Machine Learning of Rate Coefficients of S_NAr Reactions through rapid Quantum Chemical Property Prediction

Lowie Tomme[1], Florence Vermeire[2], Christian Stevens[3], Kevin M. Van Geem[1]*

1 Laboratory for Chemical Technology, Ghent University, Ghent, Belgium
2 Department of Chemical Engineering, KU Leuven, Leuven, Belgium
3 Department of Organic Chemistry, Ghent University, Ghent, Belgium
**Corresponding author: Kevin.VanGeem@UGent.be*

Highlights

- Quick calculation of quantum chemical property that relates to the reaction rate
- Machine learning method for fast prediction of rate constants
- ML models accurately capture trends in predicted rate

1. Introduction

The design of efficient processes is an essential step in drug development and production. Currently, the optimal process parameters are often obtained experimentally by trial and error. For the shift towards *in silico* design of pharmaceutical processes, fast and accurate insights in the kinetics of the studied reactions are required. Furthermore, knowledge of reaction rates can be useful in determining the optimal pathway toward the targeted drug. By selecting pathways where all reactions are fast energy, materials, and money can be saved.

The nucleophilic aromatic substitution (S_NAr) is one of the most widely used reaction families in the synthesis of pharmaceuticals. A recent study showed that this family is one of the most occurring in the field's literature [1]. This academic interest combined with the availability of experimental kinetic data makes the nucleophilic aromatic substitution an ideal reaction class to study *in silico* methods for predictive kinetics. Jorner et al. [2] constructed a literature database and predicted the rate coefficients of S_NAr reactions by using a combination of quantum chemistry and machine learning. This approach yields accurate results, but requires computationally expensive calculations for every reaction the user aims to predict.

In this work, a machine learning model is developed such that the reaction rate constants can be predicted in a fast and accurate manner. First, a quantum chemical property is proposed that is correlated with the rate constant for a wide variety of S_NAr reactions. This property is then calculated for a broad set of relevant molecules and predicted with a machine leaning model. The property is then used along with other features as input for a machine learning model to predict the rate constant. The addition of this property significantly improves the accuracy of the machine learning model and speeds up the calculation of rates for new reactions.

2. Results and discussion

A logical choice of a quantum chemical property that is related to the rate constant would be the activation energy, *i.e.* the energy difference between the transition state and the reactants. This, however, requires a computationally expensive transition state search. As the energy of the intermediate ionic complex is often close to the energy of the transition state, another reasonable property would be the energy difference between this intermediate and the reactants. The disadvantage of this approach is that some S_NAr reactions proceed via a concerted mechanism, consequently they do not have an intermediate. Therefore, a pseudo-complex is defined in which some distances and angles are constraint. These constraints ensure that an intermediate is found for every reaction.

Although this approach is faster than computing transition states, it still requires time consuming quantum chemical calculations. Therefore, a machine learning model is trained to predict this property faster. First, a large dataset containing 50000 relevant reactions is created for which this property is calculated. This data is then used to train a directed-message passing neural network followed by feed forward neural network via the open-source Chemprop package [3].

An experimental dataset of around 500 S_NAr reaction contained the reaction rate in various solvents was obtained from literature [2]. For methodical purposes, these rate constants were converted into the apparent activation energy ΔG_{TS}^{exp}, expressed in kJ/mol. The aforementioned ML model is used to predict the proposed QM property for the reactions in this dataset. This property is then used as additional input for another ML model, also using the Chemprop package [3], to predict the rate constant (or ΔG_{TS}^{exp}).

Figure 1 displays a plot showing the correlation between the ML predicted property and the rate constant for two subsets of the data. A clear linear relationship to the apparent activation energy is found, meaning that this QM property could indeed be a very relevant QM descriptor for speeding up the predictions of the rate coefficient of S_NAr reaction through the use of ML. In addition to the relevance of the property, Figure 1 also indicates that D-MPNN as implemented in Chemprop can grasp the relevant chemical effects to predict this property.

Finally, we tested whether adding this property would indeed improve the accuracy of a ML model to predict the apparent activation energy of S_NAr reaction. The performance of the model trained solely on experimental data is compared to the model using the property as an additional input. Adding the machine learned property as input improved the accuracy of the final ML model on ΔG_{TS}^{exp} from 5.1 kJ/mol to 4.6 kJ/mol, getting very close to chemical accuracy.

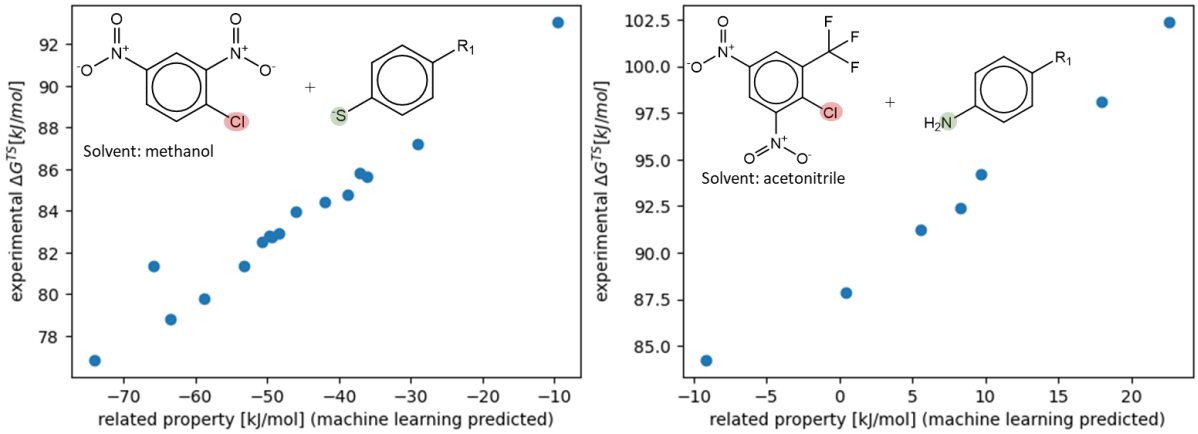

Figure 1. Plots linking the machine learning predicted related property with the apparent activation energy ΔG_{TS}^{exp}. For each subset, the leaving group is highlighted in red, the nucleophilic atom in green.

3. Conclusions

A new approach for the fast and accurate prediction of the reaction rate coefficient of nucleophilic aromatic substitution reactions was developed. The new method increased the accuracy in comparison with the benchmark. In the process of obtaining this improvement, a quantum chemical property was proposed. This property is correlated with the reaction rate constant and does not require computationally expensive transition state calculations. The use of this property and the combination with machine learning offers a novel approach for fast prediction of reaction rates that could be extended towards other reaction classes.

References

[1] Brown, D.G. and J. Boström, Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? Journal of Medicinal Chemistry, 2016. 59(10): p. 4443-4458.W.-D. Deckwer, R.W. Field, Bubble column reactors, Wiley, 1992

[2] Jorner, K., T. Brinck, P.-O. Norrby, and D. Buttar, Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. Chemical Science, 2021. 12(3): p. 1163-1175.

[3] Heid, E., et al., Chemprop: A Machine Learning Package for Chemical Property Prediction. Journal of Chemical Information and Modeling, 2024. 64(1): p. 9-17.

Keywords

Quantum Chemistry, Machine learning, Nucleophilic aromatic substitution