optibrium

Improving the Chance of Success where an Outcome Can't be Predicted

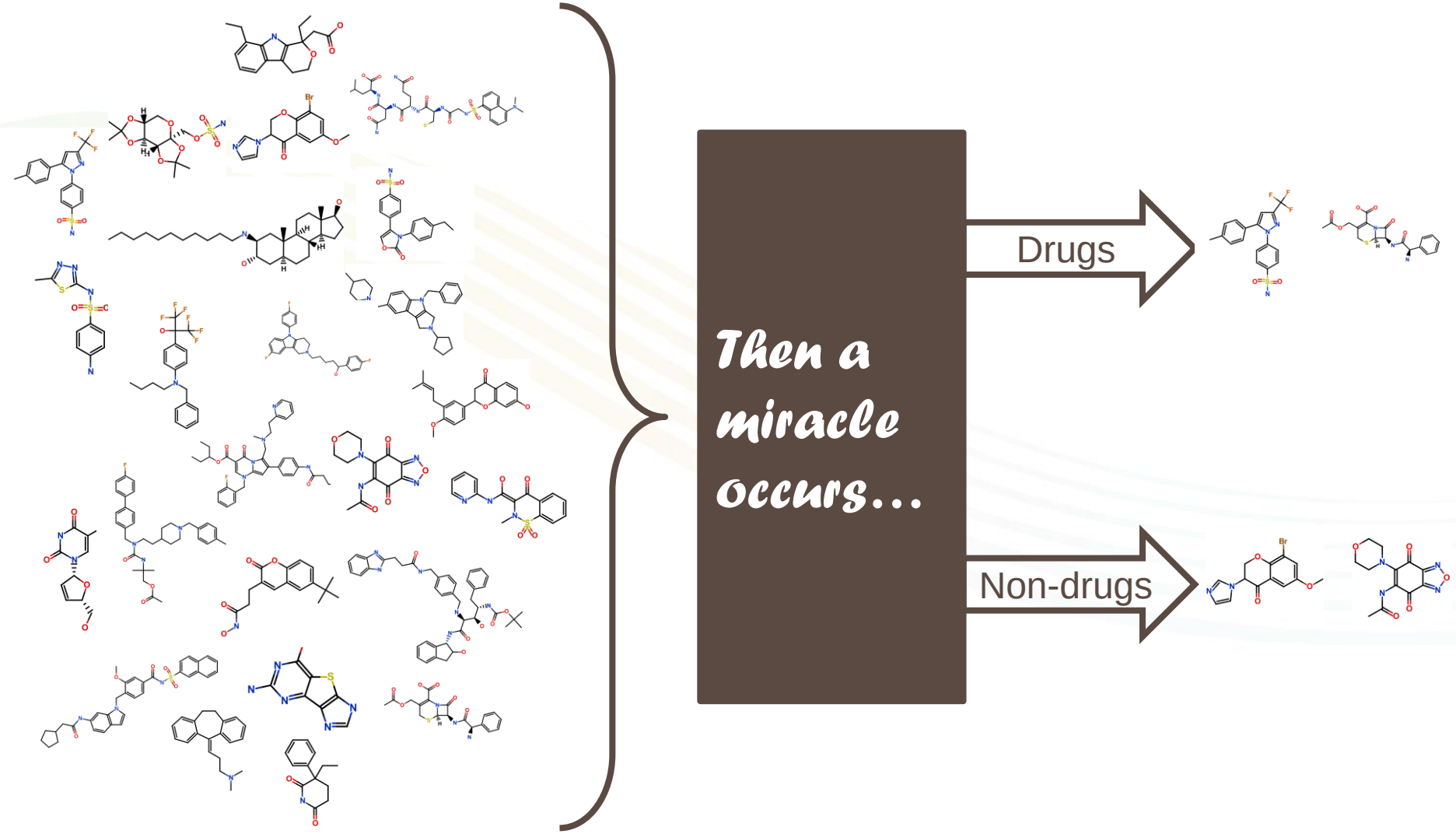
ACS Fall National Meeting, August 10th 2014

Matthew Segall, Iskander Yusof, Ed Champness

Overview

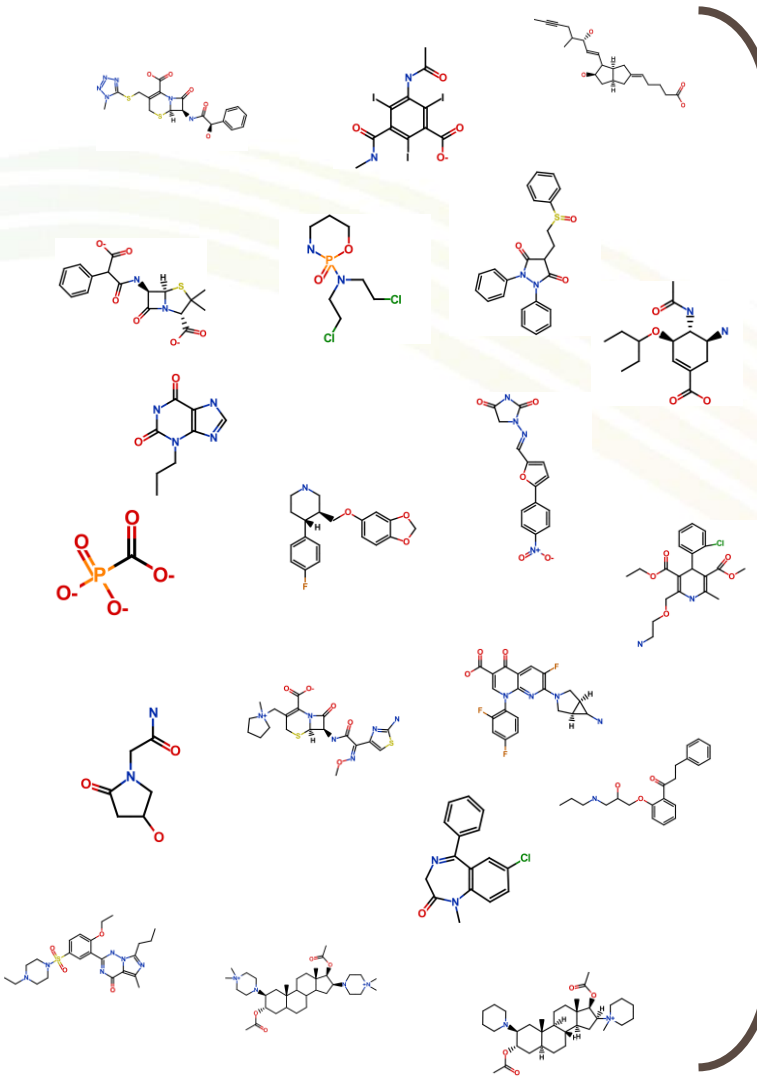
- Prediction... The ideal scenario
  - Why this is often not possible
- An alternative… rules that improve the chance of success
  - Finding rules
  - Applying rules
- Example 1: Finding CNS Drugs
- Example 2: Finding Long duration compounds
- Conclusions

