optibrium

Finding Multi-parameter Rules for Successful Optimization

ACS Spring National Meeting, April 10th 2013

Matthew Segall, Iskander Yusof, Edmund Champness

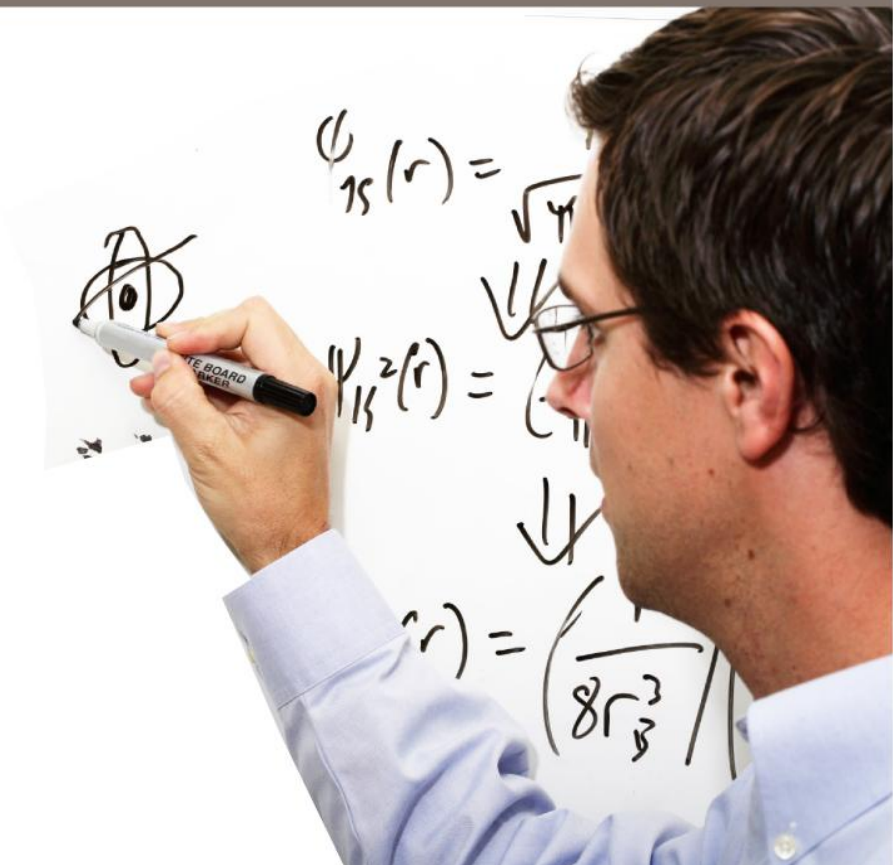
Patent pending

Overview

- Probabilistic scoring for multi-parameter optimization (MPO)
- Finding multi-parameter rules for drug discovery
- Methods
  - Rule induction
- Illustrative applications
  - ‘Drug-like’ properties
  - Oral CNS compounds
- User interaction
- Conclusions

Probabilistic Scoring for MPO

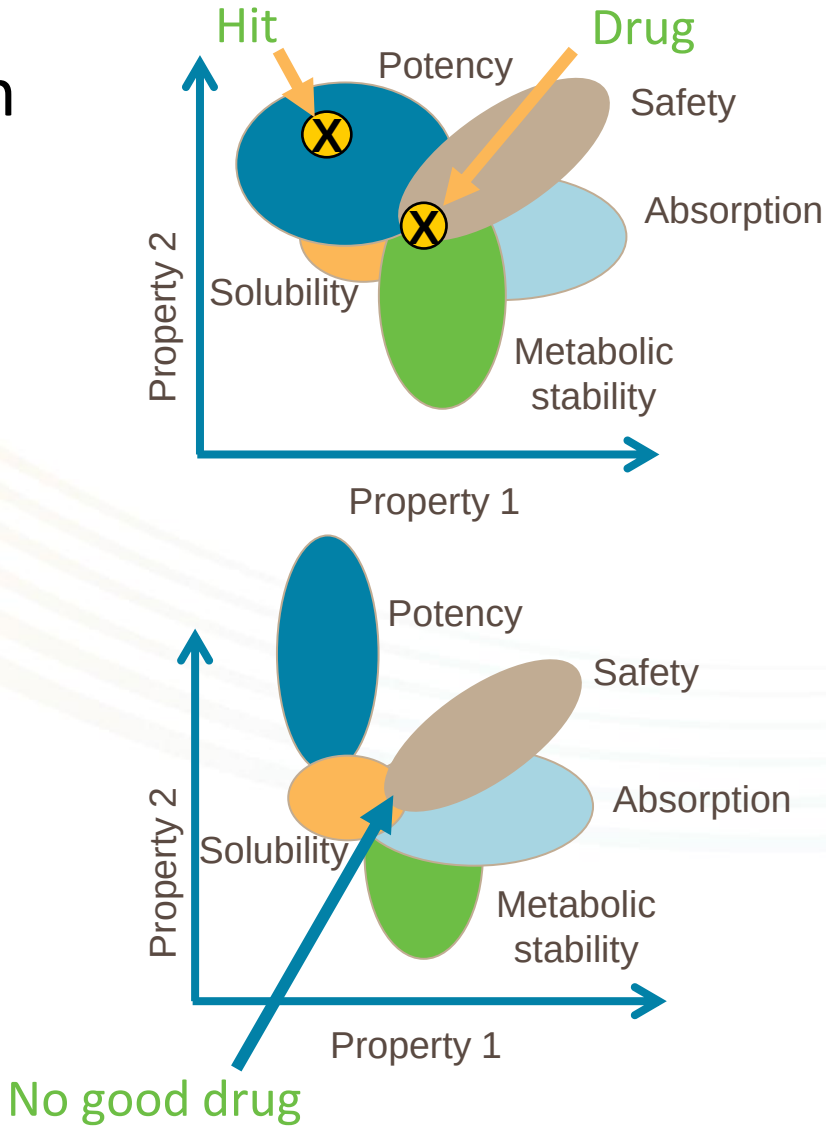
optibrium



The Objectives of MPO

- Identify chemistries with an optimal **balance** of properties

- Quickly identify situations when such a balance is not possible
  - –Fail fast, fail cheap
  - –Only when **confident**



