

## Trabalho

|                                    |                                                                                                                     |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| <b>Título em Português:</b>        | Desenvolvimento de modelo de aprendizagem de máquina para previsão de parâmetros fotofísicos de moléculas orgânicas |
| <b>Título em Inglês:</b>           | Development of a machine learning model for predicting photophysical parameters of organic molecules                |
| <b>Autor:</b>                      | João Marcos Rigon Araujo                                                                                            |
| <b>Instituição:</b>                | Universidade de São Paulo                                                                                           |
| <b>Unidade:</b>                    | Instituto de Física de São Carlos                                                                                   |
| <b>Orientador:</b>                 | Leonardo De Boni                                                                                                    |
| <b>Área de Pesquisa / SubÁrea:</b> | Física da Matéria Condensada                                                                                        |
| <b>Agência Financiadora:</b>       | USP - Programa Unificado de Bolsas                                                                                  |

## Previsão de parâmetros fotofísicos em moléculas orgânicas baseada em aprendizado de máquinas.

João Marcos Rigon Araujo

Leonardo de Boni

Universidade de São Paulo

joao\_araujo@usp.br

### Objetivos

O objetivo principal do projeto foi replicar um modelo de aprendizado de máquina com a intenção de prever parâmetros fotofísicos de moléculas orgânicas, mais precisamente o momento de dipolo e o *gap* de energia do HOMO-LUMO.

### Métodos e Procedimentos

Modelos de predição baseados em aprendizado de máquina têm sido aplicados em uma ampla diversidade de áreas e tarefas. Por meio de uma função de perda a ser minimizada pelo modelo, é possível mapear iterativamente uma característica que se deseja prever ou entender com base em outras características conhecidas. O desafio, neste caso, é traduzir eficientemente uma molécula em uma estrutura de dados que sirva de entrada para esses algoritmos. Dessa forma, é necessário um engajamento inicial com a base de dados QM9 [1]. Essa base de dados concentra um universo de 134 mil moléculas com inúmeras propriedades físico-químicas calculadas por meio de cálculos em química quântica. Para uma melhor compreensão dos dados, conceitos teóricos chave foram estudados, como o *gap* de energia HOMO-LUMO e o momento de dipolo de uma molécula. Além disso, a notação SMILES [2] foi introduzida para extrair as informações necessárias e, mais importante, transformar os dados da QM9 em entradas para o algoritmo de regressão.

Visto que o foco do deste trabalho é, utilizando o aprendizado de máquinas, fazer a predição do *gap* de energia HOMO-LUMO e do momento de dipolo de moléculas orgânicas contidas na base de dados, foi utilizada a biblioteca RDKit, mais precisamente seus módulos *Chem* e *Descriptors*, que irá auxiliar de maneira ímpar nos cálculos necessários.

Primeiramente, o módulo *Chem* irá permitir uma manipulação eficiente das estruturas moleculares, podendo ler, escrever e visualizar as moléculas de forma conveniente, além de, principalmente, realizar cálculos de propriedades moleculares. Já o módulo *Descriptors* apresenta uma ampla gama de descritores moleculares. No cálculo desses descritores, é possível obter informações importantes das características estruturais e físico-químicas das moléculas, os quais serão a base para a construção de modelos preditivos. Porém, nem todos os descritores são de interesse para o modelo, de forma que um tratamento nos dados foi efetuado, removendo dados que tenham baixa variância entre si e descritores que são altamente correlacionados. A escolha para o melhor método de regressão foi baseada no coeficiente de determinação  $R^2$  e, posteriormente, para a análise do modelo, o erro médio absoluto *MAE* que define a medida de erro entre o valor real e o previsto.

### Resultados

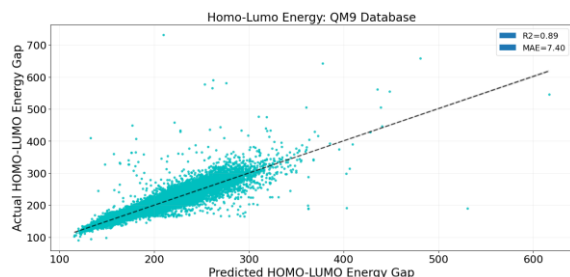


Figura 1: Gráfico do valor esperado e do valor predito para o *gap* de energia HOMO-LUMO.

