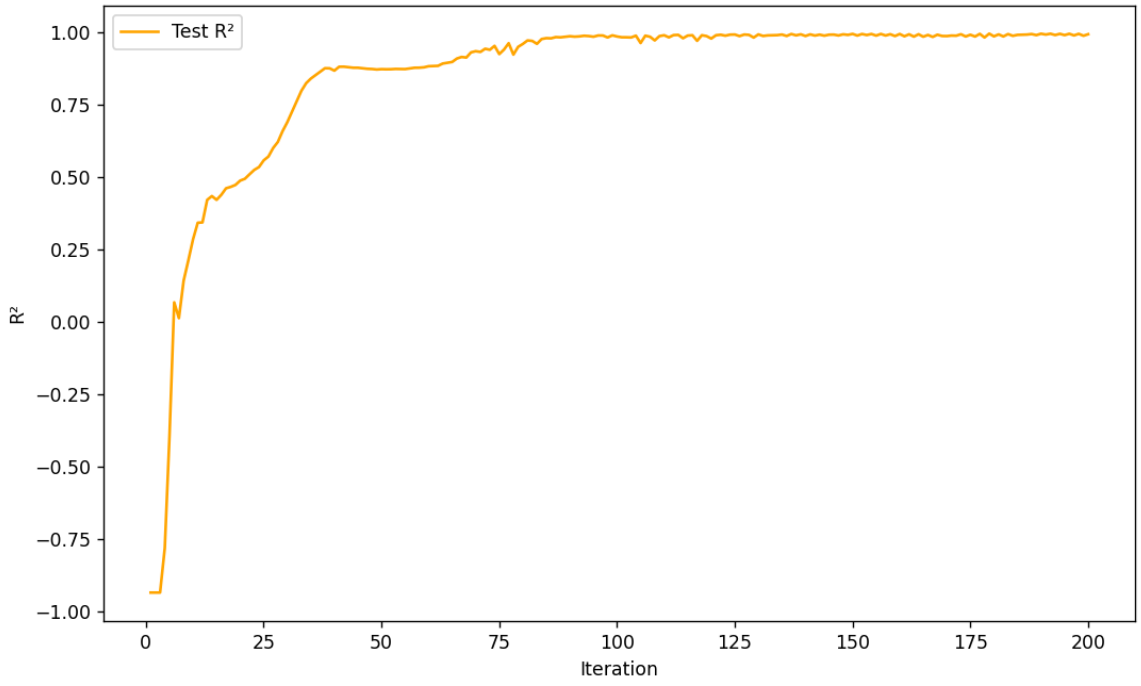


Fig.4-11 Test R^2

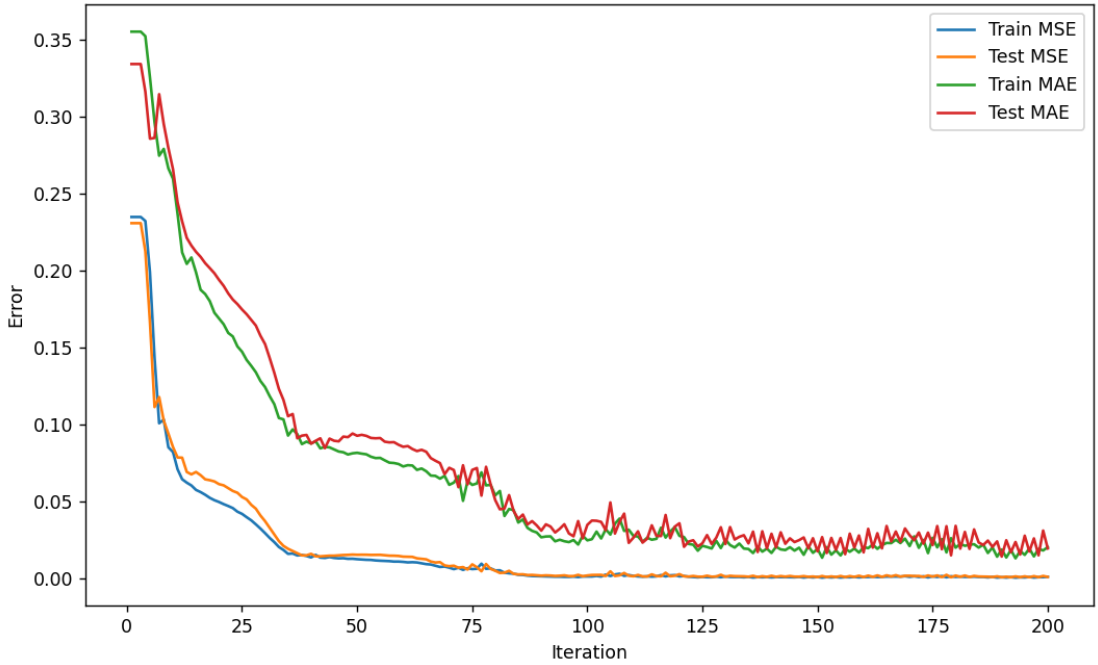


Fig.4-12 Train and Test MSE/MAE

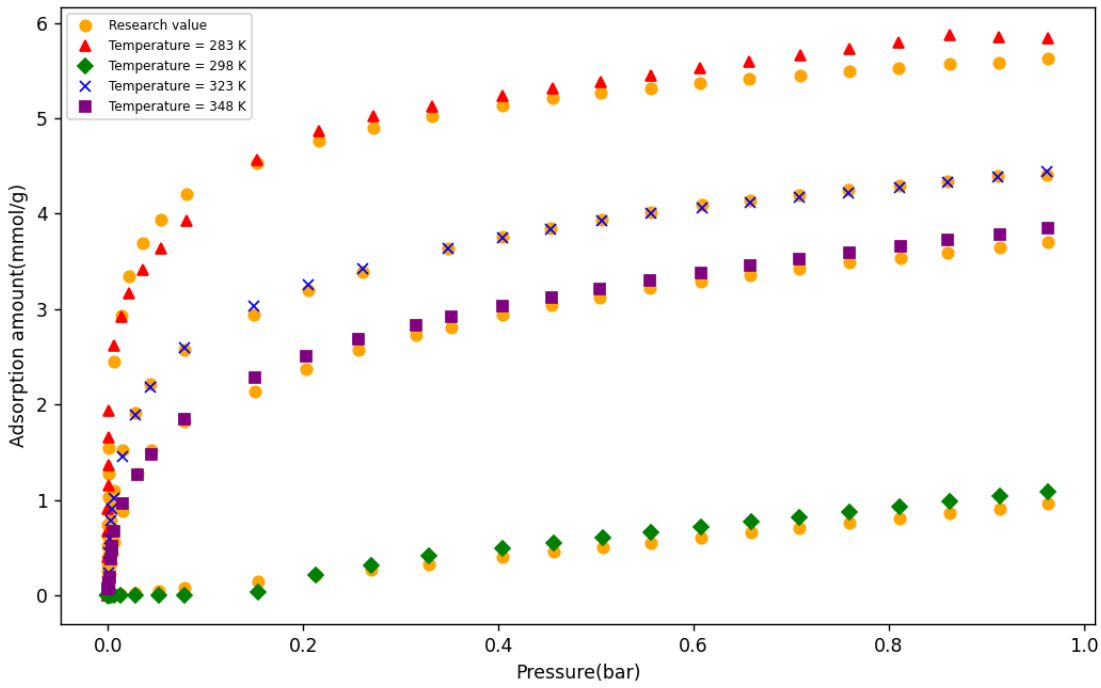


Fig.4-13 Equilibrium adsorption prediction

From the error analysis, it is evident that the fitting complexity of the DNN model surpasses that of the ANN model. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) exhibit frequent oscillations during training. Examining the predicted adsorption isotherm plot, the DNN model produces the most accurate simulation results among the four models, aligning closely with the actual values.

4.2.5 Results analysis

The evaluation metrics for each model are shown in the Table 4.9.

Table 4.9 Model evaluation metrics

	MSE	MAE	R^2	TIME
GBDT	0.00141	0.02801	0.98815	2.21654s
XGBoost	0.00119	0.02411	0.99004	3.38953s
ANN	0.00167	0.03285	0.98595	2.08822s
DNN	0.00072	0.01963	0.99395	2.16752s

From the table above, it can be observed that the DNN model exhibits the smallest prediction error. In situations where precision is of utmost importance, the DNN model is the most suitable choice. The training time for the ANN model is slightly shorter than that of the DNN model, attributed to the simpler structure of the ANN model. This holds true for the XGBoost model and the GBDT model as well. While the XGBoost model provides more accurate predictions, its complex structure necessitates a longer training time. However, due to the excellent predictive framework of the XGBoost model, fewer training iterations can be set compared to the GBDT model, reducing overall training time.

Fig.4-14 illustrates the residual plots for the four models. From the presentation of the residual plots, it is evident that all four models have effectively controlled most prediction errors within 15%, meeting the requirements for practical predictions. Regarding the types of errors, the errors in XGBoost and GBDT models fall under random error, whereas the errors in ANN and DNN models are categorized as systematic errors.

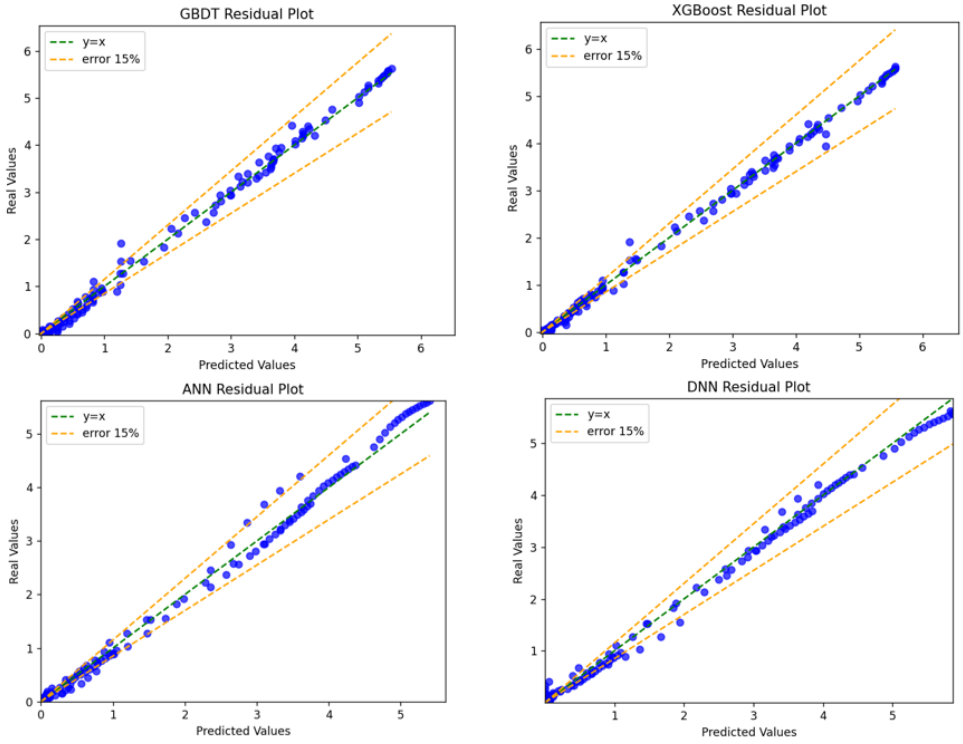


Fig.4-14 Residual plot for four models

4.3 Adsorption of different zeolites at same temperature

In the environmental field, predicting the adsorption properties of different zeolites at the same temperature can be useful for assessing their environmental impact under similar conditions, e.g., in water treatment or air purification.

This section uses five types of zeolite data on N-Butane adsorption at same temperature for machine learning.

4.3.1 GBDT prediction model

The GBDT model parameters are set as shown in the following Table 4.10.

Table 4.10 GBDT model parameters

n_estimators	100
learning_rate	0.08
max_depth	5

The change of R2, MSE and MAE when GBDT model running is shown in Fig.4-15 and 4-16. The prediction result is shown in Fig.4-17.

