



UFRJ

UNIVERSIDADE FEDERAL
DO RIO DE JANEIRO

INSTITUTO DE QUÍMICA
JOÃO MATHEUS CHAGAS SOUSA

**Explorando algoritmos de aprendizado de
máquina para previsão de isotermas de CO₂
e N₂ em materiais reticulares**

Rio de Janeiro
2025

João Matheus Chagas Sousa

**Explorando algoritmos de aprendizado de máquina para
previsão de isotermas de CO₂ e N₂ em materiais
reticulares**

Projeto final de Curso apresentado ao corpo docente do Instituto de Química da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do grau de Químico Bacharel

Universidade Federal do Rio Janeiro – UFRJ

Instituto de Química

Orientador: Dr. Pierre Mothé Esteves

Coorientador: Dr. Felipe Lopes Oliveira

Brasil

2025

João Matheus Chagas Sousa

Explorando algoritmos de aprendizado de máquina para previsão de isotermas de CO₂ e N₂ em materiais reticulares/ João Matheus Chagas Sousa. – Brasil, 2025-93p. : il. (algumas color.) ; 30 cm.

Orientador: Dr. Pierre Mothé Esteves

Trabalho de Conclusão de Curso – Universidade Federal do Rio Janeiro – UFRJ Instituto de Química, 2025.

1. Adsorção. 2. Dióxido de Carbono. 2. Machine Learning. I. Pierre Mothé Esteves. II. Universidade Federal do Rio de Janeiro. III. Instituto de Química. IV. Estudo Computacional de Machine Learning para Descoberta de Novos Materiais Nanoporosos para Captura de CO₂.

João Matheus Chagas Sousa

Explorando algoritmos de aprendizado de máquina para previsão de isotermas de CO₂ e N₂ em materiais reticulares

Projeto final de Curso apresentado ao corpo docente do Instituto de Química da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do grau de Químico Bacharel

Trabalho aprovado. Brasil, 21 de Fevereiro de 2025:

Dr. Pierre Mothé Esteves
Orientador

Dr. Felipe Lopes de Oliveira
Coorientador

Dr. João Monnerat Araújo Ribeiro de Almeida
Instituto de Química - UFRJ

Dr. Itamar Borges Júnior
Instituto Militar de Engenharia -IME

Brasil
2025

Este trabalho é dedicado à memória dos meus avôs, José Rodrigues de Lima Sousa e Francisco Honorato Chagas, em reconhecimento ao legado de amor, sabedoria e valores que deixaram.

Agradecimentos

À minha família:

Tenho um imenso orgulho da minha família. Quando ainda não conhecemos o mundo, nem sempre conseguimos compreender a verdadeira importância dela, o quanto somos amados e cercados por pessoas tão incríveis.

Minha vida foi abençoada por ter três mães que me moldaram e me tornaram quem sou hoje. À minha vó Belinha, que sempre me envolveu com amor, cuidado e zelo desde a infância. Eu amava passar o tempo com ela e meu avô no interior. À minha mãe, que sempre esteve ao meu lado, me apoiando em tudo. E à minha querida vizinha Socorro, pelo carinho incondicional e pela atenção e cuidado que sempre teve comigo. Nada disto teria acontecido sem você. A gratidão que sinto por vocês é imensurável. Aos meus queridos avós, que partiram deixando imensa saudade.

Meu avô Chagas, com seu olhar doce para o mundo, me ensinou que o bem sempre retorna de alguma forma. Seu carinho e jeito de ser é um legado que minha família vai sempre carregar. E ao meu avô José Rodrigues, sou profundamente grato por sua confiança inabalável em mim. O senhor foi a pessoa que mais ouviu sobre meus sonhos, e no brilho dos seus olhos, eu via o quanto aquilo era grandioso para o senhor. Sem seu apoio constante, eu talvez não tivesse tido a coragem de sair da minha pequena cidade e alçar voos mais altos. Obrigado por tudo!

Às minhas queridas irmãs, Mardyla e Adrya, sou imensamente grato por todo o amor, carinho e pelas preciosas trocas de afeto. Ter vocês em minha vida é um privilégio. Também não poderia deixar de expressar minha gratidão à minha madrinha, a quem carinhosamente chamo de mãe Jé. Meu muito obrigado por tudo!

À minha segunda família em São Luís:

Obrigada, tia Márcia, Marnand, Yaçana e, em especial, Madson, que foi um verdadeiro pai para mim nesse período tão único. Foram anos inesquecíveis e repletos de felicidade.

Aos meus amigos:

Aos grandes amigos que colecionei ao longo do caminho e que sempre estarão no meu coração: Ben-Hur, Lucas Moreira e, claro, meu eterno Moção, Louise. Vocês são parte essencial da minha jornada.

Aos meus amigos da faculdade e de laboratório, que tornaram essa trajetória mais leve, divertida e amável: obrigado, Arthur, Beatriz, João Vitor, Leon, Perola, Victor Lessa, Marcele, Pedro, Jéssica e Hugo.

A Eva, por ser uma amiga incrível desde o início da faculdade e hoje minha parceira

de vida, sempre ao meu lado, me apoiando em tudo.

Aos meus mentores e professores:

Sou profundamente grato aos grandes mestres que moldaram minha trajetória: Pierre Esteves, Márcio Contrucci, Ricardo Oliveira, Alexandre Rocha e Thiago Cardozo. Sem vocês, eu não teria desenvolvido a base sólida e a paixão pela ciência que me guiam até hoje.

Aos meus dois pais científicos e amigos, Felipe Lopes e Marcos Aurélio. Ao Felipe, por sempre exigir o meu melhor e, ao mesmo tempo, ser um amigo e um apoio constante. E ao Marcos, em especial, por acompanhar minha trajetória de perto, pelos conselhos, pela orientação e pelo carinho e amor inestimável.

Às famílias que me acolheram no Rio de Janeiro:

Meu agradecimento especial à tia Lourdes e à tia Cláudia, por me receberem com tanto amor e carinho.

Ao meu maior achado na graduação:

Por fim, não poderia deixar de mencionar meu grande melhor amigo, Gabriel Teixeira. Você esteve comigo em tantas situações ao longo dessa jornada. Essa conquista é sua também, pela amizade e pela conexão eterna que criamos.

*“Em todos os gestos,
uma quinta dimensão
(Dom Helder Camara, O Deserto é Fértil)*

Resumo

A intensificação das emissões de CO₂ devido às atividades humanas representa um dos maiores desafios ambientais da atualidade, exigindo o desenvolvimento de tecnologias eficazes para captura e armazenamento. Entre as estratégias disponíveis, os materiais reticulares se destacam por sua alta área específica e seletividade na adsorção de gases. Contudo, identificar os materiais mais eficientes para captura de CO₂ sob diferentes condições de pressão e temperatura ainda é um desafio, devido ao grande volume de estruturas disponíveis em bancos de dados. Neste trabalho, foi adotada uma abordagem sistemática para a geração de dados de adsorção de CO₂ e N₂ em materiais reticulares. As quantidades de adsorção foram determinadas por meio de simulações GCMC, que modelaram a interação entre os materiais e as moléculas-alvo. Os dados simulados foram ajustados ao modelo de isoterma de Langmuir de dois sítios (DSL), e os parâmetros resultantes foram utilizados para treinar diferentes modelos de aprendizado de máquina (ML). O desempenho dos modelos foi avaliado, e o mais preciso foi utilizado para prever a capacidade de adsorção de novos materiais. Essa abordagem possibilitou a triagem eficiente de materiais promissores, reduzindo a necessidade de simulações computacionais intensivas e contribuindo para a otimização de processos de captura de carbono.

Palavras-chave: Aprendizado de Máquina. Materiais Nanoporosos. Adsorção. Captura de CO₂.

Abstract

The intensification of CO₂ emissions due to human activities represents one of the greatest environmental challenges of today, requiring the development of effective technologies for capture and storage. Among the available strategies, reticular materials stand out due to their high specific surface area and selectivity in gas adsorption. However, identifying the most efficient materials for CO₂ capture under different pressure and temperature conditions remains a challenge due to the large volume of structures available in databases. In this study, a systematic approach was adopted to generate adsorption data for CO₂ and N₂ in reticular materials. The adsorption quantities were determined through GCMC simulations, which modeled the interaction between the materials and the target molecules. The simulated data were fitted to the dual-site Langmuir (DSL) isotherm model, and the resulting parameters were used to train different machine learning (ML) models. The performance of the models was evaluated, and the most accurate one was used to predict the adsorption capacity of new materials. This approach enabled an efficient screening of promising materials, reducing the need for computationally intensive simulations and contributing to the optimization of carbon capture processes.

Keywords: Machine learning. Nanoporous materials. Adsorption. CO₂ capture.

Lista de ilustrações

Figura 1 – Emissões anuais globais de CO ₂ e distribuição por setor.	18
Figura 2 – Esquema sistemas captura de CO ₂	20
Figura 3 – Classificação de isothermas de adsorção combinando propostas da IUPAC	23
Figura 4 – Potencial de interação poro-água, W , através do poro e energia interna água-água, como uma função da largura do poro (Z). Na parte a, as larguras de poro consideradas são 0,7, 1,0 e 2,0 nm. Os potenciais são calculados assumindo a posição central de uma molécula de água adsorvida.	25
Figura 5 – Diversidade estrutural de COFs e MOFs	27
Figura 6 – Fluxo tradicional de descoberta de novos materiais	28
Figura 7 – Abordagens para descoberta computacional de materiais	29
Figura 8 – Exemplo de <i>Ensemble grand-canônico</i>	33
Figura 9 – Modelos de isothermas de Langmuir	43
Figura 10 – Representação dos grupos de descritores RACs	52
Figura 11 – Comparação entre as principais propriedades geométricas das estruturas presentes no CRAFTED e no QMOF	55
Figura 12 – Isothermas de adsorção para três materiais reticulares distintos nas temperaturas de 273 K, 298 K e 323 K, juntamente com as curvas ajustadas pelo modelo <i>dual-site Langmuir</i> (DSL).	56
Figura 13 – Ajuste inicial para a isoterma de adsorção com base nos modelos de Henry e Langmuir.	58
Figura 14 – Fluxograma representando o processo de ajuste dos parâmetros do modelo dual-site Langmuir (DSL) e single-site Langmuir (SSL)	60
Figura 15 – Distribuição dos parâmetros do modelo DSL para os diferentes ajustes aplicados aos dados de simulação GCMC.	61
Figura 16 – Distribuição das métricas de avaliação para os diferentes ajustes aplica- dos aos dados de simulação GCMC.	62
Figura 17 – Isothermas de adsorção de um MOF nas temperaturas de 273 K, 298 K e 323 K, juntamente com as curvas ajustadas pelo modelo <i>dual-site</i> <i>Langmuir</i> (DSL).	62
Figura 18 – Distribuições dos ajustes realizados nos dados de simulação GCMC para o ajuste final escolhido: Coeficiente de Determinação (R^2), Erro Quadrático Médio (MSE) e Porcentagem de Variância Explicada (EVP	62
Figura 19 – Distribuição dos parâmetros ajustados da isoterma DSL	63

Figura 20 – Comparação entre os valores de quantidade adsorvida de CO ₂ e N ₂ obtidos por simulações GCMC e os valores previstos pelo modelo <i>dual-site Langmuir</i> (DSL)	65
Figura 21 – Distribuição dos parâmetros ajustados da isoterma DSL em escala Logarítmica	67
Figura 22 – Comparação dos desempenhos dos modelos de aprendizado de máquina nas métricas R^2 e MSE.	68
Figura 23 – Correlação entre Valores de Referência e Valores Preditos pelo Modelo Random Forest	70
Figura 24 – Isotermas de adsorção para três materiais reticulares distintos junto com as isotermas obtidas pelo ajuste DSL e pelo modelo de ML	71
Figura 25 – Distribuição da quantidade de gás adsorvido quando estimado pelos parâmetros ajustados via GCMC com os valores preditos pelo modelo Random Forest em comparação com os obtidos via simulação GCMC	72
Figura 26 – Taxas de erro em 14 bancos de dados de MOFs	73
Figura 27 – Distribuição de descritores geométricos do banco de dados ARC-MOF.	74
Figura 28 – Distribuição dos descritores RAC no banco de dados ARC-MOF.	75
Figura 29 – Capacidade de trabalho versus seletividade de CO ₂ para os materiais reticulares do ARC-MOF e do Banco de estruturas utilizado no treinamento dos modelos	76
Figura 30 – Comparação dos desempenhos dos modelos de aprendizado de máquina nas métricas R^2 e MSE.	88

Lista de tabelas

Tabela 1 – Parâmetros críticos for CO ₂ e N ₂	91
Tabela 2 – Parâmetros de Lennard-Jones para os campos de força UFF e DREIDING.	91
Tabela 3 – Bases de Dados de Materiais Porosos Cristalinos	92

Lista de abreviaturas e siglas

GGE	Gases do Efeito Estufa
CCS	Carbon capture and storage
MEA	Monoetanolamina Aquosa
IGCC	Integrated Gasification Combined Cycle
COF	Covalent Organic Framework
MOF	Metal organic framework
CSD	Cambridge structural database
ML	Machine Learning
GCMC	<i>Grand-canônico</i> Monte Carlo

Sumário

1	INTRODUÇÃO	17
1.1	Captura e Armazenamento de Carbono - CCS	19
1.1.1	Sistemas de Captura de CO ₂	19
1.1.2	Tecnologias de captura de CO ₂	21
1.2	Adsorção em sólidos	21
1.2.1	Isotermas de Adsorção	22
1.2.2	Efeito de Confinamento	24
1.3	Materiais Reticulares	25
1.4	A Evolução dos Paradigmas na Ciência dos Materiais	27
1.4.1	Abordagem Tradicional	27
1.4.2	Previsões Baseadas em Leis Físicas e Químicas	27
1.4.3	Simulação Computacional	28
1.4.4	Aprendizado de Máquina (<i>Machine Learning</i>)	28
1.4.4.1	Aplicação na descoberta de materiais para captura de CO ₂	29
2	OBJETIVOS	31
2.1	Objetivos específicos	31
3	METODOLOGIA	32
3.1	Banco de Dados	32
3.2	Simulações Monte Carlo no <i>Ensemble</i> Grand Canônico (GCMC)	32
3.2.1	<i>Ensemble grand-canônico</i> (μVT)	33
3.2.2	Ergodicidade	33
3.2.3	Algoritmo Monte Carlo Metropolis	34
3.2.3.1	Passos do Monte Carlo	37
3.2.4	Campo de Força e Cargas Parciais	38
3.3	Ajuste de Isotermas de adsorção	39
3.3.1	Modelo de Langmuir	40
3.3.1.1	Modelo de Langmuir de dois sítios	42
3.3.2	Ajuste dos modelos de adsorção	44
3.3.3	Modulo <code>curve_fit</code>	44
3.3.4	Métricas de Avaliação para o Ajuste	45
3.3.4.1	Coeficiente de Determinação (R^2)	45
3.3.4.2	Erro Quadrático Médio (MSE)	45
3.3.4.3	Porcentagem de Variância Explicada (EVP)	46
3.3.5	Métricas de Desempenho	46

3.3.5.1	Capacidade de Trabalho	46
3.3.5.2	Seletividade	47
3.4	Aprendizado de Máquina	47
3.4.1	Decision Tree	48
3.4.2	Random Forest	48
3.4.3	XGBoost	49
3.4.4	Redes neurais	50
3.4.5	Descritores	50
3.4.5.1	Descritores Geométricos de materiais nanoporosos	51
3.4.5.2	Descritores Químicos de materiais nanoporosos	51
4	RESULTADOS E DISCUSSÕES	53
4.1	Banco de Estruturas	54
4.1.1	QMOF	54
4.1.2	CRAFTED	54
4.1.3	Comparação dos Banco de dados	54
4.2	Ajuste ao modelo DSL	56
4.2.1	O viés estatístico das equações linearizadas	57
4.2.2	Processo de ajuste	57
4.3	Aprendizado de Máquina	65
4.3.1	Métricas para Avaliação de Modelos de Aprendizado de Máquina	65
4.3.2	Predição dos parâmetros da Isoterma	66
4.3.2.1	Escolha dos Descritores e de Modelo	67
4.3.2.2	Modelo Random Forest	69
4.4	Avaliando o modelo fora do conjunto de treinamento	72
4.4.1	Aplicando Métricas de Desempenho	74
5	CONCLUSÕES E PERSPECTIVAS	77
5.1	Conclusões	77
5.2	Perspectivas	78
	REFERÊNCIAS	79
	APÊNDICES	87
	APÊNDICE A – DADOS MODELOS DE APRENDIZADO DE MÁ- QUINA	88
A.1	Modelos de Aprendizado de Máquina - Dados sem escala Logarítmica	88

	ANEXOS	90
	APÊNDICE A – TABELAS E DADOS	91
A.1	Descritores Geométricos	92

1 INTRODUÇÃO

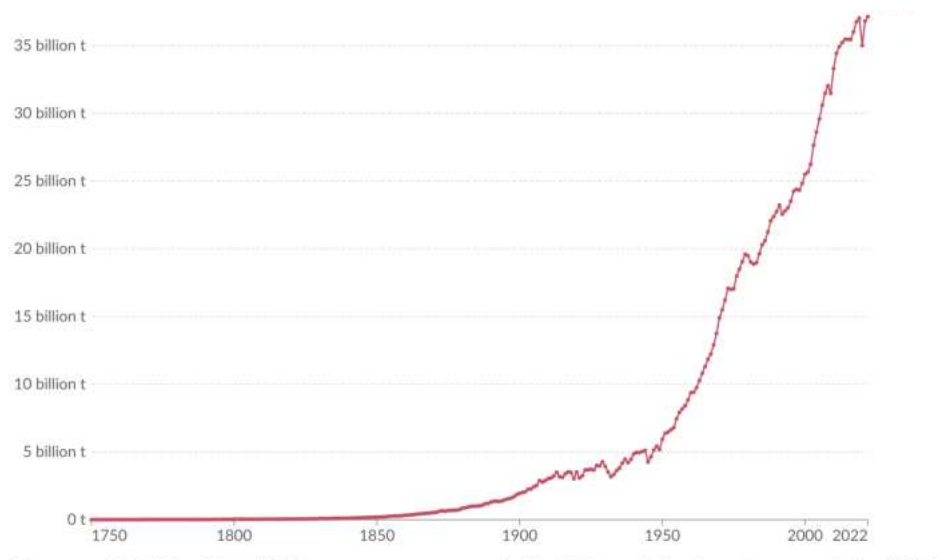
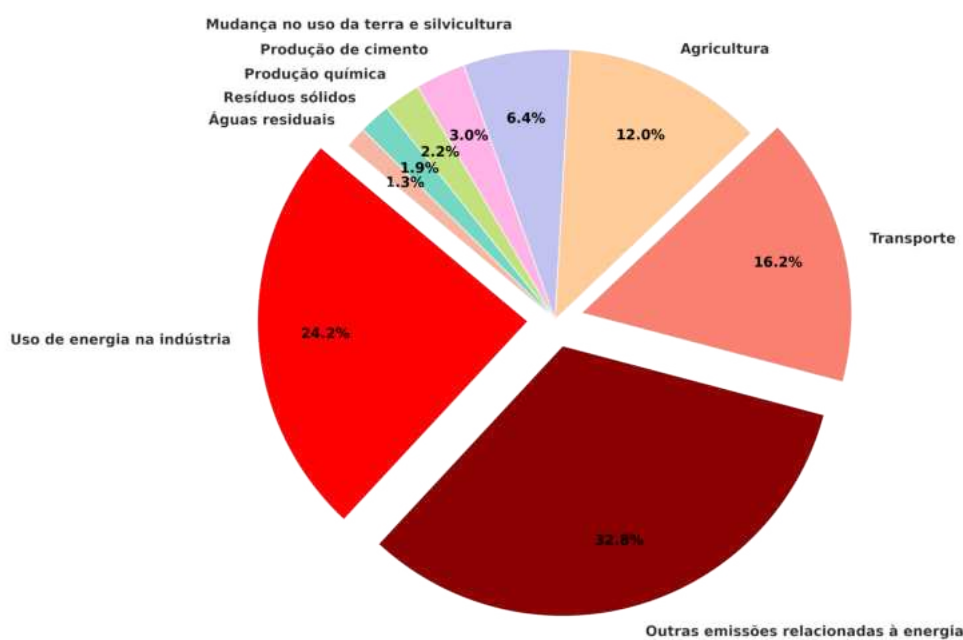
O clima e o tempo têm uma influência significativa em diversos aspectos da vida humana, como na agricultura, abastecimento de água e na manutenção dos ecossistemas que sustentam a biodiversidade planetária.^[1] No entanto, essas condições estão se alterando de forma acelerada, com o aquecimento global como fator principal. O aumento das temperaturas globais tem causado sérios problemas para a sociedade e o meio ambiente, com impactos já observáveis em diversas áreas.^[2]

A principal causa dessas mudanças climáticas é o aumento das emissões de dióxido de carbono (CO_2), um gás de efeito estufa produzido por atividades humanas como a queima de combustíveis fósseis, o desmatamento e processos industriais.^[3] O CO_2 acumula-se na atmosfera e intensifica o efeito estufa, que embora seja natural e necessário para manter a temperatura da Terra, tem sido amplificado pela ação humana, resultando no aquecimento global.^[4]

O aquecimento global, por sua vez, tem gerado uma série de desafios para a sociedade humana, incluindo eventos climáticos extremos, escassez de recursos hídricos e impactos negativos sobre a biodiversidade.^[5] A mitigação das emissões de CO_2 tornou-se portanto, uma questão central nas agendas políticas globais.

As principais fontes de emissões de CO_2 são as usinas de energia baseadas em combustíveis fósseis e as plantas indústrias de ferro, aço e cimento, que têm um grande impacto ambiental e são os principais alvos de estratégias de captura de carbono.^[2] A [Figura 1](#) ilustra as emissões anuais de CO_2 provenientes da queima de combustíveis fósseis, onde é apresentado também a contribuição de cada setor. Essa análise destaca a predominância das indústrias de produção de energia como principais contribuintes para o aumento das concentrações de CO_2 na atmosfera, em comparação com as demais atividades.

Para combater esse problema, uma solução promissora é a captura e armazenamento de CO_2 , que pode ajudar a reduzir a quantidade desse gás na atmosfera e assim, diminuir os impactos do aquecimento global.^[7] No entanto, embora a captura de CO_2 seja uma estratégia eficaz, ela enfrenta desafios significativos relacionados ao custo e à eficiência dos materiais utilizados para essa captura.^[8] As tecnologias existentes, como os materiais para a captura de carbono, têm se mostrado promissoras, mas ainda apresentam limitações quanto à capacidade de adsorção e seletividade. Isso destaca a necessidade de materiais mais eficientes para esse fim, que possam ser aplicados de maneira viável em grande escala.^[9]

Figura 1 – Emissões anuais globais de CO₂ e distribuição por setor.(a) Emissões anuais de CO₂ geradas em nível global.Fonte: Retirado de Ritchie e Roser^[6](b) Distribuição das emissões anuais de CO₂ por setor.Fonte: Adaptado de Ritchie e Roser^[6]

Uma solução em potencial para esse problema é o uso de materiais reticulares, que possuem uma grande área de superficial e seletividade, características ideais para a captura de CO₂.^[10] Entretanto, a busca por materiais com maior capacidade de adsorção é um desafio contínuo, pois envolve a necessidade de otimizar tanto as propriedades dos materiais quanto as condições operacionais.^[11]

Diante disso, a combinação de simulações computacionais com modelos de aprendizado de máquina surge como uma abordagem poderosa para prever e selecionar materiais com maior potencial para captura de CO₂. Através de simulações de adsorção é possível identificar os parâmetros que governam esse processo, permitindo o treinamento de modelos preditivos que podem guiar a seleção de materiais mais eficientes para a captura de CO₂.^[12] No entanto, mesmo com esses avanços ainda há limitações nas simulações que precisam ser superadas para garantir que os modelos preditivos possam ser aplicados de forma precisa em condições reais de operação.

Este trabalho propõe-se a investigar como simulações Monte Carlo no *Ensemble grand canônico*, combinadas com modelos de aprendizado de máquina, podem ser usados para melhorar a predição e a seleção de materiais com maior capacidade de captura de CO₂.

1.1 Captura e Armazenamento de Carbono - CCS

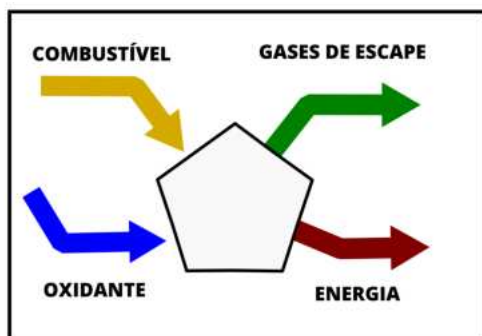
Para enfrentar o aquecimento global, é necessário reduzir as emissões de gases de efeito estufa, sendo as principais estratégias a proteção das florestas e a transição de fontes de energia não renováveis para renováveis.^[13] Contudo, essas ações embora essenciais, demandam um longo prazo para surtirem efeito significativo, o que dificulta o combate imediato ao problema.

Nesse cenário, a captura e armazenamento de CO₂ (CCS, do inglês Carbon Capture and Storage) apresenta-se como uma solução promissora para reduzir emissões atmosféricas de CO₂ provenientes de atividades humanas de forma mais rápida.^[3] Entretanto, a implementação de sistemas de captura enfrenta desafios como a complexidade técnica e os custos elevados associados às infraestruturas necessárias.

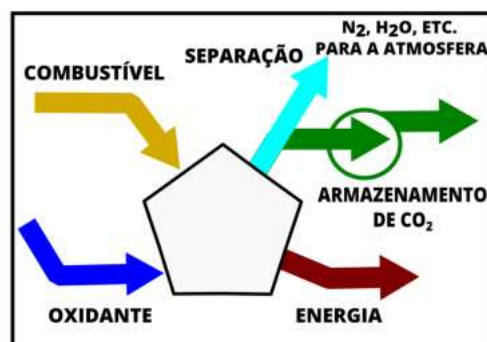
1.1.1 Sistemas de Captura de CO₂

Para superar esses desafios, os sistemas de captura de CO₂ são projetados para extrair dióxido de carbono diretamente de correntes gasosas industriais ou atmosféricas, permitindo a sua redução efetiva em setores como geração de energia, indústrias químicas e processos de refino. A captura de CO₂ pode ser realizada por meio de tecnologias específicas como a captura pós-combustão e pré-combustão, ilustradas na [Figura 2](#).^[14]

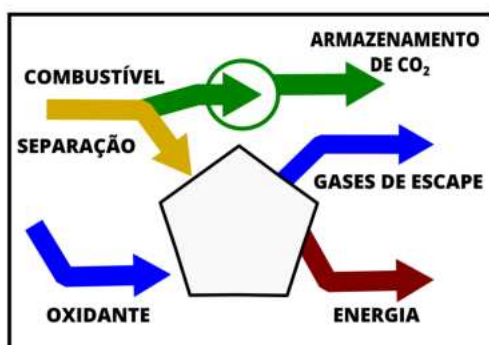
Na pré-combustão, destacam-se a oxi-combustão e o ciclo combinado de gaseificação integrada (IGCC). Na oxi-combustão, utiliza-se oxigênio puro em vez de ar atmosférico, o que resulta em um gás de combustão composto principalmente por CO₂ e H₂O, facilitando a separação do CO₂ por condensação da água.^[15] No entanto, a produção de oxigênio por separação criogênica é altamente intensiva em energia, o que limita sua aplicação em larga escala. Para superar essa limitação, alternativas como membranas e ciclos químicos

Figura 2 – Esquemas representando diferentes processos de captura de CO_2 .

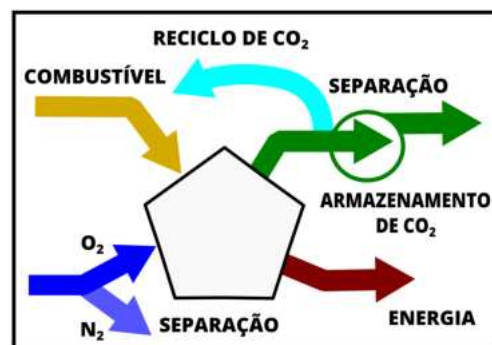
(a) Diagrama esquemático da geração de energia baseada em combustíveis fósseis



(b) Diagrama esquemático de captura pós-combustão



(c) Esquemático diagrama de captura pré-combustão



(d) Diagrama esquemático da combustão do oxi-combustão

Fonte: Próprio autor

para produção de oxigênio estão sendo investigadas, mas até o momento, nenhuma opção economicamente viável foi encontrada.^[16]

O IGCC, por sua vez, gaseifica combustíveis como carvão para produzir gás de síntese rico em CO e H_2 . Em seguida, o CO reage com vapor d'água em um reator catalítico, gerando CO_2 e H_2 , permitindo a separação do CO_2 por absorção enquanto o hidrogênio é utilizado em aplicações industriais.^[3] Apesar de eficiente, o custo elevado dos processos de gaseificação e a necessidade de catalisadores específicos representam desafios que ainda precisam ser superados.

Já no sistema de captura pós-combustão, o CO_2 é removido dos gases de exaustão antes de sua liberação na atmosfera, podendo ser armazenado em reservatórios geológicos. Este método apresenta a vantagem de ser compatível com plantas industriais existentes, dispensando a necessidade de novas infraestruturas.^[16] No entanto, o elevado consumo energético dos equipamentos de captura ainda limita sua eficiência e custo-benefício, destacando a necessidade de desenvolvimento contínuo de sistemas mais eficientes para a implementação em larga escala.