In an Ideal World

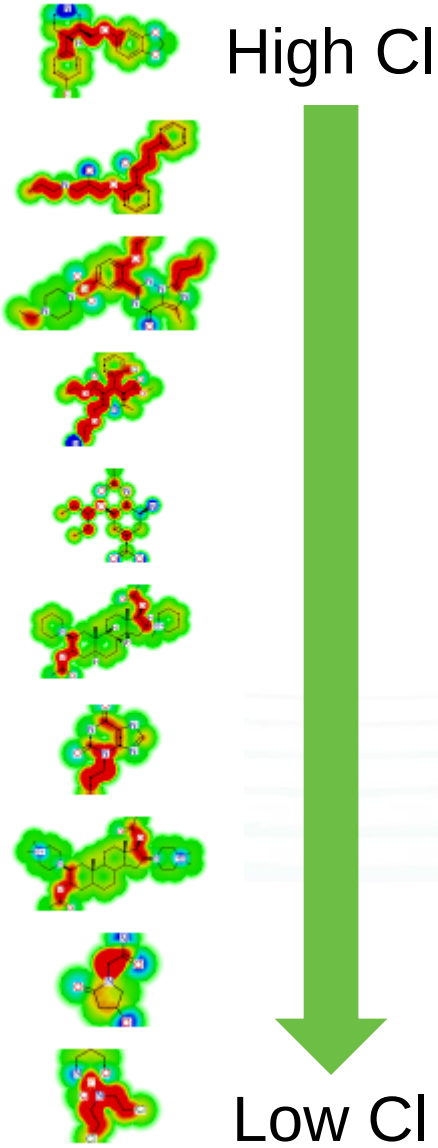


In an Ideal World

Predicting an endpoint



Then a miracle occurs…



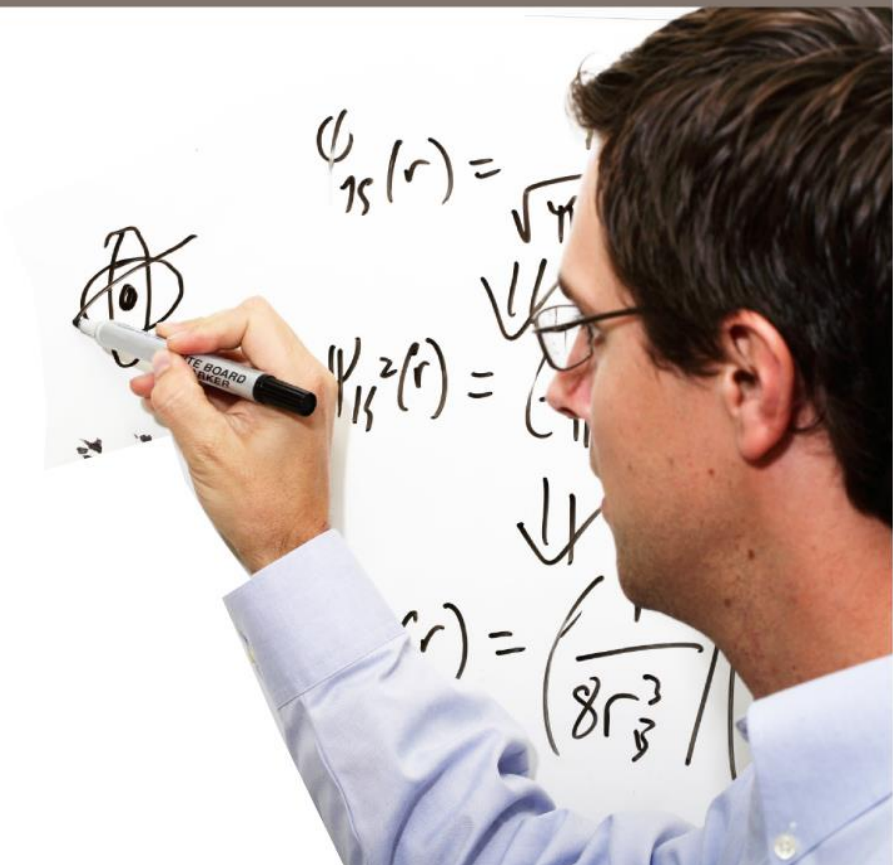
Reasons Why Prediction (often) Doesn't Work
Examples…

- Too much uncertainty
  - Variability in experimental data
  - Confidence in predictions too low
  - See talk 194, "Challenges of decision making using uncertain data" Room 2005, Moscone West, Tuesday 3.45 pm
- Process being modelled is too complex
  - E.g. multi-mechanistic, physiological processes
- Not enough data
  - Bias in available data

Finding Rules for Success

Patent pending

optibrium



What is a Rule?

- A **Rule** is a set of property criteria that in **combination** identify 'good' compounds, e.g.