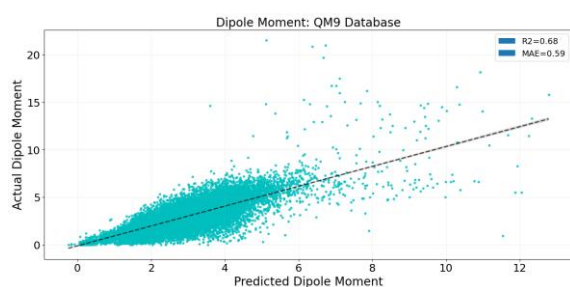


Figura 2: Gráfico do valor esperado e do valor predito para o momento de dipolo.

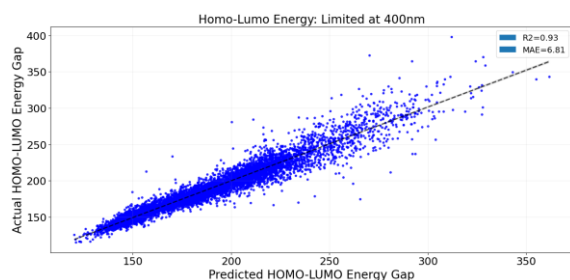


Figura 3: Gráfico do valor esperado e do valor predito para o *gap* de energia HOMO-LUMO com uma base de dados limitada a 400nm.

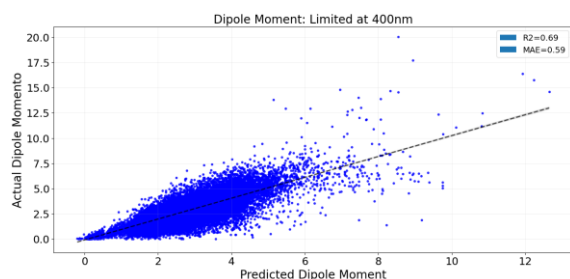


Figura 4 Gráfico do valor esperado e do valor predito para o momento de dipolo com uma base de dados limitada a 400nm.

## Conclusões

Durante o período do projeto, vários conceitos sobre aprendizado de máquinas foram introduzidos de forma que foi possível realizar a manipulação de informações contidas em base de dados de uma maneira eficiente. Além disso, é visto o poder da notação SMILES e da biblioteca RDKit na realização do modelo. Entretanto, os resultados não foram satisfatórios em relação a modelos mais atuais. É necessário um maior refinamento dos dados e uma revisão teórica mais detalhada nas moléculas em que o modelo falhou na predição.

## Agradecimentos

A realização do presente projeto foi possível em virtude do suporte do Grupo de Fotônica do Instituto de Física de São Carlos e à PUB como agência de fomento.

## Referências

- [1] NR. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Scientific Data 2014 1:1, vol. 1, no. 1, pp. 1–7, Aug. 2014, doi: 10.1038/sdata.2014.22.
- [2] Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." J. Chem. Inf. Comput. Sci. 28 (1988): 31-36.
- [3] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature 2018 559:7715, vol. 559, no. 7715, pp. 547–555, Jul. 2018, doi: 10.1038/s41586-018-0337-2.
- [4] Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, Frank Glorius, "A Structure-Based Platform for Predicting Chemical Reactivity", Chem, Volume 6, Issue 6, 2020, Pages 1379-1390.

## Machine learning-based prediction of photophysical parameters in organic molecules.

João Marcos Rigon Araujo

Leonardo de Boni

University of São Paulo

joao\_araujo@usp.br

### Objectives

The main goal of the project was to replicate a machine-learning model with the aim of predicting photophysical parameters of organic molecules, specifically the dipole moment and the HOMO-LUMO energy gap.

### Materials and Methods

Prediction models based on machine learning have been applied in a wide variety of fields and tasks. Through a loss function to be minimized by the model, it can iteratively map a characteristic that one wants to predict or understand based on other known characteristics. The challenge, in this case, is to efficiently translate a molecule into a data structure that serves as input for such algorithms. Therefore, an initial engagement with the QM9 database [1] was necessary. The QM9 dataset contains over 134 kilo molecules with various physicochemical properties calculated through quantum chemistry calculations. For a better understanding of the data, key theoretical concepts were studied, such as HOMO-LUMO energy gap and the dipole moment of a molecule. Additionally, the SMILES [2] notation was introduced to extract the necessary information and, more importantly, to transform the QM9 data into inputs for the regression algorithm.

Given that the focus of this work is to use machine learning to predict the HOMO-LUMO

energy gap and the dipole moment of organic molecules in the database, the RDKit library was used, specifically its *Chem* and *Descriptors* modules, which uniquely assist in the necessary calculations.

