FEDERAL UNIVERSITY OF SANTA CATARINA
GRADUATE PROGRAM IN CHEMICAL ENGINEERING

Luís Henrique Zimmermann Feistel

ARTIFICIAL INTELLIGENCE INTEGRATED WITH MOLECULAR SIMULATION IN THE STUDY OF THE MULTISCALE MODELING OF CARBON DIOXIDE ADSORPTION

Florianópolis/SC
2024

Luís Henrique Zimmermann Feistel

ARTIFICIAL INTELLIGENCE INTEGRATED WITH MOLECULAR SIMULATION IN THE STUDY OF THE MULTISCALE MODELING OF CARBON DIOXIDE ADSORPTION

Master thesis for the degree of Master in Chemical Engineering presented to the Graduate Program in Chemical Engineering at the Federal University of Santa Catarina.
Advisor: Prof. Dr. Cíntia Soares
Co-Advisor: Prof. Dr. Natan Padoin
Prof. Dr. Savio Leandro Bertoli

Florianópolis/SC
2024

Zimmermann Feistel, Luis Henrique
ARTIFICIAL INTELLIGENCE INTEGRATED WITH MOLECULAR SIMULATION IN THE STUDY OF THE MULTISCALE MODELING OF CARBON DIOXIDE ADSORPTION / Luis Henrique Zimmermann Feistel ; orientadora, Cintia Soarez, coorientador, Natan Padoin, coorientador, Savio Leandro Bertoli, 2024.
123 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Química, Florianópolis, 2024.

Inclui referências.

1. Engenharia Química. 2. CO2 adsorption. 3. Multi-scale modeling. 4. Molecular simulation. 5. Machine Learning. I. Soarez, Cintia. II. Padoin, Natan. III. Bertoli, Savio Leandro IV. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Química. V. Título.

Luís Henrique Zimmermann Feistel

ARTIFICIAL INTELLIGENCE INTEGRATED WITH MOLECULAR SIMULATION IN THE STUDY OF THE MULTISCALE MODELING OF CARBON DIOXIDE ADSORPTION

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Nicolas Spogis, Dr.
Universidade Estadual de Campinas (UNICAMP)

Prof. Frederico Wanderley Tavares, Dr.
Universidade Federal do Rio de Janeiro (UFRJ)

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Engenharia Química pelo Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Santa Catarina.

Documento assinado digitalmente
Debora de Oliveira
Data: 11/11/2024 07:41:24-0300
CPF: ***.453.929-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

Documento assinado digitalmente
Cintia Soares
Data: 11/11/2024 06:56:11-0300
CPF: ***.401.629-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Dr. Cíntia Soares
Advisor

Florianópolis/SC, 2024.

This work is dedicated to the ones who kept,
and to those who keep,
always,
moving forward.

ACKNOWLEDGEMENTS

Professionally, it is a privilege to develop a master's thesis in a high-quality graduate course with the mentoring of highly qualified Professors. Hence, I thank the Federal University of Santa Catarina and the Graduate Program in ChemEng for their commitment.

My thankfulness also goes to my co-advisor, Pr. Savio Leandro Bertoli for his inquisitive way of questioning, as he once said in a moment of joy. Your questions enlightened my perspective.

I am also thankful to my co-advisor, Prof. Natan Padoin. With his patience and knowledge, made me look with calmness toward my ambitions. Your composure is unique.

To my advisor, Prof. Cíntia Soares, I dedicate a special thank: For your trustfulness in developing my ideas with freedom of thought, originality, and creativity. Your patience and commitment to excellence will always be a reference for me.

I would like to thank the Human Resource Program (PRH) from the National Agency of Petroleum (ANP) for the financial support and privilege of being part of that initiative. To CNPq, the National Council for Scientific and Technological Development, a special thanks as well.

I am thankful for the love of my mother, who welcomed me when I needed love; my father, who was my best friend when I needed it; and my sister, who enlightened me when I needed inspiration; for my younger brother, who made me laugh when I was needed for joy; for my grandmother, who, by her gratitude, invested in me and my family's education; for my stepfather and stepmother, always looking upon my mom and dad with love; lastly, for my old friends, whom I am thankful for their way of friendship and presence.

To my friends of LABMAC and PósENQ, thanks for sharing moments of joy, laughter, support, and guidance. The scientific coffee moments were remarkable for me. It was priceless what we all shared.

“Friendship is like good coffee;"
(KANT, IMMANUEL, 1770)

RESUMO

A integração da modelagem multi-escala com simulação molecular e algoritmos de aprendizado de máquina representa uma abordagem computacional promissora para explorar operações físico-químicas em diversas escalas. Este estudo foca na sinergia entre algoritmos de aprendizado de máquina, simulações moleculares e modelagem determinística para investigar a adsorção de CO_2. A aplicação baseia-se na modelagem das interações em nanoescala para a adsorção de CO_2 por métodos de Monte Carlo no Grande Canônico (GCMC), onde as propriedades em nanoescala são avaliadas. Essas propriedades são usadas como entradas em Modelos de Aprendizado de Máquina para prever os indicadores de desempenho que descrevem a Curva de Ruptura de um sistema de adsorção em leito fixo (macro escala) para três materiais diferentes. Os resultados obtidos usando a metodologia proposta demonstram uma concordância satisfatória, com valores médios do Erro Quadrático Médio (MSE) e Erro Quadrático Médio da Raiz (RMSE) na validação da isoterma de adsorção em nanoescala sendo 1.0955 mol/kg e 0.8588 mol/kg, respectivamente. Na macro escala, o RMSE foi 0.0565, e o MSE ficou abaixo de 0.0032 para a carga do adsorbato. A aplicação de algoritmos de aprendizado de máquina destaca a superioridade das Redes Neurais Artificiais (0.0565, 0.0032, 1.260%, 0.9864), conforme evidenciado por indicadores como MSE, RMSE e R^2, contribuindo para avanços na compreensão da adsorção de CO_2 e seu impacto nas mudanças climáticas globais. A metodologia XAI é empreendida para verificar se os pesos associados a cada variável têm uma relação física com a operação macro simulada. A metodologia geral adotada é promissora e pode ser expandida para a integração de diferentes modelos e operações, aproveitando suas propriedades de multi-escala.

Palavras-chave:CO_2 adsorption, Multi-scale, Molecular simulation, Machine learning algorithms, Breakthrough curve.

ABSTRACT

The integration of multi-scale modeling with molecular simulation and machine learning algorithms represents a promising computational approach to explore physico-chemical operations across various scales. This study focuses on the synergy between machine learning algorithms, molecular simulations, and deterministic modeling to investigate CO_2 adsorption. The application relies on the modeling of nanoscale interactions for CO_2 adsorption by Grand Canonical Monte Carlo (GCMC) methods, where nano-scale properties are evaluated. These properties are used as inputs within Machine Learning Models to predict the performance indicators that describe the Breakthrough Curve of a fixed-bed adsorption system (macro scale) for three different materials. The results obtained using the proposed methodology demonstrate satisfactory agreement, with mean values of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) in the validation of the nano-scale adsorption isotherm being 1.0955 mol/kg and 0.8588 mol/kg, respectively. On the macro scale, the RMSE was 0.0565, and the MSE was below 0.0032 for the adsorbate load. The application of machine learning algorithms highlights the superiority of Artificial Neural Networks (0.0565, 0.0032, 1.260%, 0.9864), as evidenced by indicators such as MSE, RMSE, and R^2, contributing to advancements in the understanding of CO_2 adsorption and its impact on global climate change. XAI methodology is undertaken to verify if the weights associated with each variable have a physical relation with the macro-operation simulated. The overall methodology undertaken is promising and can be expanded towards the integration of different models and operations, taking advantage of its multi-scale properties.

Keywords:CO_2 adsorption, Multi-scale, Molecular simulation, Machine learning algorithms, Breakthrough curve.

RESUMO EXPANDIDO

Introdução

A crescente urgência das mudanças climáticas, impulsionada pelo aumento das emissões de gases de efeito estufa, especialmente o dióxido de carbono (CO_2), demanda soluções inovadoras. As tecnologias de Captura, Utilização e Armazenamento de Carbono (CCUS) têm ganhado atenção significativa para mitigar as emissões de CO_2. Apesar dos avanços, ainda existem desafios em escalonar essas tecnologias para atender às demandas globais. O presente estudo foca no desenvolvimento de métodos computacionais voltados para a adsorção de CO_2, especificamente por meio de simulações moleculares e técnicas de aprendizado de máquina (ML), visando melhorar a eficiência operacional e a escalabilidade. A pesquisa explora a modelagem multiescala, com o objetivo de conectar interações em nível molecular aos resultados em larga escala, abordando questões-chave como a eficiência de um leito de adsorção analisado por indicadores de operação. A integração de simulações moleculares in silico com ML oferece uma abordagem orientada por dados para melhorar a previsibilidade e o desempenho da adsorção de CO_2. A dissertação está estruturada em torno de três componentes principais: experimentação em nanoescala por meio de simulações moleculares, modelagem em macroescala e a aplicação de ML para a conexão entre escalas. O trabalho oferece insights sobre como ferramentas computacionais avançadas podem aprimorar as tecnologias de captura de CO_2, proporcionando uma estrutura inovadora para futuros desenvolvimentos na área.

Objetivos

Os objetivos específicos incluem, primeiramente, modelar a adsorção de CO_2 em diferentes materiais, como zeólitas e estruturas metalo-orgânicas (MOFs), utilizando simulações moleculares. Além disso, será realizado o estudo de um sistema de adsorção em leito fixo desses mesmos materiais através de modelos determinísticos, validando os resultados das simulações moleculares com dados experimentais disponíveis na literatura. Outro objetivo é desenvolver uma metodologia capaz de integrar dados de diferentes escalas referentes à adsorção de CO_2 em um único conjunto de dados. Por fim, será implementado um conjunto de algoritmos supervisionados de aprendizado de máquina, treinados com dados gerados por simulações moleculares, para prever os indicadores de desempenho em simulações em macroescala e fornecer insights científicos sobre o campo da adsorção de CO_2.

Metodologia

A metodologia desenvolvida para esta tese integra simulações em escala nanométrica e macroscópica com técnicas de Machine Learning (ML) para construir uma estrutura multiescalar consistente. O fluxo de trabalho consiste em três ramos principais, cada um dedicado a aspectos específicos do estudo. O primeiro ramo foca no desenvolvimento de dados em escala nanométrica, particularmente através de simulações de adsorção de CO_2 utilizando métodos de Monte Carlo no software RASPA. Essas simulações geram saídas em nível molecular que servem como entradas para aplicações de ML. Os procedimentos incluem a definição de arquivos de entrada para simulações, como a estrutura de adsorventes e adsorvatos, campos de força e regras de mistura,

garantindo precisão na representação das interações de van der Waals e parâmetros potenciais. As simulações são executadas por meio de scripts de shell em um ambiente de desenvolvimento integrado (IDE), neste caso, o Visual Studio Code. O segundo ramo aborda a modelagem macroscópica, que captura o comportamento em maior escala do sistema sob condições termodinâmicas semelhantes. As simulações em escala nanométrica e macroscópica estão interconectadas por restrições consistentes, particularmente temperatura e pressão, que mantêm a correspondência física entre as escalas. Embora os fenômenos ocorram em escalas de tempo muito diferentes—nanosegundos na escala nanométrica e horas na escala macroscópica—essas restrições comuns garantem que os dados sejam coerentes e estruturados para integração em um conjunto de dados unificado. Uma parte essencial deste trabalho é o contraste entre o método de Monte Carlo (GCMC) e modelos de isoterma, que descrevem a adsorção de CO_2 a partir de diferentes perspectivas — estocástica na escala nanométrica e determinística na escala macroscópica. Embora o modelo de isoterma de Langmuir tenha sido inicialmente considerado, ele foi excluído devido à sua incapacidade de descrever adequadamente as interações em nível nanométrico. Em vez disso, os modelos de Freundlich e SIPS foram empregados para ajustar os dados do GCMC. Essa abordagem evita inconsistências entre as escalas, o que poderia levar a imprecisões termodinâmicas. O terceiro ramo da metodologia detalha o uso de modelos de ML, especificamente redes neurais e algoritmos de Random Forest. Esses modelos são treinados com o conjunto de dados estruturado que resulta nas saídas das simulações em escala nanométrica (entradas) e simulações em escala macroscópica (objetivos). Essa abordagem multiescalar, baseada em dados, garante que as conexões entre fenômenos físicos em diferentes escalas sejam capturadas eficazmente. No geral, a metodologia estabelece uma estrutura onde propriedades intensivas, como temperatura e pressão, servem como variáveis-chave que conectam os domínios nanométrico e macroscópico. Essa integração facilita um modelo multiescalar confiável e consistente que impulsiona a capacidade preditiva a partir da computação intensiva.

Resultados e Discussão

Simulações de GCMC foram realizadas e comparadas com dados da literatura. Todas as simulações desenvolvidas neste estudo foram para um sistema de adsorção de um único componente (CO_2). A precisão das simulações foi verificada ao comparar os dados obtidos com as propriedades de equilíbrio termodinâmico e o modelo determinístico da curva de ruptura do leito. As simulações foram realizadas nas mesmas condições (temperatura e pressão) dos estudos de referência. As isotermas obtidos por experimentos *in silico* foram comparados utilizando erro quadrático médio (MSE), raiz do erro quadrático médio (RMSE), e raiz do erro quadrático médio relativa (RRMSE), junto do coeficiente de determinação (R^2). Os indicadores de desempenho mostraram uma boa concordância entre a experimentação *in silico* e os dados de referência, com um MSE médio de 1.0955 mol/kg e um RMSE médio de 0.8588 mol/kg, equivalente a 6.633% de erro relativo. O R^2 medio geral foi calculado como 0.994, indicando um bom desempenho geral das simulações. No entanto, foi observada uma divergência na simulação para ITQ-29 a partir de 2 bar de pressão, devido ao ajuste do sistema para o termo de fugacidade na equação de Peng-Robinson. Embora o ZIF-8 tenha mostrado uma boa concordância com os dados de referência a altas pressões, foi identificado que os campos de força representam uma barreira significativa para uma

boa correspondência entre dados experimentais e simulados. A qualidade das simulações foi avaliada pelo desempenho do SWAP, com uma diferença média de 0.0785% entre exclusões e inserções, confirmando a precisão das simulações. Para realizar as simulações em macroscala, o trabalho de Sabouni e colaboradores foi reproduzido e validado. O sistema físico em macroscala, conectado aos modelos de aprendizado de máquina com a escala nanométrica, é descrito por este trabalho de referência, essencial para o projeto atual. Os modelos de aprendizado de máquina foram treinados para prever alvos específicos para esse sistema físico. A modelagem determinística da BKC de adsorção de CO_2 foi desenvolvido em MATLAB. A simulação em macroscala foi avaliada com indicadores de desempenho estatístico semelhantes aos usados na escala nanométrica. Os resultados foram comparados com dados experimentais de referências da literature. Os resultados mostraram uma boa representação dos dados experimentais, com um RMSE de 0.0565 e um MSE menor que 0.0032 para carga de adsorvato. O RRMSE de 1.260% e o R^2 de 0.9864 confirmaram a concordância. A utilização de grupos adimensionais permitiu uma melhor descrição do sistema em macroscala, reduzindo a complexidade e evitando problemas rígidos, como em casos de alta pressão. A qualidade dos algoritmos de regressão de aprendizado de máquina foi impactada pela estrutura dos conjuntos de dados, sendo essencial que sejam bem estruturados para evitar vieses e complexidades adicionais. O procedimento de suavizacão é um passo principal para que cada característica dos isotermas dos materiais ITQ-29, IRMOF-1 e ZIF-8 seja transformada numericamente, facilitando o aprendizado de maquina. O ajuste de dados foi realizado em grande maioria o logarítico natural. Este procedimento de regressão apresentou boa concordância com os dados computacionais, com um coeficiente de determinação médio de 0.9922 e um desvio padrão de 0.00728. O coeficiente de variação é de 0.73%, indicativos de um bom ajuste. A melhoria dos modelos de RF e ANN seguiu os mesmos equivalentes: variação dos hiperparâmetros para encontrar o melhor modelo e conjunto de treinamento. Inicialmente, os resultados do RF foram obtidos com base na busca aleatória de hiperparâmetros, e a otimização da arquitetura MLP foi realizada analisando a função de ativação, o número de épocas e o tamanho do lote. Ambos os algoritmos foram aprimorados com base no tempo estequiométrico, aplicando hiperparâmetros ótimos para o tempo de saturação e o tempo de quebra diretamente. Para o TC, com um MAE de 0.00296 e um RMSE de 0.00357, o modelo se ajusta bem aos dados de treinamento. No entanto, R^2 e MSE indicam overfitting, com valores de 0.99999 e 0.00001, respectivamente. Ao aplicar o modelo no conjunto de teste, a adequação é boa, sem indicar viés de overfitting. Indicadores de desempenho mostram variações leves para MAE (0.06087) e RMSE (0.12188), e MSE e R^2 confirmam um excelente ajuste. A análise dos valores SHAP indica que S_{BET}, pressão e V_{pore}/H-A Coulomb são as principais variáveis para o modelo RF de TC, com valores médios SHAP de 3.4, 0.65 e 0.45, respectivamente, sugerindo que o modelo não utiliza todo o conjunto de dados para um ajuste capacitado. As Figuras 22 e 23 ilustram essa análise e destacam a importância de um uso mais equilibrado das distribuições de dados para melhorar o ajuste para o conjunto de teste. Para desenvolver o melhor modelo de rede neural artificial (ANN), o estudo focou na combinação ideal entre função de ativação, tamanho do lote e número de épocas, utilizando uma arquitetura de 7 camadas com 15x32x64x32x8x8x1 neurônios. A função ReLu se destacou, superando a sigmoidal e apresentando melhor desempenho. Testes mostraram que tamanhos de lote menores e mais épocas melhoram a performance, mas o melhor conjunto foi 10/90 para ReLu. O modelo ANN alcançou um MSE médio de 0.0062 e RMSE de 0.0541, com R^2 médio de

0.9993, evidenciando um ajuste adequado. A análise SHAP revelou que variáveis como S_{BET} e *Enthalpy of Adsorption* são cruciais para a previsão, intengrando informação fisica relevante da ramificação da nanoescalado junto da abordagem macroescala por intermedio da abordagem *Big data*.

Considerações Finais

Os resultados do estudo são promissores, mas precisam de uma análise crítica, especialmente em relação à adsorção de CO_2 em sistemas multicomponentes, que envolve interações complexas com componentes como O_2 e água. Essa complexidade exige uma reavaliação de todas as abordagens anteriores, particularmente no modelamento macroescala, onde as isotermas de difusão e absorção precisam ser revistas para sistemas multicomponentes, o que afetará modelos como o de leito fixo. A incorporação de campos de força com modelos de aprendizado de máquina pode aprimorar o modelamento molecular e potencialmente mudar a aplicação de aprendizado de máquina em modelagem multiescalar. A análise contínua e rigorosa das características junto de *outputs* é essencial, pois pode impactar significativamente o *framework* desenvolvido. No caso das redes neurais artificiais (ANNs), explorar diferentes arquiteturas e hiperparâmetros pode, potencialmente, melhorar o desempenho e a interpretabilidade da integração. O estudo mostrou que as ANNs superaram os modelos de floresta aleatória (RF) na previsão de indicadores como TBK, TC e TS. A integração de simulações em nanoescala com modelagem em macroescala através de aprendizado de máquina oferece uma abordagem robusta para análise e previsão da adsorção de CO_2, mas ainda passiva de melhorias.

Palavras-chave:Adsorção de CO_2. Multi-escala. Simulação Molecular. Algoritmos de Aprendizagem de Máquina. Curva de ruptura.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADAM	Adaptive moment estimation
ADCP	Average derivative of the chemical potential
ANN	Artificial Neural Network
BD	Database
BKC	Breakthrough Curve
CCUS	Carbon Capture, Utilization, and Storage
DT	Decision Tree
FF	Force Field
GCC	GNU compiler collection (C/C++ libraries)
GCMC	Grand canonical Monte Carlo
GHG	Greenhouse gas or Greenhouse gases
ICC	Intel C/C++ Compilers
IPCC	Intergovernmental Panel on Climate Change
LDF	Linear Driving Force
LJ	Leonard Jhones
MAE	Mean absolute error
MC	Monte Carlo
MD	Molecular Dynamics
ML	Machine Learning
MLA	Machine Learning Algorithms
MLP	Multilayer Perceptron
MOF	Metal organic framework
MS	Molecular Simulation
MSE	Mean Squared error
PSA	Pressure Swing Adsorption
ReLu	Rectified linear activation function
RF	Random Forest
RMSE	Root mean squared error
RRMSE	Relative root mean squared error
R^2	Determination coefficient
SMLA	Supervised Machine Learning Algorithms
SVM	Support Vector Machine
TBK	Breakthrough Time or Time of Breakthrough
TC	Stoichiometric Time
TRL	Technology Readiness Level
TS	Saturation Time
TSA	Temperature Swing Adsorption

UAT	Universal approximation theorem
UFF	Universal force field
XAI	Explainable artificial intelligence approach
ZIF	Zeolitic Imidazolate Framework

LIST OF SYMBOLS

A	General adsorbate
M	General adsorbant
AM	Adsorbant - Adsorbate complex
K	Constant of Equilíbrium
q	Moles of adsorbate adsorbed per mol of adsorbant
M_a	Molecular mass of "A"
M_m	Molecular mass of "M"
n	Number of open sites in an adsorbant
Q	Mass fraction of "A" adsorbed on "M"
Q_l	Langmuir equation coeficient - Maximum amount of "A" adsorbed in "M"
K_f	Freundlich equation equilibrium term
n_f	Freundlich equation exponential term
n_s	SIPS equation exponential term
K_s	SIPS equation equilibrium term
Q_s	SIPS equation coefficient
t	time in min
κ_c	Linear driving force coefficient
q_t	The equilibrium concentration
k_f	Film layer diffusion [m/s]
L_o	Characteristic length of the bed
r_p	Particle radius
u	Interstitial velocity
D_m	Molecular self-diffusion coefficient in m^2/s
C	Concentration of the adsorbate in mol/m^3
C_o	Initial concentration of the adsorbate in mol/m^3
D_x	Axial mass diffusion coefficient in m^2/s
γ	Coefficient laaa
β	Coefficient baaaa
d_p	Particle pore diameter in m
R	Universal gas constant $J/mol.K$
M_m	Molar mass or molar weight of a singular component in mol/kg
Ω	Average molecular velocity in m/s
λ	Free mean molecular path in m
$P_{e,\infty}$	The limitation of the Peclet number for the adsorbent particles
ε	Porosity of the framework
α	Coefficient caaa
ρ_p	Bed density in kg/m^3

$\partial q_i \partial z$	Rate of equilibrium concentration in the sorption zone in $mol/kg.s$
ε_o	The dispersion energy coefficient
r	Distance between particles or particles size
Z	The charge of a given molecule
p	The distances where the potential energy is zero
E_{index}	Potential energy associated to index
E_{vdw}	Van der Waals potential energy
r_f	Tha cut-off radius in A

CONTENTS

1 INTRODUCTION

Among all the new century challenges, one has continuously gained attention: the greenhouse effect and environmental and weather stability. Greenhouse gas (GHG) emissions are addressed as one of the driving forces towards Global warming. Anthropogenic GHG emissions have been increasing consistently in the last decade (YORO; DARAMOLA, 2020; GABRIELLI; GAZZANI; MAZZOTTI, 2020), hence the scenario that humankind faces has gained urgency characteristics. The principal component of this context, in terms of mass emissions, is carbon dioxide (CO_2). The Intergovernmental Panel on Climate Change (IPCC) is constantly related as a reference, while its content indicates different scenarios that should be achieved or avoided. Even being criticized and reviewed consistently (BURGESS et al., 2020a), the data present a continuous increase in the concentration of CO_2 reaching values of 420 ppm (BURGESS et al., 2020b). The consequences of that rate imply several changes in our way of life, resources supply and demand (PUGNAIRE et al., 2019).

To solve that growing urgent problem, Carbon Capture, Utilization, and Storage (CCUS) technologies are gaining more attention, both on the academic and the industrial fronts. At the beginning of that value chain, carbon capture has three main fronts that should be addressed (GABRIELLI; GAZZANI; MAZZOTTI, 2020): 1) Rate of capture, i.e. basically the amount of CO_2 uptake from the fonts of emissions and atmosphere; 2) time and storage, meaning the amount of CO_2 trapped until the usage of one pitfall; and 3) scale, since the growth of sources of CO_2 is higher than the growth of CCUS technologies. Those factors imply an incentive on the way and the velocity at which engineers develop technologies. In the context of CO_2, addressing capturing technologies that are still in need of improvement at their Technology Readiness Level (TRL) requires innovative methods to accelerate their development. That is the case of CO_2 adsorption (OLECHOWSKI; EPPINGER; JOGLEKAR, 2015), assigned at the demonstration phase, with TLR addressed as level 7, specifically for Post-Combustion Adsorption, and Direct Air capture technologies. That leveling indicates that scalability is still a challenge for the adsorption technologies, with several aspects to be enhanced (i.e. cost-minimization, operational efficiency, control, and operation). One of the key aspects thought, still relies on materials screening and effectiveness towards scalability (BUI et al., 2018)

Gas-solid adsorption is grounded on interaction forces that will trap a gaseous molecule in a framework. In a multicomponent adsorption system, molecules would compete for adsorption sites, tending to thermodynamic equilibrium (ZHAO et al., 2021). Several ways to enhance the technology have been presented in the last years, espe-

cially considering the so-called *in silico* methods, where experimentation is performed in a virtual system (MIRZAEI, 2020). The attention of this work follows the molecular simulation technique where in a confined system the surface phenomena are designed and modeled. The approach allows one to fundamentally measure inner properties, evaluate heterogeneous atoms in the constitution of a new framework, analyze the impact of pore size and structure fails, gradients, and so on (ZHOU, W. et al., 2019; CASTILLO, 2009; KWON et al., 2017, 2017), throughout visual and intuitive interfaces, as iRASPA and RASPA 2 software (DUBBELDAM et al., 2016).

Even more recently, another insightful and agile tool has gained attention. Machine Learning Algorithms have been presented as an opportunity to improve insightfully chemical engineering technologies in general. Regarding CO_2 adsorption, a diverse set of applications has already been displayed by academia, from operational systems, e.g., geological injection (STURLUSON et al., 2019; HU et al., 2019) and Pressure Swing Adsorption (PSA) (AN et al., 2019; GU, C. et al., 2019), to materials screening (YAMADA et al., 2019; AGHAJI et al., 2016; FERNANDEZ; BARNARD, 2016) and synergistic interactions (ZHAO et al., 2021).

Machine Learning is assigned to the data-driven engineering field (MONTÁNS et al., 2019). The molecular simulation method, though, is a piece of a bigger picture with its pros and cons (GE et al., 2019). Suppose one aims to perform detailed simulations connecting scales to enhance a system's predictability throughout the modeling. In that case, the system's degrees of freedom increase as the detailing (more minor scales) is considered. Therefore, greater detail might leads to higher and makes the trade-off between informational and precision more challenging. Connecting scales is a challenge in chemical engineering that generally finds itself in CO_2 adsorption and storage (AFAGWU et al., 2021; LE et al., 2020). If overcome, significant improvement can be made regarding problems such as GHG emissions, carbon capture technology improvement, adsorption efficiency, etc. Enlighted by the present ideas and concepts, the primary motivation of the current work is to tackle multiscale modeling by the innovative techniques mentioned and provide a glimpse of the collaborative application of MLA and MS within the field of Chemical Engineering, more precisely, CO_2 adsorption.

Finally, to improve the technological face of CO_2 adsorption operation, the current workflow focuses on developing a data-driven method that forecasts multiscale performance through the integration of Machine Learning ML and Molecular Simulation MS.

1.1 STRUCTURE OF THIS DOCUMENT

It is essential to convey the work's structure: focusing on multiscale modeling for CO_2 adsorption. Methodology details result application and implication. For clarity, readers are encouraged to grasp theoretical fundamentals from bottom-up, understanding nano-scale interactions and their macro consequences. Results presented step-by-step, rooted in methodology and theoretical background. Methodology grounded in three main contents.

- Nanoscale *in silico* experimentation, performed by molecular simulation.
- Macro-scale modeling, employed by classic deterministic modeling.
- The application of Machine Learning models building a connection with nanoscales and macroscales.

Therefore, some questions are presented that not just anchor the present work but are also dedicated to displaying an evident comprehension to the reader about the principles behind it and its physical validation.

The questions that may orient the present study's development are related to the mathematical unfolding of the modeling of the same phenomena at different scales. As the scale of the phenomena is closer and closer to the non-continuum domain, the interactions' degrees of complexity increase significantly. A critical notion can be used to elucidate its complexity. The degrees of freedom of a molecular system can be close to the unit of thousands. At the same time, a macro-scale model will be stated in the decimals unit when too complex. How can two distant scales with different complexity degrees be connected throughout the same mathematical model?

A second question that might instigate the reader toward the present study is related to the universal approximation principle for neural networks, the Monte Carlo algorithm, and the deterministic modeling of a physiochemical system. By which means do those three mathematical approaches find each other and bring to light a direct connection between scales that are separated by time and space measurements so discrepant?

From that second question, one more can be formulated. Despite the scale, the phenomena modeled will follow the dimensions of time regarding its interactions. While in the nanoscale the time frame of nanoseconds is passive to frame the interaction over there, at the macro scale, the unit of minutes, hours, and sometimes days is pertinent for adsorption. To be more precise, a calculus can be done. For example, for every

hour for the operation of a fixed bed adsorption system, 3.6×10^{15} picoseconds are computed. How, without a high computational cost, can nanoscale attributes be directly associated with a macro scale system directly and intuitively?

The questions made above are the foundations of the present work. The objective of the following document is to ground and solve those questions. The core of the present work, as presented, has three main niches of work, and its final deployment follows the direct integration of those by a data-driven method. Therefore, the connections between contents are represented by simple and didactic apparatus throughout the sections.

Furthermore, the document's structure comprehends the context, motivation, and objectives, addressed in the former section. Then, in sequence, the computational tools used are elucidated, establishing the theoretical background of the protocol developed. In the third chapter, the reader will be presented with the methodology in general and specific terms. In chapter four, the top results related to the presented procedures will be explained and analyzed in two sections, the first dedicated to a general validation of the methodology, and the second addressing specific applications. The final chapter will present the main conclusions of the study developed.

1.2 OBJECTIVES

1.2.1 General objective

To integrate Machine Learning Algorithms with Molecular Simulations to enhance the accuracy and efficiency of CO_2 adsorption studies.

1.2.2 Specific objectives

1. Model through molecular simulations the CO_2 adsorption in different materials (Zeolites and MOFs), and a fixed bed adsorption system of those same materials by deterministic models while validating those molecular simulations with experimental data from the literature.
2. Develop a methodology capable of integrating data from different scales regarding CO_2 adsorption in a single dataset.
3. Implement Supervised Machine Learning Algorithms trained on data generated by molecular simulations towards macroscale simulations performance indicators and uncover their intelligence toward scientific insights regarding CO_2 adsorption field.

2 THEORETICAL BACKGROUND

To sustain the methodology here developed and evaluate the results right away, the theoretical background is composed of three main parts. The first introduces the reader to the fundamentals of CO_2 adsorption, explaining it from the perspective of the equilibrium and dynamics of the system. The second part is dedicated to molecular simulation, where its basics are grounded. Lastly, Machine Learning Algorithms are established, giving attention to Supervised Machine Learning Algorithms (SMLA).

This approach also sustains how the methodology was developed, although from a clear and cohesive perspective. All the details that orchestrate a linear workflow are detailed in the methodology section. Therefore, since the procedure develops a strategy for multiscale modeling of CO_2 adsorption throughout the integration of deterministic modeling, Molecular Simulation (MS), and Machine Learning Algorithms (MLA), the objective of this section is to present not just the fundamental concepts that sustain the physical bases of the methodology developed, but also a mathematical comprehension and intuitive endeavor of the present work.

2.1 CO_2 ADSORPTION: FUNDAMENTALS AND MODELING

Before delving into the concepts of adsorption, it's crucial to understand why CO_2 adsorption is a relevant process for addressing climate change. Industrial activities alone contribute approximately 30% of the U.S. primary energy-related CO_2 emissions, amounting to close to 1.36 gigatons in 2020 (BEASLEY; O'KEEFE; RODGERS, 2023a). This statistic underscores the urgent need for effective strategies to mitigate industrial emissions and tackle climate change. Adsorption presents itself as a promising approach due to its modular features, allowing for scalability and adaptability (BEASLEY; O'KEEFE; RODGERS, 2023b, 2023c). As a process at the forefront of a chain (where carbon must first be captured for processing), the effectiveness of CO_2 adsorption addresses not only environmental concerns but also industrial challenges, offering potential for a range of further applications (BEASLEY; O'KEEFE; RODGERS, 2023d, 2023e, 2023f).

Despite being promising and reliable, CO_2 adsorption still faces several challenges, including high selectivity and the renewability of adsorbents (BEASLEY; O'KEEFE; RODGERS, 2023g). CO_2 competes with CH_4 for adsorption, making selectivity complex in multi-component adsorption scenarios. Additionally, the presence of humidity and other components further complicates selectivity (KOLLE; FAYAZ; SAYARI, 2021). Over time, humidity affects the adsorbent's capacity, with water clustering at high con-

centrations and competing with methane and carbon dioxide at medium percentages (BEASLEY; O'KEEFE; RODGERS, 2023g; BAHAMON; VEGA, 2016). Enhancing the adsorption approach requires leveraging fundamentals to address these challenges effectively.