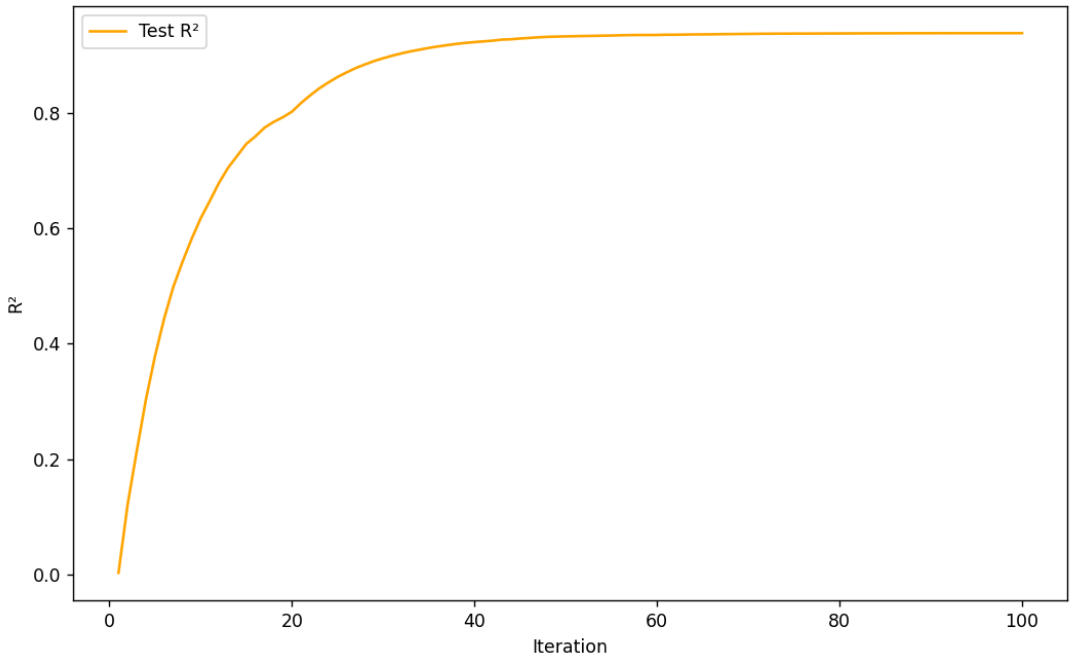


Fig.4-15 Test R^2

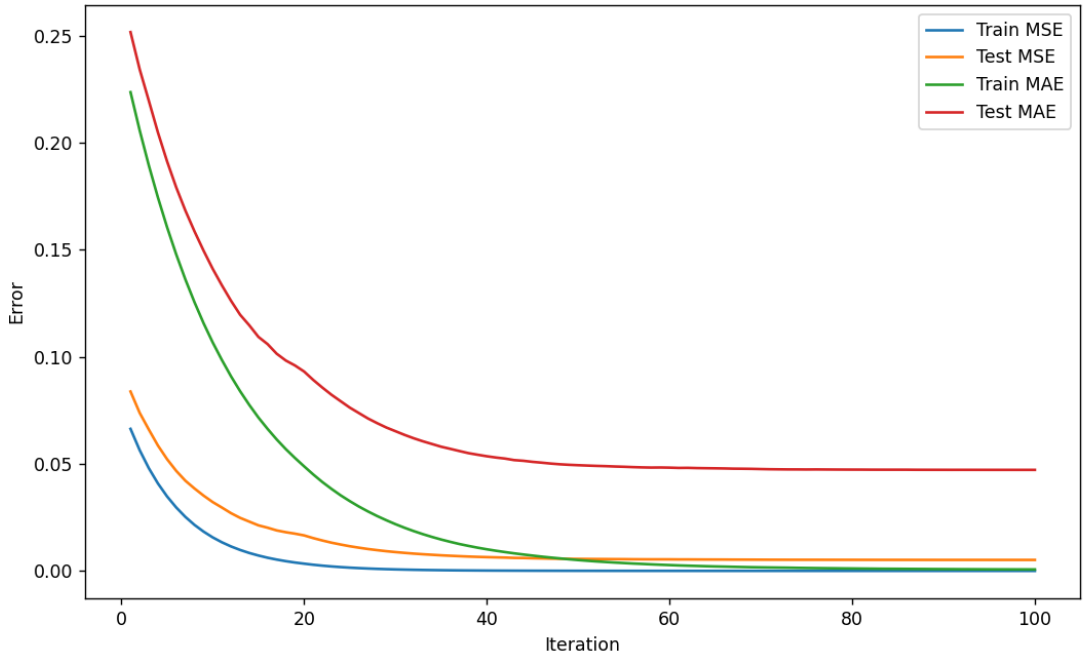


Fig.4-16 Train and Test MSE/MAE

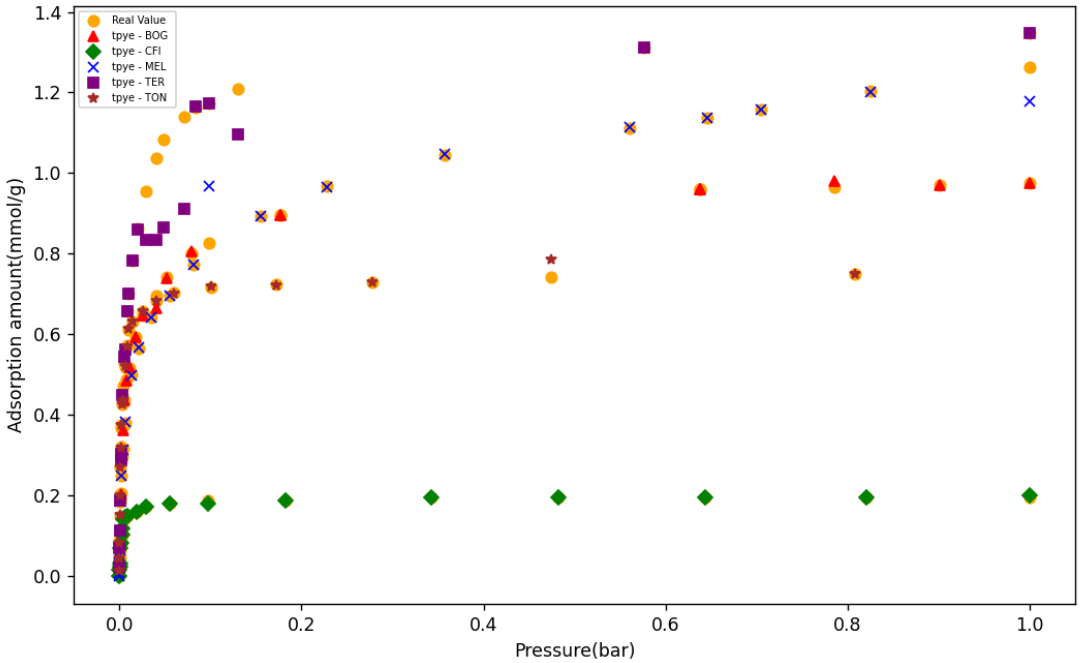


Fig.4-17 Equilibrium adsorption prediction

From error analysis, it can be observed that the model exhibits a satisfactory fitting performance. However, from the adsorption prediction figure, the adsorption isotherm prediction of Zeolite TER is not good enough.

4.3.2 XGBoost prediction model

The XGBoost model parameters are set as shown in the following Table 4.11.

Table 4.11 XGBoost model parameters

n_estimators	100
learning_rate	0.1
max_depth	3

The change of R2, MSE and MAE when XGBoost model running is shown in Fig.4-18 and 4-19. The prediction result is shown in Fig.4-20.

