DC5: Combining QM simulations and ML models for reactivity prediction

Bob van Schendel

Reactivity prediction

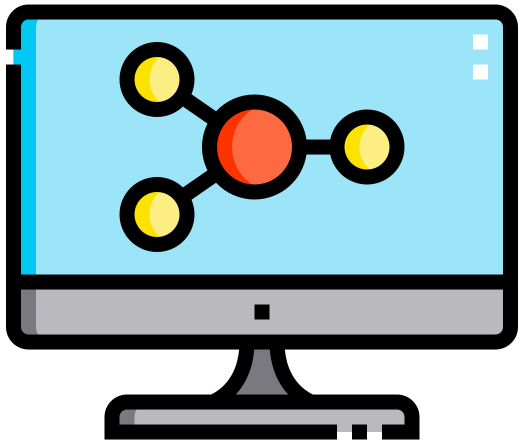
- Many regions in reactivity prediction with sparse data
- Generating big datasets with QM simulations is expensive

QM and ML interplay

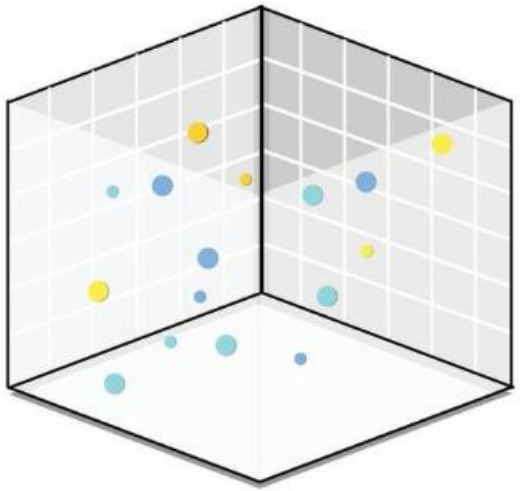
- Approximate experimental results
- Computationally expensive and slow (multiple hours to multiple days)

- Approximate QM results
- Computationally cheap and quick
- Limited by experimental + generated data

Train ML model on all available data



Sample points for QM simulation



Augment dataset with new data

2 parts:
QM data generation & ML model

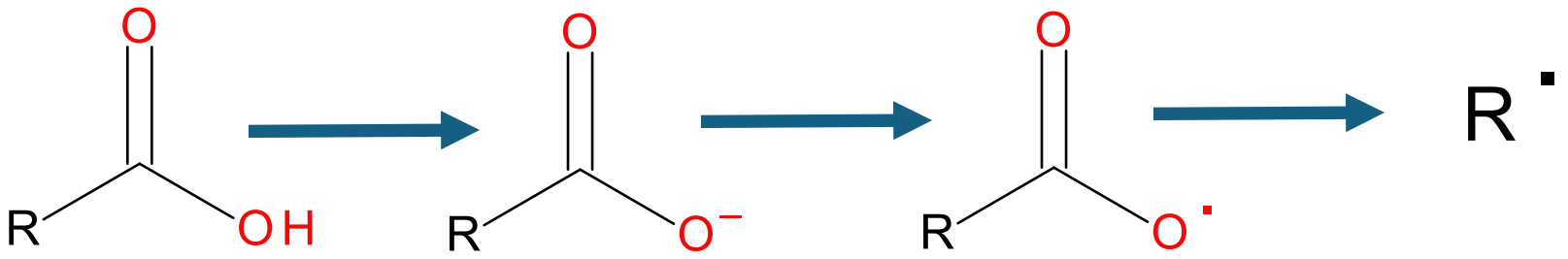
- Runs different programs as part of a QM workflow
- Long and computationally expensive

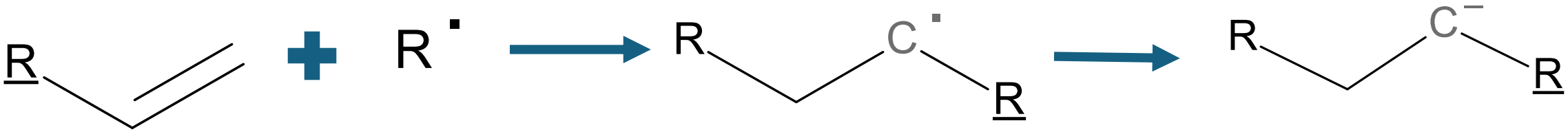
- Goal: approximate the **experimental** results

- Runs as *single model* or *hybrid model (part simulation)*
- Relatively short and computationally inexpensive

