UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

Luiz Eduardo Foschiera Couto Lopes

Aprendizado de máquina para predição de parâmetros de reações químicas

Campinas
20/06/2024

Luiz Eduardo Foschiera Couto Lopes

Aprendizado de máquina para predição de parâmetros de reações químicas*

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Carlos Eduardo Driemeier.

*Este trabalho foi financiado pela CNPQ, projeto 128616/2024.

Resumo

Redes de reações químicas associadas a métodos de aprendizado de máquina têm sido propostas para modelar meios reacionais complexos. Neste projeto, essa abordagem é empregada para fragmentos moleculares resultantes da despolimerização de lignina. Em particular, o projeto desenvolve a técnica de aprendizado por transferência, em que uma rede neural artificial é inicialmente pré-treinada com uma base de dados de reações fonte e, posteriormente, aprimorada empregando uma base de dados alvo específica para fragmentos de lignina. As técnicas de aprendizado de máquina são implementadas em linguagem Python com o pacote TensorFlow. As primeiras versões do modelo seguem uma estrutura de três camadas escondidas de 250 neurônios em cada camada e uma camada final com 50 neurônios, resultando em um Erro Absoluto Médio (MAE) de 6.28 kcal/mol, Raiz do Erro Quadrático Médio (RMSE) de 9.42 kcal/mol e R^2 de 0.95 para a predição de entalpia das reações da base fonte. Para essa mesma base de dados, o modelo resulta em MAE de 9.20 kcal/mol, RMSE de 13.67 kcal/mol e R^2 de 0.79 para a previsão da energia de ativação das reações.O modelo treinado na base de dados fonte demonstrou transferência para a predição de entalpia de reações químicas de lignina, alcançando MAE de 6.15 kcal/mol , RMSE de 6.86 kcal/mol e R^2 de 0.92, dadas somente 4 reações de treino e 50 parâmetros livres treináveis para ajuste.

Abstract

Chemical reaction networks associated with machine learning methods have been proposed to model complex reaction media. In this project, this approach is employed for molecular fragments resulting from lignin depolymerization. In particular, the project develops the transfer learning technique, where an artificial neural network is initially pre-trained with a source reaction database and subsequently fine-tuned using a target database specific to lignin fragments. The machine learning techniques are implemented in Python using the TensorFlow package. The initial versions of the model follow a structure of three hidden layers with 250 neurons in each layer and a final layers with 5 neurons, resulting in a Mean Absolute Error (MAE) of 6.28 kcal/mol, Root Mean Square Error (RMSE) of 9.42 kcal/mol, and R^2 of 0.95 for the prediction of reaction enthalpy from the source database. For this same database, the model results in an MAE of 9.20 kcal/mol, RMSE of 13.67 kcal/mol, and R^2 of 0.78 for the prediction of reaction activation energy.

The model trained on the source database demonstrated transferability to predict the enthalpy of lignin chemical reactions, achieving an MAE of 8.13 kcal/mol, RMSE of 9.23 kcal/mol, and R^2 of 0.86, given only 4 fitting reactions and 5 trainable parameters for adjustment.

Conteúdo

1 Introdução

Aprendizado de Máquina é uma área de estudo que foca em desenvolver e estudar algoritmos que aprendem a partir de grandes volumes de dados [Mitchell [1997]]. Entre os tipos de algoritmos empregados para aprendizado de máquina, as Redes Neurais Artificiais (RNAs) destacam-se por sua robustez e eficácia.

As RNAs são uma técnica robusta de aprendizado de máquina utilizada para a aproximação de funções reais, vetoriais e discretas, inspirada no funcionamento biológico de neurônios e no modo como se comportam em presença de um estímulo [Mitchell [1997]]. Essa técnica tem se mostrado particularmente promissora na análise de reações químicas e suas propriedades, pois permite codificar construções químicas em vetores numéricos que podem ser processados pelos modelos de aprendizagem.