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

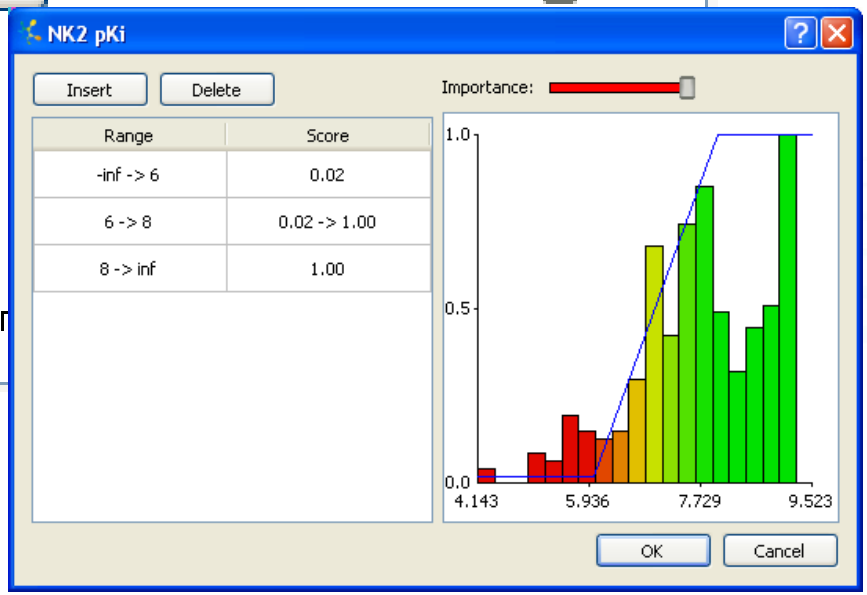
Requirements for MPO in Drug Discovery

- Interpretable
  - Easy to understand compound priority and how to improve compounds' chances of success
- Flexibility
  - Define criteria depending on therapeutic objectives of a project
- Weighting
  - Take into account relative importance of different endpoints to the success of a project
- Uncertainty
  - Take uncertainty into account, avoid missed opportunities

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

Probabilistic Scoring
Scoring Profile

Property	Desired Value	Importance
pIC50	> 8	
logS	> 1	
BBB log([brain]:[blood])	-0.2 -> 1	
Cytotox. (pLD50)	⩽ 5	
logP	0 -> 3.5	
P-gp category	no	
Ames	-	
hERG pIC50	⩽ 5	
2C9 pKi	⩽ 6	
2D6 affinity category	low mediu	

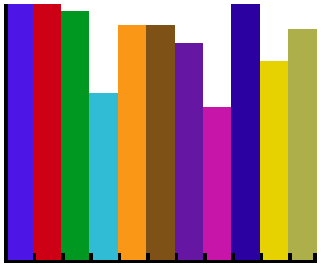


* Segall *et al.* Chem. & Biodiv. **6** p. 2144 (2009)

StarDrop Prioritisation
Probabilistic Scoring

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

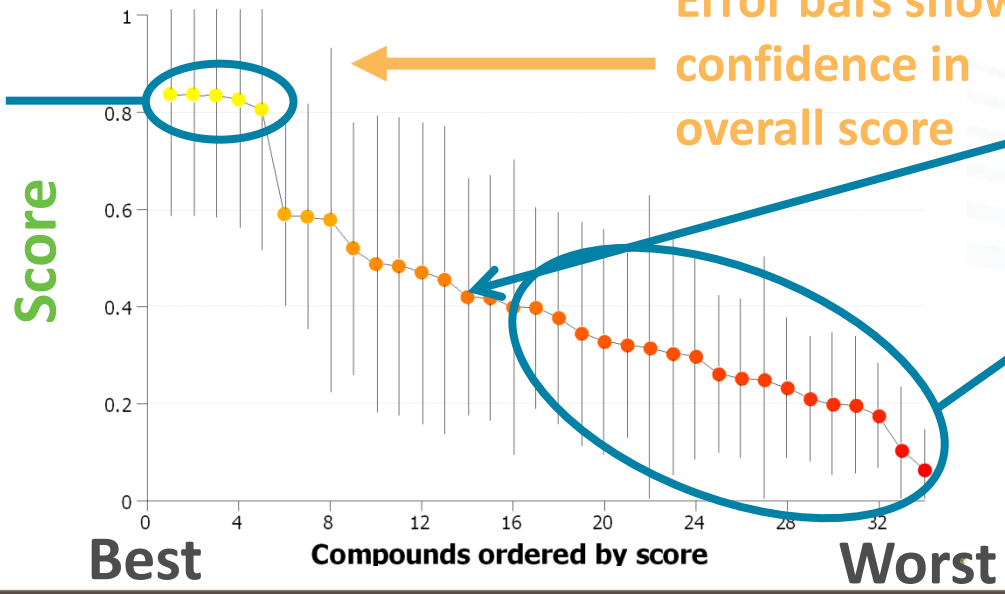
- **Score (Likelihood of Success)**
- **Confidence in score**



