



The University of
Nottingham



GlaxoSmithKline

QSAR modeller seeks meaningful relationship

Craig Bruce

UKQSAR, 6 November 2008

Current thoughts?

Pitfalls in QSAR

QSAR: dead or alive?

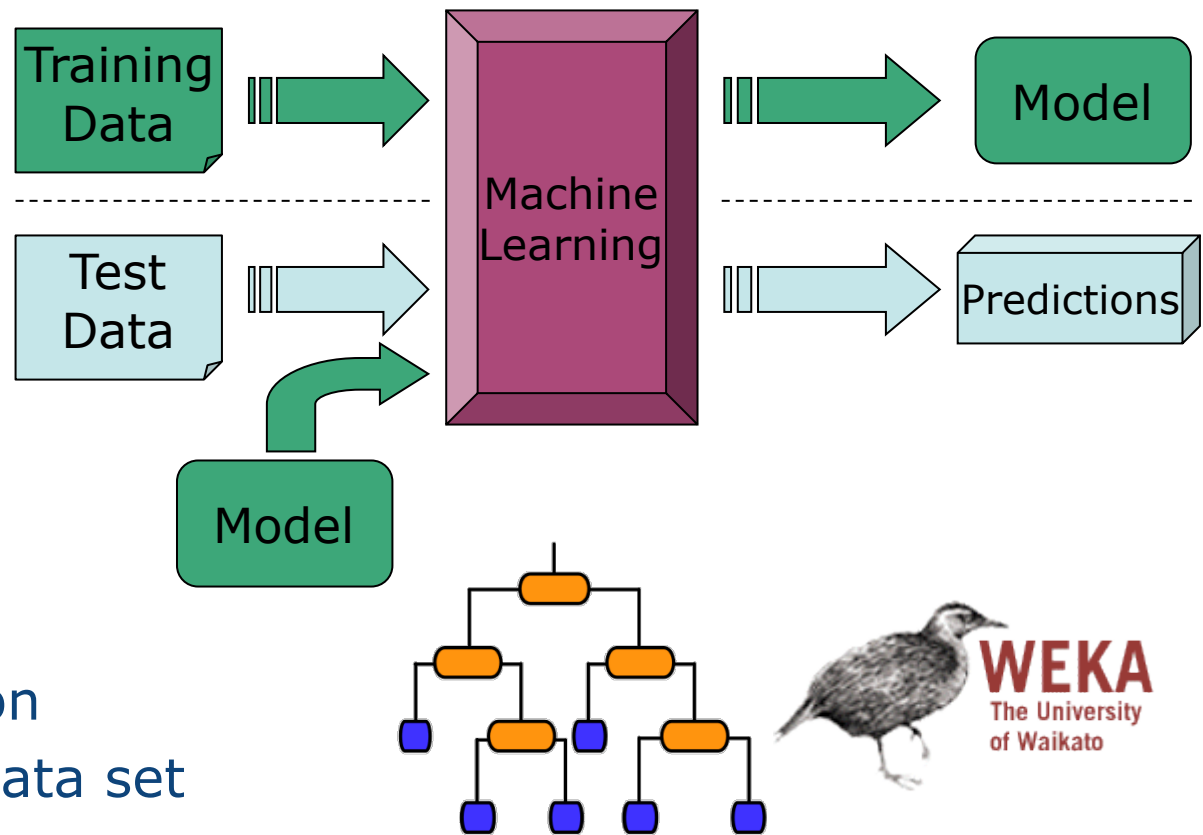
The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy)

Introduction

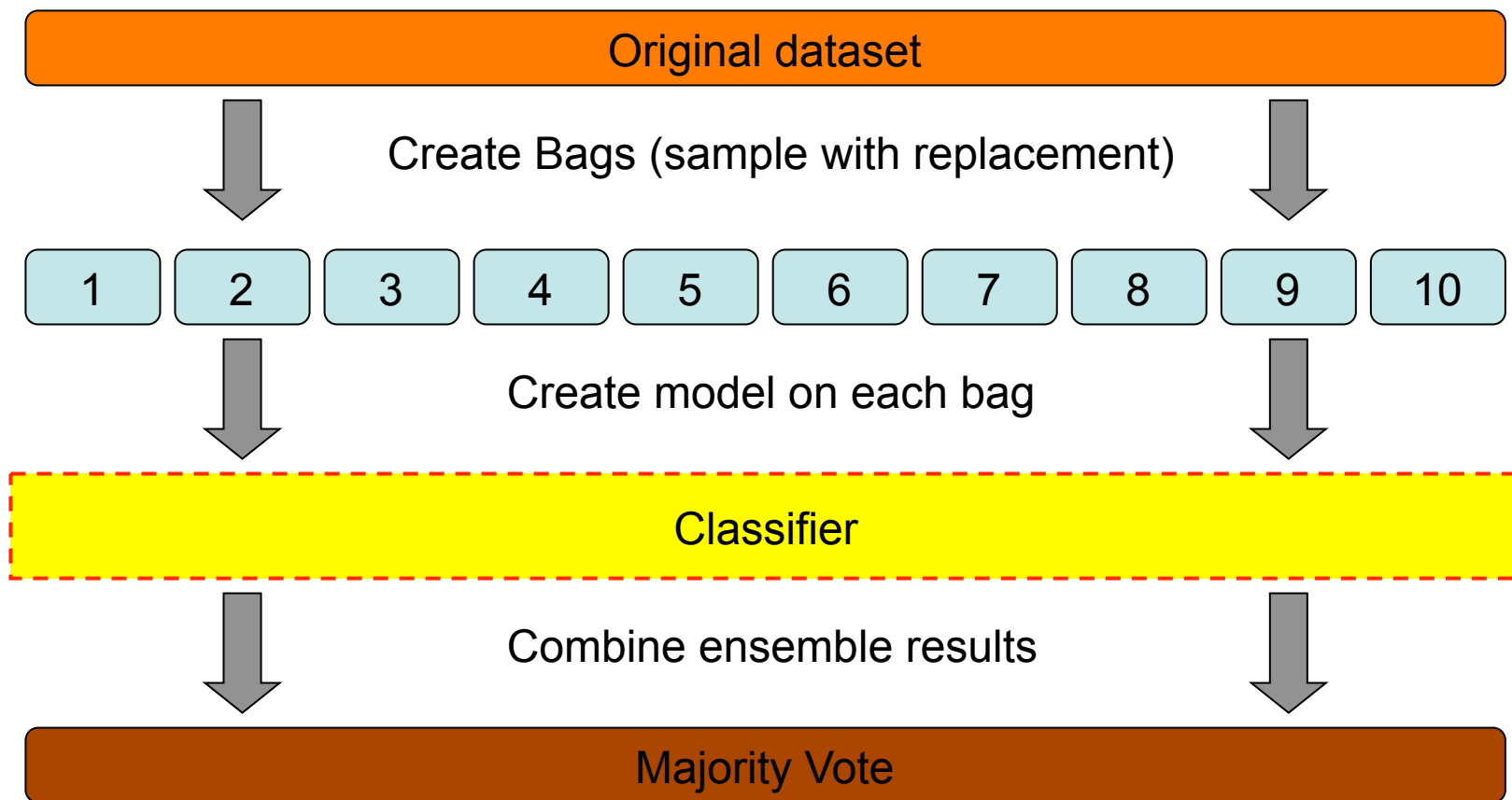
- Modern QSAR uses techniques from machine learning:
 - Neural networks
 - Decision trees
 - Support Vector Machine
- They were not designed specifically for use in the pharma domain
- Our data is very different
 - Dimensionality

