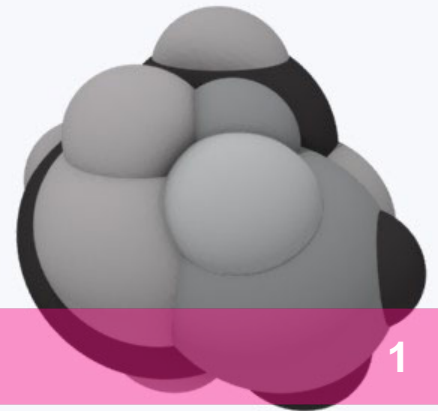


10/2008

New Lazarz Developments

A. Maunz¹⁾
C. Helma^{1), 2)}

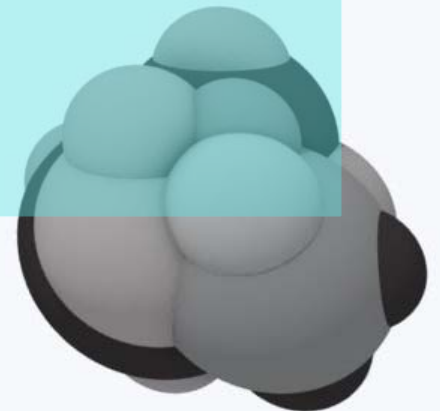
¹⁾FDM Freiburg Univ.
²⁾in silico toxicology





What is Lazar?

INTRODUCTION



Introduction

Lazar is a fully automated SAR system.

- 2D fragment-based (linear, tree-shaped under development)
- Nearest-neighbor predictions (local models)
- Confidence-weighting for single predictions

Applications include highlighting, screening, and ranking of pharmaceuticals.

- In use by industrial corporations. Regulatory acceptance request as alternative test method has been submitted.
- Part of EU research project **OpenTox**
- Public web-based prototype and source code available at: <http://lazar.in-silico.de>

DEFAULT STYLES

Introd

1. Draw a chemical structure ([help](#))

Blockieren

☺ CLR DEL D-R +/- UDO JME

C

N

O

S

F

SGOT increase

SGPT increase

Mutagenicity

Salmonella typhimurium (CPDB)

Salmonella typhimurium (Kazius/Bursi)

Carcinogenicity

Rodent carcinogenicity (multiple sex/species/sites)

Rodent carcinogenicity (single sex/species/site)

Rat carcinogenicity (both sexes)

Rat carcinogenicity (male)

Rat carcinogenicity (female)

Mouse carcinogenicity (both sexes)

Mouse carcinogenicity (male)

Mouse carcinogenicity (female)

Hamster carcinogenicity (both sexes)

Hamster carcinogenicity (male)

Hamster carcinogenicity (female)

IBIS upper bound excess lifetime cancer risk

2. **FDA Maximum Recommended Daily Dose (FDAMDD)**

Maximum recommended daily dose

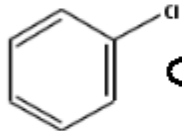
96 hr LC50

3. Predict

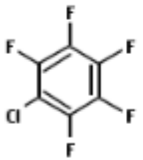
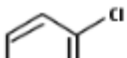


Lazar Toxicity Predictions

Mutagenicity - Salmonella typhimurium (Kazius/Bursi) [Validation and endpoint definition]

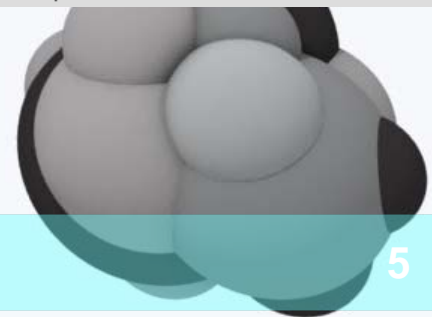
Predicted Activity (Confidence)	Structure	Measured Activity	Additional Information	SMILES InChI
inactive (-0.124294)		inactive	Relevant Fragments DSSTox database PubChem database	Clc1ccccc1 InChI=1/C6H5Cl/c7-6-4-2-1-3-5-6/h1-5H

Similar Structures (10 from 146 Neighbors)

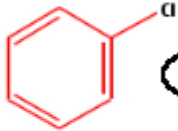
Similarity	Structure	Measured Activity	Additional Information	SMILES InChI
0.92		inactive	Original data (DSSTox) PubChem database	FC1=C(C(=C(C(=C1F)F)Cl)F)F InChI=1/C6ClF5 /c7-1-2(8)4(10)6(12)5(11)3(1)9
0.92		inactive	Original data (DSSTox)	ClC1=CC=C(C=C1)Cl

...more neighbors

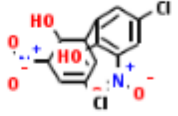

DEFAULT STYLES



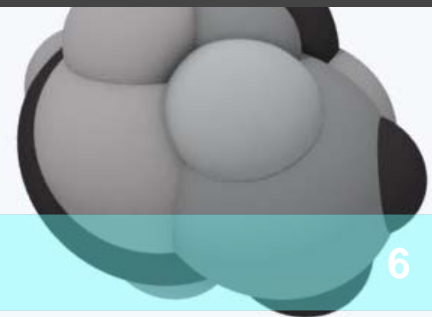
FDA Maximum Recommended Daily Dose (FDAMDD) - Maximum recommended daily dose [Validation and endpoint definition]

