Lhasa Limited
is a developer of expert knowledge based prediction software and chemical databases.

Interpretable Ames mutagenicity predictions using statistical learning techniques

S.J. Webb[ab], P. Krause[b], J.D. Vessey[a]

[a] *Lhasa Limited, 22-23 Blenheim Terrace, Woodhouse Lane, Leeds, LS2 9HD*

[b] *Computing Department, University of Surrey, Guildford, Surrey, GU2 7XH*

Knowledge Transfer Partnerships

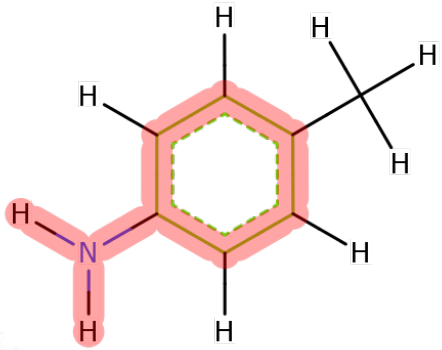
UNIVERSITY OF SURREY

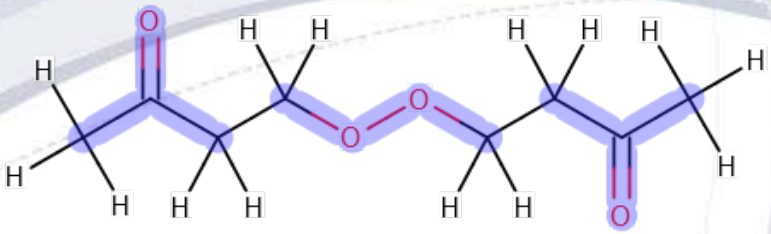
lhasa limited

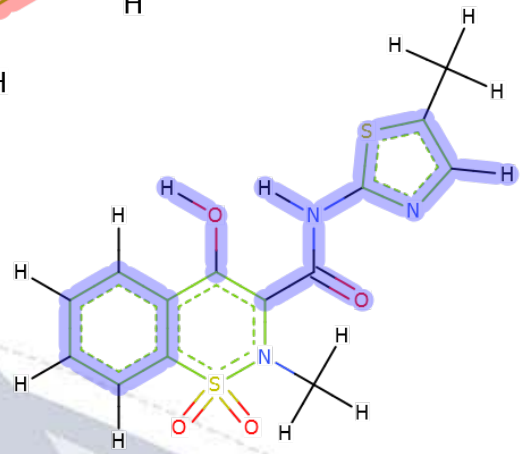
Tel: +44 (0)113 394 6020
Email: info@lhasalimited.org
Web: www.lhasalimited.org

Introduction

- Goal: produce a QSAR methodology for interpretable prediction of Ames mutagenicity providing:
- Confidence
- Cause
- Prediction
- Assessment of domain