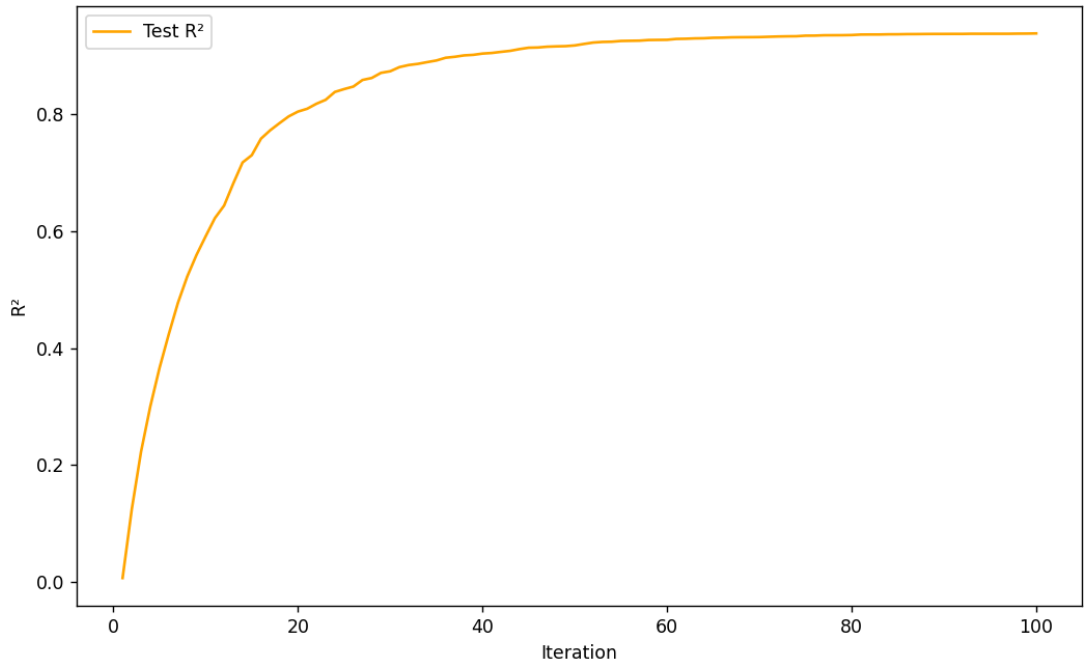


Fig. 4-18 Test R^2

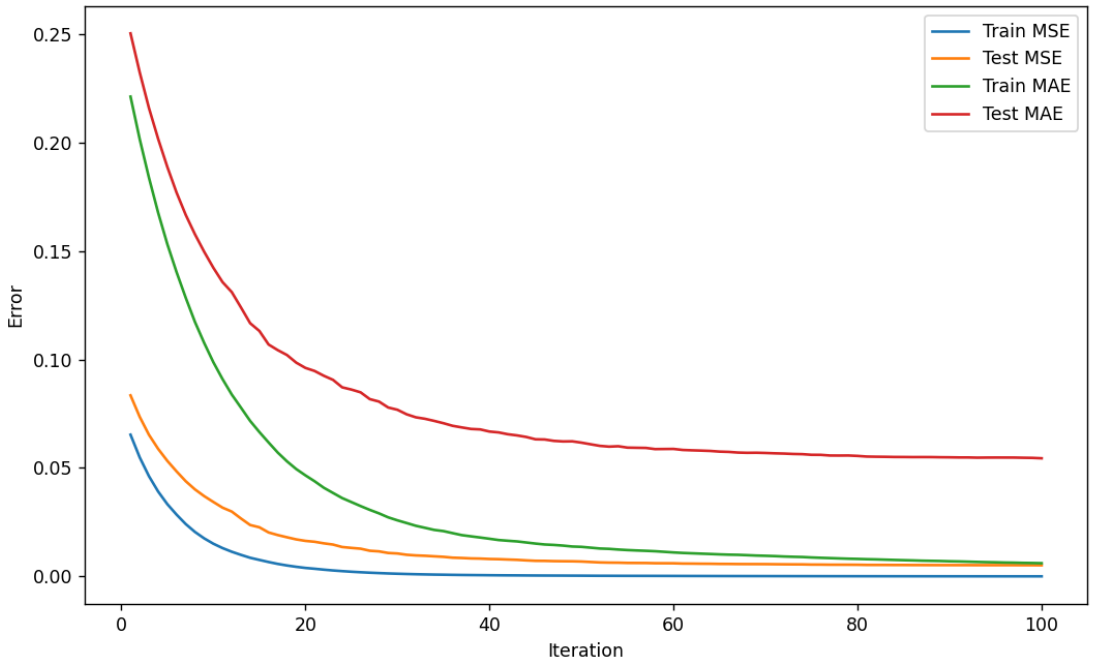


Fig. 4-19 Train and Test MSE/MAE

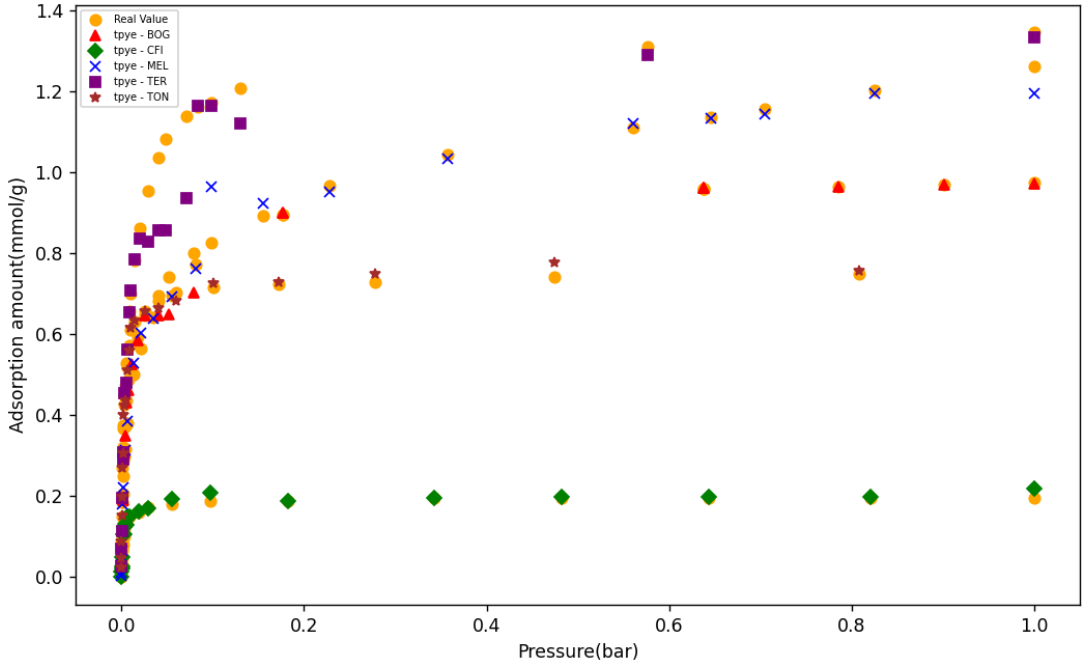


Fig.4-20 Equilibrium adsorption prediction

The prediction of xgboost is like the gbdt model. Predictions for Zeolite TER are less favorable. It can be assumed that in the Decision Tree prediction model machine learning, the model will predict adsorption of zeolites with higher adsorption capacity weaker than other zeolites.

4.3.3 ANN prediction model

The ANN model parameters are set as shown in the following Table 4.12.

Table 4.12 ANN model parameters

Input layer	(11, 8)
Hidden layer	(8, 8)
Output layer	(8, 1)
lr	0.019
weight_decay	0.0005
epoch	200

The change of R2, MSE and MAE when ANN model running is shown in Fig.4-21 and 4-22. The prediction result is shown in Fig. 4-23.

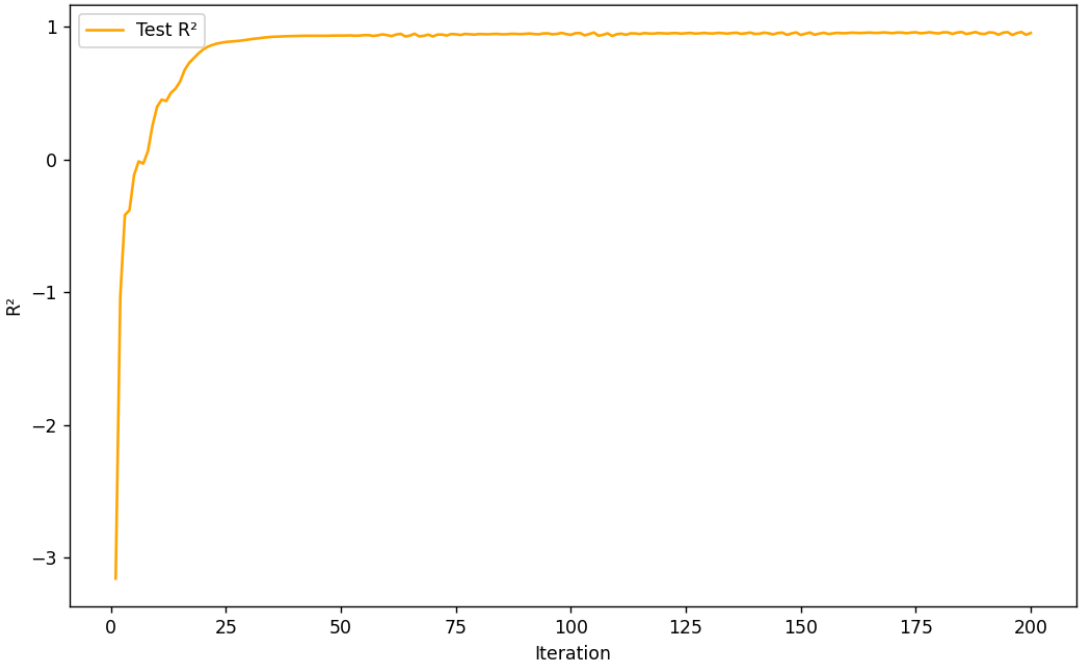


Fig.4-21 Test R^2

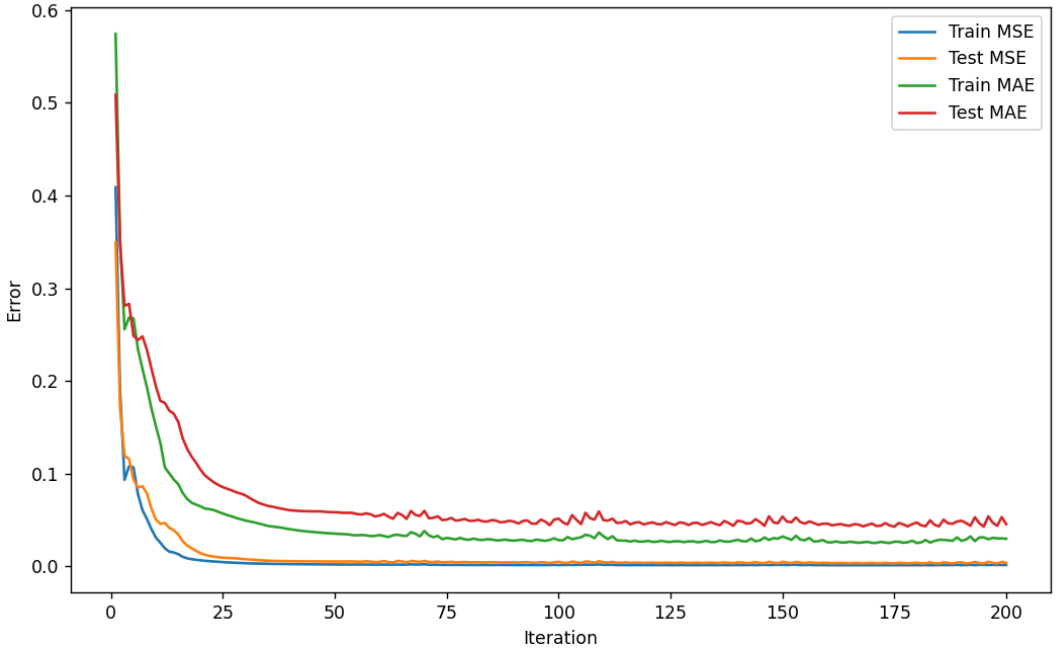


Fig.4-22 Train and Test MSE/MAE

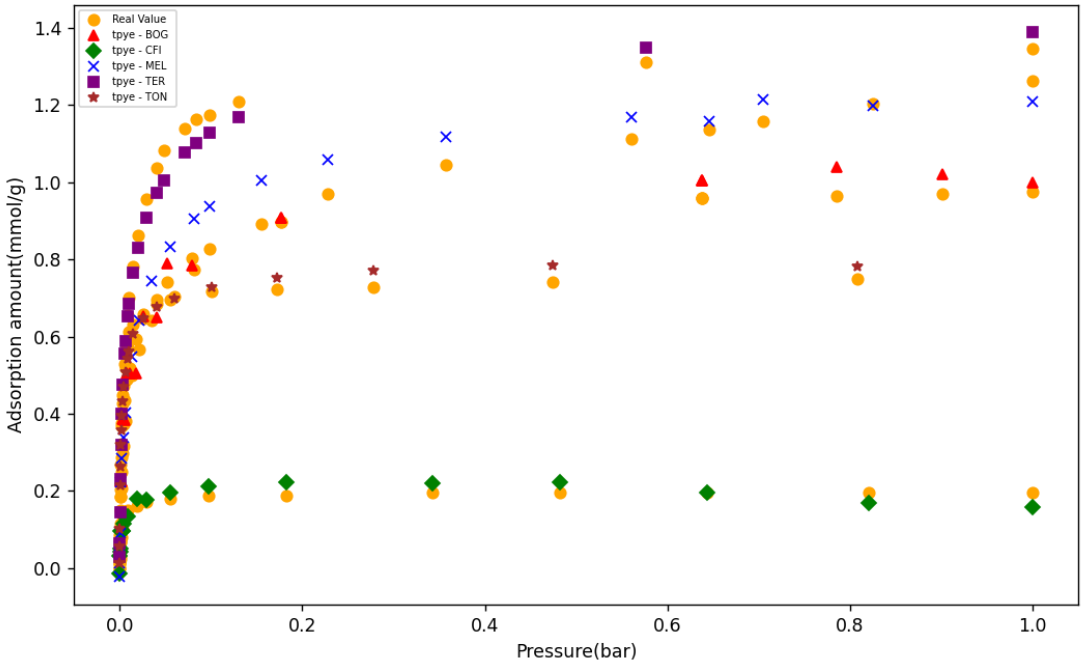


Fig.4-23 Equilibrium adsorption prediction

From the error analysis, it can be observed that the ANN model exhibits oscillations even after fitting, requiring an appropriate number of training iterations to achieve convergence. In the predicted adsorption isotherm plot, systematic biases are evident in the ANN predictions, but these biases fall within an acceptable range.

4.3.4 DNN prediction model

The DNN model parameters are set as shown in the following Table 4.13.

Table 4.13 DNN model parameters

Input layer	(11, 8)
Hidden layer	(8, 8)
Output layer	(8, 1)
lr	0.01
weight_decay	0.0005
epoch	200

The change of R2, MSE and MAE when ANN model running is shown in Fig.4-24 and 4-25. The prediction result is shown in Fig. 4-26.