- Goal: approximate the **QM workflow** results

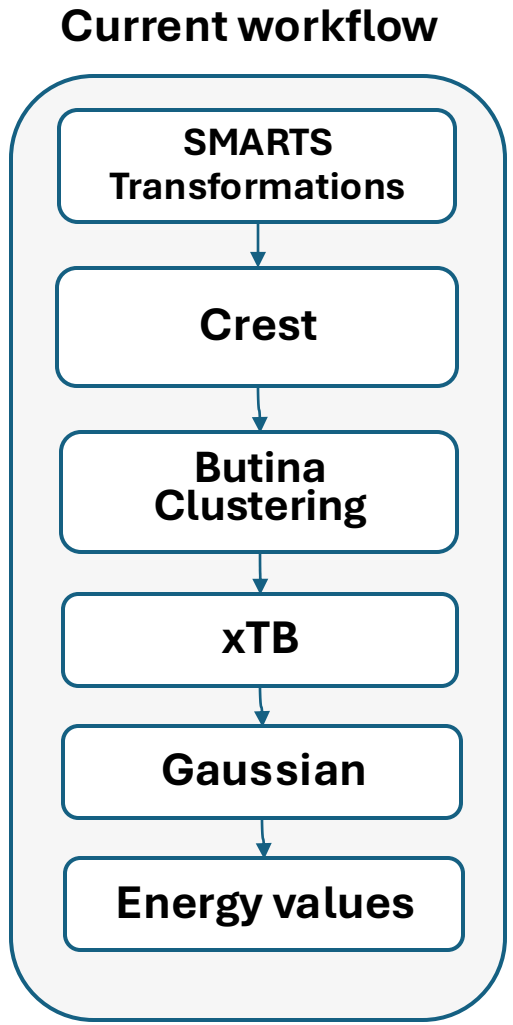
First project:
Model **Giese-like radical addition reactions** to predict *reaction feasibility*



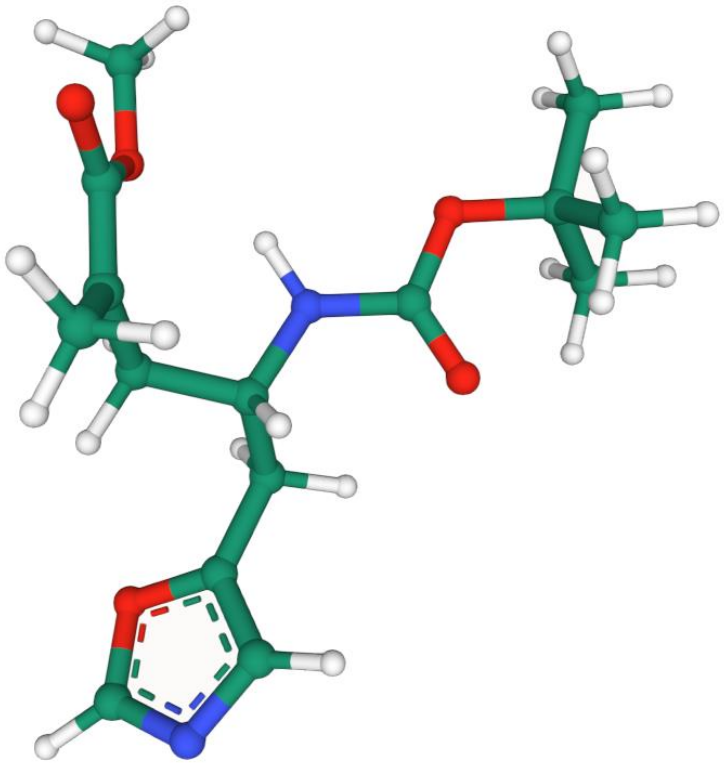


QM Workflow structure

- Generating intermediate states and product
- Conformational sampling
- Selection of diverse conformers
- Quick geometry optimization
- DFT geometry optimization