Aplicar RNAs no estudo de reações químicas é vantajoso, pois além de captar informações já contidas nas reações, a técnica pode extrapolar essas informações para predições, como a energia de ativação, a entalpia de formação ou de reação, entre outras propriedades químicas. A representação computacional de compostos químicos pode ser feita de várias formas, como SMILES (Simplified Molecular-Input Entry-Line System) ou formatos de arquivo .mol ou .XYZ. Além disso, há um extenso trabalho na conversão de compostos e reações químicas em vetores numéricos, chamados descritores moleculares, que são usados como entradas para os algoritmos de aprendizado de máquina [Lauri Himanen [2020]].

No entanto, as técnicas de aprendizado de máquina, incluindo RNAs, geralmente requerem grandes volumes de dados para o treinamento eficaz dos modelos. Para muitas aplicações, esse volume de dados não está disponível, o que limita o seu desenvolvimento. Nessas situações, técnicas de Aprendizado por Transferência podem ser empregadas. O modelo a ser treinado é inicialmente pré-treinado em uma base de dados chamada base fonte, distinta da base alvo. Em seguida, a última camada do modelo é treinada e adaptada aos novos dados alvo, enquanto as outras camadas mantêm o treinamento realizado na base fonte. Essa abordagem parte da hipótese de que o modelo treinado na base fonte pode generalizar para a base alvo e atingir uma performance superior a um modelo treinado apenas na base alvo [Pan and Yang [2009]].

Um exemplo prático dessa aplicação é o estudo da lignina, que compõe cerca de 20–30% da biomassa lignocelulósica, a fonte de carbono renovável mais abundante na Terra. Atualmente, a lignina é utilizada principalmente como combustível de caldeiras, um uso de baixo valor agregado. No entanto, há grande potencial para a utilização da lignina na produção de biocombustível líquido e bioquímicos aromáticos [Schutyser W [2018]]. A complexidade dos meios reacionais nos processos de despolimerização da lignina, devido à diversidade de compostos e reações químicas que ocorrem paralela e sequencialmente, é uma das principais barreiras nesse desenvolvimento.

Problemas semelhantes têm sido abordados usando redes de reações químicas associadas a métodos de aprendizado de máquina, como RNAs [Stocker S [2020]], mas nunca aplicados a bases de dados de reações de lignina. Este projeto visa explorar a aplicação das RNAs na previsão da entalpia e da energia de ativação de reações químicas associadas à lignina. Para isso, são empregadas técnicas de Aprendizado por Transferência, com o objetivo de avançar o conhecimento na área e identificar possíveis desafios para pesquisas futuras.

2 Metodologia

2.1 Conjunto de Dados

Os dados de reações químicas produzidos por Grambow et al. [Grambow [2020]] por meio de cálculos DFT - Density Functional Theory foram empregados para pré-treinamento da RNA. Essa base de dados, a qual denominaremos base de dados fonte, dispõe de 11961 reações químicas envolvendo compostos formados por carbono, hidrogênio, nitrogênio e oxigênio.

A biblioteca Pandas foi empregada para a limpeza dos dados, gerando uma base de dados livre de reações nitrogenadas e/ou iônicas, o que garantiu 4323 reações dentro do escopo de análise. Primeiramente, cada reagente e produto das reações foi convertido do formato SMILES canônicas para o formato .mol, o qual foi convertido para descritores SOAP (Smooth Overlap of Atomic Positions) utilizando a biblioteca DScribe [Lauri Himanen [2020]].

Nesse ponto, a base de dados gerada consiste nos descritores moleculares de

reagentes e produtos das reações químicas com suas respectivas diferenças de entalpia e energias de ativação. É necessário que sejam gerados vetores para cada reação. Portanto, para cada reação, foi gerado um vetor resultante reagente e um vetor resultante produto. Os vetores resultantes foram construídos da seguinte forma: para cada reação foi construído um vetor resultante reagente e um vetor resultante produto, dados por: considere B o número de compostos no reagente (sem perda de generalidade) e A o número de átomos em um dado composto K e J a dimensão do descritor:

$$p^i = \sum_{b=0}^{B} \sum_{j=0}^{J} \sum_{a=0}^{A} K_{bja} \tag{1}$$

Isto é, é computada a soma por colunas de cada composto. Desse modo, o vetor resultante da reação i, R_i, é dado por:

$$R_i = p^i_{produtos} - p^i_{reagentes} \tag{2}$$

onde p representa o vetor descritor molecular correspondente.