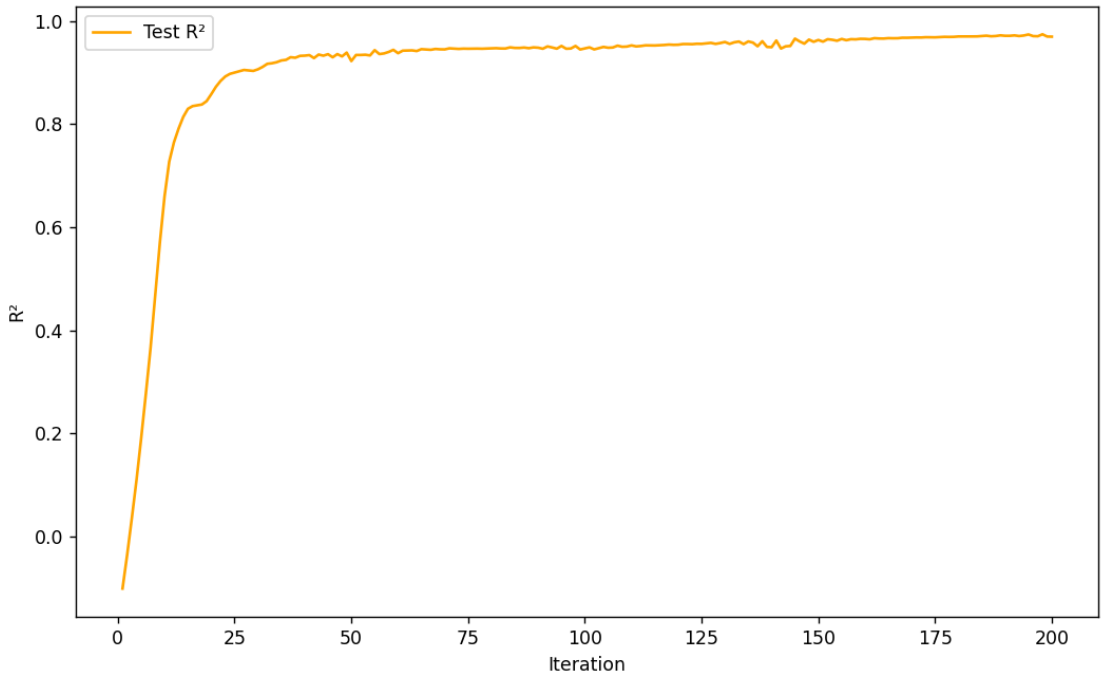


Fig.4-24 Test R^2

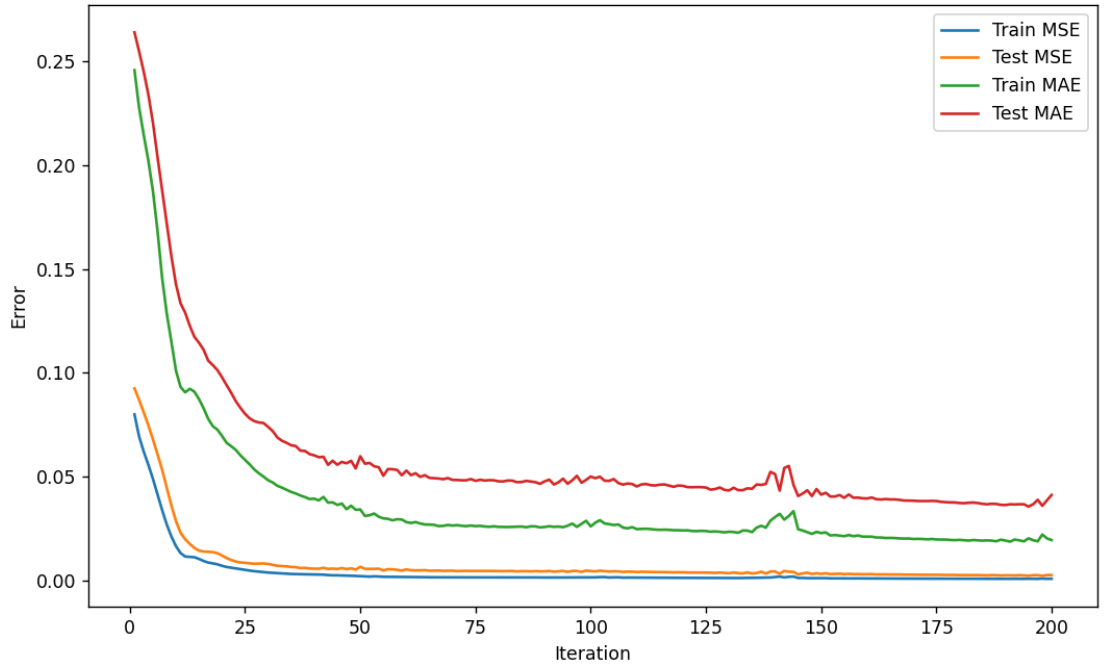


Fig.4-25 Train and Test MSE/MAE

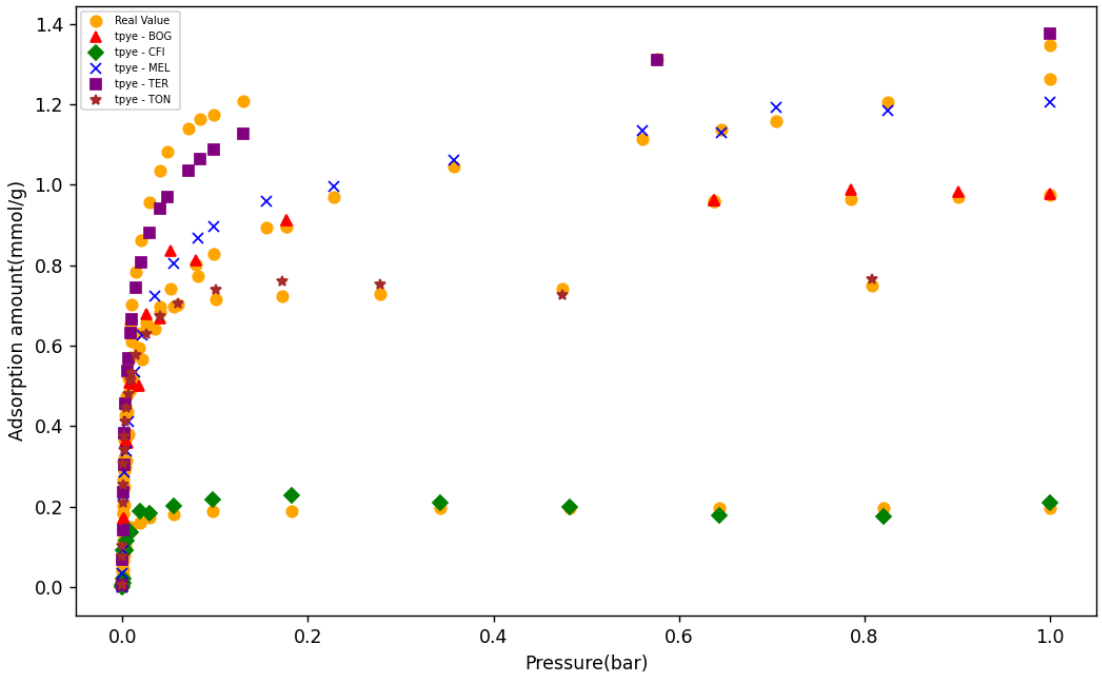


Fig.4-26 Equilibrium adsorption prediction

From the error analysis, the DNN model fit for different zeolite adsorption is better than the prediction of adsorption of the same zeolite at different temperatures. From the prediction of adsorption isotherms, the DNN model is significantly better than the ANN model.

4.3.5 Results analysis

The evaluation metrics for each model are shown in the Table 4.14.

Table 4.14 Model evaluation metrics

	MSE	MAE	R^2	TIME
GBDT	0.00513	0.04719	0.93895	2.35291s
XGBoost	0.00517	0.05445	0.93852	3.32423s
ANN	0.00211	0.03626	0.97480	1.75425s
DNN	0.00210	0.03597	0.97492	1.76046s

From the table above, the performance of Neural Networks model matches the analysis in 4.2.5. the

DNN model has the smallest prediction error. the training time of the ANN model is slightly shorter than that of the DNN model. However, in this situation, the structurally simple gbdt model outperforms xgboost, both in terms of error control and training time.

Fig.4-27 illustrates the residual plots for the four models. From the residual plots, the predictions of the XGBoost and GBDT models are in better agreement with the real values. However, the random error is larger and may exceed 15% deviation. This is the reason why the MAE, MSE, and R^2 performance of XGBoost and GBDT models is lower than that of Neural Networks. On the other hand, the errors of ANN and DNN models belong to the systematic errors, which can basically control the deviation within 15%.

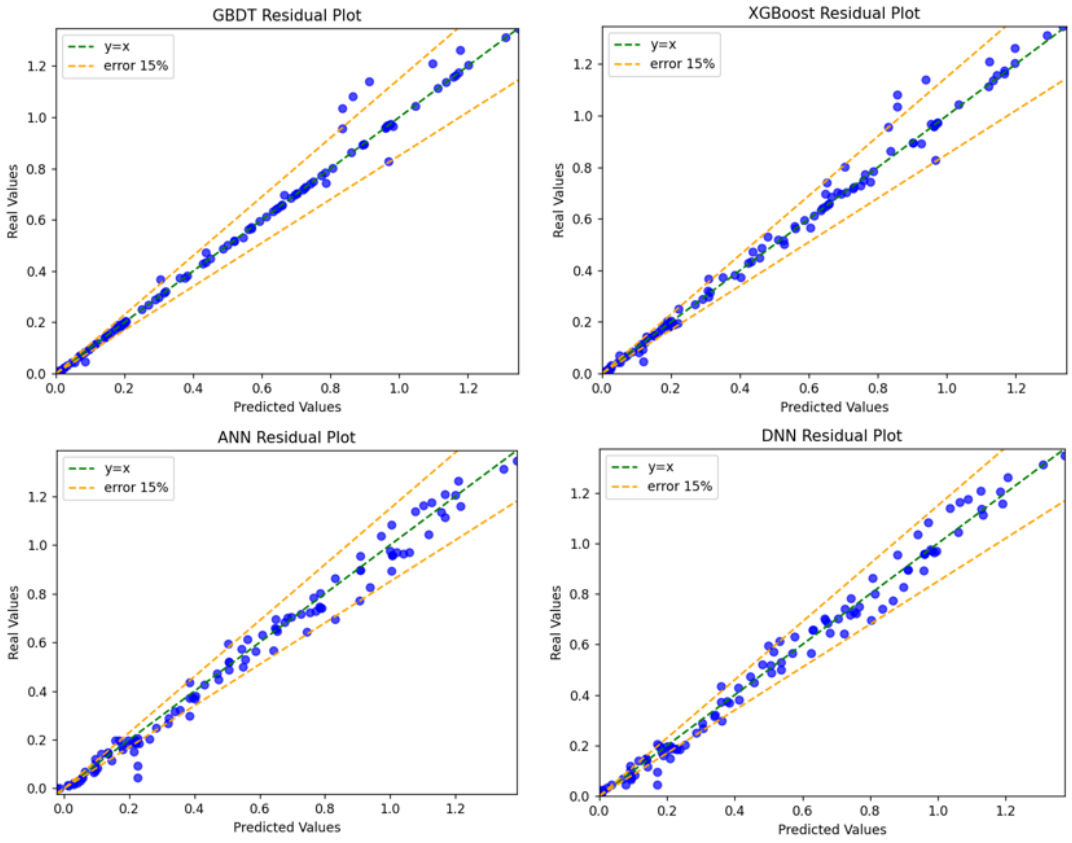


Fig.4-27 Residual plot for four models

4.4 Big zeolite data processing

To find out the most suitable algorithm for zeolite adsorption prediction in the case of big data, this section uses all the collected zeolite data for machine learning.

4.4.1 GBDT prediction model

The GBDT model parameters are set as shown in the following Table 4.14.

Table 4.14 GBDT model parameters

n_estimators	150
learning_rate	0.05
max_depth	10

The change of R2, MSE and MAE when GBDT model running is shown in Fig.4-28 and 4-29. The prediction result is shown in Fig.4-30.

