The
University
Of
Sheffield.

In-silico approaches to toxicity prediction

Val Gillet
University of Sheffield

Outline

- Background

- In silico methods for toxicity prediction
  - QSAR
  - Machine learning methods
  - Expert systems

- Use of emerging pattern mining to assist knowledge-workers in building the knowledge-base of an expert system

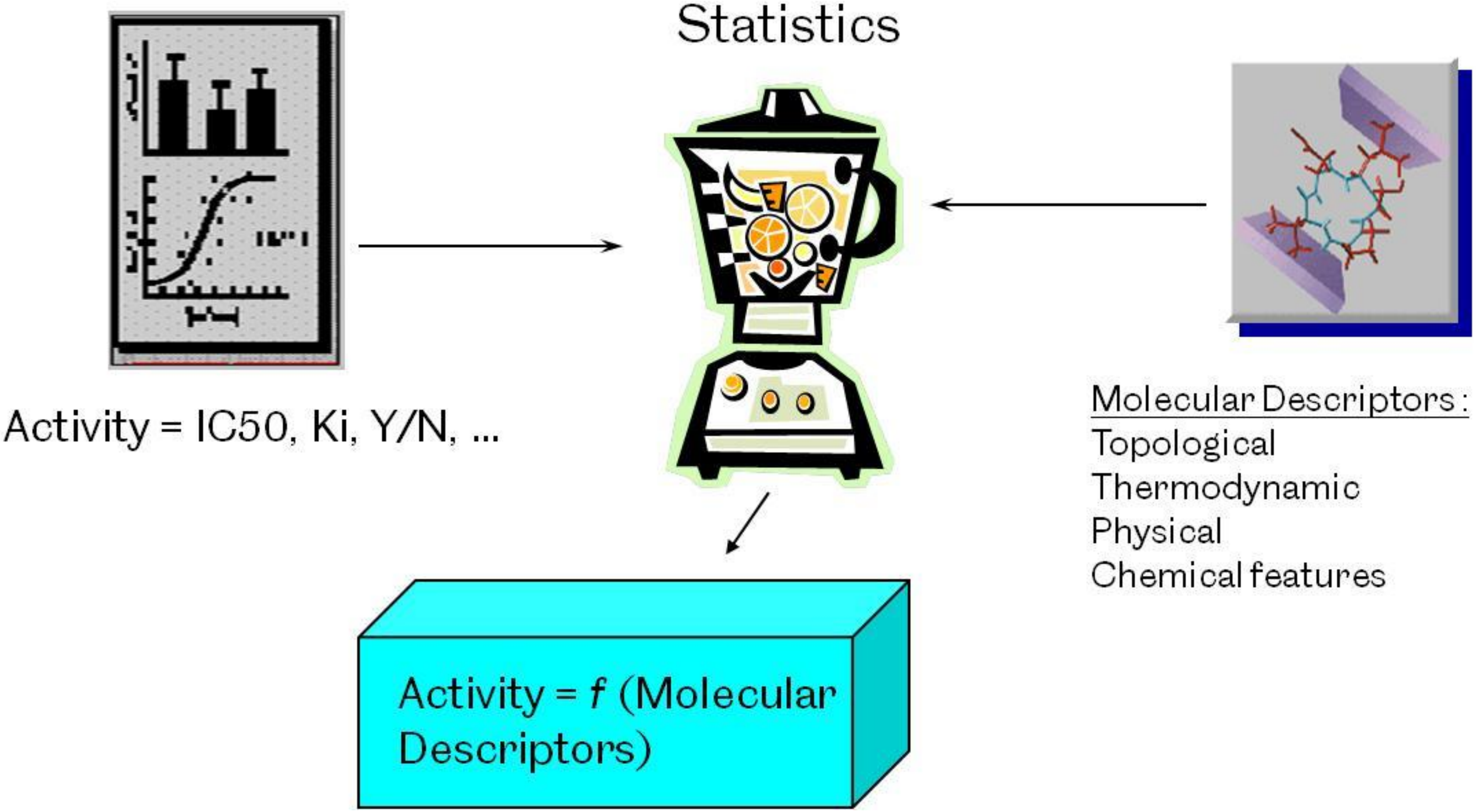
Toxicity prediction

- Avoid late stage failures in drug discovery
- Large numbers of compounds available early in drug discovery and not possible to test all
- In-silico prediction: low cost high-throughput process
  - Can be used to prioritise compounds
  - Highlight potential problems with compounds
  - Allows predictions to be made on virtual compounds as well as real compounds
  - Lead to a reduction in in-vivo tests

Toxicity prediction

- Multiple different endpoints exist
- The same endpoint can arise through multiple mechanisms
- For many endpoints, such as carcinogenicity, the mechanisms are poorly understood
- Lack of availability of reliable data

Statistical methods: QSAR



Statistics

Activity = IC50, Ki, Y/N, ...

Molecular Descriptors:
Topological
Thermodynamic
Physical
Chemical features

Activity = *f* (Molecular Descriptors)

Training set is used to develop a model of activity

Molecular descriptors

- Many thousands of descriptors

- Physicochemical properties
  - ClogP, MW, MR, PSA, ...
- 2D descriptors
  - based on the connection table
  - unweighted (MACCS eg count of the number of acids)
  - deterministic
- 3D descriptors
  - based on geometric patterns of features
  - partially subjective

Handbook of Molecular Descriptors
Roberto Todeschini, Viviana Consonni, Wiley-VCH, 2009

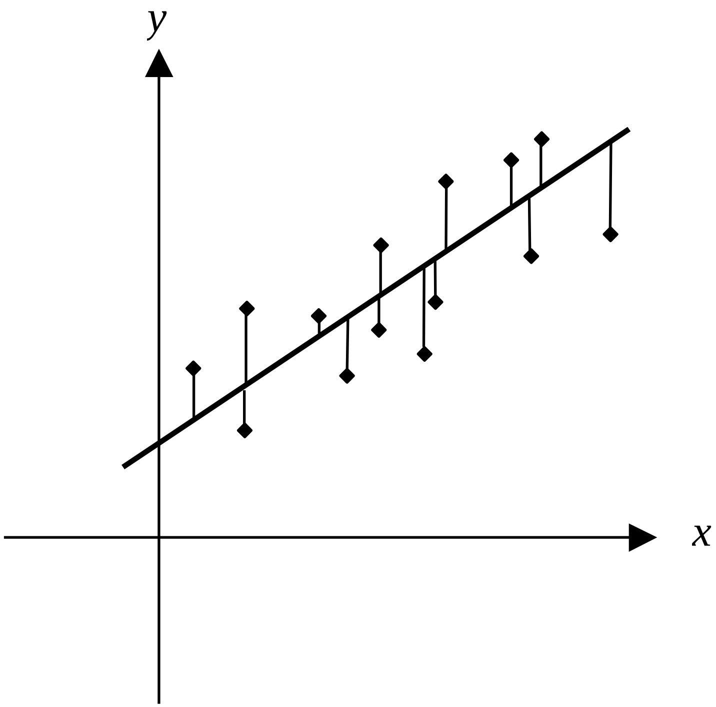
Linear Regression

- Requirements
  - Congeneric series of compounds as training set
  - High degree of similarity in structures

$$y = mx + c$$

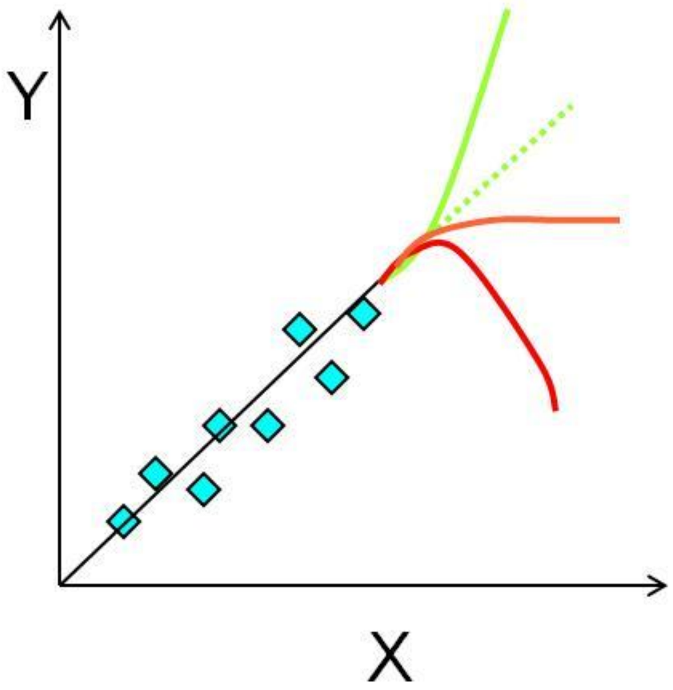
y is the dependent variable (activity)
x is the independent variable eg a molecular descriptor

Aim is to find *m* and *c* to minimise differences in predicted values and actual values



Extrapolation?