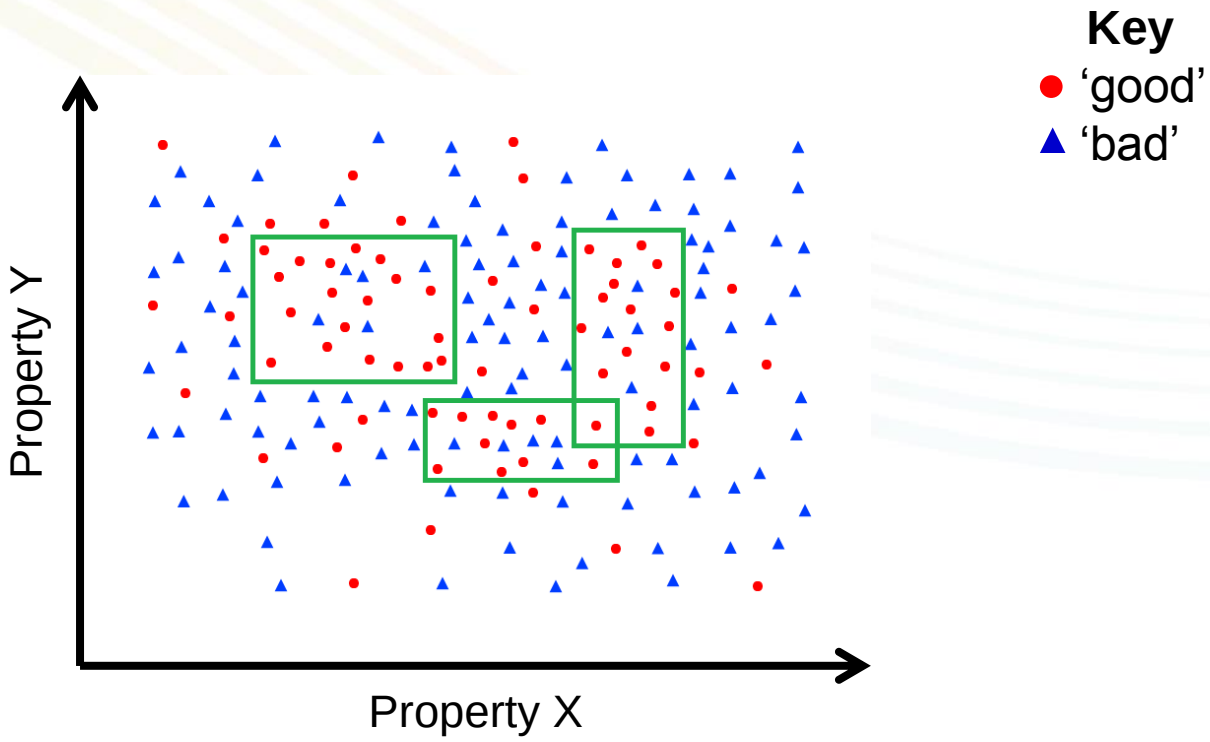
logP < 4
Ligand efficiency > 0.3
100 < MW < 450
PPB category = low

- For example, Lipinski's Rule of Five:

logP<5	MW<500
HBD<5	HBA<10

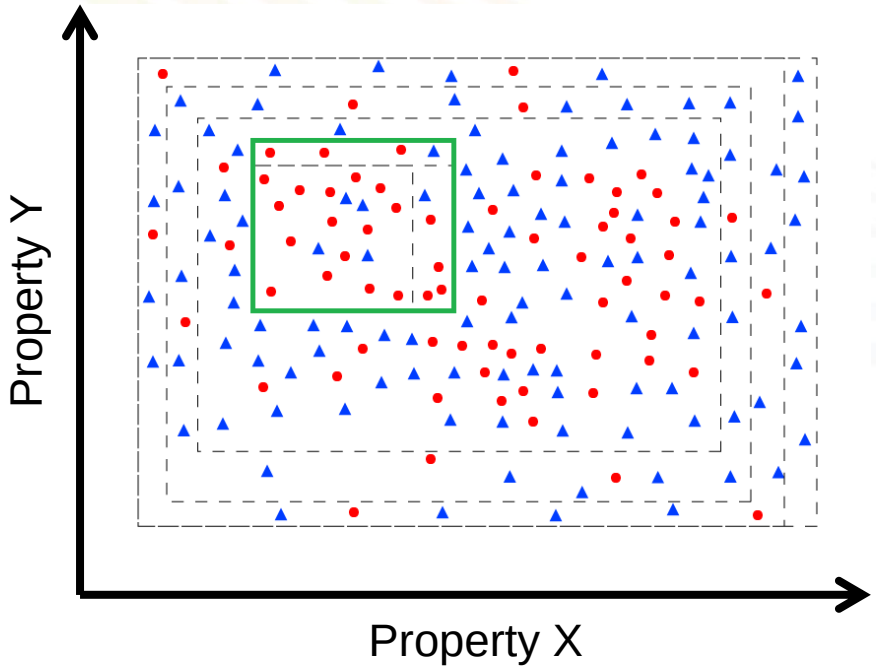
What is a Rule?

- A **Rule** is also a box in multi-dimensional property space containing significantly more ‘good’ than ‘bad’ compounds



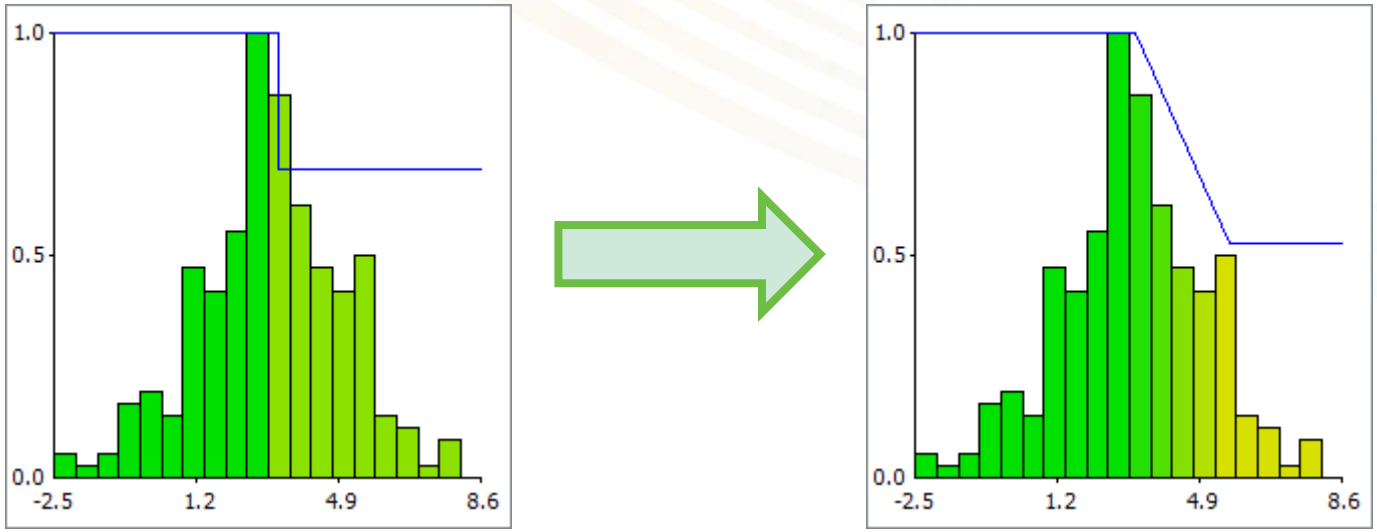
Rule Induction

- 'Rule induction' method identifies multi-parameter regions of property space with higher chance of success
  - Also known as 'bump hunting' because it can find property regions corresponding to small increases in probability distribution