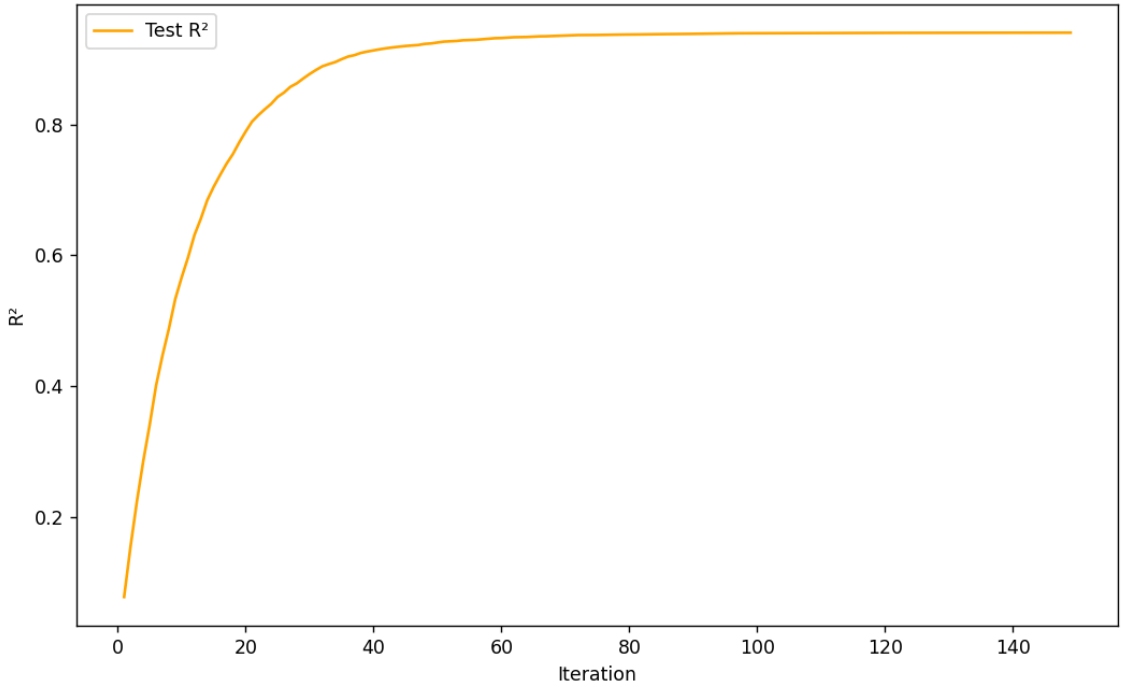


Fig.4-28 Test R^2

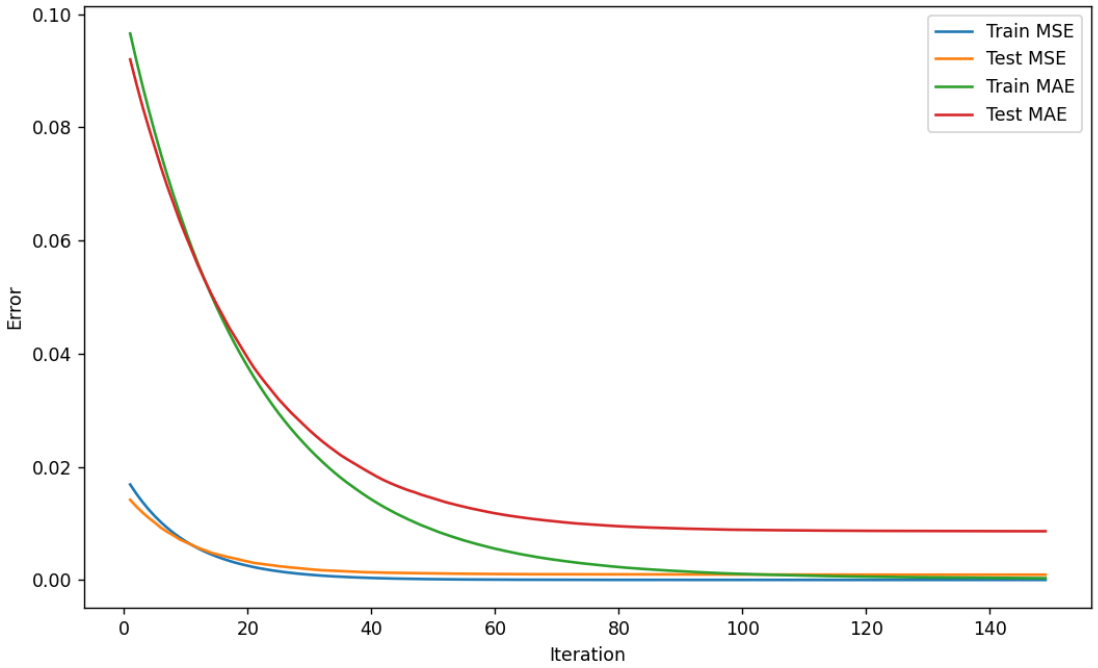


Fig.4-29 Train and Test MSE/MAE

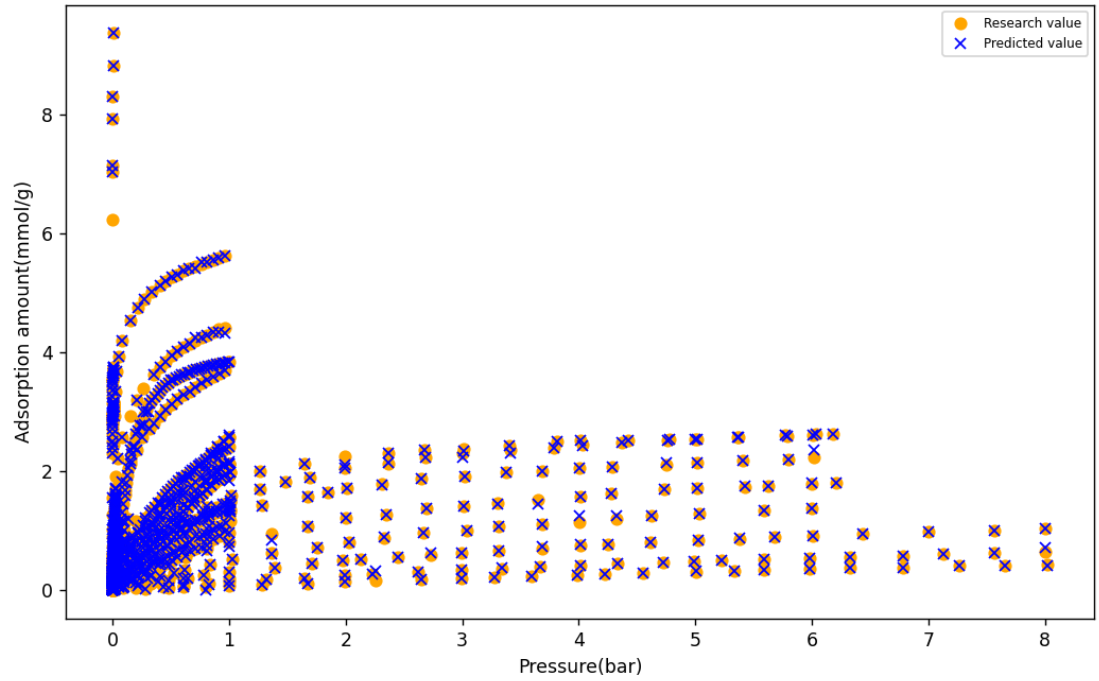


Fig.4-30 Equilibrium adsorption prediction

From error analysis, the number of training sessions was increased to 150 because of the increased amount of data. And the model was fitted well. Because of the large amount of data, it is difficult to see the prediction directly from the isotherm predictions. This study will continue to analyze the prediction effect in the residual plot analysis later.

4.4.2 XGBoost prediction model

The XGBoost model parameters are set as shown in the following Table 4.15.

Table 4.15 XGBoost model parameters

n_estimators	150
learning_rate	0.05
max_depth	10

The change of R2, MSE and MAE when XGBoost model running is shown in Fig.4-31 and 4-32. The prediction result is shown in Fig.4-33.