Data do not separate these as error bars overlap

Error bars show confidence in overall score

Bottom 50% may be rejected with confidence



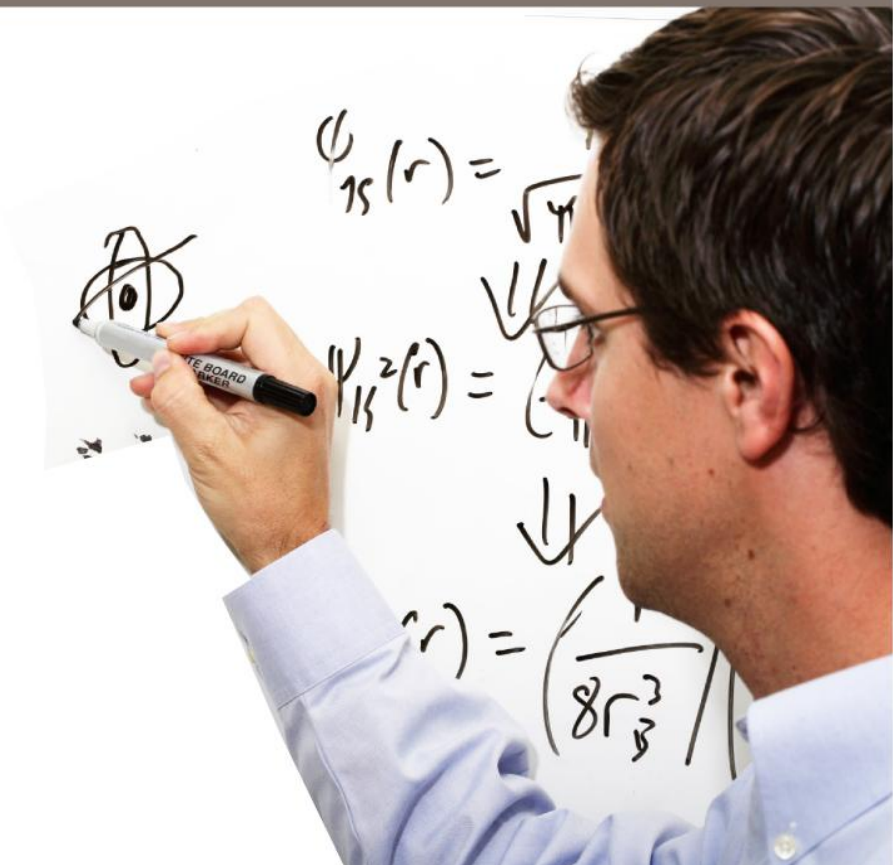
* Segall *et al.* Chem. & Biodiv. **6** p. 2144 (2009)

The Next Challenge

- How do we choose an appropriate scoring profile?
- Two approaches:
  - Domain/expert knowledge
  - Find the profile *automatically* using existing data
- Can we score compounds automatically without losing the benefits of expert knowledge?
  - Avoid ‘black boxes’
  - Maintain interpretability and interactivity

Finding Multi-Parameter Rules for Drug Discovery

optibrium



Objectives and Challenges

- Use historical data to find scoring profiles with which to identify compounds with improved chance of success
  - Any drug discovery objective, e.g. clinical, PK, toxicity...
  - Once developed, profile can be applied prospectively to find new compounds
- Identify most important data with which to distinguish between successful and unsuccessful compounds
  - Any data can be used as input, calculated or experimental
- Explore multi-parametric data
  - Consider properties simultaneously, not individually
  - Avoid ‘over counting’ of correlated factors
- Rules must be interpretable and modifiable
  - Avoid black boxes
  - Synergy between computer and experts

What is a Rule?

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

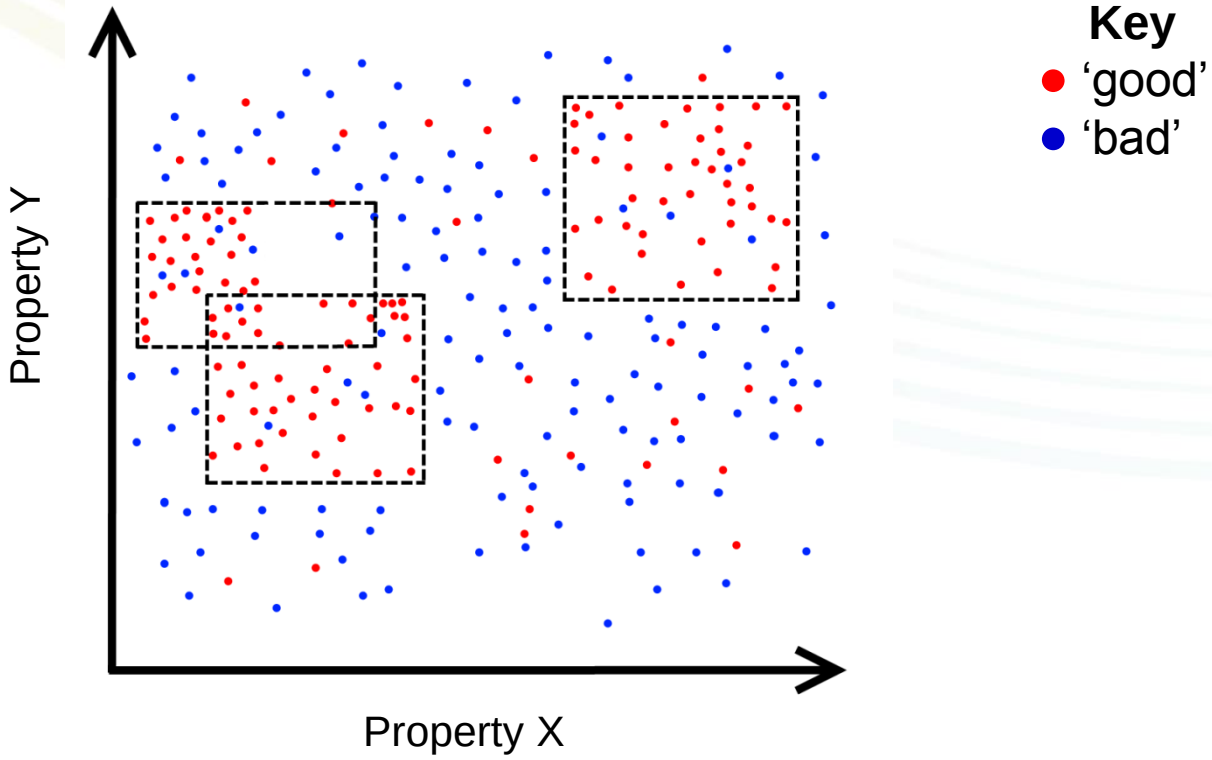
logP < 4
Ligand efficiency > 0.3
100 < MW < 450
PPB category = low

- For example, Lipinski RoF:

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

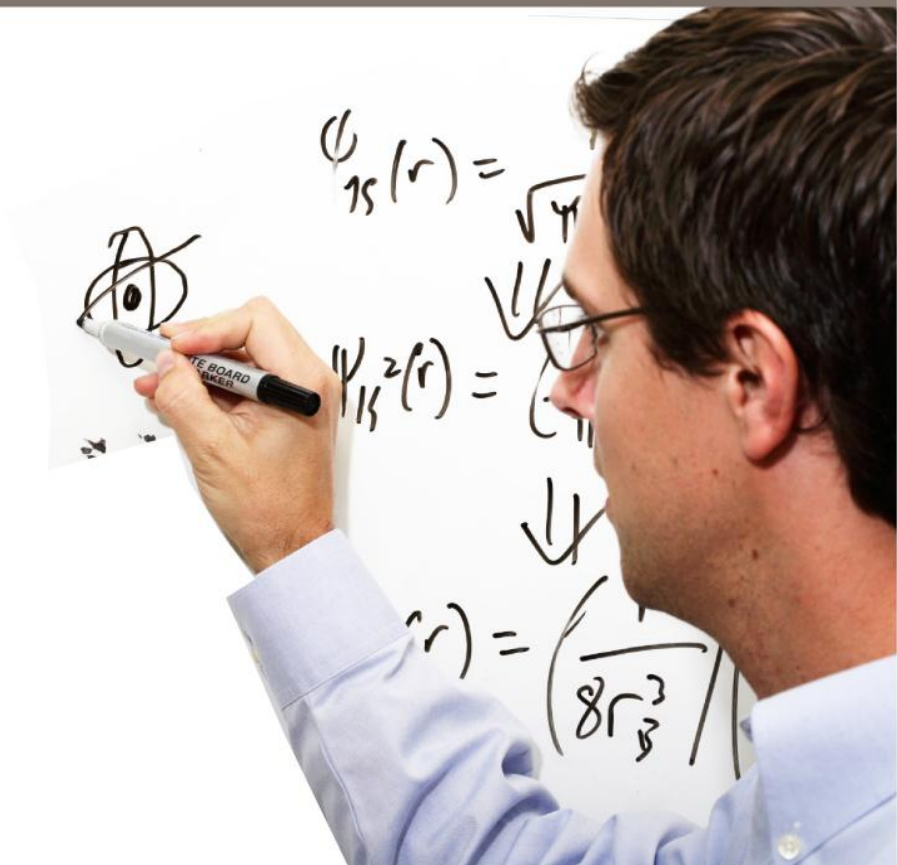
What is a Rule?

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



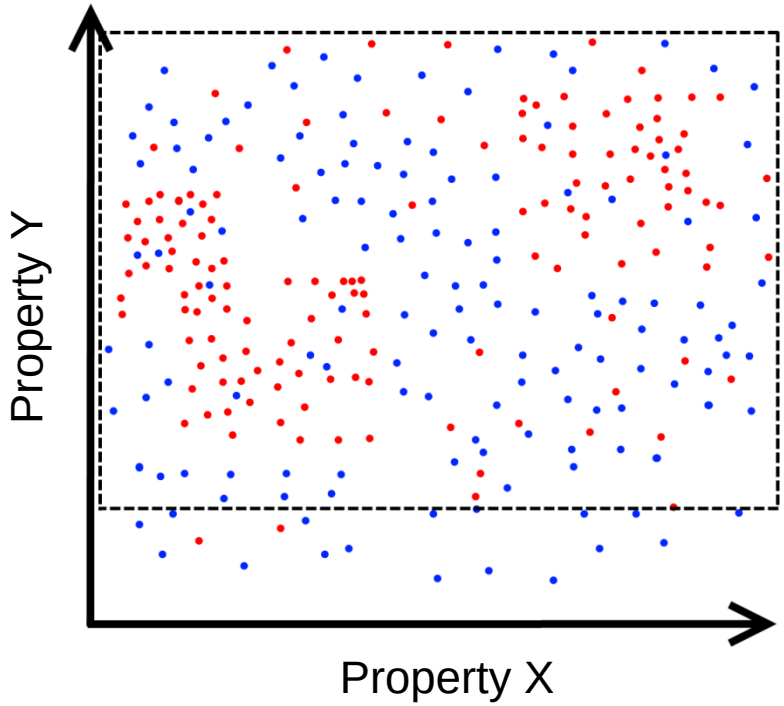
Methods

optibrium



Rule Induction with PRIM

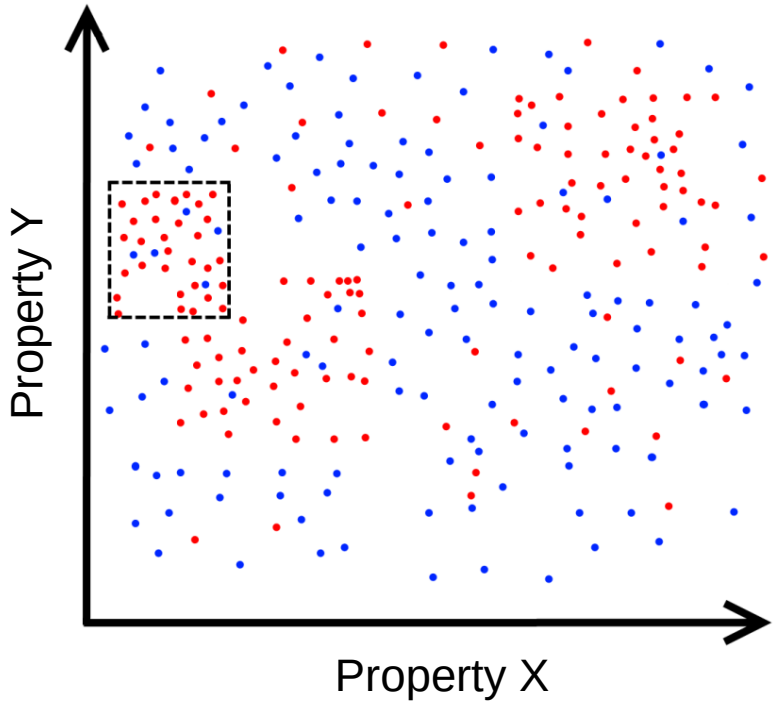
- The Patient Rule Induction Method (PRIM) by Friedman and Fisher is an effective way to find rules
- **Top-down peeling:** Start with a box covering all the compounds
- Then repeatedly peel the “worst” sides of the current box