Predicted Activity (Confidence)	Structure	Measured Activity	Additional Information	SMILES InChI
0.00476 milimol (0.055391) low confidence (<0.2)		not available	Relevant Fragments DSSTox database PubChem database	<chem>Clc1ccccc1</chem> InChI=1/C6H5Cl/c7-6-4-2-1-3-5-6/h1-5H

Similar Structures (5 from 5 Neighbors)

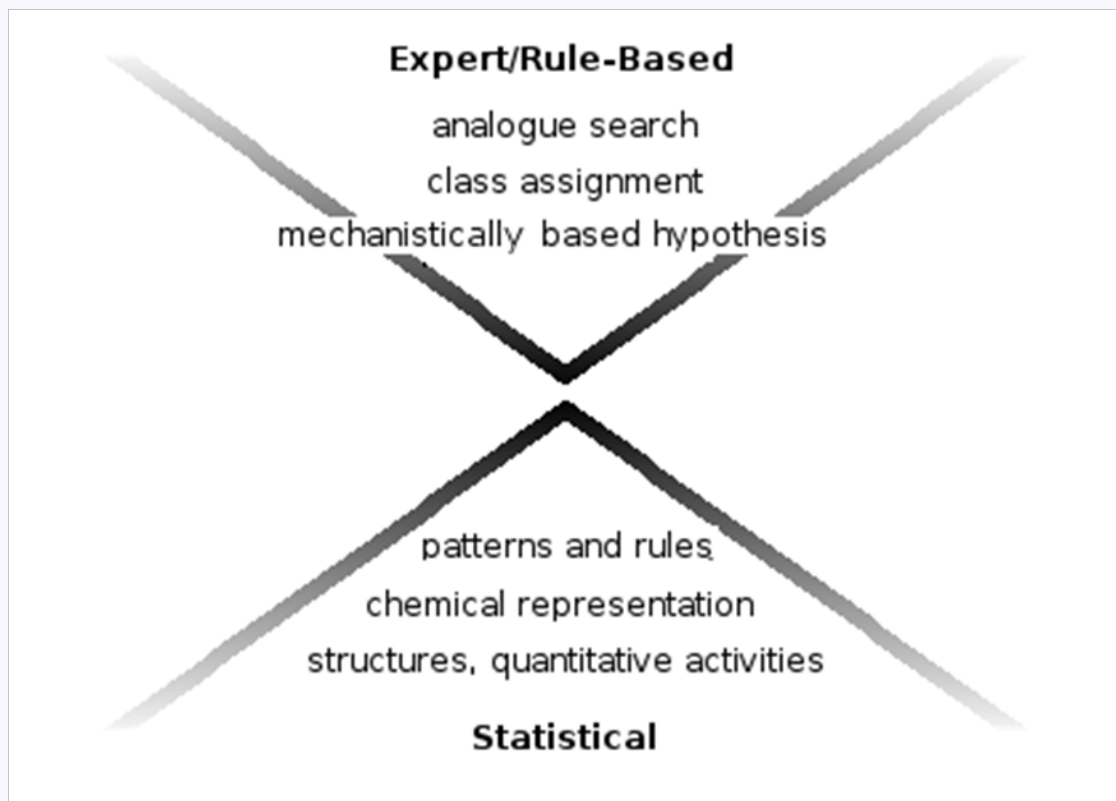
Similarity	Structure	Measured Activity	Additional Information	SMILES InChI
0.43		0.00580 milimol	Original data (DSSTox) PubChem database	<chem>Clc1cc(c(O)c(c1)c2cc(Cl)cc(c2O)[N+](=[O-])=O)[N+](=[O-])=O</chem> InChI=1/C12H6Cl2N2O6 /c13-5-1-7(11(17)9(3-5)15(19)20)8-2-6(14)4-10(12(8)18)16(21)22/h1-4,17-18H
		0.000174	Original data (DSSTox)	<chem>C2=CC=C(Cl)C(N=C1NCCN1)=C2Cl</chem>

...more neighbors

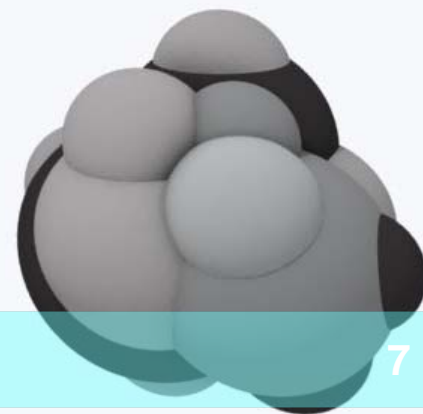


Introduction

Lazar is completely data-driven, no expert knowledge is needed.



DEFAULT STYLES



Introduction

Similarity for a specific endpoint.

Every fragment f has an assigned p -value p_f indicating its significance. Similarity of compounds x, y is the **p -weighted ratio of shared fragments**:

$$sim(x,y) = \frac{\sum \{gauss(p_f) \mid f \subseteq x \wedge f \subseteq y\}}{\sum \{gauss(p_f) \mid f \subseteq x \vee f \subseteq y\}} .$$

Note that this equals the standard Tanimoto similarity if all p -values are set to 1.0.