Adsorption generally relies on the thermodynamic equilibrium between an adsorbate (the substance or molecules that are being attracted and adhere to a surface), present in the bulk of the system, and its concentration at the surface of an adsorbent, the contact framework. Adsorption is characterized as a superficial phenomenon (PULLUMBI; BRANDANI, F.; BRANDANI, S., 2019; DĄBROWSKI, 2001). From the practical point of view, its final results are the measured capacity of a material to attach molecules at its surface at determinate conditions (CASTILLO, 2009). The interaction between the process agents determines the equilibrium concentration, the central aspect of adsorption evaluation. Several properties concerning adsorbate and adsorbent will determine the thermodynamic equilibrium (presence of ions, superficial area, open sites, etc.), as well as the nature of the interaction: physical (physisorption) or chemical (chemisorption) (DĄBROWSKI, 2001).

In the following sections, adsorption focused on CO_2 will be deepened, relying upon its equilibrium and dynamic modeling. The text of this section is divided into two main topics: equilibrium and dynamics. The first conceives the core ideas behind the nanoscale interactions, leading to the adsorbate/adsorbent complex thermodynamic equilibrium. The second topic defines the kinetics of the adsorption process in a fixed-bed system from the perspective of the interactions from small to bigger scales, supported by mechanistic models. That is the first part of the theoretical background, and the reader can then regard a deterministic model through intermolecular fundamentals.

2.1.1 Equilibrium

The nature of the CO_2 adsorption process relies on the equilibrium between forces of attraction within a system composed of a framework and the gaseous bulk molecules (DĄBROWSKI, 2001). The manipulation of temperature and pressure will imply a new state of thermodynamic equilibrium, where those forces of attraction will maintain an amount of CO_2 on the surface of the framework (DĄBROWSKI, 2001; CASTILLO, 2009). The thermodynamic equilibrium is determined when a concentration of equilibrium in the bulk and on the surface is established (WANG, J.; GUO, 2020). That is the core measurement of a CO_2 adsorption isotherm or general adsorption process.

Before stepping into the isotherms modeling, one should consider what causes the surface phenomena and its equilibrium. From a primary standpoint, the forces of attraction between adsorbate and adsorbent determine the degree of attraction of one molecule to a framework (DĄBROWSKI, 2001). That degree of attraction will lead to the concentration of equilibrium. To explain further, the force field on a framework's surface interacts with a molecule structure, reducing the potential energy of that free molecule toward its stabilization on the framework surface (CASTILLO, 2009; VLUGT et al., 2009; DUBBELDAM et al., 2016).

Adsorption is a process where a molecule's potential energy is reduced, leading to its transition into a new state known as the adsorbed state. Thermodynamically, this is associated with the transfer of energy from the adsorbed molecule to the system, which gives the fundamental explanation for the exothermic nature of the process (WANG, J.; GUO, 2020; DĄBROWSKI, 2001). That nanoscale aspect is fundamental to the adsorption equilibrium state (HOLLINGSWORTH, Scott A; DROR, Ron O, 2018a). Considering the interaction between adsorption main agents, the molecule of CO_2 has a quadrupole moment, making it easier to be adsorbed within a nonpolar adsorbent (CHEN, Cong et al., 2020), where van der Waals forces are dominant (ZHOU, W. et al., 2019). Polar surfaces will eventually create obstacles to an efficient adsorption process, especially in the presence of competitive adsorption systems (CHEN, Cong et al., 2020). Therefore, the forces associated with the composition of the framework will be significant for the final adsorption equilibrium. Notwithstanding, geometrical and textural properties will be as sensitive to the process as molecular composition (ANDERSON et al., 2018).

Regardless of the number of features that will influence the process equilibrium, the typical approach to determine the equilibrium profile between adsorbate and adsorbent is throughout an isotherm. For a gaseous system, the pressure variation will affect the equilibrium concentration, which can be described by a mechanistic model of equilibria, e.g., Langmuir isotherm, Freundlich isotherm, SIPS isotherm, and several others (WANG, J.; GUO, 2020). The adsorption process can be described as having a multiscale nature, as a significant determinant of macroscale (mechanistic) models is the interactions within non-continuous spectra, particularly force field interactions.

Isotherm data can be pursued by *in situ* approaches (laboratory experimentation) or *in silico* approaches (computational modeling) (HUANG, H. et al., 2011). The Langmuir isotherm is the most classic model for representing adsorption systems, gaseous or liquid (DĄBROWSKI, 2001; GHAEDI, 2021). The deduction of the model will be described, as well as its hypothesis. In the following, the Freundlich and SIPS model

will be explained in general terms to finally be interpreted in its physicochemical terms. The Langmuir model has the following hypothesis:

1. The adsorbent “M” has identical sites to each adsorbate molecule.
2. Every site has a binding process identical to each molecule; hence, the adsorption energy is equal to every site.
3. The gaseous adsorbate gas has an ideal behavior.
4. Once a gaseous molecule is adsorbed, it stays constant.
5. The adsorption is monolayered, meaning there is no second layer or interaction between adsorbate-adsorbate.

Considering that one has an adsorbate “A” and an adsorbent “M” with “n” open sites.

$$A + M \rightarrow AM \tag{1}$$

The binding of “A” with “M” generates the complex “AM”. “K” is the constant of equilibrium, which is determined as the following. “q” stands for the number of moles of adsorbate attached to the framework per mol of “M” and is calculated by equation (3).

$$K = \frac{[AM]}{[A][M]} \tag{2}$$

$$q = \frac{[AM]}{[M] + [AM]} \tag{3}$$

$$q = \frac{KA}{1 + KA} \tag{4}$$

The substitution of Equation (3) in Equation (4) can be made, giving Equation (5):

$$\frac{M_a}{M_m} q = n \frac{M_a}{M_m} \frac{KA}{(1 + KA)} \tag{5}$$

By the consideration that every site has one single bound, quantitatively, the isotherm is representative of the sum of the “n” sites. Multiplying the equation for the ratio of the molar mass of “A” by the molar mass of “M” Equation (6) is given:

$$Q = Q_l \frac{KA}{1 + KA} \tag{6}$$

The Freundlich equation is an empirical model that relates an adsorbed gas's mass ratio to an adsorbent's mass over the system's pressure. Equation (7) presents its model. Differently from Langmuir, the Freundlich isotherm has no mathematical background (WANG, J.; GUO, 2020). The Freundlich equation, due to its mathematical features, at higher pressure, may fail at describing the adsorption process (WANG, J.; GUO, 2020; GHAEDI, 2021). Therefore, the fitting can describe the saturation pressure of a system improperly.

$$Q = K_f [A]^{1/n_f} \tag{7}$$

This model has a non-single site consideration; hence, it can describe more complex systems even being empirical (e.g., heterogeneous surfaces, multilayered systems). In the face of the Langmuir model, one can infer from the Freundlich equation several characteristics related to the adsorption mechanism, the nature of the adsorbent surface, if monolayer or multilayer adsorption, and so on (WANG, J.; GUO, 2020; GHAEDI, 2021).

The physical meaning of the coefficients is assigned to, first, K_f as the partitioning coefficient or the adsorption affinity, and second, as n_f, assigned as the Freundlich constants characteristics of the system, an indication of the adsorption heterogeneity of the adsorbed-adsorbent system (DEMESSIE; SORIAL; SAHLE-DEMESSIE, 2022; WANG, J.; GUO, 2020).

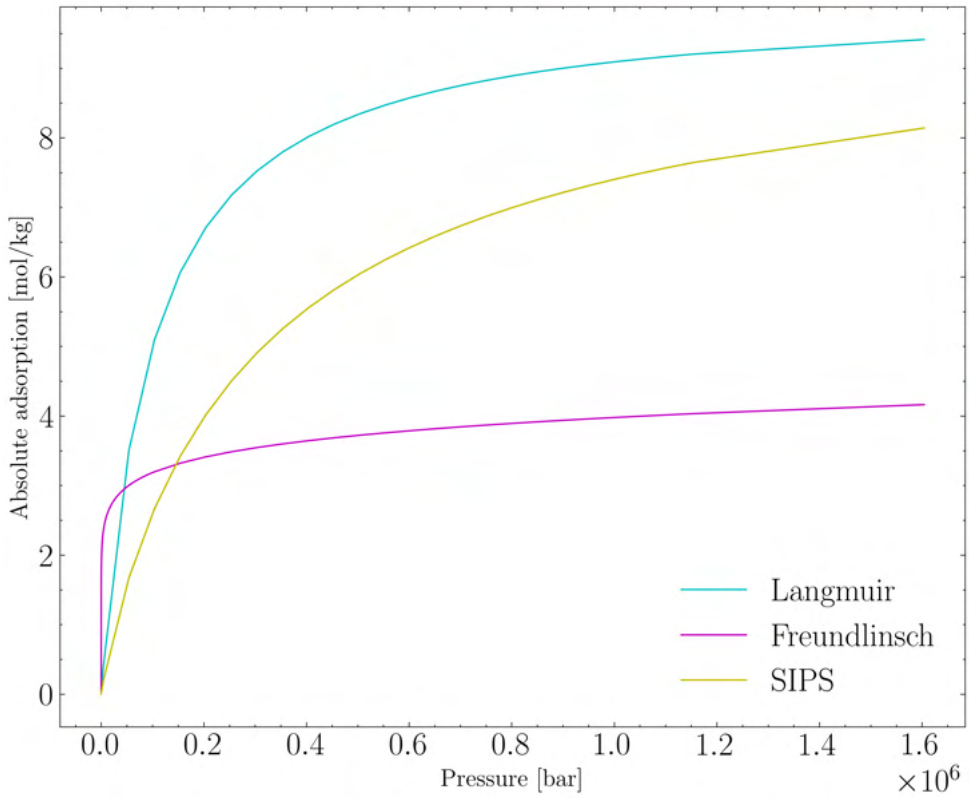
The SIPS model is the latest isotherm model of focus. Unlike the previous models, it features three adjustable parameters. Combining the Langmuir and Freundlich models, it incorporates the n_s exponent, similar to the Freundlich isotherm (THOMAS; CRITTENDEN, 1998b). This additional term (Q_s) represents a mathematical flexibility for the model, allowing the exponent parameter to compensate for the K_s parameter, the system's equilibrium constant. Q_s, measured in $mg.g^{-1}$ stands for the maximum

adsorbed amount. As a result, the SIPS model can accurately describe heterogeneous surfaces and complex interactions, overcoming limitations of previous models, such as Freundlich's model at high pressures (YANG, R. T., 1997; THOMAS; CRITTENDEN, 1998b). The following equation describes the SIPS isotherm model.

$$Q = \frac{Q_S K_S [A]^{1/n_s}}{1 + K_S [A]^{1/n_s}} \tag{8}$$

The major advantage of the above equation is its adaptability for the adsorbate concentration in the system, since it follows the Freundlich model at lower concentrations and, in the opposite context, the Langmuir model (MURPHY et al., 2023). Figure 1 indicates the profile of the models described above, and the adaptability of the SIPS equation.

Figure 1 – Langmuir, Freundlich, and SIPS profiles



Source: Author (2024). Note: SIPS parameters allow this equation to settle between Langmuir and Freundlich models since the SIPS model regards a combination of both.

With the apparatus given, one can evaluate data gathered from *in situ* measurements or, as will be further explored, by *in silico* methods. Regarding the theme of the

present work, the equilibrium between adsorbent and adsorbate is mainly driven by physisorption (THOMAS; CRITTENDEN, 1998c). Hence, temperature and pressure have a more sensitive impact on the phenomena equilibrium and its dynamics. It is important to emphasize that, for single-component and multicomponent systems, the increase in temperature implies a lower adsorption efficiency (WANG, J.; GUO, 2020) as a consequence of the exothermic adsorption process (THOMAS; CRITTENDEN, 1998b). Reversibility is accessible by the increase of temperature or the reduction of the pressure (RUTHVEN, Douglas Morris; FAROOQ, Shamsuzzaman; KNAEBEL, 1994) of the system, characterizing Pressure Swing Adsorption (PSA) and Temperature Swing Adsorption (TSA) operations, respectively (GREEN, 2007). By manipulating those variables, one restores the previous equilibrium at the initial condition, where CO_2 is detached from the framework. The following section will unfold the core of adsorption dynamics modeling.

2.1.2 Dynamics

The transition between thermodynamic states of equilibrium can embrace the dynamic of the adsorption system (YANG, R. T., 1997; LETCHER; MYERS, Alan L, 2004). To grasp the dynamics of adsorption, PSA and TSA operations can be used once conducted on a fixed-bad system, where the concentration versus time curve describes the phenomena occurring. This plot is the Breakthrough curve, a practical basis for assessing the behavior of an absorbent in a fixed-bed adsorption system (MYERS, A., 2002). The breakthrough curve (BKC) depends on the bed geometry, diffusion and transport properties, operational conditions, and, as important as those presented, the adsorption isotherm of the material present in the bed (THOMAS; CRITTENDEN, 1998d).

The adsorption equilibrium will be reached once passed through a dynamic adsorption system. What regulates this part of the process is the adsorbate's diffusion rate, evaluated as the mass transference gradient (LETCHER; MYERS, Alan L, 2004). Hence, the system's driving force is the concentration gradient, which is determined by the difference between the equilibrium (isotherm) and the system's present state (THOMAS; CRITTENDEN, 1998b). In simple terms, there will be diffusion of the component "i" (adsorbate) while there is a gradient.

The mass transference gradient in a dynamic adsorption system is relative to the degree of interaction between framework and adsorbate over time, and the inner properties of those (polar sites, quadrupole moment, heterogeneous surfaces, framework porosity, and similar (THOMAS; CRITTENDEN, 1998a; ZHOU, W. et al., 2019). These

properties lead the initial state of a single molecule within the operation to a final state of equilibrium attached to the adsorbate. With a rate of adsorption taking place, the overall coverage of the adsorbent surface by the adsorbate is now controlled by bulk properties, affecting the energy change, flux, and distribution velocity of the gaseous molecules being adsorbed (LETCHER; MYERS, Alan L, 2004; GHAEDI, 2021). From these conceptions, adsorption system dynamics can be, then, assessed by a phenomenological approach. The phenomenology of the system, directly designed by the convergence of the chemical potential of the adsorbed phase and gas phase, allows one to model the operation in a fixed bed apparatus mechanistically. Operational conditions can be evaluated once the model is stated (LE et al., 2020; AFAGWU et al., 2021).

Overall, the present work focuses on mass transport phenomenon as primary resistances, although it should be mentioned that heat transfer resistance should be kept in mind for more specific modeling (MAREK, N.; MAREK, S.; JAN, 2022). Ultimately, the mechanistic modeling that considers intrinsic kinetics and transference resistance will describe the transient adsorption process in a fixed bed, leading to the BKC curve shape (SCHILLER; WANG, F., 2018). What has been illustrated so far and the adsorption phenomena are described mathematically by the equations in the subsequent subsections.

However, before stepping into the presentation of the BKC model, the linear driving force concept (LDF) and the dimensionless numbers associated with adsorption will be assigned since those compose the BKC curve model. Diffusion coefficient correlations will also be assigned.

2.1.2.1 The LDF model

The linear driving force is an approach to describe the diffusion of the adsorption system directly to the gradient of concentration of a component (SABOUNI; KAZEMIAN; ROHANI, 2013). It relates the gradient of concentration of the component with the external and internal resistances in a linear model, hence, a linear driving force. This approach is physically consistent and has a solid literature background (NAIDU; MATHEWS, 2021; RUTHVEN, Douglas M., 2003; RAY, 1999).

$$\frac{\partial q_i}{\partial t} = \kappa_c (q_t - q_i) \tag{9}$$

The factor κ_c can be described by several correlations, and different authors present different equations to address it (SABOUNI; KAZEMIAN; ROHANI, 2013; SUN, L. M.; LE QUERÉ; LEVAN, 1996). The term q_t represents the equilibrium concentration. Considering the mass transfer resistances as the main limitation of the kinetics phenomena, one can access the LDF model by considering the external mass resistance being the limitation of the system or the internal mass transfer resistance as the primary resistance, resuming the model to describe, accurately, the fixed-bed adsorption system (NAIDU; MATHEWS, 2021; RUTHVEN, Douglas M., 2003)

The LDF correlation used is supported by the experiments of Sabouni et al. (2013) (SABOUNI; KAZEMIAN; ROHANI, 2013), applying the film mass transfer coefficient directly to the equation. Accordingly, the mass diffusion resistance limitation is associated with the external mass transfer resistance, so the pores diffusion is negligible.

$$\frac{\partial q_i}{\partial t} = \frac{3k_f L_o}{r_p u}(q_t - q_i) \tag{10}$$

In the Equation above, k_f is the film layer diffusion coefficient in m/s, defined in the equation (11); u is the interstitial velocity of the gas in the system in m/s; L_o is the characteristic length of the system in the z axis, referred to the length of the bed.

The film layer diffusion coefficient is defined accordingly to Matsumara et al. (1995) (MATSUMURA; NAYVE JR., 1995) and Sabouni et al. (2013) (SABOUNI; KAZEMIAN; ROHANI, 2013), where D_m is the molecular self-diffusion coefficient in m^2/s, μ_f is the viscosity of the fluid phase in $Pa.s$ and ρ is the density of the fluid in kg/m^3.

$$k_f = 1.09\left(\frac{u}{\varepsilon}\right)\left(\frac{\mu_f}{\rho D_m}\right)^{-\frac{2}{3}}\left(\frac{2d_p u\rho}{\mu}\right)^{-\frac{2}{3}} \tag{11}$$

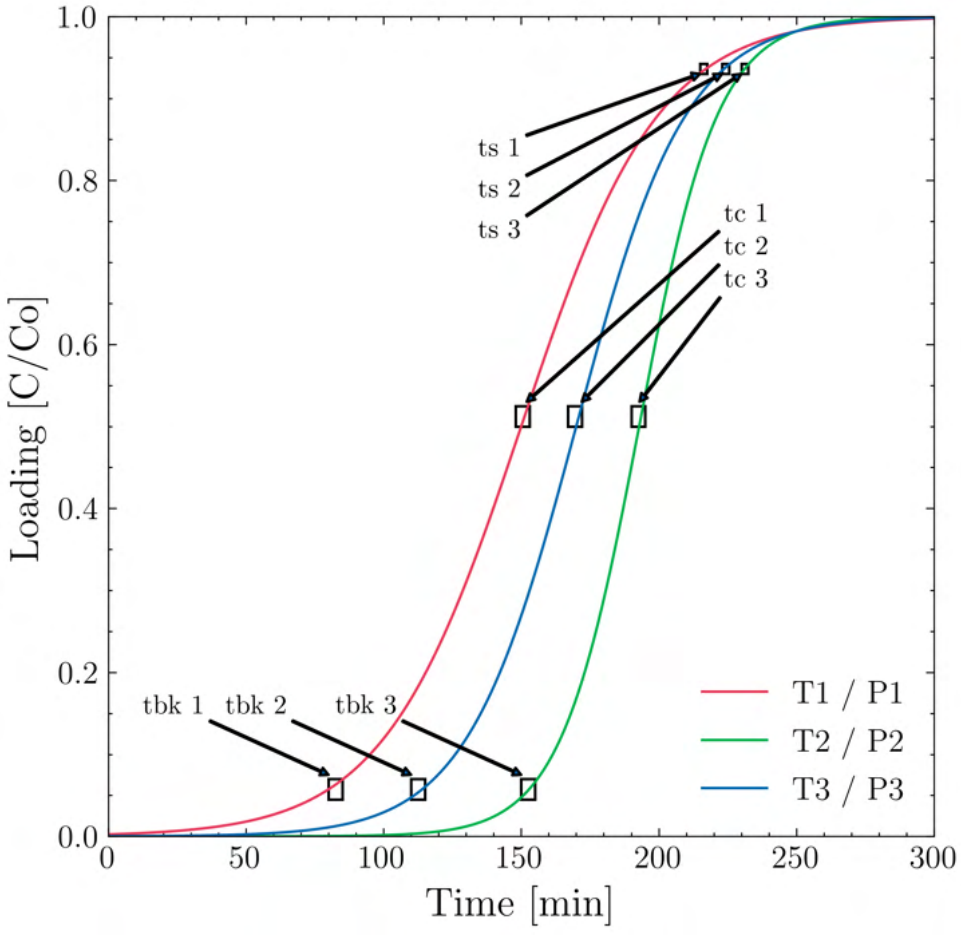
2.1.2.2 The breakthrough curve

The breakthrough curve will describe the performance of a fixed bed column and the column dynamics (THOMAS; CRITTENDEN, 1998d). It is related to several macro physical properties of the system as the flow rate of the adsorbate, its initial concentration, the absorbent's particle size, and the column's length (CHU, 2020; GHAEDI, 2021). Regardless of the scale of those characteristics, the breakthrough curve represents the

adsorbate accumulation within the bed in a time frame. The common approach to plotting the BKC is made with the y-axis representing the ratio of the outlet concentration by the inlet concentration ($\frac{C}{C_o}$) and the x-axis the time profile (t). The plot is presented in figure 2, alongside the main characteristics of a BKC related to time indicators.

As the flow of the inlet passes through the bed, a primary sorption zone develops through a mass transfer zone (KNOX et al., 2016). Considering the inlet entering at the bottom of the column, the first fraction of the adsorbate is being captured by the absorbent, and in the outlet of the column its concentration is near zero (MAREK, N.; MAREK, S.; JAN, 2022). As the mass transfer zone walks through the column and the bed constantly captures more and more carbon dioxide, the amount of carbon dioxide passing through the bed increases as the uptake rate decreases due to the saturation of every fraction of the bed. After an amount of time, the mass transfer zone will be at the edge of the column, indicating the saturation of the bed.

Figure 2 – Breakthrough curve time indicators



Source: Author (2024)

The variation of the column properties will imply different shapes for the BKC, more straight or elongated. However, regardless of the shape of it, a few features will be common to all breakthrough curves. From the above physical description of the breakthrough curve and the column dynamics, two performance indicators associated with the BKC inherently can be addressed: the breakthrough time (TBK), when the first 5% of the amount of carbon dioxide is measured in the outlet of the bed, that moment is defined for the aims of the present work, as the breakthrough time; and the saturation time (TS), when 95% of the C/C_o ratio is reached. Different authors determine the above values differently, although always in the same range (1% < TBK < 5% and 90% < TS < 100% (KNOX et al., 2016; YU, Hui et al., 2015). A third indicator can be addressed when the ratio of the outlet concentration is half of the inlet concentration, the stoichiometric time (TC) (MESFER et al., 2020; SARKAR; AROONWILAS; VEAWAB, 2017, 2017).

Once all physical features of a fixed bed column are constants, the variation of temperature and pressure of the system will have the same effect on the shape of the breakthrough curve as it will have if one changes or varies the physical features of the fixer bed (MAREK, N.; MAREK, S.; JAN, 2022; SABOUNI; KAZEMIAN; ROHANI, 2013). It will happen due to the linear driving force correlation used to define the mass transfer zone of the system (NAIDU; MATHEWS, 2021; RUTHVEN, Douglas M., 2003; RAY, 1999). Finally, changing the temperature and pressure, the mass transfer gradient by the adsorbate's local concentration, and the adsorbate's ideal concentration will inflict on several shapes of the breakthrough curve, each one with its own set of indicators, i.e. TBK, TC, and TS.

To define those performance indicators, the adsorption process can be described mechanistically. From that, dimensionless numbers will be assigned to the system equations. The Peclet number fundamentally analyses the ratio between convective transport, addressed in the numerator at Equation (12), and diffusive transport phenomena, addressed by the mass diffusion coefficient, D_x, in the denominator part of the same equation (SABOUNI; KAZEMIAN; ROHANI, 2013; MATSUMURA; NAYVE JR., 1995).

$$P_e = \frac{L_o u}{D_x} = \frac{\text{convection transport}}{\text{diffusion transport}} \tag{12}$$

From the Equation (12), one can verify that the dimensionless number depends both on the velocity of the bulk – the superficial gas velocity (u) -, and the characteristic length of the system – in this case, length of the bed (L_o) (MAREK, N.; MAREK, S.; JAN,

2022). To calculate the Peclet number, it is necessary to determine, D_x, the axial mass diffusion. To do so, correlations can be used. In the present work, D_x is calculated by the correlation of Edwards and Richardson (1970) (HARKER; BACKHURST; RICHARDSON, 2002). The axial dispersion is coupled to the molecular diffusion and considers Wicke's (1973) approximation for the coefficient γ, and Boshoff's (1969) (HARKER; BACKHURST; RICHARDSON, 2002) expression of coefficient β. The calculus follows in the equation (13) and is referenced in the papers of Nedoma et al. (2022) (MAREK, N.; MAREK, S.; JAN, 2022), and Wilkins et al. 2020 (WILKINS; RAJENDRAN; FAROOQ, S., 2020).

$$D_x = \gamma D_m + \frac{P_{e,\infty}^{-1}(ud_p)}{1 + \frac{(\beta\gamma D_m)}{(ud_p)}} \tag{13}$$

The equation above presents d_p as the pore diameter; R is the universal gas constant and M_m is the molar weight of the single component. D_m stands for the molecular diffusion coefficient for a single component system, defined by Equation (14).

$$D_m = \frac{1}{3}\lambda\Omega \tag{14}$$

Ω referes to the average molecular velocity, in m/s, and λ to the Free mean molecular path in m. It should be mentioned the determination of the $P_{e,\infty}$ for particles with a radius smaller than 0.25 cm -, is determined by Equation (15), accordingly with Langer (1978) (HARKER; BACKHURST; RICHARDSON, 2002).

$$P_{e,\infty} = 6.7 d_p \tag{15}$$

Both coefficients, γ and β are related by Wicke (1973) and Bischoff (1969) (HARKER; BACKHURST; RICHARDSON, 2002), with porosity of the framework ε. The γ coefficient is determined in the following, where one can verify the direct relation with porosity. The β coefficient has a more complex relation, determined with $P_{e,\infty}$ and α, the velocity distribution.

By determining α by Equation (18), one can determine β in Equation (17). The physical meaning of those correlations relies on the approximation of the axial dispersion by the local velocity in the fixed bed, allowing one to correlate velocity with particle diameter, which is stated in the following equations.

$$\gamma = 0.45 + 0.55\varepsilon_p \tag{16}$$

$$\frac{1}{Pe,\infty} = \frac{\beta}{\alpha} \tag{17}$$

$$\alpha = 8.1352\ln(dp) + 24.807 \tag{18}$$

Finally, considering all models, concepts, and equations above, the mechanistic modeling of the BKC can be designed throughout the following assumptions.

- The system is isothermal, adiabatic, and has equal distribution of temperature axially.
- The pressure drop of the column is negligible, as well as momentum effects.
- The adsorption equilibrium isotherm can be described by the Langmuir model, Freundlich model or SIPS model.
- The gas is axially dispersed in the bed, being radially homogeneous regarding concentration.
- The adsorbent particles are spherical and homogenous in size and density, and bed porosity is homogeneous.
- The interstitial gas velocity is constant.
- The mass transfer rate between the solid and gas phases is described by the linear driving force model.

From those hypotheses, the fixed-bed adsorption system can be assigned to the mathematical modeling of the variation of concentration within time, as follows.

$$\frac{\partial c_i}{\partial t} = D_{ax}\frac{\partial^2 c_i}{\partial z^2} - u\frac{\partial c_i}{\partial z} - \rho_p\frac{(1-\varepsilon_p)}{\varepsilon_p}\frac{\partial q_i}{\partial t} \tag{19}$$

The first left term of Equation (19) describes the concentration of component “i” in the gas phase as a consequence of the convection phenomena. The axial dispersion is described by the second term of the Equation, directly related to the axial diffusion coefficient. The amount of gas that accumulates in the packing and the gas adsorption is described by the third term of Equation (19). The last term of the main equation of this section expresses the mass transfer as a consequence of the concentration gradient with the equilibrium, closing the concept of the BKC. The physical properties are assigned by ρ_p, which stands for the bed-density; and $\partial q_i \partial z$, the concentration of equilibrium or the mass balance of adsorbed gas in the adsorption framework, determined by the linear driving force approach, which can be calculated thought correlations specified before.

2.2 MOLECULAR SIMULATION

One can endeavor the isotherm of an adsorption system from in situ methods, based on experimental fundamentals, or by *in silico* methods, where computational means set the equilibrium curve (DI BIASE; SARKISOV, 2015; HUANG, L. et al., 2018; LIU, X.-Q. et al., 2016). A molecular simulation aims to reproduce experiments at a low investment cost or promote insights through the molecular perspective for macro behaviors (VLUGT et al., 2009; CASTILLO, 2009). Regarding CO_2 adsorption, molecular simulation has empowered the academic community to distinguish the properties of adsorbents towards CO_2 adsorption, enhancing the process, identifying tendencies, and designing new materials (BURNS et al., 2020; CHEN, H. et al., 2021). The reader will be introduced to molecular simulation’s main development protocols, exploring the basics of the approach to ensure the appropriate fundamentals for the current work.

Molecular simulation is designed to measure the system’s macro properties throughout the molecular interactions of a certain system (DUBBELDAM et al., 2016). Regarding adsorption, the strategy behind this computational method is to minimize the conformational energy between adsorbate and adsorbent interaction from the perspective of the dynamics or equilibrium. Mathematically, for dynamic properties measurement, the Molecular Dynamics technique will be used. When equilibrium is aimed to be described, the Monte Carlo algorithm will be the best choice (DUBBELDAM et al., 2016; VLUGT et al., 2009). To describe those interactions physically, one last agent is necessary, leveling up the interactions and framing the thermodynamic response of the system, regardless of the algorithm. That is the force field. Therefore, to perform a molecular simulation for a determinate adsorption structure, the following points must be observed (DUBBELDAM et al., 2016; VLUGT et al., 2009):

- The adsorbent and adsorbate molecular structure.
- The force field that defines the interactions.
- The algorithm (resolving the objective of the *in silico* experiment).
- Intensive system properties determination (e.g., temperature, pressure, concentration).

The above list will be deepened into separate topics, starting with the adsorbent molecular structure.

2.2.1 The adsorbent and adsorbate molecular structure

To perform a molecular simulation, one must represent all the components constituting the adsorption system computationally. Before calculating the bulk interactions, those species have to be specified and built computationally (HOLLINGSWORTH, Scott A.; DROR, Ron O., 2018b). When defined, features essential for the phenomena will be calculated based on critic temperature, critic pressure, and concentration of a determined set of species that constitutes the bulk (DUBBELDAM et al., 2016; HOLLINGSWORTH, Scott A.; DROR, Ron O., 2018b). For the sake of an example, fugacity will be calculated by a thermodynamic package, based on the critical properties of the adsorbate molecules that are present in the system (DUBBELDAM et al., 2016).

After being computationally represented, the next step in performing an adsorption molecular simulation is to define the freedom of movement for the molecules. In other words, if those are rigid or flexible. In the present work, only rigid molecules were adopted. Non-flexible molecules indicate that inner movement terms, like torsion and bend, will not be considered in the overall interactions (VLUGT et al., 2009; DÜREN; BAE; SNURR, 2009). It does not mean, though, that the polarization of the molecules and their charge/momentum (e.g., quadrupole moment) will have a higher impact on the definition of the final system energy equilibrium (CHEN, Cong et al., 2020). Nevertheless, the charge of the built molecule is essential for the interaction response between absorbent and adsorbate (CASTILLO, 2009). Therefore, a specific force field for the adsorbate must be defined after its structure to assemble molecule charges (AIMOLI; MAGINN; ABREU, 2014; SMIT, 2008).

For instance, the Trappe forcefield (transferable potentials of face equilibrium) is an accurate quantitative method that estimates the functional form between atoms of a built molecule (BAI; TSAPATSIS; SIEPMANN, 2013). Its application allows quantifying interatomic potentials and the molecule's potential energy. Therefore, in the

case of a rigid molecule, the potential energy is determined by van der Waals forces and Coulomb forces, quantified by the Leonard Jones parameters and atoms' charge (DUBBELDAM et al., 2016). Several force fields can be used for this task. However, the advantage of the Trappe force field is that it is based on phase equilibrium data, facilitating the determination of the saturation value within an isotherm (MARTIN; SIEPMANN, 1998; DUBBELDAM et al., 2016). The following equations describe what above has been stated. One can verify by Equation (23) that the essential inputs for defining a molecule's potential energy are ε_o, r, and Z, the depth of the potential energy or the dispersion energy coefficient, distance between particles or particles size, and the charge of a given molecule, respectively. Finally, the distances where the potential energy is zero, for components i and j are addressed as p.