* Friedman & Fisher Stat. and Comp. **9**(2), p. 123 (1999)

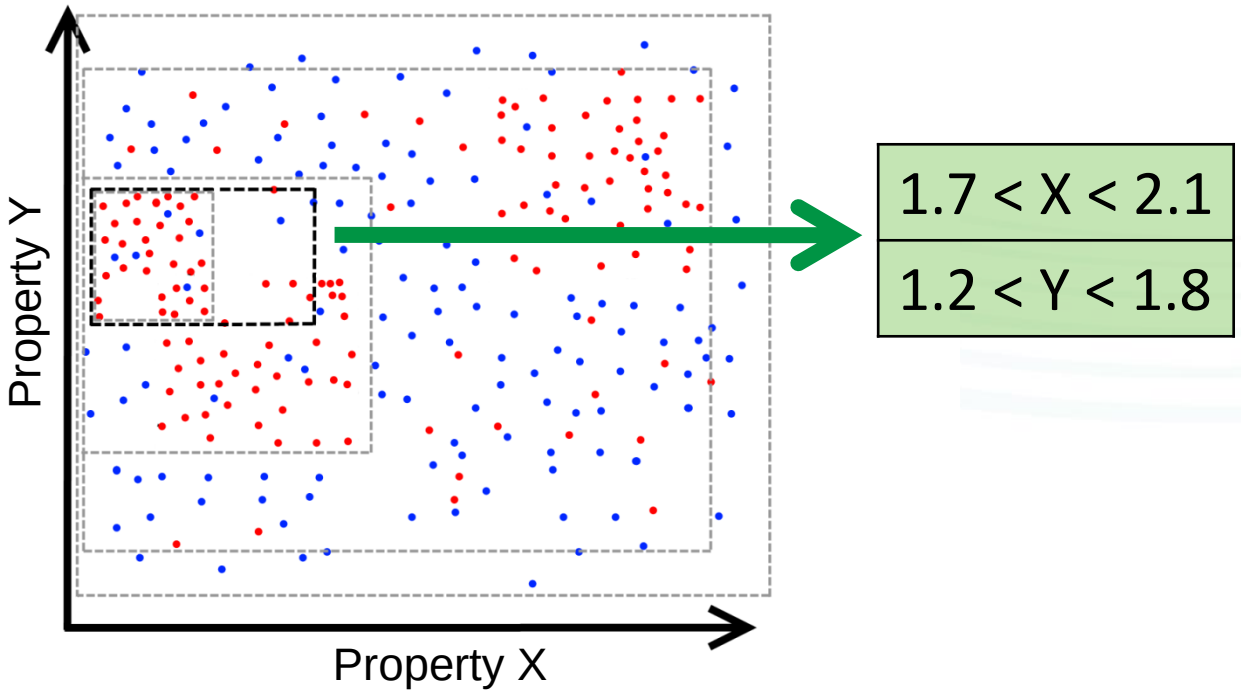
Rule Induction with PRIM

- **Bottom-up pasting:** “Paste” back regions that we overzealously peeled
- We stop when pasting provides no improvement



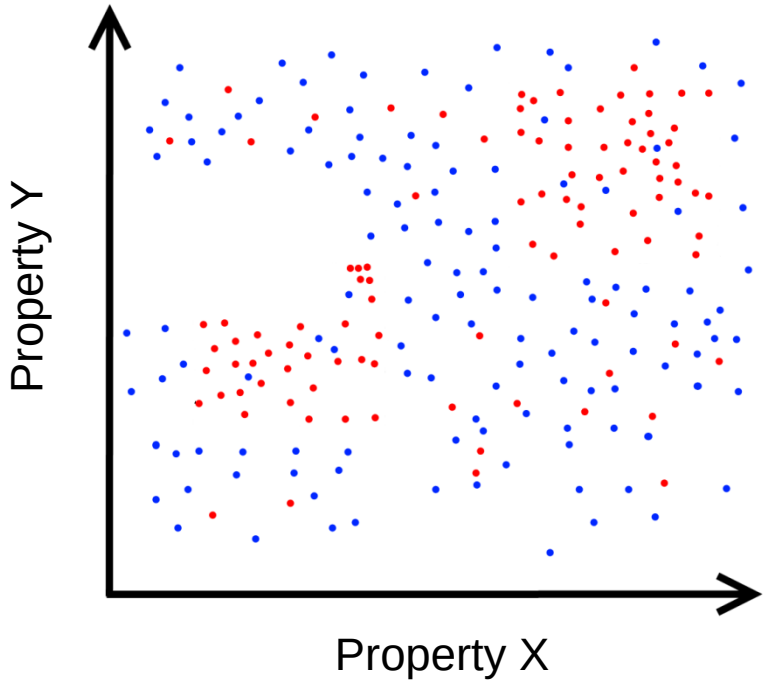
Rule Induction with PRIM

- This peeling-and-pasting process gives us a **peeling sequence** of boxes
- We select a single box from the peeling sequence based on its performance over the validation set
- Resulting box corresponds to a rule for selection of successful compounds