Determining 'Soft' Box Boundaries

- Box bounds from rule induction are hard cut-offs
- Sensitivity analysis of box bounds to data sampling
  - Particularly important for sparse data
  - Incorporate uncertainty into the generated box bounds
  - Cross validation between training/validation sets

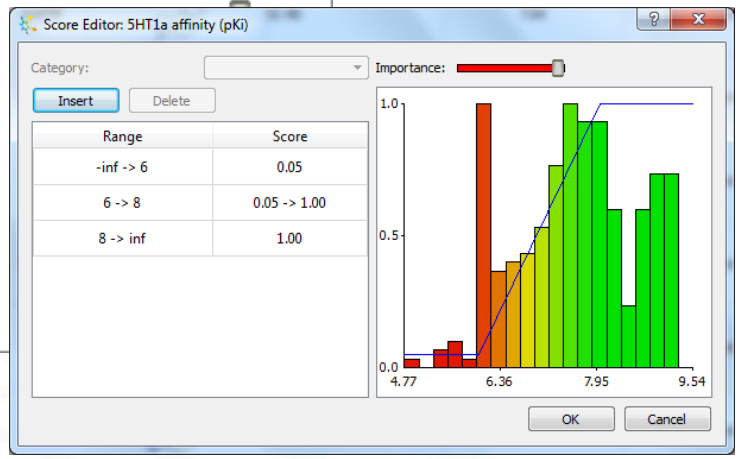


Measuring Rule Performance

- Mean = Average objective value in box
  - Reported as % increase over objective value for full set
- Support = Proportion of data set ‘covered’ by box
  - Reported as % coverage
- Specificity vs. Sensitivity trade-off
  - Specify minimum coverage to avoid overtraining
- Also reported
  - Statistical significance (p-value)
  - Odds ratio (probability of success inside the box vs. outside)

Applying Multi-Parameter Rules
Probabilistic Scoring*

Profile	Desired Value	Importance
5HT1a affinity (pKi)	8 -> inf	
logS	> 1	
HIA category	+	
logP	0 -> 3.5	
BBB log([brain]:[blood])	-0.2 -> 1	
BBB category	+	
P-gp category	no	
hERG pIC50	⩽ 5	
2C9 pKi	⩽ 6	
2D6 affinity category	low medium	
PPB90 category	low	



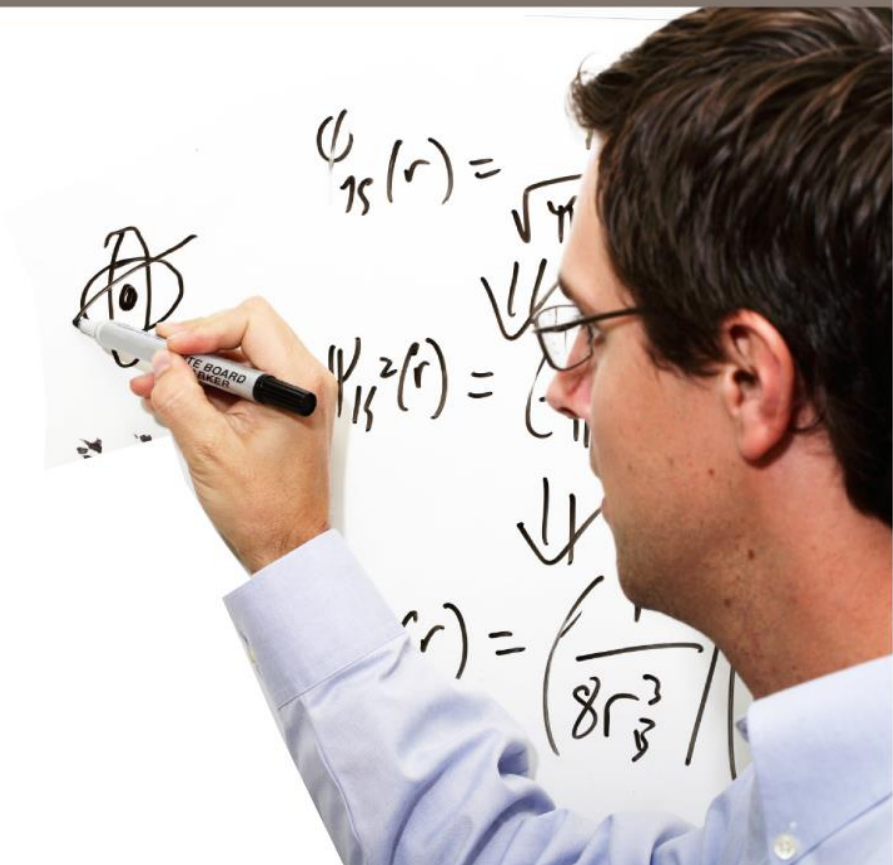
- **Property data**
  - Experimental or predicted
- **Criteria for success**
  - Relative importance
- **Uncertainties in data**
  - Experimental or statistical

- **Score (Likelihood of Success)**
- **Confidence in score**

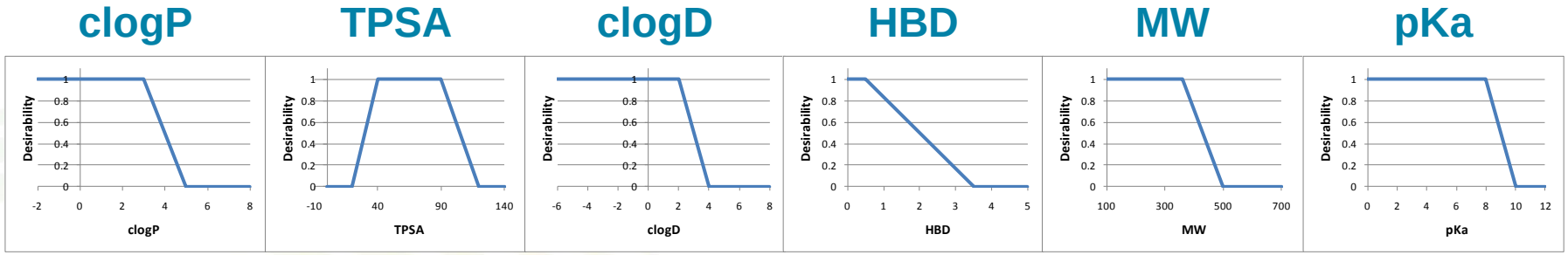
* M.D. Segall (2012) Curr. Pharm. Des. 18(9) pp. 1292-1310