Work flow

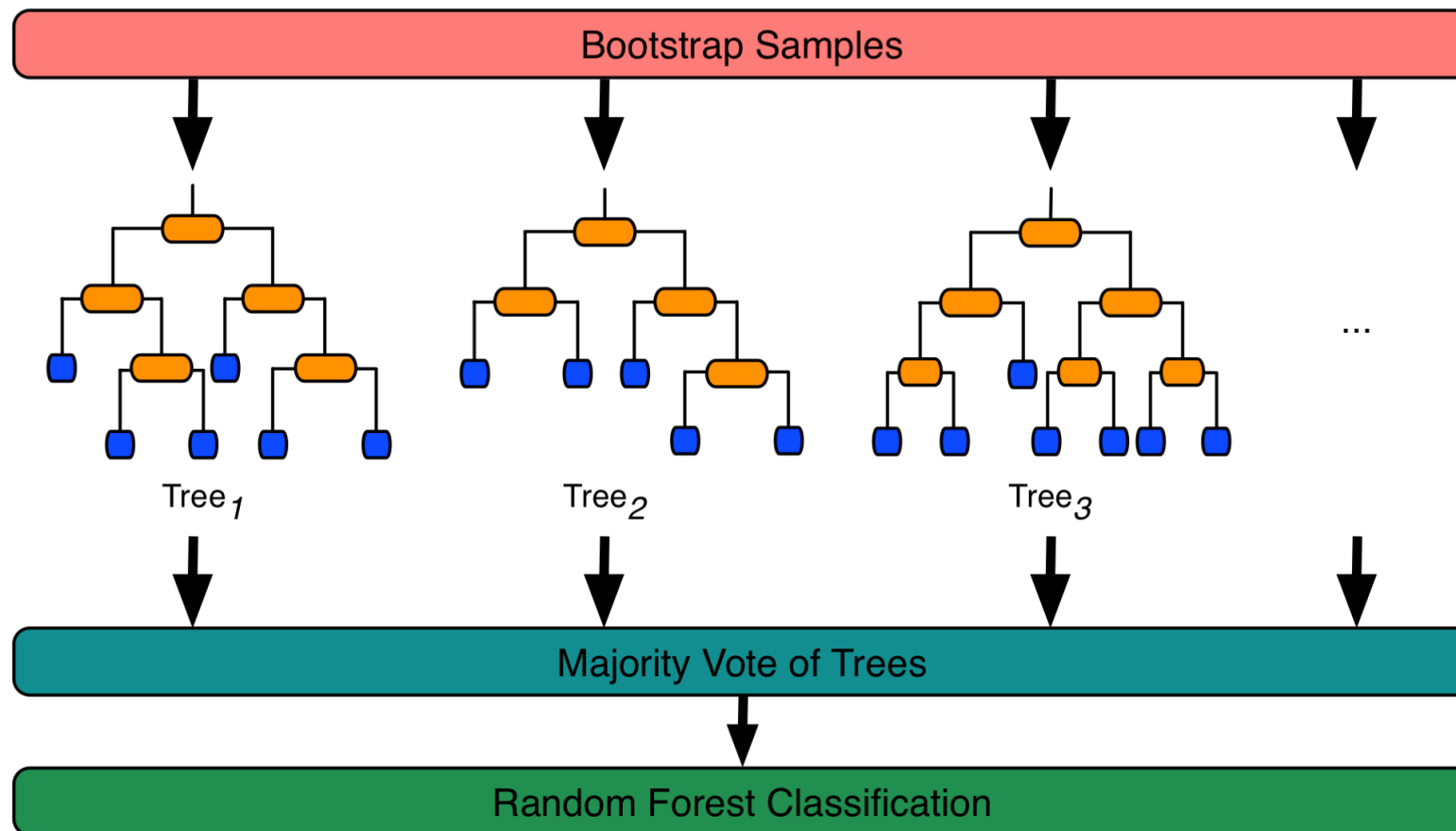
- Data sets
 - Training
 - Test
- Descriptors
- Algorithm
- Validation
 - Cross-validation
 - Independent data set



Bagging



Random forest



Why random forest?

- Why not use SVM?
 - Good predictive performance
 - Extensive optimization required
 - Interpretation not straightforward
 - Longer compute time
- Machine learning techniques not designed with interpretation in mind

Parameter tuning

- C
- ϵ
- Kernel
 - RBF
 - Polynomial
- Kernel-specific options
 - Cache
 - Exponent
 - γ
- Number of trees
- Depth of tree
- Number of features

SVM options tend to be data set specific
Random forest offers more generic defaults

Data sets for comparison



Data set	Compounds	Descriptors	
		2.5D	Fragments
ACE	114	56	1024
AchE	111	63	774
BZR	163	75	832
COX2	322	74	660
DHFR	397	70	952
GPB	66	70	692
THER	76	64	575
THR	88	66	527

Multiple comparison statistics

- Paired t -test (two-tailed)
 - 2 classifiers over multiple datasets
- Multiple Classifiers
- Multiple hypothesis is a statistical problem
 - ANOVA (parametric)
 - Friedman test (non-parametric)
- Iman and Davenport test
- Nemenyi test
- Tukey-type comparisons