1.1.2 Tecnologias de captura de CO₂

Além dos sistemas mencionados, as tecnologias de captura de CO₂ dividem-se em três principais processos: absorção em solventes, adsorção em sólidos e separação por membranas.^[16]

A absorção em solventes, como a monoetanolamina aquosa (MEA), é amplamente empregada devido à sua capacidade de separar seletivamente o CO₂. Esse processo ocorre por formação de carbamatos em uma reação química entre o CO₂ e as aminas. No entanto, apresenta desvantagens significativas, como o alto consumo energético para regeneração, corrosividade e sensibilidade a contaminantes como SO_x e NO_x, que reduzem sua eficiência.^[15, 9]

A separação por membranas utiliza materiais como polímeros ou cerâmicas para promover a permeabilização seletiva do CO₂. Embora promissora, essa tecnologia ainda não alcançou maturidade para aplicações industriais em larga escala, devido a desafios relacionados à durabilidade e eficiência de separação.^[3]

Por outro lado, a adsorção em sólidos concentra o CO₂ na superfície de materiais adsorventes. Esse processo, por ser interfacial e ocorrer entre sólidos e gases, apresenta maior controle das propriedades dos adsorventes. Por isso, a eficiência dessa abordagem é limitada pela estabilidade térmica e química dos materiais, especialmente devido ao custo energético associado à regeneração do adsorvente para um novo ciclo de captura. Nesse contexto, o desenvolvimento de materiais com melhores propriedades é fundamental para minimizar esse custo e tornar a tecnologia mais viável e sustentável.^[17, 18, 19]

1.2 Adsorção em sólidos

Embora cada uma dessas tecnologias apresente limitações, a adsorção em sólidos desponta como uma das alternativas mais promissoras devido à sua alta eficiência seletiva e à flexibilidade no design de materiais adsorventes. Este método permite maior controle sobre as propriedades dos materiais, possibilitando otimizações direcionadas à captura de CO₂ sob condições específicas de operação. No entanto, sua aplicação em larga escala ainda depende do desenvolvimento de adsorventes com alta capacidade de adsorção e seletividade por CO₂, além de estabilidade térmica e química capazes de suportar múltiplos ciclos de regeneração com eficiência energética otimizada.^[8]

No contexto da adsorção, as moléculas de CO₂ interagem com a superfície do adsorvente por meio de interações não específicas, como as de van der Waals e eletrostáticas (fisissorção), ou interações mais fortes e específicas, como as ligações químicas (quimissorção). A fisissorção é caracterizada por interações fracas e reversíveis, enquanto a quimissorção envolve a formação de ligações químicas mais fortes. A eficiência na purifica-

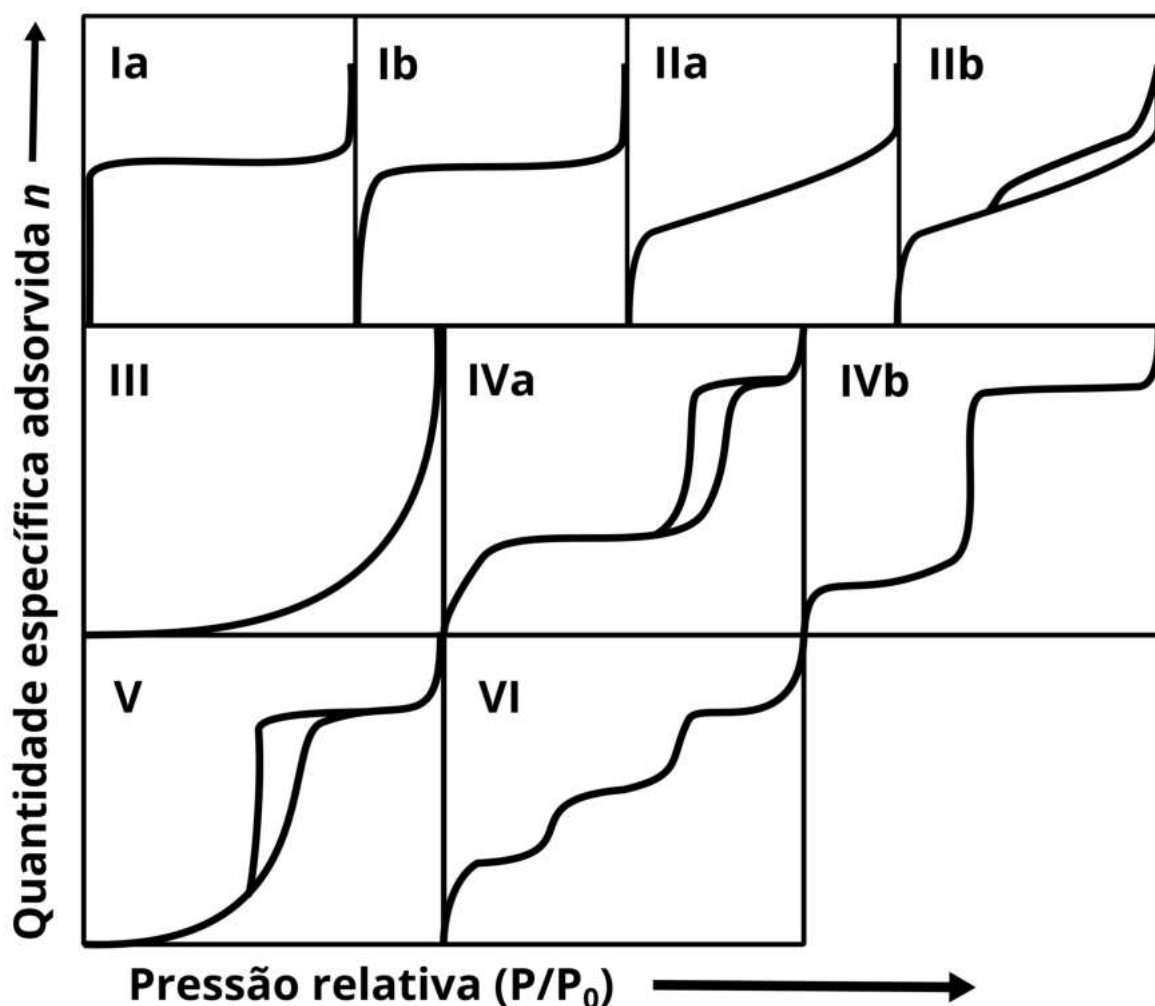
ção e concentração do CO_2 em ciclos de adsorção, utilizando variações de pressão, vácuo ou temperatura, está diretamente relacionada à natureza dessas interações. A adsorção em sólidos oferece uma abordagem eficiente para processos de separação, com a vantagem de permitir a regeneração do adsorvente, tornando o processo mais sustentável e econômico.^[9]

1.2.1 Isotermas de Adsorção

As isotermas de adsorção fornecem informações cruciais sobre a afinidade do adsorvato com a superfície do adsorvente e o comportamento do sistema em equilíbrio, em uma dada temperatura. Por sua vez, desempenham um papel fundamental na compreensão do processo de adsorção, pois descrevem a variação na quantidade de gás adsorvido por um material em função da pressão, mantendo a temperatura constante. Além disso, são amplamente aplicadas no estudo de processos industriais, como separação química, catálise e caracterização de sólidos porosos.^[20] A modelagem matemática das isotermas permite aprofundar a compreensão do fenômeno de adsorção e otimizar esses processos.^[21]

As isotermas experimentais, que variam conforme os sistemas gás-sólido, apresentam formas distintas que ajudam a revelar características preliminares da estrutura dos poros do adsorvente. A classificação estendida da IUPAC, apresentada na [Figura 3](#), divide as isotermas em nove tipos, e outras formas podem ser explicadas por combinações de diferentes tipos, sendo classificadas como isotermas compostas.^[22] Os tipos I, II, III, IV e V, inicialmente propostos por Brunauer, Deming, Deming e Teller (1940),^[23] são conhecidos como a classificação BDDT ou de Brunauer. Essas isotermas fornecem uma base importante para entender o comportamento de adsorventes e otimizar processos de adsorção.

Figura 3 – Classificação de isotermas de adsorção combinando propostas da IUPAC: O **Tipo I** é caracterizado pela adsorção em microporos. Dentro do Tipo I, a isotermas **I(a)** referem-se a microporos estreitos, enquanto **I(b)** indicam microporos mais largos. O **Tipo II(a)** envolve a adsorção multimolecular em superfícies não porosas ou macroporosas, com uma transição entre monocamada e multicamadas. A isotermas **II(b)** exibe histerese devido à condensação capilar entre partículas agregadas. O **Tipo III** é caracterizado por adsorção fraca em materiais não porosos ou macroporosos, sem formação clara de monocamada. O **Tipo IV** refere-se à adsorção em mesoporos, com platô de saturação em altas pressões. A isotermas **IV(a)** exibe histerese devido à condensação capilar nos mesoporos, enquanto a **IV(b)** é reversível, com histerese menos pronunciada, encontrada em materiais ordenados. O **Tipo V** é caracterizado pela adsorção fraca em micro ou mesoporos, com histerese associada ao preenchimento e esvaziamento dos poros. Por fim, o **Tipo VI** representa a adsorção em camadas sucessivas em superfícies altamente uniformes, como o carbono grafitizado.



Fonte: Próprio autor.

1.2.2 Efeito de Confinamento

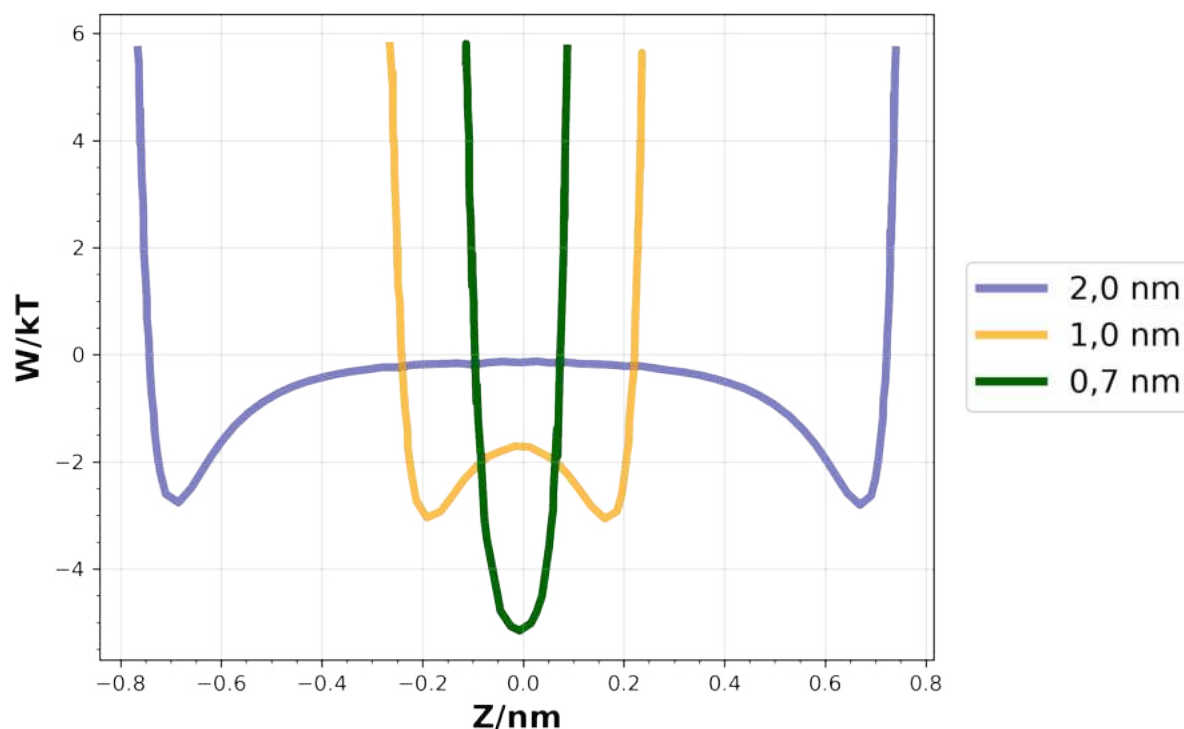
Os materiais nanoporosos apresentam uma característica fundamental: o efeito de confinamento, que influencia diretamente as propriedades de adsorção em poros de dimensões nanométricas. Esse confinamento altera as interações entre o adsorvato e o adsorvente, especialmente no processo de fissorção, que é dominado por interações físicas mais fracas, como as forças de van der Waals. A curvatura da superfície desses materiais aumenta a densidade de interações entre as moléculas adsorvidas e a superfície, o que amplia a capacidade de adsorção sob determinadas condições. Estudos mostram que a interação entre o adsorvato e o adsorvente se intensifica à medida que a largura do poro diminui, resultando em um aumento da energia de interação devido ao confinamento.^[24]

Por esse motivo, os poros na escala nanométrica geram mais interesse para maximizar a densidade de interações. Segundo a classificação da IUPAC, os materiais nanoporosos são divididos com base no tamanho dos poros: microporosos (0,2 a 2 nm), mesoporosos (2 a 50 nm) e macroporosos (50 a 1000 nm).^[25]¹

Esse efeito de confinamento é evidenciado na [Figura 4](#), que mostra como o potencial de interação entre o poro e a água (W) e a energia interna entre moléculas de água (U_{ww}) variam com a largura do poro. Quando a curvatura do material aumenta, a interação de van der Waals se intensifica, o que modifica as propriedades de adsorção em comparação com materiais menos “curvados”. Essa intensificação das interações resulta em um aumento significativo da capacidade de adsorção, o que é particularmente relevante em aplicações como a captura de CO_2 , onde o desempenho do material nanoporoso pode ser otimizado por essas interações.

¹ A utilização de materiais com poros superiores a 100 nm, entretanto, deve ser feita com cautela, pois pode comprometer o controle da porosidade, especialmente em processos que exigem alta precisão na adsorção.^[19]

Figura 4 – Potencial de interação poro-água, W , através do poro e energia interna água-água, como uma função da largura do poro (Z). Na parte a, as larguras de poro consideradas são 0,7, 1,0 e 2,0 nm. Os potenciais são calculados assumindo a posição central de uma molécula de água adsorvida.



Fonte: Adaptado de: [Striolo et al.^{\[26\]}](#)

1.3 Materiais Reticulares

A eficiência dos processos de adsorção depende diretamente das propriedades do material utilizado, especialmente da sua estrutura e área específica. Embora diversos adsorventes sólidos, como carvão ativado, zeólitas, óxidos metálicos e hidrotalcitas tenham sido amplamente investigados para aplicações de captura de CO_2 ,^[27] as zeólitas se destacam devido à sua alta seletividade e capacidade de adsorção.^[28] No entanto, essas materiais apresentam limitações, como a necessidade de pressões elevadas (acima de 2 bar) para operação eficiente e uma capacidade de adsorção reduzida na presença de umidade no gás. Além disso, a regeneração das zeólitas exige temperaturas acima de 300 °C, o que aumenta o custo e a complexidade do processo.^[29]

Uma alternativa promissora são os materiais reticulares, que têm atraído grande atenção devido ao seu alto potencial para captura de CO_2 . Com estruturas orgânicas ou inorgânicas que exibem porosidade regular em escala nanométrica, esses materiais apresentam grande área específica,² tornando-os ideais para aplicações em processos de

² O atual recordista, o MOF DUT-60, apresenta uma área superficial de 7.800 m²/g,^[30] evidenciando o potencial dessa classe de materiais para maximizar a captura de CO_2 .

de adsorção.^[11] A química reticular, que estuda a ligação de entidades químicas discretas por meio de ligações fortes para formar estruturas estendidas, é fundamental para o desenvolvimento desses materiais.^[31] Dentro dessa classe, destacam-se as estruturas metal-orgânicas (MOFs, do inglês *metal-organic frameworks*) e as estruturas orgânicas covalentes (COFs, do inglês *Covalent Organic Frameworks*).^[32]

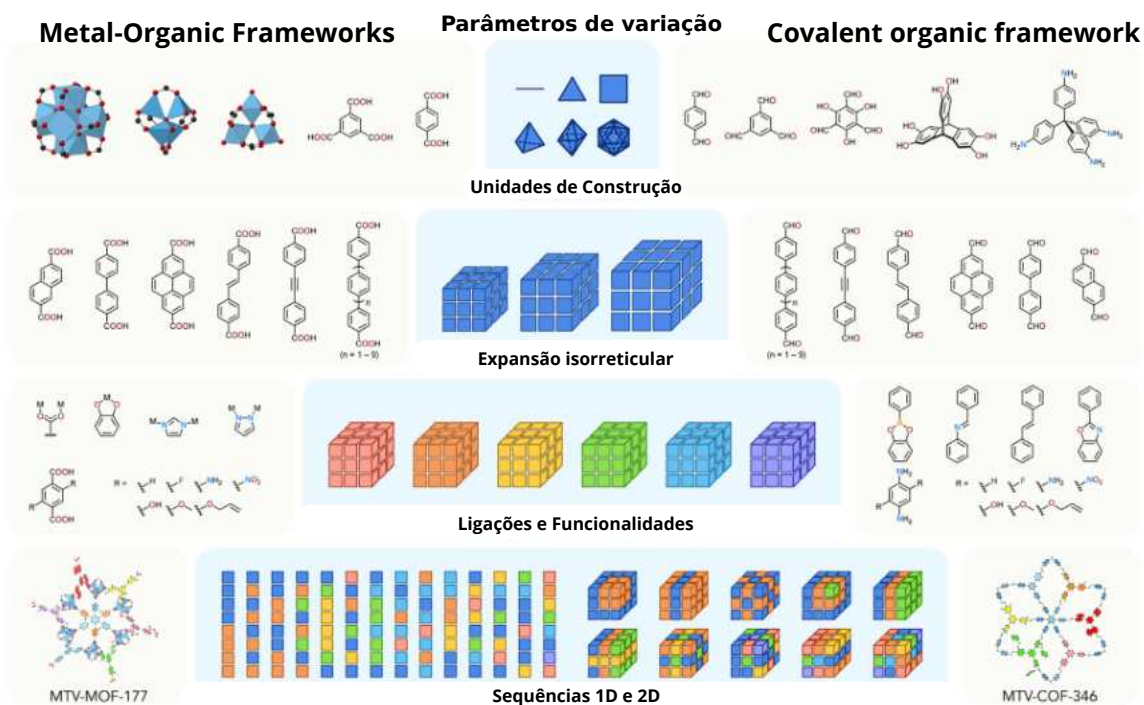
Os MOFs, também conhecidos como polímeros de coordenação porosos, são compostos por aglomerados polinucleares de metais, denominados unidades de construção secundárias (SBUs, do inglês *secondary building units*), unidos por ligantes orgânicos para formar estruturas com porosidade e cristalinidade permanentes.^[33] Esses materiais foram um dos primeiros sólidos cristalinos a serem desenvolvidos no campo da química reticular e devido à sua capacidade de sintonia estrutural e funcional, a área dos MOFs tornou-se uma das mais promissoras nesse campo.^[33]

Por outro lado, as estruturas formadas apenas por unidades de construção orgânicas ligadas covalentemente, conhecidas como COFs, também são materiais reticulares notáveis. Eles são definidos como uma classe de materiais orgânicos porosos e cristalinos com blocos de construção ligados covalentemente, formando um material com porosidade permanente, estrutura e topologia bem definidas.^[31, 34]

A quantidade de materiais reticulares disponíveis conta com mais de 100.000 estruturas de MOFs no Banco de Dados Estruturais de Cambridge (CSD, do inglês *Cambridge Structural Database*).^[35] A Tabela 3 apresenta os principais bancos de dados para MOFs e COFs, evidenciando a enorme diversidade desses materiais. Sua natureza modular sugere que há uma quantidade virtualmente infinita de estruturas possíveis,^[36] o que, por sua vez, aumenta a complexidade da investigação experimental necessária para testar todos os materiais dentro desse vasto espaço químico.^[32]

A capacidade de controle e ajuste das características químicas e estruturais desses materiais tem direcionado diversos esforços para encontrar essas propriedades de interesse.^[37] A Figura 5 ilustra a diversidade estrutural dos MOFs e COFs, destacando as múltiplas possibilidades de design e a combinação de fatores como topologia, tipos de blocos de construção e grupos químicos adicionados.^[36] Embora os materiais reticulares mostrem grande promessa, o desafio é explorar e identificar novas estruturas com propriedades superiores, como por exemplo, dentre as milhões de possibilidades disponíveis, o melhor candidato com capacidade de adsorção para captura de CO₂.

Figura 5 – diversidade estrutural de COFs e MOFs



Fonte: Adaptado de: [Lyu et al.](#)^[36]

1.4 A Evolução dos Paradigmas na Ciência dos Materiais

1.4.1 Abordagem Tradicional

A descoberta de novos materiais tradicionalmente segue um fluxo de trabalho baseado em tentativa e erro, conforme ilustrado na [Figura 6](#). Inicialmente, seleciona-se um material para testes experimentais. Caso o desempenho não atenda às expectativas, os resultados orientam a escolha de novos materiais ou ajustes nos previamente analisados.^[11]

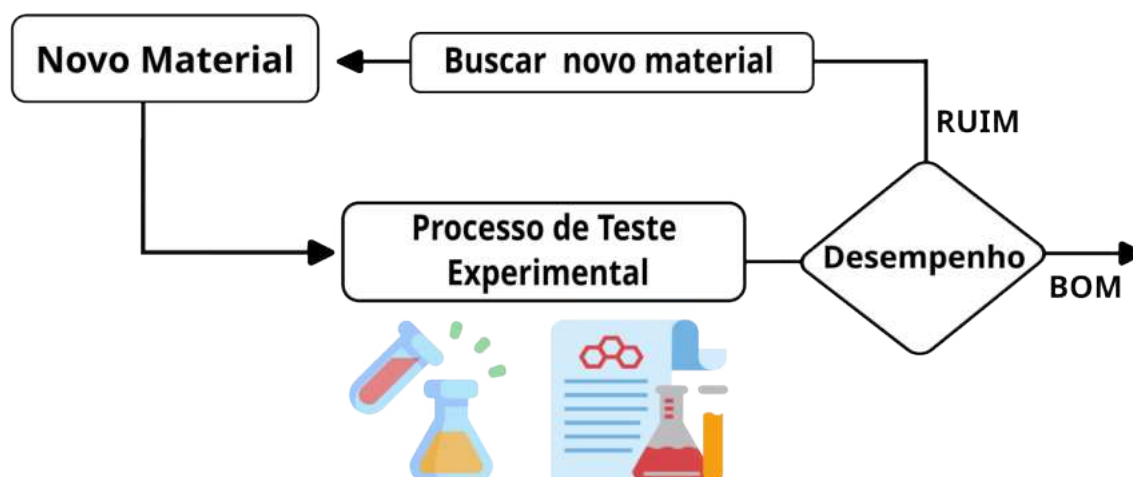
Embora eficaz para descobertas iniciais, essa abordagem enfrenta limitações significativas. O alto custo, a grande demanda por insumos e o tempo necessário para síntese e caracterização de materiais restringem sua escalabilidade.^[32] Esses desafios dificultam a exploração eficiente do vasto espaço químico, limitando o desenvolvimento de soluções tecnológicas mais avançadas.

1.4.2 Previsões Baseadas em Leis Físicas e Químicas

Para superar as limitações do paradigma tradicional, surgiu o segundo paradigma, que utiliza leis físicas e químicas para prever propriedades de materiais de forma mais fundamentada. Essa abordagem permitiu reduzir a dependência exclusiva de experimentos, fornecendo maior previsibilidade no processo de seleção de materiais.^[38]

No entanto, a aplicação prática desse paradigma enfrenta dificuldades relacionadas à

Figura 6 – Fluxo tradicional de descoberta de novos materiais.



Fonte: Próprio autor

escalabilidade. A complexidade das interações químicas e físicas muitas vezes exige modelos analíticos e numéricos altamente detalhados, que não conseguem abranger completamente o vasto espaço de possibilidades materiais. Assim, embora mais fundamentada, essa abordagem ainda é insuficiente para resolver problemas em larga escala de maneira eficiente.

1.4.3 Simulação Computacional

O terceiro paradigma trouxe o uso de simulações computacionais, que possibilitam a exploração virtual do espaço de materiais. Ferramentas como métodos de primeiros princípios, teoria do funcional da densidade (DFT), dinâmica molecular (MD) e simulações Monte Carlo Grand Canonical (GCMC) permitem investigar propriedades específicas com maior precisão e menos custos experimentais.^[38, 39] A abordagem *high-throughput* (HT), em particular, possibilita a análise de grandes bibliotecas de materiais em curto período.

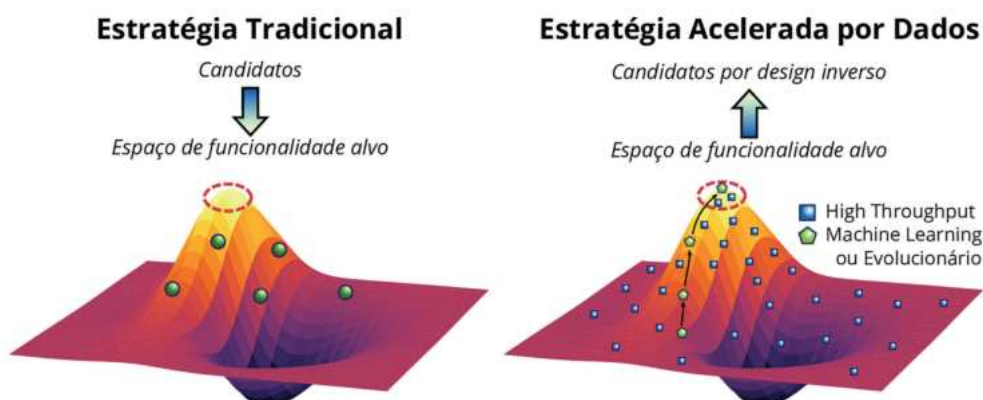
Apesar das inovações, desafios permanecem. As simulações GCMC, amplamente empregadas para prever adsorção de gases em materiais nanoporosos, podem exigir desde minutos até várias horas, dependendo da complexidade do sistema e das condições experimentais.^[40] Isso torna a triagem de milhares de materiais uma tarefa intensiva em recursos computacionais, demandando soluções mais rápidas e acessíveis.^[39]

1.4.4 Aprendizado de Máquina (*Machine Learning*)

O quarto paradigma integra *big data* e aprendizado de máquina (ML) à ciência dos materiais, revolucionando o processo de desenho, projeto, desenvolvimento e descoberta de

novos materiais. A [Figura 7](#) compara a abordagem tradicional com a baseada em dados. Com ML, é possível explorar amplos espaços químicos com maior eficiência, reduzindo custos e tempo de análise. Modelos treinados com dados de alta qualidade identificam padrões complexos e prevêem propriedades com alta precisão, otimizando a descoberta de materiais.^[41, 42, 43]

Figura 7 – Diferença entre as abordagens tradicional e baseadas em dados para o design e a descoberta de moléculas e materiais.



Fonte: Retirado de: [Schleder e Fazzio](#)^[44]

Apesar das vantagens, o sucesso do ML depende diretamente da qualidade e abrang ncia das bases de dados. Portanto, esfor os na curadoria de dados e no desenvolvimento de modelos robustos s o essenciais para consolidar seu papel na ci ncia dos materiais.^[45]

1.4.4.1 Aplica  o na descoberta de materiais para captura de CO₂

A literatura apresenta avan os significativos no uso de aprendizado de m quina (ML) para otimizar e projetar materiais para captura de gases. T cnicas como redes neurais artificiais (*Artificial Neural Networks*, ANN), m quinas de vetores de suporte (*Support Vector Machines*, SVM) e florestas aleat rias (*Random Forests*, RF) t m sido aplicadas com sucesso para prever capacidades de adsor  o,^[46] estabilidade estrutural^[47] e propriedades f sicas de MOFs.^[48]

Estudos recentes mostram que modelos de aprendizado de m quina podem prever capacidades de adsor  o com erros m dios absolutos inferiores a 0,05 mmol/g para CO₂, o que representa um avan o not vel na precis o das predi  es.^[49] Al m disso, essas t cnicas t m sido utilizadas para entender a estabilidade estrutural dos MOFs, um aspecto crucial para sua aplica  o pr tica. Modelos preditivos t m avaliado par metros como entalpia de adsor  o,^[50] resist ncia mec nica e estabilidade qu mica,^[47] ajudando a identificar materiais promissores que podem manter desempenho em condi  es adversas.

Apesar dessas contribui  es, desafios importantes permanecem. Um dos principais problemas   a escassez de dados experimentais de alta qualidade e abrang ncia

suficiente para treinar modelos que possam generalizar para além das condições específicas avaliadas.^[45] Outro obstáculo é a interpretabilidade de modelos complexos, como redes neurais profundas, que embora apresentem alta precisão, muitas vezes carecem de explicações claras sobre como as decisões são tomadas, dificultando sua aplicação em escala industrial.^[45] Além disso, a integração de predições de aprendizado de máquina com validação experimental em tempo real está em estágios iniciais de desenvolvimento, limitando sua aplicabilidade em condições práticas.