$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{20}$$

$$E_{\text{bonded}} = E_{\text{bond}} + E_{\text{bends}} + E_{\text{torsion}} \tag{21}$$

$$E_{\text{non-bonded}} = E_{\text{Coulomb}} + E_{\text{van der Waals}} \tag{22}$$

$$E_{\text{non-bonded}} = \frac{1}{4\pi\varepsilon_0} \cdot \frac{Z_i \cdot Z_j}{rij} + 4\varepsilon_0 \left[\left(\frac{p_i}{r} \right)^{12} - \left(\frac{p_j}{r} \right)^{6} \right] \tag{23}$$

The framework is also represented computationally, allowing for a thorough analysis of its properties. Given its larger size compared to the adsorbate, predefined guidelines are necessary to construct the adsorbent, addressing size-related challenges and ensuring the adsorbent meets desired specifications (TURBAN et al., 2016). The usage of pseudo-atoms is a technique where atomistic structures, repetitive or more stable (e.g., CH_3), are defined previously (DUBBELDAM et al., 2016). It is a common approach because those structures' atomic partial charges are balanced *a priori*. In the case of zeolites, for instance, the partial charges and the Lennard-Jones parameters (LJ) interaction sites (that will later interact with the adsorbent charges) are disposed of alongside the oxygen and silicon atoms. That strategy allows a better distribution of charges, representing with more fidelity electrostatic interactions (HOLLINGSWORTH, Scott A; DROR, Ron O, 2018a; MÍGUEZ et al., 2018).

The Trappe Force Field (FF) can also be applied to define the adsorbent, yet other approaches, such as the DEIDRING and UFF force field, can be used (BAI; TSAPATSIS; SIEPMANN, 2013; RAPPÉ et al., 1992). Those are notable mentions since they

can be undertaken to a broader diversity of atomistic combinations, especially in the case of MOFs (CHUNG, Yongchul G et al., 2014; STURLUSON et al., 2019).

Another vital aspect of adsorbent is its geometry. The final molecule of the adsorbent will be only correctly built once the geometrical structure of its atoms is set. Considering a three-dimensional space, all atom's relative positions must be declared, creating a framework supercell, the framework structure at a molecular level. In appendix A, one can identify those differences in Figure 41. Every atom has a positional argument within, framing the adsorbate geometry in the x, y, and z axes. The molecular simulation's framework and adsorbate building blocks are set within that last instance setting. Once the molecular structures are set, one follows the definition of how the interactions between those building blocks will be quantified, the interaction force field.

2.2.2 The Force field determination

The process to define the FF is based on the molecular energy of the system, with terms between bonded and non-bonded interactions. Determining the intermolecular interactions focuses on capturing all terms of the chemical entities and their physical properties, hence, the more terms considered, the more accurately the FF will describe the interaction (AIMOLI; MAGINN; ABREU, 2014; KOLLE; FAYAZ; SAYARI, 2021; DUBBELDAM et al., 2016).

$$E = E_{\text{bonds}} + E_{\text{bends}} + E_{\text{tortions}} + E_{\text{non-bonded}} + \quad \cdots \tag{24}$$

$$E_{\text{bonds}} = \sum_{\text{bond}} E_{\text{r}}(r) + \sum_{\text{bond-bond}} E_{\text{bb0}}(r, r_0) + \sum_{\text{bond-bend}} E_{\theta_0}(r, \theta) \tag{25}$$

$$E_{\text{bends}} = \sum_{\text{bends}} E_{\theta}(\theta) + \sum_{\text{bend-bend}} E_{\theta\theta_0}(\theta, \theta_0) \tag{26}$$

$$E_{\text{tortions}} = \sum_{\text{torsion}} E_{\varphi}(\varphi) + \sum_{\text{bond-torsion}} E_{r\varphi}(r, \varphi, r_0) + \sum_{\text{bend-torsion}} E_{\theta\varphi}(\theta, \varphi, \theta_0) \tag{27}$$

$$E_{\text{non-bonded}} = \sum_{\text{Coulomb}} E_{\text{Coulomb}}(Z_i, Z_j, r_{ij}, \varepsilon_o) + \sum_{\text{vdW}} E_{\text{vdW}}(p_i, p_j, r, \varepsilon_0) \tag{28}$$

Representing the functional form of the FF applied in the present work, Equation (24) is stated. The main term E_{index} is representative of the potential energy associated with several movements or features that an adsorbate may have within an adsorbent. The equation was broken into four other terms, for clarity. The first parameter (Equation (25)) regards the bonded energy and its derivatives (e.g., Bond-Bond and Bond-Bend potentials, which are related to the stretching and compression of the molecules' bonds, and the the cross term with bends. The second term regards the bends on their completeness (Equation (26)), considering every interaction that concerns a certain angle as a consequence or as a cause of the molecular interactions (e.g., Bend-Bend). The third term regards torsion (Equation (27)), considering the forced rotations caused by the layers of molecules alongside the process of adsorption, for example. The torsion potential itself ($E_{tortion}$) is described as a three-term Fourier expansion itself, allowing one to comprehend the complexity for an accurate approximation that regards a FF. The last term, the non-bonded therm, (Equation (28)) is used to consider not just the van der Walls potentials, but also the Coulomb potentials, as a way to discretize for the reader the variety of interactions that functional form of a FF has to capture. The units of the present Equation are referenced to a spherical coordinate system (θ, r, φ), precisely due to the best presentation for torsion and rotation, regardless of its application in the present work.

Even presenting several terms, it is reinforced that all those are fundamentally calculated by a handful of parameters, as presented in the definition of the CO_2 molecule for intermolecular interactions. Nevertheless, the LJ parameters (ε_o, r, and p) and a force constant ($\tilde{k}$), when considering bending and torsion potentials are requested to solve every term of the equation. Furthermore, the fictionalization of every term is a chapter apart due to its deduction procedure (DUBBELDAM et al., 2016). As a disclaim for the reader, it is recommended the review the RASPA Software Manual, where every term is deepened ((DUBBELDAM et al., 2016)) in a close format of the one presented in the set of equations above.

The Equation (24) represents a general force field definition. However, it has a general limitation that should be pointed out. The development of a precise FF, even for a specific application, is stated as an obstacle to scientific development (AIMOLI; MAGINN; ABREU, 2014; BOOTHROYD et al., 2023; EMELIANOVA et al., 2023). It happens because the number of degrees of freedom that a molecular system has, at the order of 10^3 degrees, and difficult to define without a numerical approach (VLUGT et al., 2009). Therefore, all simulations developed were compared with previous experimental procedures to validate the simulations developed herein.

Since non-bonded interactions have an essential role in the CO_2 adsorption (CHEN, Cong et al., 2020), the concept of cut-off distance has to be defined since it has relevant mathematical and physical implications on the final measurement of E_{vdw} (SMIT, 2008; HOLLINGSWORTH, Scott A.; DROR, Ron O., 2018b). Figure 3 demonstrates the shape of attractive and repulsive forces between two particles. By summing those potentials, one has the potential energy of interaction measurement in Joules. The potential energy lower point is representative of ε_o. The cut-off is represented by r_f, truncating or defining all interactions in a higher radius as zero. Computationally, it diminishes the operational cost and improves calculation speed. Interactions of an r higher than r_f, are negligible.

In addition to the cut-off concept, the most noteworthy aspect of the computed FF terms in this study is associated with the non-bonded terms, as illustrated in Equation (28). These terms, along with their derivatives, result from the characteristics of the ensemble employed to describe adsorption interactions. Subsequent sections will delve into a comprehensive discussion of all thermodynamic terms considered in the molecular simulation conducted in this work, from the intensive and extensive properties of the system.

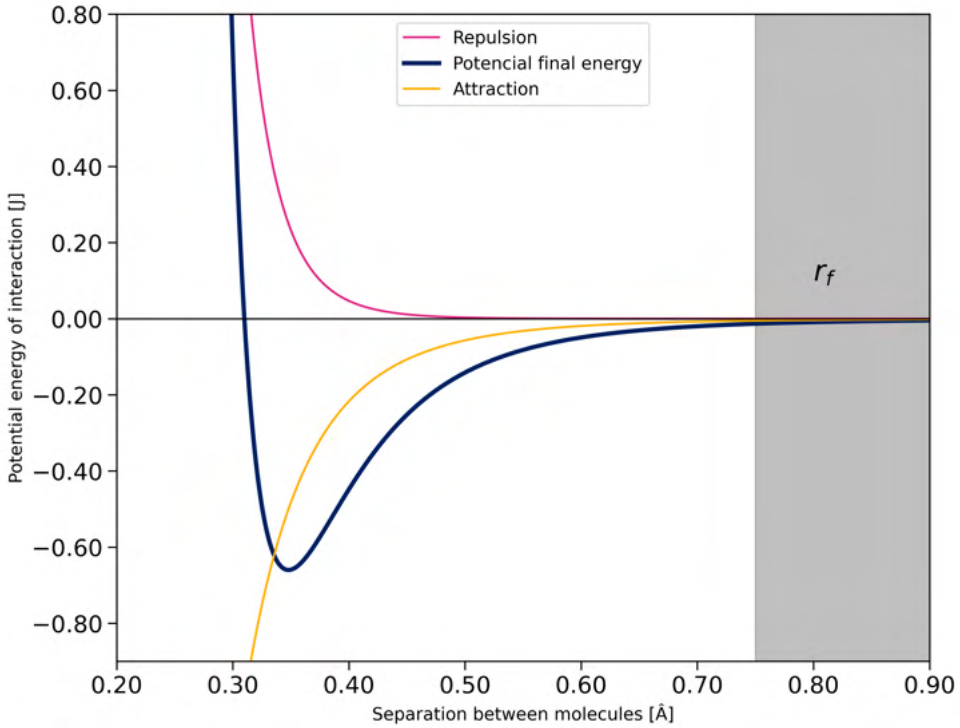
Once the interaction FF is stated, it is possible to determine the algorithm to iterate the arguments of the previously presented equations, finally defining the equilibrium or dynamic properties of the adsorption *in silico* experiment.

2.2.3 The Algorithm specifications

There are two main methods to solve a molecular simulation. Before stating those, it is necessary to establish its fundamentals, grounded on thermostatistics. A molecular simulation has a significant number of degrees of freedom. Mathematically, the theoretical apparatus to represent that physical aspect has to consider several possibilities of the system arrangement (DUBBELDAM et al., 2016). Physically, the thermostatistics principle which sustains that approximation is the ergodic principle (VLUGT et al., 2009). That proposition contemplates the feasibility of every microstate of a system after an appropriate amount of time. Those microstates have an equal possibility to occur precisely due to the high degree of freedom of the nanoscale system. Through this hypothesis, the system is physically represented and mathematically approachable by algorithms, such as the Monte Carlo (MC) method.

The MC method applied to molecular simulations of adsorption is based, as the name says, on the Monte Carlo classical simulations. Briefly, this model is used to define the probability of a different resolution of a system, representing its occurrence

Figure 3 – Repulsion and attraction energy of two molecular entities defined by the distance between molecules



Source: Author (2024)

rate (KANG et al., 2020; KUMAR, K.; KUMAR, A., 2018). Nevertheless, first, one needs to set a previous number of features to a variable of uncertainty. Multiple results will then be calculated by the convergence of that variable of uncertainty until its stabilization, allowing, by consequence, the definition of those preassigned variables by their average over iterations.

By assigning to the main variable of the Monte Carlo approach the system's potential energy defined by the interaction FF, one has the MC applied to adsorption molecular simulation. By converging the potential energy to a stable value, which indicates convergence of the algorithm, properties of the system are calculated over that estimation. Looking at the physical aspect and its mathematical representation within the MC algorithm, several physiochemical variables must be *a priori* determined to guarantee statistical equilibrium. In other words, one provides a path to derivate properties of an ideal thermodynamic system that resembles reality. The algorithm, therefore, works to find that constant value over a set of iterations. This idealization regards an ensemble, a group of thermodynamic variables determined to be constant in the system (VLUGT et al., 2009).

Those ensembles of variables combined with the stochastic methods respect the ergodicity principle by replicating the system on its several possibilities. There are more than a few ensembles that can be cited (LANDAU; BINDER, 2021; VLUGT et al., 2009). However, this work focuses on the Grand Canonical Ensemble (GCMC), where the chemical potential (μ), absolute temperature (T), and volume (V) are considered constants. The micro-canonical ensemble, for instance, considers the number of molecules (n), absolute temperature (T), and volume (V) as constants. It is precisely due to its characteristics that the Grand Canonical Monte Carlo (GCMC) method is applicable efficiently for isotherms definitions *in silico* (DUBBELDAM et al., 2016). The algorithm converges to the thermodynamic equilibrium of an adsorption system since the method will converge to a value of chemical potential that dictates equilibrium between interactions of adsorbate and adsorbent. Another important aspect is the definition of the number of molecules not needed in the ensemble. That definition allows one to determine an estimative of adsorption sites or surfaces properly, embracing the hypothesis that all sites of the framework supercell interact with the adsorbate molecules.

Another method to implement molecular simulation is Molecular Dynamics (MD). Instead of calculating the equilibrium regarding adsorption, MD is applied to define the dynamic properties of the system (HOLLINGSWORTH, Scott A.; DROR, Ron O., 2018b). Since one of the inputs for its application is the time step, diffusion, for instance, can be calculated by the mean displacement of molecules over the system. The number of molecules also has to be defined, being the ensemble, therefore, different than the GC. MD is a method where the equation of motion of the particles (in the case of adsorption, the adsorbate) is solved numerically (DUBBELDAM et al., 2016). Hence, the solution of the system is done in a time discretization when the thermodynamic properties of the system, as well as kinetics, can be determined.

It is important to emphasize that, even though it is a method that solves the equation of motion of particles numerically, it is still a thermostatistics method based on statistical mechanics. Moreover, the present work solely develops *in silico* experiments utilizing the GCMC method. Therefore, MD will not be as deepened as GCMC was.

The algorithm specification is the last building block for the execution of a molecular simulation, alongside the force field and the computation representation of molecular structures defined. With all those in hand, one is capable of performing a proper molecular simulation. With that stated, the *in silico* experiment of adsorption needs to comprehend what is measured within the molecular simulation, nothing else than the intensive and extensive properties of the system. Even though it is a fundamental aspect of physics, it is deepened with a didactic purpose. Since it helps the reader comprehend

what is measured at the nanoscale, the core of the strategy for developing the present work relies on the intensive properties of the system. Those determine what in the future will be assigned as instances, regarding Machine learning methods. A data-driven approach is then developed based on the intensive properties of the adsorption system.

2.2.4 Extensive and intensive properties of the nanoscale system

The definition of the intensive properties of the adsorption system is fundamental to calculating its extensive properties employing a molecular simulation approach. *Intensive properties* are features that do not change in every fraction of the system (BORGNAKKE; SONNTAG, 2020). These properties are independent of the mass amount in a thermal equilibrium system. Temperature, in this case, is one physically intensive property.

In the case of the Grand Canonical ensemble, one can relate that the essential properties of the simulated system are the chemical potential and temperature. Considering that, for a single temperature, several points of mass equilibrium are related for a given pressure (e.g., isotherm for an adsorption system), one can define the same for pressure.

An adsorption system at an equilibrium point of temperature and pressure has more intensive properties assigned, such as density, for example. However, the density of an adsorption system is *a posteriori-defined* value, not a priori, as temperature and pressure, since the density regards the amount of mass adsorbate added to the mass of the framework and volume. Even though not an extensive property of the system, it is a consequence of *a priori* physical definition. The same follows for the heat capacity of the adsorbate-adsorbent complex.

On the other hand, the enthalpy of adsorption depends on the number of molecules interacting with the framework, therefore it is an extensive property of the system. Furthermore, due to their basic definition, van der Waals energy, Coulomb energy, and differentials regarding the adsorbate and adsorbent within each other are also extensive.

Finally, the extensive properties of an adsorption system are a direct consequence of the number of interactions between adsorbate and adsorbent. Considering the GCMC method, those properties are calculated by the average number of iterations or cycles convenient for convergence.

Table 1 resumes the adsorption system's intensive and extensive properties *in silico* developed. More features can be defined regarding the approach and method of calculation. However, the ones present in Table 1 are applied to the present work's development, all based on the RASPA manual (DUBBELDAM et al., 2016)

Table 1 – Properties from molecular simulation

Property	**Int/ext**	**Post/Prio**
Pressure	Intensive	Priori
Temperature	Intensive	Priori
Final host/adsorbate energy	Extensive	Posteriori
Final host/adsorbate vdW energy	Extensive	Posteriori
Average volume	Extensive	Posteriori
Average density	Intensive	Posteriori
Average heat capacity	Intensive	Posteriori
Total energy	Extensive	Posteriori
Enthalpy of adsorption	Extensive	Posteriori
Average derivative of the chemical potential	Intensive	Posteriori
Average adsorbate-adsorbate energy total	Extensive	Posteriori
Average adsorbate-adsorbate energy vdW	Extensive	Posteriori
Average adsorbate-adsorbate energy coulomb	Extensive	Posteriori
Average host-adsorbate energy total	Extensive	Posteriori
Average host-adsorbate energy vdW	Extensive	Posteriori
Average host-adsorbate energy coulomb	Extensive	Posteriori
Total vdW	Extensive	Posteriori
Total coulomb	Extensive	Posteriori

Source: Author (2024)

Special attention must be given to the units associated with the variables mentioned above. In the upcoming sections, we will delve into the methodologies of data wrangling and manipulation employed to standardize the input data for our machine-learning applications. To streamline the entire process, we have opted to utilize the internal energy units from the RASPA software, referred to in this work as "U" (DUBBELDAM et al., 2016).

These internal units in the software are defined by a coefficient equivalent to the ratio of the Boltzmann factor to molar energy units (e.g., J/mol). The following example (Equation (29)) illustrates the conversion from the internal units of the RASPA software to molar energy units. For the following equation, k_b stands for the Boltzmann constant, and the 300 U value is related to the temperate of simulation, since in practical terms $U = K$.

$$U = 8.31446\frac{k_b}{J/mol} \quad \rightarrow e.g. \quad = (–2000 \;\; U – 300 \;\; U)\frac{8.314446}{1000} = 19.123\frac{kJ}{mol} \tag{29}$$

Briefly, the main topics that the reader should keep in mind from this section can be summarized in the following:

- GCMC methods can be used to define equilibrium in a molecular adsorption simulation, having an isotherm of adsorption as its main result.
- Extensive properties of the system and posterior intensive properties are determined by a priori intensive variables, such as temperature and pressure.
- The GCMC method has "μ" being defined by the convergence of the phenomena, and the convergence process expresses the ergodic principle being calculated.

Concluding the present subsection, it was presented all building blocks and properties measured for a molecular simulation of adsorption. From the knowledge above, several materials can be developed *in silico*. Hence, the materials developed in the present work will be presented next to its molecular simulation parameters.

2.2.5 Materials simulated

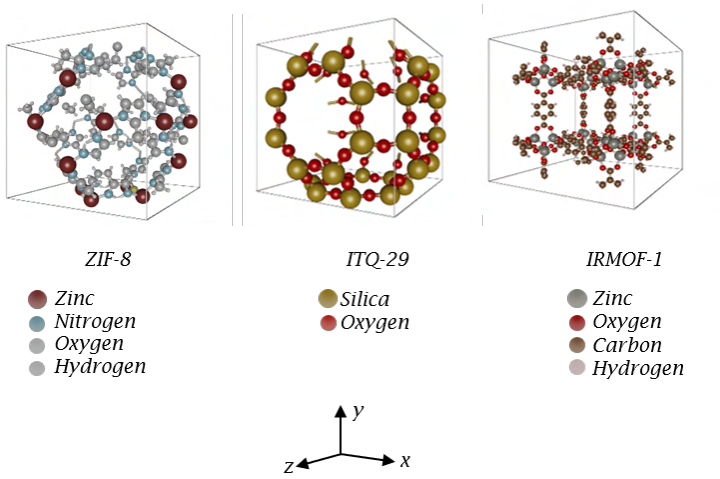
Several materials can be applied to the adsorption of CO_2, as Activated carbons, Zeolites, and Metal-Organic Frameworks (MOFs). The materials developed in the computational experiments in the present work are Zeolites and MOFs.

Zeolites are based on Alumina and Silica, being crystalline materials with microporous of a magnitude ranging from 0.5 to 1.2 nm (COLELLA; WISE, 2014). Zeolites are stable materials with a high surface area and strong adsorption sites. MOFs are porous crystalline solids alongside zeolite imidazolate frameworks ZIFs. A MOF can be conceived as a structure where organic compounds link metallic clusters (CHUNG, Yongchul G et al., 2019a). The highlighting aspect related to the MOFs materials is their commonly high superficial area and their diversity since the synthesis is accessible and the open metal sites - a fundamental aspect of its composition - present a trend to form strong links with CO_2 (GHANBARI; ABNISA; DAUD, 2020; LIU, C. et al., 2021). A ZIF framework can be addressed as a MOF derivate, being a porous crystalline material. However, its resemblance with zeolites makes them more comprehensive when one realizes its structures are built with transition metals despite silica. ZIFs can be understood as frameworks with features close to Zeolites and MOFs, since their basic

compounds have the same geometry as Zeolites but are built with metal links.

The present work follows the development of three materials, each representative of the three above classes of reviewed materials for CO_2 adsorption applications: IRMOF-1, representative of MOF's structures; ZIF-8, representatives of ZIF structures; and ITQ-29, representative of zeolites. In Figure 4, one can verify more specific differences in the structure of each material addressed in the present work. Unit cells are presented at an angle of 60° on the z-axis, in a projection perspective, rotated on the x-axis. Cells are adapted from the iRASPA visualization software for materials. Chemical composition is presented too (DUBBELDAM et al., 2016). Table 2 is presented in the following where SBET, pore volume, pore diameter, and particle density.

Figure 4 – Molecular unit cell of IRMOF-1, ITQ-29 and ZIF-8, and chemichal composition



Source: Author (2024)

IRMOF-1 has the chemical formula of $Zn_4O(BDC)_3$, where BDC stands for 1,4-benzene-dicarboxylate. IRMOF-1 exhibits interconnected channels; that feature apertures measuring 12 Å and 15 Å in size, having as base an octahedral module built with Zn_4O cores (BABARAO et al., 2007a, 2007b). Alongside carboxylate connectors, the structure presents a three-dimensional cubic form, resulting in a highly porous framework. This specific material is also known to have isoreticular structures, such as IRMOF-3, IRMOF-10, and several others, commonly related to structural flexibility (BABARAO et al., 2007a). This feature is not studied in association with IRMOF-1 in the present work.

ITQ-29 is a zeolite with a relevant channel diameter compared to other common

Table 2 – Materials simulated properties

Material	ITQ 29	ZIF 8	IRMOF1
SBET area [m^2/g]	629.000	1386.000	1810.000
Pore Volume [cm^3/g]	0.300	0.730	0.552
Pore diameter [m]	1.907	2.106	1.219
Particle density [kg/m^3]	1432.806	924.468	593.306
Reference	(TISCORNIA et al., 2008)	(SAEEDIRAD et al., 2020)	(BABARAO et al., 2007a)

Source: Author (2024)

zeolites, such as BEA and MFI (TISCORNIA et al., 2008; MARTIN-CALVO et al., 2018). ITQ-29 is a silica Linde type A zeolite with a constitution of 3d-cages connected. The result of its basic constitution is the presence of large cavities. A relevant aspect associated with zeolites is the Si/Al ratio, a feature normally used to describe several other properties. However, ITQ-29 is a Zeolite with no presence of Alumina, being replaced by Germanium, which, considering other factors, is associated with the ITQ-29 zeolite thermal and acidity resistance (MARTIN-CALVO et al., 2018).

Finally, ZIF-8 is structured by Zn sites, presenting a significant surface area. The basic constitution of ZIF-8 is imidazole bonds, which comprehend its overall structure according to its crystallographic information (SAEEDIRAD et al., 2020). Links of Zn - N and C - N are notable aspects of the ZIF-8 structure (MARTIN-CALVO et al., 2018).

Closing this section, in the last part of the Theoretical Background Chapter, the reader will be introduced to Machine Learning Algorithms, data set structures, concepts, and methods to step into the methodology section.

2.3 MACHINE LEARNING ALGORITHMS

Before starting this section, some concepts should be introduced to clarify the additional ideas that will be expressed. First, every area of science is expected to have nomenclatures for a distinguished class within. It is no different in the context of data science and machine learning. For instance, the variables of a phenomenon can be described or assigned to the name of "features." If one has more contact with deterministic and mechanics frameworks, the variables of a system (e.g., density, number of molecules, loss of charge) can be understood as "features" too. In parallel, every individual experimentation that describes a phenomenon's outcome, a particular procedure, defines the concept of "instance" for the ML context. It is important to emphasize that an instance can be a value of pressure, time, temperature, and other examples

that the reader may think. Given that, for every instance, there is a measurement of all the features of the phenomena associated once one is working with ML. In other words, a feature is a characteristic of the phenomena. At the same time, the instance is a singular execution of that phenomenon. A simple way to understand that is by looking at a spreadsheet, where every line corresponds to an instance, and every column corresponds to a feature. Finally, a designed feature, or a couple of it, can be assigned as a target or output. The rest of the features are assigned to inputs. Thus, from those inputs, MLAs are trained to predict the outputs.

Following through, machine learning applications have received significant attention in recent years (FOTOOHI et al., 2016; LEE, Y. et al., 2018; PILANIA et al., 2013). Due to their capacity for prediction, forecasting, and classification, Machine Learning Algorithms are tools applied to discover patterns or predict a target, a discrete, continuum, or class type of value. There are two main archetypes of algorithms: unsupervised and supervised models (GÉRON, 2021).

Those two paradigms, unsupervised and supervised, are assigned to unlabeled and labeled data sets. Regarding Supervised Machine Learning Algorithms (SMLA), when one has a labeled dataset, it is common sense that it was previously reviewed or verified by an expert since every feature is described or named (RASCHKA; MIRJALILI, 2019). On the other hand, unsupervised Machine Learning Algorithms (UMLA) are orientated to discover hidden data patterns, becoming specialists (RASCHKA; MIRJALILI, 2019). One can distinguish SMLA and UMLA by the perspective, where the first applies to regression and classification, and the second by grouping or clustering data. The present work uses SMLA exclusively.

Despite the archetype, the application of MLA is vast (RACCUGLIA et al., 2016; POURSAEIDESFAHANI et al., 2018). An expressive application niche is Material Science and multiscale modeling, the core of this work. Regarding chemical engineering, it is almost impossible to punctuate MLA applications in the present document.

Since SMLA are an essential component of the present work, the algorithms' fundamentals and nature will be presented and elucidated before explaining the case study. The third part of the fundamental methodologies of the present work will present the principles that regard all MLA alongside artificial neural networks (ANN) structures and famous applications. Random forest (RF) will be clarified in the following subsection. Dataset structures will be explained further regarding the importance of that aspect for the case study. Finally, a brief review of the most vital related applications will be presented within the concept of data-driven engineering.

2.3.1 Machine Learning Algorithms general specifications

MLA is predicated on statistical and mathematical methods; every model has its inner approach assigned. The capacity of an MLA is based on its learning experience and a model's training process to estimate an outcome from related inputs (GÉRON, 2021; RASCHKA; MIRJALILI, 2019). This process is based on describing a phenomenon by a set of features and instances arranged in a dataset. From that dataset, the learning process of an MLA starts.

Once it is started, it is optimized by *a priori* error standard or several iterations. To find an optimum result, two main aspects are looked upon in the initialization of the model: the optimization method and the loss function (GÉRON, 2021). To create a first prediction of the outcome, a random set of numbers is determined by the optimizer to the weights of the model. After evaluating if the initial weight can perform a good prediction, the optimizer updates the model weights to minimize the loss function (e.g., accuracy). Different optimizers (e.g., gradient descent, Adaptive Moment Estimation (ADAM)) can be set to undertake the minimum value for that task (BANGERT, 2021). Finally, a linear regression can be set between the actual and predicted outcomes. In simple terms, the weights assigned to the current model have changed according to the angle of the regression closeness to the 45º, the linear regression optimum result. That 45º angle line represents the loss function minimum.

Just as the optimization method can be set, several other steps can build a supervised machine learning algorithm (SMLA). However, some standard methods grip the final model performance, regardless of the optimization method. Those procedures are related to the data preparation phase, specifically, dataset subdivision and feature scaling (BANGERT, 2021).

The division of the dataset is done in two or, in some cases, three parts to train and validate a model. The first division is dedicated to model training when iterations for comprehending the inner patterns are developed. At this moment, the model "learns" from data. The second part is dedicated to validating the build model, and the third part, when used, is dedicated to the final result evaluation. A random splitting of the data is done within this process to avoid biases.

Data preparation is also executed by scaling the data (BANGERT, 2021). Scaling can be done by normalization or standardization. Normalization bounds the range of the values between two numbers, commonly -1 and 1, while standardization will

transform the data to have the mean at the zero point and a variance of 1 (KROESE et al., 2019). Both techniques will make data unitless; thus, regardless of normalization or standardization, that approach is relevant when the scale of the features can be discrepant. By putting them on the same scale, the SMLA is not induced to consider one more than the other, avoiding favoring one over the other based on the measurement and emphasizing the information gain the data offers to the model.

The scaling procedure improves the learning rate of the machine learning algorithm (GÉRON, 2021). For example, if X1 is much higher than X2, the optimization method may take a long time to converge. When X1 and X2 are scaled, the learning rate is accelerated. It has to be clear that, regardless of the usage of normalization or standardization, the variance and distribution of data are not affected. Another aspect is that the correlation coefficient of a scaled variable is the same as that unscaled variable, with a target or output (GÉRON, 2021). The gain is related to the learning rate of the machine learning algorithm within the optimization method.

Until now, it has been presented to the reader three main points for the development of a general MLA: optimization method, dataset division, and features scaling. Several algorithms can be used for a regression problem (e.g., Support Vector Machine SVM), decision tree (DT), logistic regression, etc.). However, what model should be used to perform a good prediction of outcomes? Artificial Neural Networks (ANNs) are highlighted in this aspect due to their fundamentals when looking toward the application in Chemical Engineering.

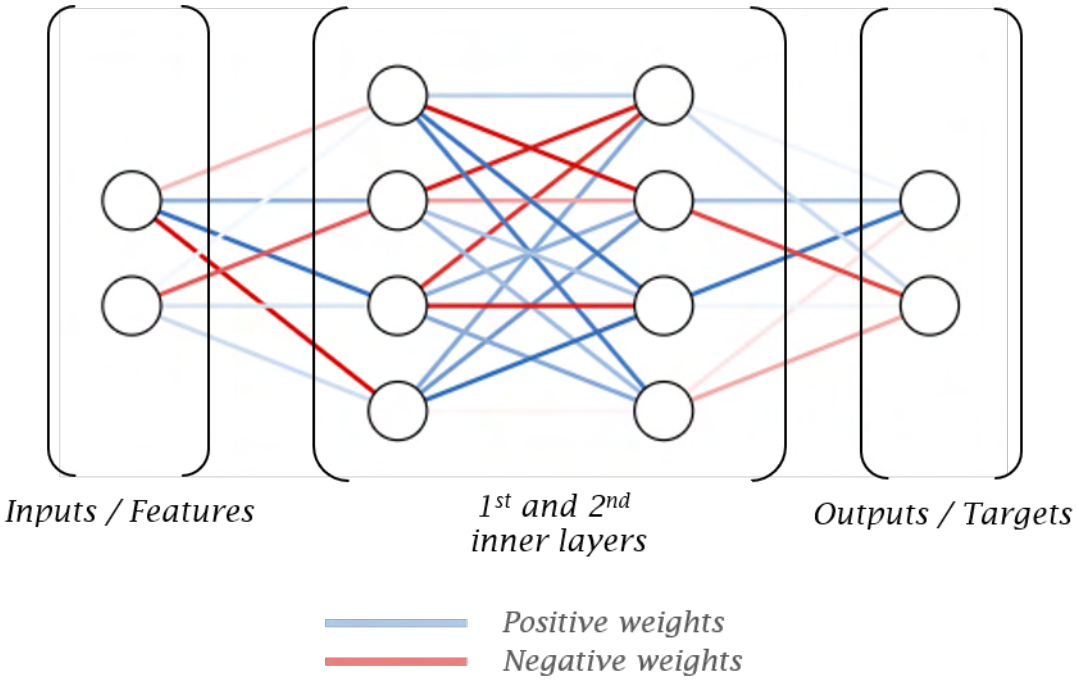
2.3.2 Artificial neural networks

ANNs are mathematically based on the Universal Approximation Theorem (UAT) (NISHIJIMA, 2021). Considering a finite number of dimensional spaces in a Euclidian domain, it states that a feedforward multi-layer perceptron can characterize an expressive number of functions, hence, universal approximators (NISHIJIMA, 2021). Regarding the present work, it is fundamental that the reader keeps in mind that neural networks can characterize functions by fitting weights within the model and approximating values from one domain to another.

Going further on the Neural Network comprehension, starting with its creation motive is vital. ANNs are inspired by the human brain's neurons, whereas income goes forward within a web of connections (AMBAW, 2005). Mathematically, those neurons are settled by graphs and their connections by vectors. Figure 5 illustrates that central conception represents multiple inputs and multiple outputs neural networks. Strong

color vectors represent absolute higher weights. Nevertheless, the aspects that made ANNs so effective on approximation tasks can be summarized by their architecture, the number of inner layers, and the initialization functions of every neuron (AMBAW, 2005).