- Choose the training set with care
- The model explains the data it was trained on (r^2)
- Validate the model (q^2, pred r^2)
- Can only reliably predict for compounds that are similar to those in the training set
- Local vs global models



Muster W, Breidenbach B, Fischer H, Kirchner S, Mueller L, Pahler A. Computational toxicology in drug development. Drug Discovery Today 13, 2008, 303-310

Machine learning methods

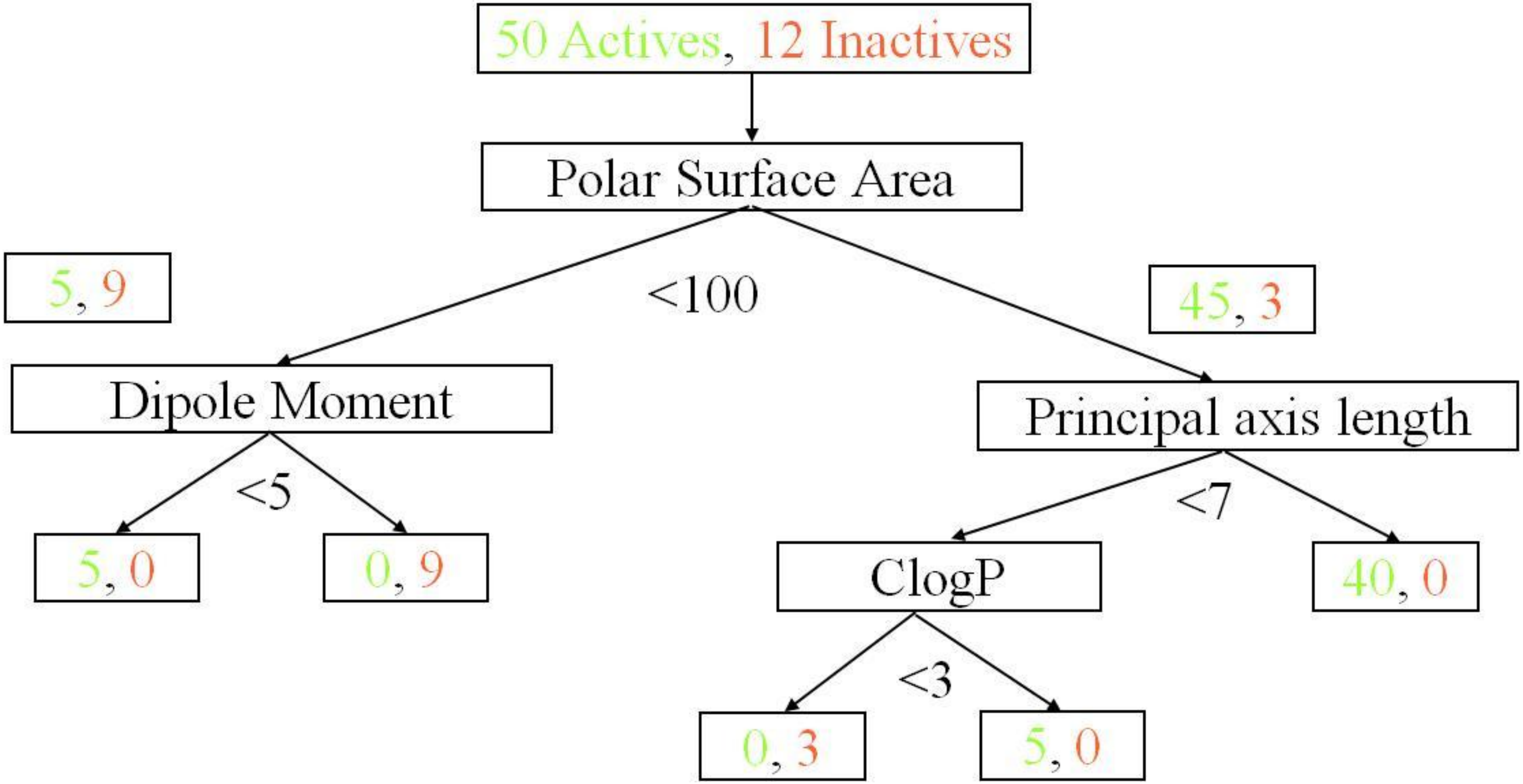
- Training set is used to develop a model of activty
- Can be used with more heterogeneous datasets
- Qualitative or quantitative predictions are possible
- Many different approaches
  - Substructural analysis
  - Recursive partitioning
  - Support vector machines
  - K nearest neighbours
  - Neural networks

Recursive Partitioning

- Classification approach that constructs a decision tree from qualitative data
  - active/inactive, soluble/insoluble, toxic/non-toxic
- Identification of a rule that gives the best statistical split into classes, with the lowest rate of misclassification
  - Example drug|non-drug: MW < 500|MW > 500
- Repeat on each set coming from the previous split until no more reasonable splits can be found
- Can generate good models but with poor predictive power if used without care
  - Use leave-many-out strategies to validate
  - Easy to interpret/drive what-next decisions

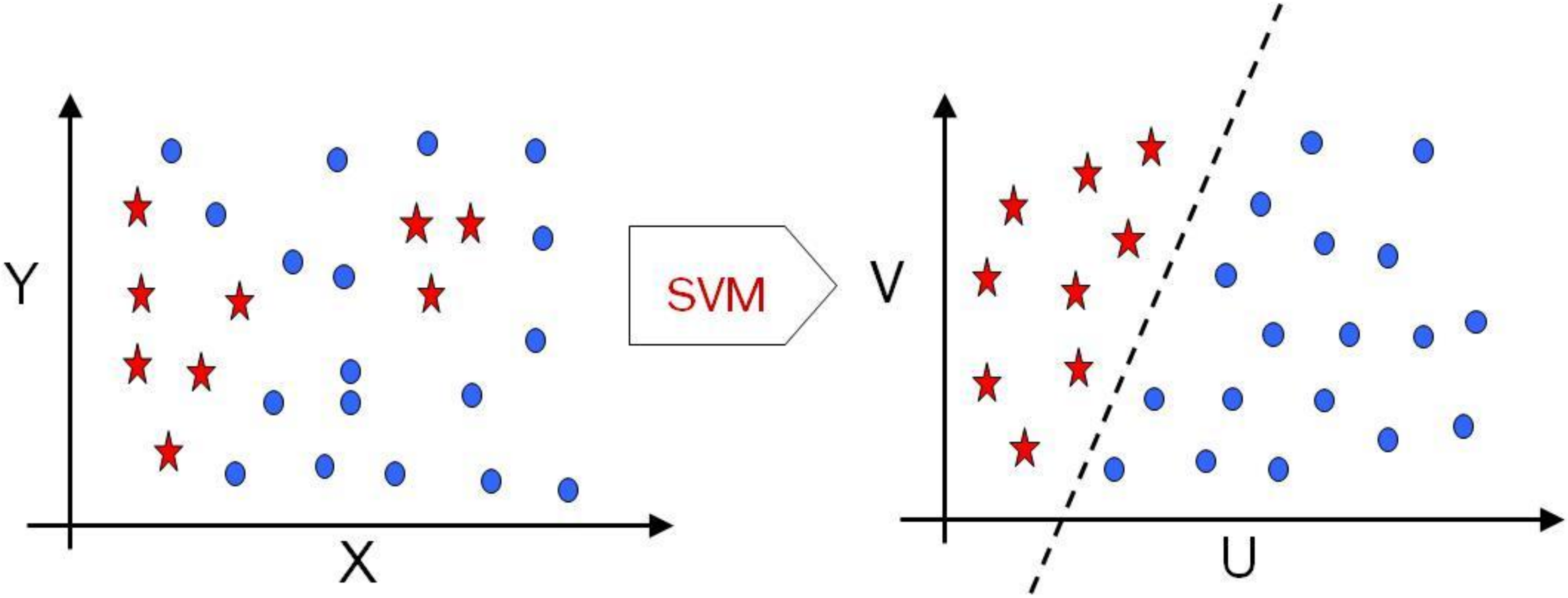
Hamman F, Gutmann H. Voigt N, Helma C, Drewe J. Prediction of adverse drug reactions using decision tree modeling. Clin Pharmacol Ther, 2010, 88, 52-59.

Example



Test compounds are dropped through the tree. Prediction depends on whether they fall into “active” or inactive nodes”

Support Vector Machines (SVMs)

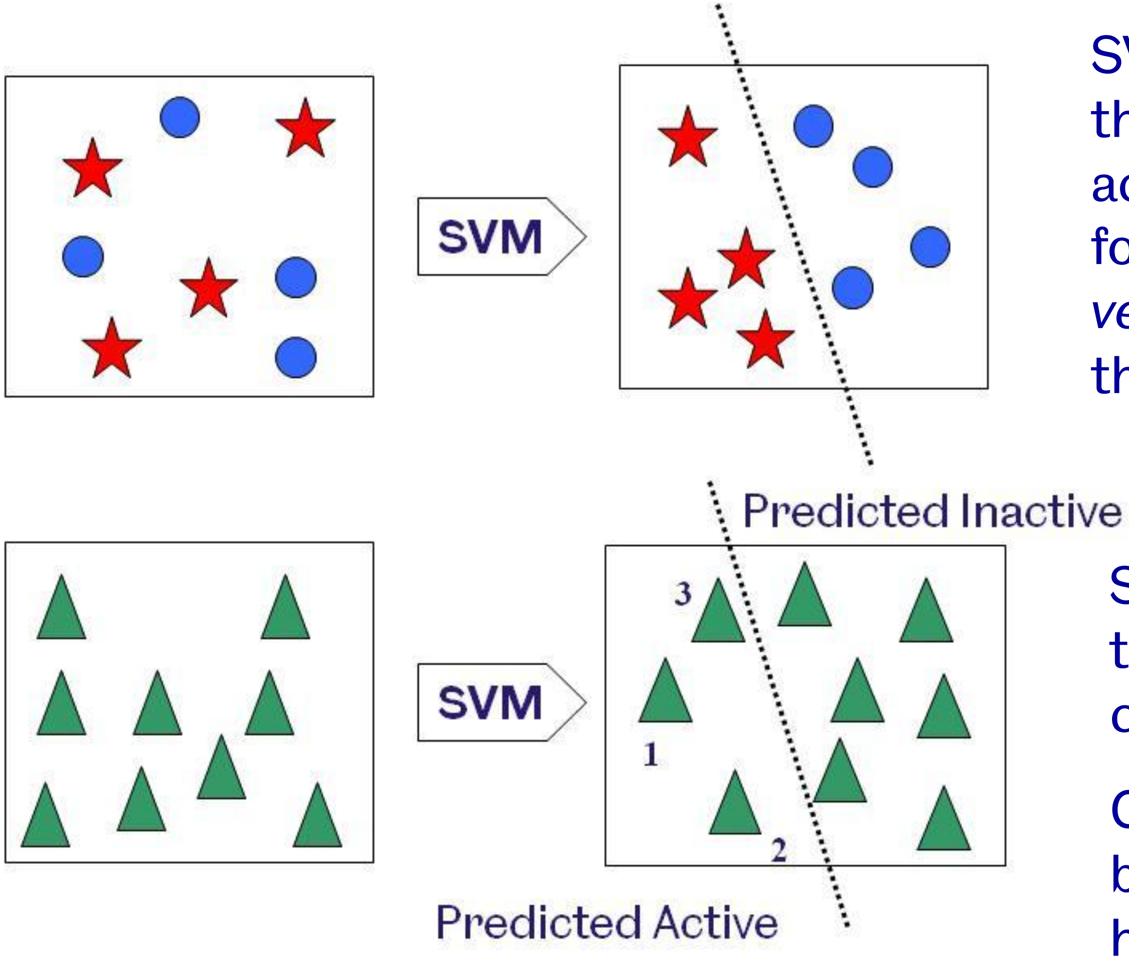


N Dimensions

N+ Dimensions

SVM transforms data into a, usually higher dimensional, space where the actives and inactives are separated by a hyperplane