Example: Finding CNS Drugs

optibrium



Finding CNS Drugs
CNS MPO Score*



CNS MPO = sum of desirabilities for each parameter

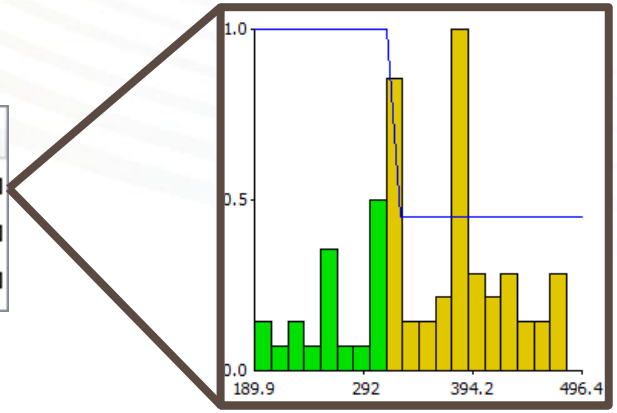
- 74% of marketed CNS drugs achieved CNS MPO > 4 vs. 60% of Pfizer candidates
- Correlations observed between high CNS MPO score and good *in vitro* ADME properties, e.g. MDCK P_{app}, HLM stability, P-gp transport

*Wager *et al.* (2010) ACS Chem. Neurosci. **1** p. 435

Finding CNS Drugs
Applying Rule Induction

- Data set of 119 CNS drugs and 108 failed candidates published by Wager *et al.**
- Divided into training and validation sets (70:30)
- Rule derived with 20% minimum coverage:

Profile	Desired Value	Importance
MW	-inf -> 312.5	
PKA	-inf -> 9.676	
CLOGP	-inf -> 2.973	



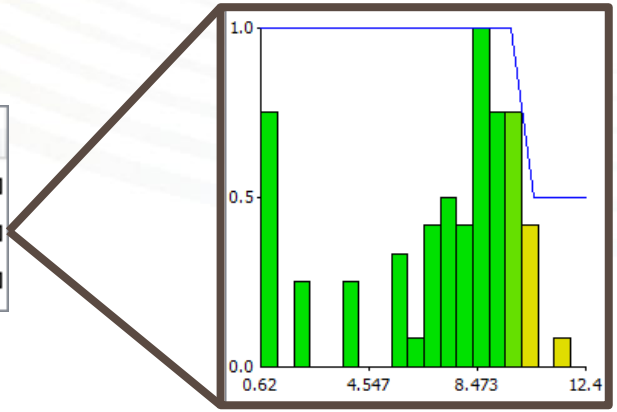
MW

*Wager *et al.* (2010) ACS Chem. Neurosci. **1** p. 435

Finding CNS Drugs
Applying Rule Induction

- Data set of 119 CNS drugs and 108 failed candidates published by Wager *et al.**
- Divided into training and validation sets (70:30)
- Rule derived with 20% coverage:

Profile	Desired Value	Importance
MW	-inf -> 312.5	
PKA	-inf -> 9.676	
CLOGP	-inf -> 2.973	



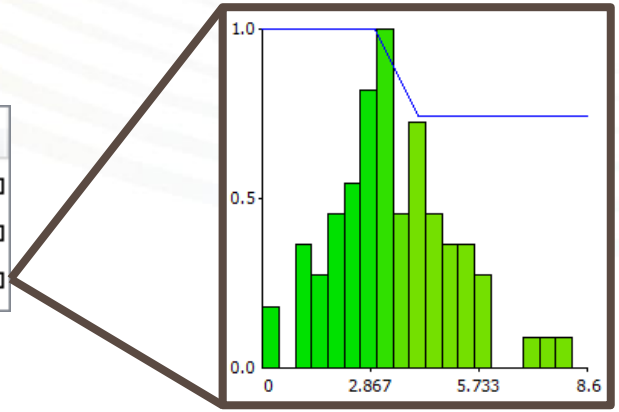
PKA

*Wager *et al.* (2010) ACS Chem. Neurosci. **1** p. 435

Finding CNS Drugs
Applying Rule Induction

- Data set of 119 CNS drugs and 108 failed candidates published by Wager *et al.**
- Divided into training and validation sets (70:30)
- Rule derived with 20% coverage:

Profile	Desired Value	Importance
MW	-inf -> 312.5	
PKA	-inf -> 9.676	
CLOGP	-inf -> 2.973	

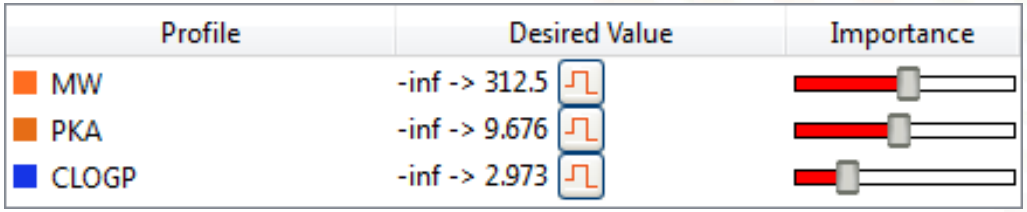


CLOGP

*Wager *et al.* (2010) ACS Chem. Neurosci. **1** p. 435

Finding CNS Drugs
Applying Rule Induction

- Data set of 119 CNS drugs and 108 failed candidates published by Wager *et al.**
- Divided into training and validation sets (70:30)
- Rule derived with 20% minimum coverage:

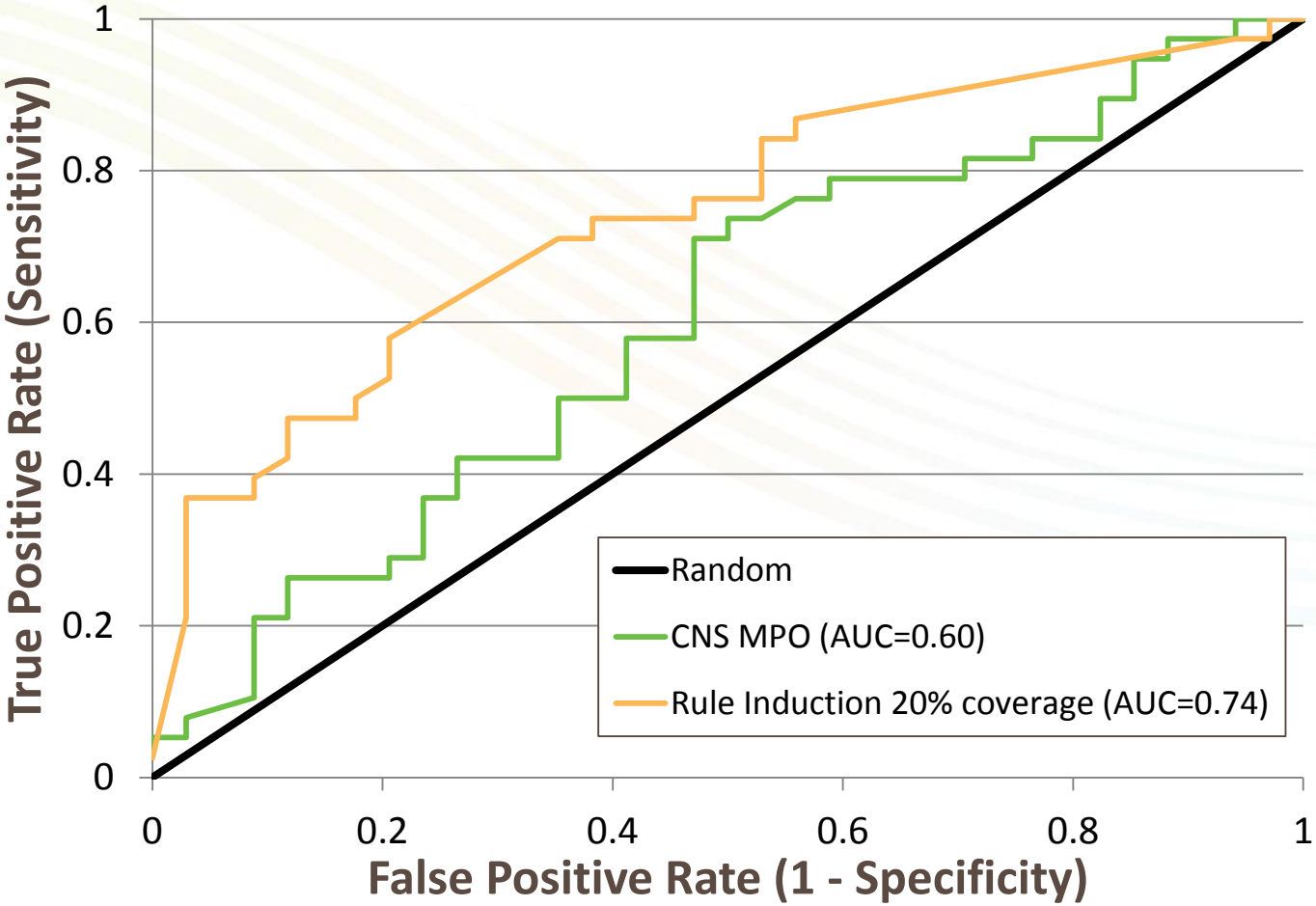


Set	Mean Improvement (%)	Support (%)	Odds Ratio	p-value
Train	34	36	3.3	3×10^{-3}
Val	67	24	10.4	2×10^{-4}

*Wager *et al.* (2010) ACS Chem. Neurosci. **1** p. 435

Finding CNS Drugs
Validation Results – ROC plot

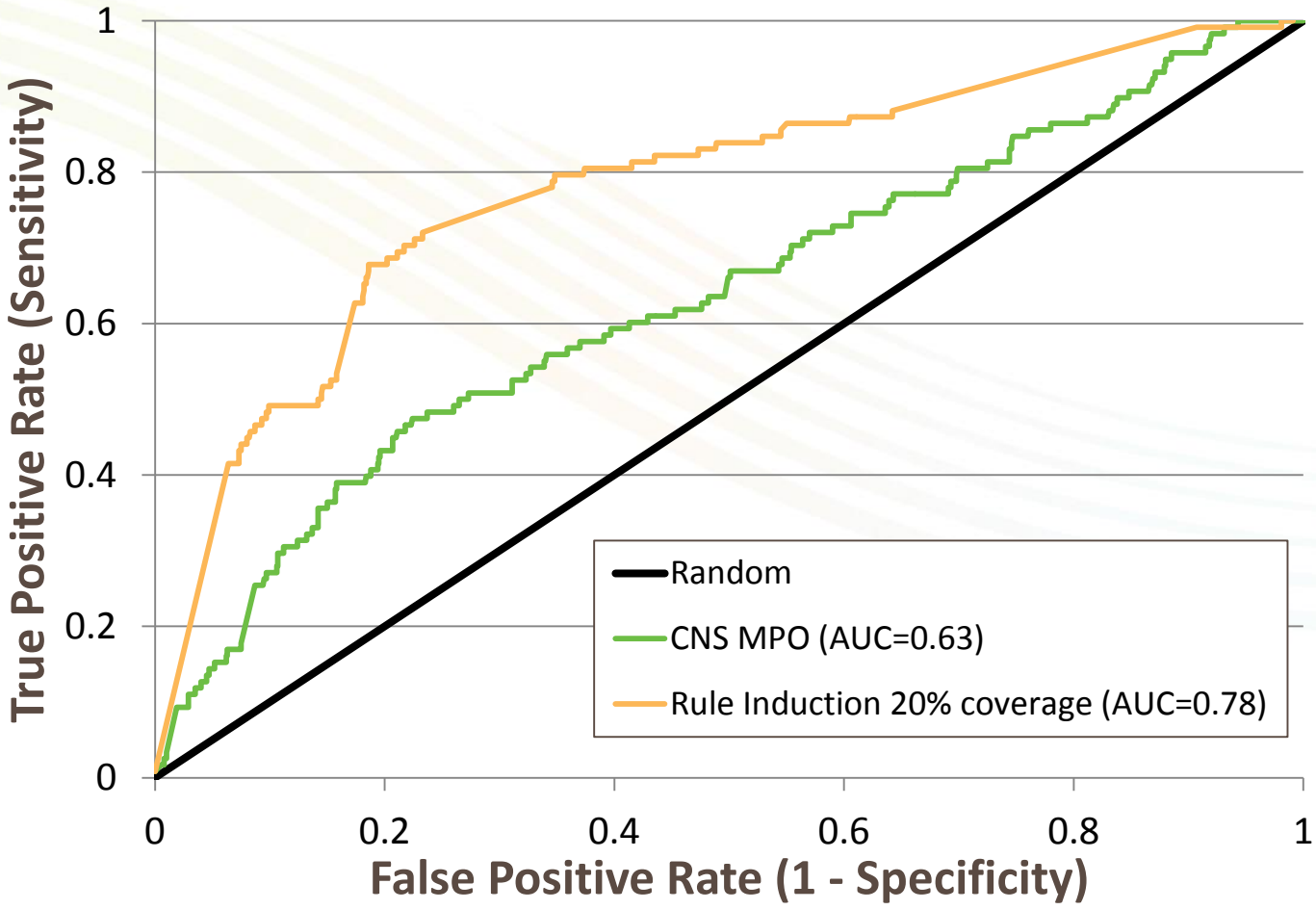
- 38 CNS drugs and 34 failed candidates from Wager dataset*