Figure 5 – Multi-layer perceptron neural network



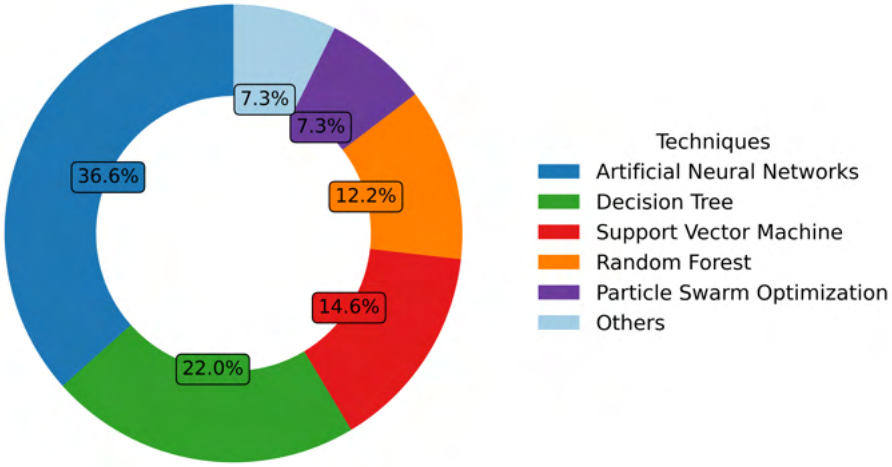
Source: Author (2024)

Efficient estimation and process modeling pose consistent challenges in the context of CO_2 adsorption (DOBBELAERE et al., 2021; WANG et al., 2017; REBELLO et al., 2022). Consequently, efforts focusing on ANN-based strategies offer feasible approaches to predict additional indicators beyond efficiency. The literature demonstrates direct applications of ANN in modeling the full PSA cycle (YE et al., 2019).

ANNs play a crucial role in comprehending CO_2 adsorption phenomena, as illustrated in Figure 6, where their usage is prevalent in the field. When considering deep learning, ANN-derived approaches encompass 52% of the techniques employed in this area, based on the reviewed literature for the present work. Notably, ANN approaches outperform other methods like decision trees and meta-heuristic. Furthermore, this technique allows for the investigation of time-dependent systems, providing a reliable approach to process modeling (YE et al., 2019; LEPERI et al., 2019). Regarding CO_2 adsorption application, the advancement of different architectures, beyond the MLP

one, is not vast, but present in the literature (WANG, Zhenguang et al., 2022; MARTINS et al., 2021; OLIVEIRA et al., 2020). Insights from ANN's are still scarce, leaving several points to be explored, connected with data and MLA intelligence, going beyond the present literature.

Figure 6 – Machine Learning Algorithms most used in the CO_2 adsorption field



Source: Author (2024); Note: The information presented is derived exclusively from the set of articles reviewed for this work

2.3.3 Random Forest

RF procedures are grounded in the Decision Tree algorithm (GÉRON, 2021; BIAU; SCORNET, 2016). The mathematical approach that bases DT divides the dataset into small groups separated by their internal resemblance and external differences. Fundamentally, a DT will learn from observations regarding heterogeneity and homogeneity from the dataset. Random Forest will follow the same, although with some enhancements, where a decision tree is developed for every prediction class instead of finding the differences and similarities between a dataset subset (GÉRON, 2021). A random forest algorithm can be summarized as a group of decision trees, adding the principle that a group of moderately independent models (trees) functioning collaboratively as a committee will outperform any individual model (GRÖMPING, 2009).

The random forest can be applied for classification and regression tasks. Despite the application, this model's hyperparameters are related to the structure of a decision tree, with some concerns related to the size of every tree inside the forest

structure (GRÖMPING, 2009). For instance, the number of trees in every node may affect prediction performance since many trees affect information discretization. At the same time, the number of features used to split a node directly affects the model's computational cost. Lastly, the number of subsets will affect the data entropy, impacting the model performance (ORNSTEIN; WEISS, 1993). Finally, a detailed description of the algorithm will be synthesized to clarify how RT approaches a dataset.

- The dataset will be split into subsets by a random subspace choice or bootstrapping, following a hyperparameter definition.
- A decision tree is trained for every subset, where every decision tree has a singular way of determining the outcome.
- All trees are aggregated by the ones with the best performance on subset outcome estimation.

Out of that list, two concepts are behind the whole procedure. First, data entropy is deeply related to the RT method since it will not just define the informational value that the dataset has (ORNSTEIN; WEISS, 1993); it will be applied directly to define the division of subsets following the principle that subsets have to be homogeneous inside but heterogeneous from the outside. Second, every decision tree within the forest is related to the subset data entry; hence, the results of best-performance decision trees are averaged to detach the model from its dependency on data entry. From that, a random forest reduces biases and avoids overfitting (KROESE et al., 2019).

Notwithstanding, the hyperparameters definition of a Random Forest will be related to the trade-off between data entropy, model capacity, and computational costs. The consequence of those choices will impact every decision tree within the random forest, where a blueprint of the subset related is stated in the format of heuristics.

2.3.4 Models interpretability: Opening the Black Box

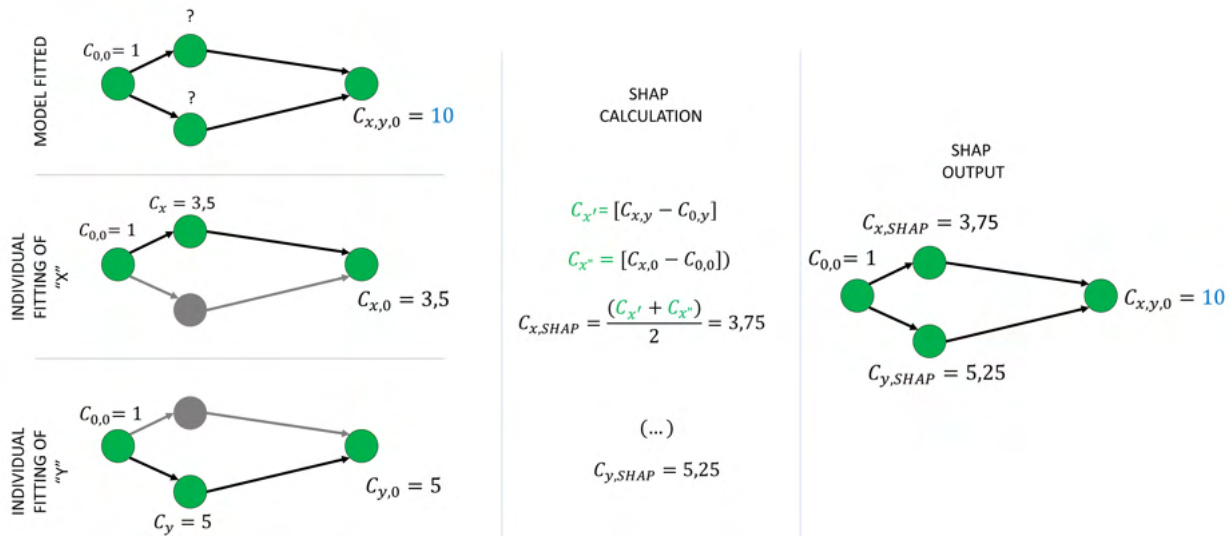
Interpretability is an important factor to be kept in mind when a mathematical approach is applied. The same follows to MLA. However, the term *black box* is commonly relatable with MLA since those can have several shells, making them complex to understand when compared to a deterministic function, for example (GUNNING et al., 2019; MESSALAS; KANELLOPOULOS; MAKRIS, 2019). A deterministic equation's applicability and physical relations are reachable by analyzing the units from its coefficients and scales. Mathematical operators will also present clearly how those units and coefficients are related and, lastly, how those affect the dependent variable. MLA

are built with several coefficients, weights, and heuristics, making those models hard to read by the same means as a classic equation. The explainable artificial intelligence approach (XAI) can open the so-called *black boxes* (DAS; RAD, 2020).

SHAP is an explainable artificial intelligence approach (XAI). It is one of the most popular due to its practical aspects accessible for Scikit Learn and Keras, the basic ML packages used in the present work (DAS; RAD, 2020). SHAP stands for Shapley's Additive explanation, and the mathematical approach under it is grounded on cooperative game theory, where the expected marginal contribution of each feature is calculated. The expected marginal contribution is the SHAP value (DAS; RAD, 2020).

The most relevant SHAP aspect regarding its application in this work is to explain how each feature has contributed to a forecast application. By doing so, one can understand if the MLA application considers features with physical meaning relatable to the target or if the most relevant features agree with the literature. The scheme entitled *Expected marginal contribution of two features*, representative of the SHAP analysis, is resented at Figure 7. Fundamentally, the illustration represents a didactic way to comprehend how the SHAP values are measured.

Figure 7 – Expected marginal contribution of two features - SHAP analysis



Source: Author (2024)

Firstly, the target $C_{x,y,0}$ is a prediction by the MLA called "x,y,0", being "x" and "y" two features that one aims to understand its contribution to the model, and "0" represents all the features standardly used in the model. When the "y" and "0" features are not present, "x" can predict the target as being 3.5. For the "y" feature solely predicting

the target, it defines it as being equal to 5.0. The model "x,y,0" defines the target value of 10.0. To calculate the marginal contribution of "x" and "y," the average between the theoretical importance of "x" to the final model and the individual contribution of "x" to a model where just "x" and "0" are present (C'_x value and C_x") defines the SHAP value of "x," hence, its contribution in the final model.

The reader should notice that the summing of the SHAP values with C_0 represents the final model forecast, which does not mean the final model makes the best prediction. One has to assume that the final model performs well, defined by different methods, being those statistical or comparative.

2.4 MULTI-SCALE MODELING THEORETICAL BACKGROUND INTEGRATED WITH MACHINE LEARNING AND MOLECULAR SIMULATION

The integration of Machine Learning Algorithms (MLA) and Molecular Simulation (MS) in the development of multi-scale modeling introduces a novel perspective. Traditionally, the multi-scale approach in chemical engineering has centered on non-dimensional numbers (e.g., Reynolds, Weber, Chyly modulus, etc.) (KEVLAHAN, 2012), respecting constitutive relations in the scales interfaces. However, when considering a macro work frame and a defined control volume, specifically adopting a Newtonian approach to the system, non-dimensional numbers capture only a facet of the system that does not encompass each particle necessarily. Instead, they reflect a common behavior associated with all entities governed by constitutive relations and boundary conditions. Musti-scale modeling, aiming to capture the individual physics of every molecule within the system linked with larger scales, introduces a new set of variables as the scale changes, with attention given to error propagation across scales (HOEKSTRA; CHOPARD; COVENEY, 2014).

The exploration of chemical reactions emerges as an intuitive pathway, particularly when attempting to depict every molecule interaction within the controlled volume. In recent years, the design of chemical reactors based on insightful molecular simulations has gained relevance (KEIL, 2018). However, the increased level of detail and discretization comes at a mathematical (and computational) cost. In the realm of molecular simulation (e.g., Grand Canonical Monte Carlo (GCMC) and Molecular Dynamics (MD)), the integration of Machine Learning (ML) is considered a strategic approach for determining GCMC potentials, thereby balancing the computational cost required for multi-scale modeling (MS) (YANG, Wuyue et al., 2020; KEIL, 2018). For illustrative purposes, although not the focus of this work, it is worth mentioning the application of ML alongside MD, especially when the time framework of the procedure can be extended

without a loss of information or biases due to Machine Learning contribution (BOTU; RAMPRASAD, 2015).

As relevant as reactor modeling, multi-scale modeling finds application in material analysis, where structural molecular assessments rely on the parametrization of material history and mechanical state space (KARAPIPERIS et al., 2021). Once again, computational cost remains a significant challenge (KARAPIPERIS et al., 2021; KEIL, 2018; YANG, Wuyue et al., 2020). However, exploring the interface between molecular-level interactions and process-level descriptions becomes relevant, encompassing considerations such as the accuracy of MS (e.g., force field fidelity) - and already considering CO_2 adsorption - precise material composition data, the impact of surface heterogeneity, crystal formation, and synergistic effects (FARMAHINI et al., 2018).

Attempts to improve force fields deserve special attention, as this is a problem extensively discussed in the theoretical background of Molecular Simulations, particularly within the Force Field subsection. The *in silico* modeling of gas-solid interactions serves to explore and complement the design of new materials, employing a combination of a priori simulations with the Density Function Theory approach—a popular method in recent years (XIANG et al., 2010; MAHAJAN; LAHTINEN, 2022; MORGANTE; PEVERATI, 2020). However, a challenge arises from the trade-off between enhanced FF accuracy through a priori approaches and the associated increase in computational cost. This imbalance becomes pronounced when aiming for precise multi-scale modeling to connect nano-scale behavior with macro-operation indicators of performance. The summation of challenges becomes larger and larger.

A term that encapsulates the solution presented in this work is "Bridge". Connecting molecular-scale data with process descriptors involves the application of ML within the methodology outlined in the following section. Upon validating MS outputs with experimental data, the generated information can be utilized to incorporate these insights into industrial-scale operations (VEGA; BAHAMON, 0000; BAHAMON; VEGA, 2016).

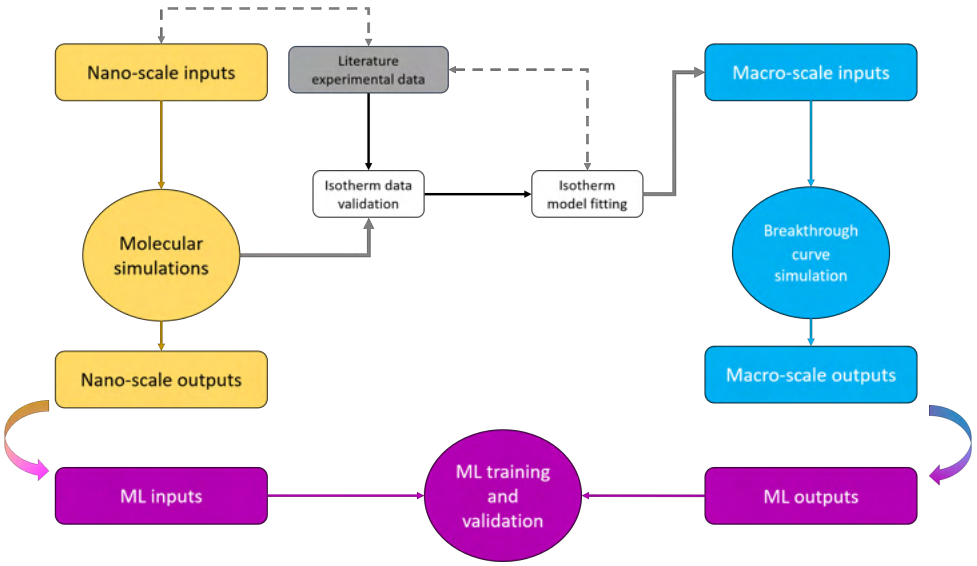
3 METHODOLOGY

The conceptual background that sustains the methodology detailed in the current section was presented. Following through, the present chapter is divided into three main branches to present the workflow developed constructively. These branches maintain a connection to the theoretical background but with a focus on practical aspects, which are crucial for building the final result. The three branches are as follows:

- The first branch outlines the data development procedures at the nanoscale.
- The second branch delves into the development of targets at the macroscale and emphasizes the reliability of physics in both scales.
- The third branch elaborates on the method used for performing artificial neural networks' multilayer perceptron, Random Forest, and the software for conducting all simulations.

To provide a visual representation of the entire protocol and simplify the understanding of the process, a schematic flowchart entitled *Simplified flowchart methodology* is included (Figure 8).

Figure 8 – Simplified flowchart of the methodology



Source: Author (2024)

Figure 8 presents an intuitive notion of how the dataset was built and illustrates an interface dedicated to interaction with the protocol as a unit. A good starting point is comprehending that nanoscale simulations have their inputs. Once performed, the

outputs from molecular simulations of CO_2 have transformed into inputs for Machine Learning (ML) applications. Similarly, macroscale simulations also have their inputs. Once those are determined through the BKC simulation, their outputs are transformed into targets in ML applications. Briefly, the outputs of nanoscale and macroscale simulations are inputs and targets, respectively, from the perspective of the Machine Learning approach. Another important aspect is that considering the interface with the ML models, the inputs for that interface are the same regarding nanoscale inputs and macroscale inputs, closing the information workflow. Lastly, it is essential to recall that the inputs for nano and macroscale approaches are detailed in the theoretical background.

Before stepping into a deeper outline of the methodology, some principles follow through every step of it. The present work's final deployment follows a direct integration of three work niches: nanoscale simulations, macroscale simulations, and Machine Learning Algorithms. Although properly related within a data-driven approach, one has to be concerned with the structure where multiscale data are assembled and the keys that allow them to be connected in a single dataset. Going straight forward, the keys that differentiate every instance are the core points of a structured dataset.

The physical connection between scales plotted in a structured dataset is related in this work to the keys of every instance. So, the phenomena modeled at the nanoscale occur in a fraction of a second, measured in nanoseconds or picoseconds. Furthermore, the same phenomena modeled on a macro scale develop itself in a time frame of hours. However, regardless of scale, Pressure and Temperature are framed equally. Intuitively, it can be understood as an absolute correspondence between scales. That principle allows one to build a structured data set, which would not occur if time was used as a granularity key (level of detailment). On that account, the level of detail of the dataset regards temperature and pressure sets, allowing the whole dataset to be structured for thermodynamic equilibrium for each material being used as the framework.

An important point should be addressed regarding the contrast that GCMC and isotherms models have. Both of them describe the same phenomena, but from different perspectives: one macro, the other micro. One should note that the macro-modeling of the isotherm considers constrains over what happens with the system at a nano-scale. Those constrains should match between scales, while the constrains of the macro model should consider all the measured interactions at nano-scale. Anything different than that would result in a multi-scale model inconsistency or, in a more detailed evaluation, a thermodynamic inconsistency. Although the Langmuir isotherm initially appears to fit well across various temperatures and materials, it was excluded from the isotherm model fitting due to its limitations in defining nano-scale interactions. Instead,

the Freundlich and SIPS models were used to fit the GCMC data.

Since both scales represent the same phenomenon (CO_2 adsorption) but use different methods — Monte Carlo stochastic algorithms at the nano-scale and deterministic modeling at the macro-scale — validating data at both scales is essential. Experimental validation is crucial for ensuring that models accurately reflect the physical causality between scales, not just statistical correlations. At the nano-scale, the isotherms developed by Monte Carlo stochastic algorithms are compared with previous works and experimental results to validate the frameworks studied. For the macro-scale data, the model used was required to describe the experimental data also from experimental data from the literature.

Summing up the theoretical approach, the core of the present thesis relies on the approach where intensive properties of a system describing a particular phenomenon can be used as instances to ensemble features from different scales in a structured dataset. This sentence resumes the principal concept of this work leading to its results.

3.1 MOLECULAR SIMULATION PROCEDURES – FIRST BRANCH

The simulations regarding CO_2 adsorption were performed in the RASPA software, a general-purpose simulation package (DUBBELDAM et al., 2016). The software has a pre-definition regarding its compilation, assembled in C++ language, requiring a few libraries and specific compilers (e.g., the GNU Compiler Collection, 'GCC', and the Intel C++ Compiler, 'ICC').

To proceed with simulations, the user needs to define the following files:

- 'simulation.input': A file where the specification of the simulation is defined alongside the primary orientation of the characteristic of the simulation.
- 'FRAMEWOKR.cif': where the structure of the adsorbent is defined in the format of a '.cif' file or '.xyz' file.
- ADSORBATE.def': where the structure of the adsorbate is defined (e.g., atoms positions, rigid/flexible, critical constants, bonds)
- 'ForceField.def': where the van der Waals potentials are listed, alongside tail-corrections, cutoffs, mixing rules definition, and, lastly, LeJs parameters and charges.
- 'ForceFieldMixingRules.def': used when the 'ForceField.def' is not present, specifying the same information from the pair's definition despite individual

atom values.

To properly run the software beyond the above-cited specification, a file named “RUN” must be called in the integrated development environment (IDE). The “run” file is a shell script directing the software to a directory where the above files are resident. At the same time, this file is also commanding the software to run by the terminal. The whole package for RASPA software can be found in a GitHub repository (http://github.com/irapa/RASPA2). The simulations performed in the present work were performed through the IDE Visual Studio Code v. 1.78.0.

3.1.1 Molecular simulation specifications

The “simulation.input” file is the document that assembles the principal specifications for the CO_2 adsorption simulation to all materials developed. This file is where the specifications for the isotherm calculation are presented, hence, being a central file. A sample of it will be detailed in the following, although the reader should keep in mind that the nature of this code, which is embedded with the RASPA software, has its syntax. The complete code concerning the “simulation.input” file is presented in Figure 9 to then be elaborated.

Figure 9 – General simulation inputs for RASPA2 software simulation framed in the "simulation.input" file

```
SimulationType                              MonteCarlo
NumberOfCycles                              25000
NumberOfIniticalizationCycles               5000
PrintEvery                                  5000
PrintPropertiesEvery                        5000

Forcefield                                  GenericMOFs
UseChargesFromCIFFile                       yes

Framework       0
FrameworkName   Cu-BTC
UnitCells       1 1 1
HeliumVoidFraction 0.77390
ExternalTermperature 313.00
ExternalPressure 1.0e4 1.0e5 1.0e6

Componemt 0     MoleculeName                CO2
                MoleculeDefinition          TraPPe
                TranslationProbability      0.5
                RotationProbability         0.5
                ReinsertionProbability      0.5
                SwapProbability             1
                CreateNumberOfMolecules     0
```

Source: Addapted from Dubbeldam, D. et al. (2015)

The first part of the code, declared in Figure 9, follows the definition of the simulation type, number of cycles, and initiation cycles. The printing definitions are essential since the RASPA software output is presented in a report format. Therefore, defining the printing parameters is crucial since it is related to the amount of information the software reports.

Figure 10 – Initial features for GCMC simulation - simulation inputs for RASPA2 software "simulation.input" file

```
SimulationType                  MonteCarlo
NumberOfCycles                  25000
NumberOfInitializationCycles    5000
PrintEvery                      5000
PrintPropertiesEvery            5000

Forcefield                      GenericMOFs
UseChargesFromCIFFile           yes
```

Source: Addapted from Dubbeldam, D. et Al. (2015)

In the same frame, going further to what is presented in Figure10, the force field and charges are specified. Further, the framework has to be embodied in the code. Some unit cells regard a cubic constitution of the final framework super-cell, where adsorption will be evaluated. Therefore, one needs to understand that a definition of 1x1x1 represents a super-cell that has 8 replicates of the original “framework.cif” file. In Appendix A, Figure 41 illustrates super-cell computation as the number of cells increases. Finally, those definitions are specified closely in Figure 11”.

Figure 11 – Framework and isotherm equilibrium points definitions - simulation inputs for RASPA2 software "simulation.input" file

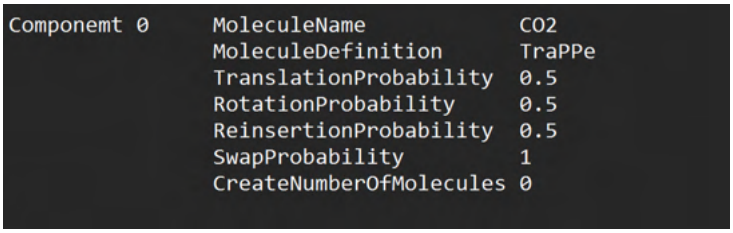
```
Framework       0
FrameworkName   Cu-BTC
UnitCells       1 1 1
HeliumVoidFraction 0.77390
ExternalTermperature 313.00
ExternalPressure 1.0e4 1.0e5 1.0e6
```

Source: Addapted from Dubbeldam, D. et al. (2015)

Lastly, the adsorbate is defined. The adsorbate is called by its file name (CO2.def), alongside the force field of its structure definition. Motion definitions are settled in the same snipped code - variables directly related to the simulation type. These variables’ probability values were studied and balanced with the number of cycles. The more

cycles the simulation has, the more efficient the convergence, although the time cost gets higher. Incrementing the probability of adsorbate motion diminishes time but raises biases, forcing the addition of initialization cycles. Small probabilities inflict late convergence, forcing the increment of cycles as well. Figure 12 indicates those definitions declaration.

Figure 12 – Adsorbate features defintions - simulation inputs for RASPA2 software "simulation.input" file



Source: Addapted from Dubbeldam, D. et al. (2015)

3.1.2 Molecular simulation evaluation

To evaluate the molecular simulation for adsorption, swap probability has particular relevance since the swap move acceptance enforces a chemical equilibrium between the system and the adsorbate (DUBBELDAM et al., 2016; HOLLINGSWORTH, Scott A.; DROR, Ron O., 2018b). It can be understood since adsorption modeling is done over a framework computationally represented in a set of cells (unit-cells), also regarding the adsorbate motion through the unit cells where an imaginary reservoir surrounds the computational system. Then, to control the addition and deletion of those molecules within the system, when an individual molecule is close to the edges of a unitary cell, it is deleted from that side and added to the other side of the cell, representing a constant number of molecules in the system towards the chemical equilibrium. The swap probability, then, considers the chance of that molecule being deleted or added since a good agreement between swap addition and swap deletion represents a good performance of the Monte Carlo simulation. For every simulation, swap deletion and addition were evaluated.

Another factor to validate the isotherm physical consistency is comparing the *in silico* experimental data representative of the present work with the previous literature. Briefly, isotherms were compared by statistical indicators, the mean squared error (MSE), residual mean squared error (RMSE) and relative residual mean squared error (RRMSE), and, finally, correlation coefficient, R^2. All are described by the following

equations, where n is the number of samples.

$$\text{MSE} = \frac{1}{\text{n}} \sum_{i=1}^{n} (x_{predicted,i} - x_{observed,i})^2 \tag{30}$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{\left(x_{predicted,i} - x_{observed,i}\right)^2}{n}} \tag{31}$$

$$\text{R-RMSE} = \frac{\sqrt{\sum_{i=1}^{n} \frac{\left(x_{predicted,i} - x_{observed,i}\right)^2}{n}}}{\sigma_{\text{observed}}} \tag{32}$$

$$\text{R}^2 = 1 - \frac{\sum_i (x_{\text{predicted,i}} - \hat{x}_{\text{predicted}})^2}{\sum_i (x_{\text{predicted,i}} - \bar{x}_{\text{observed,i}})^2} \tag{33}$$

3.2 BREAKTHROUGH CURVE SIMULATIONS – SECOND BRANCH

The breakthrough curve model solving was approached through the finite difference approximation (FDA) technique with centered differences, implemented in the MATLAB software, R2021 v.9.10.0, 64-bit, R2021a. As a solver method, the ODE23s were used, presenting good convergence.

The MATLAB code follows a simple structure where, in its first part, the variables of the porosity, length of the bed, and particle radius are declared alongside isotherm parameters – those are macroscale inputs. After, the time step and length step are declared. Concerning initial conditions, a set of matrices is stated. Matrices for the performance indicators to be extracted from the solved system's final result are also specified. The solver is called within a loop related to every joint of Pressure and Temperature settled for the BKC calculation, also used to calculate the coefficients needed for the BKC. Lastly, the discretized function for the breakthrough curve is defined and incorporated into the software, being discretized in a length step of 100 points and a time length of 40.000 points, representing Δz = 0.15 cm and Δt = 0.36 s.

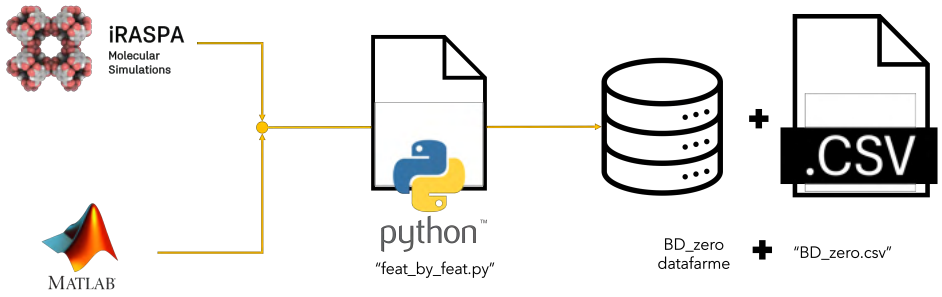
The model validation, though, was performed by the comparison between literature data and data elaborated by the above process. The same statistical indicators mentioned before (MS, RMSE, RRMSE, and R^2) used in the previous subsection, measured the accuracy between literature data and experimental data developed in the former work.

3.3 MACHINE LEARNING ALGORITHMS AND DATAFRAMING – THIRD BRENCH

3.3.1 Data wrangling and dataset assembling

Once molecular simulations and BKC simulations are established, the dataset development is prosecuted. This step of the methodology refers to a data-wrangling approach. It involves manipulating and restructuring data to make it more suitable and understandable for further exploration and analysis, as well as an organizational purpose. Data were assembled by a Python code developed uniquely for the present work. That code was entitled *feat-by-feat*, referring to the extraction of every feature of the RASPA Monte Carlo simulations report alongside the joint of those data with MATLAB simulation by the keys of Pressure, Temperature, and Material. The outputs from both software are read by the Python code, where a data frame variable is written, entitled *BDzero*. The assembling is done by the joint of the macroscale indicators with nanoscale simulations. The steps of this approach are illustrated in Figure 13.

Figure 13 – Schematic representation of the assembling of RASPA software outputs with MATLAB software outputs



Source: Author (2024)

Moreover, developing the *BD-zero* has only a scientific and didactic purpose. For practical aspects, referring to developing a single application on a unique code, the *BD-zero* does not need to be registered, since data follow straight to further methodology steps.

Once data are extracted and assembled, a crucial step must be established and justified by the number of instances related to the simulations. One may notice that the number of points for the isotherms is the same number of points for every target of the BK curve simulations - the number of instances determined by the keys of Temperature and Pressure. A dataset's data quantity is crucial for a machine learning application, being, at this point, a small dataset. Since with more data coming from a molecular simulation more computational power and time are needed, the methodology proposed does not make itself feasible. Hence, to overcome that obstacle, the "BDzero" is enhanced by smoothing its data concerning its keys (Pressure, Temperature, and Material).

Regressions serve as the smoothing process for every feature concerning the row of pressures within the specified temperature and material. This process involves applying the natural logarithm to the original data when necessary. Following the development of the *BD-zero*, the methodology proceeds with a series of regressions for each feature. Subsequently, the best-fitted equations are applied to create an enhanced dataset, termed *BD-MIP*. Figure 14 illustrates this process.

The set of equations used in the fitting process is presented in Equations (34), (35), and (36), where indexes *g*, *p*, and *l* stand for *sigmoidal*, *polynomial*, and *logarithm*, respectively. The *x* value refers to the independent variable used—in the present work, pressure—while the *y* variable represents the dependent variable fitted, corresponding to each variable listed in Table 1, except for Pressure and Temperature.

$$y = \alpha_g \left(1 - \exp\left[-\beta_g x^{\gamma_g}\right]\right) \tag{34}$$

$$y = \alpha_p x^2 + \beta_p x + \gamma_p \tag{35}$$

$$y = \alpha_l \ln x + \beta_l \tag{36}$$

The initial dataset comprises 17 points of pressure, equally distributed across 3 temperatures and 3 materials, resulting in a total of 153 data points. Each unique combination of Temperature/Pressure/Material yields 3 corresponding outputs: TBK, TC, and TS. The resulting fittings, as described, allow for working with as much data as desired. In the present case, the data smoothed concerning a fitting alongside pressure,

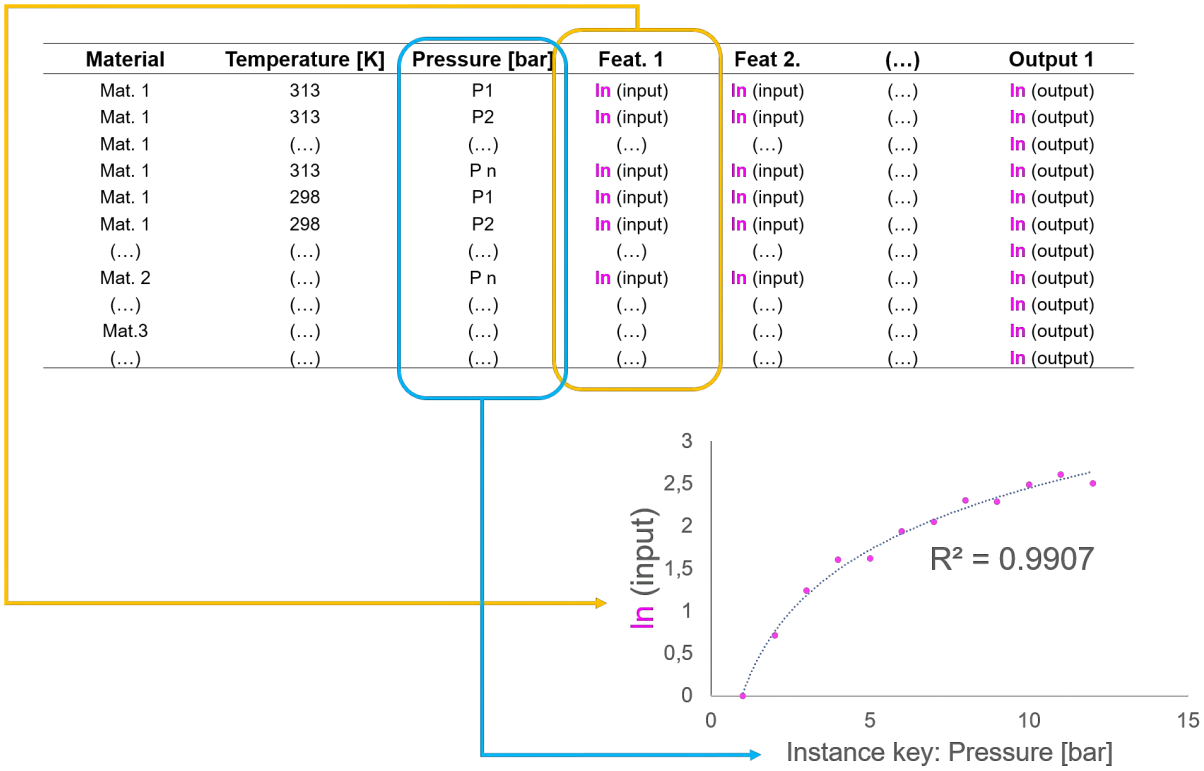
resulting in a final dataset 30 times larger than the initial one. For more detailed information about the validation, those are presented in Chapter 4, while additional information can be found on appendixes.