Applying an SVM model



SVM finds a transformation for the training set that separates actives from inactives, focussing on the *support vectors* near the borders of the two classes

SVM performs same transformation on untested compounds

Compounds can be ranked by distance from the hyperplane

Fourches D, Barnes JC, Day NC, Bradley P, Reed JZ, Tropsha A. Cheminformatics analysis of assertions mined from literature that describe dug-induced liver injury in different species. Chem Res Toxicol 2010, 23, 171-183

Nearest neighbour methods

- Select the k most similar compounds in training set to query compound
- Use the toxicological activities these to predict the activity of the query
- Lazar
  - lazy learning method – training compounds are selected at the time of processing a query compound
  - Allows models to be updated as new data become available
  - Includes models for mutagenicity and rodent carcinogenicity

Helma C. Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Mol Divers 2006, 10, 147-158

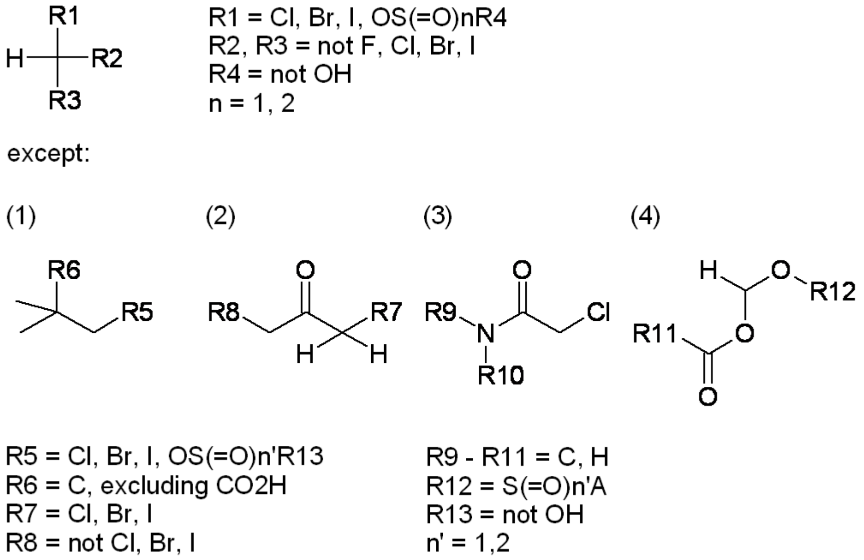
Expert systems

- Toxicological knowledge of human experts encoded as rules
- Can provide predictions about multiple mechanisms
- Include information relating to mechanism of action
- Derek for Nexus
  - Structural alerts
  - Reasoning model used to weigh up multiple arguments for and against toxicity eg using physiochemical properties, relationship between endpoints
  - Level of confidence in prediction is provided
    - Eg improbable, plausible, certain
  - Literature references are provided

Structural alerts

- Alerts: collection of substructures (toxicophores) that are associated with a toxic effect

Alkylating agent alert

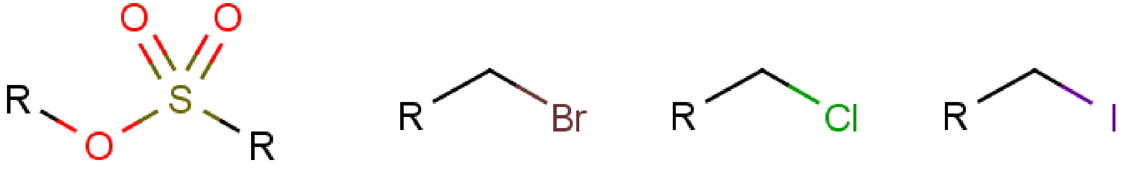


R1 = Cl, Br, I, OS(=O)nR4
R2, R3 = not F, Cl, Br, I
R4 = not OH
n = 1, 2

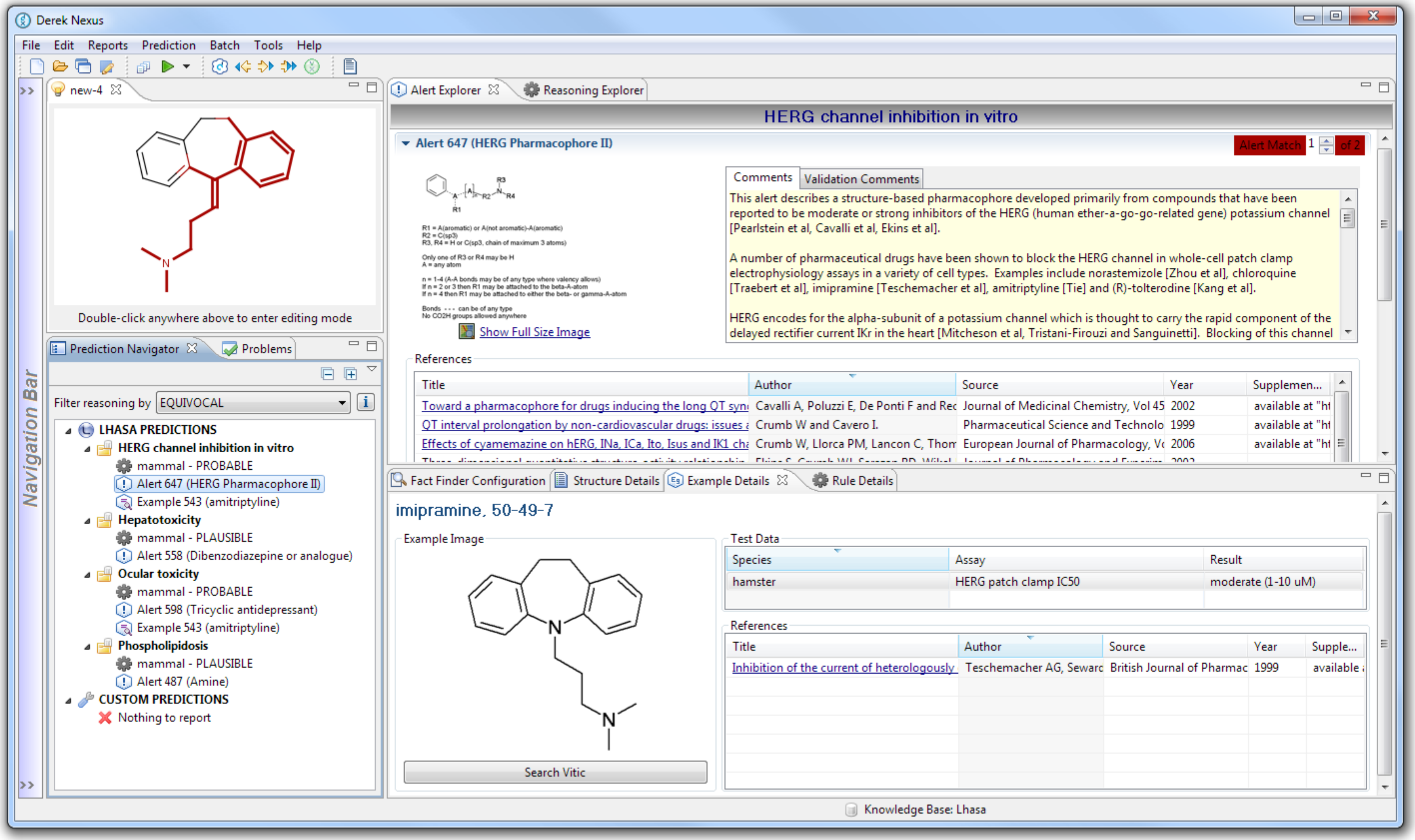
R5 = Cl, Br, I, OS(=O)n'R13
R6 = C, excluding CO2H
R7 = Cl, Br, I
R8 = not Cl, Br, I

R9 - R11 = C, H
R12 = S(=O)n'A
R13 = not OH
n' = 1,2

Alkylating agent toxicophores



Derek Nexus (www.lhasalimited.org)



Double-click anywhere above to enter editing mode

Filter reasoning by EQUIVOCAL

- LHASA PREDICTIONS
  - HERG channel inhibition in vitro
    - mammal - PROBABLE
    - Alert 647 (HERG Pharmacophore II)
    - Example 543 (amitriptyline)
  - Hepatotoxicity
    - mammal - PLAUSIBLE
    - Alert 558 (Dibenzodiazepine or analogue)
  - Ocular toxicity
    - mammal - PROBABLE
    - Alert 598 (Tricyclic antidepressant)
    - Example 543 (amitriptyline)
  - Phospholipidosis
    - mammal - PLAUSIBLE
    - Alert 487 (Amine)
- CUSTOM PREDICTIONS
  - Nothing to report

HERG channel inhibition in vitro

Alert 647 (HERG Pharmacophore II)

R1 = A(aromatic) or A(not aromatic)-A(aromatic)
R2 = C(sp3)
R3, R4 = H or C(sp3, chain of maximum 3 atoms)

Only one of R3 or R4 may be H
A = any atom

n = 1-4 (A-A bonds may be of any type where valency allows)
If n = 2 or 3 then R1 may be attached to the beta-A-atom
If n = 4 then R1 may be attached to either the beta- or gamma-A-atom

Bonds - - - can be of any type
No CO2H groups allowed anywhere

This alert describes a structure-based pharmacophore developed primarily from compounds that have been reported to be moderate or strong inhibitors of the HERG (human ether-a-go-go-related gene) potassium channel [Pearlstein et al, Cavalli et al, Ekins et al].