*Wager *et al.* (2010) ACS Chem. Neurosci. **1** p. 435

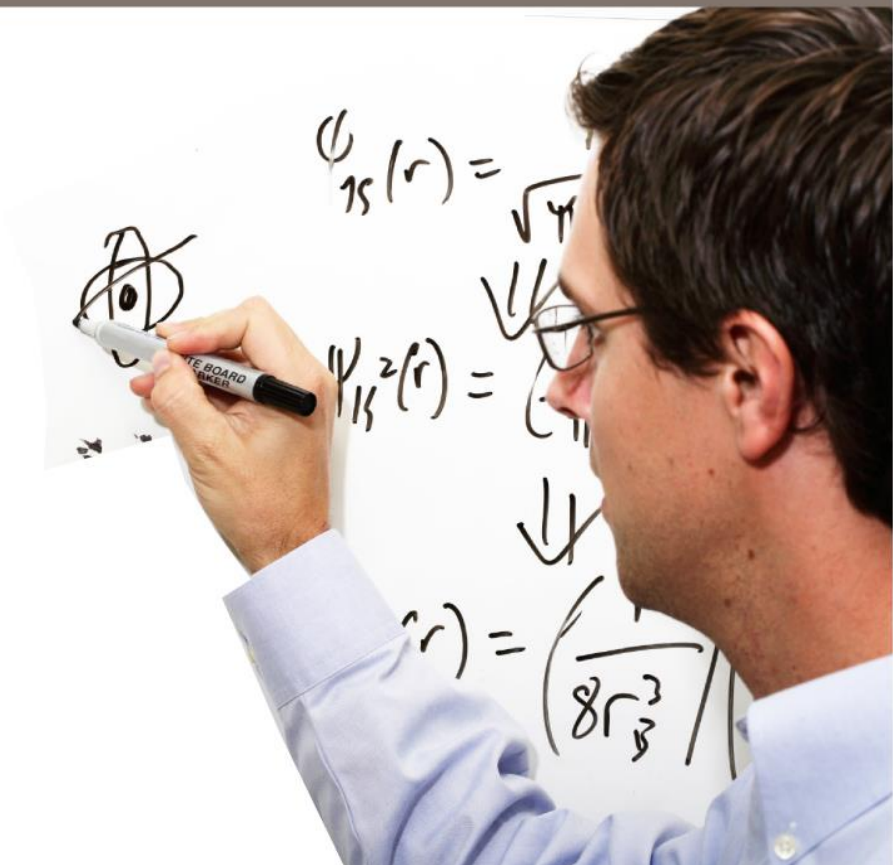
Finding CNS Drugs
A more realistic external test

- 118 (different) CNS drugs and 1000 CNS 'leads' (measured K_i < 1 µM against CNS target) from ChEMBL database



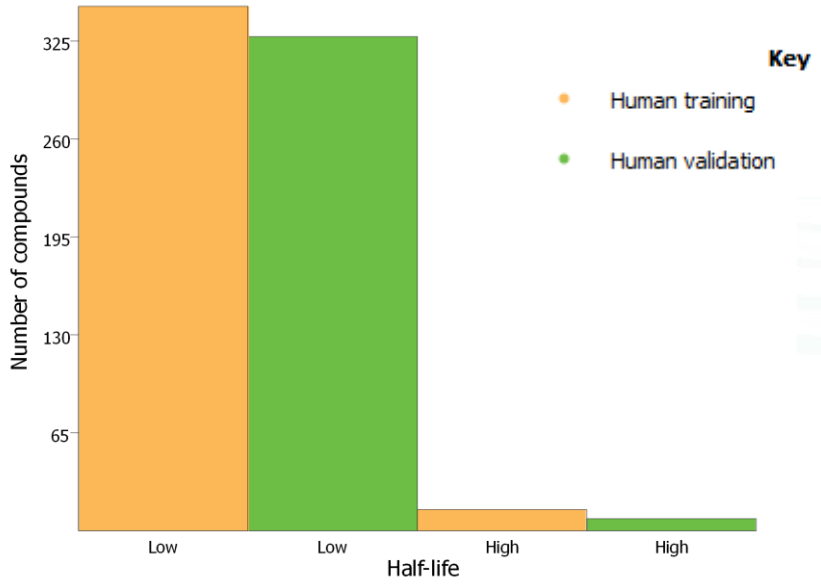
Example: Finding Long Duration Compounds

optibrium



Finding Long Duration Compounds

- Find rules to identify compounds with a half-life ($T_{½}$) in humans >100 hours
- Data set: 698 compounds with measured human $T_{½}$
  - Divided into training and validation sets (52:48) using clustering
  - Highly biased data set