It is important to elucidate the reason behind the utilization of the above equations. This process revolves around data smoothing, aiming for a precise and as accurate as possible representation of the *in silico* generated data, without loss of original information. These equations have been selected for their ability to best quantitatively describe the data with minimal computational resources. Their primary purpose is to augment the quantity of data inputted into the Machine Learning Algorithms, flitting the learning process. This step of the methodology and these fittings do not intend to provide a physical description of the data. Instead, their sole objective is to ensure reproducibility within the domain of original data, akin to how $y_1 = \cos(0.5x)$ fits almost perfectly the second order polynomial $y_2 = -0.125x^2 + 1$ within the domain of -0.5 until 0.5, the same follows for the equations above within the *in silico* data.

This process is devolved by a series of fittings by the package NumPy v. 1.22.4 within the Python Language, complemented by the Pandas package v.1.5.3 for the joint processing of data from macro and nanoscale. All procedures implemented in Python were carried out with the IDE PyCharm Community edition v. 2021.2.3 and Python v. 3.10.8. By describing each keyed feature accurately, the best-fitted equations allow one to expand the amount of data present in the *BD-zero* to design the *BD-MIP*, thereby providing a final dataset suitable for applying the ML algorithm and evaluating its performance.

Moving on to the data processing part, it includes data scaling, which was developed using the scikit-learn StandardScaler. Here, the mean and standard deviation are utilized to standardize all data to a normal distribution, normalizing them to a mean of zero and a standard deviation of one. By default, each feature was scaled individually, despite being the same input data for each output. It's important to emphasize that the dataset was designed with the interpretability of the dedicated model in mind, allowing one to verify the relevance of features for each output through the lens of XAI models. The scaling of the outputs is also performed using the same method.

Figure 14 – Data wrangling and fitting representation leading to dataset enhancement

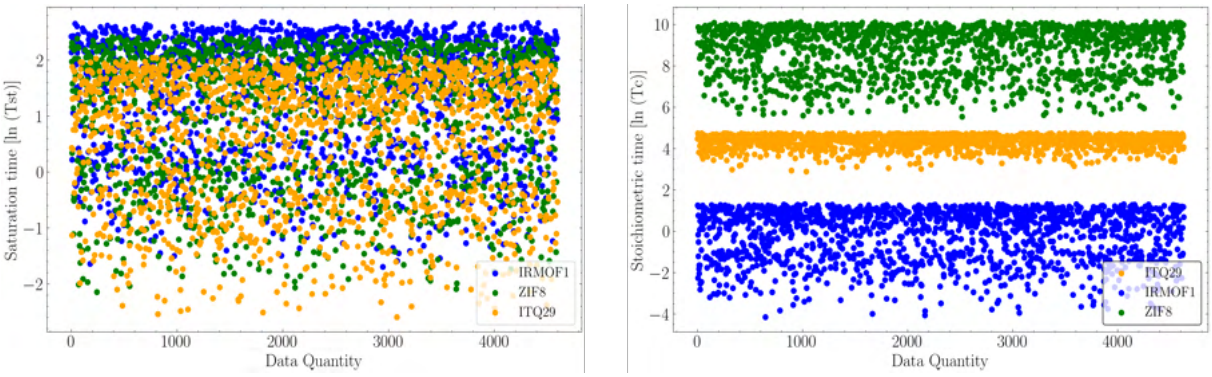


Source: Author (2024)

One should be aware that the stoichiometric time will appear differently than saturation time and breakthrough time, and, at this point (after scaling) those will present different ranges not following the trend that breakthrough curve naturally imposes (TS > TBK > TC). The Figure 15 illustrates how all outputs behave on the same scale. It is important to enphasize that, despite the procedure used to treat the data, a strong displacement of all data can be seen with saturation time.

Stoichiometric time, on the other hand, present a more dispersed display, ranging from -4 to +10 in logarithmic time units. Also, the same output variable exhibits a clustered distribution, primarily explained by the type of materials used. That trace maintains itself within the data alongside all the data treatment procedures - in another words, the data cleaning, standardization and logarithmic transformation does not erase the differentiation that Stoichiometric time has inherently. The following plot highlights that all materials utilized in the present work have distinct characteristics within an adsorption fixed bed, but it also underscores the significance of the TC indicator as a key factor in differentiating materials in terms of performance.

Figure 15 – Data inputs traits: Comparassion between Saturation time and Stoichiometric time.



Source: Author (2024)

3.3.2 Machine Learning Algorithms application and evaluation

Before implementing ML algorithms, the dataset was split into training and testing sets. A common approach was used, with 75% of the data dedicated to training and 25% to testing or validation. This resulted in 3,442 instances for training and 1,148 instances for validation. Considering all the 12 features used as inputs, the training dataset consisted of 55,072 input data points, while the validation dataset consisted of 3,444 data points for the whole 3 performance indicator.

The development of ML algorithms was done in Python within the packages of Scikit-learn v. 1.2.2, Keras 2.12.0, and Scipy v.1.10.1. The MLA models used are artificial neural networks and random forest, as stated in the theoretical background. Each model was enhanced by several approaches regarding its architecture to find the best structure. The statistical evaluation for MLA was done considering MS, RMSE, MAE, and R^2 as well.

Table 3 – RF hyperparameters range tunning

Hyperparameters	**Range**	**Hyperparameter concept**
n_estimators	200 to 2000	Number of trees in RF
max_features	auto' and 'sqrt'	Number of features to consider at splitting
max_depth	10 to 110	Max. number of levels in tree
min_samples_split	2, 5, and 10	Min. number of samples to split a node
min_samples_leaf	1, 2, and 4	Min. number of samples at each leaf node
bootstrap	True and False	Method of selecting samples for training trees

Source: Author (2024)

The hyperparameters boosting was done in different ways for each algorithm

applied, due to their natures. The hyperparameters tunning for RF was done by a grid search, which parameters range are present in Table 3. ANNs hyperparameters tunning was done by the variation of three main hyperparameters, i.e., batch size, number of epochs, and activation function, within an architecture pre-defined by the author. The specifications of the ANN hyperparameters is presented in Table 4.

Models interpretation were evaluated using the expected marginal contribution of each feature, throughout the SHAP-XAI approach. The SHAP version used was v. 0.41.0. The application of SHAP analysis is made towards the absolute feature's contribution to both models in the present work: RF and ANNs.

Table 4 – ANN hyperparameters tunning ranges

Hyperparameters	**Range**	**Hyperparameter concept**
Activation func.	ReLu, Sigmoid	Activation function on every nueron in the ANN
Batch size [bz]	10, 50, 90	Amount of data of every batch
Epochs	10, 30, 50, 70, 90	Times batches will be used for ANN trainning

Source: Author (2024)

4 RESULTS AND DISCUSSION

4.1 MOLECULAR SIMULATIONS – FIRST BRANCH RESULTS

GCMC simulations were performed and compared with literature data. All the simulations performed in the present study were developed for a single-component adsorption system. By verifying the simulation's accuracy, one can follow the development of the thermodynamic equilibrium properties dataset and the deterministic modeling of the BKC. Well-accurate simulations were developed under the same conditions as the literature used as references. Therefore, the conditions (temperature and pressure) are listed alongside the references in Table 5.

Table 5 – Statistical indicators from isotherms simulations against literature for all materials simulated

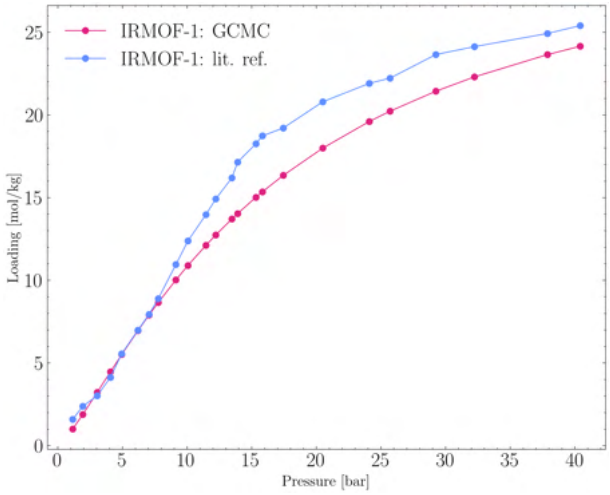
Statistic	**ITQ-29**	**ZIF 8**	**IRMOF-1**
MSE	0.111	0.050	3.609
RMSE	0.333	0.224	1.899
R-RMSE	0.083	0.024	0.040
R^2	0.998	0.998	0.986
Conditions	313 K	313 K	313 K
Reference	(MARTIN-CALVO et al., 2018)	(SAEEDIRAD et al., 2020)	(BABARAO et al., 2007b)

The isotherms developed by *in silico* experiments were compared with mean squared error (MSE), root mean square error (RMSE, relative root mean square error (RRMSE), and coefficient of determination (R^2) with literature data. Average performance indicators show a good agreement between *in silico* experimentation and references since MSE averages 1.0955 mol/kg. Root mean square error averages 0.8588 mol/kg, which equals 6.633% of relative error in non-dimensional terms.

In addition, R^2 was calculated as 0.994. The results were interpreted as an excellent overall performance of the experimentations. All performance indicators for every material developed are presented in Table 5. One can have a visual intuition from the molecular simulation accuracy within Figures 16, Figure 17, and Figure 18, at the end of this section, respectively related to IRMOF-1, ZIF-8 and ITQ-29 frameworks.

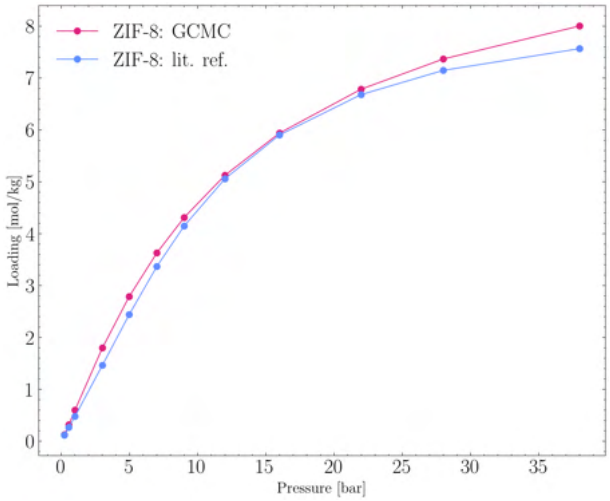
ITQ-29 molecular simulation presented an increasing deviation from the pressure of 2 bar, indicating a divergence from the representation of the system. Equivalent phenomena follow IRMOF-1 *in silico* experiments, although in the last one with a minor degree. The leading cause is correcting the system's pressure to the fugacity term in

Figure 16 – Comparative between GCMC isotherm simulation agains literature reference data for IRMOF-1, ZIF-8 and ITQ-29



Source: Author (2024)

Figure 17 – Comparative between GCMC isotherm simulation agains literature reference data for IRMOF-1, ZIF-8 and ITQ-29

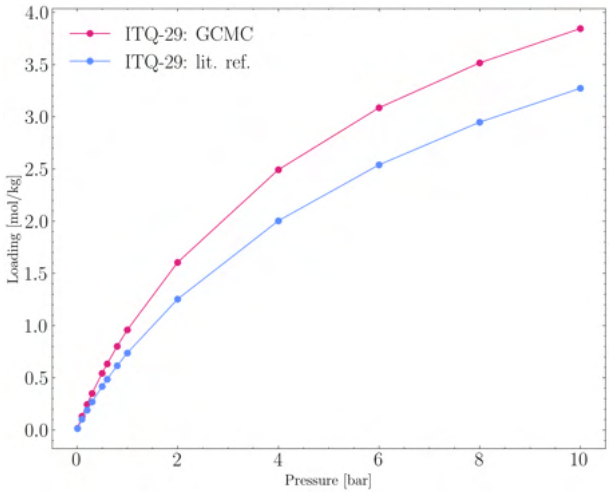


Source: Author (2024)

the Peng-Robinson equation. For small pressures, the relation between those will be close to one, and as closer to one, the better. However, as pressure increases, the ratio between those parameters diverges from the unit, causing the deviation between reference data and in silica experimentation.

For the sake of exemplification, Figure 19 presents the decaying of the fugacity coefficient with the increasing pressure in molecular simulations of ITQ-29 at 313 K.

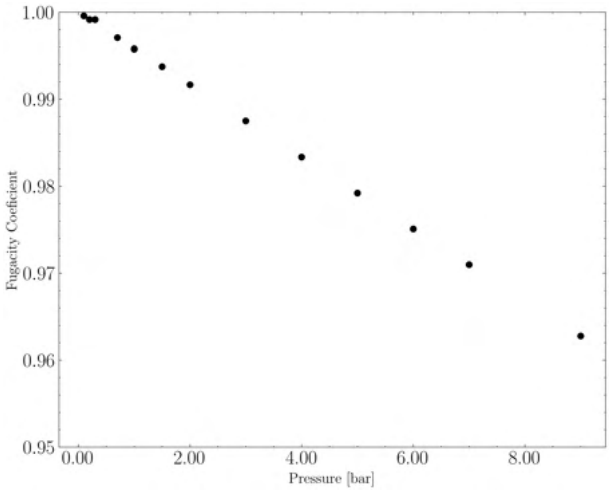
Figure 18 – Comparative between GCMC isotherm simulation agains literature reference data for IRMOF-1, ZIF-8 and ITQ-29



Source: Author (2024)

Despite this points, ZIF-8 presented a good agreement with references data at high pressures, the last one with a relative RMSE of 2.4363 %. Those parameters indicate that force fields are a relevant barrier when one aims to have a good agreement between simulation and experimental data. One of the biggest obstacles of the present work is a force field that can correctly describe the interactions at the nanoscale regarding adsorption equilibria.

Figure 19 – Fugacity coefficient decaying over pressure increase



Source: Author (2024)

Regarding the quality of the simulations within the software, SWAP performance

is an important indicator to be evaluated (DUBBELDAM et al., 2016). Considering all simulations developed, the difference between swept deletion and swap insertion averaged 0.0785%, allowing one to conclude that the simulations were performed correctly. One can see the graphical information about the Table 5.

4.2 MACROSCALE SIMULATIONS – SECOND BRANCH RESULTS

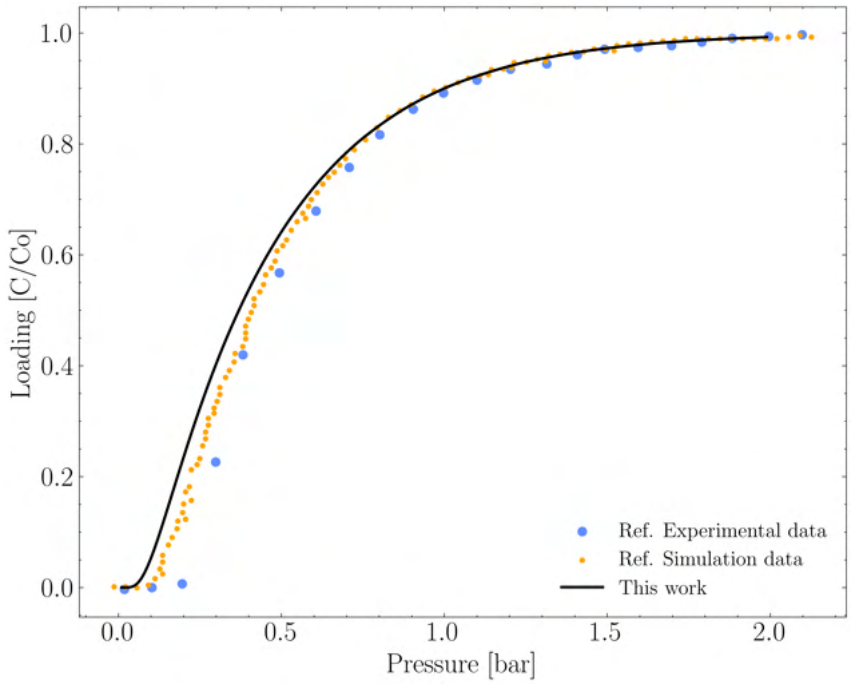
To perform macroscale simulations, the work developed by Sabouni and coworkers (SABOUNI; KAZEMIAN; ROHANI, 2013) was reproduced and validated. Moreover, due to its validated operating system, the work of Sabouni ((SABOUNI; KAZEMIAN; ROHANI, 2013) is used as a reference, where its data are developed with the frameworks studied here. The macroscale physical system that machine learning models connect with the nanoscale is the one described by the reference mentioned. That is a crucial factor associated with the current project since machine learning models were trained to predict targets for that specific system, hence, a physical system restricted.

The deterministic modeling of the BKC of CO_2 adsorption was developed in MATLAB, as previously addressed in the Methodology section. The macroscale simulation was evaluated with the same statistical performance indicators used in the nanoscale. The results generated by the methodology and developed by the author were analyzed with the experimental data from the references cited throughout the text. For the sake of clarity, experimental data from the reference article - the data developed by the experimental procedure - are addressed as *Ref. Simulation data*. The data gathered from the simulation developed by the reference article, are addressed after *Ref. Simulation data* (SABOUNI; KAZEMIAN; ROHANI, 2013). The data reproduced data, developed by the author of the present work, are addressed as *This work*. Those references are presented in the Figure 20.

By analyzing those, the results found were satisfactory regarding the model representation of the experimental data, presenting an RMSE of 0.0565 and an MS minor than 0.0032 of adsorbate loading. The RRMSE of 1.260% and an R^2 of 0.9864 also validated the agreement. Regarding the Simulation's reproducibility, the relative mean square error was 0.0433, and an MS of 0.0019 of adsorbate loading. In Figure 20 data agreements are visually represented.

The usage of dimensionless groups allowed a better description of the system on a macro-scale. As Zhao and coworkers developed - (ZHAO et al., 2021), the usage of dimensionless terms reduces the complexity of the system and provides a more comprehensive view of the importance of each term. The success of the simulations

Figure 20 – Comparative between this work fixed-bed simulation with experimental and simulated data from reference



Source: Author (2024)

is related to the dimensionlessness of the breakthrough curve model. Even being a simplistic system of equations, the dimensional analysis terms allowed the absence of stiff problems, for instance, in more extreme cases (high pressure). Stiff problems were verified with the dimensional model.

Another technical aspect should be highlighted. Determining the data set that will be used to train and validate the machine learning regression algorithms depends on its structuralism. A non- or semi-structured data set would severely decrease the algorithms‘ regression capability. Therefore, leaving the methodology more complex and passive of biases.

The dimensionless model and all the correlations presented in the theoretical background were built based on temperature and pressure variances, allowing one to connect the adsorption isotherm with the breakthrough curve in a structured data set.

4.3 SMOOTHING AND DATA WRANGLING PROCEDURES

The results of the section “Data extraction, assembling and manipulation” are presented in the current section, especially detailing the smoothing procedure. That

procedure, also entitled dataset enhancement, is the principal step that upgrades the "BDzero" to the "BDMIP". Each feature, for every isotherm of the materials ITQ-29, IRMOF-1, and ZIF-8, was transformed by the mathematical operator of the natural logarithm to then be fitted in a set of equations previously presented in the Methodology section in Equations (34), (35), and (36).

The regression presented a good agreement with the experimental data for all features, with an average determination coefficient of 0.9922, and a standard deviation of 0.00728. The variation coefficient is then measured, with the order of 0.73% (a good indication of the fitting procedure to all features). For every fitting, the MS and RMSE indicators were calculated, also reinforcing a good overall fitting. The average statistical indicators of every feature, from each material's isotherms, are presented in the Table 6.

Table 6 – Average smoothing statistical performance for each feature

Feature	**MSE**	**RMSE**	**R^2**
Average host-adsorbate energy VdW [U]	0.00775	0.06550	0.99664
Average host-adsorbate energy total [U]	0.01372	0.10114	0.99590
Average adsorbate-adsorbate energy Coulomb [U]	0.01267	0.09815	0.99595
Average adsorbate-adsorbate energy VdW [U]	0.03396	0.17439	0.99394
Average adsorbate-adsorbate energy total [U]	0.03373	0.17034	0.99630
ADCP - Average derivative of the chemical potential [U]	0.02901	0.14741	0.99725
Enthalpy of adsorption [U]	0.02541	0.14877	0.98666
Total energy [U]	0.01150	0.05205	0.99251
Average heat capacity [cal/mol/K]	0.10003	0.31627	0.97913
Average heat capacity [U]	0.00642	0.07489	0.99831
Average density [cm^3/g]	0.14083	0.27919	0.97212
Average host-adsorbate energy Coulomb [U]	0.00044	0.01838	0.99983
TBK	0.00282	0.05125	0.99211
TC	0.01260	0.14962	0.99235
TS	0.00779	0.08652	0.99322

Source: Author (2024)

4.4 MACHINE LEARNING DEVELOPMENT AND APPLICATION PERFORMANCE

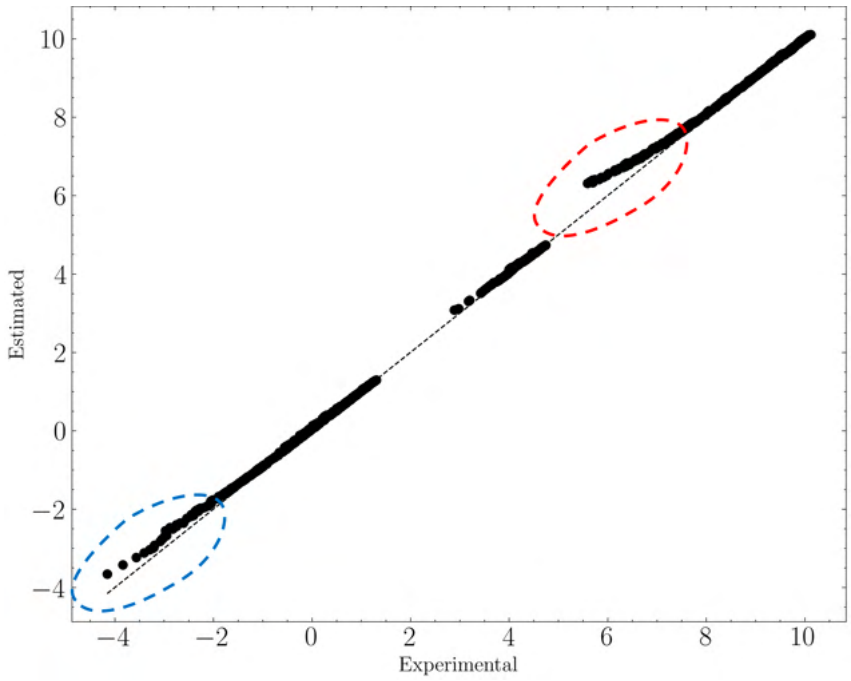
The RF and ANN model enhancement has the same criteria: hyperparameters variation to find the best model and training data. Since both algorithms have different hyperparameters, firstly, they will be presented with RF results based on the parameters grid random search. The following will present an MLP neural network architecture optimization, analyzing the activation function, the number of epochs, and the batch size impact. Both algorithms were improved based on the stochiometric time. For the

saturation and breakthrough time, optimum hyperparameters are applied directly.

4.4.1 Random forest model fitting and performance evaluation

The training set's best-fitting hyperparameters are presented in Table 3. Regarding TC, With an MAE of order 0.00296 and RMSE of equal to 0.00357, it is understood that the model fits well with training data. However, R^2 and MSE indicate overfitting, presenting the values of 0.99999 and 0.00001, respectively. Even though, when applying the model over the test dataset, a good fitting is verified, not indicating overfitting biases. Performance indicators vary slightly regarding MAE (0.06087) and RMSE (0.12188). The MSE, with a value of 0.01486 and an R^2 of 0.99897, indicates an excellent fitting alongside untrained data. Therefore, it allows to conclude that for the case of TC, RF forest has a good performance.

Figure 21 – Random forest fitting for stoichiometric time

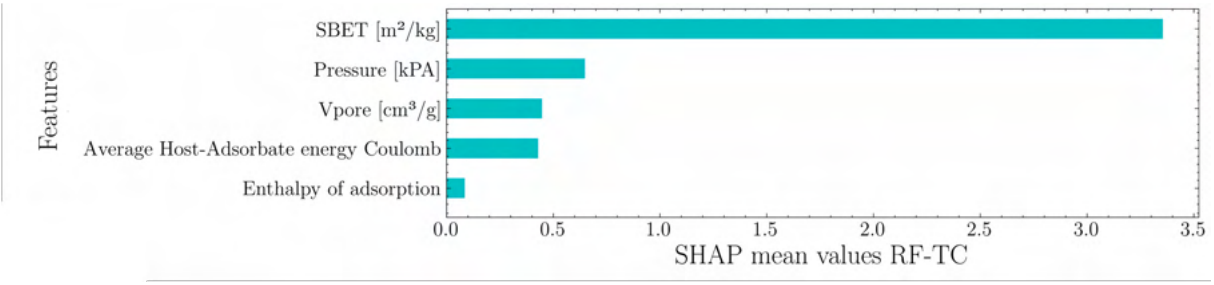


Source: Author (2024)

A few points should be emphasized, however. The above figure (Figure 21) illustrates the forecast efficiency of the RFalgorithm for the TC indicator. It indicates a tendency in the range of experimental data from 5.5 to 6.5 (red circle), where a slight drift of the fitting is verified. A slight curve tendency is also present in the values range from -4.5 to -2 (blue circle).

The cause of this drift was directly correlated to the dataset constitution alongside the model nature, and data discontinuity is one of the probes related to those phenomena. The knots on every tree within the Random Forest might have criteria capable of describing those values with local data continuity, or dataset features are not being used as a whole. Then, analyzing the variables necessary for the criteria development within the Random Forest, it is verified by the SHAP values distribution that S_{BET}, pressure, and V_{pore}/H-A Coulomb are the main variables for the TC-RF model. With a mean SHAP value of 3.4, 0.65, and 0.45, respectively, it is indicated that the model does not use all data intelligence to do a suitable fitting. Figure 22 emphasizes that statement.

Figure 22 – SHAP mean values for RF-TC fitting

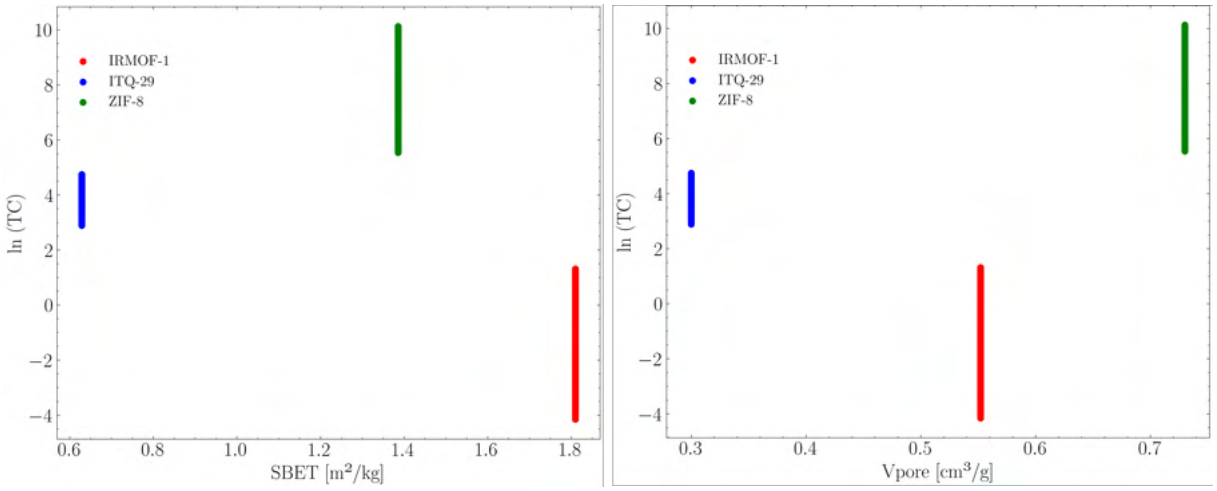


Source: Author (2024)

Another crucial aspect is given by data distribution between materials alongside the datasets. In the case of TC specifically, each range of data represents a specific material, a fact that does not follow other performance indicators. From that perspective, the overstated importance of the S_{BET} data for the discretization of the dataset is verified and, consequently, affects the unbalanced RF heuristics criteria. At the same time, since V_{pore} is also a categorical data that discretizes materials clearly, the importance that the feature has for the model is unexpected, accentuating the unbalanced weight given to S_{BET}. Not just the nature of the indicator, but also the categorical quality that V_{pore} and S_{BET} are possible causes for the drift. Figure 23 reinforces it.

Also, for the case of TC, a more balanced contribution regarding the features' dataset would enhance fitting quality. Host-Adsorbate Coulomb energy is present only in data associated with ITQ-29, for instance, due to the nature of the material adsorption process. At the same time, it is the only subset of data that does not present any drifting on its poles. These specific features differentiate this subset categorically from others, and RF heuristics take advantage of that and use it to predict data better. The remaining question is why the RF algorithm does not iterate that approach in the remaining features, regardless of being less stamped as Host-Adsorbate Coulomb Energy. A sure

Figure 23 – Textural features compared for IRMOF-1, ZIF-8 and ITQ-29 over the natural logarithm of TC



Source: Author (2024)

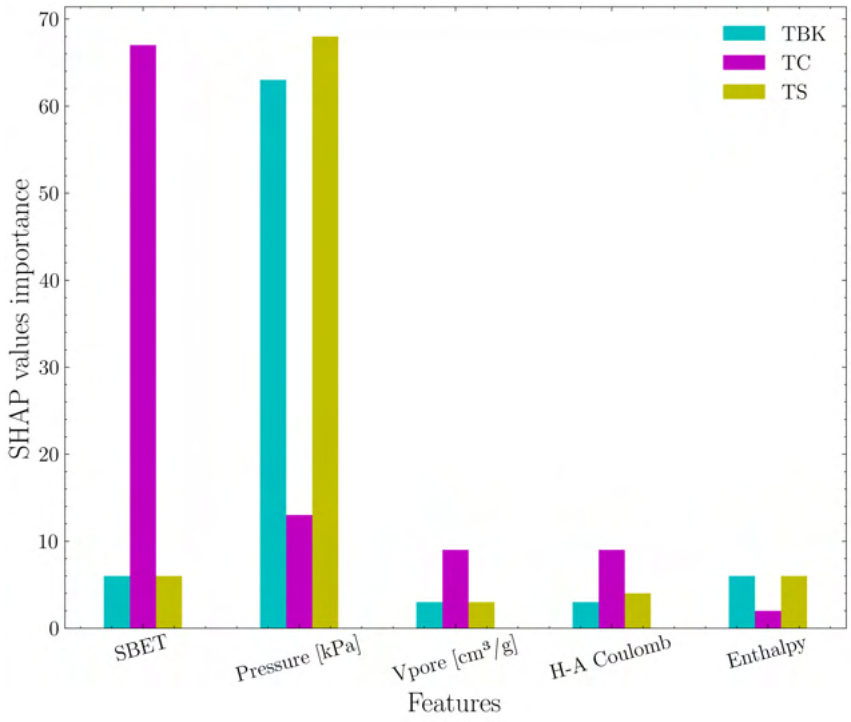
conclusion so far is that a better use of the whole dataset distributions would imply a better fitting for the test dataset.

Regarding pressure and the remaining features that have minor importance to the model, it would be interesting if the algorithm considers features such as van der Waals energy, heat capacity, and enthalpy of adsorption, for instance, which does not happen. Those features carry more of a blueprint of each material than pressure, which is common to all materials in the dataset. There is no differentiation between each material regarding pressure. Therefore, it is unexpected that pressure has more significance to the model than thermodynamic features.

Even with those considerations, the RF-TC case is representative of best fitting compared to other indicators. Breakthrough time and saturation time RF-models perform poorly considering the same hyperparameters applied to TC. For TS, even with relatively small MSE and RMSE, the fitting presents an R^2 of 0.98720. It was verified a pattern close to TC regarding the drift in the lower values range. TBK's tendency is analogous, with an MSE of 0.0009 and an RMSE of 0.00939 (significantly small values). In this case, R^2 is 0.99593. The specific and general results for every statistical indicator are presented in Table 7. Lastly, different hyperparameters were studied beyond the optimal ones regarding TS and TBK. No improvement was found.