A number of pharmaceutical drugs have been shown to block the HERG channel in whole-cell patch clamp electrophysiology assays in a variety of cell types. Examples include norastemizole [Zhou et al], chloroquine [Traebert et al], imipramine [Teschemacher et al], amitriptyline [Tie] and (R)-tolterodine [Kang et al].

HERG encodes for the alpha-subunit of a potassium channel which is thought to carry the rapid component of the delayed rectifier current IKr in the heart [Mitcheson et al, Tristani-Firouzi and Sanguinetti]. Blocking of this channel

References

Title	Author	Source	Year	Supplemen...
Toward a pharmacophore for drugs inducing the long QT synd	Cavalli A, Poluzzi E, De Ponti F and Rec	Journal of Medicinal Chemistry, Vol 45	2002	available at "ht
QT interval prolongation by non-cardiovascular drugs: issues a	Crumb W and Cavero I.	Pharmaceutical Science and Technolo	1999	available at "ht
Effects of cyamemazine on hERG, INa, ICa, Ito, Isus and IK1 cha	Crumb W, Llorca PM, Lancon C, Thom	European Journal of Pharmacology, Vo	2006	available at "ht
Three-dimensional quantitative structure-activity relationship	Ekins S, Crumb WJ, Sarazan RD, Wikel	Journal of Pharmacology and Experim	2002	

imipramine, 50-49-7

Example Image

Test Data

Species	Assay	Result
hamster	HERG patch clamp IC50	moderate (1-10 uM)

References

Title	Author	Source	Year	Supple...
Inhibition of the current of heterologously	Teschemacher AG, Seward	British Journal of Pharmac	1999	available a

Knowledge Base: Lhasa

Expert systems predict positives only - lack of prediction does not mean non-toxic!

Expert systems

- Process of knowledge discovery can be very time consuming

- Requires detailed analysis of the literature by domain experts

Towards automation of knowledge discovery

- Aim is provide an automated tool to support the process of knowledge discovery through data mining

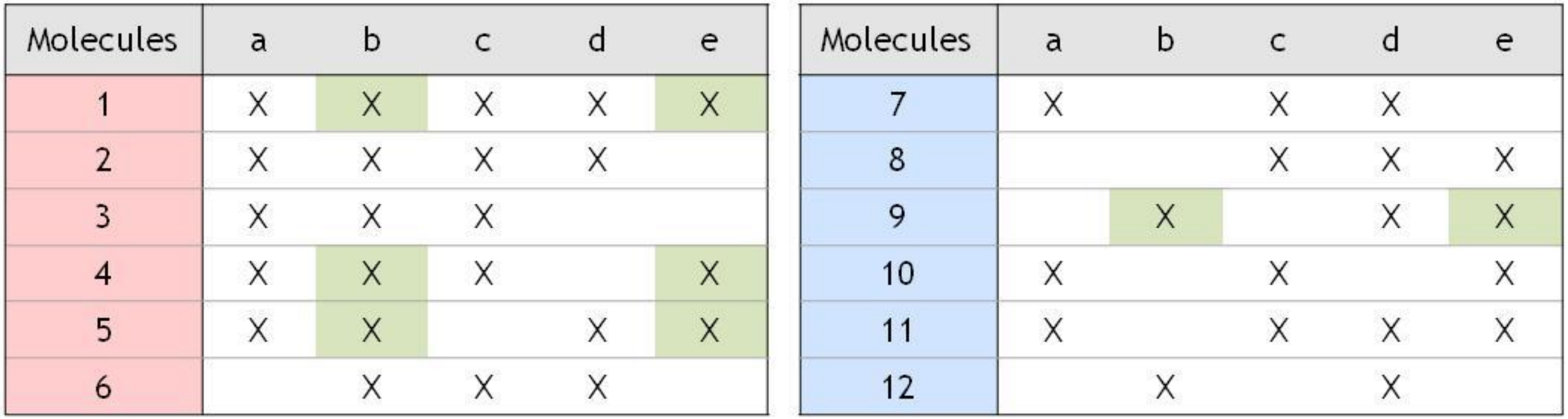
- Emerging pattern mining techniques used to identify substructural features that could be associated with toxicity

- The substructural features identified require validation through the literature by knowledge-base workers

- Collaborative project between University of Sheffield and Lhasa Limited

Emerging Patterns

- Emerging patterns are sets of properties (descriptors) that occur more often in one class compared to another



Molecules	a	b	c	d	e
1	X	X	X	X	X
2	X	X	X	X	
3	X	X	X		
4	X	X	X		X
5	X	X		X	X
6		X	X	X	

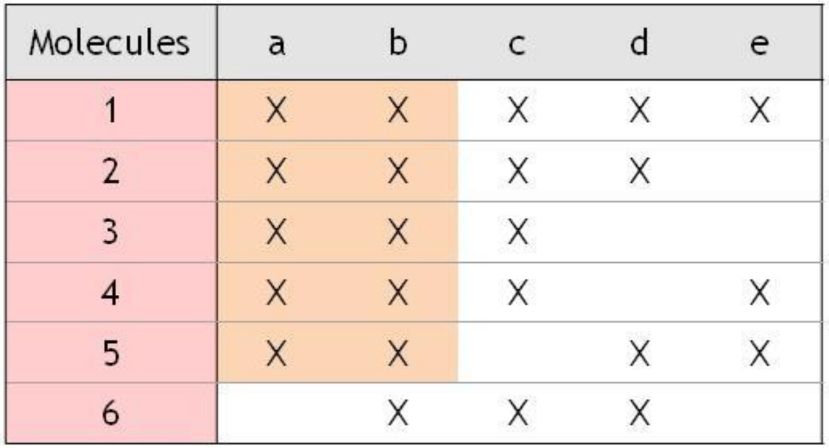
Molecules	a	b	c	d	e
7	X		X	X	
8			X	X	X
9		X		X	X
10	X		X		X
11	X		X	X	X
12		X		X	

- {b, e} is an emerging pattern supported by active molecules [1, 4, 5] and inactive molecule [9]
- Emphasis is on finding combinations of properties

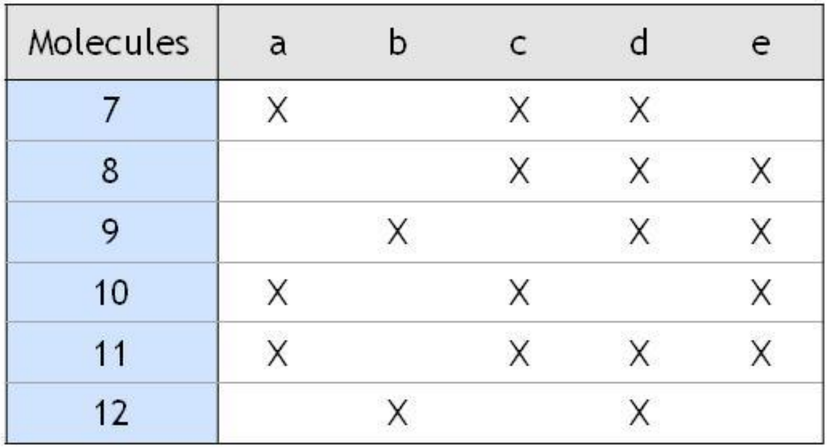
[†]Dong, G.; Li, J. In Efficient mining of emerging patterns: discovering trends and differences, The Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 1999; Association for Computing Machinery Press: San Diego, CA, USA, 1999; pp 43-52.

Jumping Emerging Patterns (JEPs)

- JEPs are patterns of properties that occur in one class *only* compared to another



Molecules	a	b	c	d	e
1	X	X	X	X	X
2	X	X	X	X	
3	X	X	X		
4	X	X	X		X
5	X	X		X	X
6		X	X	X	

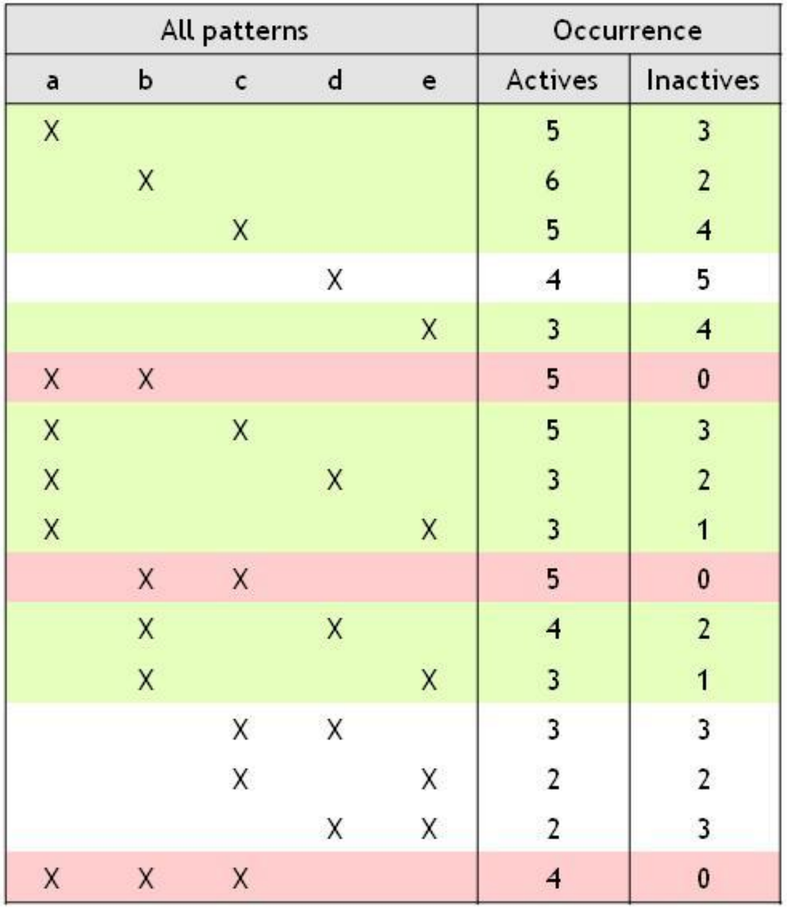


Molecules	a	b	c	d	e
7	X		X	X	
8			X	X	X
9		X		X	X
10	X		X		X
11	X		X	X	X
12		X		X	

