A machine-learned Bond Dissociation Enthalpy model to support predictions of forced degradation

Ponting, David J.; Hemingway, Rachel; Ott, Martin A.

Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS

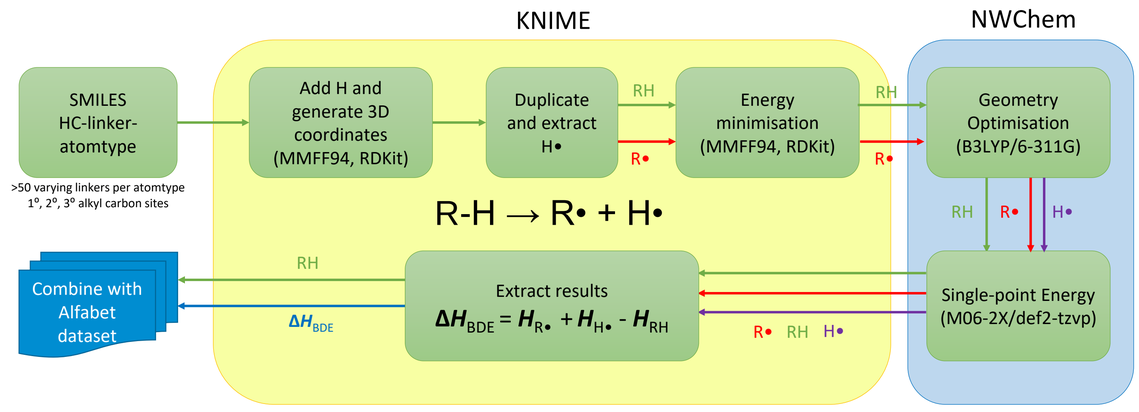
Introduction

The evaluation of forced degradation studies is a key component of drug safety assessments. These studies employ various conditions to stress-test pharmaceutical compounds, such as oxygen (air) and radical initiators to evaluate liability towards oxidative damage. Much of the degradation chemistry that occurs in these tests can be predicted *in silico* using substructural patterns encoded in the degradation prediction tool Zeneth[1] (Lhasa Limited); however, some reactivity cannot be accurately predicted from structural patterns alone since it is dependent on combinations of features that cannot be encoded in a finite number of patterns. Hydrogen abstraction by peroxyl radicals is the chain-propagating step in the homolytic cleavage of C-H bonds within the structure, a common degradation reaction for all molecules stored in air (autoxidation). The strength of the C-H bond being broken, relative to the strength of the O-H bond in a hydroperoxide (*i.e.* the bond being made) is a pivotal parameter to assess this reactivity. This strength can be quantified as the Bond Dissociation Enthalpy (BDE).

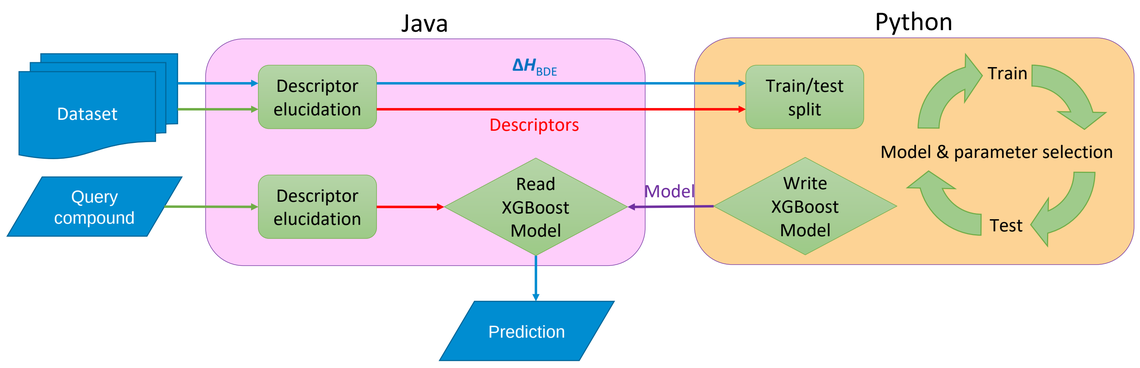
We present a gradient-boosted tree model, based on a dataset of public data[2] which we have enhanced with additional structures with BDEs calculated at a comparable level of theory (M06-2X/def2-TZVP) for additional elements that are of relevance to drug structures (F, Cl, Br, I, P, S) and additional chemical environments. This model achieves predictive performance against an external test set within the accuracy of experimentally measured BDEs[3], and can be used to improve the prediction accuracy of these radical-mediated degradation reactions by comparing the relative BDE to the hydroperoxide O-H bond, using a sigmoid curve to avoid a hard cut-off between reactive and unreactive predictions.

