Machine Learning Analysis of Boron-Lewis Base Complexation

Lead supervisor: Dr Lia Sotorrios
Affiliation: University of Manchester, Faculty of Biology, Medicine and Health, School of Health Sciences, Division of Pharmacy and Optometry
Email: lia.sotorrios@manchester.ac.uk

Boron-containing molecules are central to modern organic synthesis, catalysis, and functional materials science because of their exceptional Lewis acidity and unique electronic structure. The empty p-orbital at boron serves as an electrophilic site that forms strong dative bonds with Lewis bases, creating adducts used in diverse chemical contexts. However, designing these systems rationally requires predicting complexation behaviour, such as binding energies (ΔE), bond strengths, and stability trends across BR_3 scaffolds (R = H, alkyl, halo, OH and more) paired with bases like NH_3, phosphines, N-heterocycles or common solvents.

Our recent unpublished quantum chemical calculations (using ORCA/Gaussian at DLPNO-CCSD(T) and DFT levels) have produced a comprehensive dataset of ~2000 optimized boron–Lewis base adducts. This includes key properties: binding energies (ΔE), LUMO energies of the free boranes, Mayer bond orders, pyramidalization (strain) energies for the planar-to-pyramidal distortion at boron, and substituent-dependent electronic/steric descriptors. While offering atomic-level mechanistic insight such as how B–X π-donation modulates strain or how LUMO energy correlates with orbital mixing, this dataset's scale demands sophisticated analysis to reveal the underlying structure–property relationships.

The main goal is to use machine learning as a data analysis tool to explore this dataset and extract practical chemical design rules. Using supervised models like random forests, we will rank key descriptors that drive binding strength, check trends hold up across different boron structures, and spot promising areas worth following up on.

The project revolves around Python-based workflows in a Linux HPC environment: data curation from ORCA/Gaussian outputs, descriptor computation, exploratory statistics and visualizations to spot correlations, supervised ML model training, interpretability analyses to extract feature importance, and potentially running additional quantum chemical calculations to expand or validate the dataset. HPC resources are indispensable for success due to the dataset's scale and workflow demands. Processing thousands of DFT-optimized structures requires parallel processing and shared storage, with analysis tasks and extra calculations leveraging multi-core parallelism via SLURM jobs for 10x speedups over local machines. SLURM scripting familiarizes the student with cluster workflows, mirroring real research. Without HPC, serial processing on laptops would bottleneck analysis; with it, focus shifts to chemical insight, delivering a validated ML pipeline and structure–property maps efficiently.