Regardless of the performance related to TBK and TS, SHAP analysis was done to gather insights into the algorithm capability for the present application. The results

Figure 24 – Percentual relevance comparative of the SHAP mean values for RF fitting of TBK, TC and TS, based on TC



Source: Author (2024)

Table 7 – Random Forest total statistics for train and test dataset

		Statistical performance indicator			
Indicator	Dataset	MAE	MSE	RMSE	R^2
TBK	train	0.0004	0.0000	0.0001	1.0000
	test	0.0457	0.0009	0.0094	0.9959
TC	train	0.0609	0.0149	0.9990	0.9990
	test	0.10567	0.0149	0.1219	0.9990
TS	train	0.0589	0.0006	0.0036	0.9999
	test	0.6790	0.0547	0.1399	0.9872

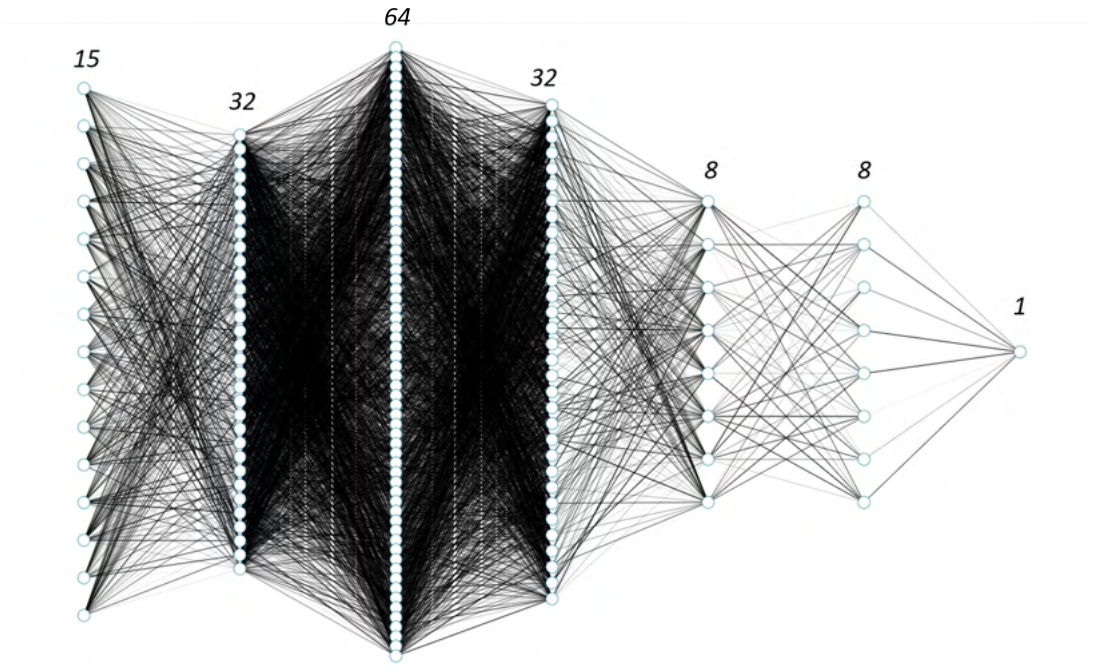
Source: Author (2024)

are synthesized in Figure 24, and complemented with Table 7. It emphasizes the difference between S_{BET} and pressure for TC and TBK or TS. While S_{BET} is the most relevant feature in the first, with 70% of the total summed SHAP values, in the other two, S_{BET} is placed in the fourth position, close to 7% of relevance. Pressure, on the other hand, while placed as the most relevant feature for TBK and TS, highlights a contrast between the informational gain that the RF algorithm considers. As stated, pressure is equally distributed for all materials and performance indicators, not solely representing a blueprint for each material or indicator.

4.4.2 ANN model fitting and performance evaluation

To develop the best model concerning an ANN application, this study focuses on the best combination of the activation function, batch size, and number of epochs and then, architechture (appendix A3). The architechure used on the present work is displayed on Figure 25 and is composed of 7 layers following the number of neurons of 15x32x64x32x8x8x1, being a Multiple-Inputs-Single-Output architecture (MISO). The specific hyperparameter variations are presented in the Table 4. Unlike the RF application, the best ANN model was not developed through a grid search. The best number of epochs and batch size combination were defined for both activation functions analyzed (ReLu and sigmoidal) to determine the best set of parameters. However, as RF, that part of the study is restricted to the forecast of the stoichiometric time and then replicated to other indicators (TBK and TS).

Figure 25 – ANN final architecture used to predict values

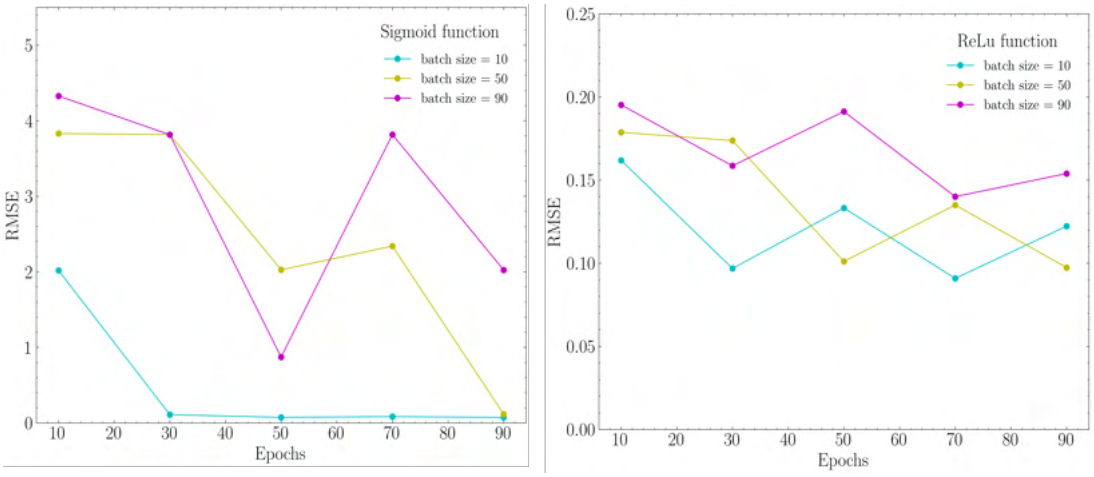


Source: Author (2024)

Following through, concerning the sigmoidal function, general performance was primarily poor, even presenting reasonable indicators for part of the batch/epochs set for big batch sizes. Hence, as expected, the smaller the batch size, the better the ANN fitting. The number of epochs follows the inverse trend since more epochs affect the optimizer performance, enabling it to converge towards a small error tolerance. Taking

as a pivot point the *Sigmoidal - 50 - 70* set from table 11, addressed in Appendix C, one can ascertain a non-convergence from the ANN model due to the lack of space for the optimizer to update the weights in its inner layers. From that standpoint, it is also possible to ascertain that the best configuration hyperparameters setting is a non-linear procedure since results that one could expect to have a better performance go otherwise.

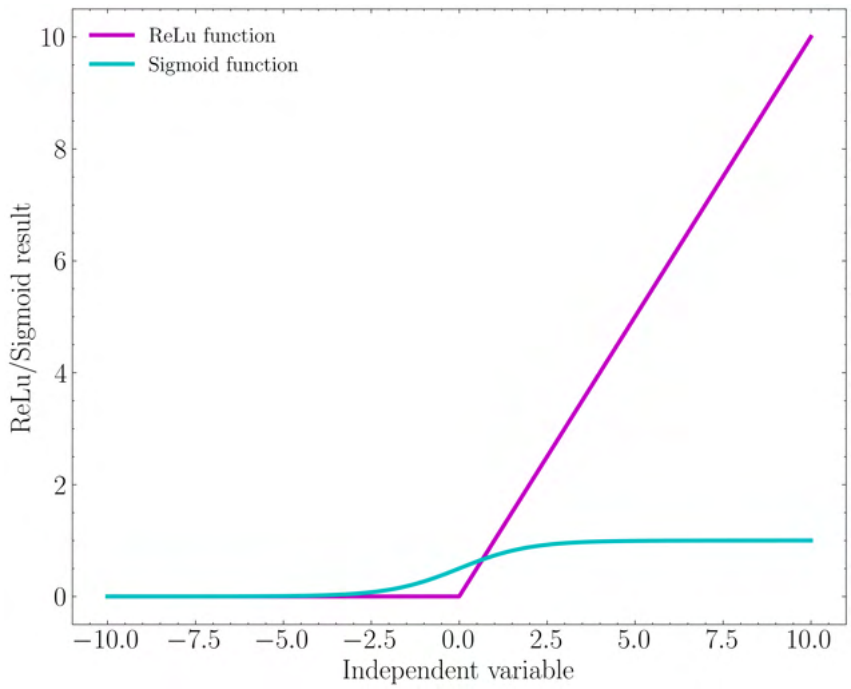
Figure 26 – Impact over the variation of epochs and batch size in RMSE statistical performance indicator



Source: Author (2024)

A good example is the set of *Sigmoidal - 50 - 50* and *Sigmoidal - 50 - 70*, sustaining that point. One can identify the overall trends for the sigmoidal function in Figure 26. The ReLu function better fits the test dataset, regardless of the batch/epoch set, converging in every combination studied (verify R^2 in the Table 11, in Appendix C). Consequently, it was chosen as the activation function for the inner layers alongside the best set of batch/epochs as 10/90.

Figure 27 – Comparative of the ReLu and Sigmoid function sensitiveness over a hypothetic independent variable



Source: Author (2024)

The ReLu function performs significantly better than the sigmoidal function mainly because of its sensitiveness regardless of the data input scale and transformation. Even with the sigmoidal function being smoother than ReLu function, the output for the sigmoid operation is always allocated in a range of 0 to 1. ReLu, on the other hand, is a linear function for values higher than zero, amplifying the importance of a neuron. Furthermore, because the natural logarithm was applied to the data previously, it is evident that the data's interpretation within the sigmoidal function was affected by the presence of the exponential term. Ultimately, this procedure performed by the sigmoidal function alongside the network will lead to neurons' deactivation since the net will transfer values less and less sensitive for the function in the next neuron. Figure 27 presents an illustrative perception of this issue, where sigmoidal and ReLu functions are compared for the same input range, providing different ranges of outputs. Tests with a linear activation function were developed but have not presented a performative convergency, with errors of 20% or more for the best parameters.

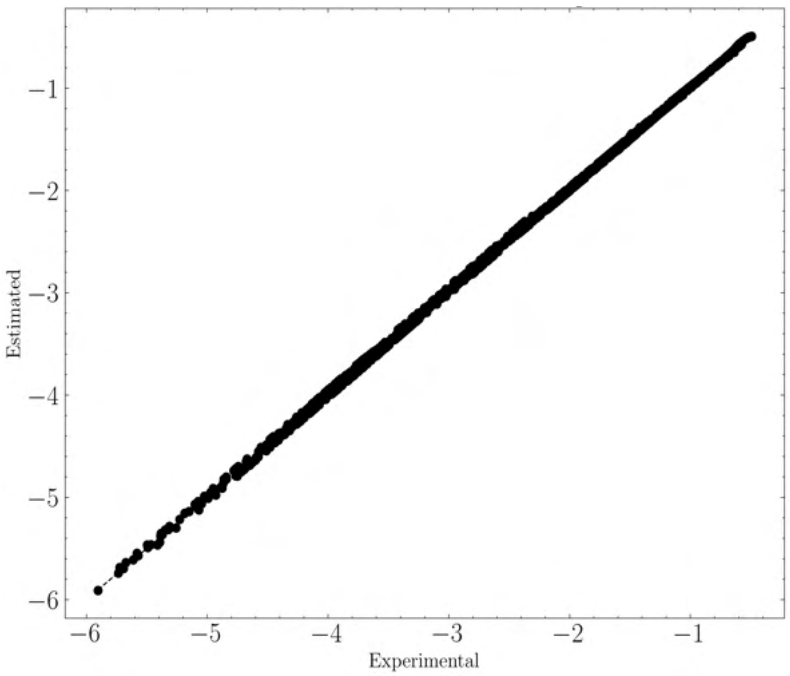
With the most relevant hyperparameters defined, the optimal ANN was applied to training dataset fitting and then test dataset forecast. Results were considered significantly better than RF applications and are individually presented in Table 8. Individually, one can verify the plotting of each one of the predicted and test data in Figures 28, 29,

Table 8 – ANN fitting between experimental and estimated values for TBK, TC and TS values

	MAE	**MSE**	**RMSE**	**R^2**
TBK	4.2285	0.0123	0.1111	0.9991
TC	0.7565	0.0005	0.0224	0.9996
TS	3.3548	0.0056	0.0288	0.9993

and 30. Generally, the forecast averaged an MSE of 0.0062 and an RMSE of 0.0541. The average R^2, regarding the proportion of the variance between experimental and estimated data, was 0.9993, considered a suitable fitting. Those performance indicators allow one to conclude that ANNs are better suited for multiscale linkage regarding CO_2 adsorption based on the presented methodology, which aligns with the universal approximation theorem (UAT) for ANNs.

Figure 28 – ANN fitting between experimental and estimated values for TBK



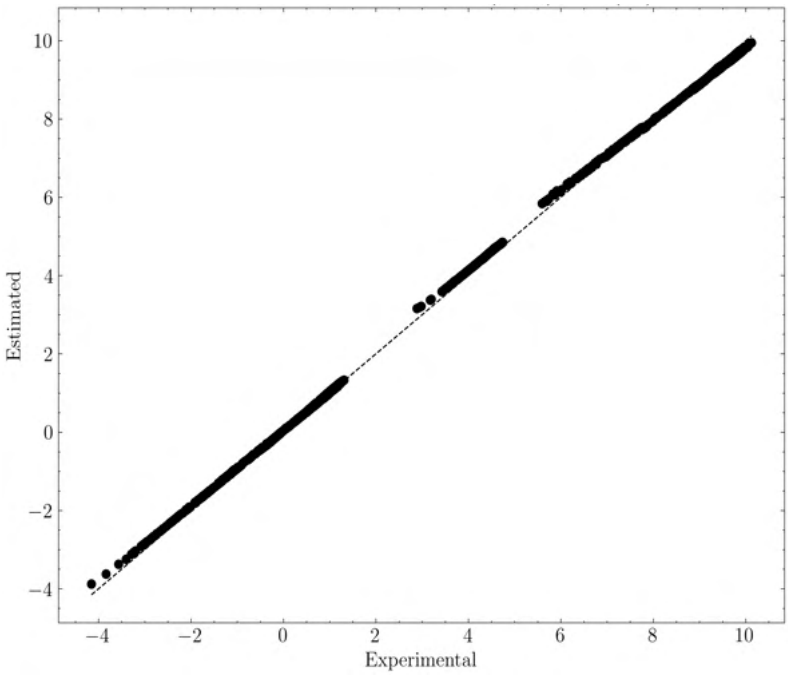
Source: Author (2024)

In front of those conclusions, SHAP evaluation on ANNs models is applied to comprehend the most informative data for the models forecast, alongside its physical meaning and physical coherence.

Concerning the time of breakthrough (TBK), the artificial neural network developed granted more distributed importance for each dataset feature, presenting a

smoother rank. The top three variables, present in the dataset are S_{BET}, Pressure, and Average Density. Even presenting S_{BET} as the most relevant feature of the model, one might comprehend that the ANN algorithm followed the same tendency as RF. Another aspect that is highlighted in the ANN-TBK case is that the top features are not, intrinsically, thermodynamic. S_{BET} can be categorized as a textural feature, while pressure is an intrinsic variable of the system. However, the average density of the system is a feature consequent to the mass adsorption process, mechanically related to mass transport phenomena.

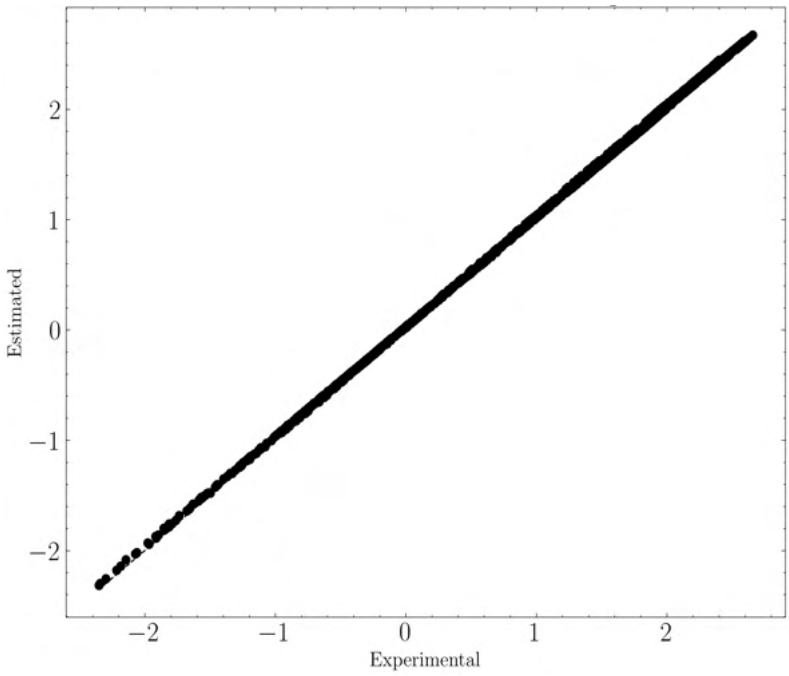
Figure 29 – ANN fitting between experimental and estimated values for TC



Source: Author (2024)

Specifically for those findings, there is agreement with previous literature, where the SHAP analysis was employed. Compared to temperature, S_{BET} and V_{pore} are assigned with more importance (positive importance) from the SHAP analysis, alongside the fact that S_{BET} has a synergic effect with Pressure for CO_2 adsorption for MOF's and Zeolites applications (LI, X. et al., 2023; OKELLO et al., 2023). Worth mentioning is the fact that the only literature where the methodology is comparable here presented in some degree, is applied with biogas adsorption, which is not just different within variables (e.g. pH), but also different regarding the nature of the process (e.g. liquid-solid adsorption). This is considered by the authors as evidence of the innovative aspect of the methodology as a whole, concerning CO_2 adsorption field (BANISHEIK-HOLESLAMI; QADERI, 2024).

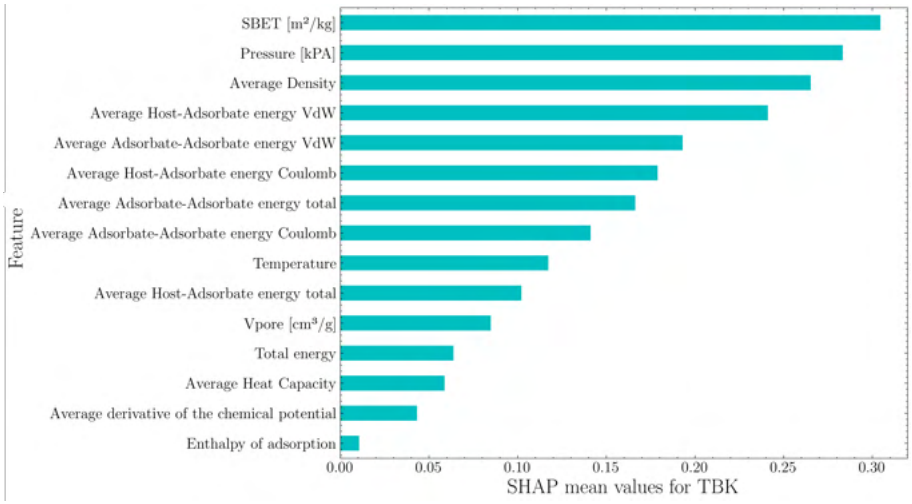
Figure 30 – ANN fitting between experimental and estimated values for TC



Source: Author (2024)

Giving sequence to the SHAP results analysis, another perspective should be added to this. Interpreting the meaning of the TBK indicator physically, the wider the area available for the adsorption to occur for the same volume, the later will be the measurement of the TBK. Since it (the material) has a more open area to the adsorption, regardless of being a monolayer or multilayer adsorption system, for a continuous inlet of CO_2, more time will be needed for the first spots to be fulfilled. Therefore, it has physical logic that S_{BET} is one of the top three most relevant features for the TBK indicator. The same follows pressure. The higher the pressure, the higher the TBK indicator. The physical logic is sustained when the adsorbate molecules are pressed into intrinsic pores, then optimally occupying the open pores of the system. The same logic follows the density of the system. Even not directly related to the superficial phenomena, the higher the mass amount in the framework volume, the higher the density. The more room the molecules have to settle on the adsorption process, the more postponed the breakthrough. Density carries more phenomenological characteristics of the process and will be dependent on the S_{BET} are of the material. The Figure 31 rank all the features.

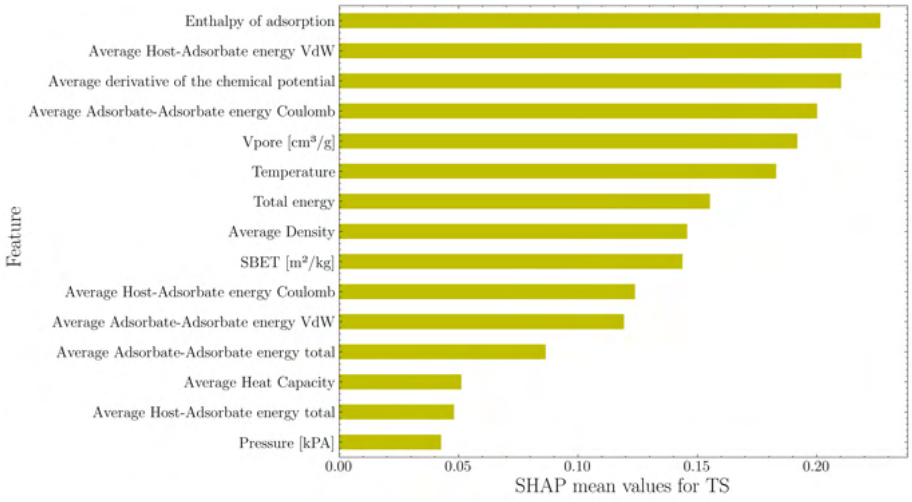
Figure 31 – SHAP mean values for ANN-TBK fitting



Source: Author (2024)

The saturation time indicator (TS) presents a different ranking for the SHAP analysis. Instead of S_{BET}, Pressure, or Average Density, the top five most relevant variables for the ANN-TS model are Enthalpy of Adsorption (0.23 SHAP), Average Host-Adsobate VDW energy (0.22 SHAP), ADCP (0.21 SHAP), Average Adsorbate-Adsorbate Coulomb energy (0.20 SHAP), and then, a textural property, V-pore (0.19 SHAP). The SHAP mean value is well distributed, although the most relevant features classes to the model deserve attention, as presented in the Figure 32.

Figure 32 – SHAP mean values for ANN-TS fitting

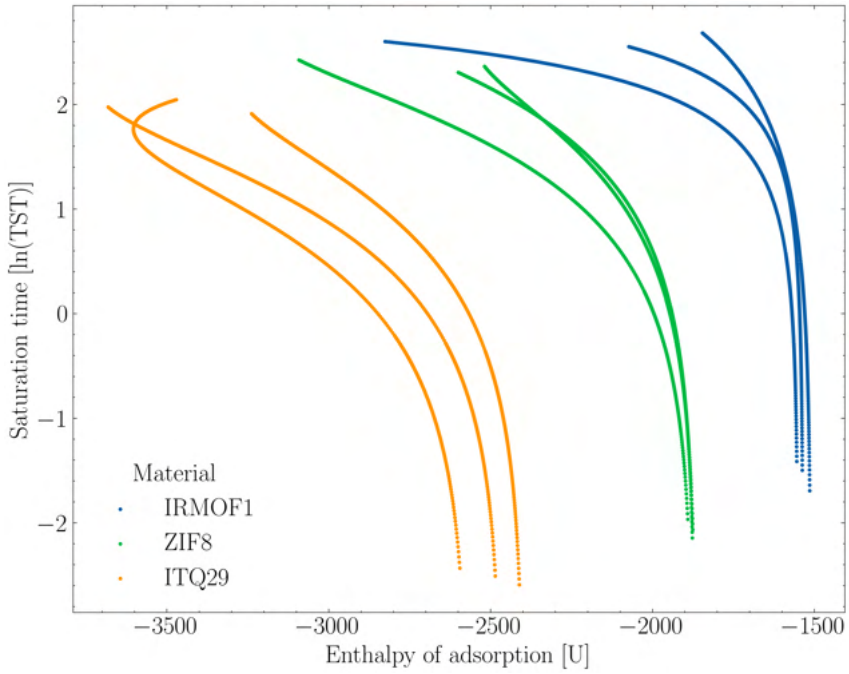


Source: Author (2024)

The enthalpy of adsorption can be physically interpreted as the affinity of a

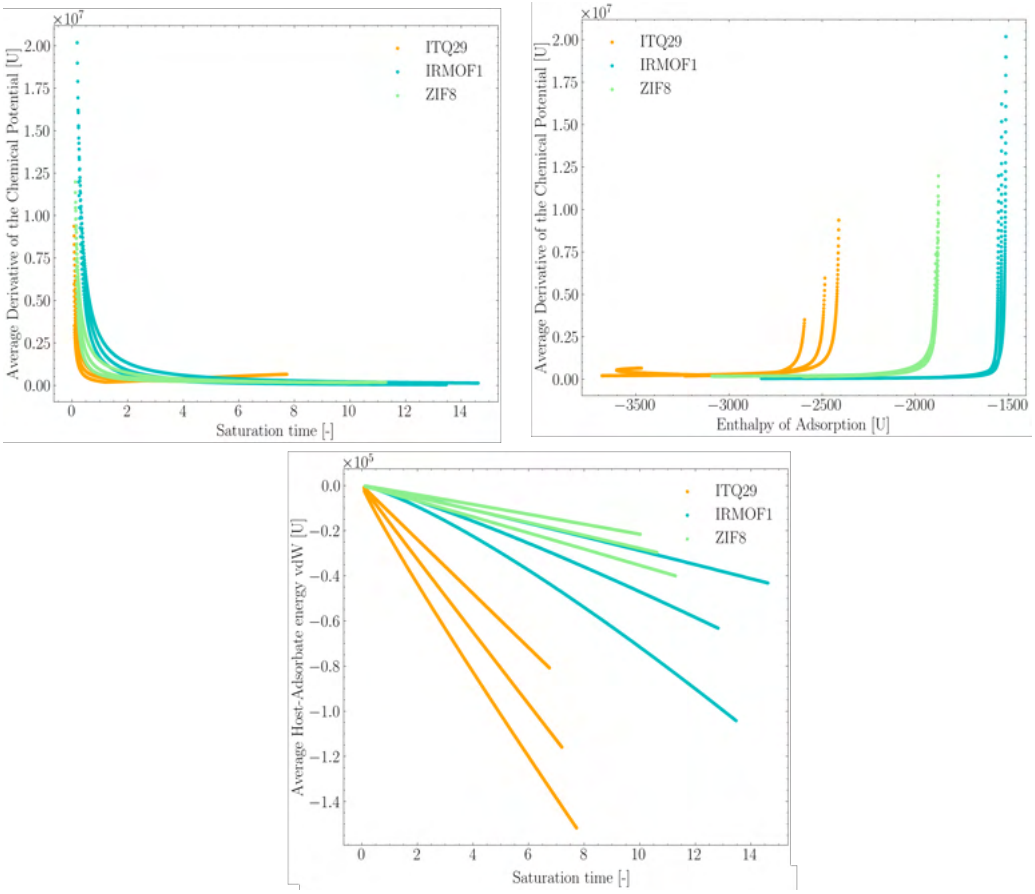
molecule with the framework, expressed as the binding energy. The stronger the binding, the higher the adsorption capacity. The relationship with saturation time follows that trend for each material individually. Hence, the stronger the binding, the higher the saturation time from the perspective of a unique material. However, the inverse trend follows when materials are compared. One can conclude it by comparing the ITQ-29 > ZIF-8 > IRMOF-1 trend from the perspective of saturation time versus enthalpy of adsorption. ITQ-29 has more vigorous bidding measured, although saturation time is lesser than other materials for all temperature cases. Hence, it can be concluded that, physically, the enthalpy of adsorption is essential to determine the saturation time once it carries more of a blueprint for the material since its *ab initio* properties of connection discretize the isotherm trend. Figure 33 summarizes the discussion above.

Figure 33 – Santuration time over Entalphy of Adsorption



Source: Author (2024)

Figure 34 – Cross-Correlation evidence from the Comparative of Average Host-Adsorbate vdW energy and ADCP over saturation time and ADCP sensitiveness over enthalpy od adsorption



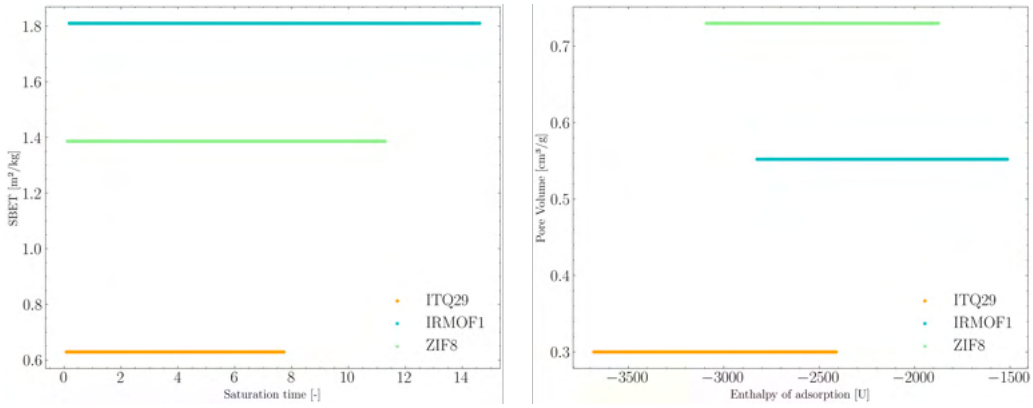
Source: Author (2024)

The ADCP and Average Host-Adsobate van der Walls energy follow the same enthalpy of adsorption prerogative. The Figure 34 illustrates both features. It is vital to emphasize that the distinction between ADCP of different materials and saturation time is subtle at first sight. However, ADCP has a strongly correlated behavior with the enthalpy of adsorption. This crossed correlation allows the model to discretize even more data towards a target. This is an important aspect concerning data intelligence for the model forecast capacity.

The most crucial aspect connecting all those features regards how those can accurately describe the saturation time physically, not just by the ANN model black-boxed. In light of the information presented above, it becomes evident that incorporating thermodynamic properties to describe the saturation time and textural properties to characterize the breakthrough time aligns with the significance attributed to the models. By comparing the most relevant features of ANN-TBK with ANN-TS, one can verify

the type of variables more relevant for each indicator. For example, in Figure 35 it is clear that the same trend regarding enthalpy of adsorption is not equivalent regarding pore volume. However, the textural property of the S_{BET} follows a direct relation with TS, even being the 9th most relevant feature for the ANN-TS model - the higher the S_{BET}, the higher the TS. Hence, the conclusion from that discussion is that sensitive temperature features are better descriptors for the saturation time for the fixed bed studied.

Figure 35 – Pore volume discontinuity over saturation time and enthalpy of adsorption for all materials simulated

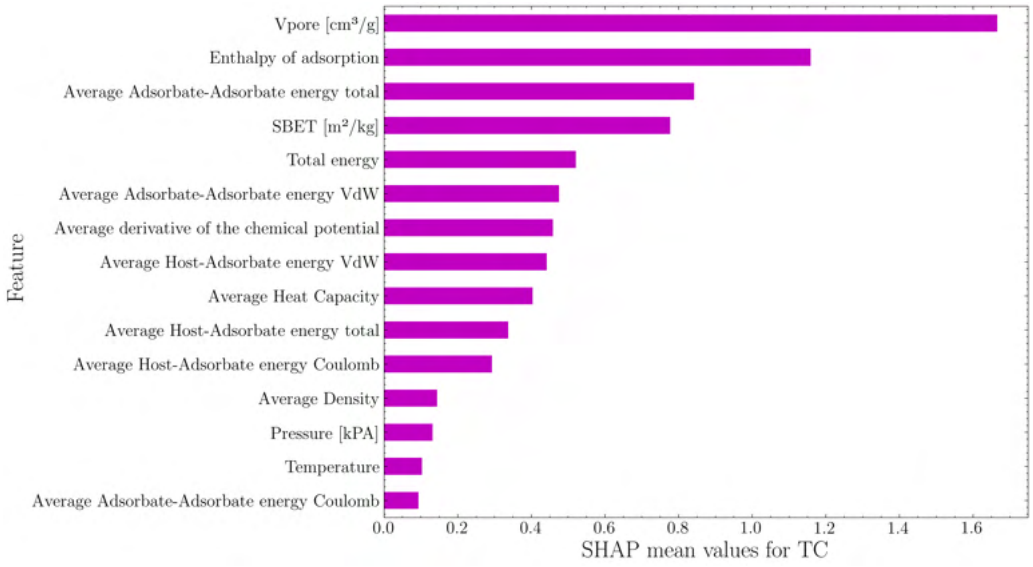


Source: Author (2024)

Summing up those trends, the TS indicator relies more on thermodynamic phenomena to be described. More textural properties or not-directly related features to temperature, are more capable of describing TBK. This last one can be conceived as a more mechanical property of the fixed bed studied.