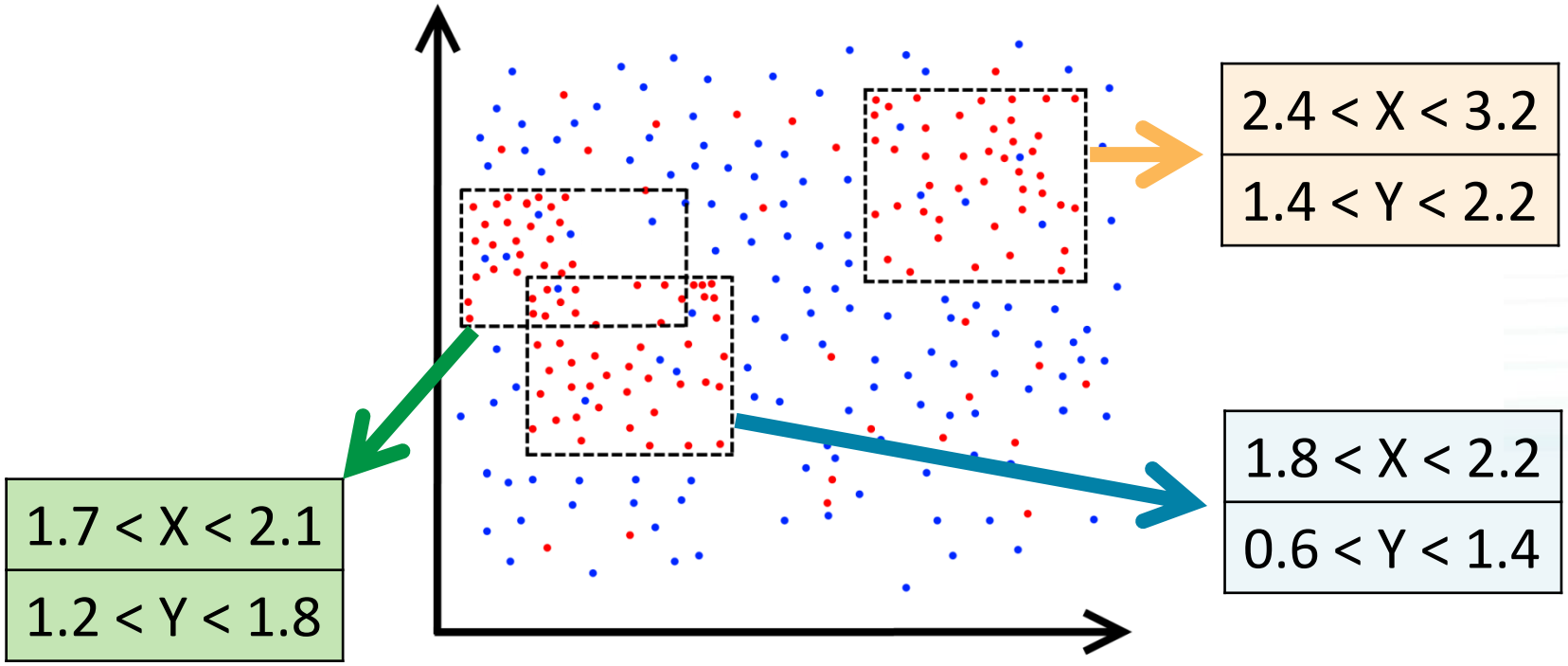
Rule Induction with PRIM

- After finding one box, we remove the box's compounds from the dataset and start over



Rule Induction with PRIM

- The result is a series of boxes, each corresponding to an individual rule



Measuring Rule Performance

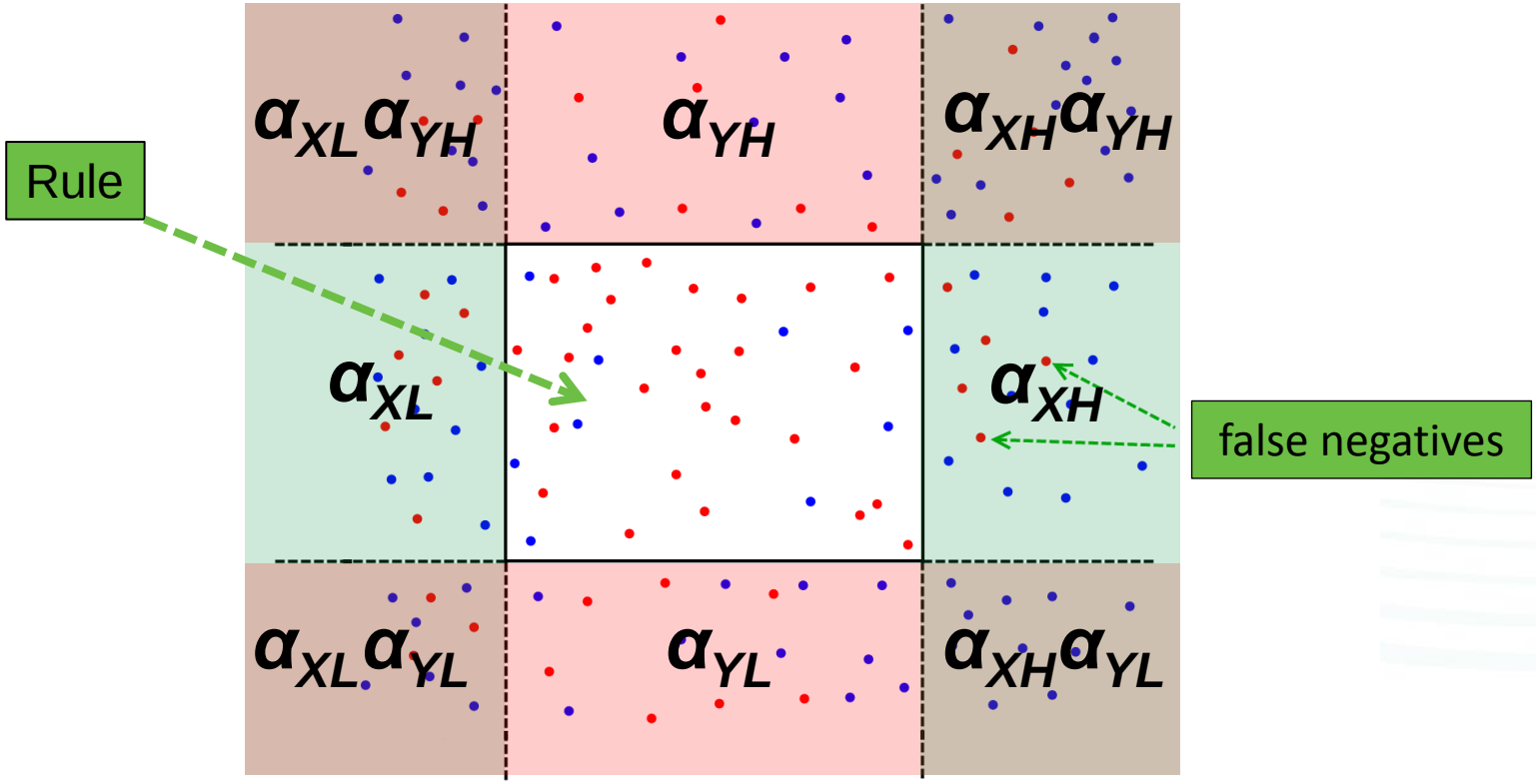
- Mean = Average objective value in box
  - Reported as % increase over objective value for full set
- Support = Proportion of data set ‘covered’ by box
  - Reported as % coverage
- Specificity vs. Sensitivity trade-off

