Discovering zeolite adsorption isotherms: a hybrid AI modeling approach

Arijit Chakrabortya, Akhilesh Gandhib, M. M. Faruque Hasanb and

Venkat Venkatasubramaniana\*

aComplex Resilient Intelligent Systems Laboratory, Department of Chemical Engineering, Columbia University, New York, NY 10027, USA

bArtie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX 77843, USA

\*venkat@columbia.edu

Abstract

Zeolites renowned for their porous structure, and adsorption capabilities, play a pivotal role in various applications across energy, health, and the environment. Despite advances in experimental techniques, the problem of predicting the adsorption isotherm of a zeolite – given its structural properties – remains inherently challenging. A key aspect of such a problem is that it is inherently a "precious data" problem, due to the limitation of having a small dataset. In this work, we propose a hybrid AI modeling approach for obtaining predictive models for zeolite adsorption isotherms. Given the structural properties of a zeolite, we are able to predict the adsorption property reasonably well, in contrast to conventional machine learning techniques which – despite their relatively complex parameterizations – are unable to provide reliable results. We rely on a combination of symbolic AI, and intelligent feature engineering, resulting in concise and interpretable adsorption isotherm models for different temperatures. The benefits of the same are that the models generalize well on zeolites that have not been encountered during the training phase, while being relatively simple in its form. This enables post-hoc analyses for gaining insights on the descriptive capabilities of the structural properties used.

**Keywords**: Zeolites, Adsorption isotherm, Artificial intelligence, Machine Learning, Hybrid AI.

* 1. Introduction

The advent of enhanced computational power along with vast amounts of data, has expanded the applicability of artificial intelligence (AI) – specifically, machine learning (ML) – across a wide range of domains. These include speech recognition, entertainment recommender systems, chat-bots like ChatGPT (Brown et al., 2020), to name a few. Unlike sciences and engineering, these domains have colossal amounts of data at their disposal, with more being acquired every day. We refer to the sciences and engineering as ”precious” data domains as opposed to ”big data”, since the cost of obtaining a new datapoint for analysis is prohibitively expensive. Accordingly, conventional AI/ML techniques are not feasible for performing analyses on such limited datasets. Further,
unlike the sciences and engineering, these domains are not governed by underlying fundamental laws and mathematical relations, nor do they have underlying expert knowledge that can be used to guide one towards a better prediction.

Thus, in order to tackle such problems, we suggest the incorporation of relevant first-principles knowledge, along with modifications in conventional AI/ML techniques to enhance the predictive capabilities of data-driven models. In this work, we tackle such a problem for the prediction of the adsorption capacity of carbon-dioxide (CO2) gas on various zeolites at 2 temperatures – 323 K and 373 K. The goal is to predict the adsorption isotherm, given the structural properties of a zeolite. We are limited by the size of the dataset (here, 188 zeolites). Accordingly, this is a *precious* data problem. Relying on the inclusion of first-principles’ based feature engineering, and evolutionary-algorithmic search for mathematical models, we successfully obtain reasonably accurate models using such a restrictive dataset. We highlight the efficacy of our approach, and how it results in more interpretable models that enable a domain-expert to gain additional insights.

* 1. Background

Zeolites and other porous structures such as metal-organic frameworks (MOFs) are known for their adsorption properties, and thus, have applications in the most crucial chemical industries. Towards the goals for sustainability, these structures play a pivotal role in gas adsorption such as applications in carbon-capture for lowering global CO2 levels, hydrogen storage for energy transition, oxygen concentrators (which helped save lives during the pandemic), and in catalysis. Prediction of adsorption properties for these materials is paramount, and has thus far been heavily reliant on experimental methods, which were time and resource intensive. With the advent of computational modeling, recent methods such as Monte Carlo (MC) simulations were developed to address the highly resource-intensive experiments. However, with the advent of computational methods, came the possibility of design of new structures and thus, several million pure-Si zeolite frameworks have been hypothesized. The computational requirement of the Grand Canonical Monte Carlo (GCMC) models are substantially high, when considering millions of such hypothetical designs.