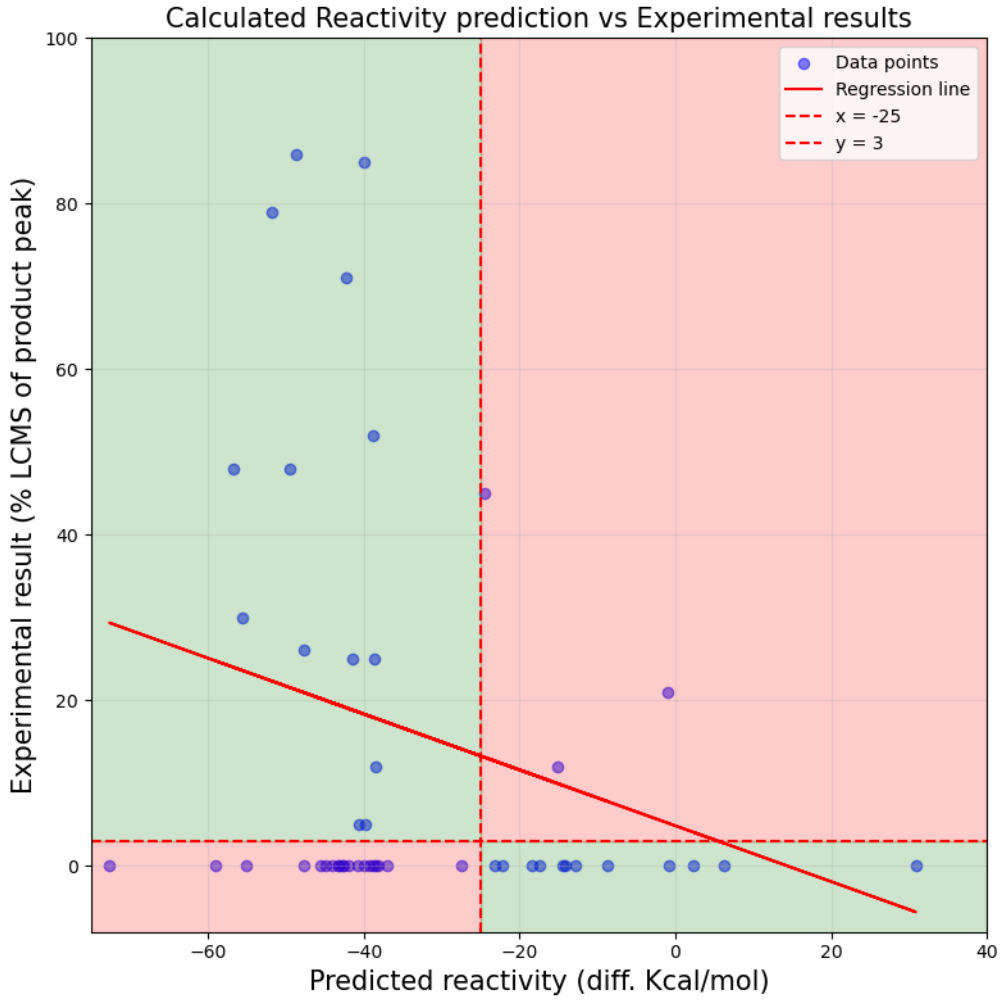
Calculated Molecular properties



- Gibbs free energy (GFE) *for each component*
- Energy *difference* between reactants and products
- Redox potentials
- Radical reactivity
- HOMO-LUMO
- Electrophilicity & Nucleophilicity

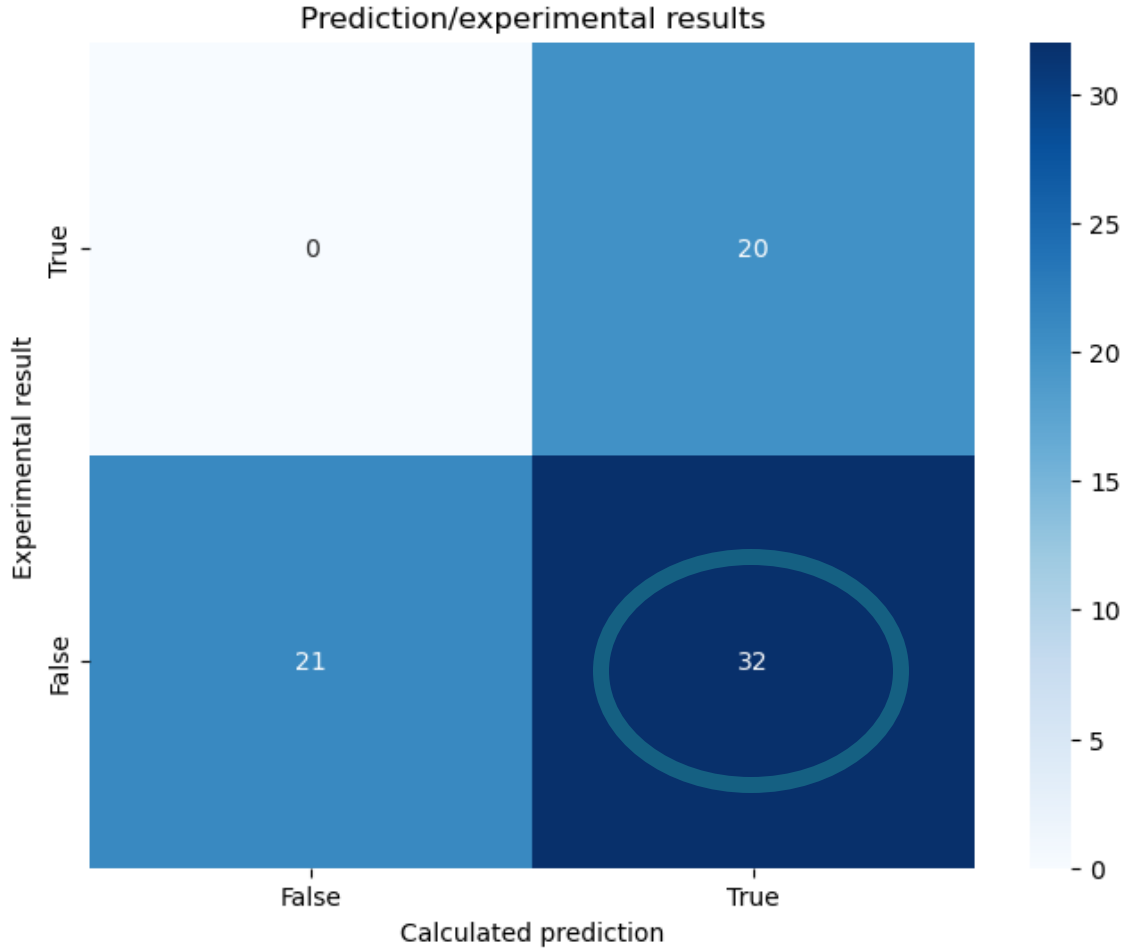
Initial validation for w/o DFT opt.