Lastly, stoichiometric time presents its own ANN model and its own ranking by the SHAP values. In the Figure 36, an interesting fact is stated. Pore volume and enthalpy of adsorption are ranked first and second as more relevant to the model. A textural property and a thermodynamic property are the most relevant features for the forecasting of the stochiometric time of the fixed bed model developed. As relevant as both cited features is the mean value given, since the most relevant features for TBK and TS average 0.25 SHAP, while V_{pore} and enthalpy of adsorption have values higher than 1.10 SHAP. Concluding that section, it is also emphasized that the ANN model regarding TC combines and interpolates textural and thermodynamic properties in the four most relevant features to forecast the target. Since TC is an intermediate between TBK and TS, it suits the physical concept given to a fixed bed model.

Figure 36 – SHAP mean values for ANN-TC fitting



Source: Author (2024)

4.5 RESULTS CRITICAL ANALYSIS

The findings of the current study are promising, and therefore, certain aspects should present a critical analysis. The study focuses on a singular model of CO_2 adsorption, where one of the most challenging aspects concerns multi-component adsorption of O_2. Simultaneously, the synergistic effects of different components, such as water, should be considered. This critical aspect affects the necessity to comprehensively review the former approach for addressing a multi-component system from all three branches of the previous work. All three branches should be revisited, at the very least. For instance, in addressing the macro-scale modeling branch, diffusion and absorption isotherms need reevaluation due to the presence of more complex systems (multi-component absorption), which will impact the fixed bed model (POURSAEIDESFAHANI et al., 2019; ZHOU, M. et al., 2021; MOREIRA et al., 2024).

Regarding the perspective that the approach can be improved for multi-component systems, this perspective also holds for machine learning models and feature engineering approaches. A promising area of research concerns the development of force fields alongside machine learning models (LIU, S. et al., 2024; WIESER; ZOJER, 2024; YU, Honglei et al., 2024), which could significantly impact the approach used for molecular modeling and potentially change how machine learning is applied to perform the multiscale modeling presented here. All features used as inputs are from the perspective of molecular modeling an outfits; however, machine learning can utilize inputs directly from molecular modeling to macro-scale outputs (making an even bigguer bridge) (CHEN, Y.

et al., 2024). Therefore, a rigorous and continuous analysis of this perspective (features and outputs) will always be relevant, since potentially has a significant impact on the application of the former framework and the re-design of it (TIAN et al., 2024; DE VOS et al., 2024).

Regarding machine learning artificial neural networks, varying the architecture could lead to performance improvements and different interpretations regarding XAI approaches. The findings of the current study demonstrate the relevance of specific features related to a specific output. However, different models may consider different sets of features, leading to varying interpretability. Another approach to improving the capability of the primary approach in the former work, since it is relatively simple compared to the full potential related to machine learning applications, is to tune the artificial networks with L1 and L2 regularizatores, naturally raising an open question of how different architectures and combinations of hyperparameters can be explored to potentially deliver better outcomes, and also, if the interpretability remains.

These aspects lead to a final consideration: the number of materials used to develop the model. With the recent MOF X database and other sources (BOBBITT et al., 2023), the number of materials used to develop a multi-scale data-driven model can be significantly increased. The Open Core MOF 2019 (CHUNG, Yongchul G. et al., 2019b), for example, has at least 13,000 workable frameworks that can be used for the present approach. Evidently, the complexity, validation and computational costs will increase tremendously, although the potential of the present framework will also improve at a similar rate.

In conclusion, this study provides a solid foundation on how to approach a multi-scale methodology, that can do more than just carbon dioxide absorption. This work is considered a stepping stone towards the development of more refined multi-scale methodology for carbon dioxide absorption. The methodology presented itself as feasible concerning the conditions applied to each scale, especially the correlations used for the macro-scale simulations. It still needs to be determined if the same methodology applies to different correlations concerning mass diffusion resistances and linear driving force determination.

The developed artificial neural network (ANN) outperformed the random forest (RF) model in forecasting BKC time indicators, including TBK, TC, and TS. The interpretation of the ANN model developed by the SHAP-XAI presented that the most relevant features for each BKC indicator have physical coherence, which does not follow the RF model developed.

5 CONCLUSION

Regarding the suitability of the ANN, it is verified that the UAT is a solid principle that fortifies the algorithm's performance. The conclusion emphasizes the UAT principle for forecast reliability, considering both statistical indicators and methodological principles. Once there are intensive system properties related to different scales, and the extensive properties of the same system can be fitted or described by polynomials, Artificial Neural Networks are suitable for multiscale linkers. This work supports and validates the hypothesis.

The expected marginal contribution analysis found that the S_{BET} area is a relevant feature toward the definition of the TBK. The thermodynamic features such as enthalpy of adsorption and ADCP were the smallest contributors to the forecast of the BKC indicator regarding the ANN intelligence. Hence, it is concluded that the definition of the TBK indicator is more sensitive to textural properties.

Regarding the saturation time of a BKC, Enthalpy of Adsorption, E_{vdw} and ADCP carry more of a blueprint of the indicator, being those thermodynamic properties. Even with the S_{BET} not being the last valuable feature, the difference between the relevance of the mentioned features over S_{BET} is significant, allowing one to conclude that thermodynamic features or energetic features are more relevant for characterizing the TS indicator.

In one single phase, the nanoscale CO_2 adsorption simulation - sustained by the statistical thermodynamics where the ergotic principle is reproduced by the stochastic method of Monte Carlo within the Grand Canonical ensemble - is passive of a direct connection alongside a data frame with the macroscale modeling of the adsorption of CO_2; that one sustained by the deterministic Newtonian physics of a couple of partial differential equations solved by the finite difference approximation technique with centered differences. Assembling those two scales by the methods described in a data frame allows a way to apply Machine Learning Algorithms and Data Analysis toward the forecast of macroscale indicators from nanoscale many-body interactions. Hence, data-driven intelligence is accessible through the methodology developed. It is concluded that the methodology presented evolves a key way to confer the basis for a data intelligence approach focused on CO_2 adsorption.

REFERENCES

AFAGWU, Clement; MAHMOUD, Mohamed; ALAFNAN, Saad; PATIL, Shirish. Multiscale storage and transport modeling in unconventional shale gas: A review. **Journal of Petroleum Science and Engineering**, v. 208, p. 109518, Sept. 2021.

AGHAJI, Mohammad; FERNANDEZ, Michael; BOYD, Peter; DAFF, Thomas; WOO, Tom. Quantitative Structure-Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2 Working Capacity and CO2/CH4 Selectivity for Methane Purification: Quantitative Structure-Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2 Working Capacity and CO2/CH4 Selecti. **European Journal of Inorganic Chemistry**, v. 2016, June 2016.

AIMOLI, Cassiano G; MAGINN, Edward J; ABREU, Charlles RA. Force field comparison and thermodynamic property calculation of supercritical CO2 and CH4 using molecular dynamics simulations. **Fluid Phase Equilibria**, Elsevier, v. 368, p. 80–90, 2014.

AMBAW, Alemayehu. **MODELING CHEMICAL ENGINEERING PROCESSES USING ARTIFICIAL NEURAL NETWORKS**. Jan. 2005. PhD thesis.

AN, Yaxiong; FU, Qiang; ZHANG, Donghui; WANG, Yayan; TANG, Zhongli. Performance evaluation of activated carbon with different pore sizes and functional groups for VOC adsorption by molecular simulation. **Chemosphere**, v. 227, p. 9–16, 2019. ISSN 0045-6535.

ANDERSON, Ryther; RODGERS, Jacob; ARGUETA, Edwin; BIONG, Achay; GÓMEZ-GUALDRÓN, Diego A. Role of Pore Chemistry and Topology in the CO2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. **Chemistry of Materials**, v. 30, n. 18, p. 6325–6337, 2018.

BABARAO, Ravichandar; HU, Zhongqiao; JIANG, Jianwen; CHEMPATH, Shaji; SANDLER, Stanley I. Storage and Separation of CO2 and CH4 in Silicalite, C168 Schwarzite, and IRMOF-1:â€‰ A Comparative Study from Monte Carlo Simulation. **Langmuir**, v. 23, n. 2, p. 659–666, 2007. PMID: 17209617.

BABARAO, Ravichandar; HU, Zhongqiao; JIANG, Jianwen; CHEMPATH, Shaji; SANDLER, Stanley I. Storage and Separation of CO2 and CH4 in Silicalite, C168

Schwarzite, and IRMOF-1:â€‰ A Comparative Study from Monte Carlo Simulation. **Langmuir**, v. 23, n. 2, p. 659–666, 2007. PMID: 17209617.

BAHAMON, Daniel; VEGA, Lourdes F. Systematic evaluation of materials for post-combustion CO2 capture in a Temperature Swing Adsorption process. **Chemical Engineering Journal**, v. 284, p. 438–447, 2016. ISSN 1385-8947.

BAI, Peng; TSAPATSIS, Michael; SIEPMANN, J. Ilja. TraPPE-zeo: Transferable Potentials for Phase Equilibria Force Field for All-Silica Zeolites. **The Journal of Physical Chemistry C**, v. 117, n. 46, p. 24375–24387, 2013. eprint: `https://doi.org/10.1021/jp4074224`.

BANGERT, Patrick. **Machine learning and data science in the oil and gas industry: Best practices, tools, and case studies**. [S.l.]: Gulf Professional Publishing, 2021.

BANISHEIKHOLESLAMI, Abolhassan; QADERI, Farhad. A novel machine learning framework for predicting biogas desulfurization breakthrough curves in a fixed bed adsorption column. **Bioresource Technology Reports**, v. 25, p. 101702, 2024. ISSN 2589-014X.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BEASLEY, William Howard; O'KEEFE, Patrick; RODGERS, Joseph Lee. Fundamentals of bootstrapping and Monte Carlo methods. American Psychological Association, 2023.

BIAU, Gérard; SCORNET, Erwan. A random forest guided tour. **Test**, Springer, v. 25, p. 197–227, 2016.

BOBBITT, N. Scott et al. MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials. **Journal of Chemical & Engineering Data**, v. 68, n. 2, p. 483–498, 2023. eprint: `https://doi.org/10.1021/acs.jced.2c00583`.

BOOTHROYD, Simon et al. Development and Benchmarking of Open Force Field 2.0. 0: The Sage Small Molecule Force Field. **Journal of Chemical Theory and Computation**, ACS Publications, 2023.

BORGNAKKE, Claus; SONNTAG, Richard E. **Fundamentals of thermodynamics**. [S.l.]: John Wiley & Sons, 2020.

BOTU, V.; RAMPRASAD, R. Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics. **International Journal of Quantum Chemistry**, v. 115, p. 1074–1083, 2015.

BUI, Mai et al. Carbon capture and storage (CCS): the way forward. **Energy & Environmental Science**, Royal Society of Chemistry, v. 11, n. 5, p. 1062–1176, 2018.

BURGESS, Matthew G; RITCHIE, Justin; SHAPLAND, John; PIELKE, Roger. IPCC baseline scenarios have over-projected CO2 emissions and economic growth. **Environmental Research Letters**, IOP Publishing, v. 16, n. 1, p. 014016, Dec. 2020.

BURGESS, Matthew G; RITCHIE, Justin; SHAPLAND, John; PIELKE, Roger. IPCC baseline scenarios have over-projected CO2 emissions and economic growth. **Environmental Research Letters**, IOP Publishing, v. 16, n. 1, p. 014016, 2020.

BURNS, Thomas D.; PAI, Kasturi Nagesh; SUBRAVETI, Sai Gokul; COLLINS, Sean P.; KRYKUNOV, Mykhaylo; RAJENDRAN, Arvind; WOO, Tom K. Prediction of MOF Performance in Vacuum Swing Adsorption Systems for Postcombustion CO2 Capture Based on Integrated Molecular Simulations, Process Optimizations, and Machine Learning Models. **Environmental Science & Technology**, v. 54, n. 7, p. 4536–4544, 2020.

CASTILLO, Juan Manuel. **Molecular simulations in microporous materials: adsorption and separation**. 2009. PhD thesis – Technische Universiteit Delft.

CHEN, Cong; SUN, Jingyue; ZHANG, Yi; MU, Jianshu; LI, Weizhong; SONG, Yongchen. Adsorption characteristics of CH4 and CO2 in organic-inorganic slit pores. **Fuel**, v. 265, p. 116969, 2020. ISSN 0016-2361.

CHEN, Hongyu; GUO, Yang; DU, Yankun; XU, Xiang; SU, Changqing; ZENG, Zheng; LI, Liqing. The synergistic effects of surface functional groups and pore sizes on CO2 adsorption by GCMC and DFT simulations. **Chemical Engineering Journal**, v. 415, p. 128824, 2021. ISSN 1385-8947.

CHEN, Yiming et al. Robust machine learning inference from x-ray absorption near edge spectra through featurization. **Chemistry of Materials**, ACS Publications, v. 36, n. 5, p. 2304–2313, 2024.

CHU, Khim Hoong. Breakthrough curve analysis by simplistic models of fixed bed adsorption: In defense of the century-old Bohart-Adams model. **Chemical Engineering Journal**, v. 380, p. 122513, 2020. ISSN 1385-8947.

CHUNG, Yongchul G et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. **Journal of Chemical & Engineering Data**, ACS Publications, v. 64, n. 12, p. 5985–5998, 2019.

CHUNG, Yongchul G et al. Computation-ready, experimental metal–organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. **Chemistry of Materials**, ACS Publications, v. 26, n. 21, p. 6185–6192, 2014.

CHUNG, Yongchul G. et al. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. **Journal of Chemical & Engineering Data**, v. 64, n. 12, p. 5985–5998, 2019. eprint: `https://doi.org/10.1021/acs.jced.9b00835`.

COLELLA, Carmine; WISE, William S. The IZA Handbook of Natural Zeolites: A tool of knowledge on the most important family of porous minerals. **Microporous and mesoporous materials**, Elsevier, v. 189, p. 4–10, 2014.

DĄBROWSKI, A. Adsorption—from theory to practice. **Advances in colloid and interface science**, Elsevier, v. 93, n. 1-3, p. 135–224, 2001.

DAS, Arun; RAD, Paul. Opportunities and challenges in explainable artificial intelligence (xai): A survey. **arXiv preprint arXiv:2006.11371**, 2020.

DE VOS, Juul S; RAVICHANDRAN, Siddharth; BORGMANS, Sander; VANDUYFHUYS, Louis; VAN DER VOORT, Pascal; ROGGE, Sven MJ; VAN SPEYBROECK, Veronique. High-Throughput Screening of Covalent Organic Frameworks for Carbon Capture Using Machine Learning. **Chemistry of Materials**, ACS Publications, 2024.

DEMESSIE, Johannes A.; SORIAL, George A.; SAHLE-DEMESSIE, Endalkachew. Chapter 9 - Removing chromium (VI) from contaminated water using a nano-chitosan–coated diatomaceous earth. In: AHUJA, Satinder (Ed.). **Separations of Water Pollutants with Nanotechnology**. [S.l.]: Academic Press, 2022. v. 15. (Separation Science and Technology). P. 163–176.

DI BIASE, Emanuela; SARKISOV, Lev. Molecular simulation of multi-component adsorption processes related to carbon capture in a high surface area, disordered activated carbon. **Carbon**, Elsevier, v. 94, p. 27–40, 2015.

DOBBELAERE, Maarten R; PLEHIERS, Pieter P; VAN DE VIJVER, Ruben; STEVENS, Christian V; VAN GEEM, Kevin M. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. **Engineering**, Elsevier, v. 7, n. 9, p. 1201–1211, 2021.

DUBBELDAM, David; CALERO, Sofía; ELLIS, Donald E; SNURR, Randall Q. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. **Molecular Simulation**, Taylor & Francis, v. 42, n. 2, p. 81–101, 2016.

DÜREN, Tina; BAE, Youn-Sang; SNURR, Randall Q. Using molecular simulation to characterise metal–organic frameworks for adsorption applications. **Chemical Society Reviews**, Royal Society of Chemistry, v. 38, n. 5, p. 1237–1247, 2009.

EGBERT, Jesse; PLONSKY, Luke. Bootstrapping techniques. In: A practical handbook of corpus linguistics. [S.l.]: Springer, 2021. P. 593–610.

EMELIANOVA, Alina; REED, Allen; BASHAROVA, Elizaveta A; KOLESNIKOV, Andrei L; GOR, Gennady Y. Closer Look at Adsorption of Sarin and Simulants on Metal–Organic Frameworks. **ACS Applied Materials & Interfaces**, ACS Publications, v. 15, n. 14, p. 18559–18567, 2023.

FARMAHINI, Amir; KRISHNAMURTHY, Shreenath; FRIEDRICH, Daniel; BRANDANI, Stefano; SARKISOV, Lev. From Crystal to Adsorption Column: Challenges in Multiscale Computational Screening of Materials for Adsorption Separation Processes. **Industrial Engineering Chemistry Research**, v. 57, Oct. 2018.

FERNANDEZ, Michael; BARNARD, Amanda S. Geometrical Properties Can Predict CO2 and N2 Adsorption Performance of Metal–Organic Frameworks (MOFs) at Low Pressure. **ACS Combinatorial Science**, v. 18, n. 5, p. 243–252, 2016.

FOTOOHI, Fatemeh; AMJAD-IRANAGH, Sepideh; GOLZAR, Karim; MODARRESS, Hamid. Predicting pure and binary gas adsorption on activated carbon with two-dimensional cubic equations of state (2-D EOSs) and artificial neural network (ANN) method. **Physics and Chemistry of Liquids**, Taylor Francis, v. 54, n. 3, p. 281–302, 2016. eprint: `https://doi.org/10.1080/00319104.2015.1084877`.

GABRIELLI, Paolo; GAZZANI, Matteo; MAZZOTTI, Marco. The role of carbon capture and utilization, carbon capture and storage, and biomass to enable a net-zero-CO2 emissions chemical industry. **Industrial and Engineering Chemistry Research**, ACS Publications, v. 59, n. 15, p. 7033–7045, 2020.

GE, Wei; CHANG, Qi; CHENGXIANG, Li; WANG, Junwu. Multiscale structures in particle–fluid systems: Characterization, modeling, and simulation. **Chemical Engineering Science**, v. 198, Apr. 2019.

GÉRON, Aurélien. Mãos à obra: aprendizado de máquina com Scikit-Learn. **Keras & TensorFlow: Conceitos, ferramentas e técnicas para a construção de sistemas inteligentes.[S. l.: sn]**, 2021.

GHAEDI, Mehrorang (Ed.). **Adsorption: Fundamental Processes and Applications**. 1st. [S.l.]: Elsevier, Mar. 2021. v. 33.

GHANBARI, Taravat; ABNISA, Faisal; DAUD, Wan Mohd Ashri Wan. A review on production of metal organic frameworks (MOF) for CO2 adsorption. **Science of The Total Environment**, Elsevier, v. 707, p. 135090, 2020.

GREEN, Don W. **Perry's Chemical Engineer's Handbook 8/E Section 16 Adsorption & Ion Exchange (POD)**. [S.l.]: McGraw Hill Professional, Oct. 2007. P. 70. ISBN 9780071542056.

GRÖMPING, Ulrike. Variable importance assessment in regression: linear regression versus random forest. **The American Statistician**, JSTOR, p. 308–319, 2009.

GU, Chenkai; LIU, Jing; HU, Jianbo; WANG, Weizhou. Metal–Organic Frameworks Grafted by Univariate and Multivariate Heterocycles for Enhancing CO2 Capture: A Molecular Simulation Study. **Industrial & Engineering Chemistry Research**, v. 58, n. 6, p. 2195–2205, 2019.

GUNNING, David; STEFIK, Mark; CHOI, Jaesik; MILLER, Timothy; STUMPF, Simone; YANG, Guang-Zhong. XAI—Explainable artificial intelligence. **Science robotics**, American Association for the Advancement of Science, v. 4, n. 37, eaay7120, 2019.

GUPTA, Surojit; LI, Lan. The potential of machine learning for enhancing CO2 sequestration, storage, transportation, and utilization-based processes: a brief perspective. **JOM**, Springer, v. 74, n. 2, p. 414–428, 2022.

HARKER, J. H.; BACKHURST, J. R.; RICHARDSON, J. F. **Coulson and Richardson's Chemical Engineering Volume 2**. 5th. Oxford: Butterworth-Heinemann, July 2002.

HOEKSTRA, Alfons; CHOPARD, Bastien; COVENEY, Peter. Multiscale modelling and simulation: a position paper. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 372, n. 2021, p. 20130377, 2014.

HOLLINGSWORTH, Scott A; DROR, Ron O. Molecular dynamics simulation for all. **Neuron**, Elsevier, v. 99, n. 6, p. 1129–1143, 2018.

HOLLINGSWORTH, Scott A.; DROR, Ron O. Molecular Dynamics Simulation for All. **Neuron**, v. 99, n. 6, p. 1129–1143, 2018. ISSN 0896-6273.

HU, Xiaofei; DENG, Hucheng; LU, Chang; TIAN, Yuanyuan; JIN, Zhehui. Characterization of CO2/CH4 Competitive Adsorption in Various Clay Minerals in Relation to Shale Gas Recovery from Molecular Simulation. **Energy & Fuels**, v. 33, n. 9, p. 8202–8214, 2019.

HUANG, Hongliang; ZHANG, Wenjuan; LIU, Dahuan; LIU, Bei; CHEN, Guangjin; ZHONG, Chongli. Effect of temperature on gas adsorption and separation in ZIF-8: A combined experimental and molecular simulation study. **Chemical engineering science**, Elsevier, v. 66, n. 23, p. 6297–6305, 2011.

HUANG, Liang; NING, Zhengfu; WANG, Qing; QI, Rongrong; ZENG, Yan; QIN, Huibo; YE, Hongtao; ZHANG, Wentong. Molecular simulation of adsorption behaviors of methane, carbon dioxide and their mixtures on kerogen: Effect of kerogen maturity and moisture content. **Fuel**, v. 211, p. 159–172, 2018. ISSN 0016-2361.

KANG, Guanxian; ZHANG, Bin; KANG, Tianhe; GUO, Junqing; ZHAO, Guofei. Effect of pressure and temperature on CO2/CH4 competitive adsorption on kaolinite by Monte Carlo simulations. **Materials**, MDPI, v. 13, n. 12, p. 2851, 2020.

KARAPIPERIS, K.; STAINIER, L.; ORTIZ, M.; ANDRADE, J.E. Data-Driven multiscale modeling in mechanics. **Journal of the Mechanics and Physics of Solids**, v. 147, p. 104239, 2021. ISSN 0022-5096.

KEIL, Frerich J. Molecular modelling for reactor design. **Annual review of chemical and biomolecular engineering**, Annual Reviews, v. 9, p. 201–227, 2018.

KEVLAHAN, Nicholas. Principles of multiscale modeling. **Physics Today**, AIP Publishing, v. 65, n. 6, p. 56–57, 2012.

KNOX, James C.; EBNER, Armin D.; LEVAN, M. Douglas; COKER, Robert F.; RITTER, James A. Limitations of Breakthrough Curve Analysis in Fixed-Bed Adsorption. **Industrial & Engineering Chemistry Research**, v. 55, n. 16,

p. 4734–4748, 2016. PMID: 31359909. eprint:
https://doi.org/10.1021/acs.iecr.6b00516.

KOLLE, Joel M.; FAYAZ, Mohammadreza; SAYARI, Abdelhamid. Understanding the Effect of Water on CO2 Adsorption. **Chemical Reviews**, v. 121, n. 13, p. 7280–7345, 2021.

KROESE, Dirk P; BOTEV, Zdravko; TAIMRE, Thomas; VAISMAN, Radislav. **Data science and machine learning: mathematical and statistical methods**. [S.l.]: CRC Press, 2019.

KUMAR, Kishant; KUMAR, Amit. Adsorptive separation of carbon dioxide from flue gas using mesoporous MCM-41: A molecular simulation study. **Korean Journal of Chemical Engineering**, v. 35, p. 535–547, 2018.

KWON, Soonchul et al. Enhanced Selectivity for CO2 Adsorption on Mesoporous Silica with Alkali Metal Halide Due to Electrostatic Field: A Molecular Simulation Approach. **ACS Applied Materials & Interfaces**, v. 9, n. 37, p. 31683–31690, 2017.

LANDAU, David; BINDER, Kurt. **A guide to Monte Carlo simulations in statistical physics**. [S.l.]: Cambridge university press, 2021.

LE, Tien Dung; QUOC DAT, Ha; PANFILOVA, Irina; MOYNE, C. Multiscale model for flow and transport in CO2-enhanced coalbed methane recovery incorporating gas mixture adsorption effects. **Advances in Water Resources**, v. 144, p. 103706, July 2020.

LEE, Yongjin; BARTHEL, Senja D; DŁOTKO, Paweł; MOOSAVI, Seyed Mohamad; HESS, Kathryn; SMIT, Berend. High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites. **Journal of chemical theory and computation**, ACS Publications, v. 14, n. 8, p. 4427–4437, 2018.

LEPERI, Karson T; YANCY-CABALLERO, Daison; SNURR, Randall Q; YOU, Fengqi. 110th anniversary: surrogate models based on artificial neural networks to simulate and optimize pressure swing adsorption cycles for CO2 capture. **Industrial & Engineering Chemistry Research**, ACS Publications, v. 58, n. 39, p. 18241–18252, 2019.

LETCHER, Trevor; MYERS, Alan L. Thermodynamics of adsorption. In: CHEMICAL thermodynamics for industry. [S.l.]: The Royal Society of Chemistry, 2004. P. 243–253.

LI, Xiaoqiang et al. Applied machine learning to analyze and predict CO2 adsorption behavior of metal-organic frameworks. **Carbon Capture Science & Technology**, Elsevier, v. 9, p. 100146, 2023.

LIU, Chao; WANG, Jing; WAN, Jingjing; YU, Chengzhong. MOF-on-MOF hybrids: Synthesis and applications. **Coordination Chemistry Reviews**, Elsevier, v. 432, p. 213743, 2021.

LIU, Shanping; DUPUIS, Romain; FAN, Dong; BENZARIA, Salma; BONNEAU, Mickaele; BHATT, Prashant; EDDAOUDI, Mohamed; MAURIN, Guillaume. Machine learning potential for modelling H 2 adsorption/diffusion in MOFs with open metal sites. **Chemical Science**, Royal Society of Chemistry, v. 15, n. 14, p. 5294–5302, 2024.

LIU, Xiao-Qiang; HE, Xu; QIU, Nian-Xiang; YANG, Xin; TIAN, Zhi-Yue; LI, Mei-Jun; XUE, Ying. Molecular simulation of CH4, CO2, H2O and N2 molecules adsorption on heterogeneous surface models of coal. **Applied Surface Science**, Elsevier, v. 389, p. 894–905, 2016.

MAHAJAN, Shreya; LAHTINEN, Manu K. Recent Progress in Metal-organic Frameworks (MOFs) for CO2 Capture At Different Pressures. **Journal of Environmental Chemical Engineering**, 2022.

MAREK, Nedoma; MAREK, Staf; JAN, Hardlika. EXPERIMENTAL AND SIMULATION STUDY OF CO2 BREAKTHROUGH CURVES IN A FIXED-BED ADSORPTION PROCESS. In.

MARTIN, Marcus G.; SIEPMANN, J. Ilja. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. **The Journal of Physical Chemistry B**, v. 102, p. 2569–2577, 1998.

MARTIN-CALVO, Ana; VAN DER PERRE, Stijn; CLAESSENS, Benjamin; CALERO, Sofia; DENAYER, Joeri F. M. Unravelling the influence of carbon dioxide on the adsorptive recovery of butanol from fermentation broth using ITQ-29 and ZIF-8. **Phys. Chem. Chem. Phys.**, The Royal Society of Chemistry, v. 20, p. 9957–9964, 15 2018.

MARTINS, Marcio AF; RODRIGUES, Alírio E; LOUREIRO, José M; RIBEIRO, Ana M; NOGUEIRA, Idelfonso BR. Artificial Intelligence-oriented economic non-linear model predictive control applied to a pressure swing adsorption unit: Syngas purification as a case study. **Separation and Purification Technology**, Elsevier, v. 276, p. 119333, 2021.

MATSUMURA, M.; NAYVE JR., F. R. P. Effects of Ammonium Ion Removal on Growth and MAb Production of Hybridoma Cells. **Cytotechnology**, v. 18, n. 1-2, p. 35–50, 1995.

MESFER, Mohammed K. Al; DANISH, Mohd; KHAN, Mohammed Ilyas; ALI, Ismat Hassan; HASAN, Mudassir; JERY, Atef El. Continuous Fixed Bed CO2 Adsorption: Breakthrough, Column Efficiency, Mass Transfer Zone. **Processes**, v. 8, n. 10, 2020. ISSN 2227-9717.

MESSALAS, Andreas; KANELLOPOULOS, Yiannis; MAKRIS, Christos. Model-agnostic interpretability with shapley values. In: IEEE. 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). [S.l.: s.n.], 2019. P. 1–7.

MÍGUEZ, JM; GÓMEZ-ÁLVAREZ, P; PIÑEIRO, MM; MENDIBOURE, Bruno; BLAS, FJ. Adsorption and interfacial phenomena of a Lennard-Jones fluid adsorbed in slit pores: DFT and GCMC simulations. **Molecular Physics**, Taylor & Francis, v. 116, n. 21-22, p. 3417–3424, 2018.

MIRZAEI, Mahmoud. Science and Engineering In Silico. v. 1, Apr. 2020.

MONTÁNS, Francisco; CHINESTA, Francisco; GÓMEZ-BOMBARELLI, Rafael; KUTZ, J. Data-driven modeling and learning in science and engineering. **Comptes Rendus Mécanique**, v. 347, Nov. 2019.

MOREIRA, Davi DS; GONÇALVES, Daniel V; COELHO, Juliana A; AZEVEDO, Diana CS de; RIOS, Rafael B; LUCENA, Sebastião MP de; BASTOS-NETO, Moises. Influence of SO2 on CO2 capture by adsorption on activated carbon: Individual pore performance via multiscale simulation. **Separation and Purification Technology**, Elsevier, v. 336, p. 126219, 2024.

MORGANTE, Pierpaolo; PEVERATI, Roberto. The devil in the details: A tutorial review on some undervalued aspects of density functional theory calculations. **International Journal of Quantum Chemistry**, Wiley Online Library, v. 120, n. 18, e26332, 2020.

MURPHY, Orla P.; VASHISHTHA, Mayank; PALANISAMY, Parimaladevi; KUMAR, K. Vasanth. A Review on the Adsorption Isotherms and Design Calculations for the Optimization of Adsorbent Mass and Contact Time. **ACS Omega**, v. 8, n. 20, p. 17407–17430, 2023. eprint: `https://doi.org/10.1021/acsomega.2c08155`.

MYERS, AL. Thermodynamics of adsorption in porous materials. **AIChE journal**, Wiley Online Library, v. 48, n. 1, p. 145–160, 2002.

NAIDU, Haripriya; MATHEWS, Alexander P. Linear driving force analysis of adsorption dynamics in stratified fixed-bed adsorbers. **Separation and Purification Technology**, v. 257, p. 117955, 2021. ISSN 1383-5866.

NISHIJIMA, Takato. Universal approximation theorem for neural networks. **arXiv preprint arXiv:2102.10993**, 2021.

OKELLO, Felix Otieno; TIZHE FIDELIS, Timothy; AGUMBA, John; MANDA, Timothy; OCHILO, Livingstone; MAHMOOD, Asif; PEMBERE, Anthony. Towards estimation and mechanism of CO2 adsorption on zeolite adsorbents using molecular simulations and machine learning. **Materials Today Communications**, v. 36, p. 106594, 2023. ISSN 2352-4928.

OLECHOWSKI, Alison; EPPINGER, Steven; JOGLEKAR, Nitin. Technology Readiness Levels at 40: A Study of State-of-the-Art Use, Challenges, and Opportunities: **SSRN Electronic Journal**, Apr. 2015.

OLIVEIRA, Luís Miguel Cunha; KOIVISTO, Hannu; IWAKIRI, Igor GI; LOUREIRO, José M; RIBEIRO, Ana M; NOGUEIRA, Idelfonso BR. Modelling of a pressure swing adsorption unit by deep learning and artificial Intelligence tools. **Chemical Engineering Science**, Elsevier, v. 224, p. 115801, 2020.

ORNSTEIN, Donald S; WEISS, Benjamin. Entropy and data compression schemes. **IEEE Transactions on information theory**, IEEE, v. 39, n. 1, p. 78–83, 1993.

PILANIA, Ghanshyam; WANG, Chenchen; JIANG, Xun; RAJASEKARAN, Sanguthevar; RAMPRASAD, Ramamurthy. Accelerating materials

property predictions using machine learning. **Scientific reports**, Nature Publishing Group, v. 3, n. 1, p. 1–6, 2013.

POURSAEIDESFAHANI, Ali et al. Prediction of adsorption isotherms from breakthrough curves. **Microporous and Mesoporous Materials**, v. 277, Nov. 2018.

POURSAEIDESFAHANI, Ali et al. Prediction of adsorption isotherms from breakthrough curves. **Microporous and Mesoporous Materials**, Elsevier, v. 277, p. 237–244, 2019.

PUGNAIRE, Francisco; MORILLO-PÉREZ, Jose Antonio; PENUELAS, Josep; REICH, Peter; BARDGETT, Richard; GAXIOLA, Aurora; WARDLE, David; PUTTEN, Wim. Climate change effects on plant-soil feedbacks and consequences for biodiversity and functioning of terrestrial ecosystems. **Science Advances**, v. 5, eaaz1834, Nov. 2019.