Results

$p = 0.05$



Classifier	Average rank score	Average prediction %
Decision Tree	2.47	73.7
Random Forest	3.03	74.3
SVM	3.91	75.5
Boosted Tree	3.94	75.0
Bagged Tree	4.38	75.5
Tuned Random Forest	5.00	76.6
Tuned SVM	5.28	77.2

Larger dataset

- GSK In-house
- 122 descriptors
 - Physical chemistry
 - E-state
- 3 class classification
 - Highly skewed

Data set size			
	Train	Hold	Test
#	7999	2217	4000
Class split			
	L	M	H
%	77	14	9

Default results

- Correctly classified: **78%**
- Kappa: 0.17

Pred	L	M	H
L	3026 (98.8)	13	23
M	516	14 (2.5)	40
H	272	17	79 (21.5)

Cost matrix

- Correctly classified: **75%**
- Kappa: 0.30
- Cost matrix:

0	2	2
20	0	5
20	10	0

Pred	L	M	H
L	2758 (90.1)	205	99
M	363	114 (20.0)	93
H	150	81	137 (37.2)

Over sampling

- Correctly classified: **75%**
- Kappa: 0.30

Pred	L	M	H
L	2741 (89.5)	214	107
M	356	121 (21.2)	93
H	144	86	138 (37.5)

Cost matrix & over sampling

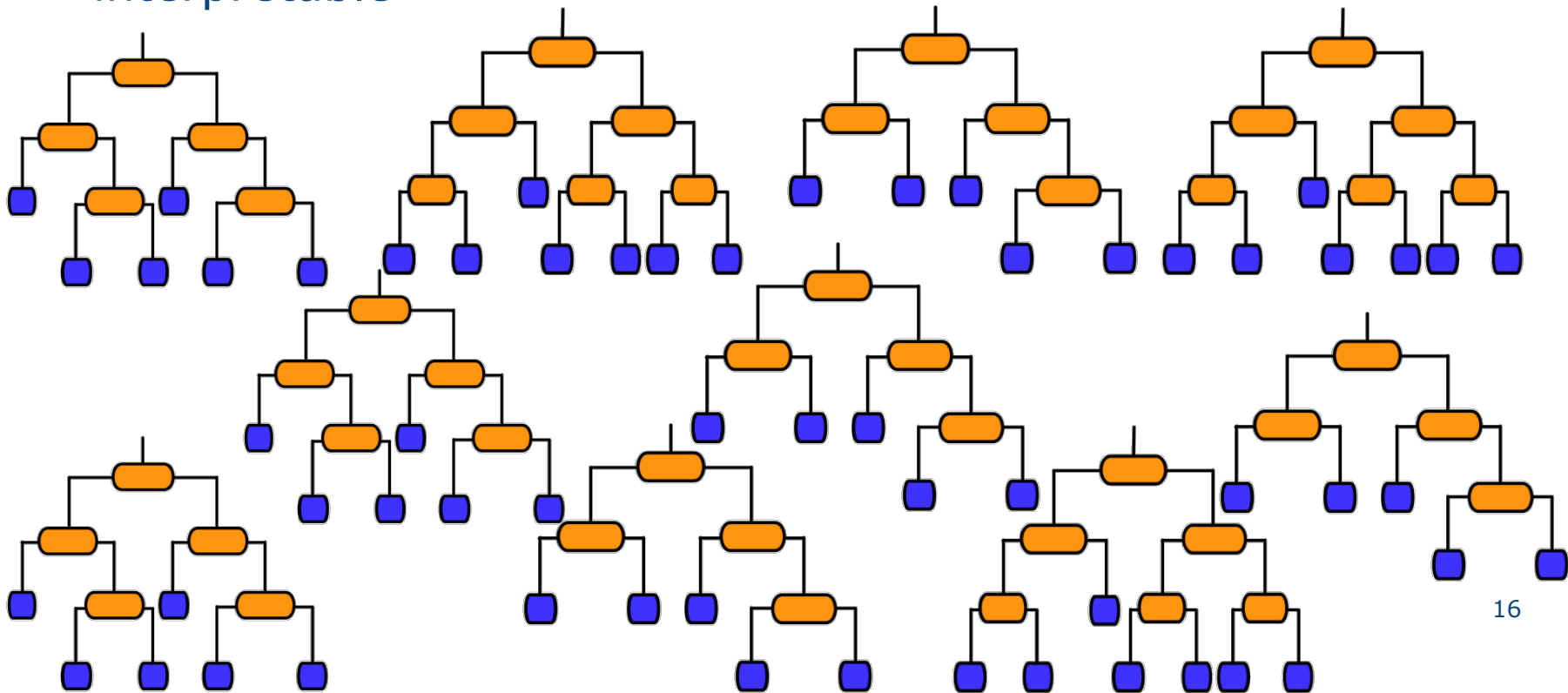
- Correctly classified: **68%**
- Kappa: 0.30
- Cost matrix:

0	10	5
20	0	5
20	15	0

Pred	L	M	H
L	2314 (75.6)	600	148
M	215	238 (41.8)	117
H	80	132	156 (42.4)

Interpretation

- Decision tree is an excellent interpretative model
- In theory this makes a random forest equally as interpretable