Variable Importance

- PRIM does not tell us the relevance of each property criterion to a given rule's predictions
- Cannot answer questions like:
  - Should we trade off solubility for potency?
  - Would it be valuable to generate data for a particular property?

Variable Importance

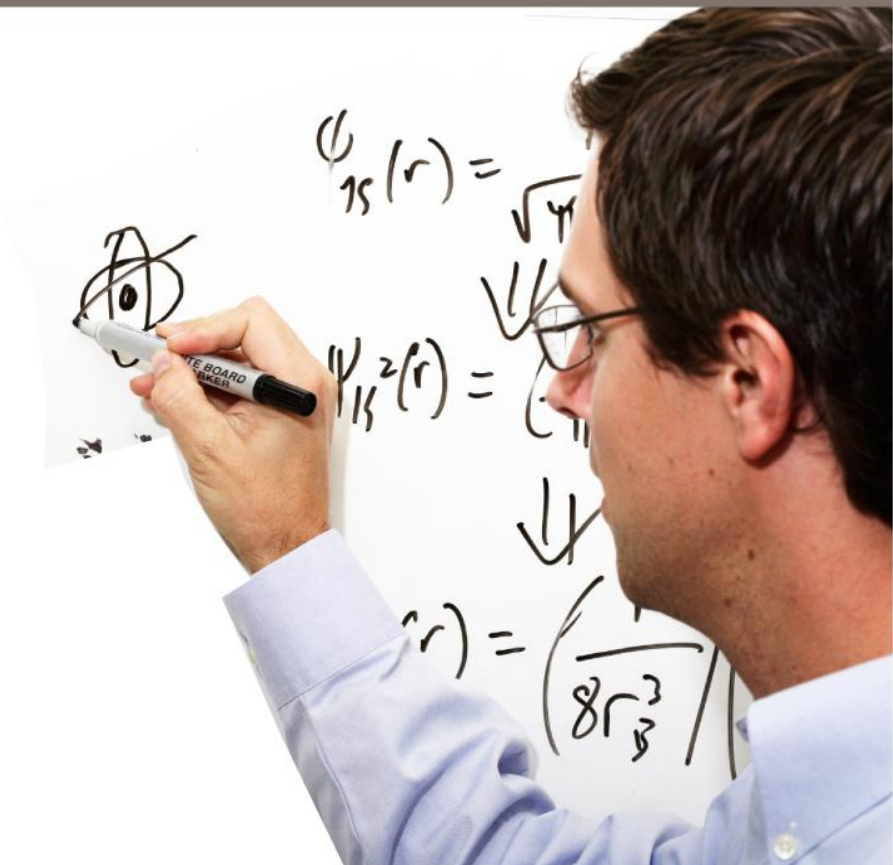
- α_i = false-negative rate of property criterion *i*



- Importance = $1 - \alpha_i$

Illustrative Results

optibrium

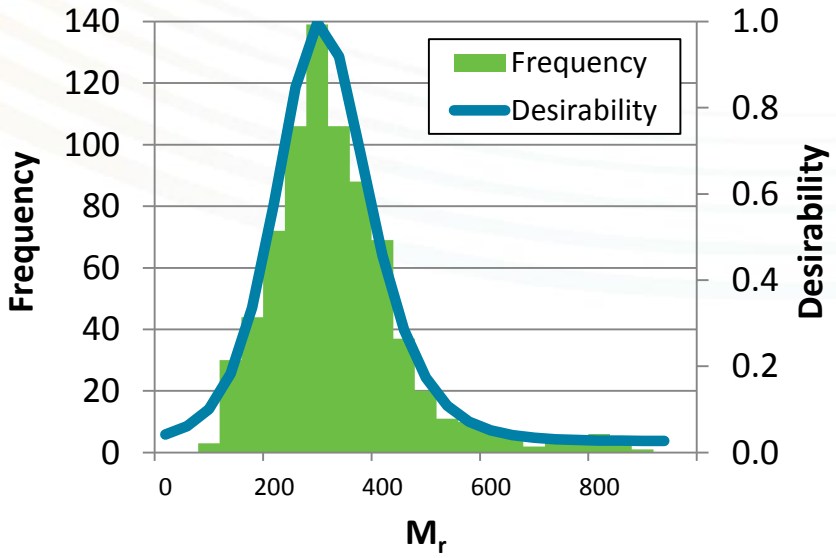


Example: Drug-Like Properties
QED*

- Quantitative Estimate of Drug-Likeness (QED)*
- Combine values for 8 properties