PULLUMBI, Pluton; BRANDANI, Federico; BRANDANI, Stefano. Gas separation by adsorption: technological drivers and opportunities for improvement. **Current Opinion in Chemical Engineering**, Elsevier, v. 24, p. 131–142, 2019.

RACCUGLIA, Paul et al. Machine-learning-assisted materials discovery using failed experiments. **Nature**, v. 533, p. 73–76, May 2016.

RAPPÉ, Anthony K; CASEWIT, Carla J; COLWELL, KS; GODDARD III, William A; SKIFF, W Mason. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. **Journal of the American chemical society**, ACS Publications, v. 114, n. 25, p. 10024–10035, 1992.

RASCHKA, Sebastian; MIRJALILI, Vahid. **Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2**. [S.l.]: Packt Publishing Ltd, 2019.

RAY, M.S. Adsorption principles, design data and adsorbent materials for industrial applications: A Bibliography (1967-1997). In: DĄBROWSKI, A. (Ed.). **Adsorption and its Applications in Industry and Environmental Protection**. [S.l.]: Elsevier, 1999. v. 120. (Studies in Surface Science and Catalysis). P. 977–1049.

REBELLO, Carine M; MARROCOS, Paulo H; COSTA, Erbet A; SANTANA, Vinicius V; RODRIGUES, Alírio E; RIBEIRO, Ana M; NOGUEIRA, Idelfonso BR. Machine

Learning-Based Dynamic Modeling for Process Engineering Applications: A Guideline for Simulation and Prediction from Perceptron to Deep Learning. **Processes**, MDPI, v. 10, n. 2, p. 250, 2022.

RUTHVEN, Douglas M. Adsorption (Chemical Engineering). In: MEYERS, Robert A. (Ed.). **Encyclopedia of Physical Science and Technology (Third Edition)**. Third Edition. New York: Academic Press, 2003. P. 251–271. ISBN 978-0-12-227410-7.

RUTHVEN, Douglas Morris; FAROOQ, Shamsuzzaman; KNAEBEL, Kent S. **Pressure Swing Adsorption**. [S.l.]: VCH Publishers, 1994. P. 352. ISBN 9781560817588.

SABOUNI, Rana; KAZEMIAN, Hossein; ROHANI, Sohrab. Mathematical Modeling and Experimental Breakthrough Curves of Carbon Dioxide Adsorption on Metal Organic Framework CPM-5. **Environmental Science & Technology**, v. 47, n. 16, p. 9372–9380, 2013. PMID: 23889136. eprint: `https://doi.org/10.1021/es401276r`.

SAEEDIRAD, Raheleh; GANJALI, Saeed Taghvaei; RASHIDI, Alimorad; BAZMI, Mansour. Experimental and Computational Study of Organic Sulfur Removal Proficiency of (Ni, Cu, Co)-Doped ZIF-8 Adsorbents. **ChemistrySelect**, v. 5, n. 1, p. 231–243, 2020.

SARKAR, A. I.; AROONWILAS, A.; VEAWAB, A. Equilibrium and Kinetic Behavior of CO2 Adsorption onto Zeolites, Carbon Molecular Sieve, and Activated Carbons. **Energy Procedia**, v. 114, p. 2450–2459, 2017.

SCHILLER, Ulf; WANG, Fang. Multiscale simulation of transport phenomena in porous media: from toy models to materials models. **MRS Communications**, v. 8, p. 1–14, Mar. 2018.

SMIT, Berend. Molecular Simulations of Zeolites: Adsorption, Diffusion, and Shape Selectivity. **Chemical Reviews**, v. 108, n. 10, p. 4125–4184, 2008. PMID: 18817356. eprint: `https://doi.org/10.1021/cr8002642`.

STURLUSON, Arni et al. The role of molecular modelling and simulation in the discovery and deployment of metal-organic frameworks for gas storage and separation. **Molecular Simulation**, Taylor Francis, v. 45, n. 14-15, p. 1082–1121, 2019.

SUN, L. M.; LE QUERÉ, P.; LEVAN, M. D. Numerical Simulation of Diffusion-Limited PSA Process Models by Finite Difference Methods. **Chemical Engineering Science**, v. 51, p. 5341–5352, 1996.

THOMAS, W. John; CRITTENDEN, Barry. 2 - Adsorbents. In: THOMAS, W. John; CRITTENDEN, Barry (Eds.). **Adsorption Technology Design**. Oxford: Butterworth-Heinemann, 1998. P. 8–30. ISBN 978-0-7506-1959-2.

THOMAS, W. John; CRITTENDEN, Barry. 3 - Fundamentals of adsorption equilibria. In: THOMAS, W. John; CRITTENDEN, Barry (Eds.). **Adsorption Technology Design**. Oxford: Butterworth-Heinemann, 1998. P. 31–65. ISBN 978-0-7506-1959-2.

THOMAS, W. John; CRITTENDEN, Barry. 4 - Rates of adsorption of gases and vapours by porous media. In: THOMAS, W. John; CRITTENDEN, Barry (Eds.). **Adsorption Technology Design**. Oxford: Butterworth-Heinemann, 1998. P. 66–95. ISBN 978-0-7506-1959-2.

THOMAS, W. John; CRITTENDEN, Barry. 5 - Processes and cycles. In: THOMAS, W. John; CRITTENDEN, Barry (Eds.). **Adsorption Technology Design**. Oxford: Butterworth-Heinemann, 1998. P. 96–134. ISBN 978-0-7506-1959-2.

TIAN, Weizhi et al. Machine-learning-assisted hydrogen adsorption descriptor design for bilayer MXenes. **Journal of Cleaner Production**, Elsevier, v. 450, p. 141953, 2024.

TISCORNIA, Inés; VALENCIA, Susana; CORMA, Avelino; TÁLLEZ, Carlos; CORONAS, Joaquin; SANTAMAR, Jesus. Preparation of ITQ-29 (Al-free zeolite A) membranes. **Microporous and Mesoporous Materials**, v. 110, n. 2, p. 303–309, 2008. ISSN 1387-1811.

TURBAN, David HP; TEOBALDI, Gilberto; O'REGAN, David D; HINE, Nicholas DM. Supercell convergence of charge-transfer energies in pentacene molecular crystals from constrained DFT. **Physical Review B**, APS, v. 93, n. 16, p. 165102, 2016.

VEGA, Lourdes F.; BAHAMON, Daniel. Importance of Bridging Molecular and Process Modeling to Design Optimal Adsorbents for Large-Scale CO2 Capture. **Accounts of Chemical Research**, v. 0, n. 0, null, 0. PMID: 38156949. eprint: https://doi.org/10.1021/acs.accounts.3c00478.

VLUGT, Thijs JH; VAN DER EERDEN, Jan PJM; DIJKSTRA, Marjolein; SMIT, Berend; FRENKEL, Daan. Introduction to molecular simulation and statistical thermodynamics. **http://homepage. tudelft. nl/v9k6y/imsst/index. html**, 2009.

WANG, Jianlong; GUO, Xuan. Adsorption isotherm models: Classification, physical meaning, application and solving method. **Chemosphere**, v. 258, p. 127279, 2020.

WANG, Z; WANG, M; YONG, Q; GUO, YH; CUI, YW. Materials informatics and its application in materials research. **Materials China**, v. 36, n. 2, p. 132–140, 2017.

WANG, Zhenguang; SHEN, Yuanhui; ZHANG, Donghui; TANG, Zhongli; LI, Wenbin. A comparative study of multi-objective optimization with ANN-based VPSA model for CO2 capture from dry flue gas. **Journal of Environmental Chemical Engineering**, Elsevier, v. 10, n. 3, p. 108031, 2022.

WIESER, Sandro; ZOJER, Egbert. Machine learned force-fields for an Ab-initio quality description of metal-organic frameworks. **npj Computational Materials**, Nature Publishing Group UK London, v. 10, n. 1, p. 18, 2024.

WILKINS, N. S.; RAJENDRAN, A.; FAROOQ, S. Dynamic Column Breakthrough Experiments for Measurement of Adsorption Equilibrium and Kinetics. **Adsorption**, v. 27, n. 3, p. 397–422, 2020.

XIANG, Zhonghua; CAO, Dapeng; LAN, Jian Hui; WANG, Wenchuan; BROOM, Darren. Multiscale Simulation and Modelling of Adsorptive Processes for Energy Gas Storage and Carbon Dioxide Capture in Porous Coordination Frameworks. **Energy and Environmental Science**, v. 3, p. 1469–1487, Oct. 2010.

YAMADA, Hironao; LIU, Chang; WU, Stephen; KOYAMA, Yukinori; JU, Shenghong; SHIOMI, Junichiro; MORIKAWA, Junko; YOSHIDA, Ryo. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. **ACS Central Science**, v. 2019, Sept. 2019.

YANG, R. T. **Gas Separation by Adsorption Processes**. Reprint. [S.l.]: Imperial College Press, Jan. 1997. P. 352. This highly popular book in the field has now been reprinted and made available in paperback form. ISBN 9781860940177.

YANG, Wuyue; PENG, Liangrong; ZHU, Yi; HONG, Liu. When Machine Learning Meets Multiscale Modeling in Chemical Reactions. **Journal of Chemical Physics**, v. 153, p. 094117, 9 2020.

YE, Feng; MA, Shuo; TONG, Liang; XIAO, Jinsheng; BÉNARD, Pierre; CHAHINE, Richard. Artificial neural network based optimization for hydrogen purification performance of pressure swing adsorption. **International Journal of Hydrogen Energy**, Elsevier, v. 44, n. 11, p. 5334–5344, 2019.

YORO, Kelvin O.; DARAMOLA, Michael O. Chapter 1 - CO2 emission sources, greenhouse gases, and the global warming effect. Ed. by Mohammad Reza Rahimpour, Mohammad Farsi and Mohammad Amin Makarem. Woodhead Publishing, p. 3–28, 2020.

YU, Honglei; WANG, Dexi; LI, Yunlong; CHEN, Gong; MA, Xueyi. Explainable molecular simulation and machine learning for carbon dioxide adsorption on magnesium oxide. **Fuel**, Elsevier, v. 357, p. 129725, 2024.

YU, Hui; WANG, Xue; XU, Chunhui; CHEN, De-Li; ZHU, Weidong; KRISHNA, Rajamani. Utilizing transient breakthroughs for evaluating the potential of Kureha carbon for CO2 capture. **Chemical Engineering Journal**, v. 269, p. 135–147, 2015. ISSN 1385-8947.

ZHAO, Jie; DENG, Shuai; ZHAO, Li; YUAN, Xiangzhou; WANG, Bin; CHEN, Lijin; WU, Kailong. Synergistic and competitive effect of H2O on CO2 adsorption capture: Mechanism explanations based on molecular dynamic simulation. **Journal of CO2 Utilization**, v. 52, p. 101662, 2021. ISSN 2212-9820.

ZHOU, Musen; WANG, Jingqi; GARCIA, Jose; LIU, Yu; WU, Jianzhong. Modeling multicomponent gas adsorption in nanoporous materials with two versions of nonlocal classical density functional theory. **Industrial & Engineering Chemistry Research**, ACS Publications, v. 60, n. 47, p. 17016–17025, 2021.

ZHOU, Wenning; WANG, Haobo; ZHANG, Zhe; CHEN, Hongxia; LIU, Xunliang. Molecular simulation of CO2/CH4/H2O competitive adsorption and diffusion in brown coal. **RSC Adv.**, The Royal Society of Chemistry, v. 9, p. 3004–3011, 6 2019.

APPENDIX A – ADDITIONAL ANALYTICAL TOPICS

A.1 PRINCIPAL COMPONENTS ANALYSIS

Given the substantial number of variables, and the emphasis on interpretability in machine learning approaches (e.g., SHAP analysis), Principal Component Analysis (PCA) can illustrate how data distribution influences the final results. To compare the PCA of datasets dedicated to each performance indicator, we examine the convergence of component distribution with SHAP analysis, providing a comprehensive visualization of variable explanations across Principal Components.

Since all performance indicators (e.g., TBK, TC, and TS) share the same input dataset, it was observed that, for all 15 variables considered, four principal components were required to describe 97.72% of the dataset variability. The contribution of the 5th to 8th components does not exceed 0.02% of the dataset variability. For formal analysis, only the first three PCs are considered, adhering to the Kaiser criteria. The table below presents eigenvalues (PCs), associated variances, and cumulative variances.

Table 9 – Principal Components Analysis Results

Eigenvalue	**Variance**	**Acumulated variance [%]**
PC 1	10.147501	0.676500
PC 2	2.887168	0.192478
PC 3	1.011803	0.067454
PC 4	0.611884	0.040792
PC 5	0.210786	0.014052
PC 6	0.062408	0.004161
PC 7	0.039218	0.002615
PC 8	0.018371	0.001225
PC 9	0.006028	0.000402
PC 10	0.003759	0.000251

Source: Author (2024)

A crucial aspect of PCA is understanding how variables manifest across PCs. Figure 37 shows that SBET and Vpore are predominantly explained by the first PC, isolated at the upper part of the graph. Meanwhile, the second PC encompasses a crowded cluster of variables, indicating their stronger association with PC-2. Notably, textural properties and thermodynamic properties exhibit different PC influences.

The case of Enthalpy of Adsorption is particularly significant, being explained with relevance by PC-1 and PC-2. The negative correlation with PC-1 suggests an inverse relationship with textural properties, while the positive correlation with PC-2 indicates a direct association with thermodynamic properties. A similar pattern is observed

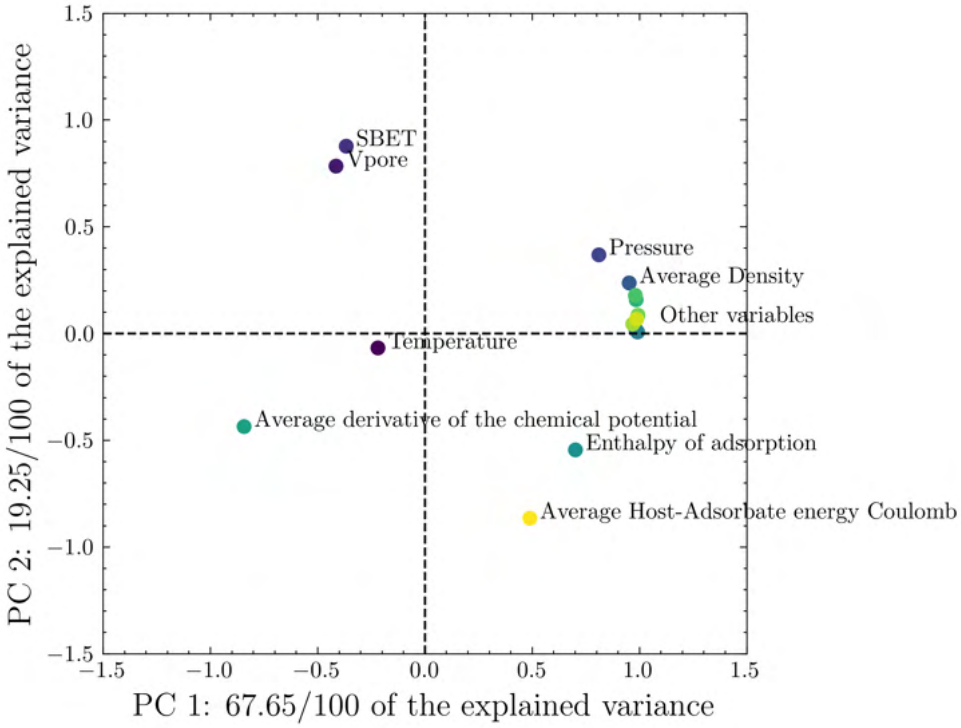


Figure 37 – Principal Components Analysis

Source: Author (2024)

for the Average Host-Adsorbate Coulomb energy. Notably, textural and thermodynamic properties are influenced by different components.

Concerning Average Host-Adsorbate Coulomb energy, it is crucial to consider data quality, as this variable relates more to mechanisms than absolute adsorption properties. The Average Derivative of the Chemical Potential consistently shows negative values for both PC-1 and PC-2, indicating an inverse relationship. Understanding the nature of this variable is essential for appropriate analysis.

Temperature stands out as a singular variable when plotting PC-1 versus PC-3, emphasizing that PC-3 almost exclusively explains the Temperature variable. A 2D plot with the respective PCs illustrates this relationship, with Temperature prominently standing out.

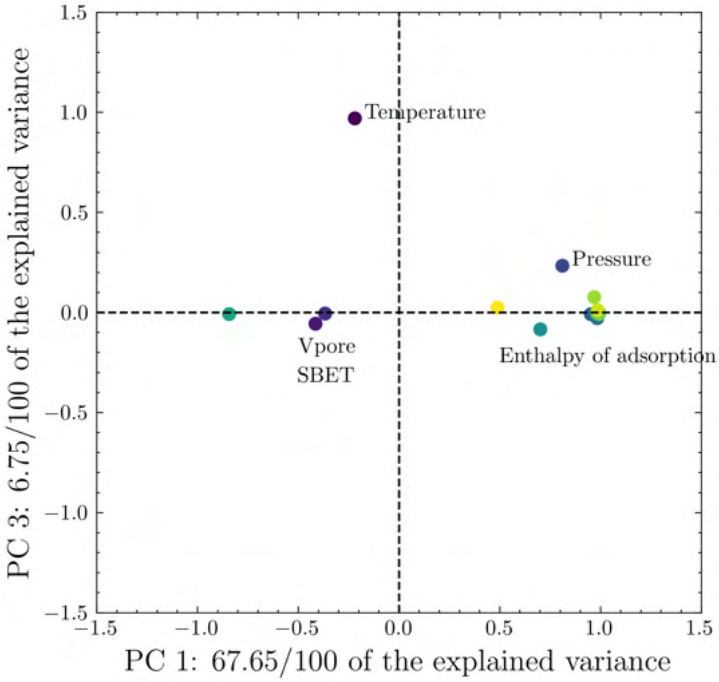


Figure 38 – 2D Plot of the first and third Principal Components

Source: Author (2024)

A.2 BOOTSTRAPPING AND FUNCTION FITTING

A common approach in data science applications is the bootstrapping technique, which aims to enhance the accuracy of statistical inference in situations where the sample size is relatively small. One notable application of bootstrapping is the development of confidence intervals based on the dataset of work (GUPTA; LI, L., 2022; EGBERT; PLONSKY, 2021). This method offers a faster approach compared to other relevant techniques, such as data function fitting and smoothing (BEASLEY; O'KEEFE; RODGERS, 2023h).

In the context of this work, the relevance of bootstrapping lies in its application during the data wrangling step, where a set of functions is used to enhance the confidence intervals of datasets. Despite potential criticisms of the applied technique, this section presents evidence that the approach used is suitable and unbiased, comparing it with the more common approach of bootstrapping.

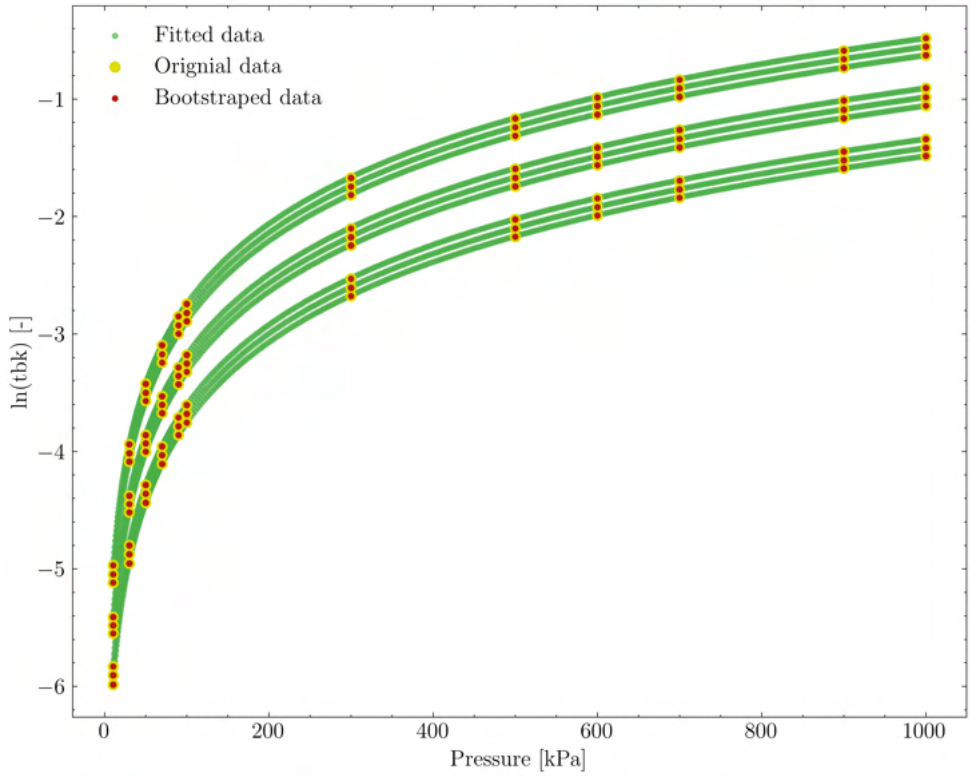
The first aspect that reinforces the use of function fitting is the behavior of the data concerning temperature and, especially, pressure. Since the physical process of CO_2 adsorption follows a predictable behavior under proper conditions, it is common to use experimentally or theoretically based equations to fit this data. The interpretation and application of these equations help not only in understanding the phenomena but also in controlling it. In the present work, the experimental data consists of a controlled

range of temperature and pressure for all sampled materials. For all these materials, the *in silico* data follows a verified *in situ* experimental approach. Function fitting describes this behavior, not by re-sampling data, but by enhancing the dataset.

Bootstrapped data will have the same distribution as the original dataset, although it does not consider the behavior of the phenomena. Regardless of how many new samples are generated by bootstrapping, they will consistently be the same data in concept. Another way to visualize it is by plotting original data and bootstrapped data, where both will be overlaid (Figure 39).

Therefore, by fitting the data through functions, the interpolated behavior of the adsorption will be considered, adding value to the new dataset, a feature that bootstrapping does not provide. Lastly, the more data generated by the molecular simulation, considering the same range of pressure and temperature, the tendency is to describe the behavior that function fitting delivers.

Figure 39 – Comparison of fitted data and bootstrapped data



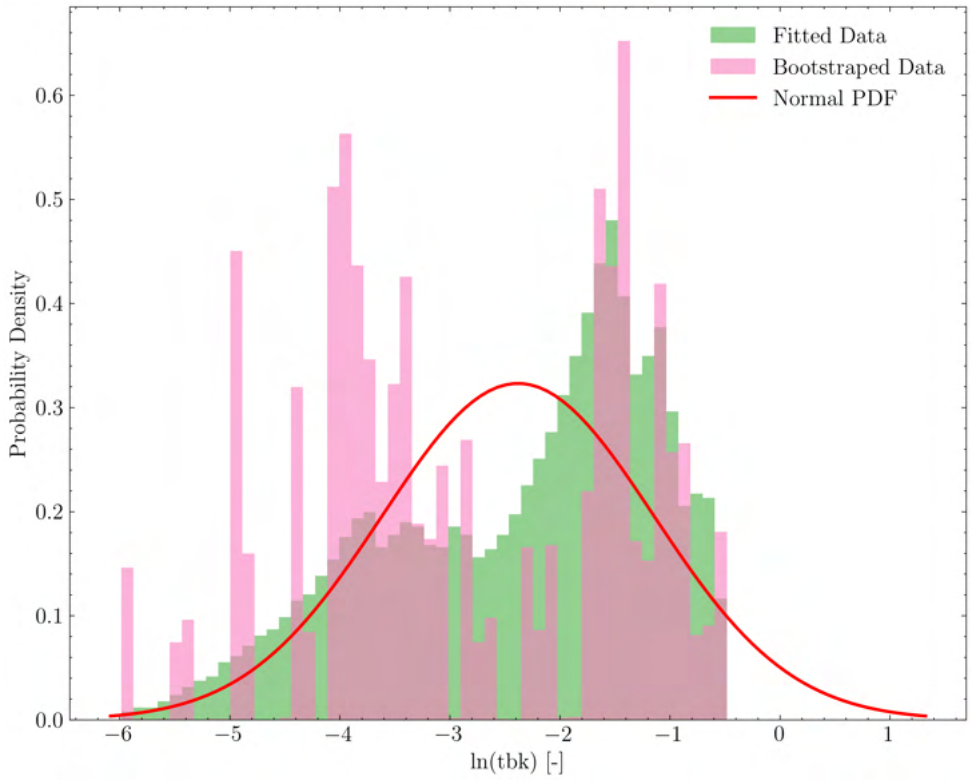
Source: Author (2024)

Therefore, a relevant aspect is the statistical significance addressing bootstrapping and function fitting. The data sampling developed from bootstrapping has a p-value significantly smaller than 0.05 (e.g., approximately null) when compared to *in silico* data.

This is strong evidence that there is a difference in data behavior when re-sampled by bootstrapping, which does not follow the function-fitting approach used in the present work. In other words, there is no statistical difference in the data fitted when compared with the original data (p-value > 0.05), and there is a statistical difference in the data bootstrapped when compared with the original data (p-value < 0.05).

The justification for this behavior is believed to be the exact point mentioned above: since the phenomenal data has a predictable behavior, the distribution of these data does not follow a normal distribution; they follow a tendency that is disrupted when bootstrapped. The CO_2 adsorption phenomena do not randomly occur concerning pressure and temperature. The following figure (Figure 40) compares the probability density for both fitted data and bootstrapped data. One can observe the difference between both and also from a normal distribution.

Figure 40 – Comparison of probability density for fitted data and bootstrapped data



Source: Author (2024)

A.3 NEURAL NETWORKS ARCHITECTURE DESIGN

The Table 10 provides a comparative analysis of various architectures used in tuning the final architecture for the former work. This section explores the variations in the number of inner layers and overall neurons. Each architecture is assessed based on Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) metrics using a training dataset. This procedure is applied only on saturation time. The study's findings extend to other variables like TBK and TC, revealing how the same model produces varied biases for different outputs using identical inputs.

To consider the impact of the number of overall neurons, the number of inner layers was fixed. One can verify that model number 5 has an architecture that maintains a consistent number of overall neurons. By analysing the "Overall neurons" column, one will verify a notable trend: as the number of overall neurons increases, there is a general decrease in error metrics. Higher numbers of overall neurons correlate with lower MAE, MSE, and RMSE values, suggesting an enhanced capacity to capture the complex data patterns within the prediction of Saturation Time. However, it's worth noting that this trend isn't universally applicable, particularly evident in instances of excessively high neuron counts, such as the 64-neuron layers in model 2, which exhibits an efficiency loss, particularly noticeable in the MSE.

By analysing the significance of an inner layer with a higher number of neurons, a consistent improvement in the training dataset statistical indicators is observed in model 3, when compared to model 2. By reducing the number of neurons in the 4th, 5th, and 6th layers, a performance close to that of model 1 is achieved, although with five times fewer connections, a trait that avoids the vanishing gradient problem. Notably, model 5 outperforms model 4 and significantly surpasses model 6, indicating that a gradual decrease in performance is observed beyond the 64-neuron inner layer. In essence, model 5 exhibits a 30% reduction in connections compared to model 6 while still having superior performance for the training dataset.

Among the considered architectures, finally, architecture number 5 presents itself as the most promising candidate based on the evaluation of the mentioned statistical parameters. It achieves the lowest values across all error metrics. As a direct consequence of that, model/architecture number 5 was selected for prediction and validation within the test dataset.

Table 10 – Statistical performance of different architectures for ANN

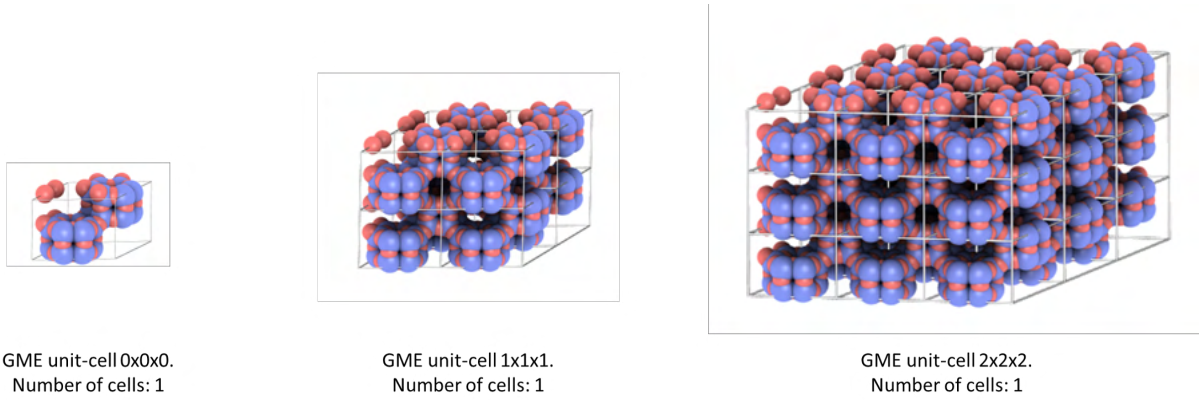
			Training dataset		
Architecture	Inner layers	Overall neurons	MAE	MSE	RMSE
Zero	5	15x15x15x15x15x15x1	3.96E-02	4.60E-03	6.29E-02
1	5	15x32x32x32x32x32x1	1.73E-02	6.88E-04	2.62E-02
2	5	15x32x64x64x64x64x1	2.33E-02	1.30E-03	3.63E-02
3	5	15x32x64x32x32x32x1	1.32E-02	4.04E-04	2.01E-02
4	5	15x32x64x15x15x15x1	1.58E-02	8.15E-04	2.85E-02
5	5	15x32x64x32x8x8x1	1.33E-02	3.74E-04	1.93E-02
6	5	15x32x64x32x15x8x1	3.14E-02	4.00E-03	4.57E-02

Source: Author (2024)

APPENDIX B – ADDITIONAL EXPLANATORY CONTENT

B.1 SUPERCELL CONCEPT VISUALIZATION

Figure 41 – Supercell concenpt illustrated



Source: Addapted from iRASPA software (DUBBELDAM et al., 2016)

B.2 ANN HYPERPARAMETER TUNING

Table 11 – Statistical indicators for the hyperparameter tuning of the ANN model

Function - Batch size - Epochs	**MAE**	**MSE**	**RMSE**	**R2**
Sigmoidal - 10 / 90	1.040	0.005	0.072	1.000
Sigmoidal - 10 / 50	1.050	0.005	0.073	1.000
Sigmoidal - 10 / 70	1.040	0.007	0.083	1.000
Sigmoidal - 10 / 30	1.050	0.012	0.109	0.999
Sigmoidal - 50 / 90	1.040	0.013	0.115	0.999
Sigmoidal - 70 / 90	1.030	0.098	0.313	0.993
Sigmoidal - 90 / 50	0.911	0.759	0.871	0.936
Sigmoidal - 10 / 10	0.961	4.070	2.022	0.578
Sigmoidal - 90 / 70	0.987	4.090	2.022	0.593
Sigmoidal - 50 / 50	0.974	4.110	2.033	0.580
Sigmoidal - 50 / 70	0.672	5.470	2.340	-0.162
Sigmoidal - 90 / 30	0.623	14.500	3.810	NAN
Sigmoidal - 70 / 10	0.626	14.600	3.810	NAN
Sigmoidal - 70 / 70	0.629	14.600	3.820	NAN
Sigmoidal - 70 / 30	0.628	14.600	3.820	NAN
Sigmoidal - 70 / 50	0.629	14.600	3.820	NAN
Sigmoidal - 50 / 30	0.630	14.600	3.820	NAN
Sigmoidal - 90 / 70	0.630	14.600	3.820	NAN
Sigmoidal - 50 / 10	0.573	14.700	3.830	NAN
Sigmoidal - 90 / 10	0.443	18.700	4.330	NAN
ReLu - 10 / 70	1.040	0.008	0.090	0.999
ReLu - 10 / 30	1.040	0.009	0.096	0.999
ReLu - 50 / 90	1.040	0.009	0.097	0.999
ReLu - 50 / 50	1.040	0.010	0.101	0.999
ReLu - 30 / 50	1.050	0.013	0.114	0.999
ReLu - 10 / 90	1.060	0.015	0.122	0.999
ReLu - 30 / 70	1.050	0.015	0.123	0.999
ReLu - 30 / 30	1.040	0.017	0.129	0.999
ReLu - 10 / 50	1.040	0.018	0.133	0.999
ReLu - 50 / 70	1.050	0.018	0.135	0.999
ReLu - 90 / 70	1.050	0.020	0.140	0.999
ReLu - 30 / 90	1.040	0.021	0.145	0.999
ReLu - 90 / 90	1.060	0.024	0.154	0.998
ReLu - 90 / 30	1.040	0.025	0.159	0.998
ReLu - 10 / 10	1.040	0.026	0.162	0.998
ReLu - 30 / 10	1.060	0.028	0.167	0.998
ReLu - 50 / 30	1.050	0.030	0.174	0.998
ReLu - 50 / 10	1.060	0.032	0.179	0.998
ReLu - 90 / 50	1.060	0.037	0.191	0.998
ReLu - 90 / 10	1.050	0.038	0.195	0.997

Source: Author (2024)