Data Curation and Enrichment

The data generated and made public by the Alfabet[2] (https://bde.ml.nrel.gov/) project was used as an initial starting point for the creation of an in-house model, and subsequently augmented to add additional atom types[4] such as the halogens, sulfur and phosphorus. Restriction to C-H BDEs allows for a simplification of the BDE calculation, since one of the formed radicals is always H•. Calculations can thus be performed on RH and R•.

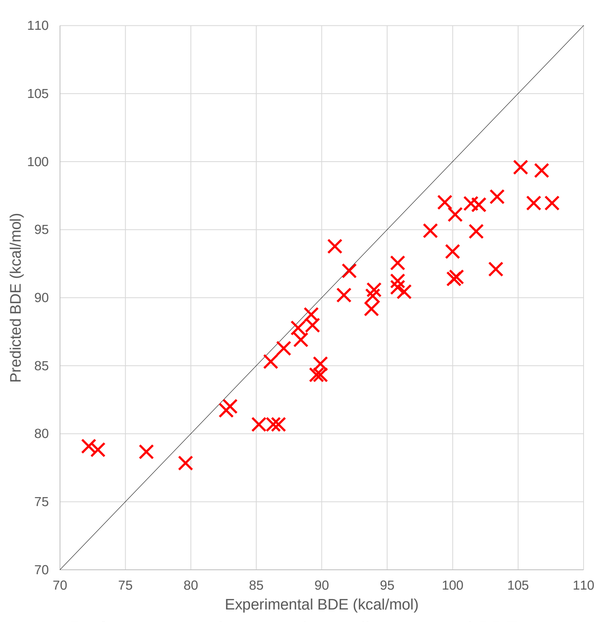


Implementation



Following exploratory approaches with a wide variety of model types and hyperparameters, tree-based models were shown to be most predictive. The descriptors used were a circular fingerprint of atom types, as previously published[4]. These atom types take both the steric and electronic nature of the atom and its substituents into account. The XGBoost library[5] was used in Python (3.7) to develop an exportable model file which can be read into Java to be combined with our internal cheminformatics toolkit, allowing consistency in performance between otherwise-incompatible languages, rather than a model that has been trained in one language having to be extrapolated to another without further validation.

Performance against worst-case test set (tertiary sites) is within typical error of BDE measurements, with MAE of 1.03 kcal/mol, RMSE of 1.82 kcal/mol and R^2 of 0.91.

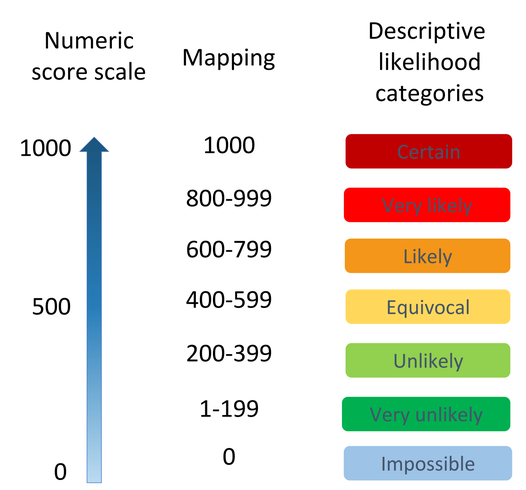


Performance against experimentally-measured BDE values[2]. A slight underprediction is observed, which is both acceptable (additional conservatism) and expected (model is trained on calculated, not experimental, data).

Degradation prediction

Zeneth[1] (currently on version 8.1.1) is a knowledge-based expert system for the prediction of forced degradation. The software incorporates a knowledgebase of transformations, which, if the patterns and conditions match the user input, predicts the structures of potential degradants.

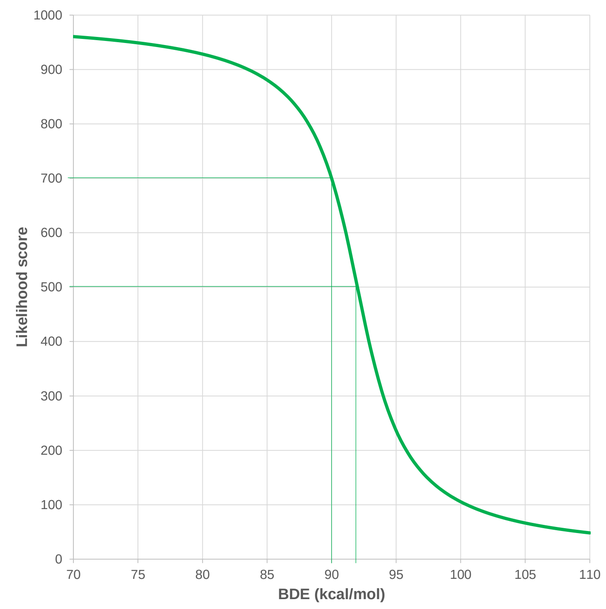
Degradants in Zeneth are associated with likelihood scores, which take the range from 0-1000; these reflect expert understanding of how likely the degradant is to be formed, based on the proportion of similar substructures that are observed to have comparable reactivity (or not).