Several works have tried to address the adsorption prediction for these zeolites due to their key importance in the materials design space (Raji et al., 2022; Okello et al., 2023; Alizadeh et al., 2022). However, most of the current literature is focused on building models to predict properties of known zeolites, and to enhance the predictions of molecular simulations. Some approaches also focus on the high throughput screening of zeolites (Moliner et al., 2019). However, property prediction for unknown zeolites adsorption data for which is not readily available – has remained an open challenge for exploration.

Towards property prediction of hypothetical structures, one of the key limitations is the representation for these crystal structures. These are cumbersome, in contrast to molecular representations which are studied in detail such as SMILES, SELFIES, molecular fingerprints etc. This limitation was overcome recently in a work by Gandhi et al. (Gandhi and Hasan, 2021), where a graph-theoretic representation for crystal structures such as zeolites, was developed. Specifically for zeolites, the architecture of the process for inverse design and property prediction has been laid out in literature (Gandhi and Hasan, 2022). A key limitation however in the case of adsorption is the lack of data to build autonomous ML models. Towards this end, we propose a hybrid AI model for the prediction of adsorption in zeolites based on structural features that can be computed easily, and can be modeled with the physical knowledge of the system to overcome the data limitations.

In this section, we have introduced the importance of adsorption isotherm modeling, and enumerated some of the pertinent challenges faced by researchers in this domain. We emphasize the need for an alternate modeling approach – one that combines the benefits of first-principles knowledge, with a data-driven approach – resulting in a hybrid AI (Chakraborty et al., 2022) model. In the next section, we discuss the approach used for modeling the adsorption isotherm of zeolites for CO2 gas. We subsequently present the results of our approach, and highlight the efficacy of domain-knowledge-driven feature engineering on the limited dataset available (188 zeolites). Finally, we conclude this manuscript by summarizing our work, and provide future directions that
are currently being undertaken, and for guiding researchers in this field.

* 1. Methods

In this section, we discuss the data used for the problem at hand, the approach used to obtain an interpretable data-driven model for the prediction of CO2 adsorption isotherm of zeolites based on their structural properties.

* + 1. Data

Adsorption of CO2 on pure-silica zeolites is a resource-intensive computation for different pressures and temperatures. Also, experimental methods are limited by synthesis and availability of these zeolite structures limiting the scope to only computational molecular simulations. In this case, we used the GCMC simulations to determine the adsorption capacity of each of the zeolite structures at different P-T conditions (Iyer and Hasan, 2019). Furthermore, the data validation only makes sense for observed zeolites where we can validate the GCMC parameters. Thus, expanding the dataset over 188 zeolites is extremely resource-intensive and challenging to validate.

The descriptors in the dataset used in this work capture the physical structure of the zeolites. Some of these descriptors, including density, Si-O-Si bond angle (average, harmonic mean, min, max, std, etc), statistics on the bond distance, and details on accessible and non-accessible surface area, among others, are computable from the location of the Si and O atoms. Some of these descriptors are already pre-computed in the IZA database and have been used in other studies regarding ML-based modeling of mechanical properties of zeolites. The data used in our work is taken from a previous study (Evans and Coudert, 2017) where the data is freely available. Accordingly, we proceed with 16 descriptor variables for each zeolite, which are further transformed by the modelling tool, as discussed in the next section.

In our work, we split the dataset into a train and test set with 80% zeolites (150 zeolites) categorized into the train set, with the remaining 20% (38 zeolites) categorized into the test set.