A lacuna que este trabalho busca superar está na falta de integração eficiente entre dados simulados e predições generalizáveis. Propomos combinar simulações computacionais de adsorção e ajustes de modelos de isothermas com algoritmos de aprendizado de máquina, a fim de identificar materiais com maior capacidade de captura de CO₂ em bancos de dados com milhares de estruturas. Essa abordagem pretende não apenas prever os parâmetros de adsorção com precisão, mas também validar o desempenho dos materiais em condições práticas, contribuindo para o avanço das tecnologias de captura e armazenamento de carbono de forma mais eficaz e confiável.

2 OBJETIVOS

Este trabalho tem como objetivo avaliar diferentes modelos de aprendizado de máquina e sua capacidade de prever isotermas de adsorção de CO_2 e N_2 para materiais reticulares.

2.1 Objetivos específicos

- Construir um banco de dados de adsorção de CO_2 e N_2 para materiais reticulares.
- Avaliar a adequação do modelo dual-site Langmuir (DSL) para o ajuste dos dados de adsorção simulados.
- Analisar a influência dos descritores químicos e geométricos no desempenho dos diferentes modelos avaliados.
- Identificar o modelo com maior capacidade preditiva para os parâmetros do modelo DSL.
- Testar a capacidade de generalização dos modelos em um banco de dados com novas estruturas.

3 METODOLOGIA

Um conjunto de materiais reticulares compostos por COFs e MOFs foram utilizados para simular a adsorção de CO_2 e N_2 sob diferentes pressões e temperaturas. As simulações, realizadas pelo método Monte Carlo no *Ensemble Grand Canônico*, forneceram dados de adsorção, que foram posteriormente ajustados ao modelo de isoterma de Langmuir de dois sítios (DSL, *dual-site Langmuir*). Os parâmetros resultantes desse ajuste foram aplicados em modelos de aprendizado de máquina com o objetivo de identificar os mais eficientes na predição das isotermas. Em seguida, o modelo mais preciso foi utilizado para avaliar novos materiais, selecionando aqueles com maior capacidade de captura de CO_2 .

3.1 Banco de Dados

Os materiais utilizados nesse trabalho foram retirados do banco de dados do CRAFTED^[51] e do QMOF,^[52] no qual os modelos das estruturas cristalinas foram obtidos experimentalmente por difração de raios-x. As informações cristalográficas são obtidas em arquivos CIF (do inglês *Crystallographic Information File*), um formato de arquivo padrão utilizado para descrever a estrutura cristalina de materiais, contendo informações sobre os parâmetros da célula unitária, coordenadas atômicas e fatores de ocupação, além de outros dados relevantes para a caracterização de estruturas cristalinas. Ao todo usamos 7375 materiais, com 1358 estruturas do CRAFTED e 6017 do QMOF.

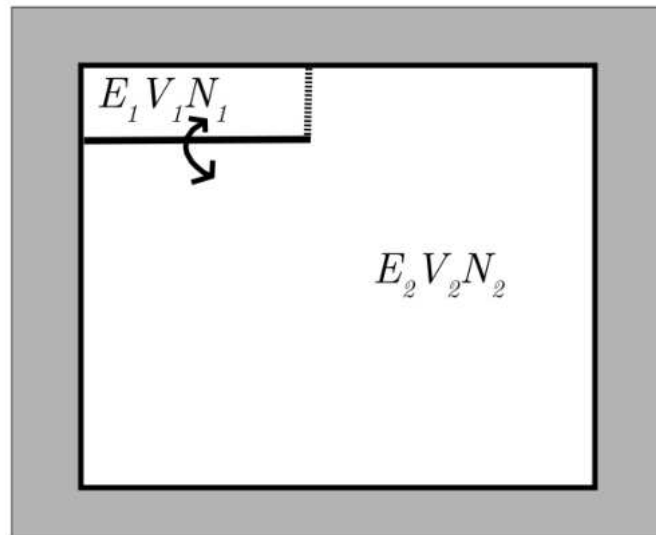
3.2 Simulações Monte Carlo no *Ensemble Grand Canônico* (GCMC)

As simulações de adsorção foram realizadas utilizando o método Monte Carlo no ensemble *grand-canônico* (GCMC). As posições dos átomos nas estruturas reticulares foram mantidas fixas e o efeito das vibrações térmicas nas estruturais foi negligenciado. Essa aproximação permite que o foco da simulação permaneça nas interações entre o adsorvato e o adsorvente, já que, especialmente em baixas temperaturas, as vibrações térmicas têm um impacto menor nas propriedades de adsorção. Além disso, manter as posições atômicas fixas reduz significativamente a complexidade dos cálculos. Outro fator limitante é a falta de campos de força aplicáveis à ampla variedade de ambientes químicos definidos pelas estruturas reticulares.^[53] De qualquer forma, essa suposição pode ser usada com confiabilidade na simulação de GCMC para a maioria dos materiais reticulares.^[54] É importante observar que pode haver algumas flexibilidades intrínsecas em certos materiais com estrutura quase rígida,^[55] que não foram consideradas neste trabalho.

3.2.1 Ensemble grand-canônico (μVT)

O *ensemble* do tipo *grand-canônico* descreve um conjunto de partículas em um sistema termodinâmico com volume fixo (V) em contato com um reservatório que mantém a temperatura (T) e o número total de partículas (N) constantes. Esse número constante de partículas faz com que o potencial químico μ não se altere. A Figura 8 ilustra o ensemble simulado. Consideramos o material nanoporoso no subsistema 1, com volume fixo, permitindo a troca de partículas e calor com o reservatório fictício (subsistema 2), que serve como fonte para as moléculas de gás. O sistema 2, sendo muito maior que o sistema 1, também atua como reservatório térmico para o subsistema 1.^[56]

Figura 8 – Um sistema isolado composto por dois subsistemas 1 e 2 que podem trocar calor e trocar partículas de forma que a energia total e o número de partículas permaneçam constantes.



Fonte: Retirado de Frenkel e Smit^[56]

3.2.2 Ergodicidade

A hipótese da ergodicidade descreve sistemas que possuem propriedades estatísticas invariantes no tempo. Em termos gerais, a ergodicidade postula que, ao longo de um período de tempo suficientemente longo, um sistema explorará todas as configurações acessíveis compatíveis com suas restrições físicas. Isso implica que o comportamento médio de um sistema ergódico, quando observado por um tempo suficientemente longo, refletirá suas propriedades estatísticas intrínsecas.^[56] Dessa forma, pode-se calcular uma determinada propriedade tanto pela média temporal quanto pela média de um conjunto de configurações acessíveis, uma vez que, para sistemas ergódicos, a média temporal é equivalente à média sobre o conjunto de configurações.^[57]

Matematicamente, para uma propriedade X , a hipótese ergódica pode ser expressa como:

$$\langle A \rangle = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau X(t) dt = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{i=1}^M X_i \quad (3.1)$$

Na equação (3.1), $\langle A \rangle$ representa o valor médio esperado da propriedade A . O termo $\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau X(t) dt$ refere-se à média temporal de $X(t)$, obtida integrando-se a propriedade ao longo de um intervalo de tempo τ que tende ao infinito. Por outro lado, o termo $\lim_{M \rightarrow \infty} \frac{1}{M} \sum_{i=1}^M X_i$ expressa a média sobre M diferentes estados independentes de X_i , ou seja, uma média de ensemble. Essa equivalência entre as duas formas de cálculo é um dos pilares da hipótese ergódica.

3.2.3 Algoritmo Monte Carlo Metropolis

O método Monte Carlo (MC) é uma técnica matemática que utiliza um conjunto de sequências numéricas aleatórias para obter uma estimativa quantitativa de uma propriedade.^[58] Garantido pela hipótese ergódica, podemos utilizar esse método para gerar o conjunto de configurações acessíveis na qual possamos calcular as propriedades de interesse em um sistema. Com o objetivo encontrar um conjunto de macro-estados de equilíbrio do sistema físico, torna-se necessário utilizar o algoritmo de Metrópolis, uma versão particular do MC. O algoritmo de Metrópolis é projetado para amostrar estados de uma distribuição de probabilidade desejada (a distribuição de Boltzmann, por exemplo) de forma mais eficiente. O algoritmo de Metrópolis corrige automaticamente o viés introduzido por diferentes probabilidades de aceitação de novos estados. Em contraste, o Monte Carlo simples pode ter dificuldade em corrigir esse viés, especialmente em sistemas com muitos estados possíveis ou quando a função de probabilidade de transição não é bem ajustada.

As configurações acessíveis de sistema físico, que também é conhecido como micro-estados, ocorrem de acordo com a distribuição de probabilidade de Boltzmann. Os micro-estados e as variáveis macroscópicas em equilíbrio podem então ser obtidas pela média de todos esses microestados ponderados pela distribuição de Boltzmann. Na prática, calcular essa média é uma tarefa quase impossível, tendo em vista a quantidade enorme de micro-estados do sistema. Com isso, a ideia do algoritmo de Metrópolis é encontrar o macro-estado de equilíbrio a partir de um micro-estado inicial qualquer, efetuando varias transições aleatórias até que permaneça constante a média das variáveis microscópicas sobre esses micro-estados percorridos.^[57]

O software RASPA^[59] realiza a cada passo Monte Carlo, uma sequência de 20 ou N (o que for maior) tentativas de movimento, sendo N o número de moléculas de adsorvato dentro da caixa de simulação. Em um passo, cada molécula experimenta em pelo menos uma tentativa de movimento, se é aceita ou rejeitada. Nos sistemas simulados via

GCMC, chamamos de microestado a configuração com um número específico de moléculas e suas posições. Esses microestados são gerados perturbando estocasticamente o estado do sistema, podendo adicionar moléculas ao sistema, removê-las ou alterar sua posição e orientação. Para garantir a distribuição de Boltzmann dos microestados, a probabilidade de aceitar um movimento é calculada de acordo com as equações abaixo:

Translação:

$$P_{ACC}(S \rightarrow S') = \min(1, \exp(-\beta\Delta U)) \quad (3.2)$$

Rotação

$$P_{ACC}(S \rightarrow S') = \min \left\{ 1, \exp \left(-\beta\Delta U \frac{\sin\theta_s}{\sin\theta_{s'}} \right) \right\} \quad (3.3)$$

Inserção:

$$P_{ACC}(N_a \rightarrow N_a + 1) = \min \left\{ 1, \frac{\beta f V}{N_a + 1} \exp(-\beta\Delta U) \right\} \quad (3.4)$$

Deleção:

$$P_{ACC}(N_a \rightarrow N_a - 1) = \min \left\{ 1, \frac{N_a}{\beta f V} \exp(-\beta\Delta U) \right\} \quad (3.5)$$

onde U representa a energia potencial de interação, N_a e V são o número de moléculas e o volume, respectivamente, β é a temperatura termodinâmica recíproca, $1/k_B T$, com k_B sendo a constante de Boltzmann, e θ é o ângulo de Euler da rotação de um corpo rígido. Aqui, f é a fugacidade da espécie adsorvente, que está relacionada ao potencial químico da seguinte forma:

$$f = \frac{q_{rot}}{\beta \Lambda^3} \exp(\beta\mu) \quad (3.6)$$

onde q_{rot} é a função de partição rotacional para uma única molécula rígida e Λ é o comprimento de onda térmico de Broglie. A adsorção é calculada como uma função da fugacidade em vez da pressão, com a equação de estado de Peng-Robinson (equação 3.7) sendo usada para converter pressão em fugacidade.

$$P = \frac{RT}{V - b} - \frac{a\alpha}{(V + (1 + \sqrt{2})b)(V + (1 - \sqrt{2})b)} \quad (3.7)$$

Nessa equação, os termos são definidos como:

- P : Pressão do sistema.
- R : Constante universal dos gases.

- T : Temperatura do sistema.
- V : Volume molar do sistema.
- a : Parâmetro que depende da interação entre as moléculas, dado por:

$$a = 0.45724 \frac{R^2 T_c^2}{P_c} \quad (3.8)$$

onde T_c é a temperatura crítica e P_c é a pressão crítica.

- b : Parâmetro que representa o volume ocupado pelas moléculas, dado por:

$$b = 0.07780 \frac{RT_c}{P_c} \quad (3.9)$$

- α : Fator de correção que depende da temperatura reduzida $Tr = \frac{T}{T_c}$, dado por:

$$\alpha = \left[1 + k \left(1 - \sqrt{Tr} \right) \right]^2 \quad (3.10)$$

onde k é uma constante ajustada empiricamente (geralmente $k = 0.37464 + 1.54226\omega - 0.26992\omega^2$, sendo ω o fator acêntrico).

Uma distinção importante que deve ser feita aqui é entre *quantidade adsorvida absoluta* e a *quantidade adsorvida em excesso*. A quantidade absoluta adsorvida é o número real de moléculas presentes nos microporos em uma fugacidade específica. A quantidade em excesso é a diferença entre a quantidade absoluta adsorvida e o número de moléculas que estaria presente no volume do microporo de acordo com a densidade do gás na pressão e temperatura de adsorção. As simulações de Monte Carlo relatam a quantidade absoluta adsorvida, enquanto as medidas experimentais são mais frequentemente apresentadas como a quantidade em excesso.^[60] Para recuperar o uptake absoluto a partir do uptake em excesso, é necessário levar em consideração a densidade do adsorvente e o volume do material. O uptake em excesso (n_{excesso}) é definido como a quantidade de gás adsorvido além da quantidade que ocuparia o mesmo volume na fase gasosa. Já o uptake absoluto (n_{absoluto}) é a quantidade total de gás adsorvido no material, incluindo o gás que está ocupando o volume poroso do adsorvente.

A relação entre o uptake absoluto e o uptake em excesso pode ser expressa pela seguinte equação:

$$n_{\text{absoluto}} = n_{\text{excesso}} + \rho_{\text{gás}} V_{\text{poroso}} \quad (3.11)$$

Onde:

- n_{absoluto} é o uptake absoluto (quantidade total de gás adsorvido).

- n_{excesso} é o uptake em excesso (quantidade de gás adsorvido além do volume poroso).
- $\rho_{\text{gás}}$ é a densidade do gás na fase gasosa.
- V_{poroso} é o volume poroso do material adsorvente.

O termo $\rho_{\text{gás}}V_{\text{poroso}}$ representa a quantidade de gás que estaria presente no volume poroso do material, se esse volume estivesse ocupado apenas pela fase gasosa.

A adsorção é a propriedade de interesse nas simulações e ela é dada como o número médio de moléculas presentes no sistema (material reticular). Para um único valor de potencial químico ou fugacidade, a simulação produzirá uma densidade média adsorvida e uma série de simulações com potenciais químicos crescentes e em uma temperatura constante irá produzir uma isoterma de adsorção.

3.2.3.1 Passos do Monte Carlo

A adsorção, sendo uma medida macroscópica, nas simulações GCMC (Grand Canonical Monte Carlo) é determinada por meio da média de uma quantidade microscópica correspondente. O valor médio da adsorção, denotado por $\langle U \rangle$, está associado a uma incerteza estatística $\sigma(U)$, que é definida pela raiz quadrada da variância $\sigma^2(U)$. A expressão da variância é dada por:

$$\sigma^2 = \frac{1}{M-1} \sum_{i=1}^M (x_i - \bar{x})^2 \quad (3.12)$$

Nesta equação, σ^2 representa a variância da amostra, que quantifica a dispersão dos valores amostrais x_i em torno da média $\bar{x}$. O fator de correção $\frac{1}{M-1}$ garante que a variância amostral seja uma estimativa não viesada da variância populacional, sendo M o número total de amostras. A soma dos quadrados das diferenças $\sum_{i=1}^M (x_i - \bar{x})^2$ reflete o grau de flutuação das amostras em relação à média.

Para grandes amostras, pode-se fazer a aproximação $M-1 \approx M$, o que simplifica a equação da variância para:

$$\sigma^2(U) = \frac{1}{M} \sum_{i=1}^M (U_i - \langle U \rangle)^2 \quad (3.13)$$

Conforme o número de amostras M aumenta, a incerteza estatística $\sigma(U)$ torna-se inversamente proporcional à raiz quadrada de M :

$$\sigma(U) \propto \frac{1}{\sqrt{M}} \quad (3.14)$$

No entanto, embora o aumento no número de amostras reduza a incerteza estatística, é crucial garantir que a amostragem seja representativa do espaço de fase. Coletar um grande número de amostras de uma região limitada do espaço de fase pode resultar em um erro sistemático significativo, mesmo que a incerteza estatística seja pequena. Nesses casos, a média calculada pode ser precisa em termos estatísticos, mas não refletirá de maneira fidedigna a propriedade macroscópica do sistema. Portanto, a qualidade da média calculada depende não apenas do número de amostras, mas também da adequação da exploração do espaço de fase. Como consequência, determinar *a priori* o número ideal de passos de simulação que equilibrem precisão e representatividade é uma tarefa complexa e geralmente empírica.

Em um estudo recente, Oliveira et al.^[61] introduziram um novo método automatizado, conhecido como Regra do Erro Padrão Marginal, para identificar o ponto de truncamento ideal nas simulações GCMC. Este procedimento garante achar um estado de equilíbrio para que as propriedades de interesse sejam extraídas com mínimo de erros sistemáticos, fornecendo resultados estatisticamente robustos e reproduzíveis.

No presente trabalho, as simulações foram conduzidas utilizando 10.000 passos de Monte Carlo para a adsorção de N₂ e 15.000 passos para CO₂. Para as simulações realizadas a 50 bar, foram adicionados mais 5.000 passos em ambos os casos. As tentativas de movimentos durante as simulações incluíram inserções e exclusões de moléculas do adsorvato, com uma probabilidade de 50% para cada tipo de troca, além de translações, rotações e reinserções, com probabilidades de 0,5, 0,3, 0,1 e 0,1, respectivamente.

As isotermas de adsorção foram geradas a partir de uma única sequência de simulações, evitando longos ciclos de inicialização. Cada ponto de pressão foi obtido sequencialmente a partir do estado final da simulação na pressão anterior, garantindo consistência entre as simulações. Os valores de captação para cada pressão foram extraídos da fase de equilíbrio da simulação GCMC, utilizando o programa pyMSER.^[61]

3.2.4 Campo de Força e Cargas Parciais

Os fatores que afetam o peso da probabilidade nos passos MC estão relacionados energia potencial (ΔU) do sistema antes e depois da tentativa de movimento. No caso dos movimentos de inserção e remoção, eles também dependem do potencial químico ou fugacidade da espécie adsorvente. A posição atômica do material adsorvente são fixas e a molécula do adsorvato é considerada rígida. Com isso, sobra apenas as interações intermoleculares para computar a energia potencial, que incluem as interações de van der Waals e eletrostáticas. Estas, por sua vez, são calculadas via uma combinação de potenciais mostrado na equação (3.15). As interação de van der Waals são calculadas usando o potencial de Lennard-Jones (LJ), enquanto que a interação eletrostática é calculada pelo potencial de Coulomb.

$$U_{ij}(r_{ij}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{1}{4\pi\epsilon} \frac{q_i q_j}{r_{ij}} \quad (3.15)$$

Nessa equação, i e j são diferentes pares de átomos que estão interagindo e r_{ij} é distância interatômica entre eles. O ϵ_{ij} e σ_{ij} são respectivamente a profundidade e o diâmetro do poço, e q_i e q_j são as cargas parciais dos átomos i e j . Os parâmetros LJ entre átomos de diferentes tipos são calculados usando as regras de mistura de Lorentz-Berthelot, mostrados na equação (3.16),

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}, \quad \sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \quad (3.16)$$

Os parâmetros do potencial de Lennard-Jones são determinados por meio de um campo de força clássico. Assim, o campo de força utilizado é o que governa o estado do sistema, controlando a energia de interação. Como consequência, o campo de força torna-se uma das principais informações de entrada necessárias para configurar a descrição física do sistema.^[11] Nas simulações utilizamos o campo de força universal (UFF, do inglês *Universal Force Field*),^[62] com os parâmetros disponíveis no [Apêndice A](#). Para as moléculas adsorvidas os parâmetros LJ foram retirados do campo de força TraPPE^[63] ([Apêndice A](#)). Um raio de corte de 12.8 Å foi usado para interações de curto alcance LJ e carga-carga, com a técnica de soma de Ewald com uma precisão relativa de 10^{-6} sendo utilizada para calcular as interações eletrostáticas de longo alcance. O potencial de Lennard-Jones foi deslocado para zero no raio de corte.

Em relação ao potencial de Coulomb, os parâmetros são definidos com base nas cargas atômicas parciais, que representam a carga atribuída a cada átomo no material. Existem diversas abordagens para calcular essas cargas, que vão desde métodos de equilíbrio de carga até derivações com a teoria do funcional da densidade (DFT). Como esse tipo de interação influencia a aceitação ou rejeição de movimentos em simulações de Monte Carlo, a escolha do esquema para as cargas parciais é crucial para a simulação GCMC, especialmente quando o adsorvato possui um momento de dipolo^[64]. Neste trabalho, foi adotado o esquema de cargas parciais tipo DDEC^[65], obtido a partir dos bancos de dados de estruturas QMOF e CRAFTED.

3.3 Ajuste de Isotermas de adsorção

Existem uma ampla variedade de modelos de isotermas, como Langmuir, Freundlich, Dubinin-Radushkevich, Temkin, Toth e várias outras. Esses modelos são derivados de abordagens cinéticas, termodinâmicas ou na simples geração de uma curva característica do fenômeno de adsorção de interesse.^[66]

3.3.1 Modelo de Langmuir

A isoterma de Langmuir foi apresentada por Irving Langmuir e é um dos modelos teóricos mais simples de adsorção.^[67] Ela é derivada originalmente de uma visão puramente cinética, com o processo da adsorção de gás em uma interface sólida sendo semelhante ao mecanismo da condensação em fase gasosa em uma superfície líquida. Ou seja, à medida que as moléculas de gás atingem a superfície, elas são mantidas por forças intermoleculares atrativas (semelhantes à condensação) por um tempo curto, até saírem novamente para a fase gasosa, semelhante à evaporação.

Segunda a teoria cinética dos gases, é assumindo que a pressão é proporcional a taxa de incidência das moléculas do adsorvato em uma área unitária da superfície do adsorvente. Com isso, a taxa de adsorção depende da pressão junto com a probabilidade de incidência em um sítio de adsorção vazio.

Ela é baseada nas seguintes hipóteses:

- A superfície do sólido é coberta por um número constante de sítios de adsorção, sendo que cada sítio só pode ser ocupado por uma única molécula adsorvida.
- Todos os sítios são equivalentes e a adsorção se completa quando todos os sítios forem ocupados, correspondendo a uma monocamada de adsorvato.
- Considera-se que as moléculas adsorvidas não interagem umas com as outras nem saltam de um sítio para outro.

A isoterma de Langmuir pode ser deduzida admitindo um processo de equilíbrio dinâmico (cinético). Para um tipo de molécula e um tipo de sítio, temos:



onde A representa uma molécula não adsorvida, S representa um sítio de adsorção livre e AS uma molécula adsorvida ou um sítio ocupado.

A reação direta representa a adsorção e admitindo-se um processo de primeira ordem para ambos componentes da “reação”, a molécula não adsorvida e o sítio livre. Temos então:

$$v_{ads} = K_a[A][S] \quad (3.18)$$

onde o K_a é a constante da velocidade de adsorção.

A reação inversa representa a dessorção e admitindo-se um processo de 1ª ordem para o “produto”, que é a molécula adsorvida, temos:

$$v_{des} = K_d[AS] \quad (3.19)$$

A concentração das moléculas não adsorvidas, $[A]$, é proporcional a pressão P no caso da adsorção de um gás. Definimos agora θ como a fração da superfície recoberta pelo adsorvato ou grau de cobertura da superfície,

$$[AS] \propto \theta \quad e \quad [S] \propto 1 - \theta \quad (3.20)$$

A partir dessas proporcionalidades podemos reescrever as velocidades de adsorção e dessorção (considerando a adsorção de um gás) como sendo

$$v_{ads} = K_a(1 - \theta)P \quad (3.21)$$

e

$$v_{des} = K_d\theta \quad (3.22)$$

Isso diz que a fração da superfície ocupada variando como o tempo será:

$$\frac{d\theta}{dt} = v_{ads} - v_{des} = K_a(1 - \theta)P - K_d\theta \quad (3.23)$$

Admitindo o equilíbrio dinâmico entre os processos de adsorção e de dessorção, a fração da superfície recoberta deixa de variar no tempo. Com isso, podemos dizer que as taxas de adsorção e dessorção são iguais ($d\theta/dt = 0$), sendo então $v_{ads} = v_{des}$. Temos então que:

$$K_a(1 - \theta)P = K_d\theta \quad (3.24)$$

Sabendo que a razão entre a velocidade específica direta (K_a) e a velocidade específica inversa (K_d) define a constante de equilíbrio para o processo

$$\frac{K_a}{K_d} = K \quad (3.25)$$

Logo

$$K = \frac{\theta}{(1 - \theta)P} \quad (3.26)$$

ou seja

$$\theta = \frac{KP}{1 + KP} \quad (3.27)$$

Sendo equação 3.27 a forma da isoterma de Langmuir em termos da fração de superfície ocupada θ .

De forma geral podemos dizer que quanto maior o valor de K , mais intensa a adsorção, ou seja, mais forte as ligações entre o adsorvato e o adsorvente. Isto nos leva a concluir que o valor de K está associado com a liberação de calor no processo de adsorção, ou seja, K está relacionado com a entalpia de adsorção.¹ Com isso, quanto maior a entalpia de adsorção, maior o calor liberado no processo de adsorção e maior o valor de K .

Uma outra forma de representar a isoterma de Langmuir é através da extensão de adsorção. Como

$$\frac{x}{m} = \theta \quad (3.28)$$

a equação correspondente à isoterma de Langmuir em termos da extensão de adsorção é dada por

$$\frac{x}{m} = \frac{KbP}{1 + KP} \quad (3.29)$$

onde b é uma constante de proporcionalidade.

3.3.1.1 Modelo de Langmuir de dois sítios

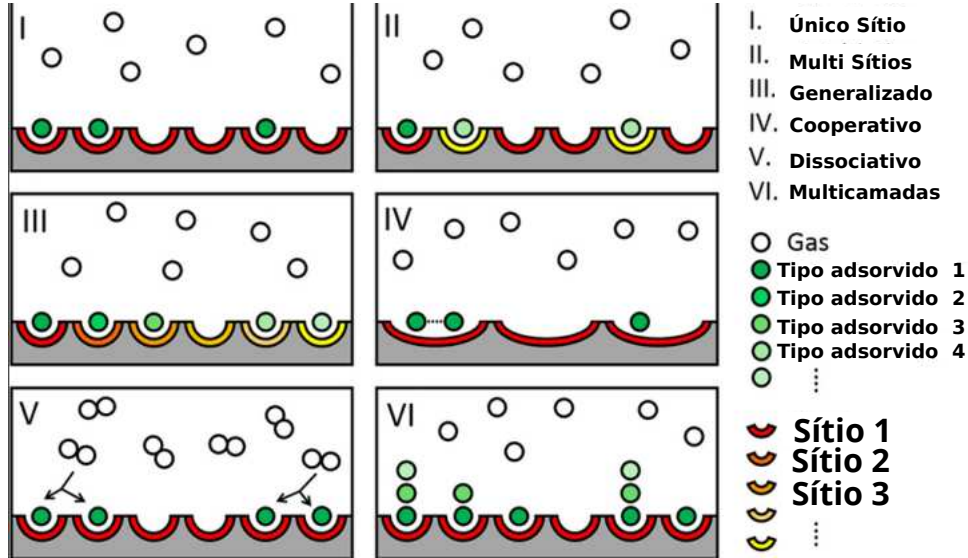
Langmuir reconheceu que a diversidade da química da superfície e da geometria estrutural exibida pelos materiais sólidos, em conjunto com a diversidade em tamanho e caráter das espécies de adsorbato, se manifesta em mecanismos físicos de adsorção amplamente diferentes.^[68] Ele identificou e classificou seis mecanismos simples e distintos de adsorção em Langmuir 1918 e então derivou a equação que rege a quantidade adsorvida, em função da pressão (P) a temperatura constante (T) para cada caso. Essas classificações são descritas na Figura 9.

A forma mais simples de adsorção por múltiplos sítios é quando a superfície oferece dois tipos distintos de sítios de adsorção, que é conhecido como o modelo de Langmuir de dois sítios (DSL, do inglês *Dual-Site Langmuir*).

Esse modelo é dado por:

¹ $\Delta G = \Delta H - T\Delta S = -RT\ln K$

Figura 9 – Seis classificações de propostas por Langmuir em 1918: adsorção de único sítio ou "simples"(SSL), adsorção de múltiplos sítios (MSL) (da qual um subcaso é adsorção de sítios duplos (DSL)), adsorção de múltiplos sítios generalizados (GL), adsorção cooperativa (CA) (da qual um subcaso é adsorção quadrática (QA)), adsorção dissociativa (DA) e adsorção em multicamadas (MLA).



Restirado de: Swenson e Stadie^[68]

$$q_i^* = \left[q_{1,i}^s \frac{b_{1,i}P}{1+b_{1,i}P} \right] + \left[q_{2,i}^s \frac{b_{2,i}P}{1+b_{2,i}P} \right] \quad (3.30)$$

ou

$$q_i^* = \sum_{j=1}^2 \left[q_{j,i}^s \frac{b_{j,i}P}{1+b_{j,i}P} \right]$$

onde $q_{j,i}$ é a capacidade de saturação do sítio j com respeito à espécie i , $b_{j,i}$ é a afinidade do sítio j com respeito à espécie i dado pela equação de van't Hoff:

$$b_{i,j} = e^{\frac{-\Delta H_{i,j}}{RT}} \cdot e^{\frac{\Delta S_{i,j}}{R}} \quad (3.31)$$

onde $\Delta H_{i,j}$ é a entalpia de adsorção do sítio j para a espécie i e $S_{j,i}$ é a entropia de adsorção do sítio j para a espécie i .