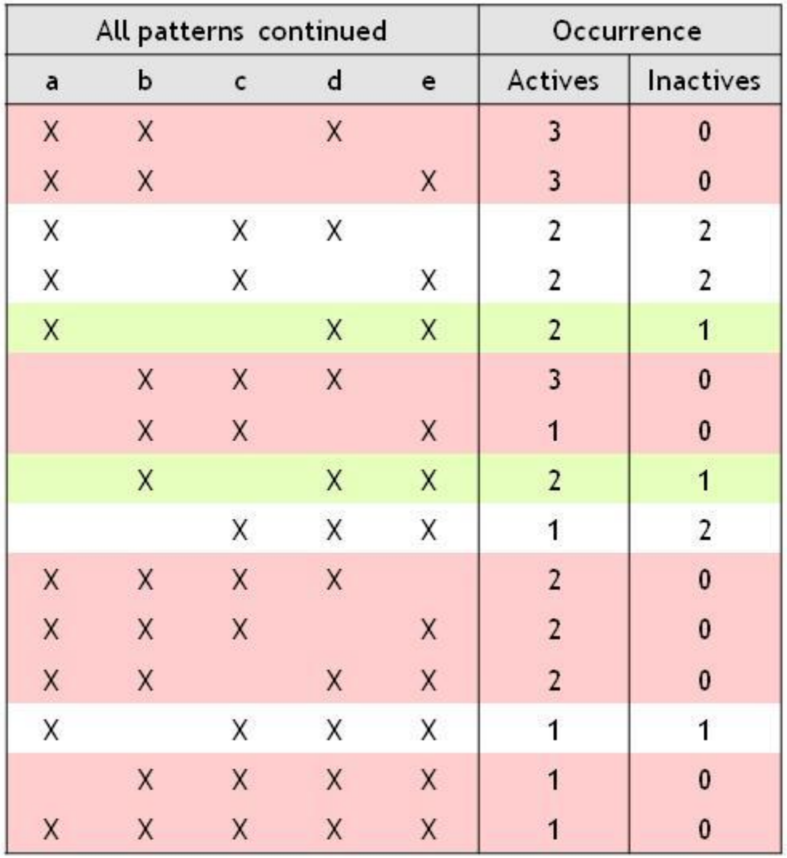
- {a, b} is a JEP supported by actives [1, 2, 3, 4, 5] and no inactives

†**Dong, G.; Li, J. In Efficient mining of emerging patterns: discovering trends and differences, The Fifth International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 1999; Association for Computing Machinery Press: San Diego, CA, USA, 1999; pp 43-52.**

JEP mining by enumeration



All patterns					Occurrence	
a	b	c	d	e	Actives	Inactives
X					5	3
	X				6	2
		X			5	4
			X		4	5
				X	3	4
X	X				5	0
X		X			5	3
X			X		3	2
X				X	3	1
	X	X			5	0
	X		X		4	2
	X			X	3	1
		X	X		3	3
		X		X	2	2
			X	X	2	3
X	X	X			4	0



All patterns continued					Occurrence	
a	b	c	d	e	Actives	Inactives
X	X		X		3	0
X	X			X	3	0
X		X	X		2	2
X		X		X	2	2
X			X	X	2	1
	X	X	X		3	0
	X	X		X	1	0
	X		X	X	2	1
		X	X	X	1	2
X	X	X	X		2	0
X	X	X		X	2	0
X	X		X	X	2	0
X		X	X	X	1	1
	X	X	X	X	1	0
X	X	X	X	X	1	0

More efficient algorithms are available!

Applications of EPs in Chemoinformatics

- Auer & Bajorath
  - Physicochemical property ranges mapped to a binary bit string

Auer, J.; Bajorath, J. Emerging chemical patterns: a new methodology for molecular classification and compound selection. Journal of Chemical Information and Modeling 2006, 46, (6), 2502-2514.

- Lozano et al.
  - "Jumping fragments" in toxicity dataset
  - Subgraphs are enumerated in actives and searched for in inactives

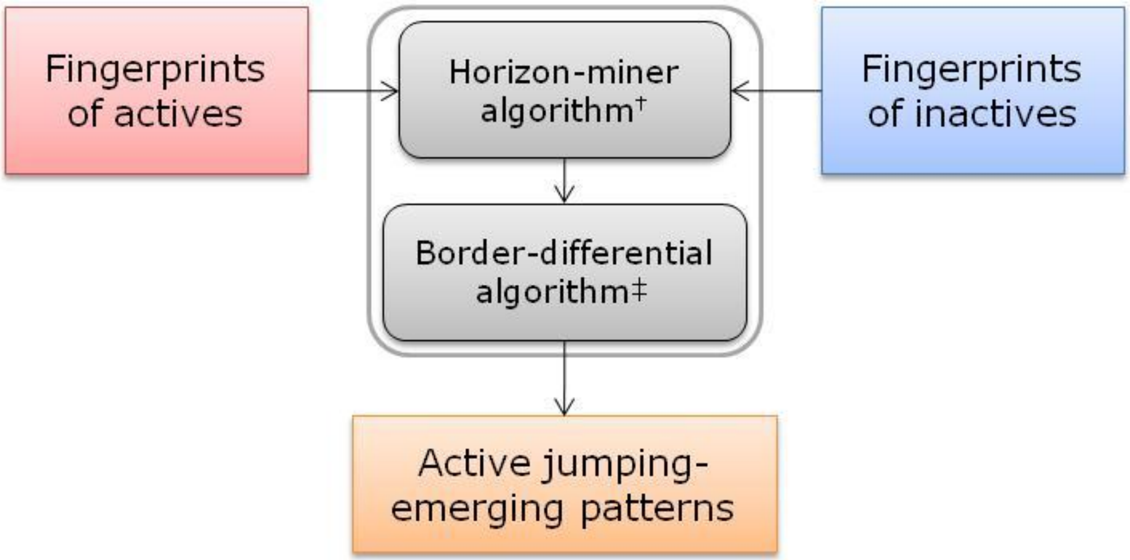
Lozano, S.; Poezevara, G.; Halm-Lemeille, M. P.; Lescot-Fontaine, E.; Lepailleur, A.; Bissell-Siders, R.; Crémilleux, B.; Rault, S.; Cuissart, B.; Bureau, R. Introduction of jumping fragments in combination with QSARs for the assessment of classification in ecotoxicology. . Journal of Chemical Information and Modeling, 2010, 50, 1330–1339.

Mining JEPs in toxicity data

- Aim is to identify patterns (combinations of structural descriptors) that are present in toxic molecules but absent from non-toxic molecules
- Use the patterns to suggest substructural features to knowledge-base workers for validation through the literature
- Applied to small structural fragments
  - Atom pairs, circular fps, etc
  - Allows combinations of descriptors to be identified
  - Potential toxicphores can be constructed from the descriptors
  - Allows hierarchical relationships to be built that represent more detailed (but lower supported) substructural features

Mining JEPs in toxicity data

Given a dataset of toxic (active) and non-toxic (inactive) compounds


Fingerprints of actives
Horizon-miner algorithm†
Fingerprints of inactives
Border-differential algorithm‡
Active jumping-emerging patterns

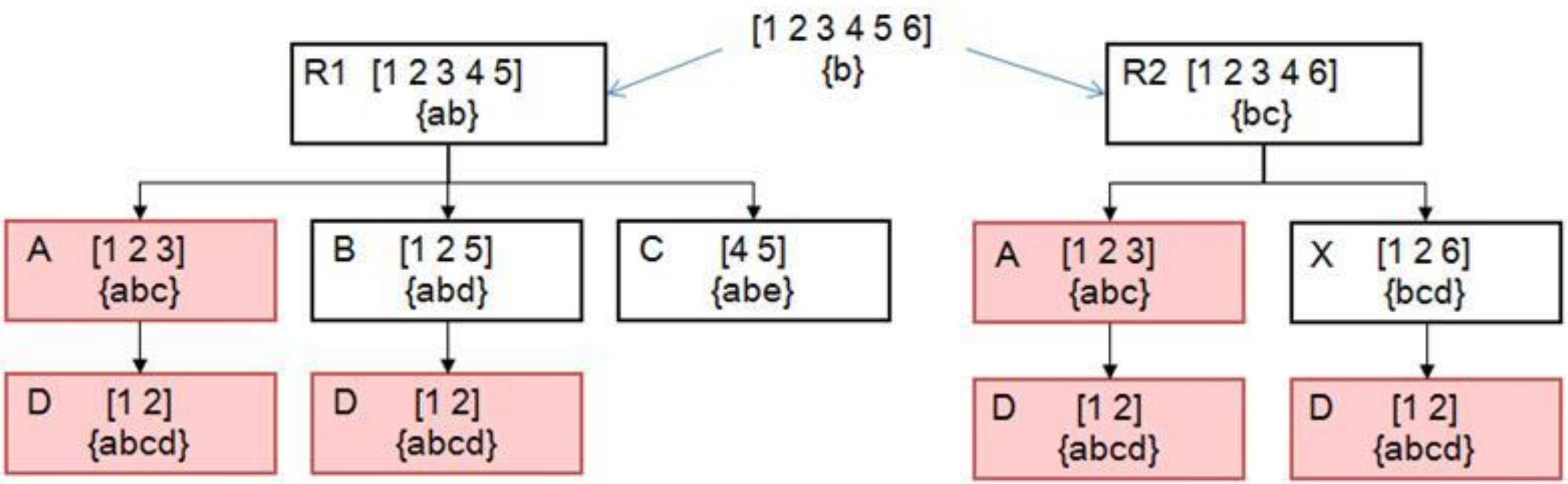
The set of toxic molecules that support a JEP are formed around a common sets of bits which describe a potential toxicophore