DEFAULT STYLES

Introduction

Confidence index

For every prediction, calculate the confidence as:

$$conf = gauss(\bar{s})$$

\bar{s} : median similarity of neighbors (to the query structure)

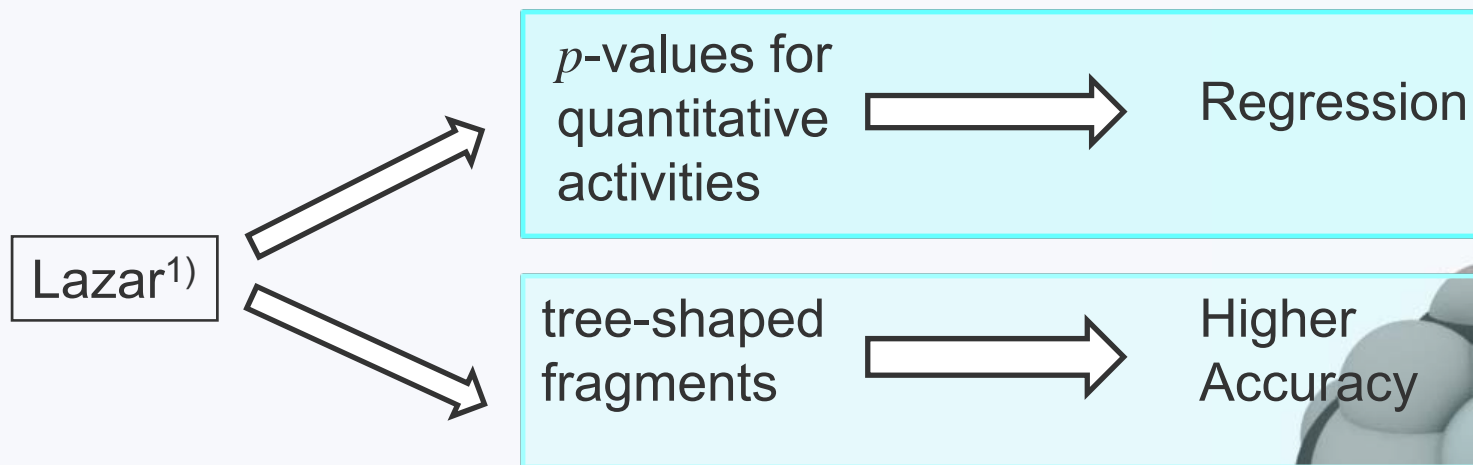


Introduction

Features and p -values

Similarity concept is vital for nearest-neighbor approaches.

- Weight of neighbor contribution to the prediction
- Confidence for individual predictions



¹⁾C. Helma (2006): "Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and Salmonella Mutagenicity", *Molecular Diversity*, 10(2), 147–158

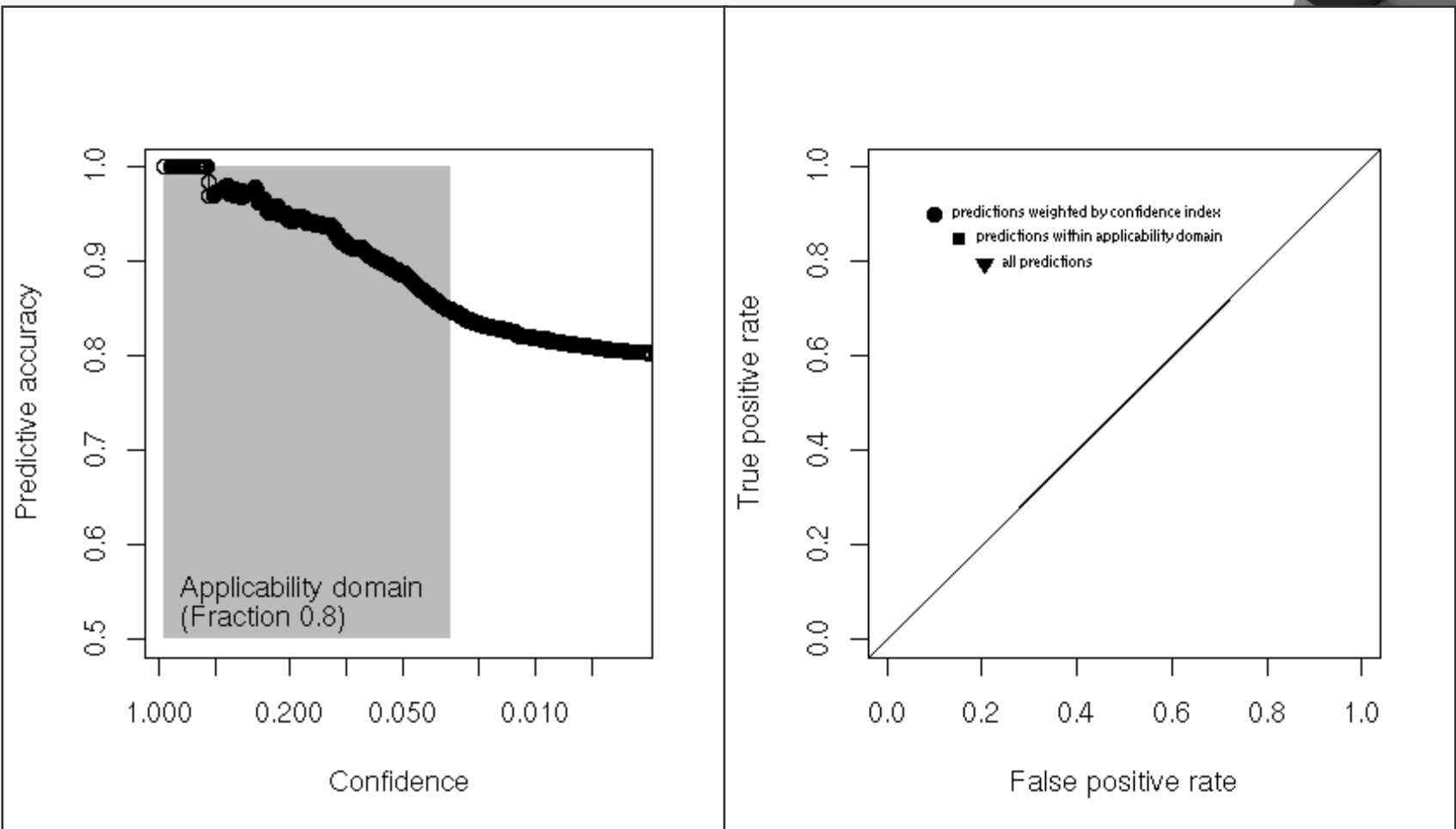
Introduction

To date, Lazar has been a classifier for binary endpoints only.