- Enthalpy change (*ΔH) is* predicted reactivity indicator
- Subpar accuracy
- *Many false positives*

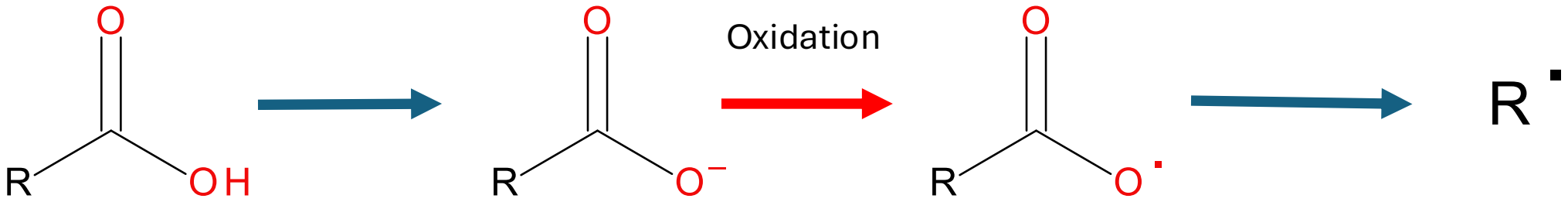


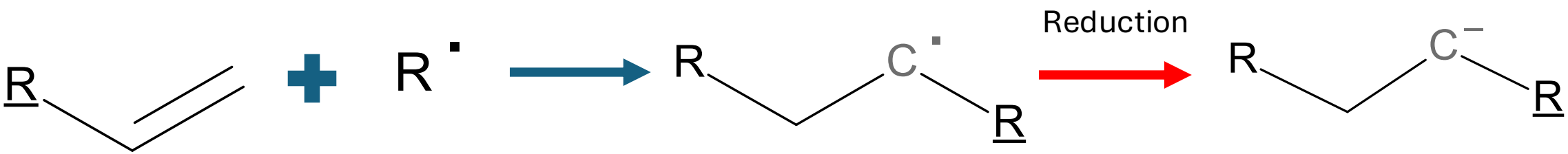
Initial validation for w/o DFT opt.

- Using only predicted reactivity not sufficient
- *Need to explain/predict the false positives*
- *Possible explanation:* ***redox potential*** *are too large/small*