O modelo Dual-Site Langmuir (DSL) é amplamente utilizado para prever com precisão o comportamento de equilíbrio em sistemas de adsorção binária, como misturas de CO_2/N_2 , ao longo de uma ampla faixa de composições.^[69] A adsorção da espécie A em uma mistura binária de A e B , a uma temperatura fixa, pode ser descrita como uma extensão do modelo DSL,^[70] dado pela seguinte expressão:

$$q_A^* = \left[q_{1,A}^s \frac{b_{1,A}Py_A}{1 + (b_{1,A}Py_A) + (b_{1,B}Py_B)} \right] + \left[q_{2,A}^s \frac{b_{2,A}Py_A}{1 + (b_{2,A}Py_A) + (b_{2,B}Py_B)} \right], \quad (3.32)$$

onde y_A e y_B representam as frações molares dos componentes A e B na fase gasosa, respectivamente. Esse modelo permite a reconstrução precisa da isoterma de adsorção de misturas binárias para qualquer composição desejada, fornecendo um ajuste adequado para uma ampla gama de materiais adsorventes. A flexibilidade do modelo DSL em capturar o comportamento de adsorção de diferentes materiais faz dele uma escolha ideal para ajustes de valores de adsorção simulados, oferecendo tanto precisão quanto versatilidade na modelagem de sistemas multicomponentes.

3.3.2 Ajuste dos modelos de adsorção

Para obter um modelo fisicamente consistente o algorítmico de ajuste foi sujeito à restrições para os valores ajustados. Além disso, para garantir a consistência termodinâmica o $q_{1,N_2} = q_{1,CO_2} + q_{2,CO_2}$. Por fim, como não é esperado que a molécula de N_2 apresente sítio de adsorção específica nesses materiais sua isoterma de adsorção será considerada como sendo uma *single-site* Langmuir, *i.e.*, $q_{2,N_2} = 0$. Dessa forma, após o ajuste cada material terá 8 parâmetros (q_{1,CO_2} , $\Delta H_{1,CO_2}$, $\Delta S_{1,CO_2}$, q_{2,CO_2} , $\Delta H_{2,CO_2}$, $\Delta S_{2,CO_2}$, $\Delta H_{1,N_2}$, $\Delta S_{1,N_2}$) que capazes de reconstruir a isoterma de adsorção em qualquer temperatura e pressão desejada.

3.3.3 Modulo `curve_fit`

O módulo `curve_fit`, parte da biblioteca `scipy.optimize`^[71] em Python,^[72] foi utilizado para realizar o ajuste de curvas aos dados simulados. O método `curve_fit` baseia-se na minimização da soma dos quadrados das diferenças entre os valores observados e os valores previstos por uma função modelo específica. A função modelo, definida previamente de acordo com a natureza do fenômeno estudado, foi ajustada aos dados com o objetivo de determinar os parâmetros que melhor descrevem o comportamento observado. O `curve_fit` não apenas retorna os valores ajustados dos parâmetros, mas também fornece a matriz de covariância, permitindo a avaliação da incerteza associada aos ajustes.

Uma das principais vantagens do `curve_fit` é sua flexibilidade. Ele permite definir a função a ser ajustada de acordo com as características específicas do problema em questão. Seja uma função linear, exponencial, logarítmica, ou até mesmo uma função personalizada e não-linear, o `curve_fit` pode ser configurado para ajustar a curva conforme necessário. Além disso, o módulo oferece a possibilidade de incluir incertezas nas medições, ponderando os dados de acordo com sua variabilidade, o que resulta em ajustes mais robustos e confiáveis.

No presente estudo, optou-se por utilizar o parâmetro `absolute_sigma=True` no ajuste dos dados experimentais utilizando a função `curve_fit` da biblioteca `scipy.optimize`.

Essa escolha se justifica pelo fato de que as incertezas associadas aos valores medidos foram determinadas a partir do desvio padrão da média de múltiplas medições independentes.

3.3.4 Métricas de Avaliação para o Ajuste

Para avaliar a qualidade do ajuste do modelo DSL aos dados simulados, utilizou-se diversas métricas que quantificam a eficácia com que o modelo captura as características dos dados. As escolhidas foram: Coeficiente de Determinação (R^2), o Erro Quadrático Médio (MSE, do inglês *Mean Squared Error*), e a Porcentagem de Variância Explicada (EVP, do inglês *Explained Variance Percentage*).

3.3.4.1 Coeficiente de Determinação (R^2)

O coeficiente de determinação, denotado por R^2 , é uma métrica que indica a proporção da variância dos dados que é explicada pelo modelo. Ele é definido como:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.33)$$

onde y_i são os valores observados, $\hat{y}_i$ são os valores preditos pelo modelo, e $\bar{y}$ é a média dos valores observados.

O valor de R^2 varia entre 0 e 1, onde valores mais próximos de 1 indicam que o modelo explica uma maior proporção da variância dos dados. Em casos onde o valor de R^2 é negativo, isso indica que o modelo é pior do que um modelo trivial que simplesmente prediz a média dos dados.

3.3.4.2 Erro Quadrático Médio (MSE)

O Erro Quadrático Médio (MSE, do inglês *Mean Squared Error*) é uma métrica que quantifica a média dos quadrados dos erros entre os valores preditos e os valores observados. É calculado pela seguinte expressão:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.34)$$

onde n é o número de observações, y_i são os valores reais e $\hat{y}_i$ são os valores preditos pelo modelo.

O MSE sempre assume valores não-negativos, onde valores mais próximos de 0 indicam um modelo com melhor desempenho. Ele é sensível a outliers, já que erros maiores são amplificados devido ao quadrado das diferenças.

3.3.4.3 Porcentagem de Variância Explicada (EVP)

A Porcentagem de Variância Explicada (EVP, do inglês *Explained Variance Percentage*) é uma métrica que indica a proporção da variância dos valores alvo que é capturada pelo modelo, expressa em termos percentuais. É calculada da seguinte forma:

$$\text{EVP} = \left(1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)} \right) \times 100 \quad (3.35)$$

onde $\text{Var}(y)$ é a variância dos valores reais e $\text{Var}(y - \hat{y})$ é a variância dos resíduos (diferença entre os valores reais e preditos).

O valor de EVP varia de 0% a 100%, onde valores mais próximos de 100% indicam que o modelo explica uma maior proporção da variância nos dados. Um valor de EVP igual a 0% indica que o modelo não consegue capturar nenhuma variância.

3.3.5 Métricas de Desempenho

Na busca por métricas de triagem eficientes para selecionar os melhores candidatos à adsorção de CO₂, diferentes abordagens têm sido desenvolvidas com o objetivo de integrar características do material, otimização de processos e considerações econômicas. Inicialmente, a seletividade e a capacidade de trabalho foram amplamente adotadas como indicadores de desempenho. Outras métricas, como energia regenerativa, estabilidade térmica, e custo de síntese, também têm sido consideradas para fornecer uma visão mais abrangente da viabilidade do material para aplicações reais.

3.3.5.1 Capacidade de Trabalho

A capacidade de trabalho, também conhecida como "working capacity", é uma métrica essencial para avaliar o desempenho de materiais em processos cíclicos de adsorção e dessorção, como PSA (Pressure Swing Adsorption) e TSA (Temperature Swing Adsorption). Ela é definida como a diferença entre a quantidade de adsorvato armazenada sob condições de alta pressão (ou baixa temperatura) e a quantidade restante sob condições de baixa pressão (ou alta temperatura).

Matematicamente, pode ser expressa como:

$$W = q(P_{\text{ads}}) - q(P_{\text{des}}),$$

onde $q(P_{\text{ads}})$ é a quantidade adsorvida em condições de carga máxima (alta pressão) e $q(P_{\text{des}})$ é a quantidade remanescente após a regeneração (baixa pressão).

Essa métrica reflete diretamente o volume de gás que pode ser efetivamente capturado e liberado durante cada ciclo de operação. Para otimização prática, materiais com

alta capacidade de trabalho são preferidos, pois maximizam a eficiência operacional e minimizam os custos associados ao ciclo de adsorção/dessorção.

3.3.5.2 Seletividade

Em aplicações de separação, a adsorção é um processo competitivo entre duas ou mais espécies adsorvantes. Naturalmente, para caracterizar essa competição, precisamos de uma métrica que possa comparar o comportamento do material com relação às espécies concorrentes. Por exemplo, na definição mais geral, a seletividade é a razão das cargas molares de dois componentes concorrentes na fase adsorvida (q_i^{ads}) em equilíbrio com uma mistura de fase fluida a granel com relação às suas fugacidades parciais.

Matematicamente, a seletividade S_{ij} pode ser definida como:

$$S_{ij} = \frac{q_i^{\text{ads}}/q_j^{\text{ads}}}{f_i/f_j},$$

onde f_i e f_j são as fugacidades parciais das espécies i e j , respectivamente. Essa definição permite avaliar o quão preferencialmente uma espécie é adsorvida em relação a outra.

Em baixa pressão, a seletividade pode ser expressa simplesmente como a razão das constantes de Henry:

$$S_{ij} = \frac{K_H^i}{K_H^j}.$$

Essa forma simplificada é particularmente útil para prever o desempenho de materiais em sistemas de baixa concentração, como a captura de CO₂ em correntes de exaustão ou atmosferas diluídas.

3.4 Aprendizado de Máquina

O aprendizado de máquina é uma área da inteligência artificial que desenvolve algoritmos capazes de identificar padrões e fazer previsões a partir de dados. Ele pode ser classificado em duas categorias principais: supervisionado e não supervisionado.^[73]

O aprendizado supervisionado envolve o treinamento de um modelo com dados rotulados. Isso significa que cada exemplo no conjunto de dados está associado a uma saída ou resposta esperada. O objetivo é que o modelo aprenda a mapear os dados de entrada para os rótulos de saída correspondentes, permitindo que ele faça previsões sobre novos dados nunca vistos antes. Aplicações comuns incluem classificação de imagens, reconhecimento de fala e detecção de spam.^[74]

O aprendizado não supervisionado envolve o treinamento de um modelo com dados não rotulados. O modelo é encarregado de descobrir padrões, relacionamentos ou agrupamentos nos dados sem qualquer conhecimento prévio da saída esperada. Aplicações

comuns incluem clusterização (por exemplo, agrupar clientes semelhantes), redução de dimensionalidade (por exemplo, reduzir o número de características em um conjunto de dados) e detecção de anomalias.^[75]

Neste trabalho, empregamos exclusivamente modelos de aprendizado supervisionado, uma vez que nossos dados contam com rótulos bem definidos. Dessa maneira, treinamos algoritmos capazes de estabelecer relações entre descritores químicos e/ou geométricos e os parâmetros de um modelo de isoterma.

Os códigos de aprendizado de máquina foram desenvolvidos utilizando a biblioteca Scikit-Learn, que forneceu ferramentas para a implementação e otimização de modelos como árvores de decisão, florestas aleatórias e XGBoost. Para as redes neurais, utilizamos o PyTorch, permitindo maior flexibilidade na construção e treinamento dos modelos, além da possibilidade de realizar ajustes finos nos hiperparâmetros e arquiteturas.

3.4.1 Decision Tree

A árvore de decisão ou *Decision Tree* é uma ferramenta no campo do aprendizado de máquina que utiliza uma estrutura hierárquica em forma de árvore, permitindo representar decisões e suas possíveis consequências de maneira clara e organizada.^[76]

A simplicidade e a facilidade de interpretação das árvores de decisão são características que as tornam extremamente valiosas. Diferente de outros modelos de aprendizado de máquina, que podem ser considerados “caixas pretas”, as árvores de decisão permitem que os usuários compreendam exatamente como uma decisão foi tomada. Cada nó interno representa um teste em um atributo, cada ramo representa o resultado do teste e cada nó folha representa uma decisão final ou um rótulo de classe. Além disso, as árvores de decisão são versáteis, podendo lidar tanto com dados numéricos quanto categóricos.^[76] Elas requerem pouca preparação de dados, o que as torna práticas e eficientes em termos de tempo.

É importante reconhecer as limitações das árvores de decisão. Uma das principais desvantagens é a tendência ao overfitting, onde a árvore se torna excessivamente complexa e ajustada aos dados de treinamento, perdendo a capacidade de generalizar para novos dados.^[77] Outra limitação é a instabilidade das árvores de decisão. Pequenas variações nos dados podem resultar em árvores completamente diferentes, o que pode ser problemático em aplicações onde a consistência é crucial.

3.4.2 Random Forest

O Random Forest é um algoritmo de aprendizado de máquina amplamente utilizado para tarefas de classificação e regressão. Ele opera através da construção de uma coleção de árvores de decisão independentes, onde cada árvore é gerada a partir de uma amostra

aleatória com reposição dos dados originais (método bootstrap) e uma seleção aleatória de características. A previsão final do modelo é obtida pela média das previsões individuais das árvores (no caso de regressão) ou pelo voto da maioria (no caso de classificação). Essa abordagem de agregação diminui a variância do modelo, reduzindo o risco de overfitting, uma vez que as árvores individuais tendem a ser altamente correlacionadas e apresentam alta variância. Ao combinar as previsões de múltiplas árvores, o Random Forest melhora a precisão e a robustez do modelo.^[78]

Outro aspecto importante do Random Forest é sua capacidade de lidar com grandes conjuntos de dados e de fornecer uma estimativa da importância das variáveis. Durante a construção das árvores, o algoritmo avalia o impacto de cada característica na redução da impureza dos nós, permitindo a obtenção de uma medida de importância das variáveis para a previsão final. Essa característica torna o Random Forest uma ferramenta valiosa para a seleção de características em cenários com um grande número de variáveis.^[79]

3.4.3 XGBoost

O XGBoost (Extreme Gradient Boosting) é um algoritmo de aprendizado de máquina que se destaca pela sua eficiência e precisão em tarefas de classificação e regressão. Baseado no princípio do gradient boosting, o XGBoost constrói modelos aditivos em forma de árvores de decisão, onde cada nova árvore corrige os erros residuais das árvores anteriores ^[80]. O processo de aprendizado é iterativo, ajustando gradualmente os pesos das observações para minimizar uma função de perda, como o erro quadrático médio ou a entropia cruzada.

O XGBoost utiliza uma regularização explícita para penalizar a complexidade do modelo, incorporando tanto a regularização L_1 (também conhecida como regularização de Lasso), que adiciona um termo de penalização baseado na soma dos valores absolutos dos coeficientes dos parâmetros, quanto a regularização L_2 (ou regularização de Ridge), que penaliza a soma dos quadrados dos coeficientes. Essa abordagem reduz a variância do modelo e contribui para a prevenção de *overfitting*, melhorando a capacidade de generalização do modelo, o que torna o XGBoost altamente eficaz em competições de aprendizado de máquina.

Um dos principais diferenciais do XGBoost em relação a outras implementações de *boosting* é sua capacidade de paralelização e escalabilidade, o que o torna ideal para grandes volumes de dados. O algoritmo incorpora técnicas avançadas, como a manipulação de dados esparsos, o cálculo eficiente de estimativas e a pré-*pruning* das árvores, que permitem economizar recursos computacionais sem sacrificar o desempenho. Além disso, o XGBoost oferece uma vasta gama de hiperparâmetros que podem ser ajustados para otimizar o desempenho do modelo, incluindo a taxa de aprendizado, o número de árvores, a profundidade máxima das árvores e a fração de amostras utilizadas em cada iteração.

3.4.4 Redes neurais

Redes Neurais são modelos computacionais inspirados no funcionamento do cérebro humano, amplamente utilizadas em tarefas de aprendizado supervisionado e não supervisionado. Uma rede neural é composta por camadas de neurônios artificiais, onde cada neurônio executa uma função de ativação não linear que transforma a entrada recebida. As camadas são organizadas em uma estrutura hierárquica, com uma camada de entrada que recebe os dados brutos, uma ou mais camadas ocultas que processam as informações intermediárias e uma camada de saída que gera as previsões. Cada conexão entre neurônios é associada a um peso que é ajustado durante o treinamento da rede, utilizando algoritmos de otimização, como o gradiente descendente. O processo de ajuste desses pesos visa minimizar uma função de perda, permitindo que a rede aprenda padrões complexos a partir dos dados.

Um aspecto central das redes neurais é a sua capacidade de modelar relações não lineares em grandes conjuntos de dados, o que as torna extremamente eficazes em problemas como reconhecimento de imagens, processamento de linguagem natural e previsão de séries temporais. A introdução de técnicas como a regularização L_2 , dropout e normalização por lotes (batch normalization) ajuda a mitigar problemas de overfitting, aumentando a capacidade de generalização do modelo. Além disso, redes neurais profundas, conhecidas como *deep learning*, que contêm múltiplas camadas ocultas, são particularmente poderosas para extrair representações hierárquicas dos dados, permitindo uma análise mais aprofundada e precisa. Com o avanço das capacidades computacionais e o aumento da disponibilidade de dados, redes neurais têm se tornado uma ferramenta essencial no campo do aprendizado de máquina e da inteligência artificial.

3.4.5 Descritores

Descritores são características ou atributos que representam os dados em um formato que pode ser utilizado por algoritmos de aprendizado de máquina. Eles desempenham um papel crucial na construção de modelos preditivos, pois determinam como os dados serão interpretados e processados. Em aprendizado de máquina, os descritores são usados para converter dados brutos em valores numéricos que permitem a análise e modelagem matemática. Por exemplo, em problemas de processamento de imagens, descritores como histograma de gradientes orientados (HOG) ou características de textura podem ser extraídos para representar a informação visual de maneira compacta e informativa. Da mesma forma, em problemas de processamento de texto, técnicas como TF-IDF (*Term Frequency-Inverse Document Frequency*) ou *embeddings* de palavras são usadas para converter texto em vetores numéricos que capturam a semântica e o contexto das palavras.

A seleção e a engenharia de descritores têm um impacto significativo no desempenho dos modelos de aprendizado de máquina. A escolha de descritores relevantes pode melhorar

a precisão e a eficiência do modelo, enquanto a inclusão de descritores irrelevantes ou redundantes pode levar a problemas como overfitting e aumento do tempo de treinamento.

3.4.5.1 Descritores Geométricos de materiais nanoporosos

O espaço químico dos materiais reticulares pode ser caracterizado pela combinação de suas propriedades geométricas, resultantes de sua natureza porosa e cristalina, e pela química de sua estrutura. Esta química inclui os nós metálicos (no caso dos MOFs), os ligantes ou blocos de construção (no caso dos COFs) e os grupos funcionais.

Para descrever a geometria dos poros, foram utilizados descritores geométricos como o tamanho do poro o volume de vazio, entre outros. Esses descritores foram obtidos utilizando o pacote de software Zeo++,^[81] especializado na análise de materiais porosos cristalinos. O Zeo++ permite a execução de análises baseadas na geometria da estrutura e na topologia do espaço vazio dentro do material, calculando parâmetros geométricos que descrevem os poros. A ferramenta utiliza a decomposição de Voronoi, que, para um dado arranjo de átomos em um domínio periódico, fornece uma representação gráfica do espaço vazio. Os detalhes sobre os descritores geométricos estão presentes na [seção A.1](#).

3.4.5.2 Descritores Químicos de materiais nanoporosos

Os descritores de autocorrelação revisados (RACs, do inglês *Revised Autocorrelation Descriptors*) foram originalmente desenvolvidos para compostos de coordenação. Eles consistem em uma série de funções de autocorrelação^[82] que codificam as correlações entre propriedades atômicas heurísticas, como tamanho, conectividade e eletronegatividade, em um grafo molecular. Posteriormente, os RACs foram adaptados para descrever a química dos materiais reticulares, permitindo correlacionar as propriedades atômicas dos átomos em um grafo cristalino. Para os materiais reticulares, os RACs foram expandidos para incluir descritores que capturam todos os aspectos relevantes do material, como a química dos metais, a química dos ligantes e os grupos funcionais, proporcionando uma representação mais completa e detalhada das características estruturais e químicas desses materiais.^[83]

A implementação dos RACs foi realizada utilizando o conjunto de ferramentas de código aberto molSimplify.^[84] Para a aplicação em COFs, o código foi adaptado para atender às características específicas desses materiais.

Os RACs incorporam tanto produtos quanto diferenças entre os nós do grafo molecular. Essa abordagem visa capturar a centralidade do metal e o seu ambiente de coordenação na determinação das propriedades de complexos de metais de transição. Para esse fim, a convenção dos RACs introduz a ideia de seleção de escopo,^[85] permitindo a escolha de quais átomos considerar na soma, como ligantes equatoriais ou axiais.

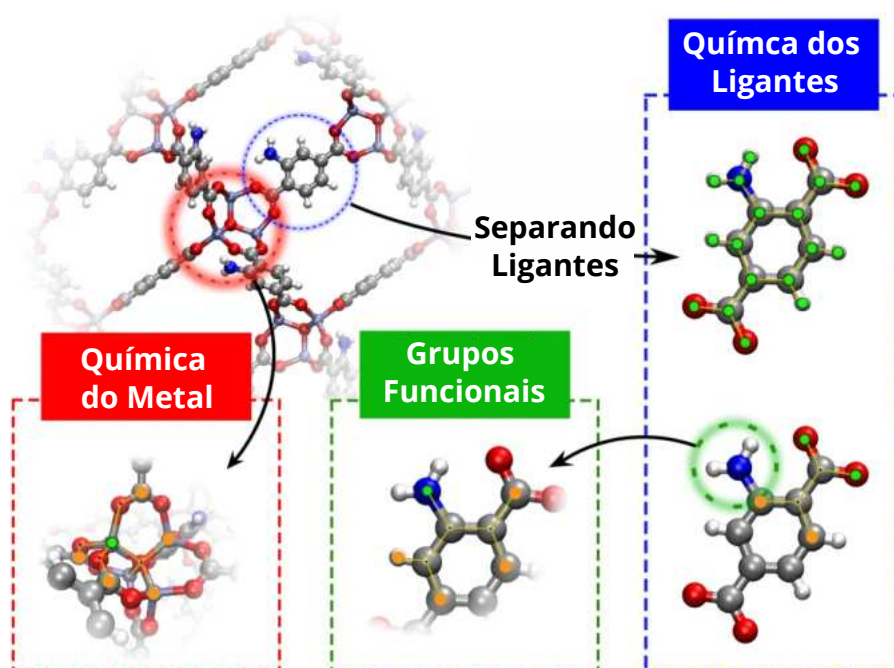
Além disso, os RACs são classificados em duas categorias: RACs centrados no

metal (mc) e RACs centrados no ligante (lc). Nos RACs mc, um dos átomos na expressão do AC é sempre o centro metálico, enquanto nos RACs lc, o átomo de referência é um ligante coordenador (Figura 10). Uma característica fundamental dos RACs é a inclusão de diferenças de propriedades em vez de produtos, o que só é não trivial para um valor mínimo de $d = 1$ e se aplica exclusivamente aos RACs lc ou mc:

$$\text{RAC}(d) = \sum_{\text{ligantes}} (\text{propriedade}_{\text{ligante}} - \text{propriedade}_{\text{metal}}) \quad (3.36)$$

Nessa equação, o escopo é definido como todos os ligantes, ou a média sobre um subconjunto específico, como ligantes equatoriais ou axiais. No total, existem seis tipos padrão de ACs e RACs (ou seja, média total sobre todos os átomos, axial ou equatorial; lc média axial ou equatorial; e mc média sobre todos). Para um determinado valor de corte d e cinco propriedades heurísticas,² há $42d + 10$ RACs.

Figura 10 – Representação dos grupos de descritores *RACs*. Os *RACs* do centro metálico são calculados no gráfico de cristal, enquanto os do ligante e do grupo funcional são calculados no gráfico molecular do ligante correspondente. A química do ligante inclui dois tipos de *RACs*: aqueles baseados no ligante completo e aqueles definidos pelos átomos de ligação do ligante. Os gráficos ilustram o átomo inicial (em verde) e o átomo próximo (em laranja) utilizados na definição dos descritores *RACs*.



Fonte: Adaptado de Moosavi et al.^[86]

² Propriedades heurísticas são características ou critérios que não derivam de princípios fundamentais rigorosos, mas sim de regras empíricas, intuições ou aproximações baseadas em observações experimentais e experiência prática. Por exemplo, na química, a eletronegatividade de Pauling é uma propriedade heurística, pois foi definida empiricamente com base em tendências observadas na ligação química, e não a partir de um modelo teórico exato

4 RESULTADOS E DISCUSSÕES

Neste capítulo é apresentado uma abordagem sistemática para a geração de dados de adsorção para o treinamento de modelos de aprendizado de máquina (ML). Para isso, foi construído um banco de dados contendo informações sobre a adsorção de CO_2 e N_2 em estruturas de MOFs e COFs. As propriedades de adsorção foram determinadas por meio de simulações GCMC, que modelaram a interação entre os materiais e as moléculas-alvo. Os dados simulados foram então ajustados ao modelo de Langmuir de dois sítios (DSL) e utilizados na análise preditiva com diferentes algoritmos de ML.

4.1 Banco de Estruturas

4.1.1 QMOF

O banco de dados QMOF, proveniente de Quantum MOF,^[87] disponibiliza informações detalhadas sobre estruturas de MOFs e suas propriedades químicas, calculadas para mais de 14.000 estruturas sintetizadas experimentalmente. Esse banco de estruturas foi desenvolvido a partir de uma seleção criteriosa de materiais contidos no subconjunto de MOFs do Cambridge Structural Database (CSD)^[88] e no banco de dados MOF Experimental (CoRE) de 2019.^[89]

Durante o processo de curadoria, foram removidas estruturas com problemas estruturais, tais como hidrogênios omitidos, ocupações atômicas fracionárias, exclusão de átomos não ligados, átomos sobrepostos, número impróprio de íons de balanceamento de carga, e outros erros que poderiam comprometer a confiabilidade das análises. Todas as estruturas do QMOF foram posteriormente otimizadas em nível de Teoria do Funcional da Densidade (DFT) utilizando funcional de troca e correlação PBE,^[90] empregando a correção de dispersão D3 de Grimme^[91] e o amortecimento de Becke-Johnson (BJ).^[92]

Para a realização deste trabalho, selecionamos apenas as estruturas com diâmetro de poro maior que 3,3 Å, valor correspondente ao raio cinético do CO₂. Após essa filtragem, o número de estruturas disponíveis foi reduzido para 6017. Essa seleção garantiu que as estruturas analisadas fossem compatíveis com a difusão do CO₂ em seus poros, otimizando a relevância dos resultados obtidos

4.1.2 CRAFTED

O banco de dados CRAFTED (***C**harge-dependent, **R**eproducible, **A**ccessible, **F**orcefield-dependent, and **T**emperature-dependent **E**xploratory **D**atabase*)^[51] foi criado com o objetivo de avaliar o impacto da escolha do campo de força e cargas parciais na simulação de adsorção de CO₂ e N₂. Ele se destaca como um recurso abrangente para o estudo de isotermas de adsorção simuladas em MOFs e COFs.

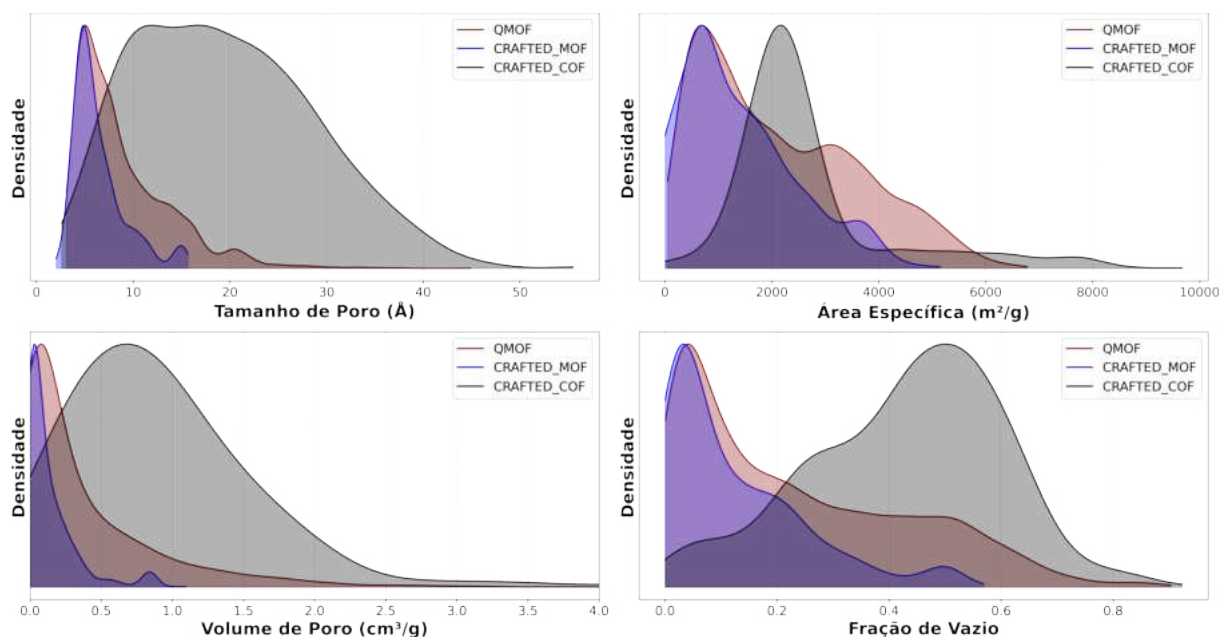
O critério para a inclusão de estruturas no CRAFTED foi a compatibilidade com os campos de forças UFF (Universal Force Field) e DREIDING, assegurando que todas as estruturas selecionadas pudessem ser modeladas com precisão. Após a curadoria, o banco CRAFTED contém 691 MOFs, provenientes do banco CoRE MOF 2014^[93], e 667 COFs, retirados do banco CURATED^[94], totalizando 1358 estruturas.