Published results include:

Dataset	Weighted accuracy
Kazius Salmonella Mutagenicity ²⁾	90%
CPDB Multicell Call	81%

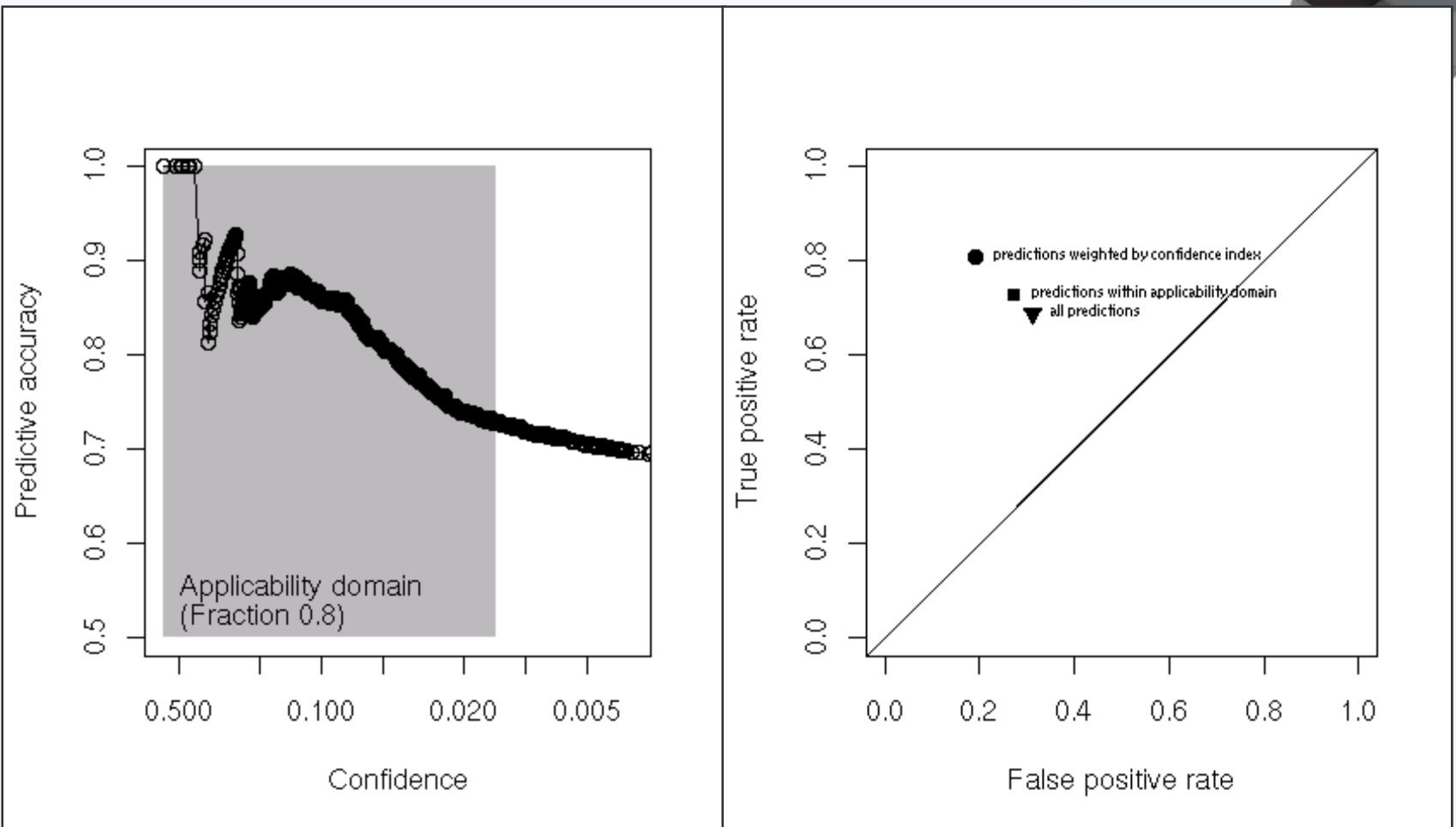
²⁾Kazius, J., Nijssen, S., Kok, J., Back, T., & IJzerman, A.P. (2006): "Substructure Mining Using Elaborate Chemical Representation", J. Chem. Inf. Model., 46(2), 597 - 605



Kazius: Salmonella Mutagenicity

Left: Confidence vs. true prediction rate

Right: ROC analysis



CPDB: Multicell Call

Left: Confidence vs. true prediction rate

Right: ROC analysis

Extension by

QUANTITATIVE PREDICTIONS



Quantitative Predictions²⁾

- Enable prediction of quantitative values (regression)
 - p -values as determined by KS test:
- Support vector regression on neighbors
 - Activity-specific similarity as a kernel function:
 - Superior to Tanimoto index
- Standard deviation of neighbor activities influence confidence
 - Applicability Domain estimation: based on new confidence values considering dependent **and** independent variables
 - Gaussian smoothed

²⁾ Maunz, A. & Helma, C. (2008): "Prediction of chemical toxicity with local support vector regression and activity-specific kernels", SAR and QSAR in Environmental Research, 19 (5), 413-431.