Form of supervised clustering

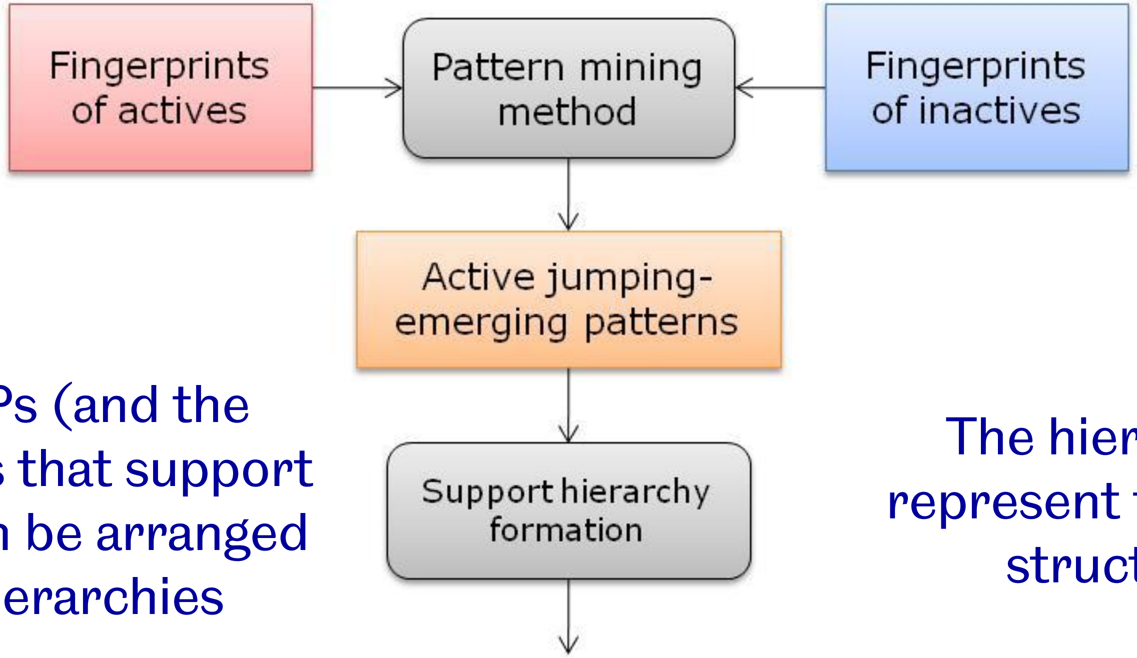
†Li, J.; Dong, G.; Ramamohanarao, K., Making use of the most expressive jumping emerging patterns for classification. Knowledge and Information Systems 2001, 3, (2), 131-145.

‡Dong, G.; Li, J., Mining border descriptions of emerging patterns from dataset pairs. Knowledge and Information Systems 2005, 8, (2), 178-202.

Hierarchies of JEPs



Hierarchies of JEPs



Fingerprints of actives

Pattern mining method

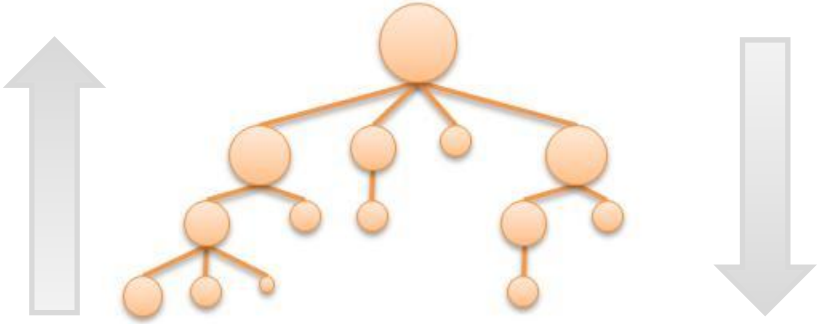
Fingerprints of inactives

Active jumping-emerging patterns

Support hierarchy formation

The JEPs (and the molecules that support them) can be arranged into hierarchies

The hierarchies represent families of structures



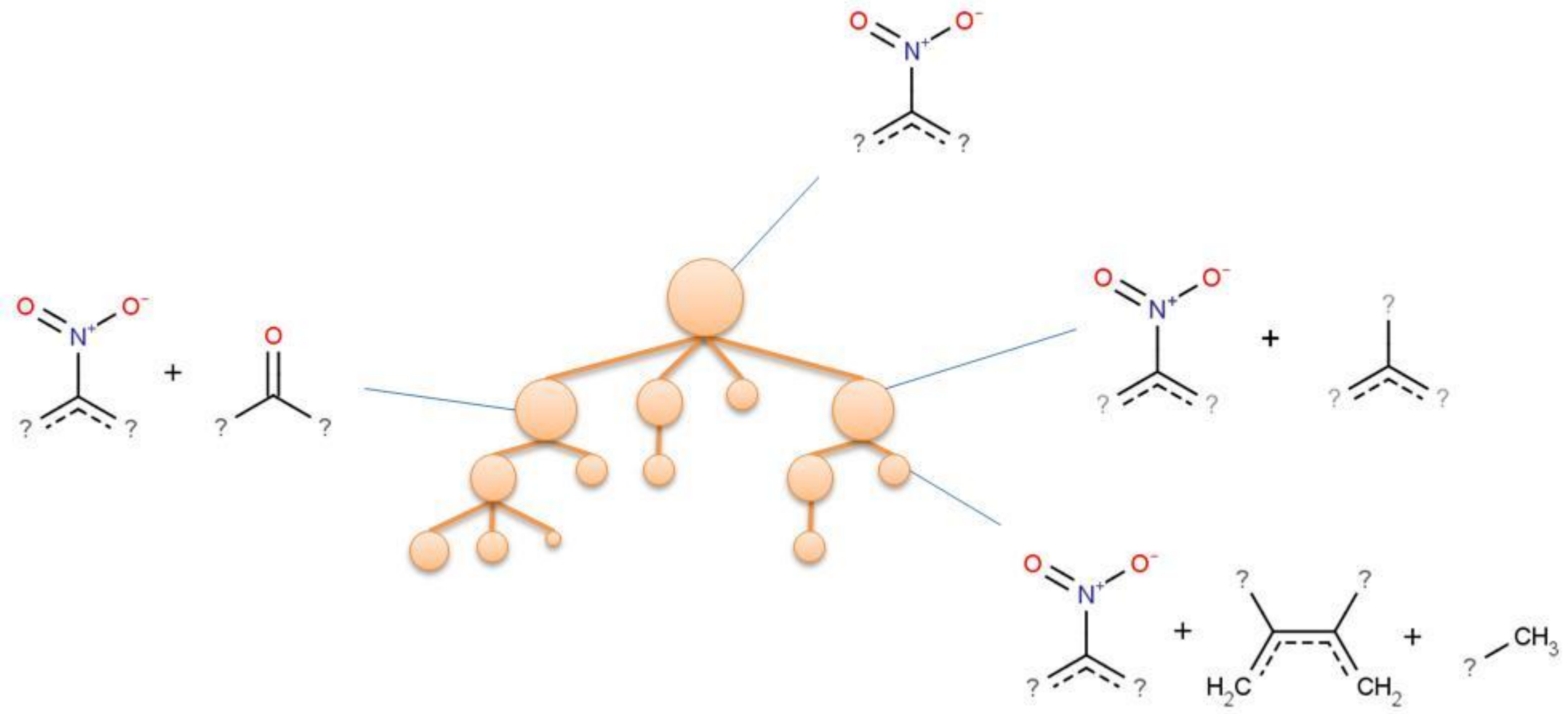
Higher support

Lower support

More generic patterns

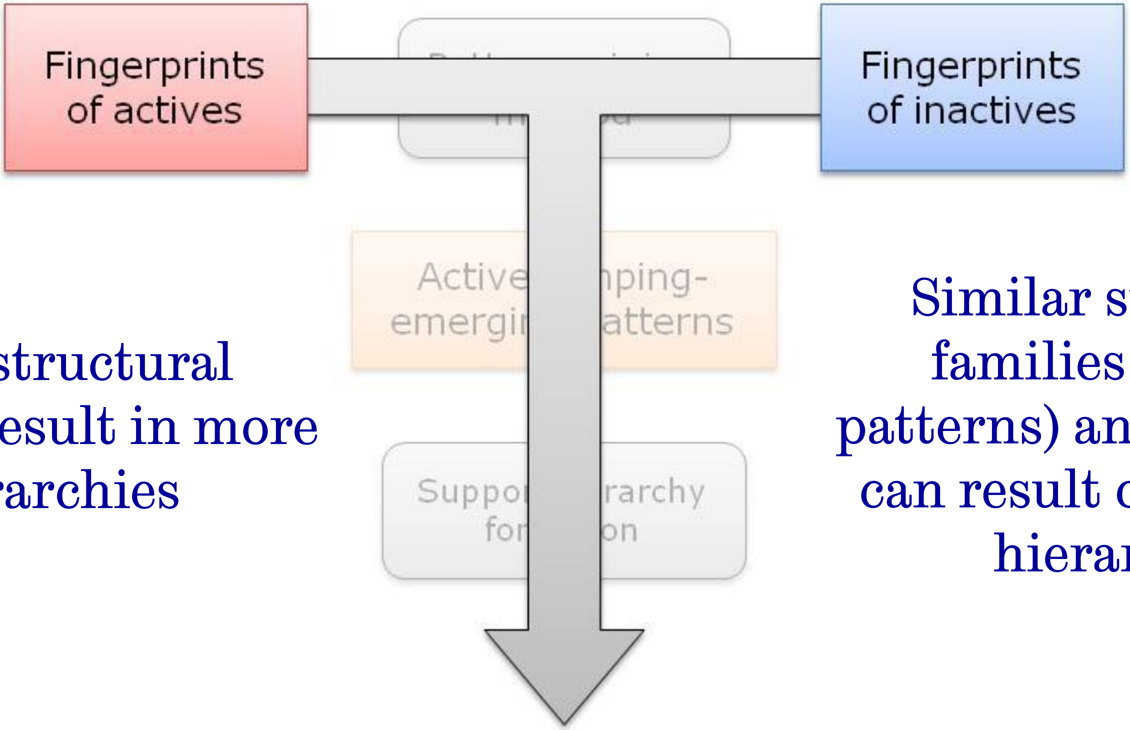
More specific patterns

Support hierarchies



Exploring the hierarchies allows relationships between structures to be analysed

Support hierarchies



More structural families result in more hierarchies

Similar structural families (similar patterns) and noisy data can result overlapping hierarchies