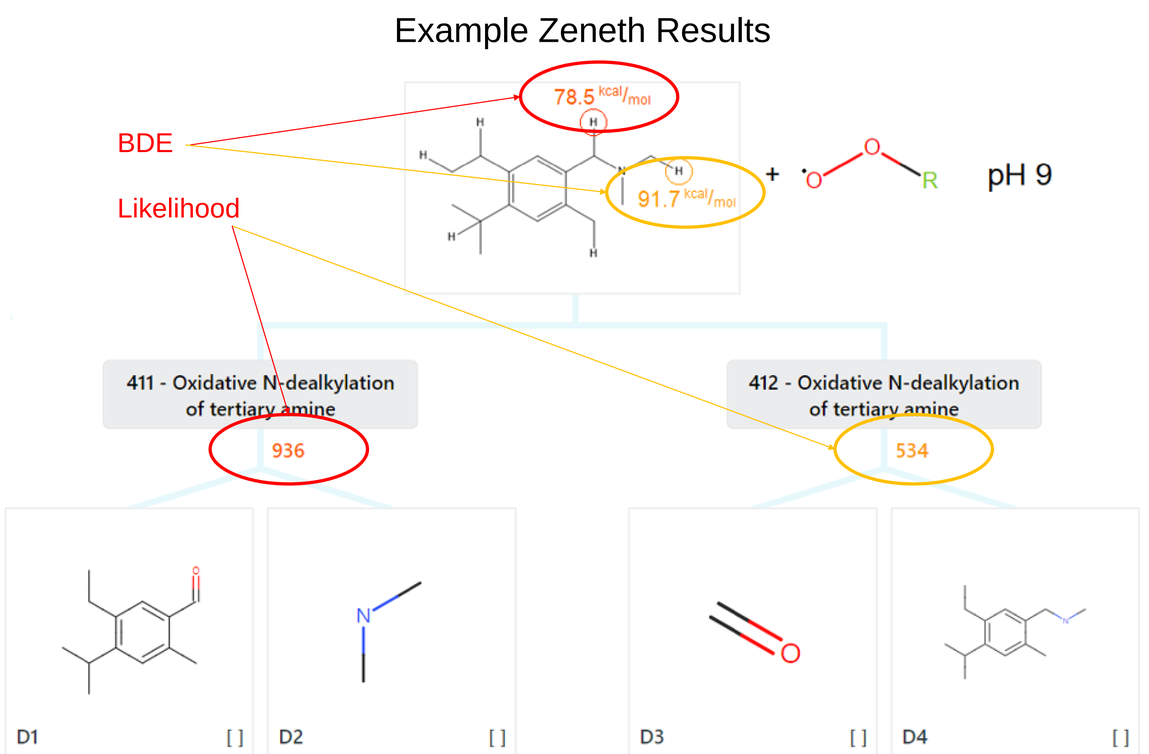


Likelihood modulation

The BDE predictor is used to influence a predicted degradant's likelihood score. Instead of a rigid likelihood for a transformation, such as "likely", likelihoods will vary depending on the strength of the C-H BDE for the relevant bond in the parent structure.
A sigmoid curve was chosen to avoid values too close to 1000 or zero, and parametrized such that a score of 700 ("likely") corresponds to 90 kcal/mol, and 500 ("equivocal") to 92 kcal/mol. This is based on a reported O-H bond strength of 87 kcal/mol.

$$score = 1000\,\frac{\arctan(0.3632713 * (92 - BDE))}{\pi} + 500$$





Conclusions

- Machine learning can be used, given a sufficiently large dataset, to predict BDE with accuracy comparable to experiment.
- Inter-language compatibility of archived ML models enhances repeatability.
- The use of a predictive BDE model can be valuable for refining expert, pattern-based predictions of radical-mediated degradation pathways.

References

1: Parenty *et al* (2013), *Mol. Pharmaceutics*, **10**, 2962-2974
2: St John *et al* (2020), *Nat. Commun*., **11**, 2328; St John *et al* (2020), *Sci. Data*., **7**, 244
3: Lienard *et al* (2015), *Pharm. Res*., **32**, 300-310
4: Plante *et al* (2018), *J. Cheminform.*, **10**, 61
5: Chen & Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794)

shared **knowledge** • shared **progress**