First, the *Chem* module allows efficient manipulation of molecular structures, enabling reading, writing and visualizing molecules conveniently, as well as performing molecular property calculations. The *Descriptors* module provides a wide range of molecular descriptors. By calculating these descriptors, it is possible to obtain important information about the structural and physicochemical characteristics of molecules, which will serve as the basis for building the predictive model. However, not all descriptors are of interest, so a data treatment was performed. Removing data with low variance and descriptors that are highly correlated with each other. The choice of the best regression method was based on the coefficient of determination  $R^2$ , and later, for the model analysis, the mean absolute error *MAE* was used, which defines the error measure between the actual and predicted values.

### Results

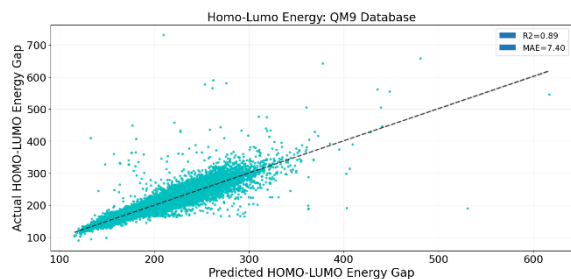


Figure 1: Graph of the expected value and the predicted value for the HOMO-LUMO energy gap for the total dataset.

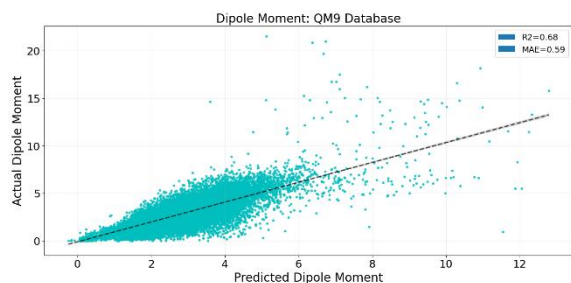


Figure 2: Graph of the expected value and the predicted value for the dipole moment for the total dataset.

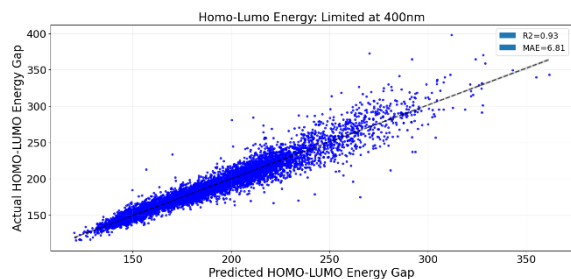


Figure 3: Graph of the expected value and the predicted value for the HOMO-LUMO energy gap for a database limited at 400nm.

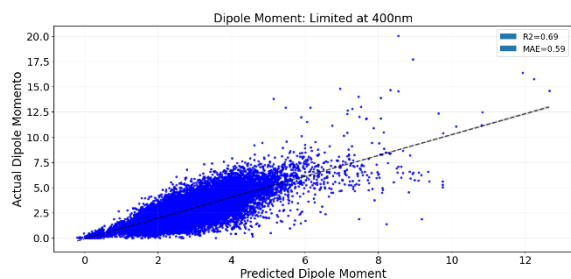


Figure 4: Graph of the expected value and the predicted value for the dipole moment for a dataset limited at 400nm.

## Conclusions

During the course of the project, various machine-learning concepts were introduced, allowing for efficient manipulation of information contained in databases. Additionally, the power of the SMILES notation and the RDKit library in building the model became evident. However, the results were not satisfactory compared to models that are more current. Further data refinement is necessary, as well as a more detailed theoretical review of the molecules where the model failed in prediction.

## Acknowledgements

The realization of this project was made possible thanks to the support of the Photonics Group at the São Carlos Institute of Physics and the PUB as a funding agency.

## References

- [1] NR. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Scientific Data 2014 1:1, vol. 1, no. 1, pp. 1–7, Aug. 2014, doi: 10.1038/sdata.2014.22.
- [2] Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." J. Chem. Inf. Comput. Sci. 28 (1988): 31–36.
- [3] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature 2018 559:7715, vol. 559, no. 7715, pp. 547–555, Jul. 2018, doi: 10.1038/s41586-018-0337-2.
- [4] Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, Frank Glorius, "A Structure-Based Platform for Predicting Chemical Reactivity", Chem, Volume 6, Issue 6, 2020, Pages 1379-1390.