Além disso, a técnica de Aumento de dados (Data Augmentation) foi empregada adicionando as reações inversas da base de dados para duplicar o número de reações. O processo de aumento consistiu na multiplicação de cada descritor e diferença de entalpia por (-1) e as novas energias de ativação foram calculadas pela seguinte expressão:

$$EA_{novo} = EA_{inicial} - DH_{incial} \tag{3}$$

Esse passo garantiu, pois, a base de dados fonte, composta por 8646 vetores descritores de reações químicas e suas respectivas propriedades. Além da base de dados fonte, foi necessário construir e manipular a base de dados de reações de lignina , a qual denominaremos base de dados alvo. Essas reações foram investigadas e parametrizadas a partir de cálculos DFT realizados pela equipe do LNBR/CNPEM. O formato original das reações de lignina consistia em arquivos do tipo .XYZ dos reagentes, estados de transição e produtos de cada reação e suas respectivas energias de ativação e entalpia de reação. A partir da conversão dos arquivos de .XYZ para formato SMILES canônicas, utilizando o conversor OpenBabel, os descritores moleculares de cada reação foram calculados e organizados na base de dados alvo final.

A técnica de Aumento de Dados anteriormente descrita também foi aplicada à base de dados alvo. Após a filtragem de reações não factíveis, resultou-se uma base de dados alvo com 16 reações de lignina e suas respectivas propriedades.

2.2 Descritores Moleculares

Descritores moleculares foram calculados com o método SOAP (Smooth Overlap of Atomic Positions) [Lauri Himanen [2020]], que codifica regiões de geometria atômica por meio da expansão local de funções de densidade atômica gaussianas com funções ortonormais baseadas em esféricos harmônicos e funções de base radial. Ele possui algumas parâmetros que definem a dimensão do vetor final, a saber: r_{cut} (corte de região local) ; n_{max} (número de funções de base radial); l_{max} (o máximo grau dos harmônicos esféricos) e; rbf (o tipo de base radial – gto ou polynomial); entre outros. Este projeto utiliza r_{cut} = 6, n_{max} = 4, l_{max} = 4 e rbf = 'gto', gerando um vetor de 390 entradas para cada composto químico.

2.3 Aplicação de Redes Neurais Artificiais

Em posse da base dados fonte, essa foi repartida em dados de treino (90%), dados de teste (5%) e dados de validação (5%). Para a configuração do modelo, após diversos testes, foi utilizado o otimizador Adam [Diederik P. Kingma [2017]] com taxa de aprendizado $9.1x10^{-4}$. O modelo é composto por três camadas escondidas de 250 neurônios cada e uma camada oculta de 5 neurônios (futuramente utilizada para a transferência), intercaladas com camadas de Dropout com taxa de 10^{-1} , função de ativação ReLU (Rectified Linear Unit) e regularização L2 de valor 10^{-1} . A função de perda Erro Absoluto Médio (MAE) foi definida e o modelo foi compilado e treinado, por 5000 épocas, utilizando o método de EarlyStopping aplicado ao MAE dos dados de validação, com paciência de 300 épocas.

2.4 Aprendizado por Transferência

Em posse dos modelos treinados na base de dados fonte, esse foi aplicado a 16 reações de lignina disponíveis na base de dados do CNPEM, geradas por cálculos DFT,

no intuito de visualizar como o modelo generaliza para uma base de dados de distribuição diferente da qual foi treinado. Desse montante, 4 reações foram separadas para re-treinamento da última camada do modelo, isto é, ajuste de 5 parâmetros treináveis, enquanto as outras 12 reações foram separadas para teste.

3 Resultados

Dois modelos foram arquitetados e treinados ,na base de dados fonte, de acordo com as informações descritas na última seção: um modelo possui como variável-alvo a diferença de entalpia e, o outro, a energia de ativação das reações químicas.