4.1.3 Comparação dos Banco de dados

A comparação entre os bancos de dados QMOF e CRAFTED revela tanto diferenças quanto semelhanças nas distribuições das propriedades texturais de MOFs e COFs, conforme

ilustrado na Figura 11. Os COFs do CRAFTED apresentam uma ampla variação em tamanho de poro (2-5 nm), volume de poro (0,5-2,5 cm³/g), área específica (500-2500 m²/g) e fração de vazio (0,3-0,8), indicando sua diversidade estrutural. Já os MOFs do QMOF e do CRAFTED exibem distribuições mais próximas, com áreas específicas geralmente entre 1000 e 4000 m²/g e tamanhos de poro variando entre 0,5 e 2,5 nm.

Figura 11 – Comparação entre as principais propriedades geométricas das estruturas presentes no CRAFTED e no QMOF



Fonte: Próprio autor.

Essas diferenças impactam diretamente a adsorção de gases. Materiais com áreas específicas elevadas tendem a favorecer uma maior capacidade de adsorção devido ao aumento da superfície disponível para interação com as moléculas adsorvidas. No entanto, tamanhos de poro muito grandes podem reduzir a seletividade, especialmente para separação de gases como CO₂/N₂, pois dificultam interações específicas e promovem difusão não seletiva. Assim, a ampla gama de propriedades texturais abrangida pelos bancos de dados os torna valiosos para modelagem e descoberta de novos materiais otimizados para processos de adsorção seletiva.

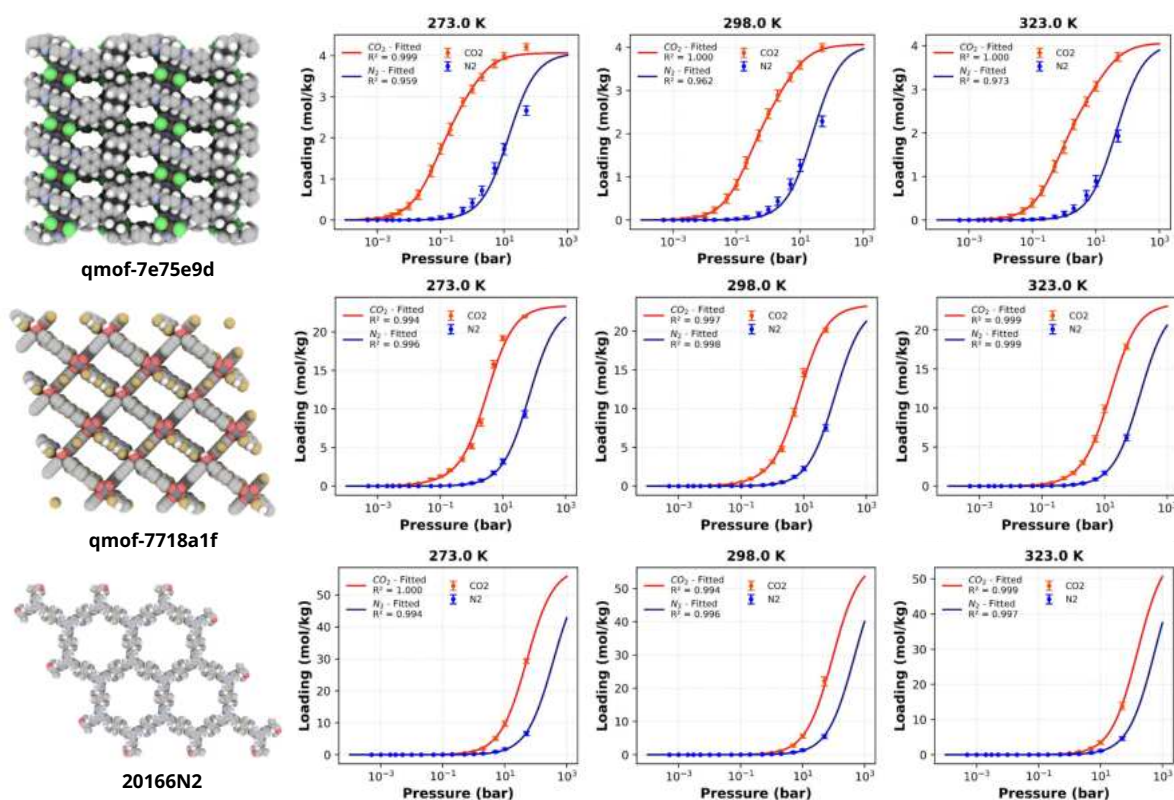
Além das propriedades geométricas dos materiais, a modelagem das isotermas de adsorção é essencial para compreender a capacidade de captura de gases sob diferentes condições experimentais.

4.2 Ajuste ao modelo DSL

As simulações de Monte Carlo no Ensemble Grand Canônico (GCMC) são amplamente utilizadas para estudar a adsorção de gases em materiais porosos, permitindo a obtenção de isotermas sob diversas condições de pressão e temperatura. Neste estudo, foram realizadas cerca de 8.000 simulações GCMC, abrangendo uma ampla gama de pressões e temperaturas para os materiais reticulares selecionados. Como resultado, foi gerado um banco de dados com aproximadamente 45.000 isotermas de CO₂ e N₂ em diferentes materiais, fornecendo uma base robusta para análises posteriores.

A Figura 12 apresenta exemplos de isotermas de adsorção simuladas para três materiais distintos nas temperaturas de 273 K, 298 K e 323 K. As diferenças observadas entre os materiais destacam a diversidade estrutural do banco de dados construído, além de evidenciar a qualidade do ajuste do modelo DSL aos dados de adsorção simulados.

Figura 12 – Isotermas de adsorção para três materiais reticulares distintos nas temperaturas de 273 K, 298 K e 323 K, juntamente com as curvas ajustadas pelo modelo *dual-site Langmuir* (DSL).



Fonte: Próprio autor.

A escolha do ajuste dos dados de adsorção pelo modelo de Langmuir de dois sítios (DSL) é que ele oferece uma vantagem significativa em termos de generalização, permitindo a estimativa da capacidade de adsorção em condições de pressão e temperatura não exploradas no conjunto de treinamento. Como a equação DSL descreve a interação

entre o gás e os sítios adsorção do material, é possível extrapolar para outras temperaturas e pressões os resultados sem a necessidade de novos experimentos ou simulações, desde que os fenômenos de adsorção sigam a mesma tendência física.

Essa capacidade preditiva é particularmente útil para otimizar processos industriais e projetar materiais adsorventes para aplicações específicas. Dessa forma, o uso do ajuste DSL não apenas reduz a necessidade de dados adicionais, mas também amplia a aplicabilidade dos modelos de adsorção em diferentes cenários industriais e ambientais.

No entanto, a escolha da abordagem para estimar os parâmetros do modelo pode influenciar significativamente a qualidade do ajuste. Tradicionalmente, equações linearizadas foram empregadas para simplificar a obtenção dos parâmetros, mas esse método pode introduzir viés estatístico, afetando a precisão dos resultados.

4.2.1 O viés estatístico das equações linearizadas

Com o avanço dos recursos computacionais, a aplicação da regressão não linear por mínimos quadrados tornou-se mais acessível, proporcionando maior robustez e menor sensibilidade a erros.^[95] Esse método minimiza a soma dos erros quadrados, estimando inicialmente os parâmetros do modelo e refinando a solução por meio de um procedimento iterativo.^[96]

Embora a linearização de modelos, que associa os parâmetros à inclinação e intercepto de uma reta, seja uma alternativa popular, especialmente em ajustes de modelos de isotermas a dados experimentais, optamos pela abordagem de regressão não linear neste estudo. Tal decisão se baseia em evidências robustas na literatura, que apontam a regressão não linear como mais adequada em comparação a regressão linear.^[97]

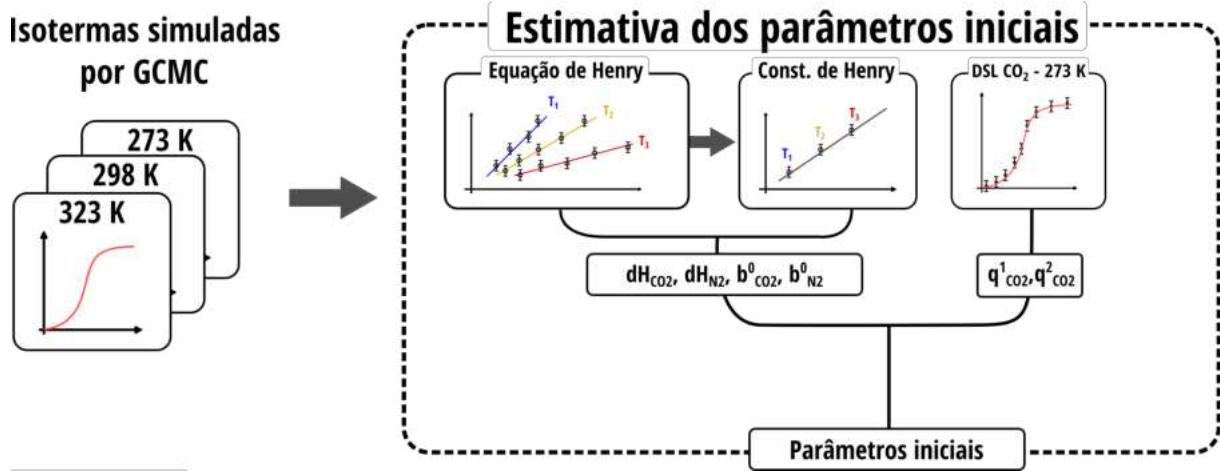
Ainda assim, existem diversos estudos recentes que recorrem ao uso da linearização, o que gera questionamento sobre sua validade, uma vez que a transformação linear pode distorcer a distribuição dos erros, introduzindo um viés estatístico na análise dos dados experimentais.^[97]

4.2.2 Processo de ajuste

O processo de ajuste dos dados simulados de adsorção foi conduzido de maneira sistemática, utilizando um fluxo de trabalho automatizado e robusto, garantindo que o ajuste fosse consistente e representasse adequadamente os fenômenos físico-químicos envolvidos na adsorção. A [Figura 13](#) ilustra o procedimento para a construção dos parâmetros iniciais, baseado nos modelos da equação de Henry e em um ajuste da equação de Langmuir simples, no qual o parâmetro b não possui dependência com a equação de Van't Hoff.

O modelo simples da isoterma de Henry foi inicialmente aplicado para obter uma estimativa inicial da constante K . Essa constante descreve a relação entre a quantidade de

Figura 13 – Ajuste inicial para a isoterma de adsorção com base nos modelos de Henry e Langmuir.



Fonte: Próprio autor.

gás adsorvido e a pressão, em condições de baixa pressão, onde a quantidade adsorvida é diretamente proporcional à pressão parcial. Essa relação é representada pela equação:

$$C = K_H \cdot P \quad (4.1)$$

onde C representa a concentração do gás adsorvido no material, K_H é a constante de Henry, que depende da natureza do gás e do material adsorvente, e P é a pressão parcial do gás. A constante de Henry pode ser relacionada à constante de equilíbrio (K_e), uma vez que descreve a distribuição de equilíbrio de uma espécie química entre a fase gasosa e a superfície do sólido, em condições de diluição infinita, conforme a lei de Henry. A partir dessa relação, podemos calcular a constante de equilíbrio utilizando a equação de Van't Hoff.

Para uma reação reversível, a constante de equilíbrio é determinada em várias temperaturas, e esses dados são esboçados em um gráfico de Van't Hoff, com $\ln K_{eq}$ no eixo y e $\frac{1}{T}$ no eixo x. A linearidade desses dados permite ajustar a equação na forma linear de Van't Hoff:

$$\ln K_{eq} = -\frac{\Delta H}{R} \frac{1}{T} + \frac{\Delta S}{R}. \quad (4.2)$$

Essa equação é amplamente utilizada para estimar a entalpia (ΔH) e a entropia (ΔS) de um processo de adsorção. A inclinação da linha ajustada corresponde a $-\frac{\Delta H}{R}$, enquanto a interceptação com o eixo y é equivalente a $\frac{\Delta S}{R}$. Assim, as equações para determinar a entalpia e a entropia são dadas por:

$$\Delta H = -R \cdot \text{inclinação}, \quad \Delta S = R \cdot \text{interceptação}. \quad (4.3)$$

A partir dos dados simulados de adsorção nas temperaturas de 273 K, 298 K e 323 K, foi possível ajustar a isoterma de Henry e realizar a linearização dos dados para estimar ΔH tanto para CO_2 quanto para N_2 . Em seguida, esses dados foram ajustados utilizando a isoterma de Langmuir simples de dois sítios, onde o parâmetro b não apresenta dependências adicionais e assim obter os valores para o chute inicial de q_s .

Ao ajustar a equação de Van't Hoff para determinar a constante de Langmuir, os parâmetros obtidos, como a entalpia de adsorção (ΔH) e o fator pré-exponencial (b_0), devem refletir valores compatíveis com a natureza exotérmica do processo de adsorção.

Devido ao número elevado de parâmetros envolvidos, é possível que haja uma considerável degenerescência na obtenção de valores capazes de reproduzir a isoterma. Por essa razão, foi fundamental adotar uma abordagem sistemática no ajuste, utilizando valores iniciais bem fundamentados. Isso direciona o ajuste para soluções que possuam maior relevância físico-química, evitando a degenerescência numérica. Essa necessidade justifica a atenção especial dada ao processo de definição dos valores iniciais ao longo do estudo.

Após a obtenção dos parâmetros iniciais, foi iniciado o processo de ajuste conforme ilustrado no fluxograma da [Figura 14](#). Com base nos valores iniciais estimados para q_{1,CO_2}^S e q_{2,CO_2}^S , foi definido um critério de decisão: se o valor de q_{2,CO_2}^S fosse inferior a 30% de q_{1,CO_2}^S , optou-se pelo modelo de isoterma de Langmuir de um sítio (SSL), em vez do modelo de dois sítios (DSL).

Após a escolha do modelo apropriado, foi realizada uma verificação dos parâmetros iniciais em busca de valores problemáticos, como infinitos ou Not-a-number (resultado obtido quando o processo de ajuste não converge). Quando tais valores eram detectados, o módulo `curve_fit` era chamado com o método de ajuste 'lm', que não requer valores iniciais definidos. Caso contrário, os métodos 'trf' e 'dogbox' foram testados.¹

Para todos os métodos, o critério de sucesso do ajuste foi baseado no coeficiente de determinação (R^2). O método que apresentasse o maior valor de R^2 era selecionado. Caso nenhum dos métodos conseguisse realizar o ajuste ou se o ajuste obtivesse um $R^2 < 0.7$, novos parâmetros iniciais eram testados. Ajustes com $R^2 \geq 0.7$ eram considerados bem-sucedidos e aceitos como resultados finais, caso o contrário, o parâmetros iniciais eram tidos como os finais, mas com $R^2 > 0$ a fim de contabilizar quais materiais tiveram um bom ajuste.

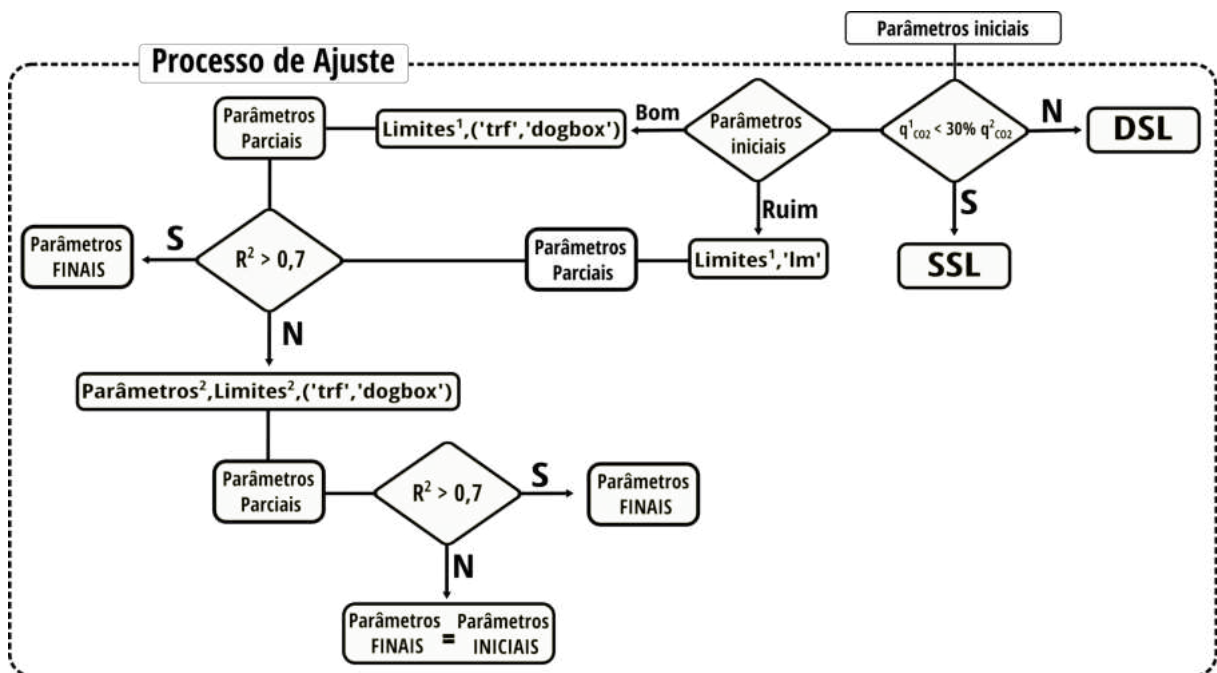
Apesar de adotarmos $R^2 \geq 0.7$ como critério de aceitação, a maioria dos ajustes

¹ Esses métodos representam abordagens distintas para realizar o processo de ajuste, cada um com características específicas para lidar com diferentes tipos de dados e condições iniciais.

apresentou um coeficiente de determinação superior a 0.9, como ilustrado na [Figura 18](#). No total, 7185 dos 7375 materiais analisados atenderam a esse critério.

A coerência dos parâmetros ajustados com a realidade física é essencial para a sua validação enquanto modelo. Parâmetros que não fazem sentido físico indicam que o modelo pode não estar capturando adequadamente os processos subjacentes do sistema em estudo. A interpretação dos resultados do modelo ajustado depende diretamente da significância física dos parâmetros obtidos, especialmente considerando que queremos treinar modelos de ML, que buscam padrões nos dados existentes

Figura 14 – Fluxograma representando o processo de ajuste dos parâmetros do modelo dual-site Langmuir (DSL) e single-site Langmuir (SSL). O processo inicia com a avaliação dos parâmetros q_{1,CO_2}^S e q_{2,CO_2}^S , e o critério de decisão para a escolha do modelo mais adequado. Em seguida, são testados diferentes métodos de ajuste (*trf*, *dogbox*, *lm*) com base na qualidade do ajuste (R^2), utilizando parâmetros iniciais ou novos valores em caso de falhas.

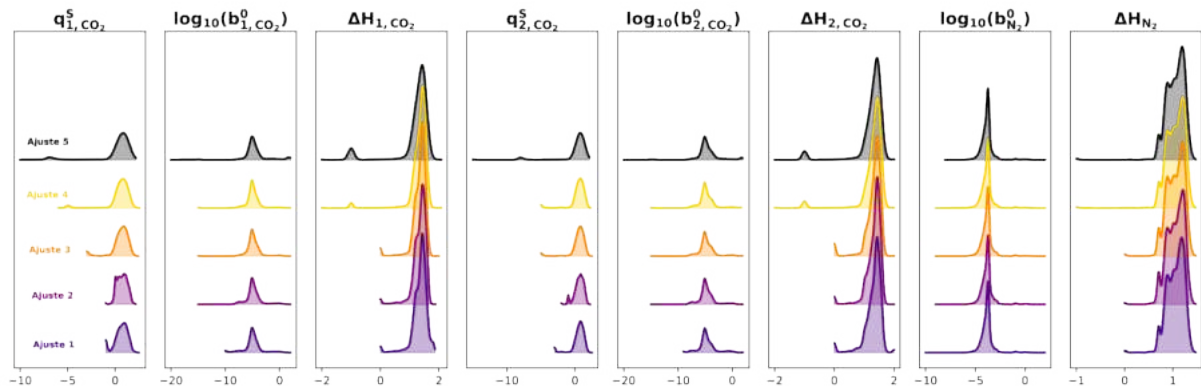


Fonte: Próprio autor.

A [Figura 15](#) apresenta a distribuição dos parâmetros ajustados para diferentes condições de limites de ajuste. Ao analisar os histogramas, observa-se que ajustes realizados com limites mais amplos resultaram em distribuições de valores mais uniformes e próximas de uma distribuição normal. No entanto, ajustes com limites mais restritos geraram distribuições concentradas ao redor dos próprios limites, resultando em 'picos' de distribuição. Esses picos indicam que, quando os limites de ajuste são muito apertados, os parâmetros ajustados tendem a se agrupar próximo a esses limites, o que distorce a normalidade esperada e pode afetar a precisão do modelo. Esse comportamento ajudou a entender a importância de escolher limites de ajuste adequados para evitar distorções nas

distribuições dos parâmetros e garantir um modelo mais robusto.

Figura 15 – Distribuição dos parâmetros do modelo DSL para os diferentes ajustes aplicados aos dados de simulação GCMC.



Fonte: Próprio autor.

Além disso, foi analisado o impacto dos diferentes ajustes nas distribuições das métricas de avaliação, como o coeficiente de determinação (R^2), o erro quadrático médio (MSE) e a porcentagem de variância explicada (EVP), conforme apresentado na Figura 16. Observou-se que, independentemente dos limites de ajuste adotados, não houve mudanças significativas nas distribuições dessas métricas. Isso sugere que, embora as distribuições dos parâmetros ajustados variem conforme os limites, a performance geral do modelo, medida por essas métricas, permaneceu estável. Para ilustrar esse ponto, a Figura 17 apresenta uma isoterma calculada via CGMC para o MOF qmof-d9ca2be com os valores dos ajustes testados, onde é possível perceber que a curva resultante é idêntica para todos os ajustes. Isso indica que, mesmo com diferentes limites, os valores encontrados para os ajustes não apresentaram diferenças relevantes para recuperar a isoterma. No entanto, vale destacar que essa degenerescência nos parâmetros ajustados, embora não tenha mostrado impactos imediatos nas métricas de avaliação, pode se tornar problemática no futuro, principalmente na distribuição dos valores desses parâmetros quando forem submetidos aos modelos de aprendizado de máquina (ML).

O resultado do ajuste final utilizado é apresentado na Figura 18, que exibe a R^2 , MSE e EVP. Observa-se que os valores de R^2 próximos de 1 indicam que os parâmetros ajustados fornecem uma boa qualidade de ajuste de forma global, o que também é corroborado pela distribuição de valores obtidos para o MSE próximo de 0 e pelo EVP em torno de 100%.

Na Figura 19, podemos observar a distribuição dos parâmetros obtidos pelo ajuste. Os parâmetros q_{1,CO_2}^S e q_{2,CO_2}^S , expressos em mmol/g, representam a capacidade de adsorção em estado de saturação. Os valores de b^0 (b_{1,CO_2}^0 , b_{2,CO_2}^0 e $b_{N_2}^0$) são adimensionais. Já os parâmetros ΔH ($\Delta H_{1,CO_2}$, $\Delta H_{2,CO_2}$ e ΔH_{N_2}), medidos em kcal/mol, representam a energia

Figura 16 – Distribuição das métricas de avaliação para os diferentes ajustes aplicados aos dados de simulação GCMC.

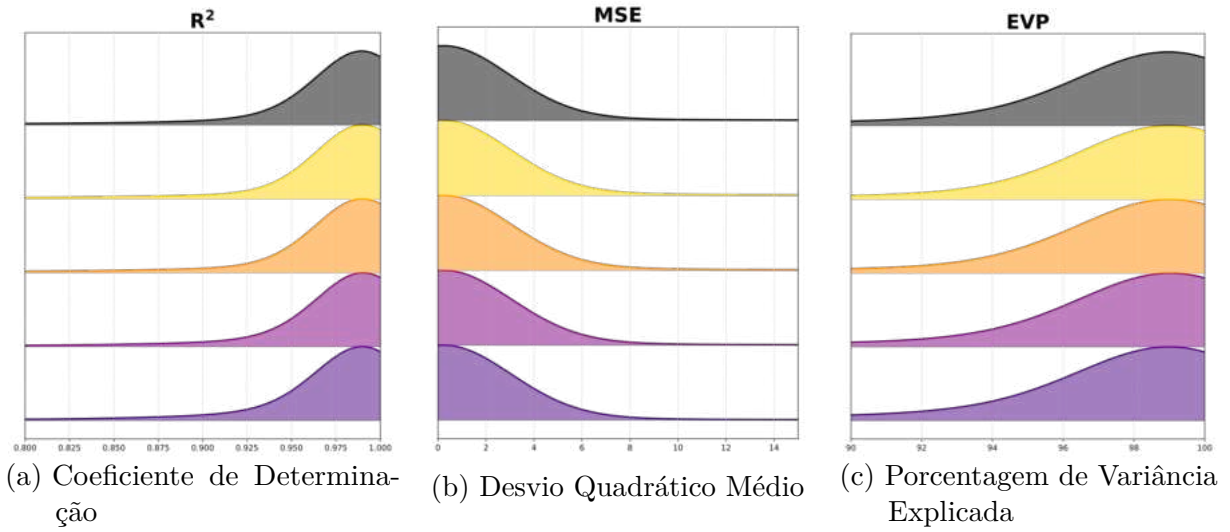


Figura 17 – Isotermas de adsorção de um MOF nas temperaturas de 273 K, 298 K e 323 K, juntamente com as curvas ajustadas pelo modelo *dual-site Langmuir* (DSL).