Quantitative Predictions

Confidence index

For every prediction, calculate the confidence as:

$$conf = gauss(\bar{s}) \cdot e^{-\sigma_a}$$

\bar{s} : median similarity of neighbors (to the query structure)

σ_a : standard deviation of neighbor's activities

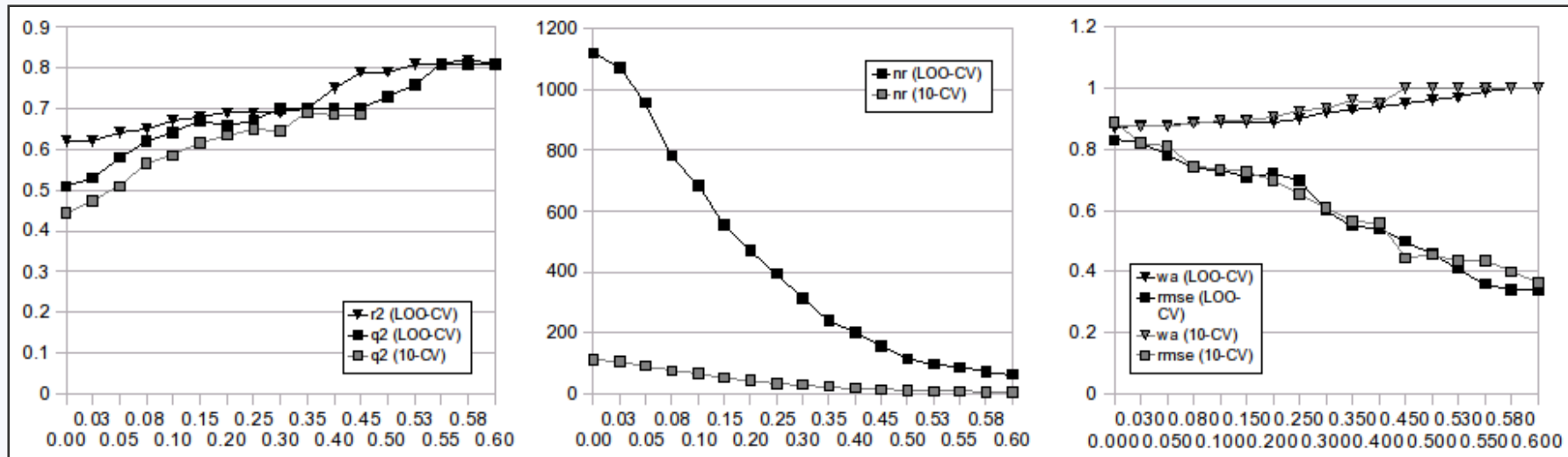
Quantitative Predictions

Validation results on DSSTox project data include:

- EPAFHM Fathead Minnow Acute Toxicity (*Ic50 mmol*, 573 compounds)
- FDAMDD Maximum Recommended Therapeutic Dose based on clinical trial data (*dose mrdd mmol*, 1215 pharmaceutical compounds)
- IRIS Upper-bound excess lifetime cancer risk from continuous exposure to 1 $\mu\text{g/L}$ in drinking water (*drinking water unit risk micromol per L*, 68 compounds)

Previously, FDAMDD and IRIS were not included in (Q)SAR studies.