3.1 Pré-treinamento para predição da entalpia das reações

O modelo com variável-alvo diferença de entalpia, quando utilizado para prever as reações separadas nos dados de teste, apresentou MAE de 6.28 kcal/mol, RMSE de 9.42 kcal/mol e R^2 de 0.95. A Figura 2 apresenta as métricas de erro, para os dados de treino e dados de validação, durante o treinamento. Enquanto a Figura 1, abaixo, relata, visualmente, o desempenho do modelo em questão nos dados de teste.

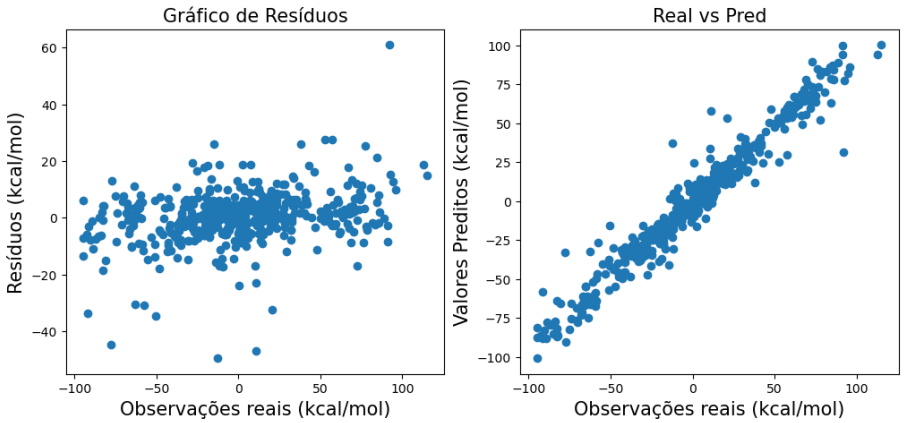

Figura 1: Dados de teste x Resíduos (à esquerda); Dados de Teste x Valores preditos (à direita);

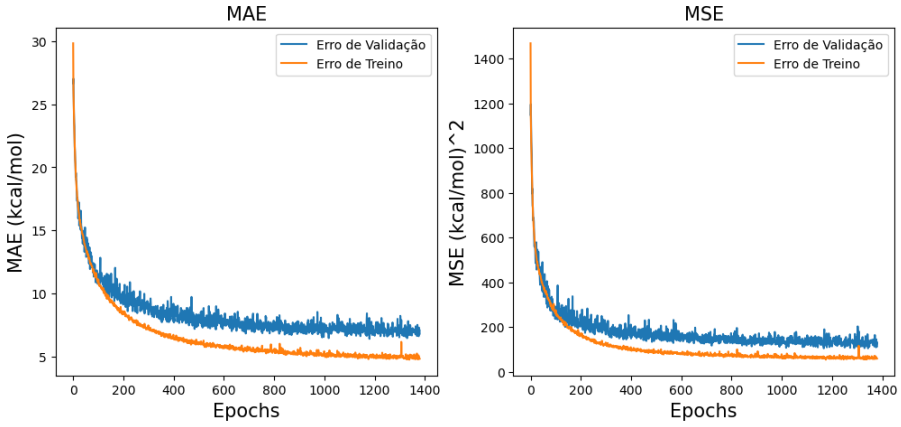

Figura 2: Erro Absoluto Médio (MAE), em kcal/mol, à esquerda, e Erro Quadrático Médio (MSE) , à direita, em $(kcal/mol)^2$, durante as épocas de treino do modelo para a previsão da diferença de entalpia;

3.2 Pré-treinamento para a predição da energia de ativação das reações

O modelo com variável-alvo energia de ativação, quando utilizado para prever as reações separadas nos dados de teste, apresentou MAE de 9.20 kcal/mol, RMSE de 13.67 kcal/mol e R^2 de 0.78. A Figura 4 apresenta as métricas de erro, para os dados de treino e dados de validação, durante o treinamento. Enquanto a Figura 3 relata visualmente o desempenho do modelo em questão nos dados de teste.