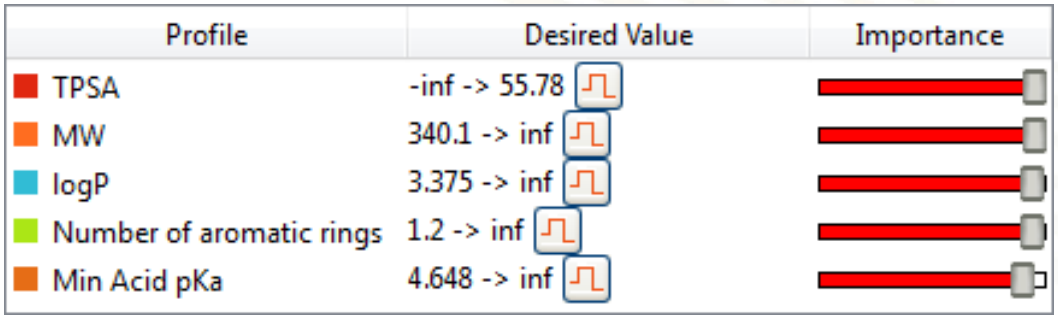


Finding Long Duration Compounds
'Conventional' modelling techniques

- Classification models:
  - Random forest, decision trees, Gaussian process classifier
- Descriptors including:
  - Simple compound properties: MW, logP, TPSA, HBA/D, ROTB, AROM
  - Ionisation: pKa*, Acid/Base/Zwitterion/Neutral indicator
  - QSAR predictions: logS, PPB, P-gp transport, BBB…
- No 'High' validation set compound correctly predicted
  - N.B. Accuracy is 97%... Beware accuracy as metric for biased data!

Finding Long Duration Compounds
Rule induction

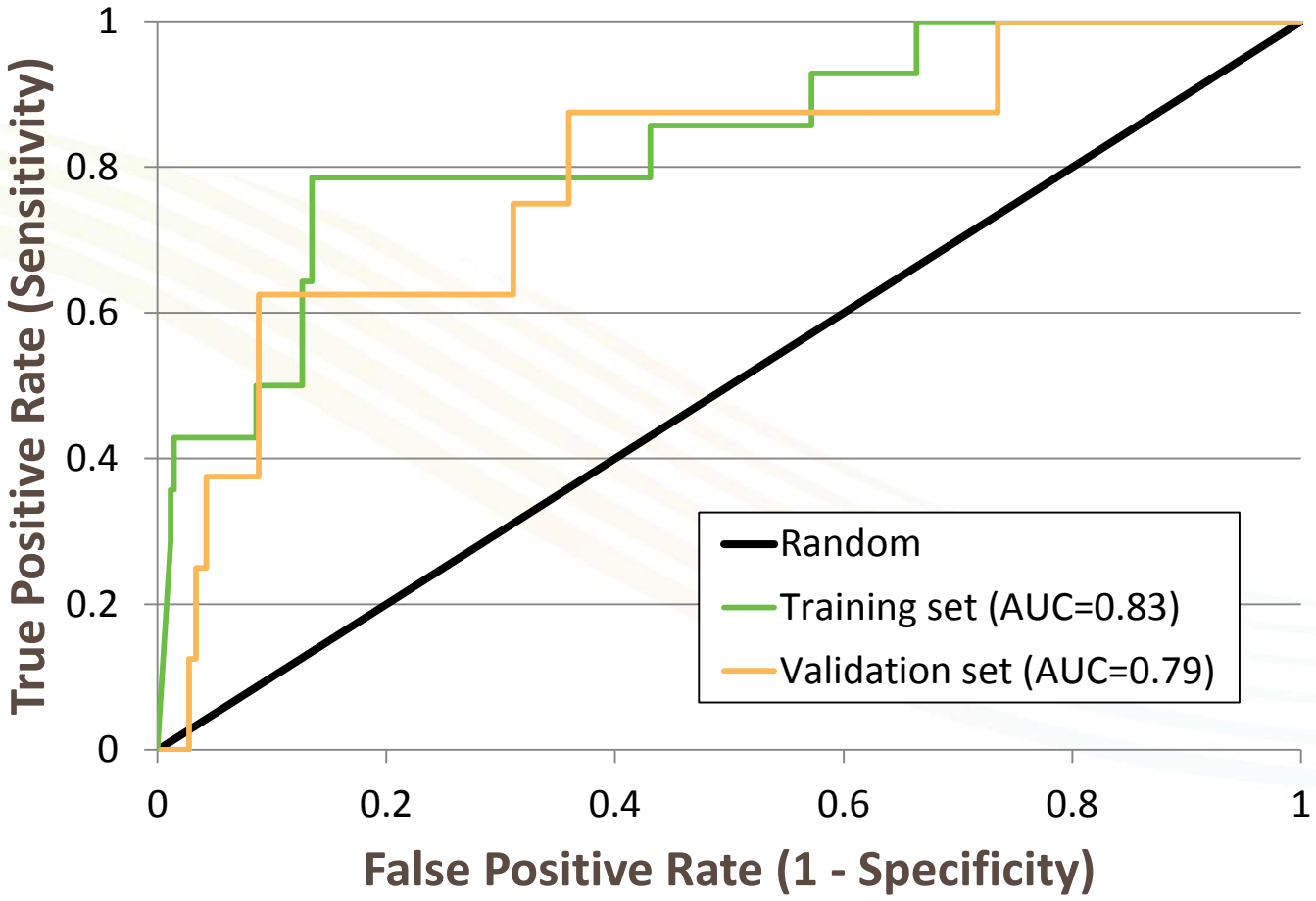
- Descriptors:
  - Simple compound properties: MW, logP, TPSA, HBA/D, ROTB, AROM
  - Ionisation: pKa*, Acid/Base/Zwitterion/Neutral indicator
  - QSAR predictions: logS, PPB, P-gp transport, BBB…
- Rule derived with 2% minimum coverage:



Set	Mean Improvement (%)	Support (%)	Odds Ratio	p-value
Train	269.4	19.3	12	8×10^{-3}
Val	425.0	11.9	14	8×10^{-2}

Finding Long Duration Compounds

ROC plot



Conclusions

- In many cases it is not possible to predict an outcome with confidence
  - Often due to sparse or biased data
- Rule induction provides a way to find multi-parameter selection criteria that improve the chance of success
- Multi-parameter optimisation provides a robust way to apply these rules and bias the odds in our favour
- For more details, see:
  - I. Yusof et al. (2014) Drug Discov. Today 19(5) pp. 680-687
  - Download (p)reprint from www.optibrium.com/community/publications
- Or visit **Booth 1324** for a demo

Acknowledgements

- Tatsu Hashimoto - MIT
- The Optibrium team, including:
  - Chris Leeding
  - James Chisholm
  - Nick Foster
  - Peter Hunt
  - Alex Elliott
  - Jon Tyzack
  - Sam Dowling
- For the long duration project:
  - Joelle Gola
  - Mark Gardner – AMG Consultants