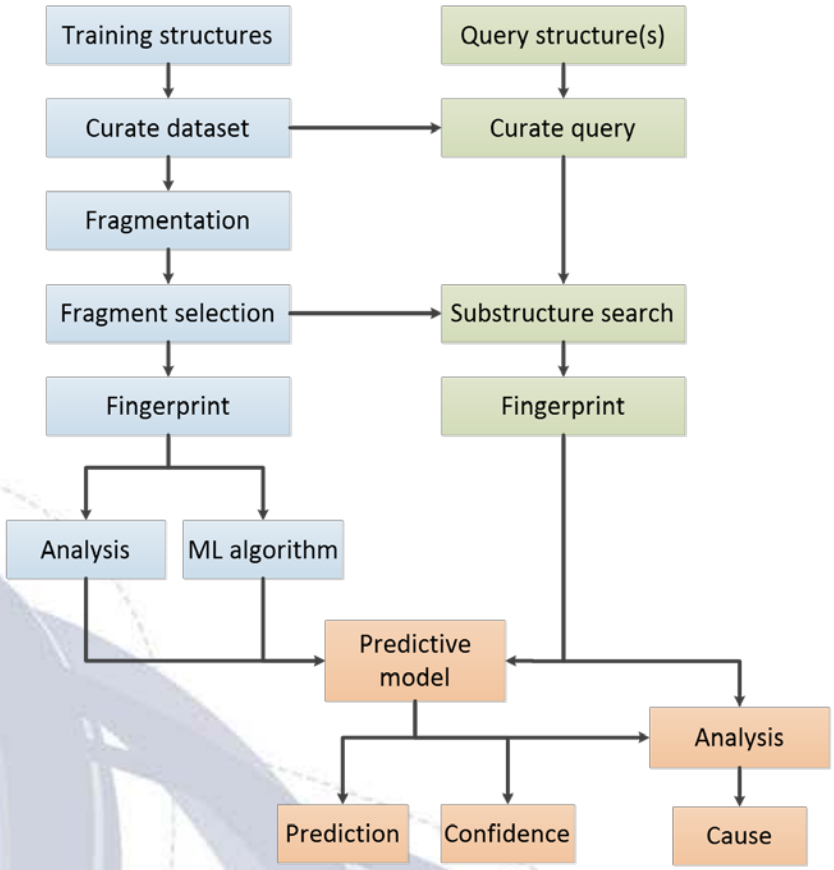






Methodology

- KNIME utilised for managing the workflow
- R integration for machine learning algorithm Random Forest
- New node developments for integration with in house chemical engine
  - Descriptor generation
  - Analysis and confidence
  - Confidence



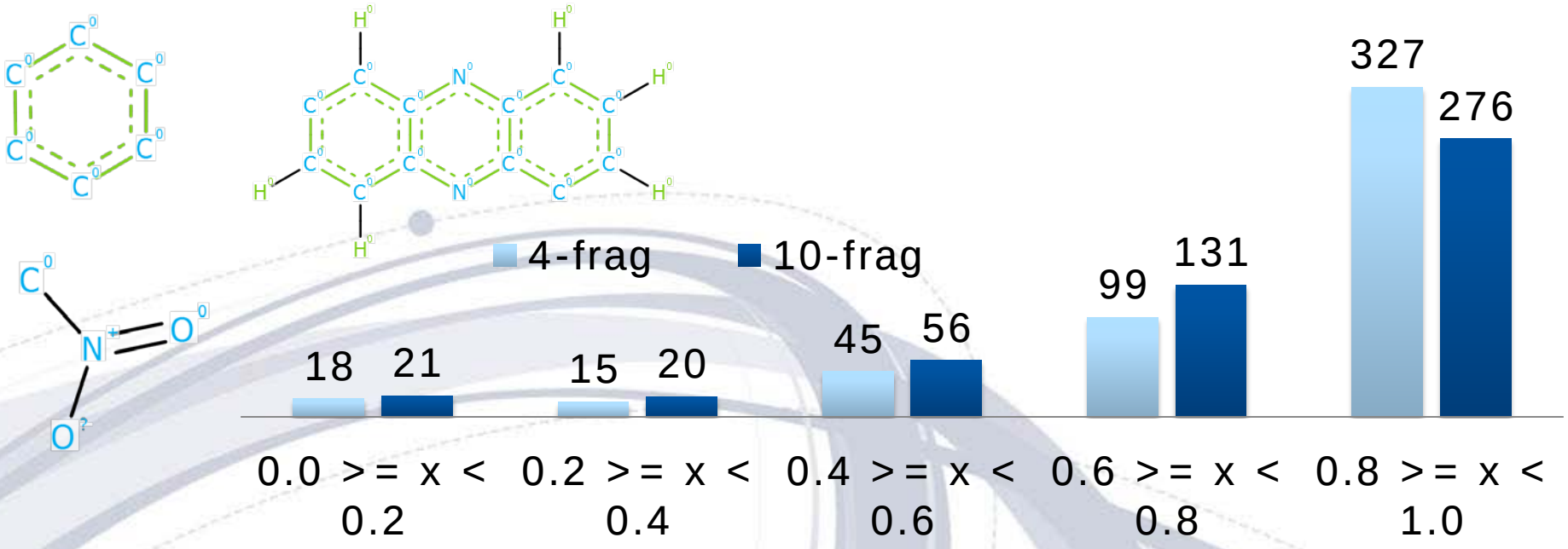
Data and descriptors

- Examples shown within the presentation use a curated version of the benchmark mutagenicity dataset [1] for training utilising a Random Forest
- Validation is performed on an external validation sets consisting of publicly available data not present in the training set as well as in house data
- Structural fragments for the basis of the descriptors and are produced by fragmentation of the training set

[1] Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich and Klaus-Robert Müller *J. Chem. Inf. Model.*, 2009, 49 (9), 2077–2081

Data and descriptors

- This model consists of functional groups, ring substitution patters and ring scaffolds
- Fragments are chosen above a threshold count