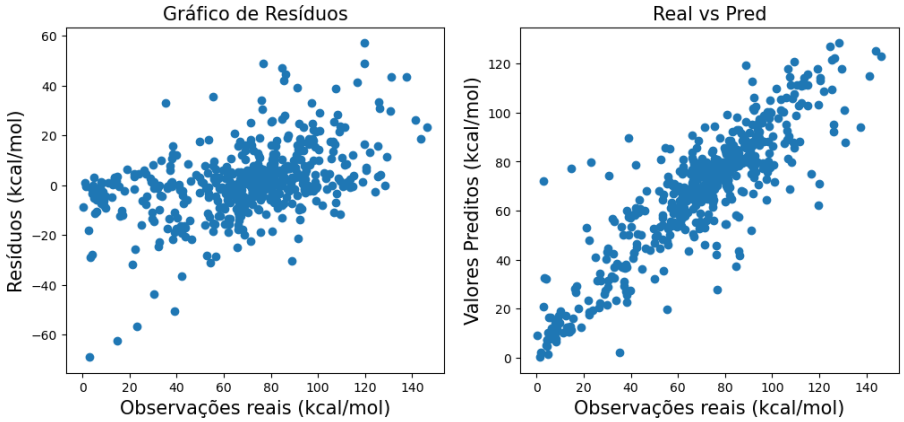

Figura 3: Dados de teste x Resíduos (à esquerda); Dados de Teste x Valores preditos (à direita);

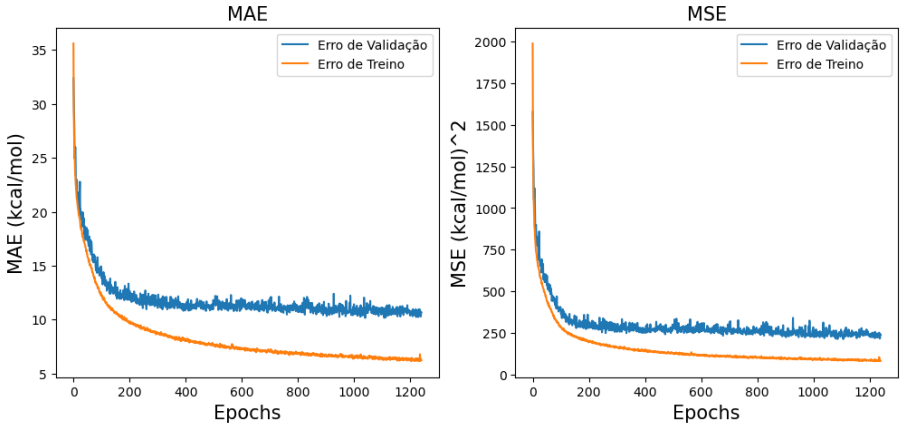

Figura 4: Erro Absoluto Médio (MAE), em kcal/mol, à esquerda, e Erro Quadrático Médio (MSE) , à direita, em $(kcal/mol)^2$, durante as épocas de treino do modelo para a previsão da energia de ativação;

3.3 Transferência para reações de lignina

O modelo treinado para a predição da diferneça de entalpia na base de dados fonte foi armazenado. Em posse das reações químicas de lignina, somente a última camada - 50 parâmetros treináveis - foi retreinada em 4 reações de treino, e então, testada nas 12 reações restantes. A Figura 5 relata visualmente o desempenho do modelo em questão nos dados de teste.

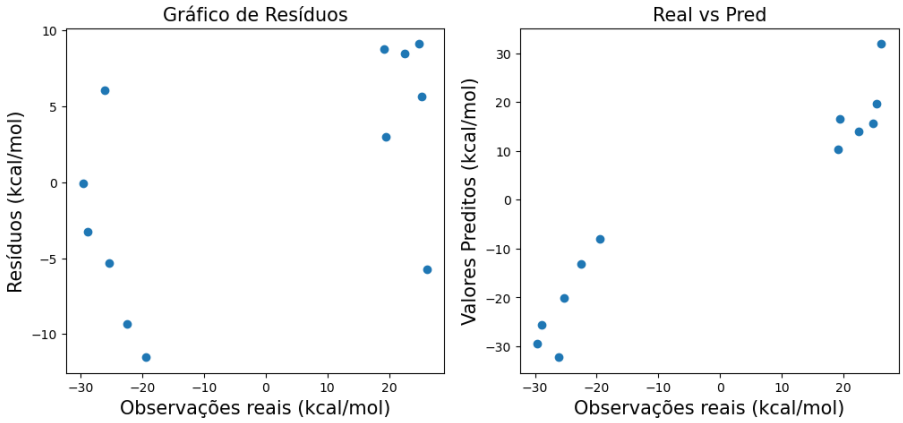

Figura 5: Dados de teste x Resíduos (à esquerda); Dados de Teste x Valores preditos (à direita);