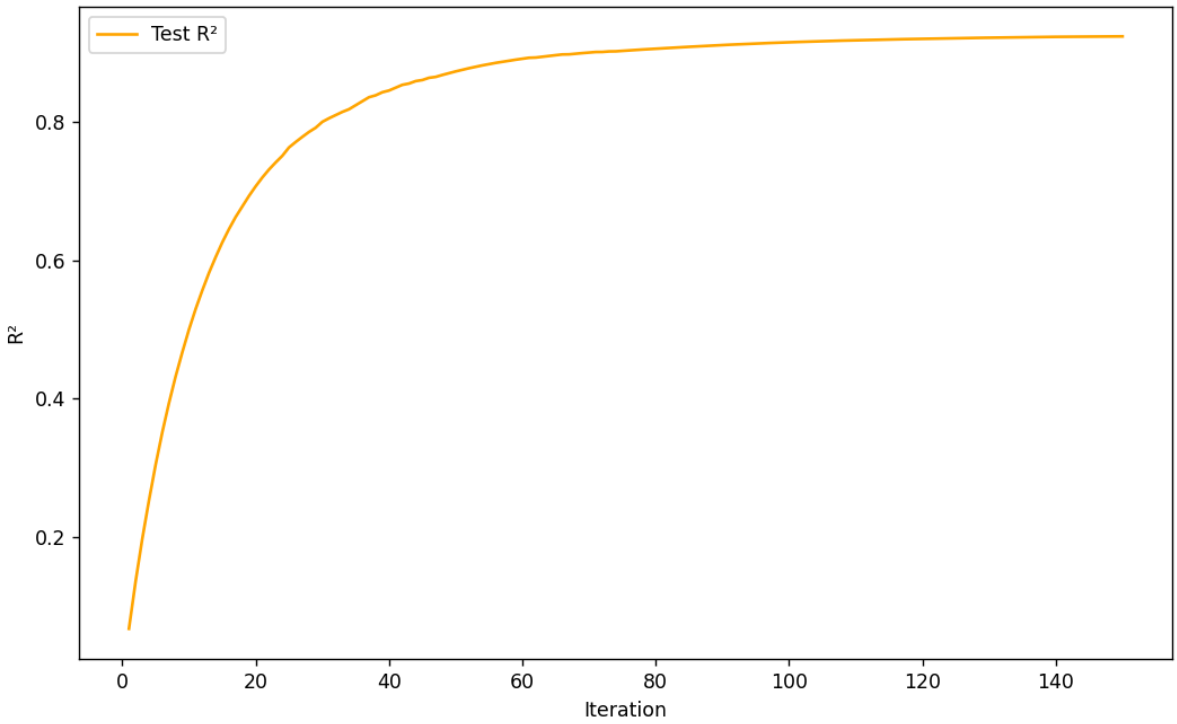


Fig.4-31 Test R^2

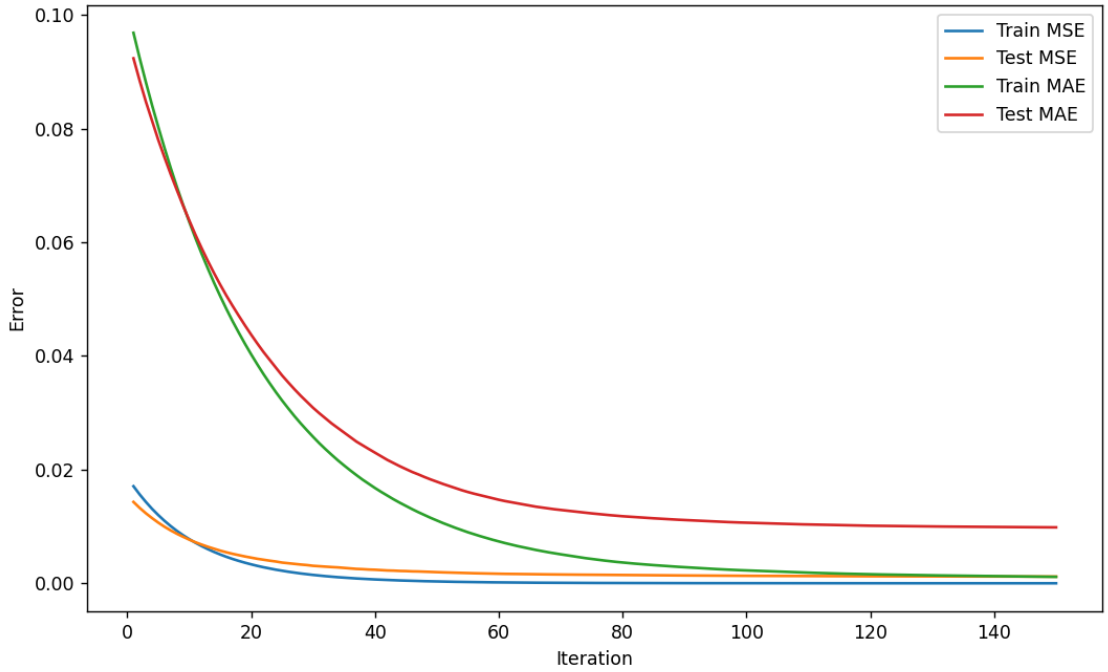


Fig.4-32 Train and Test MSE/MAE

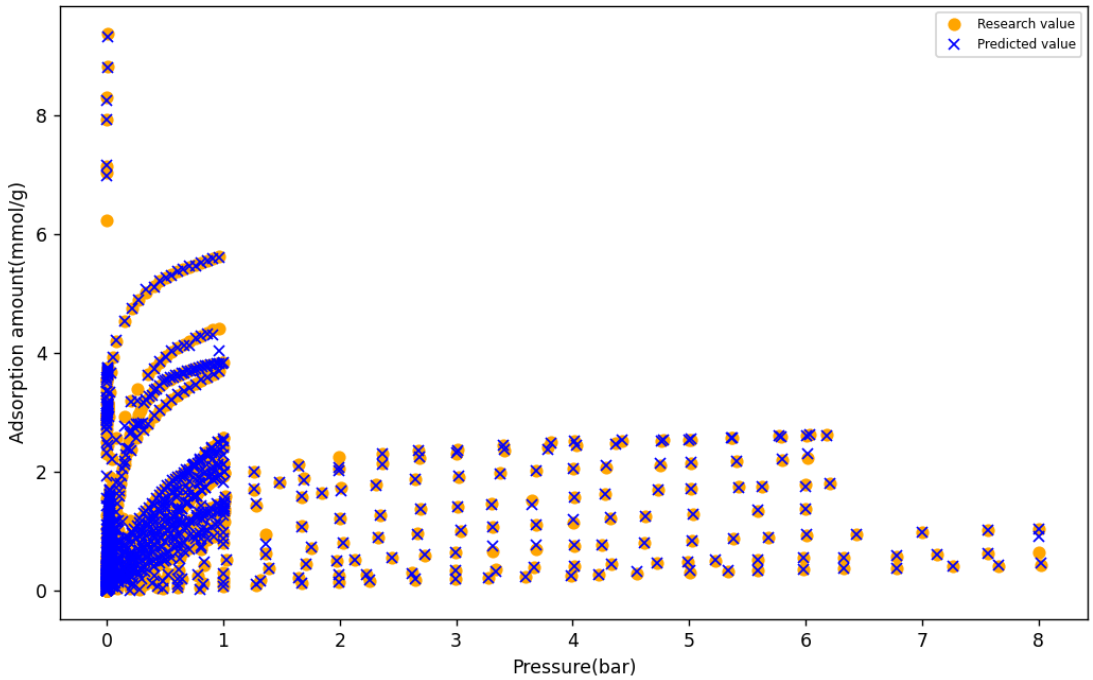


Fig.4-33 Equilibrium adsorption prediction

From the error analysis, Same as the GBDT model, the number of XGBoost model trainings was increased to 150. And a good fit was obtained. The prediction results will be discussed and analyzed in the residual plots.

4.4.3 ANN prediction model

The ANN model parameters are set as shown in the following Table 4.16.

Table 4.16 ANN model parameters

Input layer	(10, 64)
Hidden layer	(64, 64)
Output layer	(64, 1)
lr	0.0005
weight_decay	0.0005
epoch	200

The change of R2, MSE and MAE when ANN model running is shown in Fig.4-34 and 4-35. The prediction result is shown in Fig.4-36.

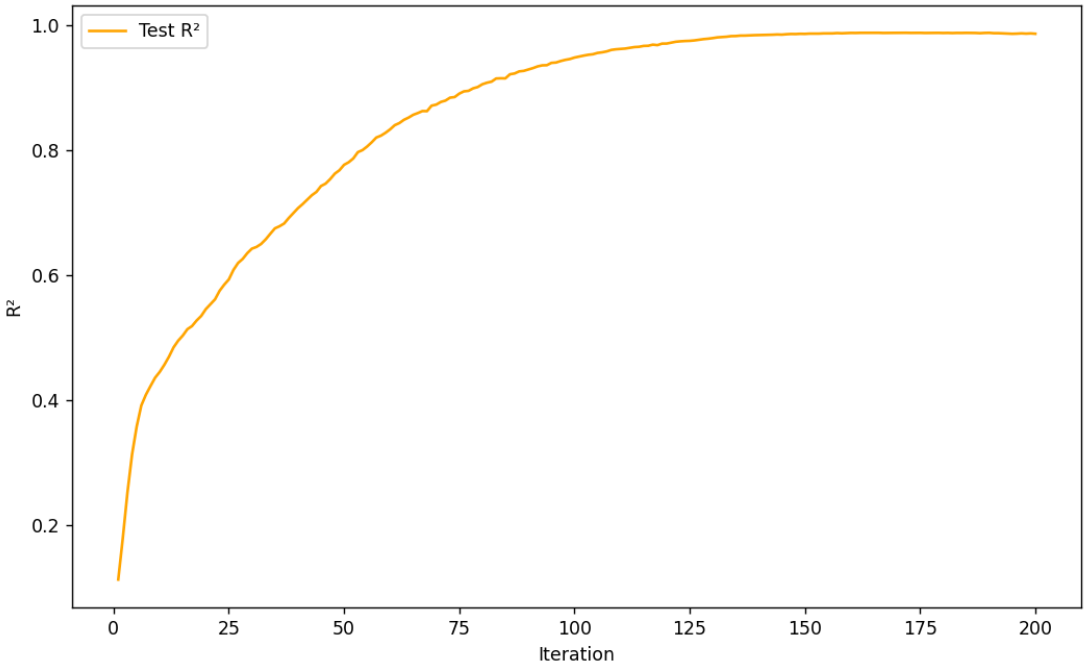


Fig.4-34 Test R^2

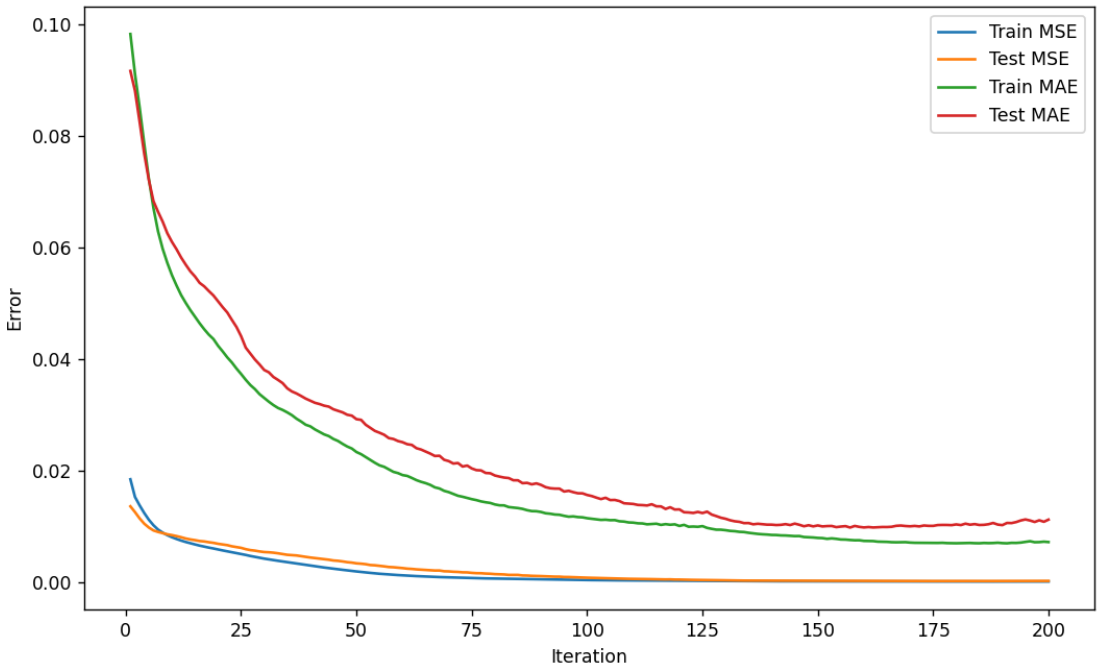


Fig.4-35 Train and Test MSE/MAE

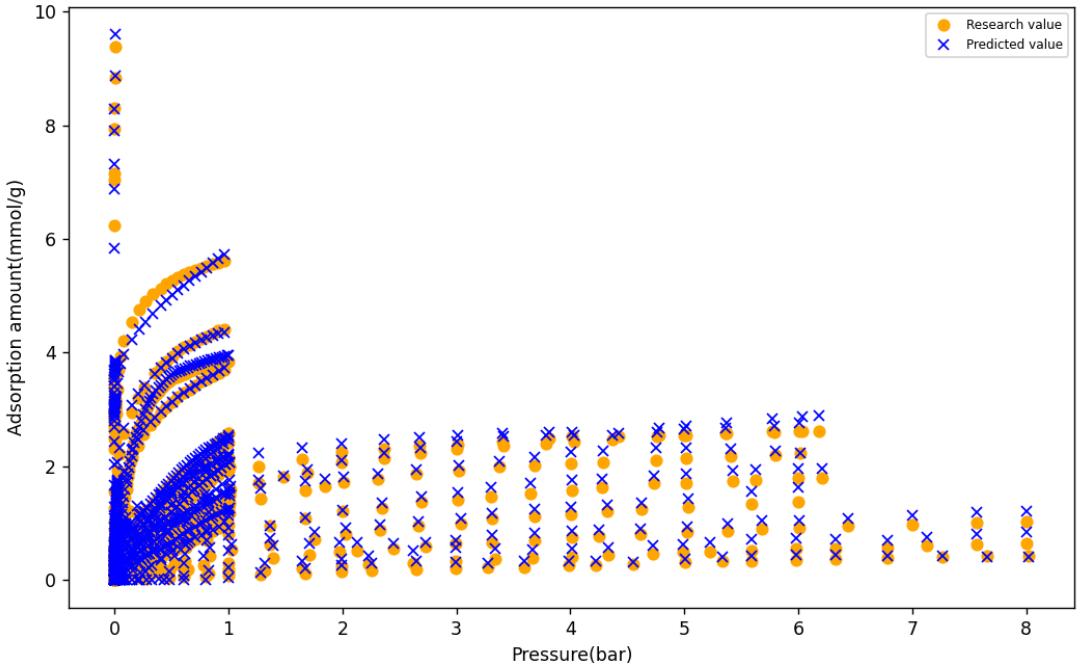


Fig.4-36 Equilibrium adsorption prediction

From the error analysis, ANN got a good fit without the training fluctuations that have been seen in the previous two situation. The ANN model trains better when dealing with large data. The prediction results will be discussed and analyzed in the residual plots.

4.4.4 DNN prediction model

The DNN model parameters are set as shown in the following Table 4.17.

Table 4.17 DNN model parameters

Input layer	(4, 8)
Hidden layer	(8, 8)
Output layer	(8, 1)
lr	0.0005
weight_decay	0.0005
epoch	200

The change of R2, MSE and MAE when ANN model running is shown in Fig.4-37 and 4-38. The prediction result is shown in Fig.4-39.