* + 1. Modeling tool – AI-DARWIN

Relying on the formulation of obtaining a mathematical model as a search problem, we build on the genetic algorithmic-based model discovery engine AI-DARWIN (Chakraborty et al., 2021), which is a more generalized extension of a linear model discovery engine titled GFEST (Chakraborty et al., 2020). Previously, this approach has yielded success in the prediction of a bubble column aeration process (Jul-Rasmussen et al., 2023), which relied on real and noisy data from a pilot plant. It relies on the use of genetic algorithm (GA), coupled with a function library where user-specified function transformations act upon the input variables, to generate features for subsequent regression. For a detailed description of the algorithm, the reader is directed to the original research articles discussing this method, and its efficacy on a variety of case-studies.

* + 1. First-principles’ based feature engineering

Upon utilizing this approach without any modifications and permitting all function transformations, we obtained unsatisfactory results. A common result from these inaccurate models was the prevalence of pressure as a major contributing factor to the prediction of the adsorption loading, which overshadowed the effect of the structural properties of the zeolite. Such a model relied entirely on the pressure of adsorption process, and failed to capture the variability caused due to a different zeolite structure, and by extension, its properties. Accordingly, it would be incorrect to apply such an approach for different zeolites.

Empirical models of adsorption include the Langmuir adsorption model, given as

|  |  |
| --- | --- |
| $$θ=\frac{K\_{eq}P}{1+K\_{eq}P}$$ | (1) |

Here, $θ$ is the fraction of the adsorption sites occupied, $K\_{eq}$ is the equilibrium constant, and $P$ is the pressure at which adsorption is taking place. Here, the effect of pressure is accounted for as a direct effect. Inspired by the form of the Langmuir adsorption model, we choose to separate the effects of the structural properties from pressure, and account for both, such that the effect of pressure does not overwhelm the predictive capability of the zeolite’s physical properties. Our proposed model takes the following form:

|  |  |
| --- | --- |
| $$θ\_{AI-DARWIN}=\sum\_{i}^{14}\frac{β\_{i}f\_{i}P}{1+β\_{i}f\_{i}P} $$ | (2) |

Here, $f\_{i}$ represents the features generated by the AI-DARWIN model, and $β\_{i}$ is the parameters obtained upon performing parameter estimation on the model. It must be noted that the model form does not have an intercept. It was formulated as such, to impose the physical constraint that at zero pressure, the absolute loading (adsorption capacity) be zero.

* 1. Results and Discussion

The results are depicted for 323 K in Figure 1, and 373 K in Figure 2. The model for 323 K (R2 of 0.65 on training set, and 0.87 on test set) was less accurate as compared to the model for 373 K (R2 of 0.87 on training set, and 0.94 on test set), likely due to the training data for numerous zeolites in the dataset being limited to low and medium pressures (up to approximately 4 MPa). Thus, there was not sufficient variability in the training data to yield more accurate predictions.

We see that by virtue of having chosen a model without an intercept, we can enforce that the absolute loading would be zero, at a pressure of zero Pa. The model is able to fit accurately (R2 > 0.9) for most zeolites. There are some errors where the model predicts adsorption isotherms which are vastly different from the true isotherms in some cases, such as for zeolites SBS at 323 K and GOO at 373 K (depicted in Figure 1 (a) and Figure 2 (a) respectively). This is likely due to reduced variability in the training data, for the model to recognize salient features effectively. In future research, this can be addressed by using alternate parametrizations, and/or enforcing constraints which account for the smoothness of the curves.



Figure 1: Adsorption isotherms at 323 K for (a) SBS zeolite, and (b) UTL zeolite.



Figure 2: Adsorption isotherms at 373 K for (a) GOO zeolite, and (b) IWW zeolite

We obtained high test R2 values for both the temperatures, which is a success of our hybrid AI approach. While it might seem that these scores could be a result of the small size of the test dataset (38 zeolites), we tested the model multiple times with different randomized splits of zeolites, with 80% used for training, and the remaining 20% used for testing, such that our testing metrics are not biased. Accordingly, we obtained similar performance with negligible variability across the different random splits of the dataset.