Giese-like radical addition reaction





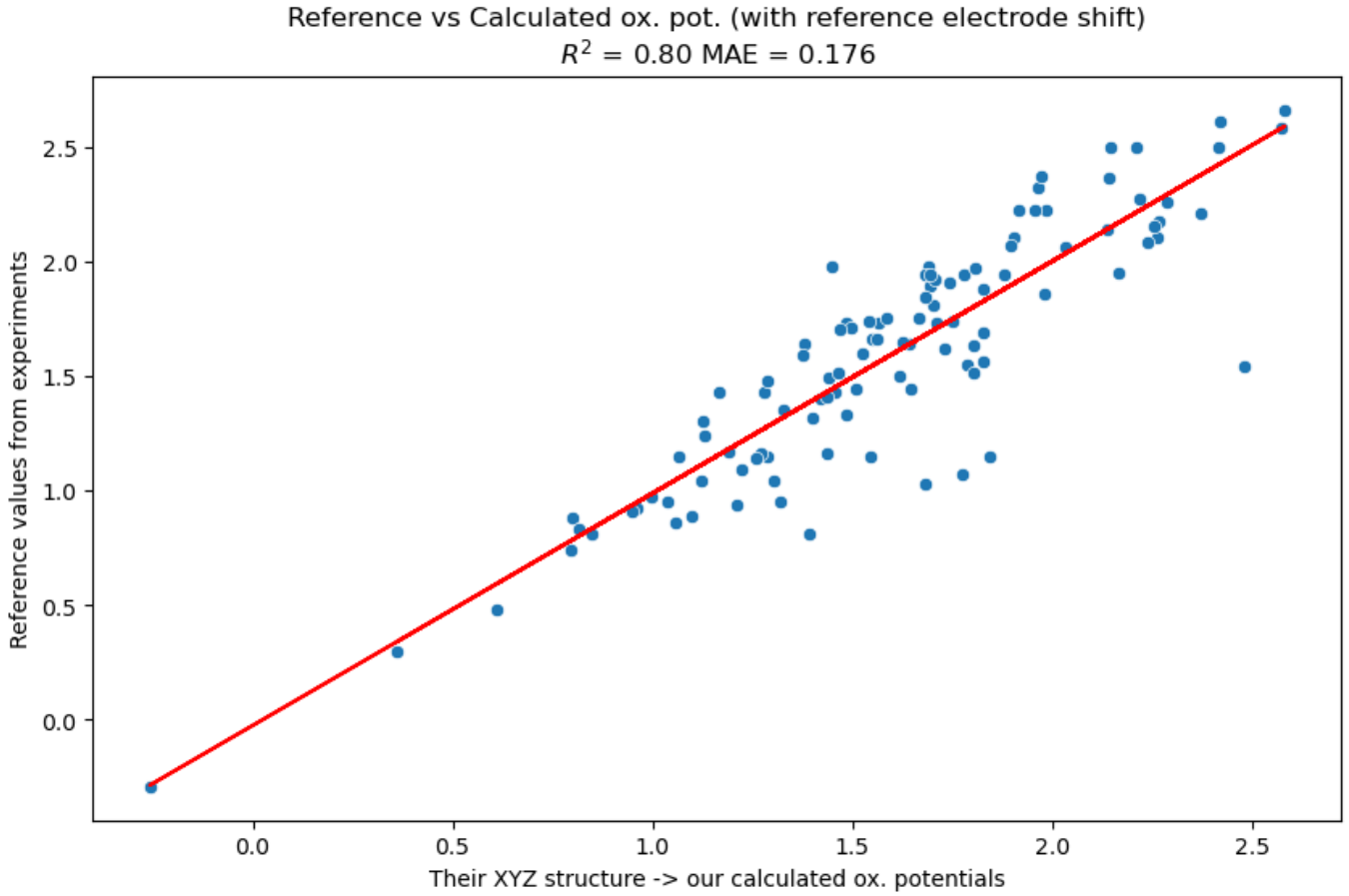
Redox potential validation with ROP313

Setup:

- Only take the organic molecules from the ROP313 dataset
- Calculate oxidation potential for organic compounds
- Same setup as the QM workflow

Goal:

- Calculating the *Gaussian redox potential offset (-4.26V)*
- Benchmarking our redox potential calculations (MAE = 0.18)