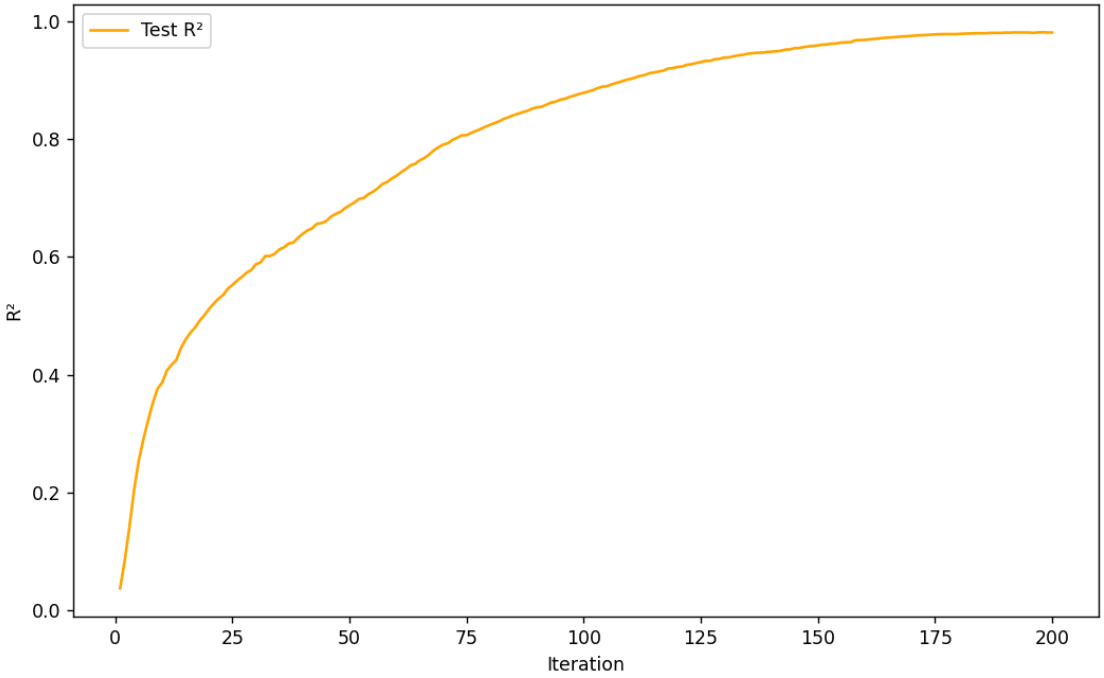


Fig.4-37 Test R^2

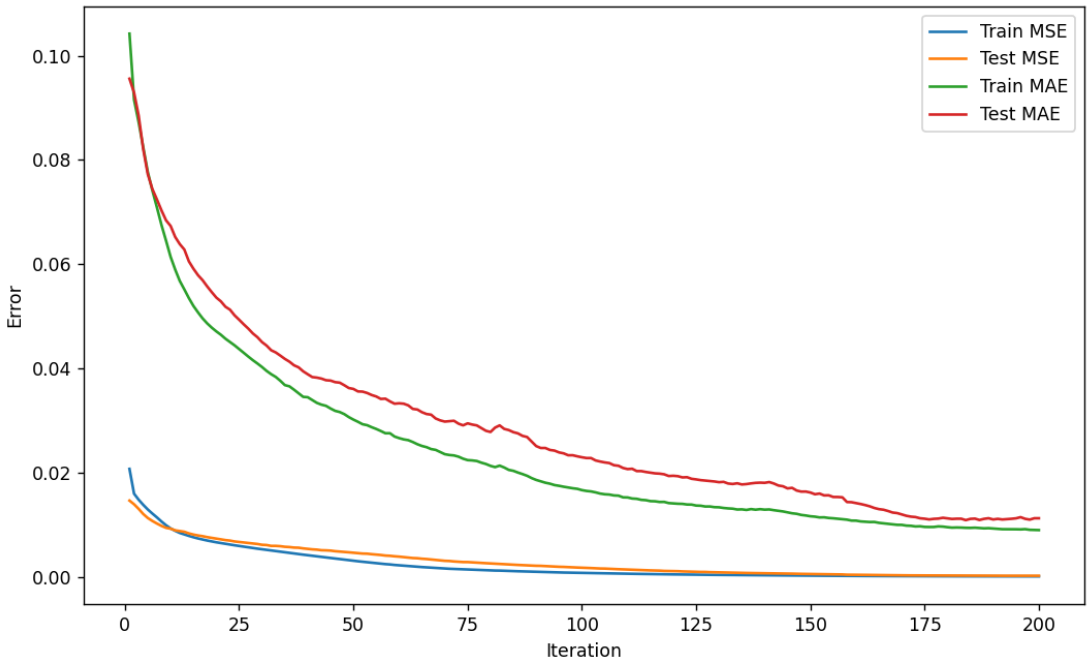


Fig.4-38 Train and Test MSE/MAE

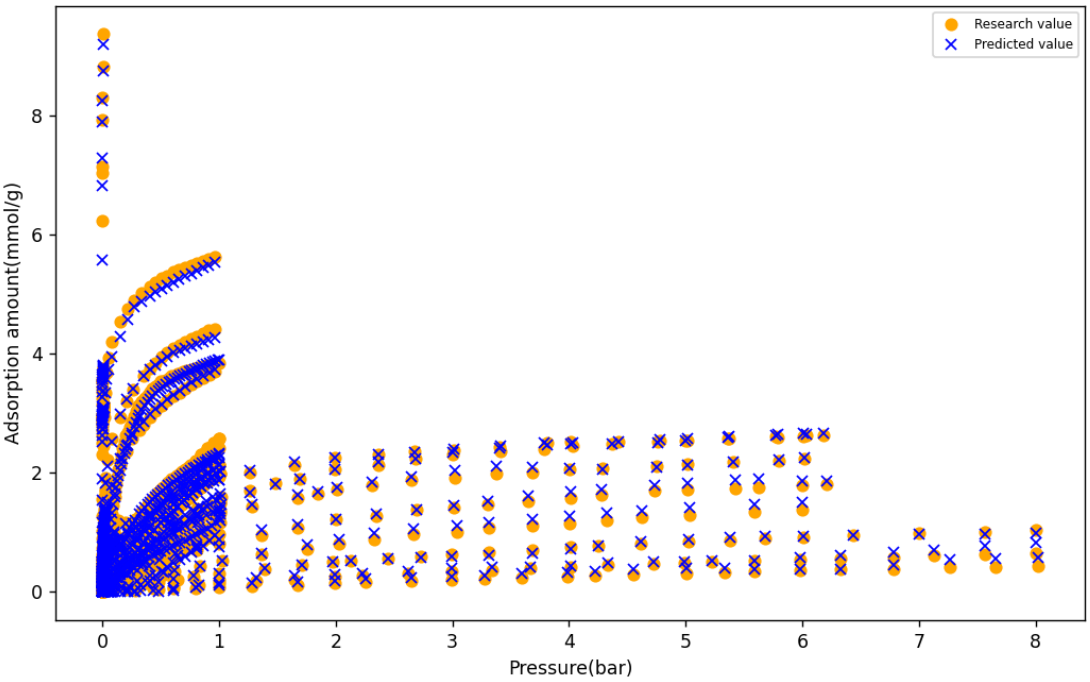


Fig.4-39 Equilibrium adsorption prediction

From the error analysis, we can find the DNN model also did not suffer from training fluctuations as in the previous two situations, but DNN required more training sessions to fit than ANN model. The prediction results will be discussed and analyzed in the residual plots.

4.4.5 Results analysis

he evaluation metrics for each model are shown in the Table 4.18.

Table 4.18 Model evaluation metrics

	MSE	MAE	R^2	TIME
GBDT	0.00092	0.00856	0.94034	45.3202s
XGBoost	0.00115	0.00965	0.92512	24.2477s
ANN	0.00022	0.01163	0.98556	14.6693s
DNN	0.00018	0.00920	0.98799	14.3087s

From the table above, it is evident that the DNN model performs most effectively in handling large-scale data. It also exhibits quicker training times compared to the traditional ANN model. While the ANN model lags the DNN model, it still outperforms the Decision Tree model in processing efficiency. For the XGBoost and GBDT models, XGBoost demonstrates significantly superior learning capabilities compared to the GBDT model. It can be observed that, as a traditional Decision Tree model, GBDT may excel in handling small-scale data, but in terms of learning from large-scale data, it falls notably behind the XGBoost model, which is also a Decision Tree model.

Fig.4-40 illustrates the residual plots for the four models. From the presentation of the residual plots, it is evident that all four models have effectively controlled most prediction errors within 15%, meeting the requirements for practical predictions. Regarding the types of errors, the errors in XGBoost and GBDT models fall under random error, whereas the errors in ANN and DNN models are categorized as systematic errors.

Figure 4-40 shows the residual plots of the four models. From the residual plots, the DNN performs as well as the evaluation parameters. And compared with the first two small data learning, DNN performs much better in large data learning, and the predictions almost match the true values. ANN predictions have some systematic biases, but they are all within 15% bias. XGBoost model and GBDT model are also good, but still have random error, and this bias is more obvious in large data learning, and the biggest error point in the figure reaches an error of 300%.

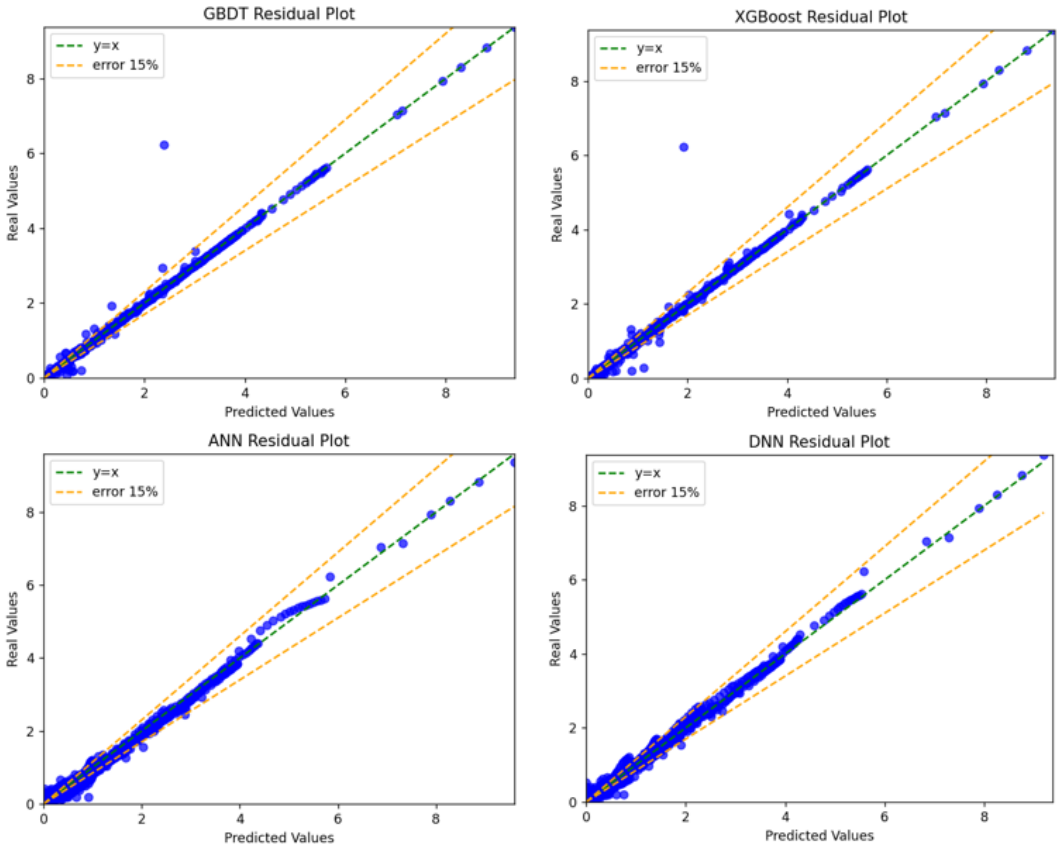


Fig.4-40 Residual plot for four models

Chapter 5: Overall conclusions

The main targets of this study are to investigate the use of machine learning models to predict the adsorption isotherms of zeolites under various operating conditions. Through a comparative analysis of learning with multiple models, the aim is to identify the most suitable model for different operating conditions. To achieve the target, the following studies were conducted.

1. A review of the past research has been done to understand the study of machine learning in zeolite adsorption prediction. In the past, Decision Tree models, represented by GBDT, have been considered as effective predictive models and have been validated through experimentation. Meanwhile, neural networks, such as deep learning, are regarded as the most advanced and outstanding machine learning models currently available. Therefore, this study selects two Decision Tree models, GBDT and XGBoost, and two Neural Network models, ANN and DNN, for investigation.
2. To get more features, CIF files of zeolite are collected in this paper and processed with software packages. A large amount of zeolite adsorption data was also collected from NIST database.
3. Machine learning using 4 models GBDT, XGBoost, ANN and DNN. Constantly optimize the model parameters during the training process to get the best model for comparison.
4. Evaluation metrics and residual plots were used to evaluate and draw conclusions about the four model predictions for each situation.

The overall conclusions summarized from a lot of efforts for above mentioned studies are given as follows:

- Deep learning neural network models are more suitable for zeolite adsorption isotherm prediction than traditional Decision Tree models.
- Regarding the type of error, the error of the Decision Tree model is mainly random error, and most of the predicted values coincide with the true values, which should be due to the regression tree structure of the decision tree model. Errors of the ANN model and DNN model are mainly systematic errors, resulting from the fact that the results fitted by the neural network model are function models.

- When dealing with small zeolite data, ANN model with simple structure is the most excellent choice. This is because the ANN training is fast compared to DNN, and the ANN model is more stable during training in terms of error fluctuations. When dealing with large zeolite data, the DNN model is the best choice, as an upgraded model of ANN, the training time, error, and stability of DNN are better than the ANN model when learning large data.
- For the GBDT model and the XGBoost model, there is not much difference between their performance when learning small data. When learning large data, the XGBoost model significantly outperforms the GBDT model, especially in terms of training time and the number of training sessions needed to reach a good fit.