JEP mining algorithm

- Generate a set of binary fingerprints using the active compounds in the dataset and use these to form fingerprints for both the actives and inactives
- Apply the Horizon-Miner algorithm to extract the maximal patterns for both the actives and the inactives using the binary fingerprints
- Apply the border-differential algorithm to mine the set of all possible minimal JEPs in the actives compared to the inactives
- Reduce the set of minimal JEPs to those that occur in distinct sets of actives
- Identify relationships between the supporting actives of minimal JEPs, and arrange them into hierarchies
- Extract the maximum set of commonly occurring descriptors from the set of actives that support each minimal JEP, to form the largest and most descriptive representation of their common structural features.

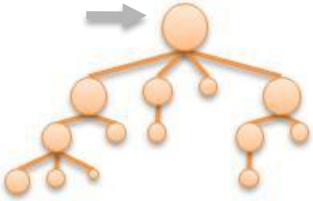
Example: Ames mutagenicity

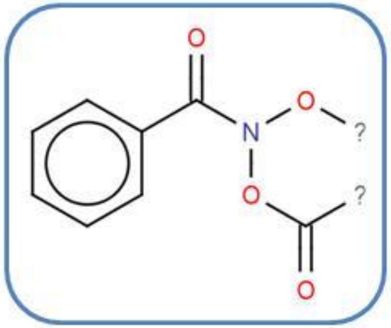
- Endpoint
  - Known to be caused by a diverse set of small activating substructures
- Dataset
  - Hansen[†] ames mutagenicity dataset was encoded as fingerprints using an in-house naïve fragmentation process
    - i.e. breaking all C-C, C-H and non-heterocyclic bonds
- Interpretable substructure fingerprints

[†] Hansen, K. Mika, S.; Schroeter, T.; Sutter, A.; Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K. R.; Benchmark data set for in silico prediction of Ames mutagenicity. Journal of Chemical Information and Modeling 2009, 49, (9), 2077.

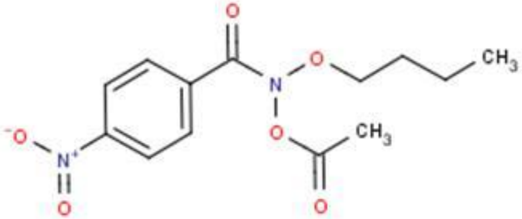
Ames mutagenicity

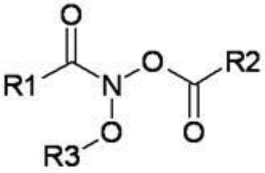
Root patterns with highest support are the most interesting



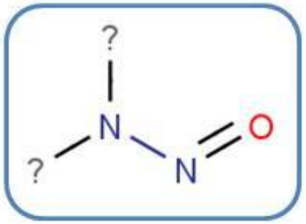


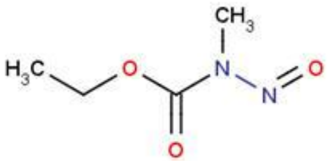
N-Acyloxy-N-alkoxyamides





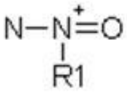
R1 = C (aromatic)
R2 = C, H
R3 = C (alkyl)





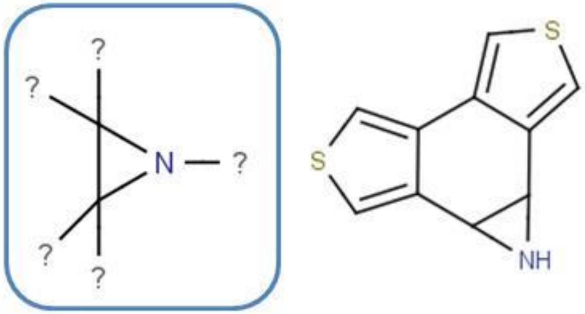
N-nitro or N-nitroso

$N-N=O$ or

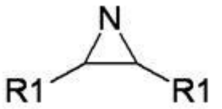


R1 = H, O^-

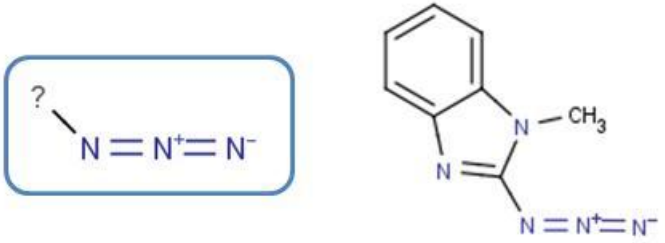
Ames mutagenicity



Azerine or Aziridine



R1 = any except =O, =S



Azide, hydrazoic acid or azide salt

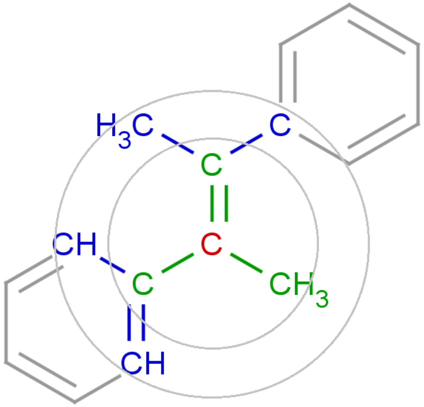
$C{-}N{=}\overset{+}{N}{=}N^-$ (I)

$N{\equiv}\overset{+}{N}{=}N^-$ or $H{-}N{=}\overset{+}{N}{=}N^-$ (II)

Found substructures that closely match existing alerts in Derek Nexus

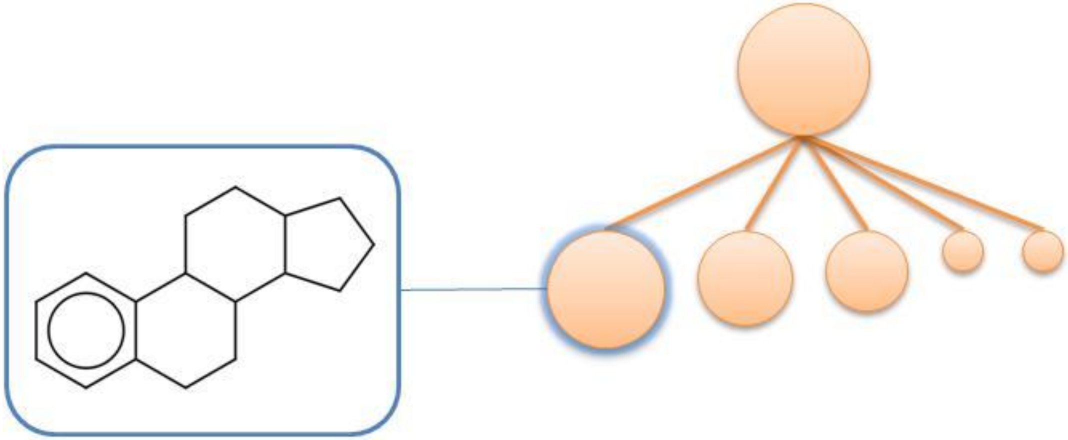
Example: Oestrogenicity

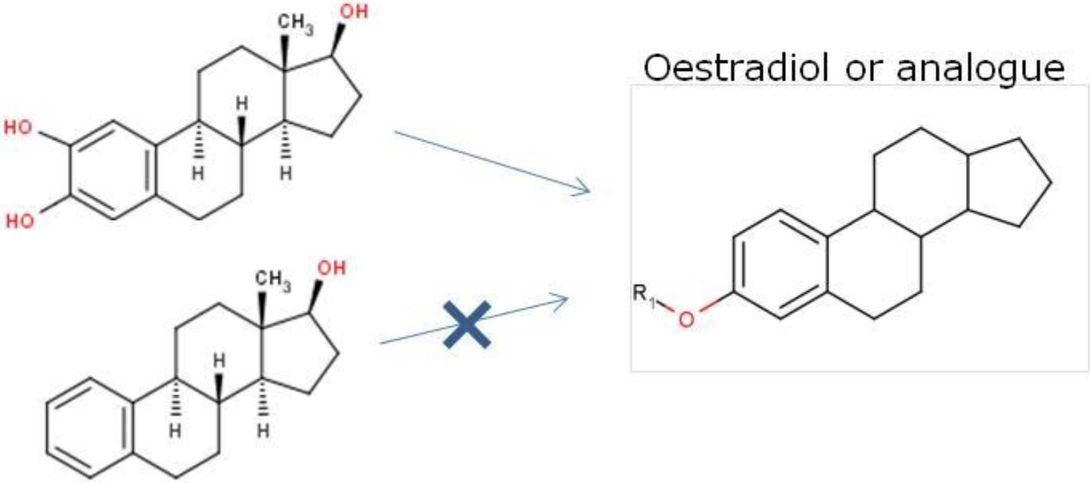
- Endpoint
  - Known to result from a small number of loosely defined toxicophores
- The oestrogenicity dataset* was encoded as circular fingerprints