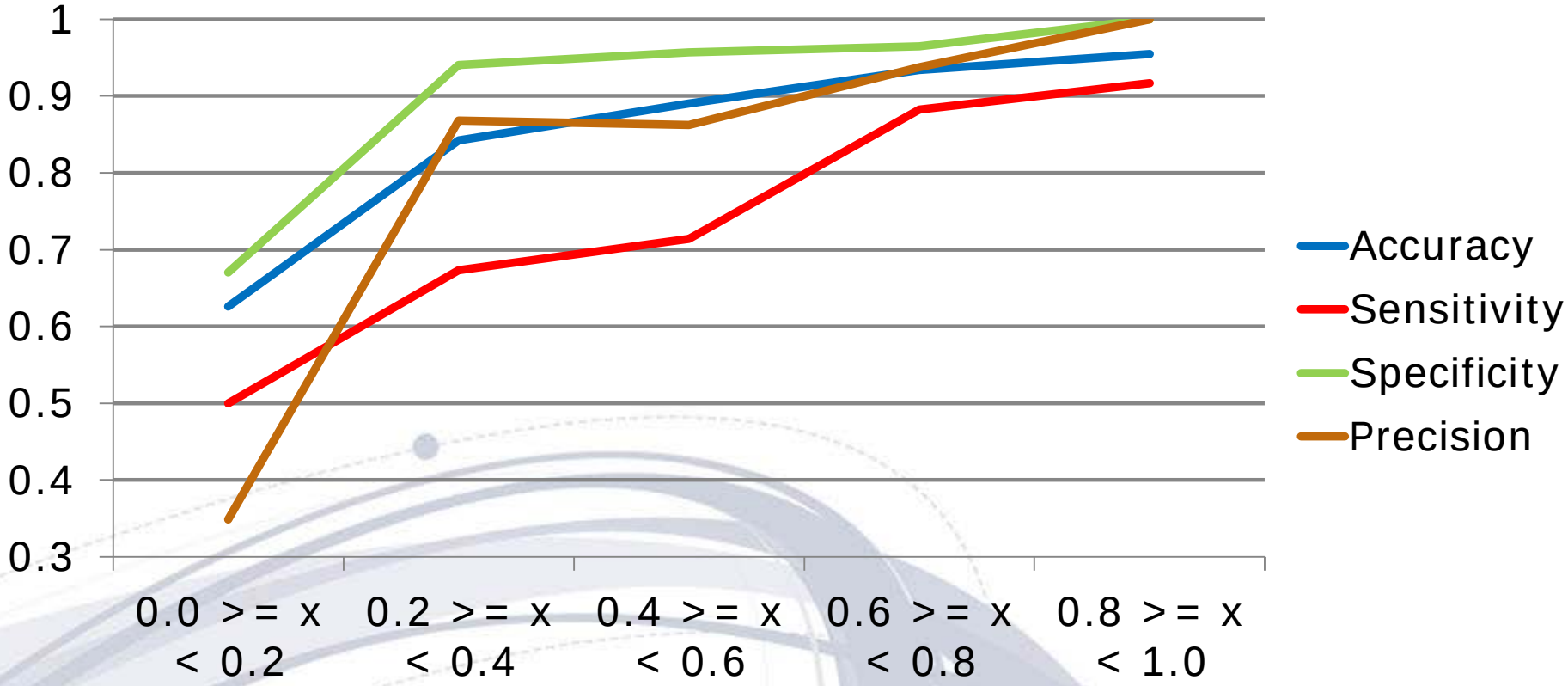
Validation coverage at two fragment count cut-off points

Performance

- Published models for global Ames mutagenicity on public data have a reported accuracy of 80-90%
- Our results for two external validation sets shown

	4 fragment cut-off			
Validation	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
BHN	0.817	0.683	0.884	0.745
Internal	0.769	0.608	0.859	0.706
	10 fragment cut-off			
BHN	0.808	0.653	0.884	0.736
Internal	0.761	0.569	0.868	0.706