Forest Fingerprints

- Binary
- Prediction
 - Correct
 - Incorrect

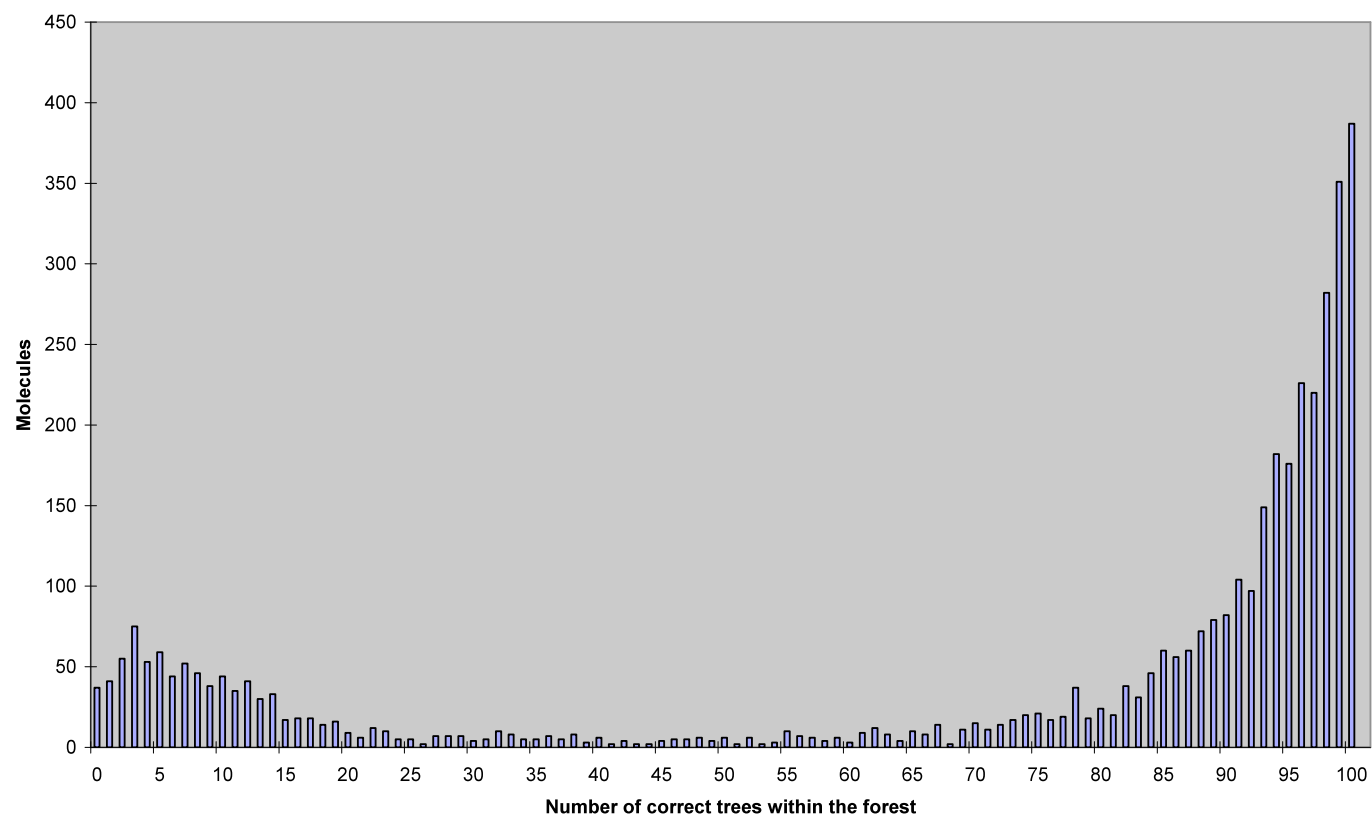
Trees: 1 - 100

0	0	0	0	0
1	1	1	0	1
1	0	0	0	1

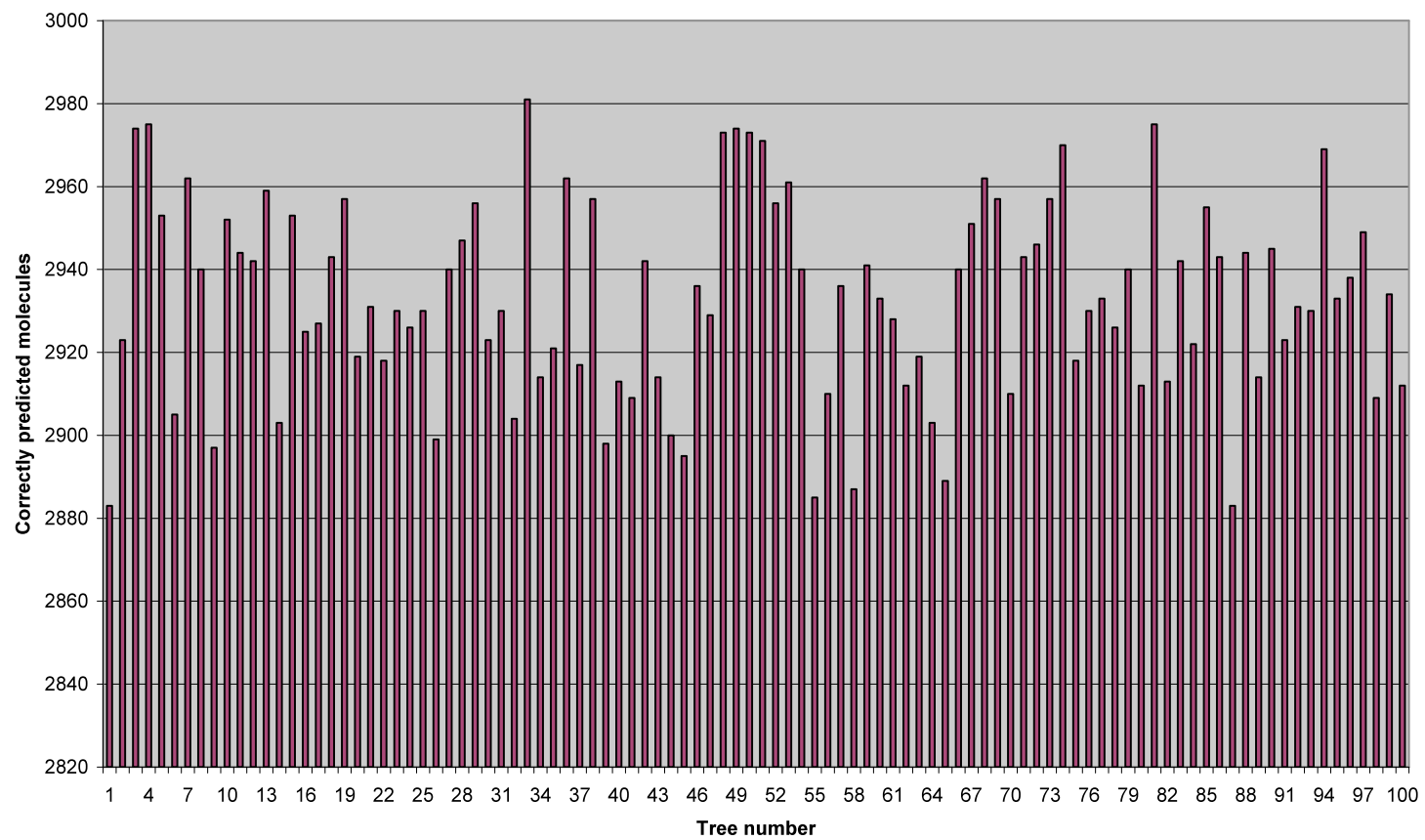
Molecules: 1 - 4000

0	1	1	1	1
0	1	0	1	1
0	1	0	1	1

Tree fingerprint



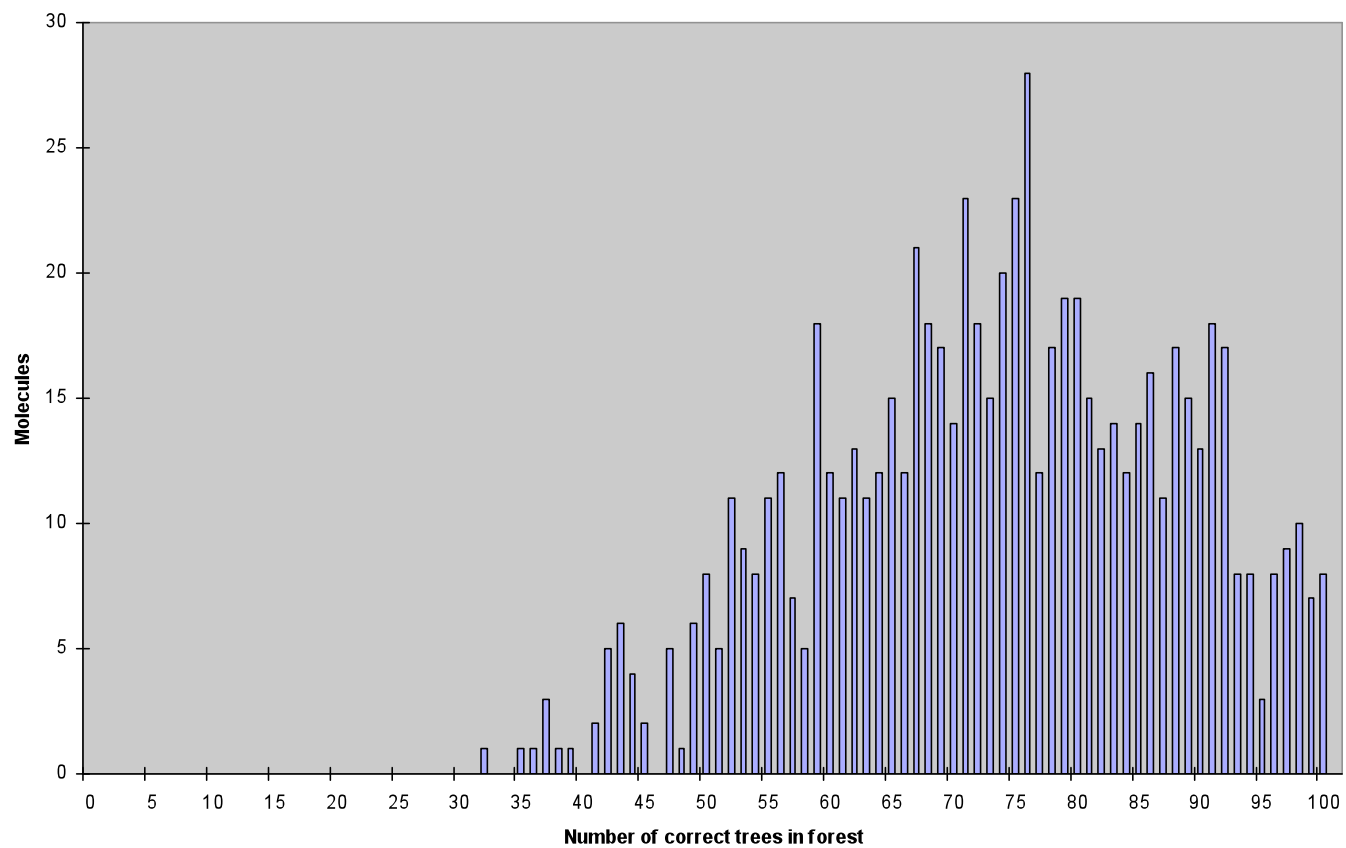