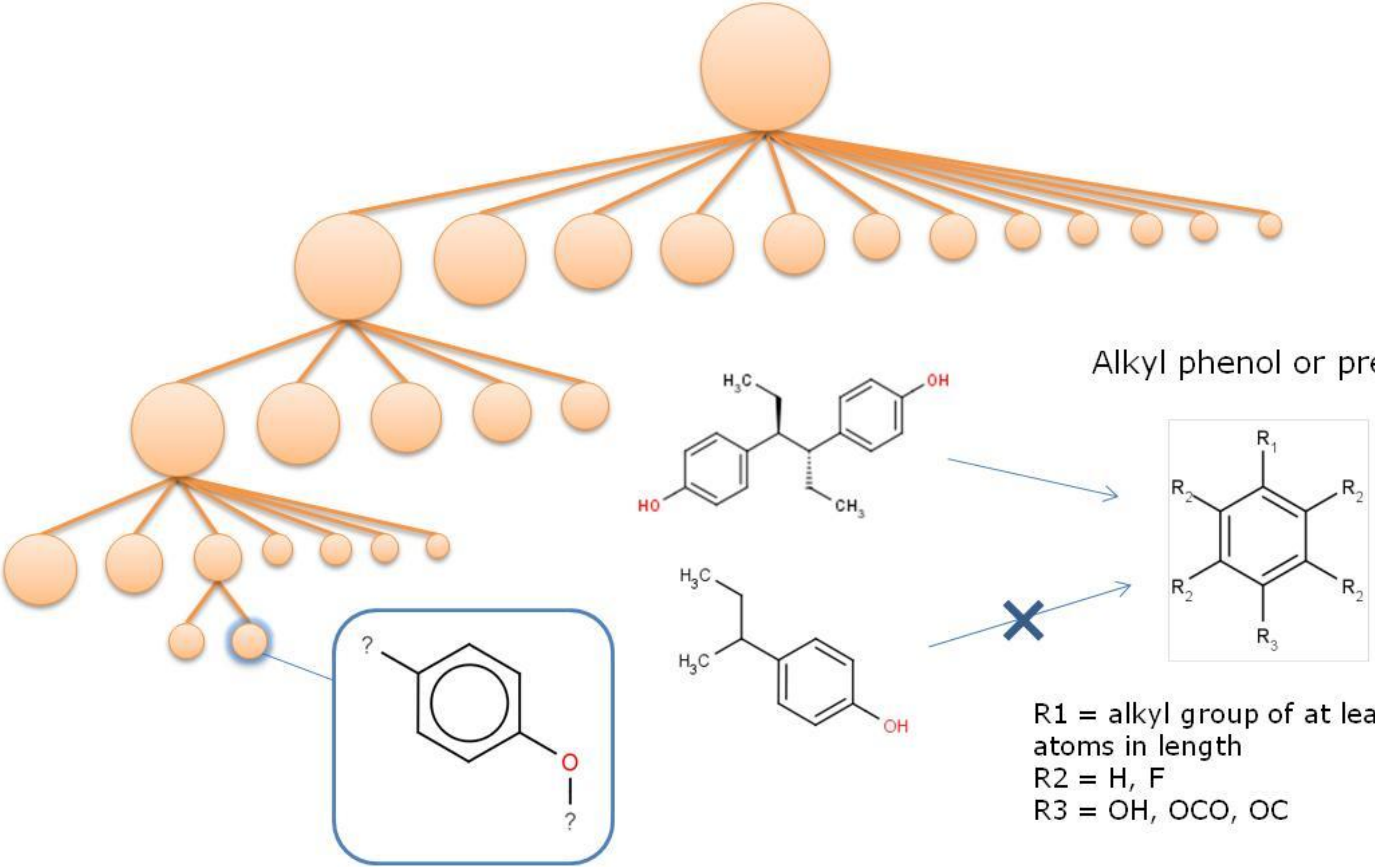
*The FDA National Center for Toxicological Research – Estrogen Receptor Binding (NCTRER) database obtained from the Distributed Structure-Searchable (DSSTox) network, hosted by the US EPA.

Oestrogenicity





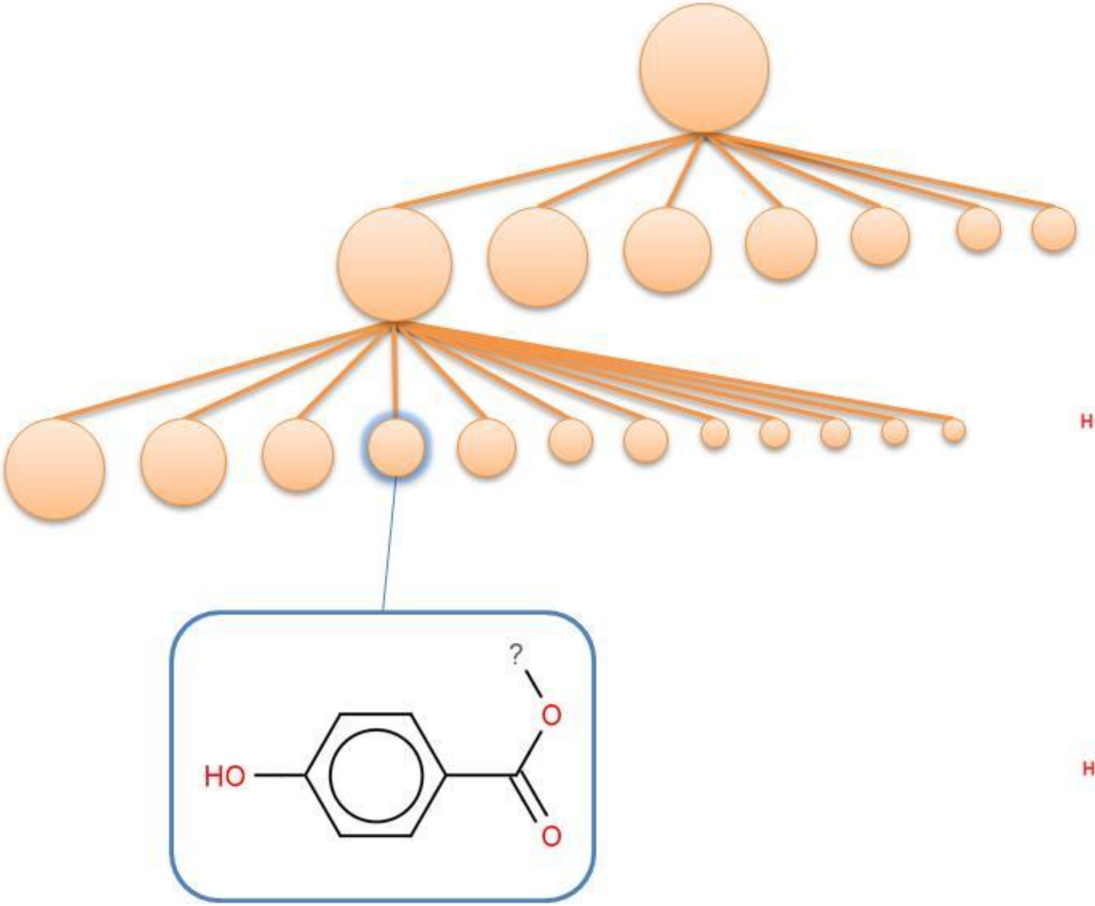
Oestrogenicity

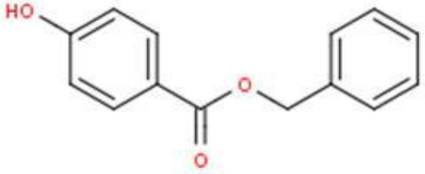


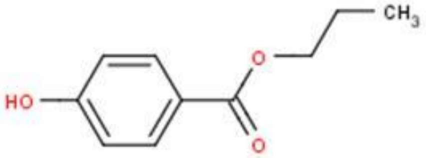
Alkyl phenol or precursor

R1 = alkyl group of at least 4 carbon atoms in length
R2 = H, F
R3 = OH, OCO, OC

Oestrogenicity







Found substructures that are not known to Derek Nexus and which may be worth further investigation

Conclusions: JEPs

- The aim of the JEP mining described here is to assist knowledge-based workers in discovering new alerts to augment the knowledge-base
- Substructural features have been identified that are similar to known toxicophores
- Substructural features not already present in the knowledge-base have also been identified
- JEP mining could be used predictively (not explored here)
- Currently focused on EP mining
  - Improved handling of noisy data
  - Preliminary work has shown that a more manageable number of patterns is found

Acknowledgements

- Richard Sherhod
  - University of Sheffield & Lhasa Limited
- Jonathan Vessey, Philip Judson
  - Lhasa Limited
- Funding
  - Lhasa Limited
  - Technology Strategy Board
  - Engineering and Physical Sciences Research Council

Further Reading

- Marchant C. Computation Toxicology: A tool for all industries. WIREs Computational Molecular Sciences, 2011, doi: 10.1002/wcms.100
- Merlot C. Computational toxicology—a tool for early safety evaluation. Drug Discovery Today, 2010, 15, 16-22
- Modi S, Hughes M, Garrow A, White A. The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. Drug Discovery Today, 2012, 17, 135-142.
- Muster W, Breidenbach B, Fischer H, Kirchner S, Mueller L, Pahler A. Computational toxicology in drug development. Drug Discovery Today , 2008, 13, 303-310
- Valerio Jr. L.G. In silico toxicology for the pharmaceutical sciences. Toxicology and Applied Pharmacology 2009, 241, 356–370
- Varnek A, Baskin I. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? Journal of Chemical Information and Modeling 2012, 52, 1413−1437