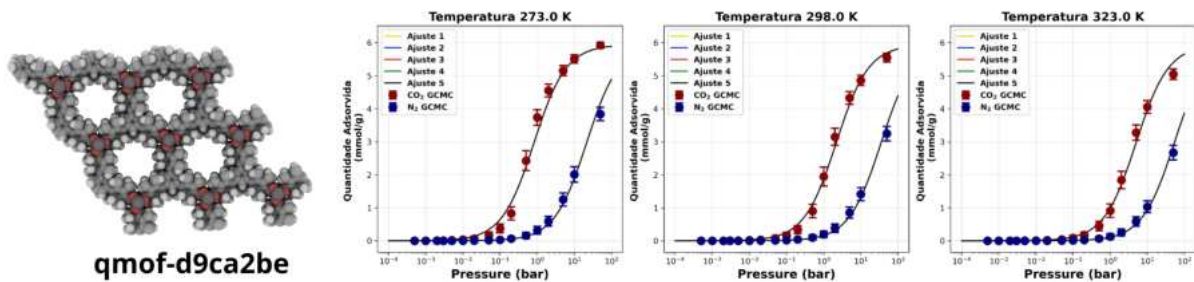
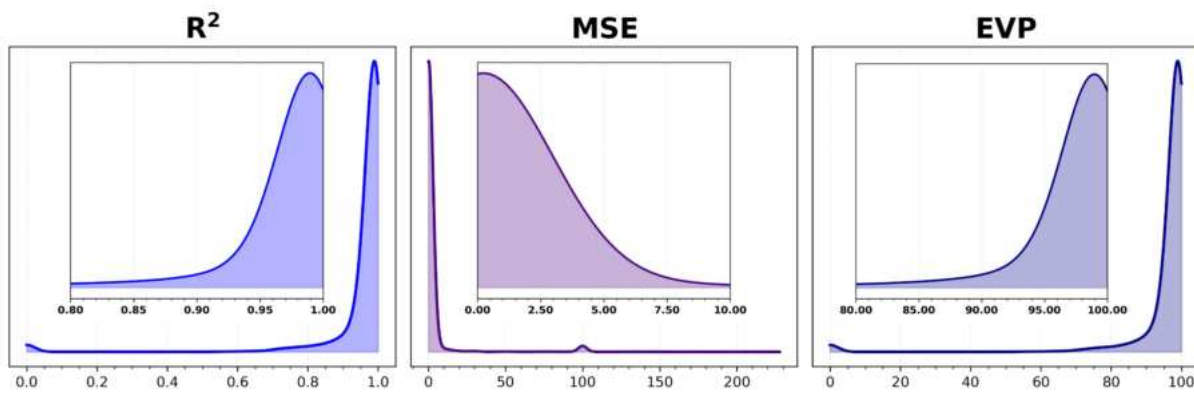


Figura 18 – Distribuições dos ajustes realizados nos dados de simulação GCMC para o ajuste final escolhido: Coeficiente de Determinação (R^2), Erro Quadrático Médio (MSE) e Porcentagem de Variância Explicada (EVP)

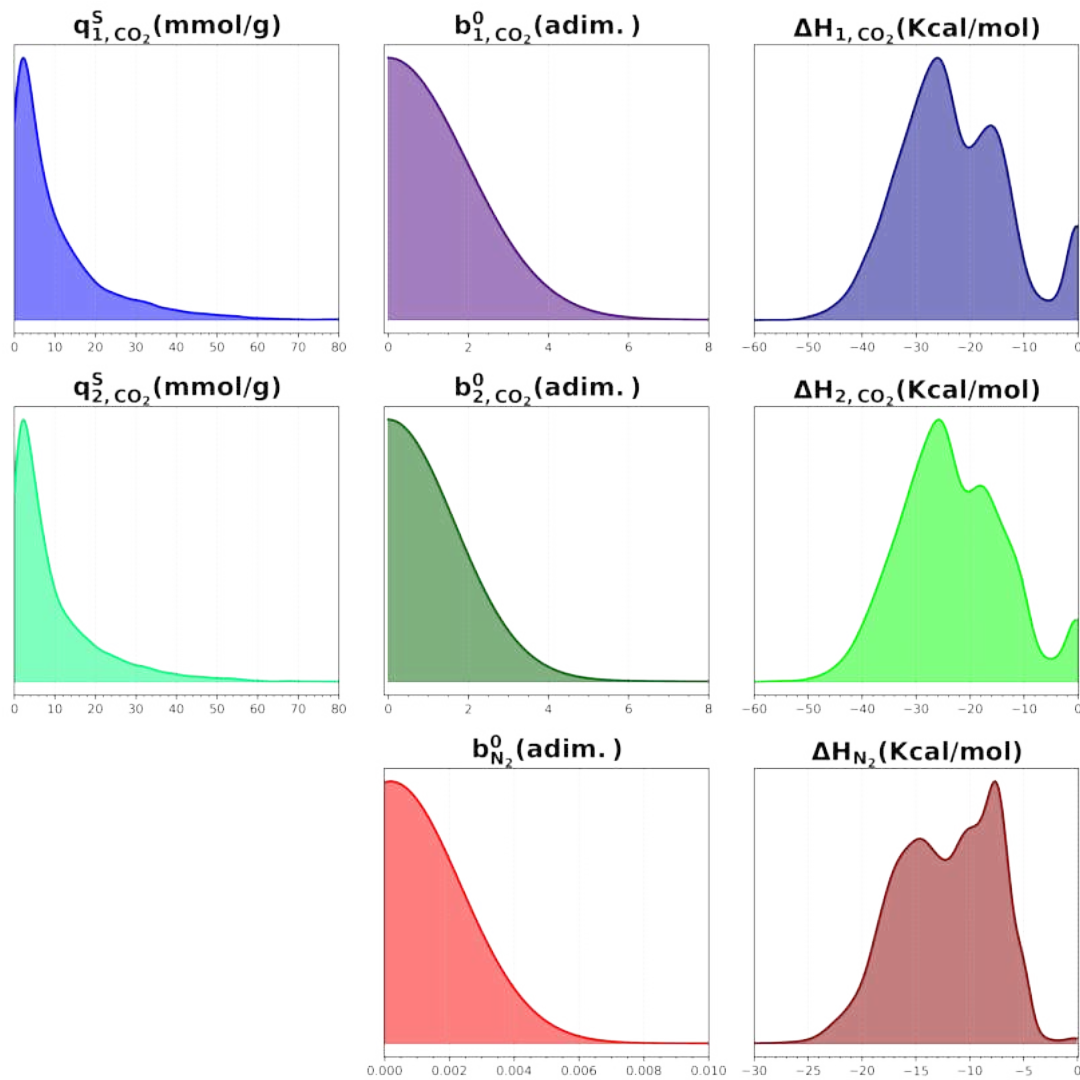


Fonte: Próprio autor.

de adsorção, que pode fornecer informações sobre a força da interação entre as moléculas de gás e a superfície do material.

É possível notar que os valores dos parâmetros q^S e b^0 não seguem uma distribuição normal (Gaussiana), enquanto os valores de ΔH exibem um comportamento mais próximo dessa distribuição. Essa diferença nas distribuições dos parâmetros pode impactar a aplicação dos modelos de aprendizado de máquina (ML), influenciando a maneira como os algoritmos interpretam e processam os dados durante o treinamento e a predição.

Figura 19 – Distribuição dos parâmetros ajustados da isoterma DSL para o conjunto de materiais analisados, destacando a variação entre os diferentes valores obtidos.



Fonte: Próprio autor.

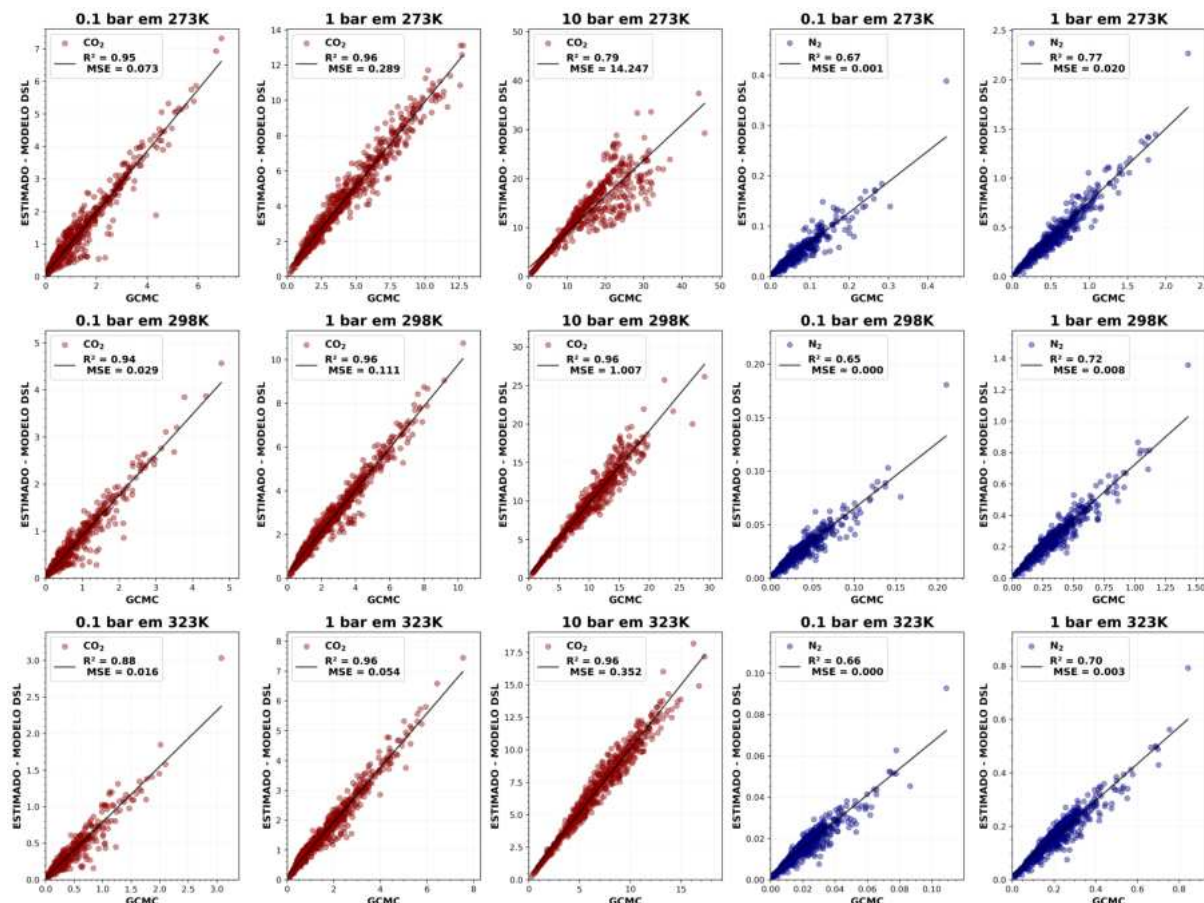
A Figura 20 apresenta a comparação entre os valores de capacidade de adsorção de CO₂ e N₂ obtidos por meio de simulações GCMC e aqueles estimados pela equação de Langmuir de dois sítios (DSL), ajustada com base nos dados simulados. A análise foi realizada para as temperaturas de 273 K, 298 K e 323 K, considerando pressões de 0,1 e 1

bar para o N_2 , e de 0,1, 1 e 10 bar para o CO_2 . Esse intervalo de pressões permite avaliar o comportamento dos parâmetros ajustados sob diferentes condições experimentais.

Os coeficientes de determinação (R^2) indicam uma excelente correlação entre os valores de adsorção simulados e estimados para o CO_2 , com R^2 superiores a 0,95 para 1 bar e 0,80 para 0,1 bar e 10 bar, evidenciando a adequação da equação DSL para descrever sua adsorção. Para o N_2 , a correlação, embora menor, ainda é satisfatória, com R^2 superiores a 0,65. Essa diferença pode estar associada ao menor número de parâmetros ajustados para o N_2 em comparação ao CO_2 . Apesar disso, a análise do erro médio quadrático (MSE) revela valores mais próximos de zero para o N_2 , indicando menor dispersão nos dados ajustados em comparação ao CO_2 .

Esses resultados destacam a importância da escolha adequada do modelo de adsorção para diferentes gases e faixas de pressão. No entanto, para expandir a capacidade preditiva das isotermas e reduzir o tempo computacional necessário para ajustes, exploramos a aplicação de modelos de aprendizado de máquina. Essa abordagem permite prever parâmetros de adsorção com base em descritores físicos e estruturais dos materiais, dispensando simulações adicionais e possibilitando a análise de um número significativamente maior de materiais.

Figura 20 – Comparação entre os valores de quantidade adsorvida de CO₂ e N₂ obtidos por simulações GCMC e os valores previstos pelo modelo *dual-site Langmuir* (DSL) nas temperaturas de 273 K, 298 K e 323 K. As pressões consideradas foram 0,1, 1 e 10 bar para CO₂ e 0,1 e 1 bar para N₂.



Fonte: Próprio autor.

4.3 Aprendizado de Máquina

4.3.1 Métricas para Avaliação de Modelos de Aprendizado de Máquina

A escolha adequada das métricas de desempenho é fundamental para avaliar modelos de regressão, garantindo que os resultados estejam alinhados com os objetivos da predição dos parâmetros das isotermas. Diferentes métricas podem ser utilizadas, cada uma com características que a tornam mais ou menos apropriada conforme o contexto do problema.

Uma das métricas mais utilizadas é o erro quadrático médio (MSE) (3.34), que calcula a média dos quadrados das diferenças entre os valores previstos e os observados. O MSE é especialmente relevante em situações onde grandes desvios devem ser minimizados, pois penaliza erros maiores de forma mais intensa. Assim, valores menores de MSE indicam um melhor desempenho do modelo, sendo que o valor ideal é zero, representando previsões

de grande acurácia.^[98]

Outra métrica amplamente empregada é o coeficiente de determinação (R^2) (3.33), que quantifica a proporção da variabilidade dos dados explicada pelo modelo. Valores de R^2 próximos de 1 indicam que o modelo explica grande parte da variação nos dados, enquanto valores mais baixos sugerem um ajuste menos eficaz. Essa métrica complementa o MSE ao fornecer uma avaliação global da adequação do modelo.

Dessa forma, a avaliação dos modelos de aprendizado de máquina neste estudo é realizada com base no MSE e no R^2 , permitindo uma análise abrangente do desempenho preditivo.

4.3.2 Predição dos parâmetros da Isoterma

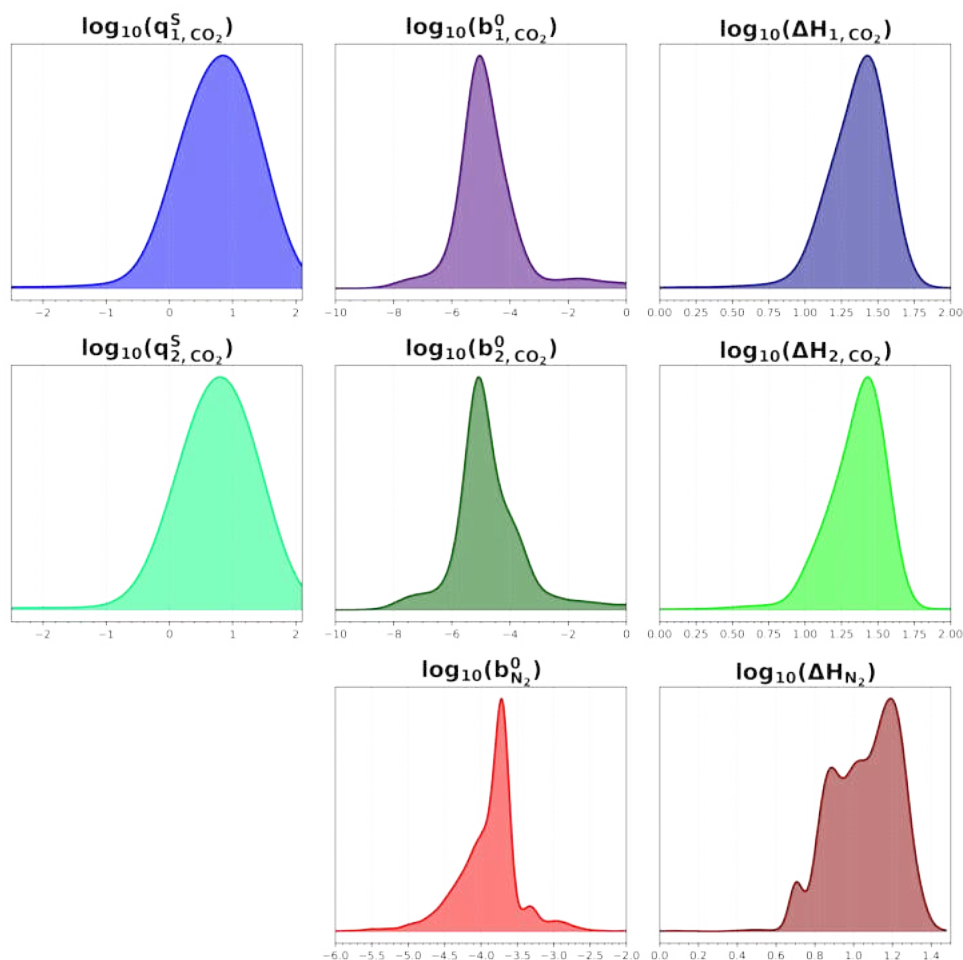
Para cada modelo, utilizou-se uma arquitetura do tipo *multiobjetiva*, na qual um único modelo foi empregado para prever simultaneamente todos os parâmetros responsáveis pela geração das isotermas de adsorção. Inicialmente, foi realizada uma etapa de pré-processamento dos dados, envolvendo a limpeza e a normalização dos valores do conjunto de treinamento, a fim de preparar os dados para os modelos de aprendizado de máquina.

Os resultados obtidos demonstram que a normalização dos dados, quando realizada em uma escala logarítmica (na base 10), proporcionou um melhor desempenho dos modelos. Esta transformação permitiu uma distribuição mais uniforme dos valores, como podemos observar na [Figura 21](#) em comparação com a distribuição sem aplicação da escala log na [Figura 19](#).

Essa característica de uma distribuição normal é relevante para os algoritmos de aprendizado de máquina, especialmente aqueles baseados em métodos estatísticos, como a regressão linear e a inferência Bayesiana, que assumem implícita ou explicitamente que os dados seguem uma distribuição normal.^[99] Tal suposição não apenas simplifica os cálculos e a modelagem, mas também contribui para que os modelos façam previsões mais robustas e generalizáveis. A distribuição normal desempenha um papel crucial na estimativa de incertezas e na construção de intervalos de confiança. Temos então não só boa qualidade no ajuste, mas também uma distribuição consistente para ser utilizado nos modelos de aprendizado de máquina.

Os resultados dos modelos aplicados aos dados sem escala logarítmica estão apresentados na [Apêndice A](#). Nessa seção, é possível observar a perda de desempenho dos modelos que não utilizaram a normalização adequada dos dados, em comparação com a abordagem de normalização, que será discutida posteriormente. O conjunto final foi dividido em treino (70%), teste (15%) e validação (15%), assegurando qualidade e confiabilidade para a modelagem.

Figura 21 – Distribuição dos parâmetros ajustados da isoterma DSL para o conjunto de materiais analisados, destacando a variação entre os diferentes valores obtidos. A aplicação do logaritmo remove as unidades dos parâmetros



Fonte: Próprio autor.

4.3.2.1 Escolha dos Descritores e de Modelo

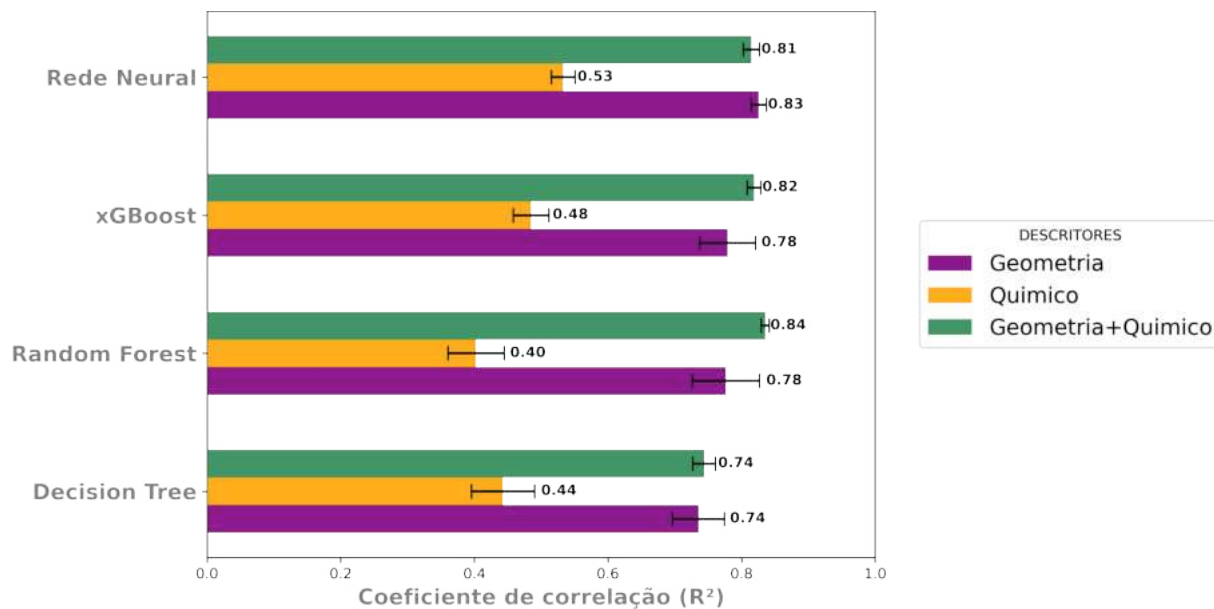
A adsorção de moléculas em materiais reticulares é um fenômeno complexo, influenciado pela interação entre características geométricas e químicas do material. Os descritores geométricos fornecem informações sobre o tamanho e o volume dos poros, enquanto os descritores químicos representam a natureza química da superfície.

O uso isolado de um único conjunto de descritores pode limitar a compreensão do fenômeno. Assim, a combinação de ambos permite uma abordagem mais abrangente, proporcionando uma representação mais precisa do comportamento da adsorção e tornando a modelagem mais robusta.

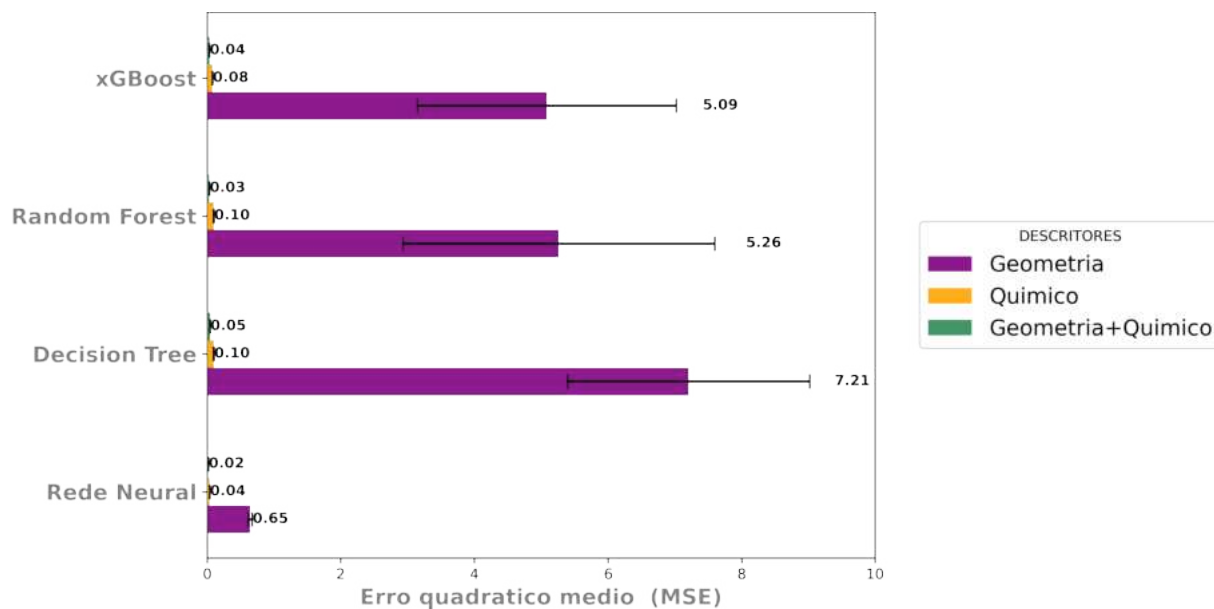
Os resultados apresentados na [Figura 22a](#) comparam os valores de R^2 , enquanto a [Figura 22b](#) exibe os valores de MSE para diferentes modelos de aprendizado de máquina (ML). As análises consideram descritores geométricos e químicos, tanto isoladamente quanto em conjunto.

Figura 22 – Comparação dos desempenhos dos modelos de aprendizado de máquina nas métricas R^2 e MSE.

(a) Comparação dos valores de R^2 entre os modelos de aprendizado de máquina. Quanto mais próximo de 1, melhor é a correlação dos resultados preditos com o verdadeiros



(b) Comparação do erro quadrático médio (MSE) entre os modelos de aprendizado de máquina. Quanto mais próximo de 0, menor é o desvio dos resultados preditos com o verdadeiros



Fonte: Próprio autor.

Para garantir a robustez dos resultados e evitar *overfitting*, cada modelo foi ajustado por meio da busca em grade (*Grid Search*) combinada com validação cruzada de cinco *folds* (*5-Fold Cross-Validation*). Além disso, a otimização dos hiperparâmetros foi realizada individualmente para cada conjunto de descritores, permitindo avaliar seu impacto no desempenho dos modelos.

Os dados indicam que o modelo Random Forest apresentou o melhor desempenho, seguido pela rede neural e pelo XGBoost, enquanto a árvore de decisão obteve os piores resultados. Além disso, a combinação de descritores geométricos e químicos resultou em previsões mais precisas, evidenciada por valores mais elevados de R^2 e menores MSE em comparação ao uso isolado de cada conjunto de descritores.

A análise confirma que a combinação de características químicas e geométricas melhora a capacidade dos modelos em capturar a complexidade dos parâmetros avaliados. Isso reforça que uma abordagem integrada, considerando simultaneamente esses aspectos, é mais eficaz para representar as propriedades de adsorção. Portanto, integrar essas informações é essencial para compreender com precisão a capacidade de adsorção e desenvolver modelos mais representativos do comportamento dos materiais sob diferentes condições operacionais.

4.3.2.2 Modelo Random Forest

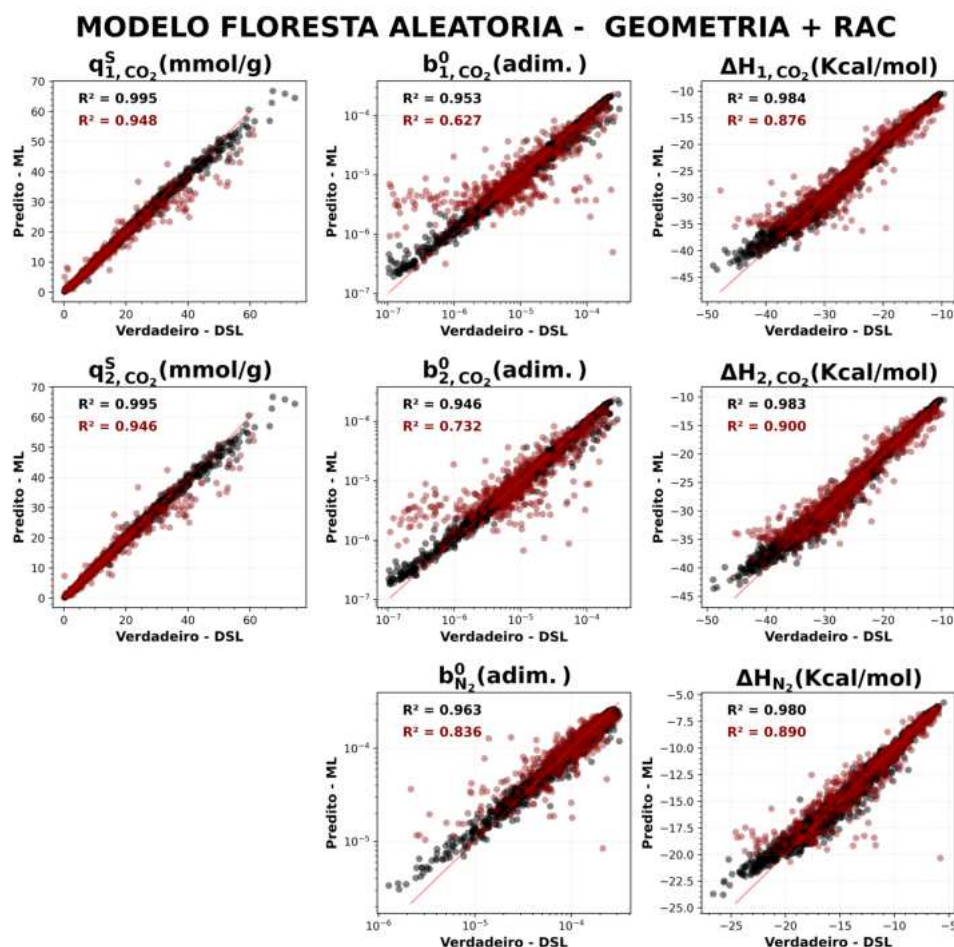
Dentre os modelos avaliados, o Random Forest, utilizando a combinação de descritores geométricos e químicos (RAC), apresentou o melhor desempenho, justificando uma análise mais aprofundada de seus limites e aplicações. A [Figura 23](#) ilustra a correlação entre os valores de referência e os valores preditos pelo modelo na predição, demonstrando sua capacidade de capturar a complexidade dos parâmetros analisados.

A [Figura 24](#) apresenta as isotermas de adsorção de CO_2 e N_2 de 3 MOFs distintos para as temperaturas de 273 K, 298 K e 323 K. Além disso, são exibidas as curvas ajustadas pelo modelo DSL, tanto a partir dos parâmetros obtidos diretamente do ajuste quanto daqueles preditos pelo modelo Random Forest.

Podemos observar que há uma boa correlação entre os dados de GCMC, os ajustes pelo modelo DSL e as previsões do modelo de aprendizado de máquina, indicando a capacidade do ML em reproduzir com precisão o comportamento da adsorção. No entanto, nota-se um leve desvio nas previsões para temperaturas mais elevadas, o que pode estar relacionado a efeitos não completamente capturados pelo modelo.

Esses resultados demonstram o potencial da abordagem baseada em ML para a obtenção de isotermas de adsorção sem a necessidade de novas simulações, permitindo a extrapolação do comportamento do material para diferentes condições de temperatura e pressão.

Figura 23 – Correlação entre Valores de Referência e Valores Preditos pelo Modelo Random Forest. Os resultados do conjunto de treinamento estão em preto, enquanto os do conjunto de validação estão em vermelho.