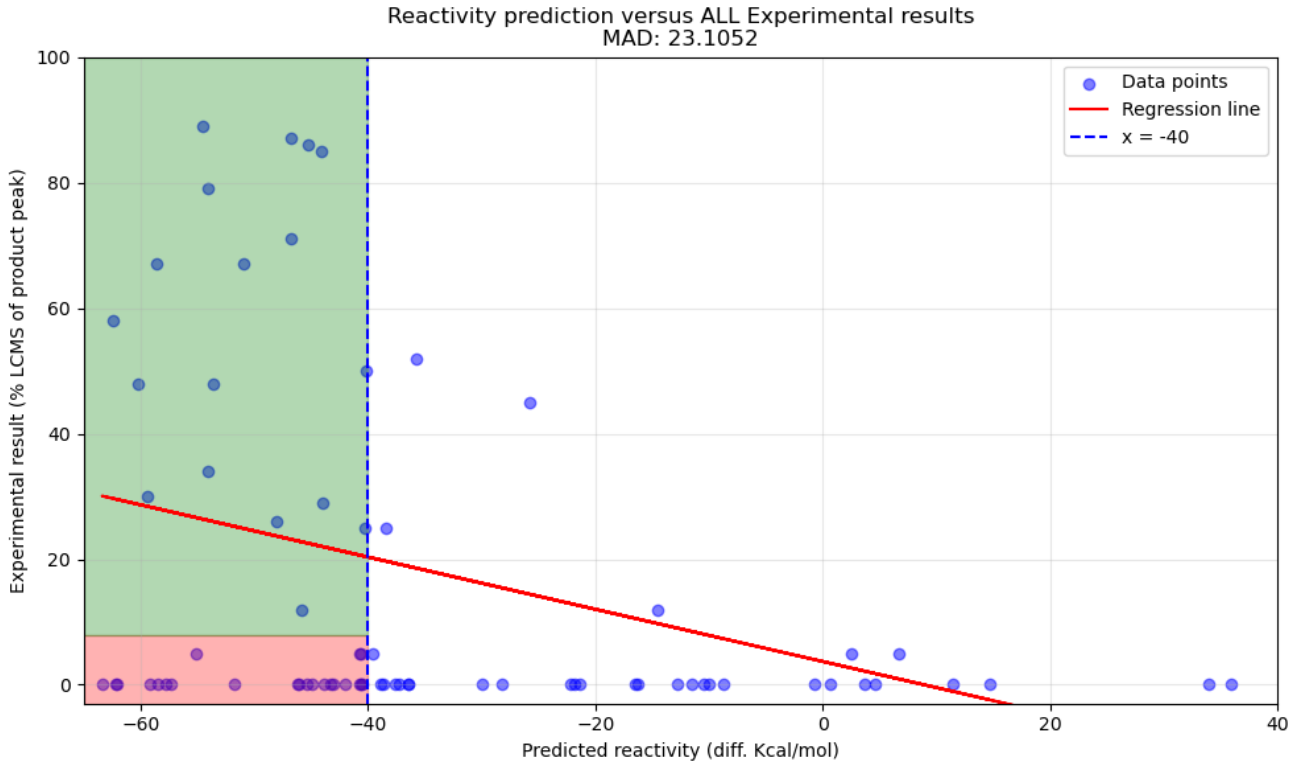


Hagen Neugebauer et al., "Benchmark Study of Electrochemical Redox Potentials Calculated with Semiempirical and DFT Methods," *The Journal of Physical Chemistry A* 124, no. 35 (2020): 7166–7176, https://doi.org/10.1021/acs.jpca.0c05052.

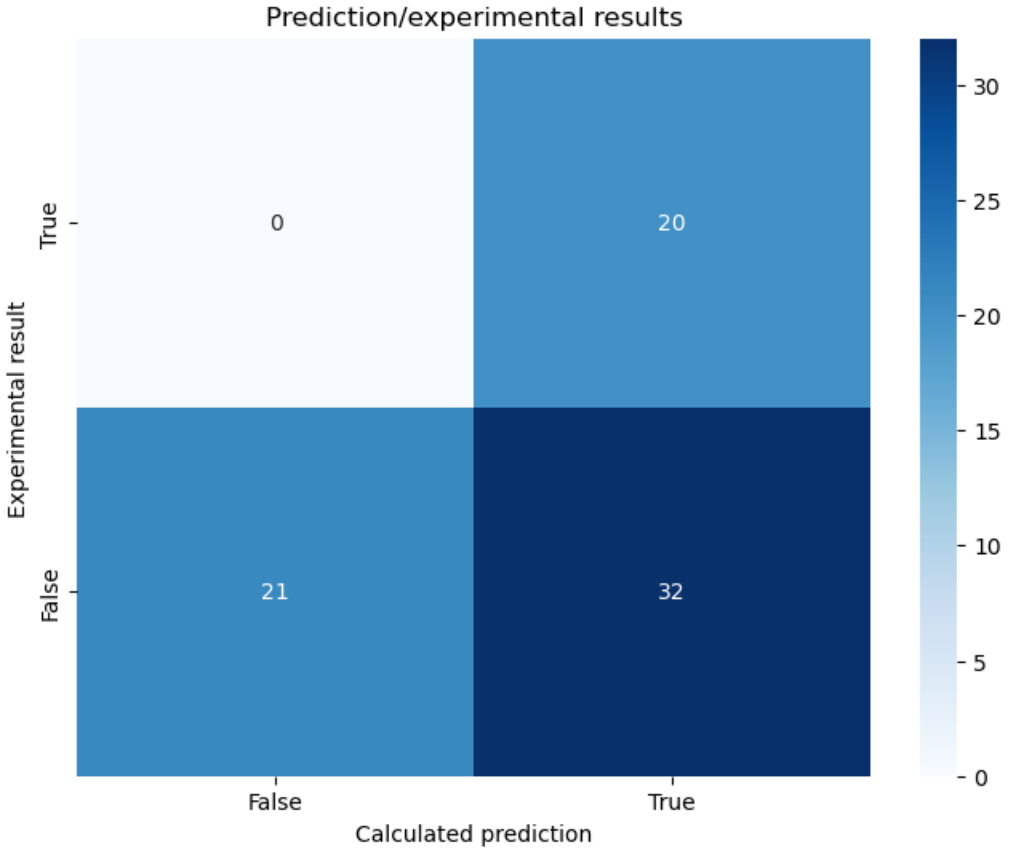
Redox potential distribution comparison

False positives:

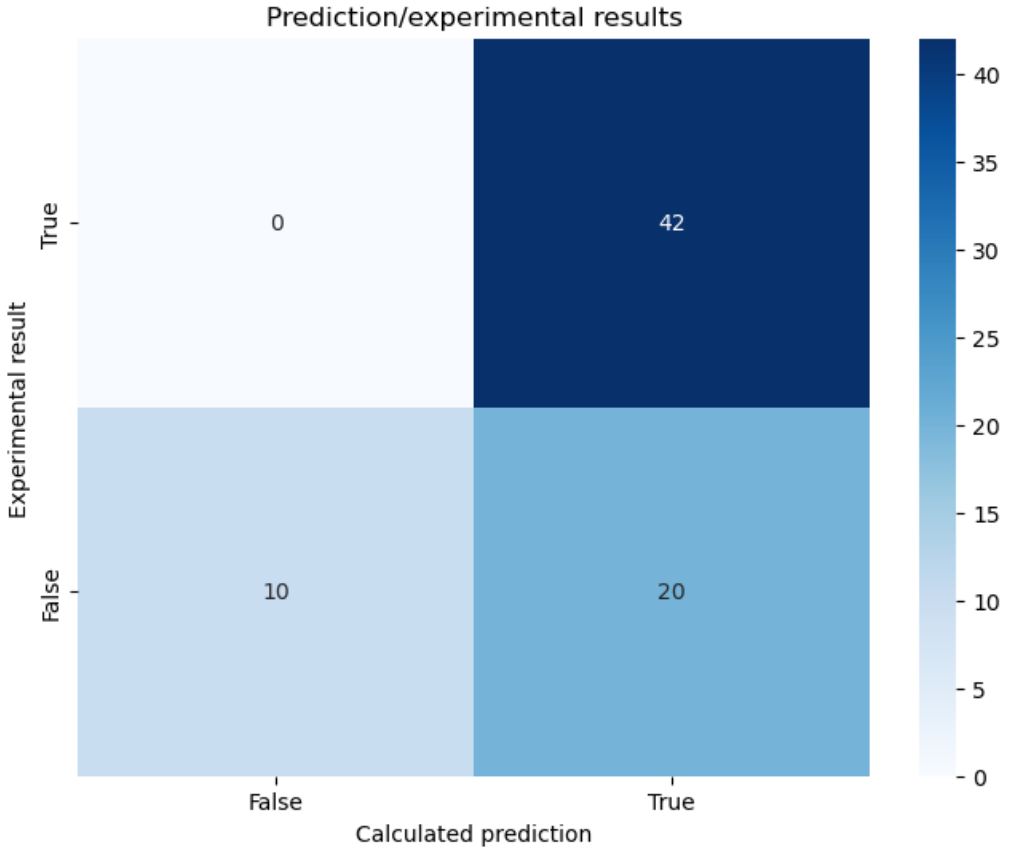
- Difficult to chemically explain failures
- Need to improve experimental setups