Quantitative Predictions



Predictivity

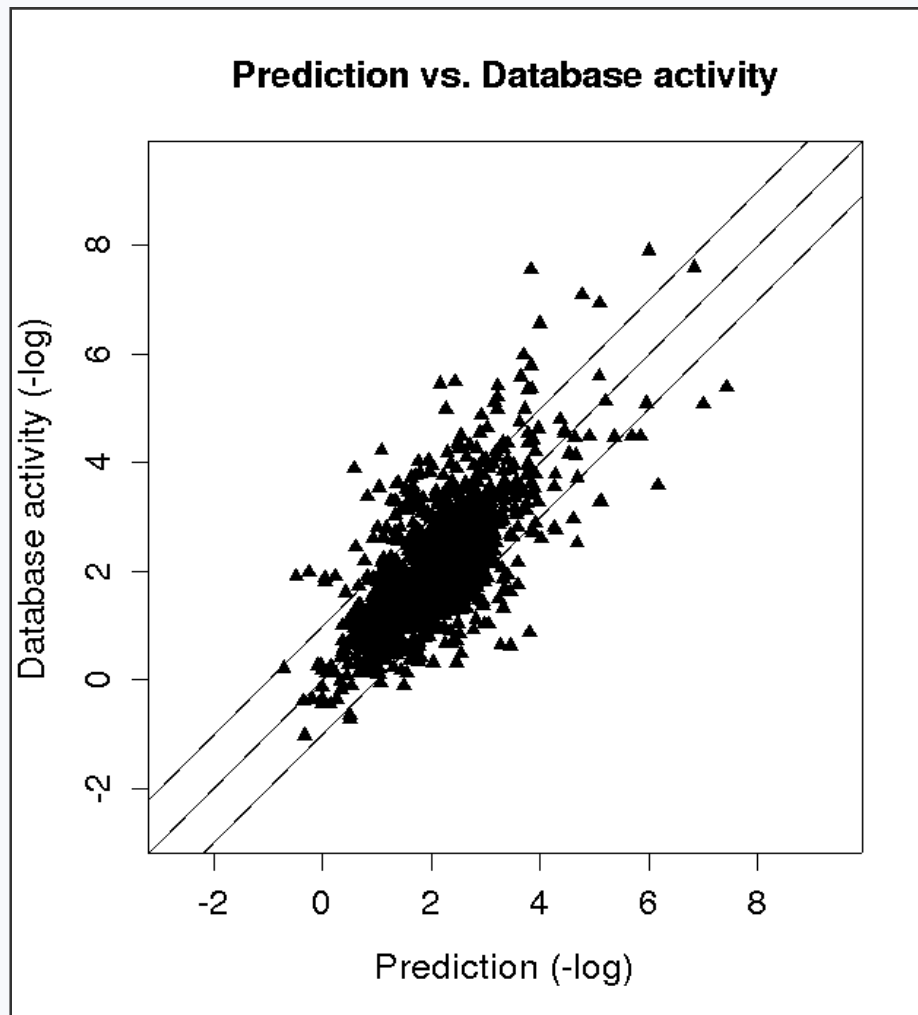
Number of Predictions

RMSE / Weighted accuracy

Validation (FDAMDD)

Effect of confidence levels 0.0 to 0.6

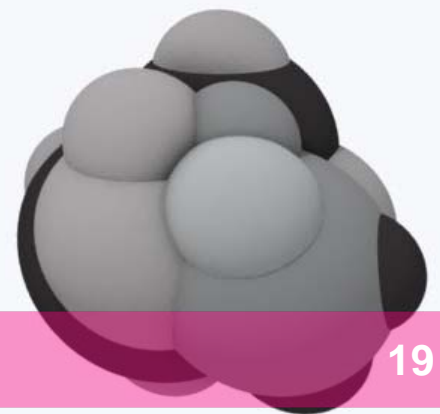
Step: 0.025



Applicability Domain (FDAMDD)

Effect of confidence levels 0.0 to 0.6

Step: 0.025





Mechanistic descriptors

BACKBONE MINING



Better descriptors




Ideal: Few highly descriptive patterns that are easy to mine and allow for mechanistical reasoning in toxicity predictions.

We have been using linear fragments as descriptors. Better descriptors would

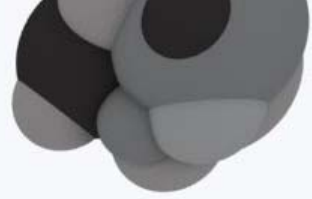
- consider stereochemistry and include branched substructures
- be less intercorrelated and fewer in numbers
- be better correlated to target classes

Tree-shaped fragments

Problem: ~80% of subgraphs in typical databases are trees!
A method to cut down on correlated fragments is needed.



Backbones and classes



The backbone of a tree is defined as its longest path with the lexicographically lowest sequence. Each backbone identifies a (disjunct) set of tree-shaped fragments that grow from this backbone.

Definition:

A Backbone Refinement Class (BBRC) consists of tree refinements with identical backbones.

Example on next slide



BBR classes (1 step Ex.)



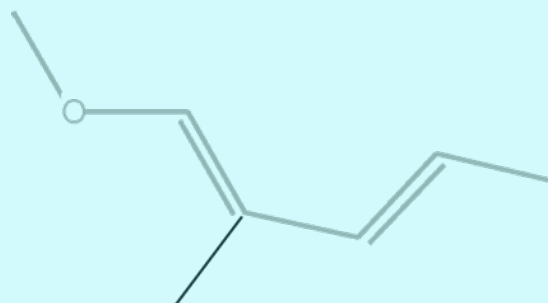
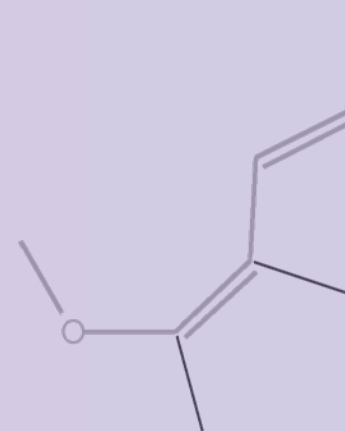
C-C(-O-C)(=C-c:c:c)

Refinement

Backbone:
c:c:c-C=C-O-C

Refinement

C-C(=C(-O-C)(-C))(-c:c:c)



C-C(=C-O-C)(-c:c:c)