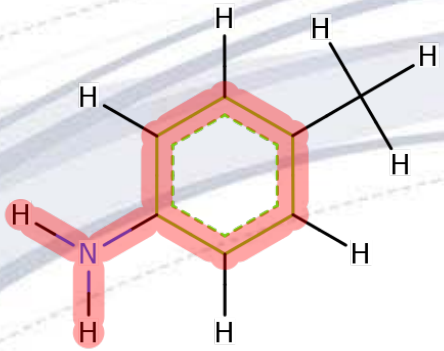
Key results

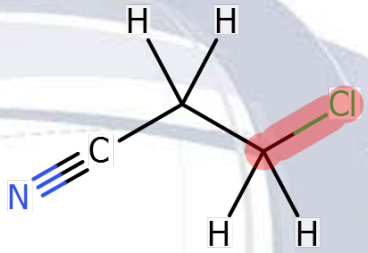


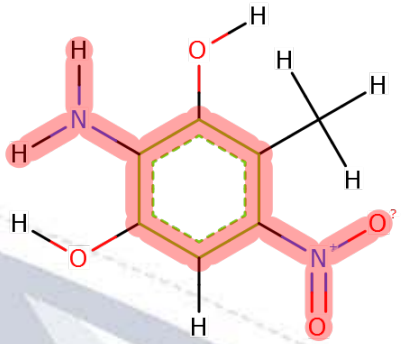
External validation set confidence bin performance for 4 fragment cut-off model using 200 trees

Key results

- The coverage of the model is represented on a fragment level, identifying what parts of a query structure the model is naïve too
- Analysis of the model and the query allows for an assessment of the prediction and an assignment of each known fragments relationship to the prediction





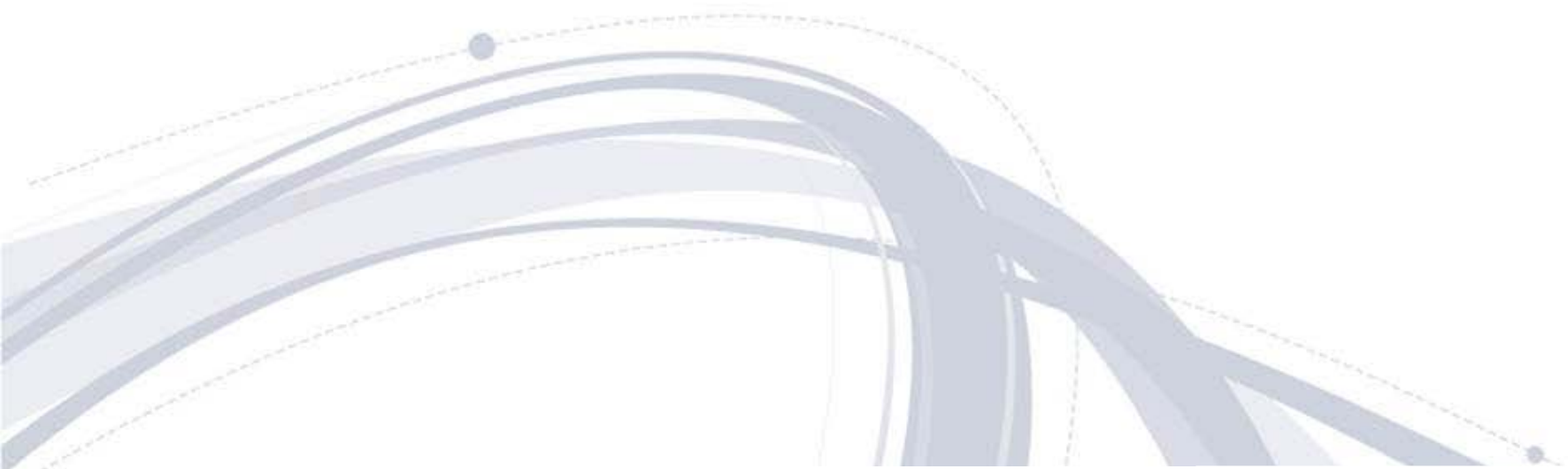


Conclusions

- Simple descriptors have been utilised in a new modelling methodology to produce interpretable predictions for Ames mutagenicity prediction
- Performance remains high, especially for the predictions with a large confidence > 0.6
- Identification of in/out of domain achieved and open to user interpretation, no prediction is thrown out unless a threshold is set
- Method provides information on the training molecules contributing to the prediction

Acknowledgements

- Lhasa research group
- Brendan Howlin (University of Surrey)
- KTP organisation for financial contribution



lhasa limited

Knowledge Transfer Partnerships

UNIVERSITY OF SURREY

Lhasa Limited
is a developer of expert knowledge based prediction software and chemical databases.

Thank you!

samuel.webb@lhasalimited.org

Knowledge Transfer Partnerships

UNIVERSITY OF SURREY

lhasa limited

Tel: +44 (0)113 394 6020
Email: info@lhasalimited.org
Web: www.lhasalimited.org