Research proposal

Exploring the potential of machine learning as a tool for chemical synthesis using SHAP

Summary of the research project:

SHAP (SHapley Additive exPlanations) is a game theory inspired method proposed by Lundberg and Lee in 2017 to explain predictions made by machine learning models. It uses the concept of Shapley values introduced by Lloyd Shapley in 1953 which is a mathematical way of fairly dividing payout among players in a cooperative game. Following the basis that a prediction made by a computer can be viewed as a coalitional game by considering each feature value of an instance (i.e. parameter) as a "player" in a game, with the "payout" becoming the value predicted by the machine, you can apply the same mathematical rules used for Shapley values to interpret the outputs of an artificial intelligence.

Complex machine learning models have in general higher predictive performances than trivial ones when it comes to applications in chemistry; however they are not inherently interpretable which can discourage scientists from using them and lead to distrust despite their potential. The goal of this project is to train different machine learning models with the experimental data from a case study before using SHAP to understand what the machine has learned and how it applies it in its predictions to bridge the gap between data and modeler.

Description of the work that you will be specifically undertaken in this project:

The experimental data used in this project comes from the article "Bayesian Optimization as a Sustainable Strategy for Early-Stage Process Development? A Case Study of Cu-Catalyzed C–N Coupling of Sterically Hindered Pyrazines" which provides the results of an effort to identify sustainable reaction conditions with satisfying yields by using Bayesian optimizers with various acquisition functions. We get an initial understanding of the data by looking at the influence of certain parameters on yield before transforming it using one hot encoding or other representations depending on the machine learning model they are given to.

Then, we experiment. For this project we will mainly be using tree-based models on python like CatBoost or XGBoost that can be trained, be tested and have SHAP values computed from the data

they are using. In addition to these tools, other python packages will have to be understood to manipulate data (pandas, torch,...), visualize results (seaborn, RDKit,...), and evaluate the performance of a model's outputs (sklearn…).

With all the tools at our disposition we are aiming to understand why certain models give accurate or inaccurate predictions and to gain a further understanding of the data given by the article we are using as a base.

Expected planned research impact

One of the biggest challenges in chemistry, especially in synthesis, is to find ways to maximize the yield of specific reactions by changing their parameters (reagents, temperature, catalysis, etc.). Artificial intelligence is the somewhat "magic" solution to reduce the amount of wasteful experimenting that would be done if a brute force approach were chosen; as it could theoretically model and predict which parameters would lead to the best results, it would help pointing research in the right direction from the start. However, as powerful as a tool it may be, it needs to be properly understood and trusted before it can be applied at larger scales: that is what this project aims to do with SHAP.

At this time, we are still not fully satisfied with the outputs of even the most complex machine learning models in chemistry and we hope that by being able to accurately interpret them we can improve their efficiency. You can probably come up with an exhaustive list of all the fields that chemistry is involved in (pharmacy, agriculture, sanitation, food—just to name a few): accelerating chemical research with artificial intelligence will undoubtedly be beneficial for all of them.

References

- Molnar, C. (2023). *Interpreting Machine Learning Models With SHAP*. leanbook.com
- Braconi E., Godineau E. (2023). *Bayesian Optimization as a Sustainable Strategy for Early-Stage Process Development? A Case Study of Cu-Catalyzed C–N Coupling of Sterically Hindered Pyrazines. ACS Sustainable Chemistry & Engineering* **2023** *11* (28), 10545-10554. DOI: 10.1021/acssuschemeng.3c02455
- Lundberg S., Lee S.-I. (2017). *A Unified Approach to Interpreting Model Predictions.* DOI: https://doi.org/10.48550/arXiv.1705.07874
- Shapley, L. (1953). *A Value for n-Person Games*. In: Kuhn, H. and Tucker, A., Eds., Contributions to the *Theory of Games II*, Princeton University Press, Princeton, 307-317. DOI: https://doi.org/10.1515/9781400881970-018