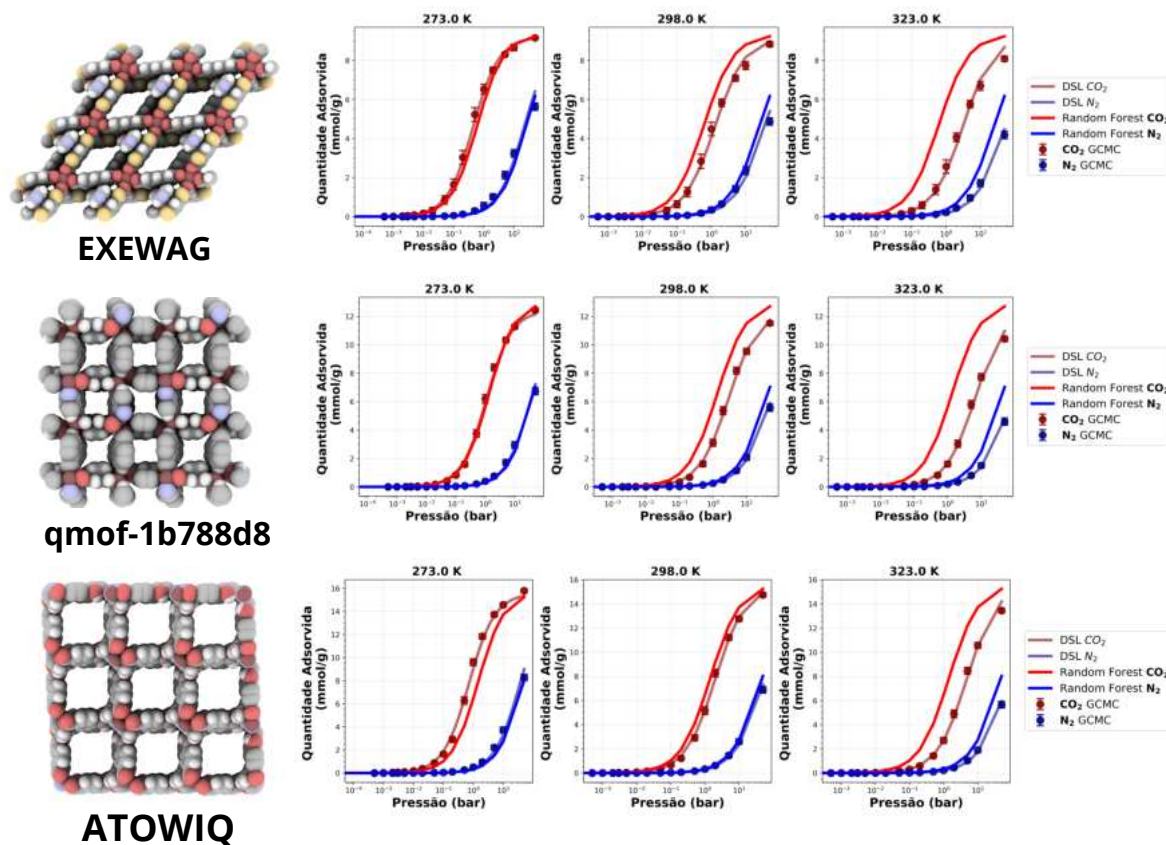


A [Figura 25](#) compara a quantidade de gás adsorvido obtida a partir dos parâmetros DSL (estimados via GCMC) e dos valores preditos pelo modelo Random Forest com os resultados das simulações GCMC, sob diferentes condições de pressão e temperatura. Observa-se uma excelente entre os valores obtidos com DSL e os de referência (GCMC), especialmente para o CO_2 . Além disso, essa correlação melhora em condições de menor temperatura e maior pressão, conforme indicado pelos valores de R^2 e MSE ajustados.

Para o N_2 , o desempenho dos valores obtidos via DSL foi ligeiramente inferior, possivelmente devido ao menor número de parâmetros disponíveis para descrever sua adsorção com precisão. Esse efeito também é observado na [Figura 20](#).

A redução nos valores de R^2 observada na comparação entre os valores de gás adsorvido estimados pelo modelo Random Forest e os obtidos pelas simulações GCMC pode ser atribuída à natureza do modelo. Como o Random Forest foi treinado com parâmetros ajustados pelo modelo DSL, que já apresentam um pequeno desvio, especialmente para o N_2 (conforme mostrado na [Figura 20](#)), a correlação final tende a ser ligeiramente reduzida.

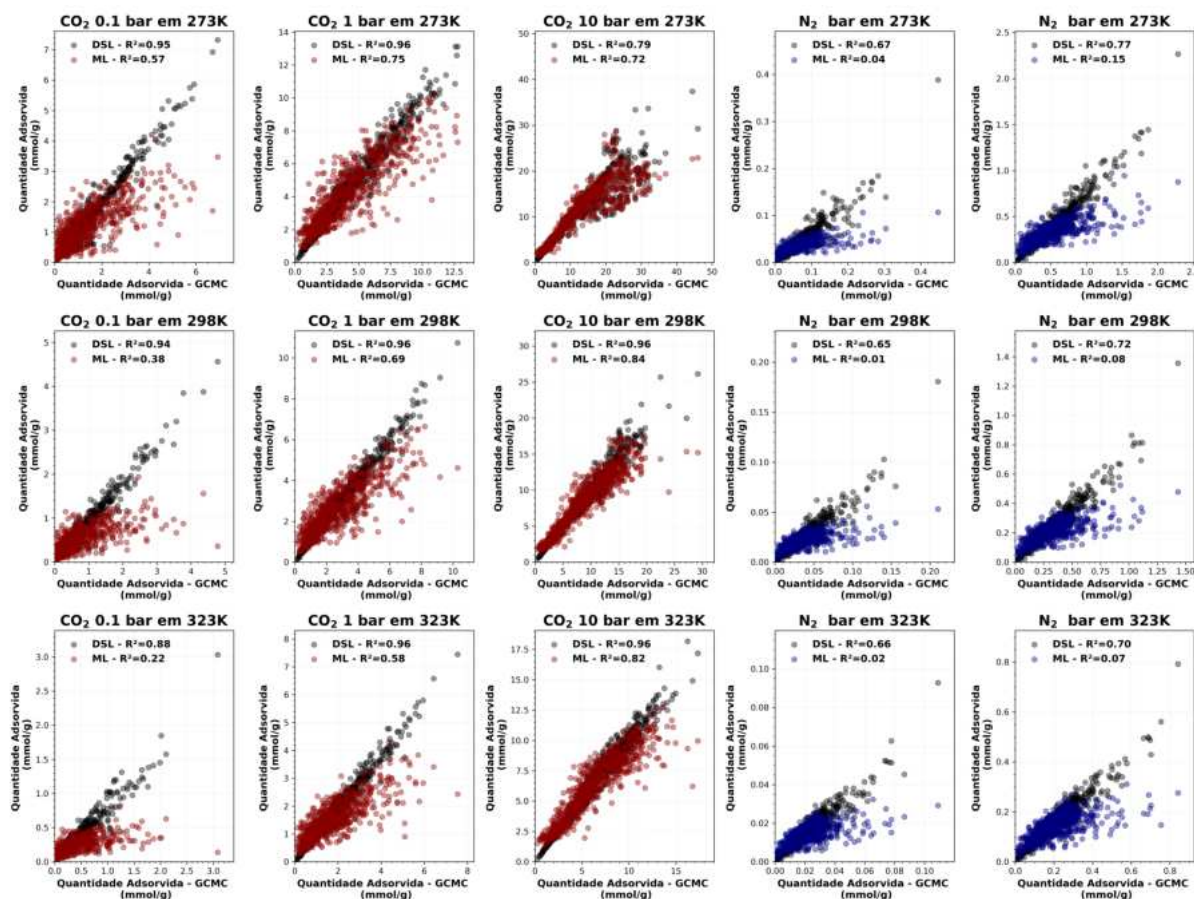
Figura 24 – Isotermas de adsorção para três materiais reticulares distintos nas temperaturas de 273, 298 e 323 K, juntamente com os dados de adsorção calculados por simulação GCMC. Também são apresentadas as curvas da isoterma obtidas pelo modelo *dual-site Langmuir* (DSL) e pelo modelo de aprendizado de máquina *Random Forest*. Esses materiais fazem parte do conjunto de teste.



Fonte: Próprio autor.

No entanto, esse comportamento é esperado, especialmente considerando que o modelo não visa decorar os dados, mas sim criar uma representação generalizada dos mesmos. Portanto, a variação no erro para o N₂ pode ser vista como uma consequência natural da capacidade do modelo de generalizar, em vez de simplesmente ajustar-se aos dados de treinamento. Apesar disso, a correlação entre os valores preditos e os valores de referência ainda se mantém satisfatória, indicando que o modelo tem uma capacidade preditiva adequada, mesmo diante dessas pequenas discrepâncias.

Figura 25 – Distribuição da quantidade de gás adsorvido quando estimado pelos parâmetros ajustados via GCMC com os valores preditos pelo modelo *Random Forest* em comparação com os obtidos via simulação GCMC. Os resultados são apresentados para CO₂ nas temperaturas de 273 K, 298 K e 323 K, com pressões de 0.1, 1 e 10 bar, e para N₂ nas pressões de 0.1 e 1 bar.



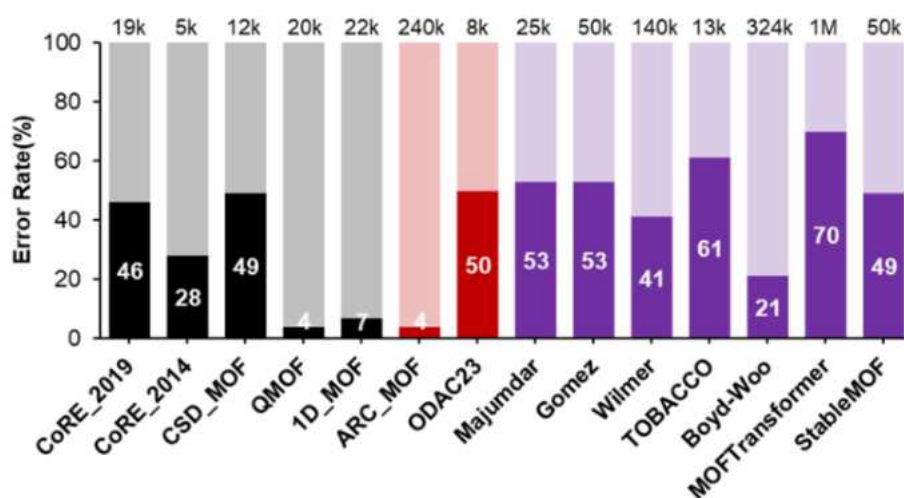
4.4 Avaliando o modelo fora do conjunto de treinamento

O banco de dados ARC-MOF ^[100] é uma das coleções mais abrangentes de estruturas de MOFs disponíveis, reunindo aproximadamente 280.000 estruturas, das quais muitas são caracterizadas experimentalmente, enquanto outras são modelos computacionais de estruturas hipotéticas. Destaca-se por integrar cargas atômicas parciais derivadas de cálculos de Teoria do Funcional da Densidade (DFT) e uma vasta gama de descritores químicos e geométricos, tornando-o uma excelente fonte para aplicação em modelos de aprendizado de máquina.

Este banco de dados foi cuidadosamente projetado para minimizar erros estruturais, utilizando uma versão inicial do algoritmo MOSAEC para identificar e descartar estruturas químicas inválidas durante sua construção. ^[101] A análise de taxas de erro (apresentado na Figura 26) revela a consistência e precisão do ARC-MOF, destacando-o entre outros bancos de dados, particularmente aqueles compostos por hMOFs. A taxa de erro de 4% é um indicativo de quão rigorosa e eficaz foi a curadoria dos dados, o que é crucial para garantir

que os modelos preditivos não sejam influenciados por estruturas defeituosas ou errôneas. Essa qualidade do ARC-MOF é uma vantagem significativa quando se trabalha com modelos de aprendizado de máquina que exigem dados de alta fidelidade para generalizar bem.

Figura 26 – Taxas de erro em 14 bancos de dados de MOFs. Bancos de dados compostos exclusivamente ou principalmente por estruturas experimentais estão coloridos em preto, bancos de dados de hMOFs em roxo, e bancos de dados híbridos (com estruturas experimentais e hipotéticas) estão coloridos em vermelho. O número total de MOFs por conjunto de dados é indicado acima da barra correspondente.

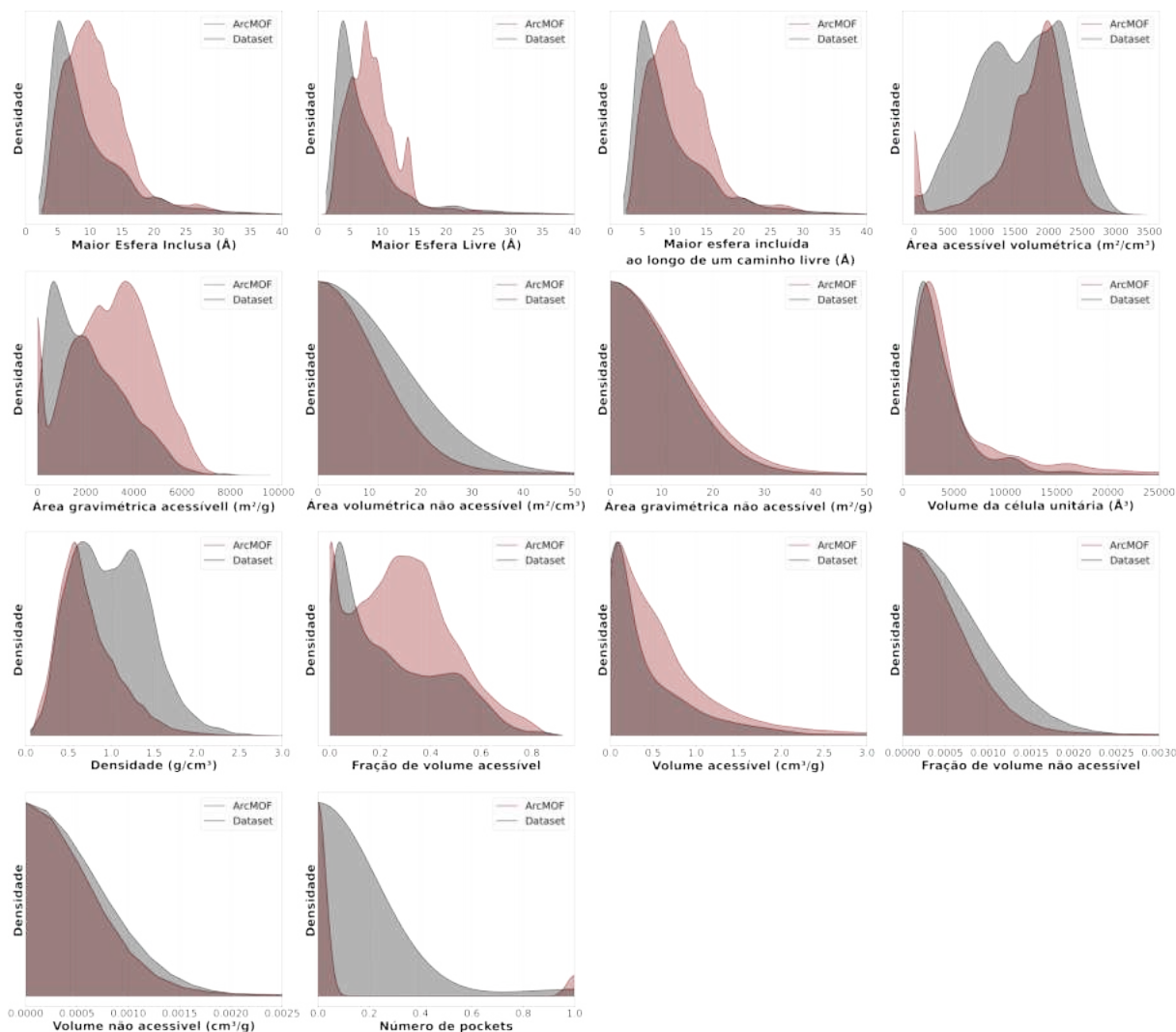


Fonte: Retirado de [White et al.](#)^[101]

A aplicação do modelo Random Forest foi estendida ao banco de dados ARC-MOF, utilizando descritores geométricos e químicos escalados na base logarítmica. A [Figura 27](#) apresenta a distribuição dos descritores geométricos tanto no banco de dados ARC-MOF quanto no conjunto de dados utilizado para o treinamento do modelo. A análise dessa distribuição revela que os valores dos descritores no ARC-MOF possuem padrões semelhantes aos observados nos dados de treinamento, o que é fundamental para assegurar a robustez e a confiabilidade do modelo quando aplicado ao ARC-MOF. Essa correspondência nas distribuições dos descritores reforça a capacidade do modelo de generalizar e fornecer previsões precisas para as estruturas presentes no ARC-MOF.

A [Figura 28](#) mostra a distribuição dos descritores RAC, que caracterizam as propriedades químicas dos MOFs. Para facilitar a visualização, somou-se cada subgrupo desses descritores, reduzindo sua dimensionalidade. Nota-se que a distribuição dos valores no banco de dados ARC-MOF é semelhante à do conjunto de treinamento, sugerindo uma correspondência química entre as estruturas. Essa similaridade indica que o modelo pode realizar previsões precisas para os descritores do ARC-MOF.

Figura 27 – Distribuição de descritores geométricos do banco de dados ARC-MOF.

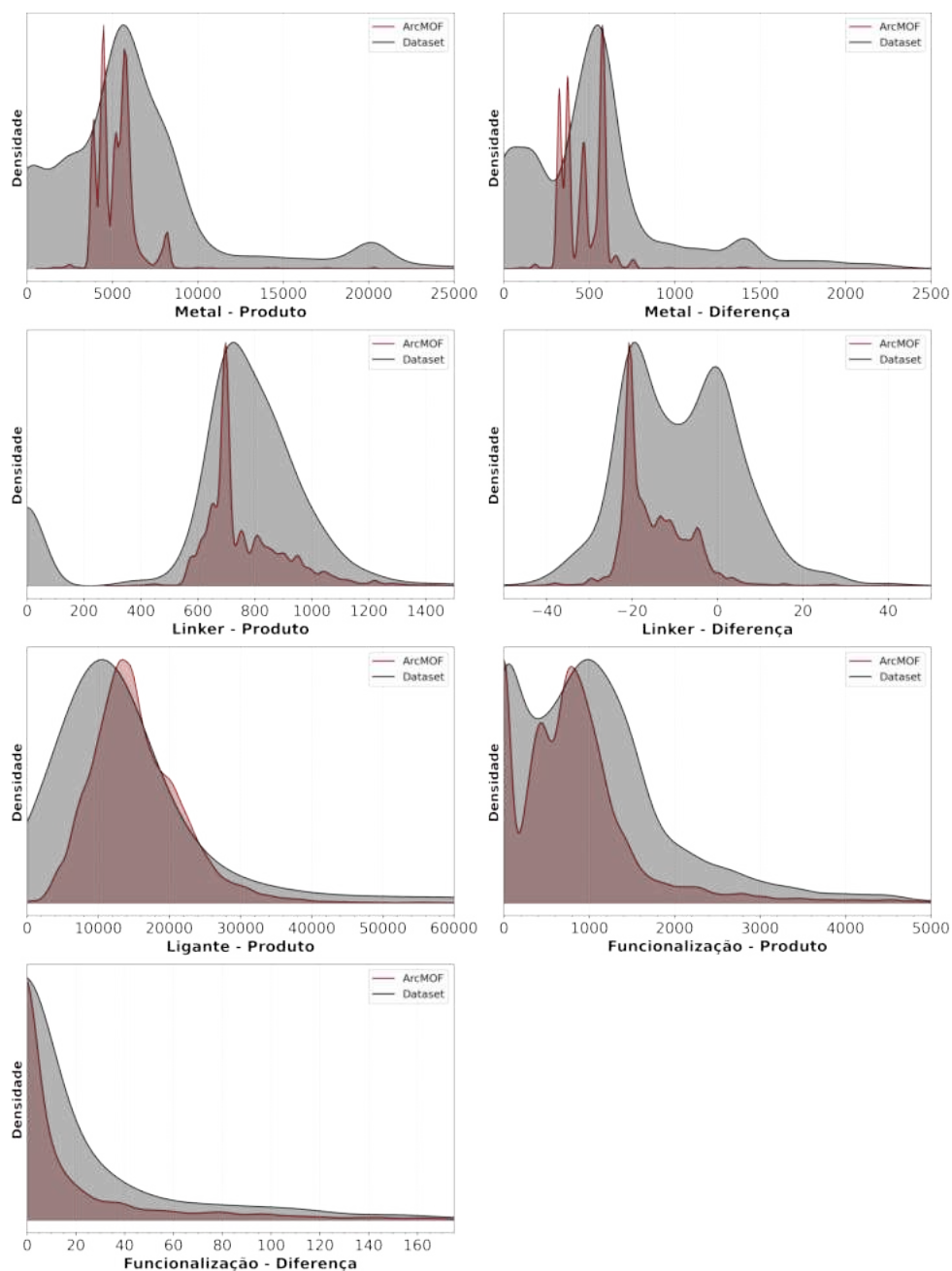


4.4.1 Aplicando Métricas de Desempenho

Os parâmetros gerados pelo modelo Random Forest, aplicados aos materiais do ARC-MOF, possibilitam a construção das isotermas de adsorção de CO_2 e N_2 em diferentes condições de pressão e temperatura. Com base nessas isotermas, métricas de desempenho, como a *working capacity* e a seletividade de CO_2/N_2 , foram calculadas para avaliar a viabilidade dos materiais em processos de captura de carbono. A *working capacity* foi estimada como a diferença na quantidade adsorvida de CO_2 entre pressões típicas de operação, refletindo a eficiência dos materiais em ciclos de adsorção e dessorção. Já a seletividade de CO_2/N_2 foi determinada com base na razão das quantidades adsorvidas dos dois gases, um parâmetro crítico para processos seletivos em misturas gasosas.

Os resultados, apresentados na Figura 29, mostram a relação entre a capacidade de trabalho de CO_2 (nas condições de adsorção de 298 K e variação de pressão de 1 bar para 0,1 bar) e a seletividade de CO_2 (nas condições de adsorção de 298 K e 1 bar) tanto para as estruturas do banco de estruturas usadas no treinamento dos modelos (CURATED +

Figura 28 – Distribuição dos descritores RAC no banco de dados ARC-MOF.

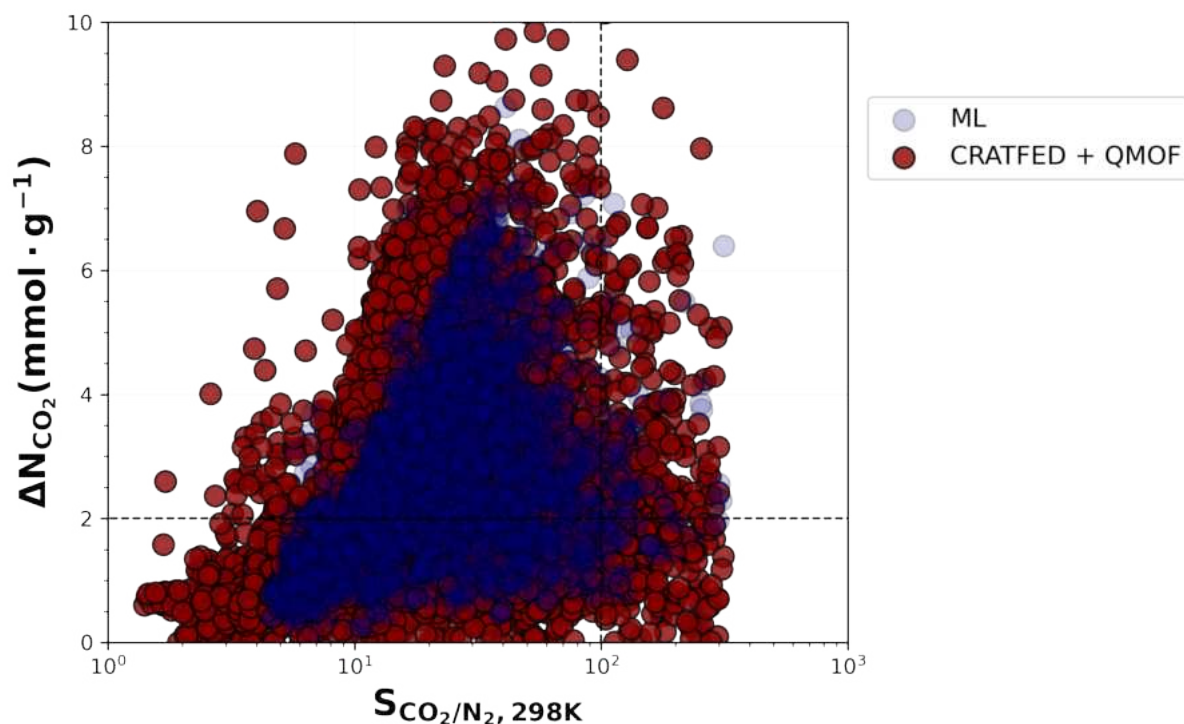


Fonte: Próprio autor.

QMOF) quanto do ARC-MOF. Os pontos acima das linhas tracejadas vertical e horizontal indicam seletividade ≥ 100 e capacidade de trabalho ≥ 2 mmol/g, respectivamente. O triângulo vermelho e o quadrado azul representam, respectivamente, o CALF-20 e o melhor MOF identificado.

Embora o banco de dados ARC-MOF seja significativamente maior que o conjunto de treinamento (CURATED + QMOF), os resultados mostraram uma distribuição de seletividade e capacidade de trabalho bastante restrita para as estruturas do ARC-MOF, em comparação com a distribuição mais ampla observada no conjunto de treinamento. Este

Figura 29 – Capacidade de trabalho de CO_2 (em condições de adsorção de 298 K e oscilação de 1 bar para 0,1 bar) versus seletividade de CO_2 (em condições de adsorção de 298 K e 1 bar) para as estruturas do ARC-MOF (azul) e banco de estruturas usado nesse trabalho (vermelho). Cada ponto corresponde a um único MOF. Os pontos acima das linhas tracejadas verticais e horizontais indicam os MOFs com seletividade ≥ 11 e trabalho $\geq 2 \text{ mmol CO}_2 \cdot \text{g}^{-1}$



Fonte: Próprio autor.

contraste é particularmente preocupante, pois sugere que, apesar de se ter um número maior de estruturas no ARC-MOF, a diversidade estrutural não foi refletida de forma proporcional nos resultados de seletividade e capacidade de trabalho. Esse comportamento indica que, para alcançar melhores predições, uma abordagem de treinamento mais robusta ou um refinamento adicional no conjunto de dados de treinamento, com maior representatividade das variabilidades presentes em bancos de dados maiores, pode ser necessário.

5 CONCLUSÕES E PERSPECTIVAS

5.1 Conclusões

Este trabalho teve como objetivo avaliar a eficácia de diferentes modelos de aprendizado de máquina na previsão de isotermas de adsorção de CO_2 e N_2 em materiais reticulares, buscando entender o impacto da combinação de descritores químicos e físicos no desempenho preditivo. Ao longo do estudo, foi possível observar que o modelo dual-site Langmuir (DSL) demonstrou uma boa adequação para o ajuste dos dados de adsorção simulados.

A investigação dos modelos de aprendizado de máquina permitiu identificar que o Random Forest apresentou a maior capacidade preditiva para os parâmetros do modelo DSL, superando os demais modelos avaliados em termos de precisão e robustez. Contudo, ao testar a capacidade de generalização dos modelos em um outro banco de estruturas, observou-se que a generalização não foi ideal, o que indicou a necessidade de ajustes e refinamentos. No entanto, isso não comprometeu o poder de generalização do Random Forest, que continuou a demonstrar um bom desempenho, embora com variações em diferentes contextos.

Por fim, os resultados obtidos fornecem *insights* para o aprimoramento de estratégias de previsão em simulações de adsorção, demonstrando a relevância do uso de aprendizado de máquina na área de materiais reticulares para a captura de gases como CO_2 , com implicações significativas para o desenvolvimento de novos materiais e tecnologias sustentáveis.

5.2 Perspectivas

A modelagem da adsorção de CO_2 em materiais porosos apresenta diversas limitações intrínsecas que devem ser consideradas na interpretação dos resultados. Os modelos utilizados, como o dual-site Langmuir (DSL), são amplamente empregados devido à sua simplicidade e capacidade de representar com boa precisão a adsorção física. No entanto, esses modelos não contemplam fenômenos como a quimissorção e a presença de sítios metálicos abertos (open-metal sites) em estruturas como nos MOFs. A interação forte entre o CO_2 e esses sítios pode modificar significativamente a capacidade e seletividade da adsorção, exigindo abordagens mais complexas para uma descrição mais fiel do fenômeno.

Além disso, a alta dimensionalidade dos dados gerados por simulações e experimentos representa um desafio para a análise e interpretação dos resultados. A aplicação de técnicas de redução de dimensionalidade, como *Principal Component Analysis* (PCA) e *t-Distributed Stochastic Neighbor Embedding* (t-SNE), pode auxiliar na identificação dos principais fatores que influenciam a adsorção. Essas técnicas permitem visualizar padrões ocultos nos dados, melhorar a interpretabilidade dos modelos de aprendizado de máquina e reduzir a complexidade computacional sem perda significativa de informação relevante.

Acima das discussões sobre captura e armazenamento de CO_2 (CCS, do inglês *Carbon Capture and Storage*), surge um questionamento fundamental: o que fazer com o CO_2 após a sua captura? Nesse contexto, os processos de captura e utilização de CO_2 (CCU, do inglês *Carbon Capture and Utilization*) emergem como uma abordagem promissora. O CCU visa converter o CO_2 capturado em produtos de valor agregado, como combustíveis sintéticos, polímeros e produtos químicos, reduzindo sua liberação na atmosfera e promovendo um ciclo mais sustentável para esse gás.

Apesar do seu potencial, o CCU ainda enfrenta desafios relacionados à viabilidade técnica e econômica, além da necessidade de avaliar os impactos ambientais ao longo do ciclo de vida desses processos. Estudos recentes destacam que, embora o CCU represente um avanço significativo em termos de segurança energética e sustentabilidade ambiental, sua implementação em larga escala requer o desenvolvimento de catalisadores mais eficientes, fontes de energia renovável economicamente viáveis e políticas regulatórias que incentivem sua adoção.

Portanto, a pesquisa em CCS e CCU deve avançar simultaneamente, buscando tanto a otimização da captura quanto a viabilização do uso do CO_2 em aplicações comerciais. A integração de novos materiais adsorventes, modelagem avançada e técnicas de aprendizado de máquina pode contribuir para tornar esses processos mais eficientes e economicamente viáveis, impulsionando soluções sustentáveis para a mitigação das emissões de CO_2 .

