USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA PREDIÇÃO DE PROPRIEDADES DE MOLÉCULAS DE CORANTES EM CÉLULAS SOLARES

GIOVANI LUIS VOLOSKI[1,2]*, CLÓVIS CAETANO[2,3]

1 INTRODUÇÃO

Dado o cenário energético mundial atual, novas fontes de energia renováveis estão sendo buscadas, sendo a energia solar uma das mais promissoras. Embora as células fotovoltaicas de silício sejam as mais utilizadas, esse tipo de célula teoricamente não pode ultrapassar os 30% de eficiência de conversão energética (SHOCKLEY; QUEISSER, 1961). Existem outros tipos de células que não compartilham do mesmo limite teórico, como é o caso da célula solar sensibilizada por corante (DSSC, do inglês: *dye sensitized solar cell*). O baixo custo de produção e a expectativa de se poder atingir altas eficiências de conversão energética têm tornado a tecnologia uma das mais estudadas atualmente. Uma DSSC é constituída por vários componentes: corante, filme de óxido semicondutor, filme transparente condutor, catalisador e solução eletrolítica. Cada componente, aliado à forma como está relacionado aos demais, influencia a eficiência de conversão energética de uma DSSC. O corante, em especial, é um componente decisivo. Para que seja um fotossensitizador eficiente, um corante deve satisfazer algumas condições, entre elas, ter alta absorbância para comprimentos de onda da região visível do espectro (COUTINHO, 2014).

Existem milhares de corantes possíveis de serem utilizados em DSSC. Entretanto, o uso de métodos experimentais na busca por corantes que satisfaçam tais condições pode não ser trivial, exigindo dispêndio de tempo, recursos financeiros, recursos materiais e geração de resíduos. Para estas situações, o uso de métodos de ciência de dados e inteligência artificial têm se mostrado promissores. Um exemplo desses métodos na área de quimioinformática é o estudo das relações quantitativas entre estrutura e atividade/propriedade, que gera modelos preditivos a partir de milhares de compostos e descritores moleculares (ALVES *et al*., 2018). Assim, dispondo de uma base de dados e técnicas de aprendizado de máquina (ML, do inglês: *machine learning*), o projeto se propôs a desenvolver metodologias para predizer o comprimento de onda de máxima absorção de moléculas de corante empregadas em DSSC.

1 Acadêmico do curso de Física - Licenciatura, **Universidade Federal da Fronteira Sul**, **campus Realeza**, **contato: giovanivoloski@hotmail.com**
2 Grupo de Pesquisa: Grupo de Pesquisa em Energias Renováveis e Sustentabilidade - GPERS
3 Docente, Universidade Federal da Fronteira Sul, **Orientador**.

2 OBJETIVOS

Utilizar métodos de aprendizado de máquina no estudo do comprimento de onda de máxima absorção de moléculas de corante (λ_{max}) para DSSC.

3 METODOLOGIA

Em um primeiro momento, utilizando a base DSSCDB (VENKATRAMAN, 2020), realizou-se o tratamento dos dados. A base contava com 4043 entradas de dados, 3685 corantes (em notação smiles), 37 solventes (18 puros e 19 misturas) e 4 semicondutores diferentes utilizados em DSSC. O tratamento dos dados incluiu a localização e exclusão de possíveis duplicatas da base. Para isso, utilizou-se a biblioteca RDKit (LANDRUM, 2006) que, entre outras funções, permite a representação gráfica das moléculas e o cálculo de similaridade entre cada par delas.

Em um segundo momento, realizou-se o cálculo de diferentes descritores moleculares. Esses descritores foram usados como *features* nos modelos de aprendizado de máquina. Um descritor molecular é definido como "o resultado final de um procedimento lógico e matemático, que transforma a informação química codificada dentro de uma representação simbólica de uma molécula em um número útil ou o resultado de algum experimento padronizado" (ALVES *et al*., 2018, apud CONSONNI; TODESCHINI, 2010, p. 204). Neste trabalho os descritores foram gerados através da biblioteca Mordred (MORIWAKI, 2018). Como o número de descritores gerados foi grande (1613), foram descartados aqueles com baixa variância e também altamente correlacionados. Além disso, foi usado o algoritmo SFS (*sequential forward selection*) para determinar quais descritores eram mais relevantes para a construção dos modelos. Este método é capaz de formar um subconjunto de *features* relevantes com base na pontuação de validação cruzada de um estimador. Desse modo, foram selecionados apenas 8 descritores.

No terceiro e último momento, realizou-se a construção dos modelos a partir de diferentes métodos de regressão (linear, polinomial, de k-vizinhos mais próximos e rede neural), utilizando a biblioteca scikit-learn (PEDREGOSA *et al*., 2011). Tais modelos buscam estabelecer uma relação entre os descritores e os comprimentos de onda de máxima absorção. Para a construção e validação dos modelos, os dados foram divididos em conjunto de treinamento e conjunto de teste (80% e 20%, respectivamente). O conjunto de treinamento, por sua vez, foi subdividido em 5 partes para realização de validação cruzada. O coeficiente

de determinação (R^2) e a raiz do erro quadrático médio (RMSE) foram utilizados como medidas do desempenho dos modelos.