Molecular fingerprint



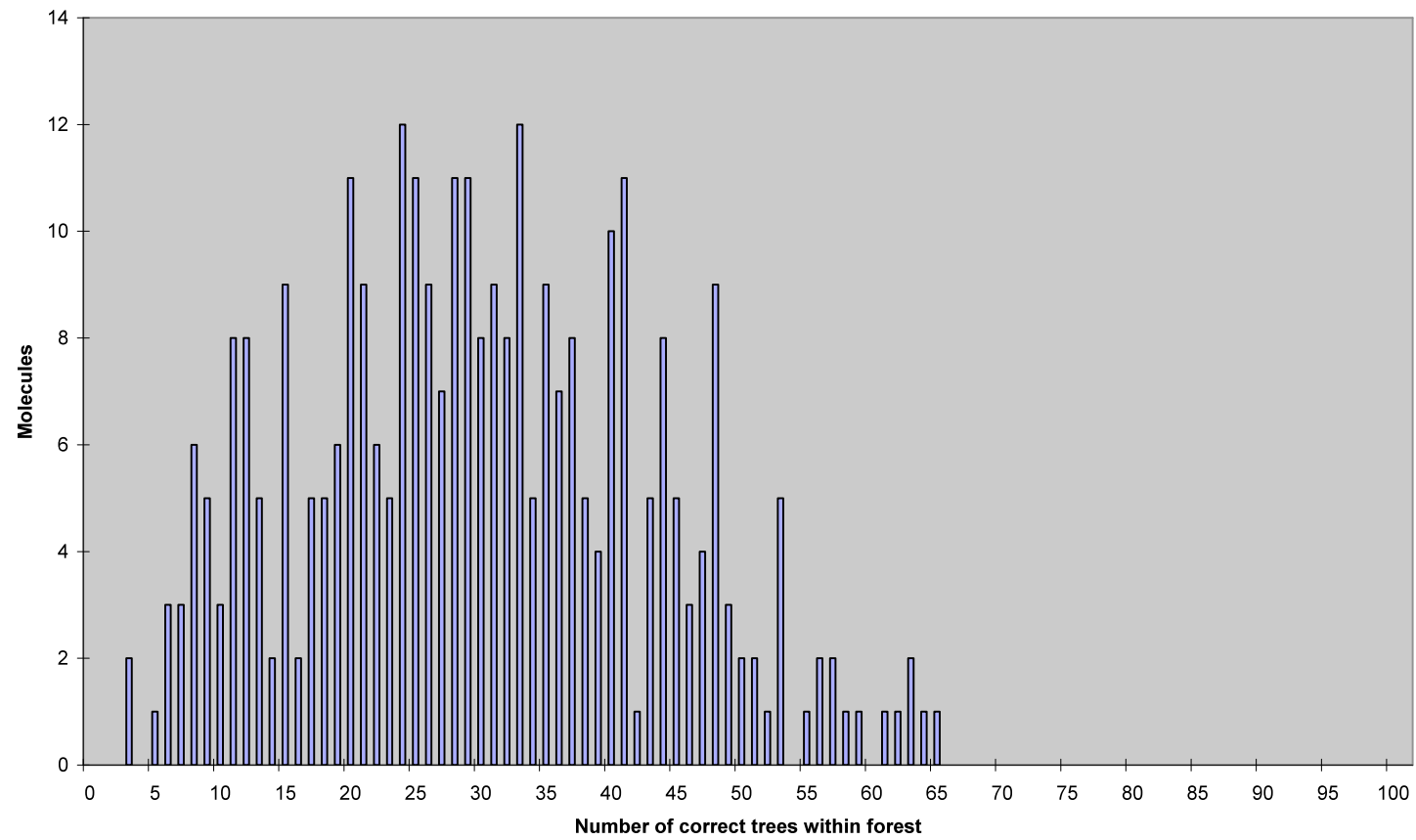
Multiple modes of action

- Can the forest detect them?
- Combined COX2 & DHFR data set

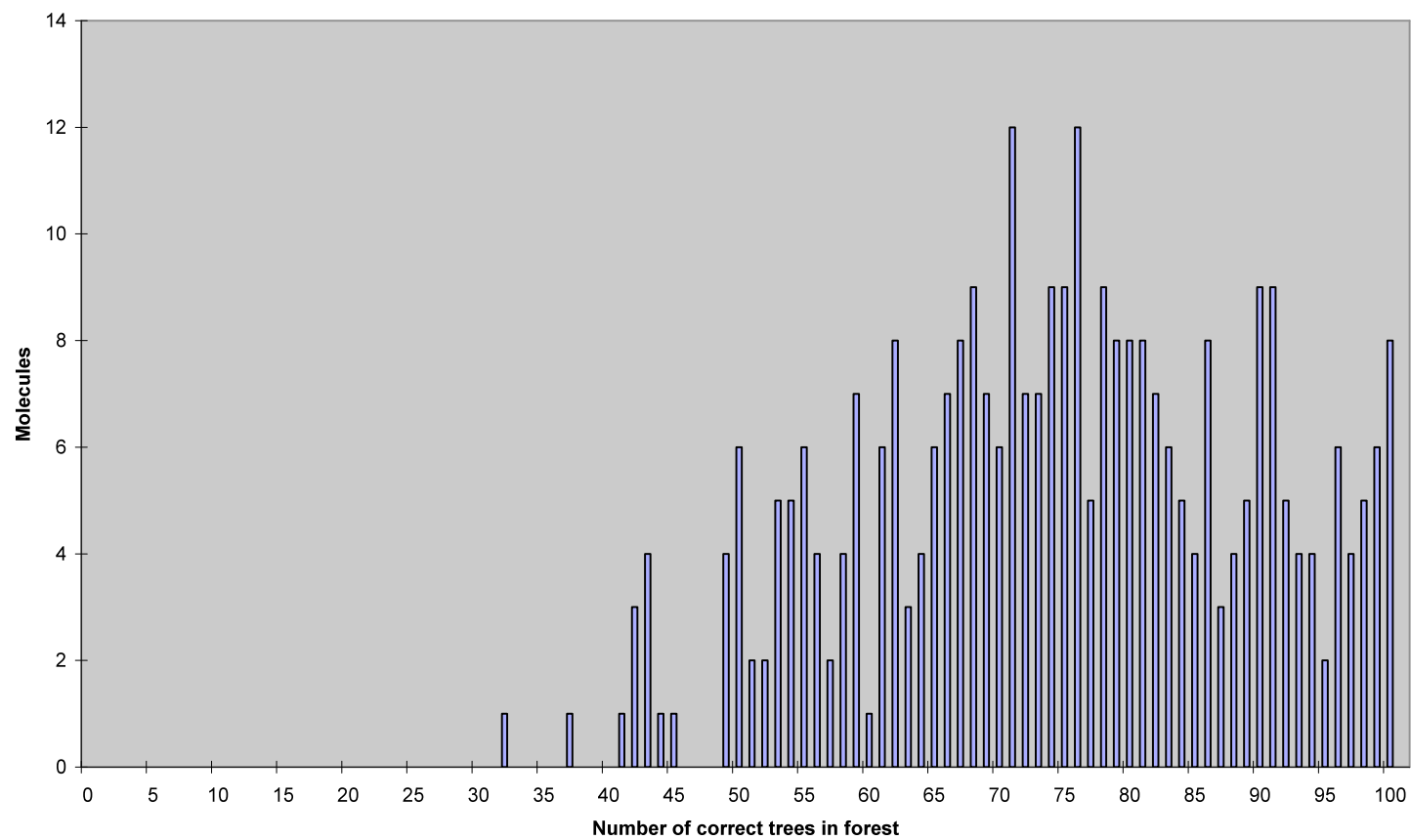
Fingerprint



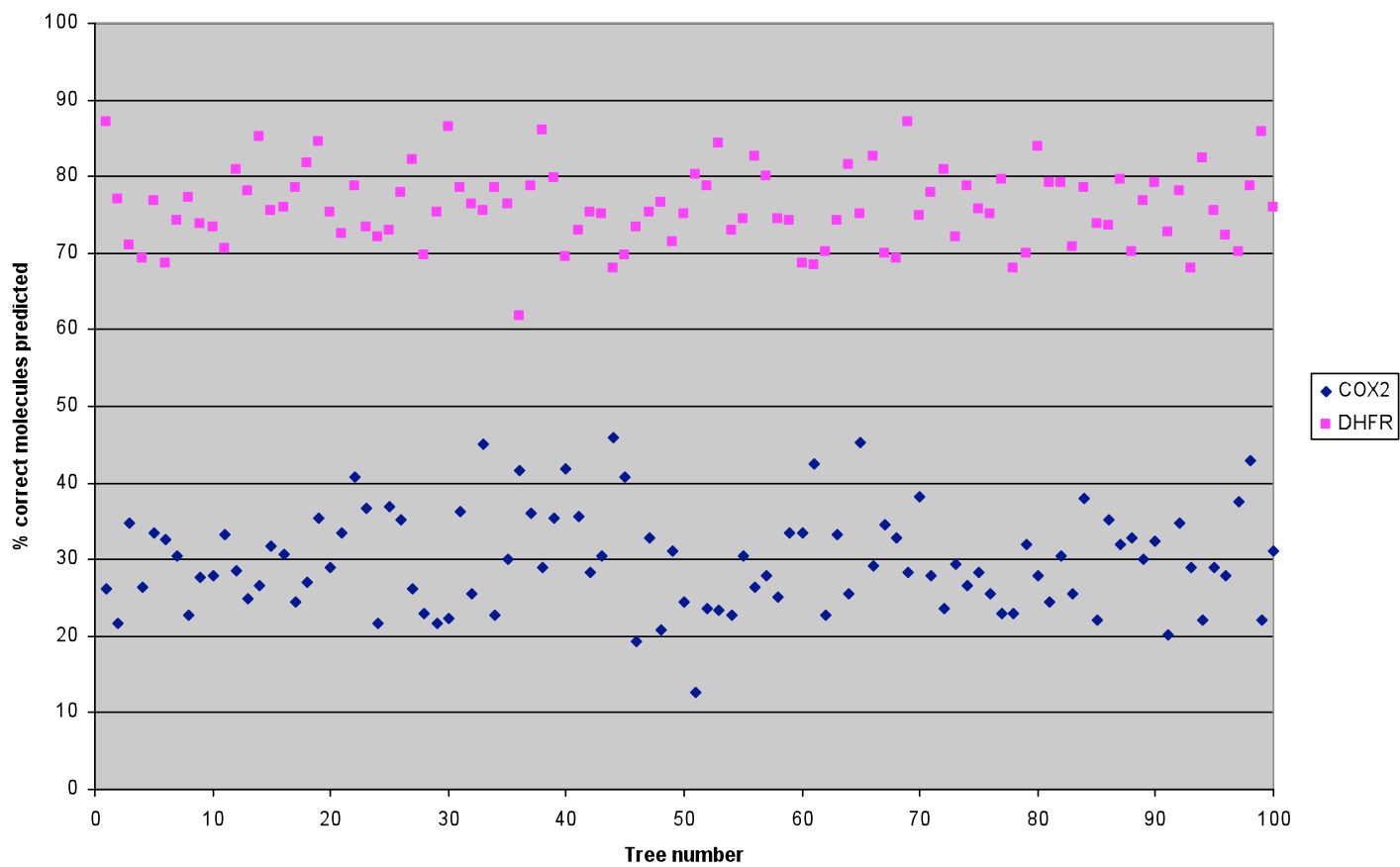
COX2



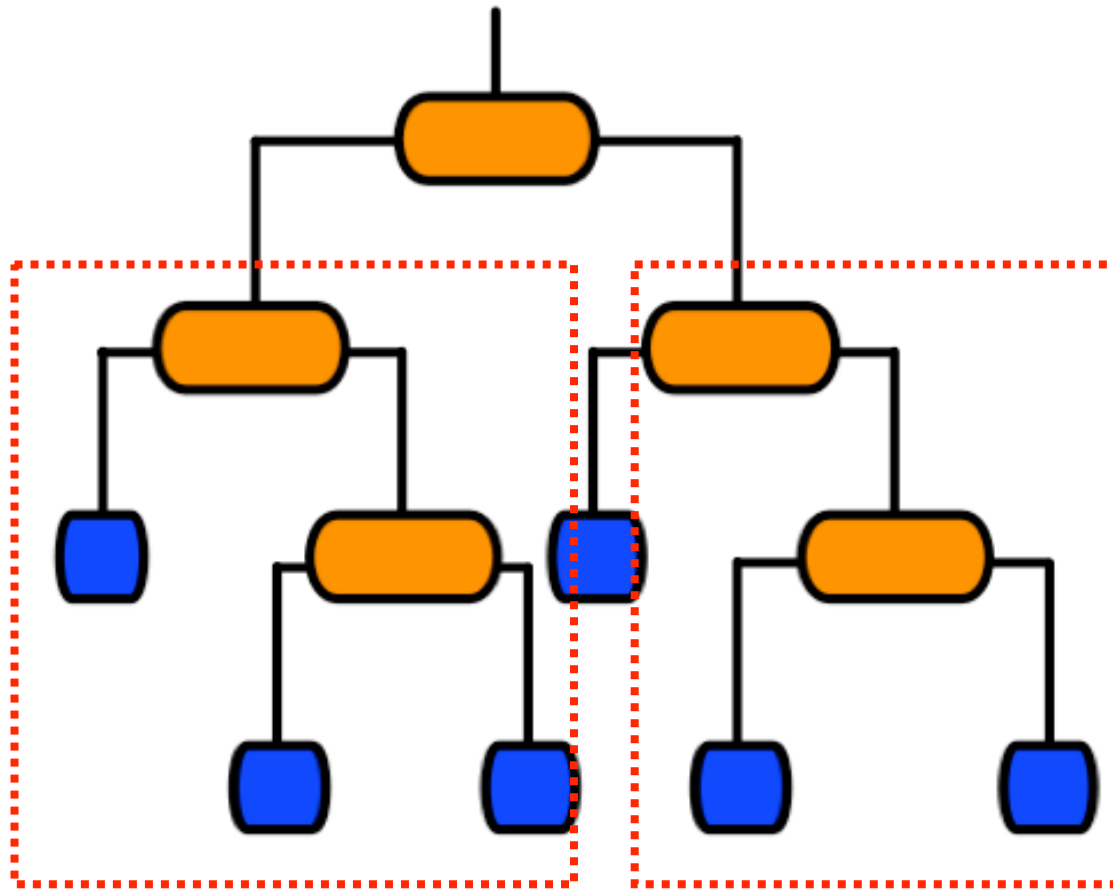
DHFR



Individual tree performance