Class 1

Class 2

BBR classes

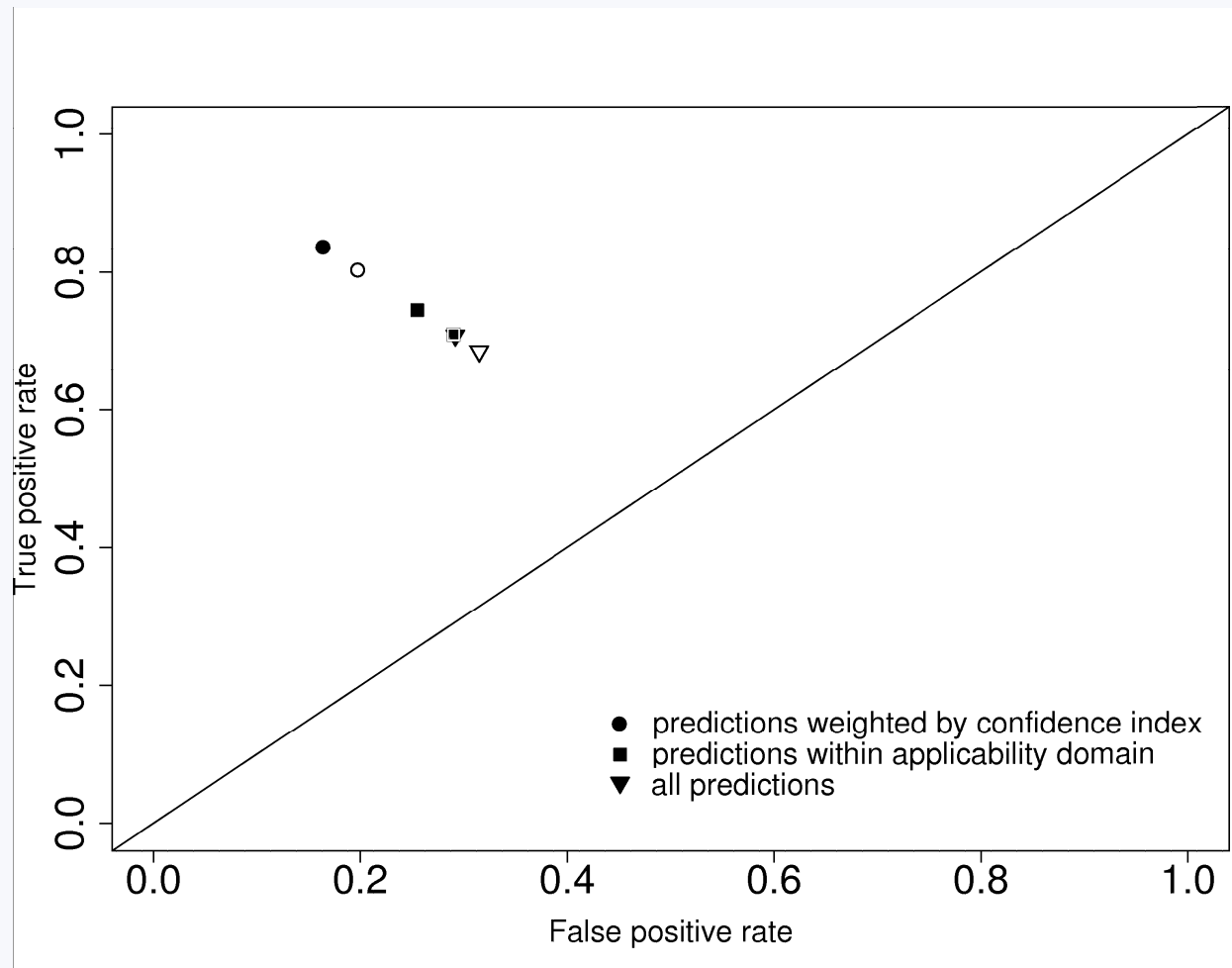
Idea: Represent each BBRC by a single feature

- We use a modified version of the graph miner Gaston³⁾.
 - Double-free enumeration through embedding lists and canonical depth sequences
- Our extension:
 - Efficient mining of most significant BBRC representatives
 - Supervised refinement based on χ^2 values by statistical metrical pruning⁴⁾ and dynamic upper bound adjustment

³⁾Nijssen S. & Kok J.N.: “A quickstart in frequent structure mining can make a difference”, KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA: ACM 2004: 647–652.

⁴⁾Bringmann B., Zimmermann A., de Raedt L., Nijssen S.: “Dont be afraid of simpler patterns”, Proceedings 10th PKDD, Springer-Verlag 2006: 55–66.