Adding DFT geometry optimization to workflow



Without DFT

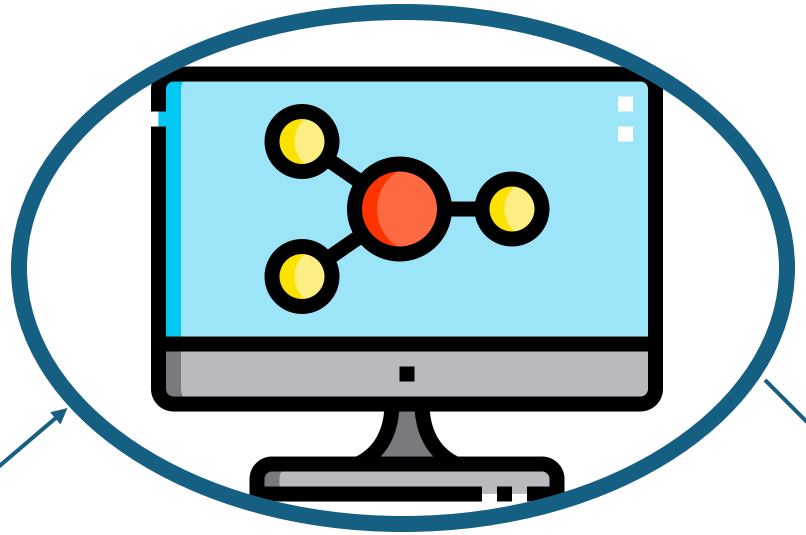


With DFT

Effect of additional molecular descriptors on prediction of experimental success

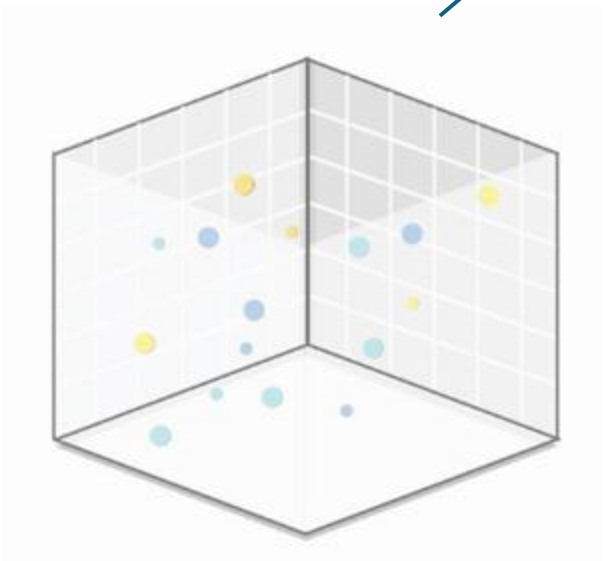
	Reactivity + redox potentials			All + HLG + nucleo/electrophilicity		
Model	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Logistic regression	**0.76**	**0.70**	**0.70**	0.67	0.66	0.66
Random Forest	0.63	0.62	0.62	0.68	0.68	0.67
Decision tree	0.57	0.56	0.57	0.70	0.69	0.68
XGBoost	0.68	0.66	0.66	**0.73**	**0.71**	**0.71**
Neural network	0.65	0.57	0.49	0.60	0.58	0.57

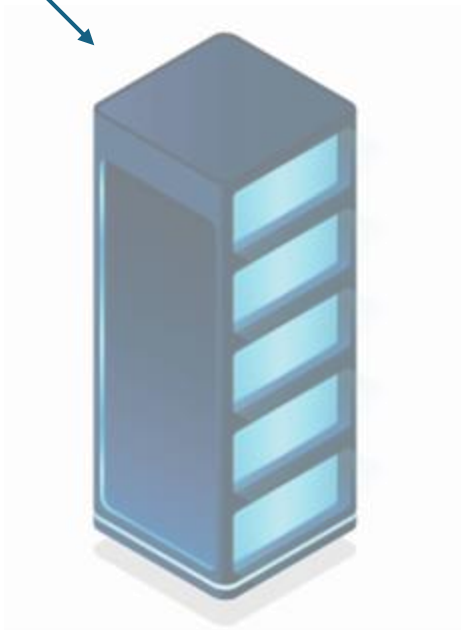
Train ML model on all available data



Sample points for QM simulation

Online model

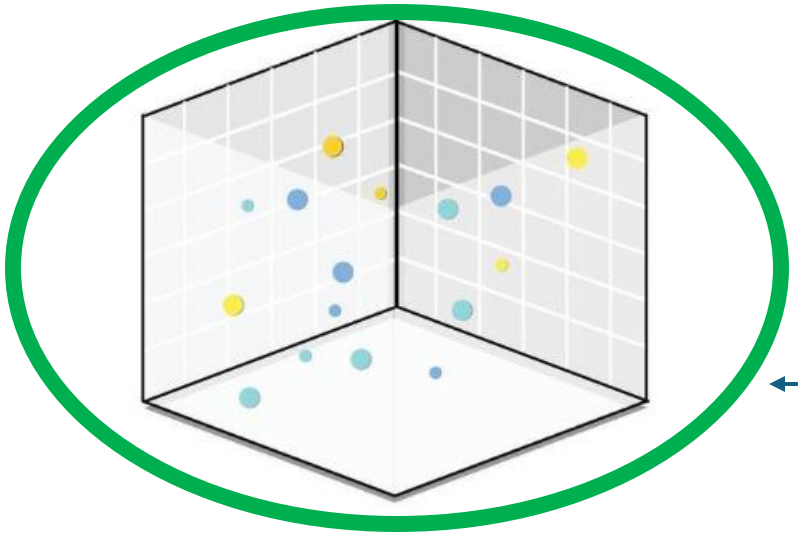




Augment dataset with new data

Train ML model on all available data

Sample points for QM simulation



Open dataset

Augment dataset with new data

The goal/idea of my PhD

- Many areas of reactivity prediction need more accurate models
- Many or these have *sparse data* = can't train ML models
- We develop QM-workflows to generate high-quality data
- Then train hybrid ML models with active learning

Supplementary slides

Workflow expanded