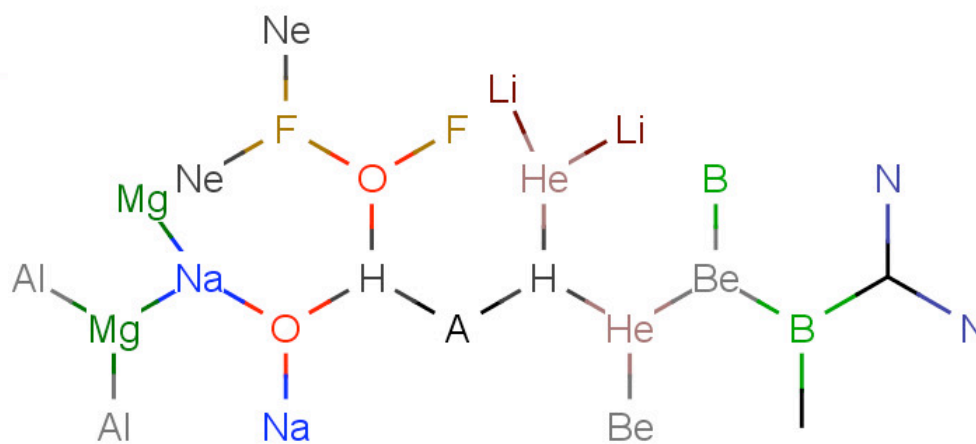
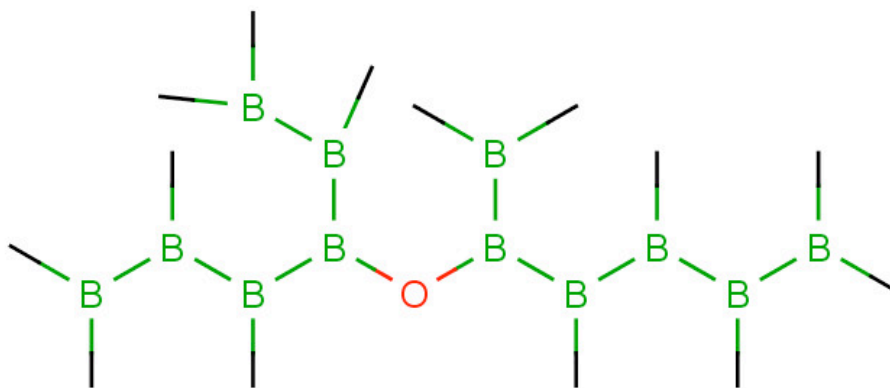


Trees are not class specific

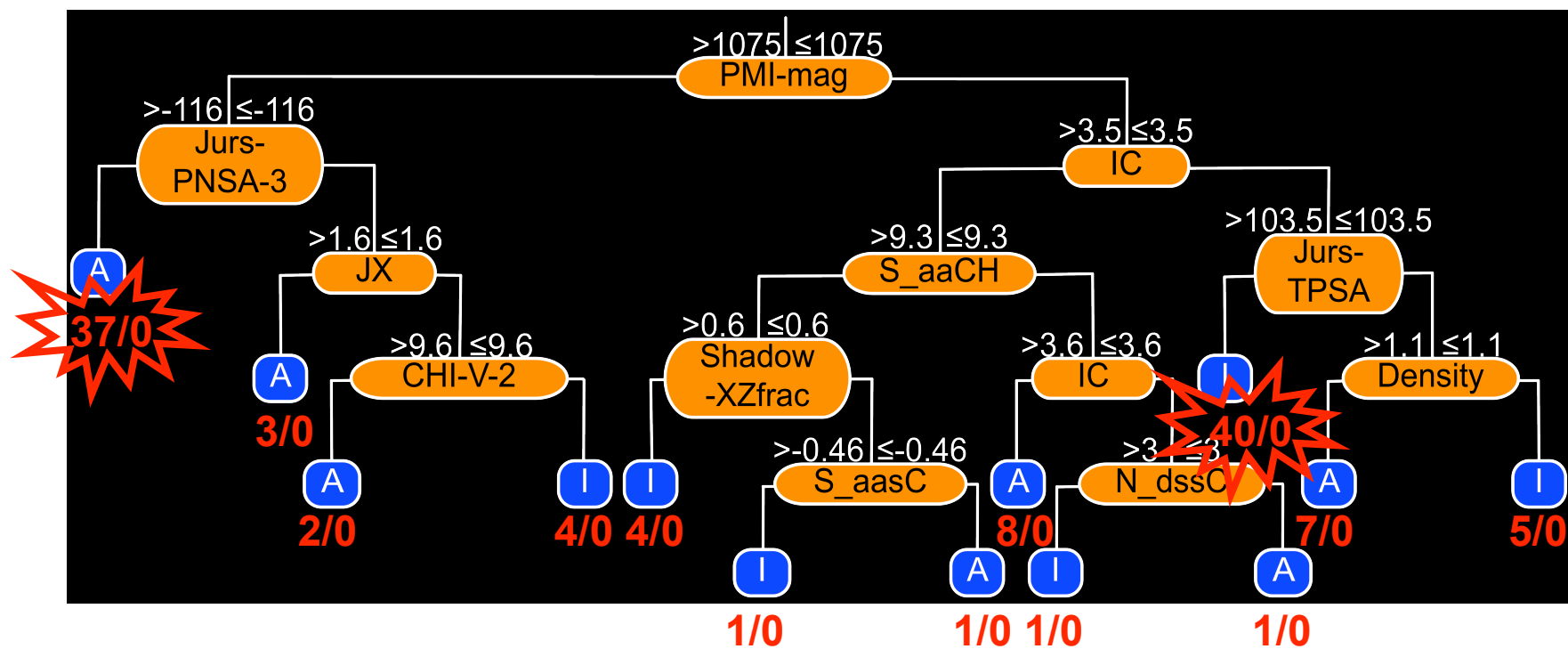


Tree representation

- SMILES
- SMARTS
- Sub graphs



Pathway analysis



Extract rules

- Active
 - PMI-mag > 1075
 - Jurs-PNSA-3 > -116
- Inactive
 - PMI-mag ≤ 1075
 - IC ≤ 3.5
 - Jurs-TPSA > 103.5

Conclusions

- Good QSAR requires good practise
- Random forest is a competitive classifier
- Suitable for real pharma data sets
- Interpretation is possible, but more complex than a single tree
- Building blocks present
- More predictive model with additional chemical insight

Acknowledgements



The University of
Nottingham

- Jonathan Hirst
- James Melville

- University of Nottingham HPC



GlaxoSmithKline

- Stephen Pickett
- Chris Luscombe
- Gavin Harper

EPSRC



The University of
Nottingham



GlaxoSmithKline

Thank you!

pcxcb1@nottingham.ac.uk
<http://comp.chem.nottingham.ac.uk>