Referências

- 1 BOLIN, B. A history of the science and politics of climate change: the role of the intergovernmental panel on climate change. Cambridge University Press, New York, NY (United States), 2007. Citado na página 17.
- 2 HOUGHTON, J. T. et al. *Climate change 2001: the scientific basis*. [S.l.]: Cambridge university press Cambridge, 2001. v. 881. Citado na página 17.
- 3 METZ, B. et al. *IPCC special report on carbon dioxide capture and storage*. [S.l.]: Cambridge: Cambridge University Press, 2005. Citado 4 vezes nas páginas 17, 19, 20 e 21.
- 4 HESTER, R. E.; HARRISON, R. M. *Carbon capture: sequestration and storage*. [S.l.]: Royal Society of Chemistry, 2010. v. 29. Citado na página 17.
- 5 MATIAS, E. F. P. *A humanidade contra as cordas: a luta da sociedade global pela sustentabilidade*. [S.l.]: Editora Paz e Terra, 2014. Citado na página 17.
- 6 RITCHIE, H.; ROSER, M. Co₂ emissions. *Our World in Data*, 2020. <https://ourworldindata.org/co2-emissions>. Citado na página 18.
- 7 MARQUES, V. S. O efeito estufa eo aquecimento global. *Anuário do Instituto de Geociências*, v. 15, p. 93–106, 1992. Citado na página 17.
- 8 YOUNAS, M. et al. Recent progress and remaining challenges in post-combustion co₂ capture using metal-organic frameworks (mofs). *Progress in Energy and Combustion Science*, Elsevier, v. 80, p. 100849, 2020. Citado 2 vezes nas páginas 17 e 21.
- 9 D’ALESSANDRO, D. M.; SMIT, B.; LONG, J. R. Carbon dioxide capture: prospects for new materials. *Angewandte Chemie International Edition*, Wiley Online Library, v. 49, n. 35, p. 6058–6082, 2010. Citado 3 vezes nas páginas 17, 21 e 22.
- 10 FREUND, R. et al. The current status of mof and cof applications. *Angewandte Chemie International Edition*, Wiley Online Library, v. 60, n. 45, p. 23975–24001, 2021. Citado na página 18.
- 11 FARMAHINI, A. H. et al. Performance-based screening of porous materials for carbon capture. *Chemical Reviews*, ACS Publications, v. 121, n. 17, p. 10666–10741, 2021. Citado 4 vezes nas páginas 18, 26, 27 e 39.
- 12 ORHAN, I. B. et al. Accelerating the prediction of co₂ capture at low partial pressures in metal-organic frameworks using new machine learning descriptors. *Communications chemistry*, Nature Publishing Group UK London, v. 6, n. 1, p. 214, 2023. Citado na página 19.
- 13 WANG, M.; OKO, E. *Special issue on carbon capture in the context of carbon capture, utilisation and storage (CCUS)*. [S.l.]: Springer, 2017. 1–4 p. Citado na página 19.
- 14 CUÉLLAR-FRANCA, R. M.; AZAPAGIC, A. Carbon capture, storage and utilisation technologies: A critical analysis and comparison of their life cycle environmental impacts. *Journal of CO₂ utilization*, Elsevier, v. 9, p. 82–102, 2015. Citado na página 19.

- 15 BUI, M. 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. Citado 2 vezes nas páginas 19 e 21.
- 16 SMIT, B. et al. *Introduction to carbon capture and sequestration*. [S.l.]: World Scientific, 2014. v. 1. Citado 2 vezes nas páginas 20 e 21.
- 17 NASCIMENTO, R. F. d. et al. Adsorção: aspectos teóricos e aplicações ambientais. 2020. Citado na página 21.
- 18 MYERS, D. et al. *Surfaces, interfaces, and colloids*. [S.l.]: Wiley New York, 1999. v. 415. Citado na página 21.
- 19 ROUQUEROL, J. et al. *Adsorption by powders and porous solids: principles, methodology and applications*. [S.l.]: Academic press, 2013. Citado 2 vezes nas páginas 21 e 24.
- 20 KONG, L.; ADIDHARMA, H. A new adsorption model based on generalized van der waals partition function for the description of all types of adsorption isotherms. *Chemical Engineering Journal*, Elsevier, v. 375, p. 122112, 2019. Citado na página 22.
- 21 ALOTHMAN, Z. A. A review: fundamental aspects of silicate mesoporous materials. *Materials*, MDPI, v. 5, n. 12, p. 2874–2902, 2012. Citado na página 22.
- 22 THOMMES, M. et al. Physisorption of gases, with special reference to the evaluation of surface area and pore size distribution (iupac technical report). *Pure and applied chemistry*, De Gruyter, v. 87, n. 9-10, p. 1051–1069, 2015. Citado na página 22.
- 23 BRUNAUER, S. et al. On a theory of the van der waals adsorption of gases. *Journal of the American Chemical society*, ACS Publications, v. 62, n. 7, p. 1723–1732, 1940. Citado na página 22.
- 24 FIRLEJ, L. et al. How dense is the gas adsorbed in nanopores? *Microporous and Mesoporous Materials*, Elsevier, v. 304, p. 109240, 2020. Citado na página 24.
- 25 SING, K. S. Reporting physisorption data for gas/solid systems with special reference to the determination of surface area and porosity (recommendations 1984). *Pure and applied chemistry*, De Gruyter, v. 57, n. 4, p. 603–619, 1985. Citado na página 24.
- 26 STRIOLO, A. et al. Water adsorption in carbon-slit nanopores. *Langmuir*, ACS Publications, v. 19, n. 20, p. 8583–8591, 2003. Citado na página 25.
- 27 WANG, J. et al. Recent advances in solid sorbents for co₂ capture and new development trends. *Energy & Environmental Science*, Royal Society of Chemistry, v. 7, n. 11, p. 3478–3518, 2014. Citado na página 25.
- 28 CHOI, S.; DRESE, J. H.; JONES, C. W. Adsorbent materials for carbon dioxide capture from large anthropogenic point sources. *ChemSusChem: Chemistry & Sustainability Energy & Materials*, Wiley Online Library, v. 2, n. 9, p. 796–854, 2009. Citado na página 25.
- 29 HARLICK, P. J.; SAYARI, A. Applications of pore-expanded mesoporous silicas. 3. triamine silane grafting for enhanced co₂ adsorption. *Industrial & Engineering Chemistry Research*, ACS Publications, v. 45, n. 9, p. 3248–3255, 2006. Citado na página 25.

- 30 HÖNICKE, I. M. et al. Balancing mechanical stability and ultrahigh porosity in crystalline framework materials. *Angewandte Chemie International Edition*, v. 57, n. 42, p. 13780–13783, 2018. Citado na página 25.
- 31 YAGHI, O. M.; KALMUTZKI, M. J.; DIERCKS, C. S. *Introduction to reticular chemistry: metal-organic frameworks and covalent organic frameworks*. [S.l.]: John Wiley & Sons, 2019. Citado na página 26.
- 32 YAGHI, O. M. *Reticular chemistry in all dimensions*. [S.l.]: ACS Publications, 2019. 1295–1300 p. Citado 2 vezes nas páginas 26 e 27.
- 33 KITAGAWA, S. et al. Metal–organic frameworks (mofs). *Chemical Society Reviews*, Royal Society of Chemistry, v. 43, n. 16, p. 5415–5418, 2014. Citado na página 26.
- 34 OLIVEIRA, F. L. et al. Enzyme immobilization in covalent organic frameworks: Strategies and applications in biocatalysis. *ChemPlusChem*, Wiley Online Library, v. 85, n. 9, p. 2051–2066, 2020. Citado na página 26.
- 35 LI, A. et al. Enabling efficient exploration of metal–organic frameworks in the cambridge structural database. *CrystEngComm*, Royal Society of Chemistry, v. 22, n. 43, p. 7152–7161, 2020. Citado na página 26.
- 36 LYU, H. et al. Digital reticular chemistry. *Chem*, Elsevier, v. 6, n. 9, p. 2219–2241, 2020. Citado 2 vezes nas páginas 26 e 27.
- 37 HAASE, F. et al. Beyond frameworks: Structuring reticular materials across nano-, meso-, and bulk regimes. *Angewandte Chemie International Edition*, Wiley Online Library, v. 59, n. 50, p. 22350–22370, 2020. Citado na página 26.
- 38 SCHLEDER, G. R. et al. Ab initio simulations and materials chemistry in the age of big data. *Journal of chemical information and modeling*, ACS Publications, v. 60, n. 2, p. 452–459, 2019. Citado 2 vezes nas páginas 27 e 28.
- 39 REN, E.; COUDERT, F.-X. Enhancing gas separation selectivity prediction through novel descriptors. 2023. Citado na página 28.
- 40 TSAMARDINOS, I. et al. An automated machine learning architecture for the accelerated prediction of metal-organic frameworks performance in energy and environmental applications. *Microporous and Mesoporous Materials*, Elsevier, v. 300, p. 110160, 2020. Citado na página 28.
- 41 NAQA, I. E.; MURPHY, M. J. *What is machine learning?* [S.l.]: Springer, 2015. Citado na página 29.
- 42 SCHLEDER, G. R. et al. From dft to machine learning: recent approaches to materials science—a review. *Journal of Physics: Materials*, IOP Publishing, v. 2, n. 3, p. 032001, 2019. Citado na página 29.
- 43 YANG, P. et al. Accelerating the selection of covalent organic frameworks with automated machine learning. *ACS omega*, ACS Publications, v. 6, n. 27, p. 17149–17161, 2021. Citado na página 29.

- 44 SCHLEDER, G. R.; FAZZIO, A. Machine learning na física, química, e ciência de materiais: Descoberta e design de materiais. *Revista Brasileira de Ensino de Física*, SciELO Brasil, v. 43, n. Supl 1, p. e20200407, 2021. Citado na página 29.
- 45 BUDACH, L. et al. The effects of data quality on machine learning performance. *arXiv preprint arXiv:2207.14529*, 2022. Citado 2 vezes nas páginas 29 e 30.
- 46 MAI, H. et al. Machine learning in the development of adsorbents for clean energy application and greenhouse gas capture. *Advanced Science*, Wiley Online Library, v. 9, n. 36, p. 2203899, 2022. Citado na página 29.
- 47 MOGHADAM, P. Z. et al. Structure-mechanical stability relations of metal-organic frameworks via machine learning. *Matter*, Elsevier, v. 1, n. 1, p. 219–234, 2019. Citado na página 29.
- 48 AHM, N. A. Machine learning approaches for evaluating the properties of materials. *Journal of Computational Intelligence in Materials Science*, AnaPub Publications, v. 1, p. 067–076, 2023. Citado na página 29.
- 49 DELPISHEH, M. et al. Leveraging machine learning in porous media. *Journal of Materials Chemistry A*, Royal Society of Chemistry, 2024. Citado na página 29.
- 50 BENITEZ, J. G. *Benchmarking Machine Learning Algorithms for the Prediction of Gas Adsorption in Metal-Organic Framework Materials*. Tese (Doutorado) — The University of Texas Rio Grande Valley, 2024. Citado na página 29.
- 51 OLIVEIRA, F. L. et al. Crafted: An exploratory database of simulated adsorption isotherms of metal-organic frameworks. *Scientific Data*, Nature Publishing Group UK London, v. 10, n. 1, p. 230, 2023. Citado 2 vezes nas páginas 32 e 54.
- 52 ROSEN, A. S. et al. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter*, Elsevier, v. 4, n. 5, p. 1578–1597, 2021. Citado na página 32.
- 53 WATANABE, T.; SHOLL, D. S. Accelerating applications of metal-organic frameworks for gas adsorption and separation by computational screening of materials. *Langmuir*, ACS Publications, v. 28, n. 40, p. 14114–14128, 2012. Citado na página 32.
- 54 WITMAN, M. et al. The influence of intrinsic framework flexibility on adsorption in nanoporous materials. *Journal of the American Chemical Society*, ACS Publications, v. 139, n. 15, p. 5547–5557, 2017. Citado na página 32.
- 55 ZHANG, J.-P. et al. Controlling flexibility of metal-organic frameworks. *National Science Review*, Oxford University Press, v. 5, n. 6, p. 907–919, 2018. Citado na página 32.
- 56 FRENKEL, D.; SMIT, B. *Understanding molecular simulation: from algorithms to applications*. [S.l.]: Elsevier, 2023. Citado na página 33.
- 57 SCHERER, C. *Métodos computacionais da Física*. [S.l.]: Editora Livraria da Física, 2005. Citado 2 vezes nas páginas 33 e 34.
- 58 LANDAU, D.; BINDER, K. *A guide to Monte Carlo simulations in statistical physics*. [S.l.]: Cambridge university press, 2021. Citado na página 34.

- 59 DUBBELDAM, D. et al. Raspa: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Molecular Simulation*, Taylor & Francis, v. 42, n. 2, p. 81–101, 2016. Citado na página 34.
- 60 BRANDANI, S.; MANGANO, E.; SARKISOV, L. Net, excess and absolute adsorption and adsorption of helium. *Adsorption*, Springer, v. 22, p. 261–276, 2016. Citado na página 36.
- 61 OLIVEIRA, F. L. et al. pymser—an open-source library for automatic equilibration detection in molecular simulations. *Journal of Chemical Theory and Computation*, 2024. Citado na página 38.
- 62 RAPPÉ, A. K. et al. 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. Citado na página 39.
- 63 POTOFF, J. J.; SIEPMANN, J. I. Vapor–liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AIChE journal*, Wiley Online Library, v. 47, n. 7, p. 1676–1682, 2001. Citado na página 39.
- 64 CLEETON, C. et al. A process-level perspective of the impact of molecular force fields on the computational screening of mofs for carbon capture. *Energy & Environmental Science*, Royal Society of Chemistry, v. 16, n. 9, p. 3899–3918, 2023. Citado na página 39.
- 65 LIMAS, N. G.; MANZ, T. A. Introducing ddec6 atomic population analysis: part 2. computed results for a wide range of periodic and nonperiodic materials. *RSC advances*, Royal Society of Chemistry, v. 6, n. 51, p. 45727–45747, 2016. Citado na página 39.
- 66 FOO, K. Y.; HAMEED, B. H. Insights into the modeling of adsorption isotherm systems. *Chemical engineering journal*, Elsevier, v. 156, n. 1, p. 2–10, 2010. Citado na página 39.
- 67 LANGMUIR, I. The adsorption of gases on plane surfaces of glass, mica and platinum. *Journal of the American Chemical society*, ACS Publications, v. 40, n. 9, p. 1361–1403, 1918. Citado na página 40.
- 68 SWENSON, H.; STADIE, N. P. Langmuir’s theory of adsorption: A centennial review. *Langmuir*, ACS Publications, v. 35, n. 16, p. 5409–5426, 2019. Citado 2 vezes nas páginas 42 e 43.
- 69 NALAPARAJU, A. et al. Co2 capture in cation-exchanged metal–organic frameworks: Holistic modeling from molecular simulation to process optimization. *Chemical Engineering Science*, Elsevier, v. 124, p. 70–78, 2015. Citado na página 43.
- 70 MYERS, A. Activity coefficients of mixtures adsorbed on heterogeneous surfaces. *AIChE journal*, Wiley Online Library, v. 29, n. 4, p. 691–693, 1983. Citado na página 43.
- 71 GOMMERS, R. et al. scipy/scipy: Scipy 1.9.0rc2. *Zenodo*, 2022. Citado na página 44.
- 72 ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado na página 44.

- 73 ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado na página 47.
- 74 NASTESKI, V. An overview of the supervised machine learning methods. *Horizons*, v. 4, n. 51-62, p. 56, 2017. Citado na página 47.
- 75 USAMA, M. et al. Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, IEEE, v. 7, p. 65579–65615, 2019. Citado na página 48.
- 76 QUINLAN, J. R. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 28, n. 1, p. 71–72, 1996. Citado na página 48.
- 77 SUTHAHARAN, S.; SUTHAHARAN, S. Decision tree learning. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, Springer, p. 237–269, 2016. Citado na página 48.
- 78 RIGATTI, S. J. Random forest. *Journal of Insurance Medicine*, American Academy of Insurance Medicine 1700 Magnavox Way, Fort Wayne, IN 46804, v. 47, n. 1, p. 31–39, 2017. Citado na página 49.
- 79 BIAU, G.; SCORNET, E. A random forest guided tour. *Test*, Springer, v. 25, p. 197–227, 2016. Citado na página 49.
- 80 FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000. Citado na página 49.
- 81 WILLEMS, T. F. et al. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, Elsevier, v. 149, n. 1, p. 134–141, 2012. Citado na página 51.
- 82 BROTO, P.; MOREAU, G.; VANDYCKE, C. Molecular structures: perception, autocorrelation descriptor and sar studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *European journal of medicinal chemistry*, v. 19, n. 1, p. 71–78, 1984. Citado na página 51.
- 83 MOOSAVI, S. M. et al. Understanding the diversity of the metal-organic framework ecosystem. *Nature communications*, Nature Publishing Group, v. 11, n. 1, p. 1–10, 2020. Citado na página 51.
- 84 LIU, F.; DUAN, C.; KULIK, H. New tools for detecting strong correlation in automated transition metal complex screening. *Bulletin of the American Physical Society*, APS, v. 65, 2020. Citado na página 51.
- 85 JANET, J. P.; KULIK, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chemical science*, Royal Society of Chemistry, v. 8, n. 7, p. 5137–5152, 2017. Citado na página 51.
- 86 MOOSAVI, S. M. et al. Understanding the diversity of the metal-organic framework ecosystem. *Nature communications*, Nature Publishing Group, v. 11, n. 1, p. 1–10, 2020. Citado na página 52.

- 87 ROSEN, A. S. et al. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter*, Elsevier, v. 4, n. 5, p. 1578–1597, 2021. Citado na página 54.
- 88 GROOM, C. R. et al. The cambridge structural database. *Structural Science*, International Union of Crystallography, v. 72, n. 2, p. 171–179, 2016. Citado na página 54.
- 89 CHUNG, Y. 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. Citado na página 54.
- 90 KRESSE, G.; FURTHMÜLLER, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review Letters*, v. 77, n. 20, p. 3865–3868, 1996. Citado na página 54.
- 91 HERMANN, J. et al. A consistent and accurate ab initio calculation of molecular properties using the density functional theory and hybrid density functionals. *The Journal of Chemical Physics*, v. 132, n. 15, p. 154104, 2010. Citado na página 54.
- 92 GRIMME, S.; EHRLICH, S.; GOERIGK, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, v. 32, n. 9, p. 1456–1465, 2011. Citado na página 54.
- 93 CHUNG, Y. 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. Citado na página 54.
- 94 ONGARI, D. et al. Building a consistent and reproducible database for adsorption evaluation in covalent-organic frameworks. *ACS central science*, ACS Publications, v. 5, n. 10, p. 1663–1675, 2019. Citado na página 54.
- 95 EL-KHAIARY, M. I.; MALASH, G. F.; HO, Y.-S. On the use of linearized pseudo-second-order kinetic equations for modeling adsorption systems. *Desalination*, Elsevier, v. 257, n. 1-3, p. 93–101, 2010. Citado na página 57.
- 96 HUBBE, M. A. Insisting upon meaningful results from adsorption experiments. *Separation & Purification Reviews*, Taylor & Francis, v. 51, n. 2, p. 212–225, 2022. Citado na página 57.
- 97 MOUSSOUT, H. et al. Critical of linear and nonlinear equations of pseudo-first order and pseudo-second order kinetic models. *Karbala International Journal of Modern Science*, Elsevier, v. 4, n. 2, p. 244–254, 2018. Citado na página 57.
- 98 ALEXANDER, S. T.; ALEXANDER, S. T. The mean square error (mse) performance criteria. *Adaptive Signal Processing: Theory and Applications*, Springer, p. 8–33, 1986. Citado na página 66.
- 99 IOFFE, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. Citado na página 66.

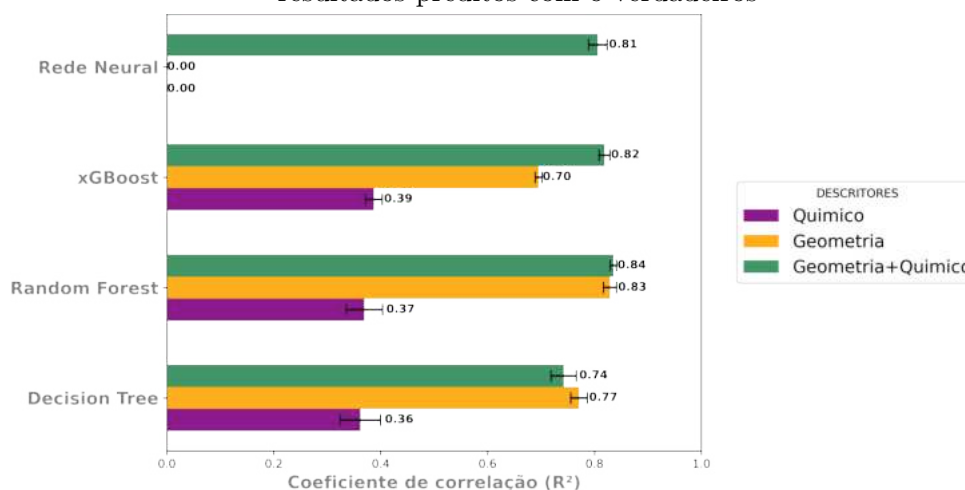
-
- 100 BURNER, J. et al. Arc-mof: a diverse database of metal-organic frameworks with dft-derived partial atomic charges and descriptors for machine learning. *Chemistry of Materials*, ACS Publications, v. 35, n. 3, p. 900–916, 2023. Citado na página 72.
- 101 WHITE, A. et al. Alarming structural error rates in mof databases used in data driven workflows identified via a novel metal oxidation state-based method. 2024. Citado 2 vezes nas páginas 72 e 73.

Apêndices

APÊNDICE A – Dados Modelos de Aprendizizado de Máquina

A.1 Modelos de Aprendizizado de Máquina - Dados sem escala Logarítmica

(a) Comparação dos valores de R^2 entre os modelos de aprendizado de máquina. Quanto mais próximo de 1, melhor é a correlação dos resultados preditos com o verdadeiros



(b) Comparação do erro quadrático médio (MSE) entre os modelos de aprendizado de máquina. Quanto mais próximo de 0, menor é o desvio dos resultados preditos com o verdadeiros

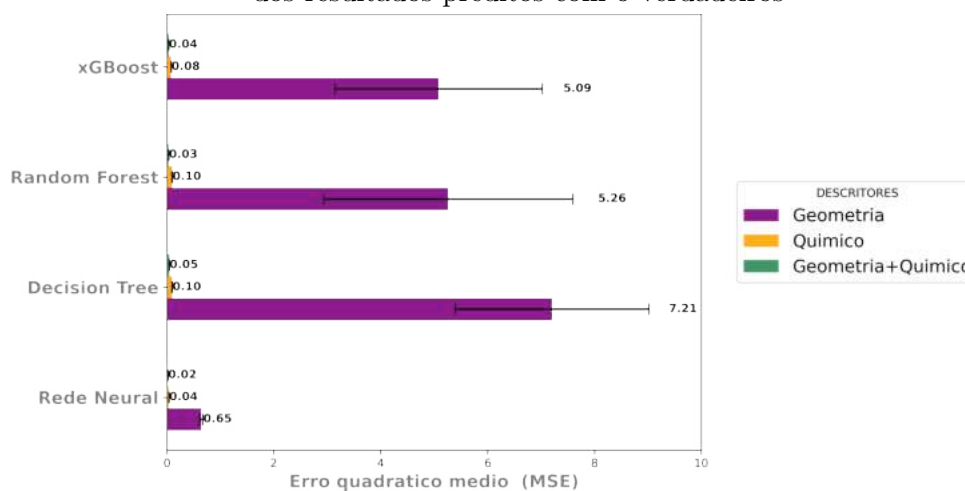


Figura 30 – Comparação dos desempenhos dos modelos de aprendizado de máquina nas métricas R^2 e MSE.

O comportamento dos modelos após a normalização logarítmica indica que esta etapa é crucial para otimizar a capacidade preditiva dos modelos, especialmente no contexto de dados com grande variabilidade nos parâmetros de adsorção.

Anexos

ANEXO A – Tabelas e Dados

Tabela 1 – Parâmetros críticos for CO₂ e N₂.

Gás	Temperatura Crítica (K)	Pressão Crítica (Pa)	Fator Acêntrico
CO ₂	304.12	7.3773×10^6	0.22394
N ₂	126.19	3.3958×10^6	0.0372

Tabela 2 – Parâmetros de Lennard-Jones para os campos de força UFF e DREIDING.

Tipo de Átomo	UFF		DREIDING	
	σ (Å)	ϵ (K)	σ (Å)	ϵ (K)
H	2.571	22.14	2.8464	27.64936
B	3.6389	0.51	3.5814	147.80852
C	3.4315	2.8	3.47299	47.85884
N	3.261	4.7	3.26256	38.95136
O	3.118	30.2	3.03315	48.16079
F	2.997	25.14	3.09323	36.48545
Cl	3.516	114.15	3.51932	142.57003
Br	3.732	126.22	3.51905	186.20159
Al	4.008	253.94	3.91105	156.00674
Si	3.826	202.15	3.80414	156.00674
P	3.695	153.37	3.69723	161.03921
S	3.595	137.78	3.59032	173.11715
Ga	3.905	208.69	3.91105	201.29901
Ge	3.813	190.58	3.80414	201.29901
As	3.769	155.38	3.69723	206.33149
Se	3.746	146.33	3.59032	216.39644
In	3.976	301.21	4.08923	276.78614
Sn	3.913	285.12	3.98232	276.78614
Sb	3.938	225.78	3.87541	276.78614
Te	3.982	200.14	3.76852	286.85109
Na	2.658	15.09	2.80099	251.62377
Ca	3.028	119.68	3.09322	5.16238
Fe	2.594	6.54	4.04468	27.67861
Zn	2.462	62.38	4.04468	27.67861
Ti	2.829	8.55	4.04468	27.67861
Tc	2.671	24.14	4.04468	27.67861
Ru	2.642	8.16	4.04468	27.67861

Tabela 3 – Bases de Dados de Materiais Porosos Cristalinos

Base de Dados	Número de Entradas	Origem	Limpo	Otimizado	Cargas Incluídas
Wilmer-et-al.	100	Simulação	Sim	Sim	Não
Boyd-and-Woo	157	Simulação	Sim	Sim	Não
ToBaCCo	152,160	Simulação	Sim	Sim	Não
CSD	>1M	Experimental	Não	Não	Parcialmente
Goldsmith et al.	4000	Experimental	Sim	Sim	Sim
CoRE-MOF-2019	14 000	Experimental	Sim	Sim	Sim
CSD-MOF-subsetc	96 000	Experimental	Sim	Sim	Sim
hZeo	2.6M	Simulação	Sim	Sim	Não
IZA	253	Experimental	Sim	Sim	Não
hPPN	18 000	Simulação	Sim	Sim	Sim
hCOF	69 840	Simulação	Sim	Sim	Sim
CoRE-COF	449	Experimental	Sim	Sim	Não
CURATED-COFs	482	Experimental	Sim	Sim	Sim
hZIFs	86	Simulação	Sim	Sim	Sim
NE-DB2	530 243	Simulação e Experimental	Não	Não	Não
NMG	>3M	Simulação e Experimental	Sim	Sim	Sim
PRAM-DB	205	Simulação e Experimental	Não	Não	Não

A.1 Descritores Geométricos

Descritores geométricos utilizados e suas definições:

- **Número de Canais:**

- Número total de canais na estrutura.

- **Dimensionalidade dos Canais:**

- Indica a dimensão dos canais.

- **Tamanho dos Canais:**

- Dimensões dos canais na estrutura, geralmente representadas em Angstroms.

- **Volume da Célula Unitária:**

- Volume total da célula unitária da estrutura cristalina.

- **Densidade:**

- Densidade do material, que indica a massa por unidade de volume.

- **Fração de Volume Acessível (AV):**

- Proporção do volume total da estrutura que é acessível para adsorção.

- **Volume Acessível Específico (AV):**

- Volume acessível por unidade de massa do material.

- **Fração de Volume Não Acessível (NAV):**

- Proporção do volume total da estrutura que não é acessível para adsorção.
- **Volume Não Acessível Específico (NAV):**
 - Volume não acessível por unidade de massa do material.
- **Número de Cavidades (Pockets):**
 - Número de cavidades ou espaços internos dentro da estrutura.
- **Área Superficial Específica (ASA):**
 - Área superficial total acessível do material por unidade de volume.
- **Área Superficial Específica por Massa (ASA):**
 - Área superficial total acessível do material por unidade de massa.
- **Área Superficial Não Acessível (NASA):**
 - Área superficial que não é acessível para adsorção, por unidade de volume.
- **Área Superficial Não Acessível por Massa (NASA):**
 - Área superficial que não é acessível para adsorção, por unidade de massa.
- **Diâmetro Interno (D_{is}):**
 - Diâmetro dos canais medidos internamente.
- **Diâmetro de Fendas (D_{fs}):**
 - Diâmetro das fendas ou poros na estrutura.
- **Diâmetro de Interseção (D_{isfs}):**
 - Diâmetro das interseções entre canais ou poros.