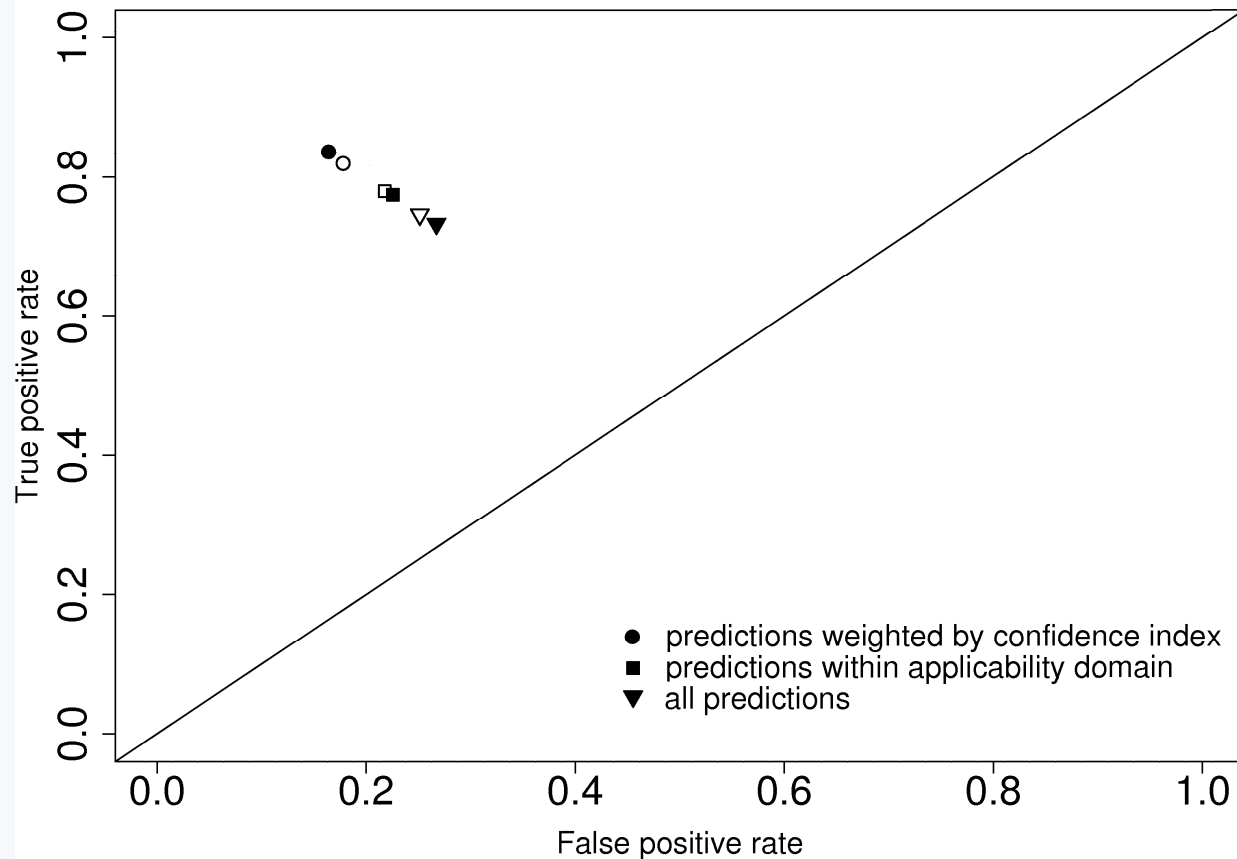
BBRC validation



CPDB
Multicell call
Comparison to
linear fragments

Filled: BBRC representatives
Hollow: linear fragments

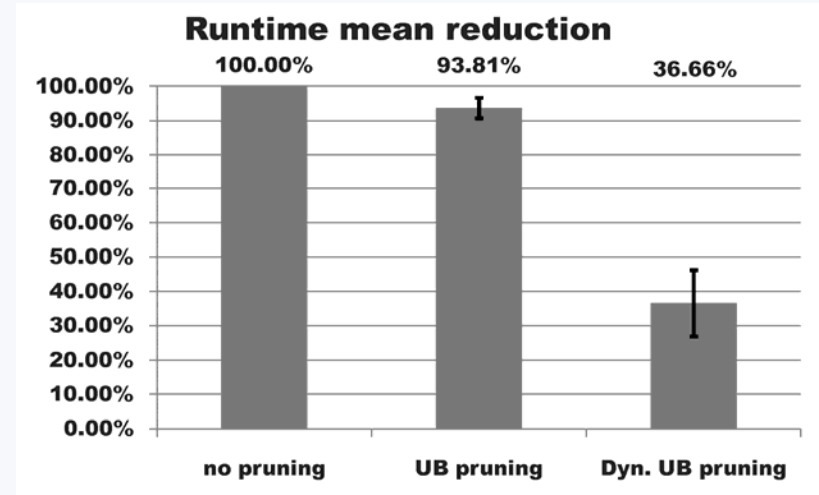
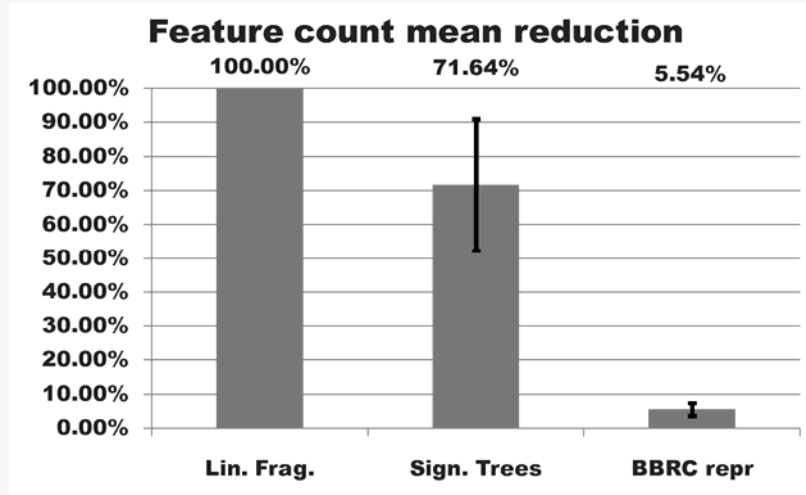
BBRC validation



CPDB
Salmonella
mutagenicity
Comparison to
linear fragments

Filled: BBRC representatives
Hollow: linear fragments

BBRC validation



Remarks:

- Linear fragments prev. used in Lazar
- Minimum frequency: lin. frag. 1, trees 6
- Minimum correlation: lin. frag. none, trees $p_{\chi^2} > 0.95$
- BBRC representatives
- time as measured on a lab workstation





SUMMARY



Summary

Lazar for quantitative predictions

- Reliable, automatic Applicability Domain estimation for individual predictions
- Includes both dependent and independent variables
- Significance-weighted kernel function

Summary

Backbone refinement classes

(*Salmonella* mutagenicity)

- Highly heterogeneous
- Suitable for identification of structural alerts
- 94.5 % less fragments compared to linear fragments
- Mining time reduced by 84.3 %
- Descriptive power equal or superior to that of linear fragments

Acknowledgements

Ann M. Richards (EPA)

Jeroen Kazius (Leiden Univ.)

Thank you!