For future work, due to the limitation of data collection, only zeolite adsorption was studied in this research. For the study of other materials such as MOFs (Metal Organic Frameworks), activated carbon, etc., machine learning also has a wide range of research prospects. It is anticipated that future investigations, involving the analysis of data from diverse materials, will deepen the achievements of this study. This exploration aims to delve into the application of machine learning techniques within the domain of material adsorption.

References

[1] Wang, D. C., Z. Z. Xia, and J. Y. Wu. "Design and performance prediction of a novel zeolite–water adsorption air conditioner." Energy conversion and management 47.5 (2006): 590-610.China Academy of Building Sciences. GB/T51350-2019 Nearly zero Energy Technical Standard [S]. 2019.

[2] Lu, Shuaihua, et al. "Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning." Nature communications 9.1 (2018): 3405.

[3] Rabe, Michael, Dorinel Verdes, and Stefan Seeger. "Understanding protein adsorption phenomena at solid surfaces." Advances in colloid and interface science 162.1-2 (2011): 87-106.

[4] Xia, Mengdan, et al. "Removal of Hg (II) in aqueous solutions through physical and chemical adsorption principles." RSC advances 9.36 (2019): 20941-20953.

[5] Auerbach, Scott M., Kathleen A. Carrado, and Prabir K. Dutta. Handbook of layered materials. CRC press, 2004.

[6] Wach, Wolfgang, et al. "Adsorption and HPLC of carbohydrates and related hydroxy compounds on zeolites." Analytical Methods 10.16 (2018): 1817-1832.

[7] Kunpeng, W. A. N. G., et al. "Adsorption of vanadium on 13X molecular sieve in aqueous solution." Chinese Journal of Environmental Engineering 10.11 (2016): 6249-6254.

[8] Ratel, Larysa, Frédéric Kuznik, and Kévyn Johannes. "Open Sorption Systems." (2022): 526-541.

[9] Sabzehmeidani, Mohammad Mehdi, et al. "Carbon based materials: A review of adsorbents for inorganic and organic compounds." Materials Advances 2.2 (2021): 598-627.

[10] Majd, Mahdieh Mozaffari, et al. "Adsorption isotherm models: A comprehensive and systematic review (2010− 2020)." Science of The Total Environment 812 (2022): 151334.

[11] Thu, Kyaw, et al. "Adsorption characteristics of methane on Maxsorb III by gravimetric method." Applied thermal engineering 72.2 (2014): 200-205.

[12] Keller, J., Dreisbach, F., Rave, H. et al. Measurement of Gas Mixture Adsorption Equilibria of Natural Gas Compounds on Microporous Sorbents. Adsorption 5, 199–214 (1999).

[13] LI ZHOU, JUNSHE ZHANG, YAPING ZHOU. A Simple Isotherm Equation for Modeling the Adsorption Equilibria on Porous Solids over Wide Temperature Ranges[J]. Langmuir: The ACS

Journal of Surfaces and Colloids,2001,17(18):5503-5507.

[14] Zhou, Hongjun, et al. "Molecular simulation of methane adsorption in activated carbon: the impact of pore structure and surface chemistry." Molecular Simulation 42.9 (2016): 776-782.

[15] Woo, Hyung-June, Aaron R. Dinner, and Benoît Roux. "Grand canonical Monte Carlo simulations of water in protein environments." The Journal of chemical physics 121.13 (2004): 6392-6400.

[16] Rahmati, Mahmoud, and Hamid Modarress. "Grand canonical Monte Carlo simulation of isotherm for hydrogen adsorption on nanoporous siliceous zeolites at room temperature." Applied Surface Science 255.9 (2009): 4773-4778.

[17] Lu, Shuaihua, et al. "Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning." Nature communications 9.1 (2018): 3405.

[18] Li, Yi, Jihong Yu, and Ruren Xu. "Criteria for zeolite frameworks realizable for target synthesis." Angewandte Chemie 125.6 (2013): 1717-1721.

[19] Li, Yi, et al. "In silico prediction and screening of modular crystal structures via a high-throughput genomic approach." Nature communications 6.1 (2015): 8328.

[20] Lin, Shiru, et al. "Machine-learning-assisted screening of pure-silica zeolites for effective removal of linear siloxanes and derivatives." Journal of Materials Chemistry A 8.6 (2020): 3228-3237.

[21] Lee, Sangwon, Baekjun Kim, and Jihan Kim. "Predicting performance limits of methane gas storage in zeolites with an artificial neural network." Journal of Materials Chemistry A 7.6 (2019): 2709-2716.

[22] Jaffari, Zeeshan Haider, et al. "Machine-learning-based prediction and optimization of emerging contaminants' adsorption capacity on biochar materials." Chemical Engineering Journal 466 (2023): 143073.

[23] Zhao, Fangzhou, et al. "Prediction of heavy metals adsorption by hydrochars and identification of critical factors using machine learning algorithms." Bioresource Technology 383 (2023): 129223.

[24] Yuan, Xiangzhou, et al. "Applied machine learning for prediction of CO2 adsorption on biomass waste-derived porous carbons." Environmental Science & Technology 55.17 (2021): 11925-11936.

[25] LI Bo-ran, LI Hui, TIAN Zi-qi. Progress in application of machine learning technology to study of porous materials for gas adsorption. Modern Chemical Industry, 2022, 42(4): 17-22.

[26] Sultan, Muhammad, Takahiko Miyazaki, and Shigeru Koyama. "Optimization of adsorption

isotherm types for desiccant air-conditioning applications." Renewable Energy 121 (2018): 441-450.

[27] NIST Data Resources for Adsorption Science and Technology [DB]. http://adsorption.nist.gov/.

[28] Hall, Sydney R., Frank H. Allen, and I. David Brown. "The crystallographic information file (CIF): a new standard archive file for crystallography." Acta Crystallographica Section A: Foundations of Crystallography 47.6 (1991): 655-685.

[29] Database of Zeolite Structures [DB]. http://www.iza-structure.org/databases/.

[30] Martin, Richard Luis, and Maciej Haranczyk. "Construction and characterization of structure models of crystalline porous polymers." Crystal growth & design 14.5 (2014): 2431-2440.

[31] Thommes, Matthias, and Katie A. Cychosz. "Physical adsorption characterization of nanoporous materials: progress and challenges." Adsorption 20.2-3 (2014): 233-250.

[32] Hung, Ting-Hsiang, et al. "Transport-Relevant Pore Limiting Diameter for Molecular Separations in Metal–Organic Framework Membranes." The Journal of Physical Chemistry C 125.37 (2021): 20416-20425.

[33] Andreucci, Carlos Aurelio, Elza MM Fonseca, and Renato N. Jorge. "Increased Material Density within a New Biomechanism." Mathematical and Computational Applications 27.6 (2022): 90.

[34] Melián-Cabrera, Ignacio, and Valeriya Zarubina. "Surface area per volumetric loading and its practical significance." Microporous and Mesoporous Materials 354 (2023): 112549.

[35] Kondor, Anett, et al. "On the BET surface area of nanocellulose determined using volumetric, gravimetric and chromatographic adsorption methods." Frontiers in Chemical Engineering 3 (2021): 738995.

[36] Seymour, Alexis J., Sungchul Shin, and Sarah C. Heilshorn. "3D printing of microgel scaffolds with tunable void fraction to promote cell infiltration." Advanced healthcare materials 10.18 (2021): 2100644.

[37] Deedar, N. A. B. I., and Irfan Aslam. "Evaluation of the adsorption potential of titanium dioxide nanoparticles for arsenic removal." Journal of Environmental Sciences 21.3 (2009): 402-408.

[38] Shabir, Faizan, et al. "Recent updates on the adsorption capacities of adsorbent-adsorbate pairs for heat transformation applications." Renewable and Sustainable Energy Reviews 119 (2020): 109630.

[39] Jebli, Imane, et al. "Prediction of solar energy guided by pearson correlation using machine learning." Energy 224 (2021): 120109.

[40] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." Journal of Applied Science and Technology Trends 2.01 (2021): 20-28.

[41] Zhao, Fangzhou, et al. "Prediction of heavy metals adsorption by hydrochars and identification of critical factors using machine learning algorithms." Bioresource Technology 383 (2023): 129223.

[42] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[43] Prakash, Anand, et al. "Model-aware XGBoost Method towards Optimum Performance of Flexible Distributed Raman Amplifier." IEEE Photonics Journal (2023).

[44] Rosenblatt, Frank. The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory, 1957.

[45] Malekian, Arash, and Nastaran Chitsaz. "Concepts, procedures, and applications of artificial neural network models in streamflow forecasting." Advances in streamflow forecasting. Elsevier, 2021. 115-147.

[46] Werbos, Paul J. "Applications of advances in nonlinear sensitivity analysis." System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[47] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.

[48] Karaci, Abdulkadir, et al. "Estimating the properties of ground-waste-brick mortars using DNN and ANN." Computer Modeling in Engineering & Sciences 118.1 (2018): 207-228.

[49] Galimberti, Clara Lucía, et al. "Hamiltonian Deep Neural Networks Guaranteeing Nonvanishing Gradients by Design." IEEE Transactions on Automatic Control 68.5 (2023): 3155-3162.

[50] Steurer, Miriam, Robert J. Hill, and Norbert Pfeifer. "Metrics for evaluating the performance of machine learning based automated valuation models." Journal of Property Research 38.2 (2021): 99-129.

[51] Naidu, Gireen, Tranos Zuva, and Elias Mmbongeni Sibanda. "A Review of Evaluation Metrics in Machine Learning Algorithms." Computer Science On-line Conference. Cham: Springer

International Publishing, 2023.

[52] Mathew, Amitha, P. Amudha, and S. Sivakumari. "Deep learning techniques: an overview." Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2020 (2021): 599-608.

[53] Espinheira, Patrícia L., Luana CM Silva, and Francisco Cribari-Neto. "Bias and variance residuals for machine learning nonlinear simplex regressions." Expert Systems with Applications 185 (2021): 115656.

Acknowledgement

My deepest gratitude is to my supervisors, Professor Takahiko Miyazaki, Associate Professor Kyaw Thu of Energy and Environmental Engineering, IGSES at Kyushu University, for their invaluable and indispensable guidance and inspiration throughout not only my research work but also life as master student.

I am also indebted to Mr. Nobuo Takada and Mr. Mojtaba Purjam for their valuable suggestion as well as helping me to do experiments smoothly.

Special appreciation to my classmates Mr. Nakatsu Hideaki, Mr. Mizobe Kouki, Mr. Eguchi Masatsugu and Mr. Matsuo Kosuke for their grateful support and helping me not only study but also throughout my entire life in master program.

I would like to express my gratitude to Academic Researcher Dr. Chairunnisa, 3rd year PhD student Mr. Mansoor Abdul Aziz, Mr. Motaz Mabrok Mohamed Salama, Mr. Yang Zhaosheng, Mr. Yu Hao, Mr. Seo Sang Won, Mr. Marco Reyes Lao, Mr. Haoulomou Pepe, Mr. Sannan Salabat Butt, Mr. Liu Xuetao, Mr. Denis Mutebi, 2nd year PhD student, Mr. Chen Haonan, Mr. Yang Cheng, Mr. You Cheng, Mr. Li Guangyao, 1st year PhD student Mr. Chai Jiazi, Mr. Zhao Peng, Ms. Wu Yitong, Mr. Fujita Daido, 2nd year MSc student Ms. Zhou Yingxin, Mr. Xing Junxi, Ms. Alaa Itaiwi, 1st year MSc student Mr. Arima Taiki, Mr. Isonaga Takashi, Mr. Kudou Norito, Mr. Wakigami Atsushi, Ms. Sameen Isha, 4th year undergraduate student Mr. Nagase Taiki Mr. Fujimoto Daiki and all other members who have graduated and left the research lab.

I also would like to express my sincere gratitude to Ms. Yuri Yamato, secretary of the laboratory, who helped me with many administrative procedures and helped me a lot in my student life. I would like to express my deepest gratitude to her.

Finally, I would like to express my deepest gratitude to my family for their warm support and encouragement during my student life.