A key advantage of our method is its inherent interpretability since we know which feature combinations (and their weights) affect the adsorption capacity. These can be used to estimate the differentiating factor for adsorption at a higher temperature (373 K), when compared to one at a lower temperature (323 K). Such analysis is not possible when using a purely black-box modeling technique such as a neural network.

* 1. Conclusions and future work

Having obtained a reasonably accurate predictive model for adsorption isotherm from a dataset with less than 200 zeolites, it is important for future research to explore the explainability of such models. Particularly, researchers must probe into how the variables and their interactions captured in the model, relate to the adsorption capacity. Such directions have the potential to provide novel insights, previously unknown by the respective subject matter experts. While quantitative estimates of these interactions and their resultant effects are possible, for e.g., using Shapley values (Shapley, 1953)), they do not provide us any insight into the fundamental process(es) that take place. Hence, these explanations will need to be guided by first-principles, since there is ample theory on the topic of adsorption in zeolites. This includes, but is not limited to, the development of causal models, and creating a knowledge base of the underlying theory such that inferences can be automatically made when new information is presented. A first step in this direction, is the use of hybrid AI models which enable post-hoc interpretability of the obtained models.

References

S. M. S. Alizadeh, Z. Parhizi, A. H. Alibak, B. Vaferi, S. Hosseini, 2022. Predicting the hydrogen uptake ability of a wide range of zeolites utilizing supervised machine learning methods. International Journal of Hydrogen Energy 47 (51), 21782–21793.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

A. Chakraborty, S. Serneels, H. Claussen, V. Venkatasubramanian, 2022. Hybrid ai models in chemical engineering–a purpose-driven perspective. Computer Aided Chemical Engineering 51, 1507–1512.

A. Chakraborty, A. Sivaram, L. Samavedham, V. Venkatasubramanian, 2020. Mechanism discovery and model identification using genetic feature extraction and statistical testing. Computers & Chemical Engineering 140, 106900.

A. Chakraborty, A. Sivaram, V. Venkatasubramanian, 2021. Ai-darwin: A first principles-based model discovery engine using machine learning. Computers & Chemical Engineering 154, 107470.

J. D. Evans, F.-X. Coudert, 2017. Predicting the mechanical properties of zeolite frameworks by machine learning. Chemistry of Materials 29 (18), 7833–7839.

A. Gandhi, M. F. Hasan, 2021. A graph theoretic representation and analysis of zeolite frameworks. Computers & Chemical Engineering 155, 107548.

A. Gandhi, M. F. Hasan, 2022. Machine learning for the design and discovery of zeolites and porous crystalline materials. Current Opinion in Chemical Engineering 35, 100739.

S. S. Iyer, M. F. Hasan, 2019. Mapping the material-property space for feasible process operation: application to combined natural-gas separation and storage. Industrial & Engineering Chemistry Research 58 (24), 10455–10465.

P. Jul-Rasmussen, A. Chakraborty, V. Venkatasubramanian, X. Liang, J. K. Huusom, 2023. Identifying first-principles models for bubble column aeration using machine learning. In: Computer Aided Chemical Engineering. Vol. 52. Elsevier, pp. 1089–1094.

M. Moliner, Y. Rom ́an-Leshkov, A. Corma, 2019. Machine learning applied to zeolite synthesis: the missing link for realizing high-throughput discovery. Accounts of chemical research 52 (10), 2971–2980.

F. O. Okello, T. T. Fidelis, J. Agumba, T. Manda, L. Ochilo, A. Mahmood, A. Pembere, 2023. Towards estimation and mechanism of co2 adsorption on zeolite adsorbents using molecular simulations and machine learning. Materials Today Communications 36, 106594.

M. Raji, A. Dashti, M. S. Alivand, M. Asghari, 2022. Novel prosperous computational estimations for greenhouse gas adsorptive control by zeolites using machine learning methods. Journal of Environmental Management 307, 114478.

L. S. Shapley, 1953. A value for n-person games 2, 307–317.