4 RESULTADOS E DISCUSSÃO

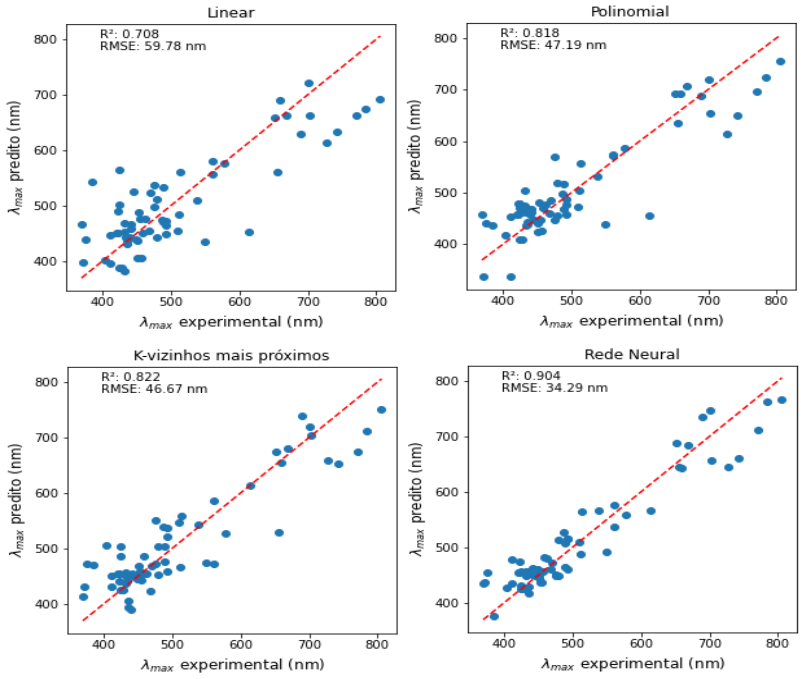
Na Tabela 1 são mostrados os resultados das validações cruzadas para o conjunto de treinamento.

Tabela 1: Métodos de regressão utilizados para construção do modelo e os respectivos coeficientes de determinação e raiz do erro quadrático médio resultantes da validação cruzada.

Método	R^2	RMSE (nm)
linear	$0{,}698 \pm 0{,}072$	$52{,}503 \pm 6{,}124$
polinomial	$0{,}809 \pm 0{,}058$	$41{,}217 \pm 4{,}073$
k-vizinhos mais próximos	$0{,}999 \pm 0{,}001$	$2{,}403 \pm 1{,}202$
rede neural	$0{,}949 \pm 0{,}024$	$21{,}245 \pm 4{,}229$

Na Figura 1 são mostrados os resultados dos valores preditos de λ_{max} para as moléculas do conjunto de teste em comparação com os resultados experimentais.

Figura 1. Comparação entre valores experimentais e valores preditos para o comprimento de onda de absorção máxima das moléculas de corante. A linha tracejada representa a relação ideal entre os valores. Os coeficientes de determinação e raiz do erro quadrático médio das validações externas são indicados.



A comparação entre os resultados da validação externa indica que, para os fins de predizer o λ_{max} dos corantes, o modelo gerado a partir da regressão linear apresentou o menor desempenho, seguido pelos modelos gerados a partir da regressão polinomial e k-vizinhos mais próximos, ambos com desempenho parecido. Por fim, o modelo gerado a partir da rede neural apresentou o melhor desempenho entre todos.

5 CONCLUSÃO

De forma geral, pode-se dizer que o objetivo foi contemplado. Entre outros, conseguiu-se desenvolver metodologias para a construção de modelos que estimassem o λ_{max} de moléculas de corantes a partir de dados catalogados na base de dados. Também fez-se possível comparar o desempenho de diferentes métodos de aprendizado de máquina. A experimentação permite concluir que os métodos de aprendizado de máquina são promissores para a descoberta de relações na predição do λ_{max} de moléculas de corante.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, Vinicius M. et al. QUIMIOINFORMÁTICA: UMA INTRODUÇÃO. ***Quím. Nova***. 2018, vol.41, n.2, pp.202-212.

COUTINHO, Natália de Faria. **Células solares sensibilizadas por corante**. Dissertação (Mestrado em Física) - Universidade Estadual de Campinas, Campinas, 2014.

CONSONNI, V.; TODESCHINI, R. In: PUZYN, T.; LESZCZYNSKI, J.; CRONIN, M. T. **Recent Advances in QSAR Studies**. Dordrecht: Springer, 2010, cap. 3.

VENKATRAMAN, Vishwesh; CHELLAPPAN, Lethesh Kallidanthiyil. An Open Access Data Set Highlighting Aggregation of Dyes on Metal Oxides, **Data Descriptor**, 5, 45, 1-8, maio. 2020.

VENKATRAMAN, V. et al. The dye sensitized solar cell database sensitized solar cells. **Journal of Cheminformatics**, v. 10, p. 18, 2018.

SHOCKLEY, W; QUEISSER, H. J. Detailed balance limit of efficiency of p-n junction solar cells. **J. Appl. Phys**, 32:510–519, 1961.

LANDRUM, G. **RDKit**: open-source cheminformatics. 2006. Disponível em <http://www.rdkit.org/>. Acessado em: 18/08/2021.

MORIWAKI, H. et al. Mordred: a molecular descriptor calculator. 2018. **Journal of Cheminformatics** 10:4. 6 fev. 2018.

PEDREGOSA, F. *et al.* Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research** 12, p. 2825-2830. 10 nov. 2011.

Palavras-chave: Células solares; aprendizado de máquina; quimioinformática.

Nº de Registro no sistema Prisma: PES 2020-0411.

Financiamento: Fundação Araucária.