O modelo testado para a previsão da energia de ativação demonstrou comportamentos mais imprevisíveis, que necessitam de análise adicional. A principal dificuldade de aplicação das técnicas citadas a uma base de dados de reações de lignina é que os pro-

cessos de geração de reações químicas factíveis, junto com suas diferenças de entalpia e energia de ativação, são computacionalmente caros. Em consequência, a aplicação direta de RNAs a reações de lignina é limitada pela baixa quantidade de dados. A hipótese gerada é que a utilização de técnicas de Aprendizado por Transferência entre uma base de dados de reações químicas maior e a base de dados de reações químicas de lignina do CNPEM, garanta a construção de um modelo de previsão de entalpia de reação de reações químicas de lignina com precisão química (1 kcal/mol), tal qual demonstrado por Grambow et al [Colin A. Grambow and Green [2020]].

4 Conclusão

O modelo arquitetado, quando testado na base de dados fonte, apresentou fortes sinais de aprendizado, demonstrando capacidade de generalização para dados fora da partição de treino, tanto no objetivo de predizer a energia de ativação quanto a diferença de entalpia das reações. É notável o superior desempenho do modelo na predição da diferença de entalpia em comparação com a energia de ativação . A diferença de entalpia é uma propriedade termodinâmica que está diretamente relacionada às energias de ligação dos reagentes e produtos, esses dados estruturais e energéticos são relativamente mais fáceis de incorporar em modelos preditivos, permitindo que a rede neural aprenda padrões consistentes e faça previsões mais precisas. Em contraste, a energia de ativação é uma propriedade cinética que depende do mecanismo de reação e do estado de transição, os quais são frequentemente complexos e menos acessíveis. Além disso, a energia de ativação é altamente sensível às condições experimentais, como temperatura e pressão, aumentando a variabilidade e a dificuldade de previsão.

Quando aplicado à base de dados de reações de lignina do CNPEM, os modelos de previsão de diferença de entalpia demonstraram que a transferência é positiva, mesmo com um número reduzido de reações de treino.

Desse modo, é explícito o potencial de utilizar técnicas de aprendizado de máquina, como transfer learning, para a predição de propriedades de reações químicas. Transfer learning permite que modelos treinados em um conjunto de dados maior e mais geral sejam ajustados para tarefas específicas com dados limitados, melhorando a precisão

das previsões. No contexto da predição de propriedades químicas, essa técnica se mostrou particularmente eficaz, permitindo que o modelo aproveite conhecimento prévio para obter melhores resultados em novos domínios, como foi observado . A utilização de transfer learning não só acelera o processo de treinamento, mas também melhora a robustez e a generalização dos modelos preditivos em cenários com dados escassos.

Referências

Lagnajit Pattanaik Colin A. Grambow and William H. Green. Deep learning of activation energies. 2020.

Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. 2017.

Pattanaik L. Green W.H Grambow, C.A. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. 2020.

Eiaki V. Morooka Filippo Federici Canova Yashasvi S. Ranawat David Z. Gao Patrick Rinke Adam S. Foster Lauri Himanen, Marc O.J. Jäger. Dscribe: Library of descriptors for machine learning in materials science. 2020.

Machine Learninng; Mitchell, Tom M. Machine learning. 1997. doi: ISBN:0070428077.

Sinno Jilalin Pan and Qiang Yang. A survey on transfer learning. 2009.

Van Den Bosch S et al Schutyser W, Renders T. Chemicals from lignin: an interplay of lignocellulose fractionation, depolymerisation, and upgrading. 2018.

Reuter K Margraf JT Stocker S, Csányi G. Machine learning in chemical reaction space. 2020.