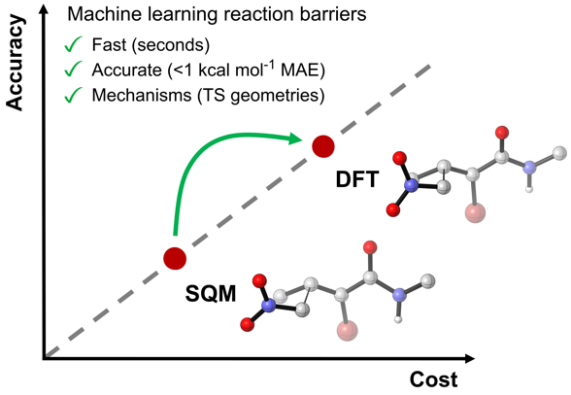
Machine Learning and Molecular Modelling: A Synergistic Approach to Rapid Reactivity Design

Matthew Grayson (M.N.Grayson@bath.ac.uk)
www.thegraysongroup.co.uk
University of Bath, UK

The computational design of new reactions is one of the "*Holy Grails*" of computational organic chemistry and biochemistry. Accurate and fast computational approaches to predicting chemical reactivity would provide cost-effective alternatives to experimental screening, and in some cases animal testing, in drug discovery, toxicology, and synthetic route design. However, current modelling (e.g., DFT) and machine learning (ML) methods all fail to simultaneously offer fast, accurate predictions with precise mechanistic insight derived from transition state (TS) geometries. This talk will discuss our recent work towards the development of new methods that combine the best elements of molecular modelling and ML to deliver rapid, accurate, and mechanism-based reactivity prediction.

In our initial proof-of-concept study,[1] we trained ML models that corrected rapidly obtained but approximate semi-empirical quantum mechanical (SQM) reaction barriers to DFT-quality barriers for a dataset of 1000 nitro-Michael reactions. After training, this approach enabled barrier predictions to be made in seconds, instead of hours or days with DFT, with errors below the widely accepted chemical accuracy threshold of 1 kcal mol^{-1}. Rapid mechanistic insight was also available from the SQM TSs which were found to be very good approximations for the DFT-level geometries.



We have since shown that through the use of transfer[2] and active learning[3], models can be built that offer the same speed, accuracy, and mechanistic insight with much lower amounts of training data (tens of data points). We then applied this methodology to the optimization of a transition metal-catalyzed dihydrogen activation reaction, which required just 12 rapidly computed new data points, to demonstrate our progress towards the "*Holy Grail*".

References

[1] *Chem. Sci.* **2022**, *13*, 7594.
[2] *Digital Discovery* **2023**, *2*, 941.
[3] *ACS Catal.* **2023**, *13*, 13506.