MW	logP	HBD	HBA
PSA	ROTB	AROM	ALERT

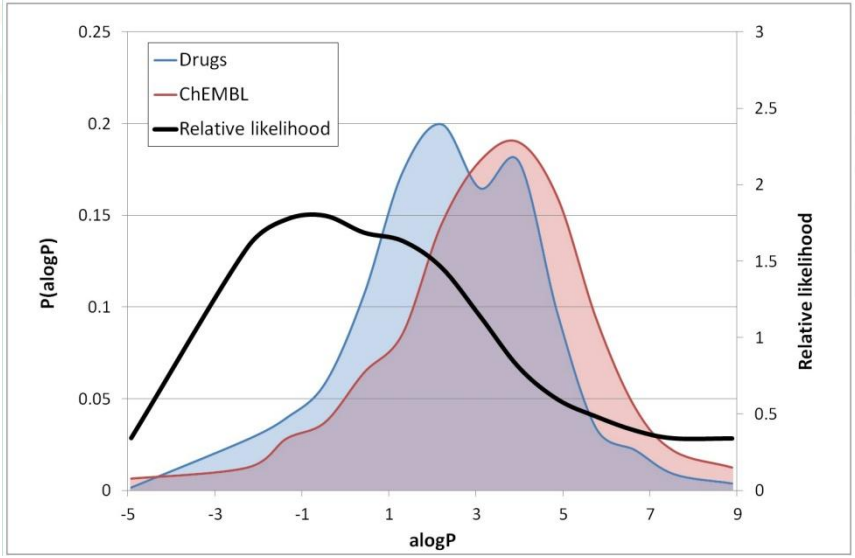
- For each individual property desirability function fitted to distribution for 771 **oral** drugs
- QED calculated as geometric mean of **individual** desirabilities



*Bickerton *et al.* Nature Chem. 4, pp. 90-98 (2012)

Example: Drug-Like Properties
RDL*

- Relative Drug Likelihood (RDL)*
- Compare characteristics of 771 **oral** drugs with 1000 randomly selected compounds from ChEMBL database
  - What property values increase likelihood of compound being an oral drug?
- Used same 8 properties as QED
- RDL calculated as geometric mean of **individual** likelihoods



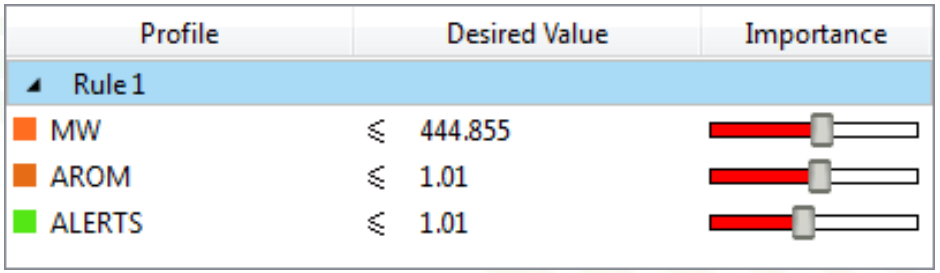
*Yusof and Segall, Drug Discov. Today DOI: 10.1016/j.drudis.2013.02.008

Example: Drug-Like Properties
Rule Induction

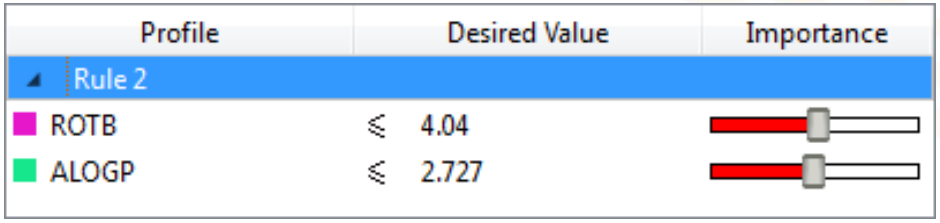
- Rule induction applied to data set of 771 **oral** drugs and 1000 randomly selected compounds from ChEMBL
  - Random split 70:30 training:validation sets
- Used same 8 properties as QED and RDL as inputs
- Minimum coverage values compared
  - 20%, 30%, 40%, 50%

Example: Drug-Like Properties
Rule Induction

- Minimum coverage 20% - 2 Rules



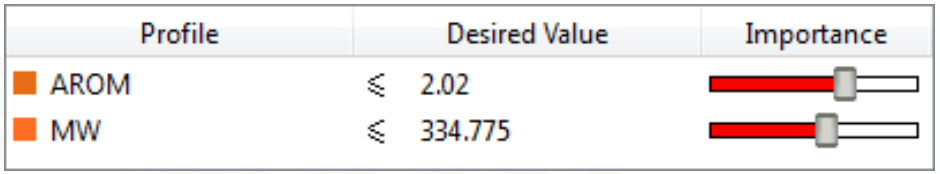
Set	Mean Improvement (%)	Coverage (%)
Train	60	22
Val	57	24



Set	Mean Improvement (%)	Coverage (%)
Train	51	23
Val	46	23

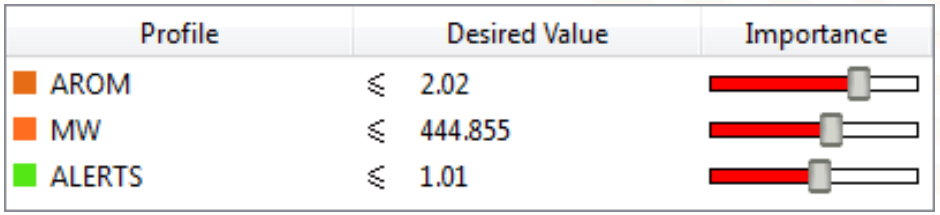
Example: Drug-Like Properties
Rule Induction

- Minimum coverage 30%



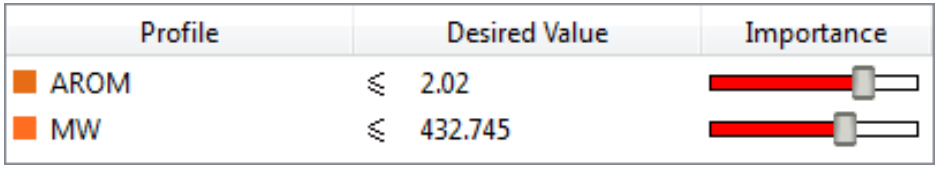
Set	Mean Improvement (%)	Coverage (%)
Train	51	36
Val	45	37

- Minimum coverage 40%



Set	Mean Improvement (%)	Coverage (%)
Train	44	44
Val	42	46

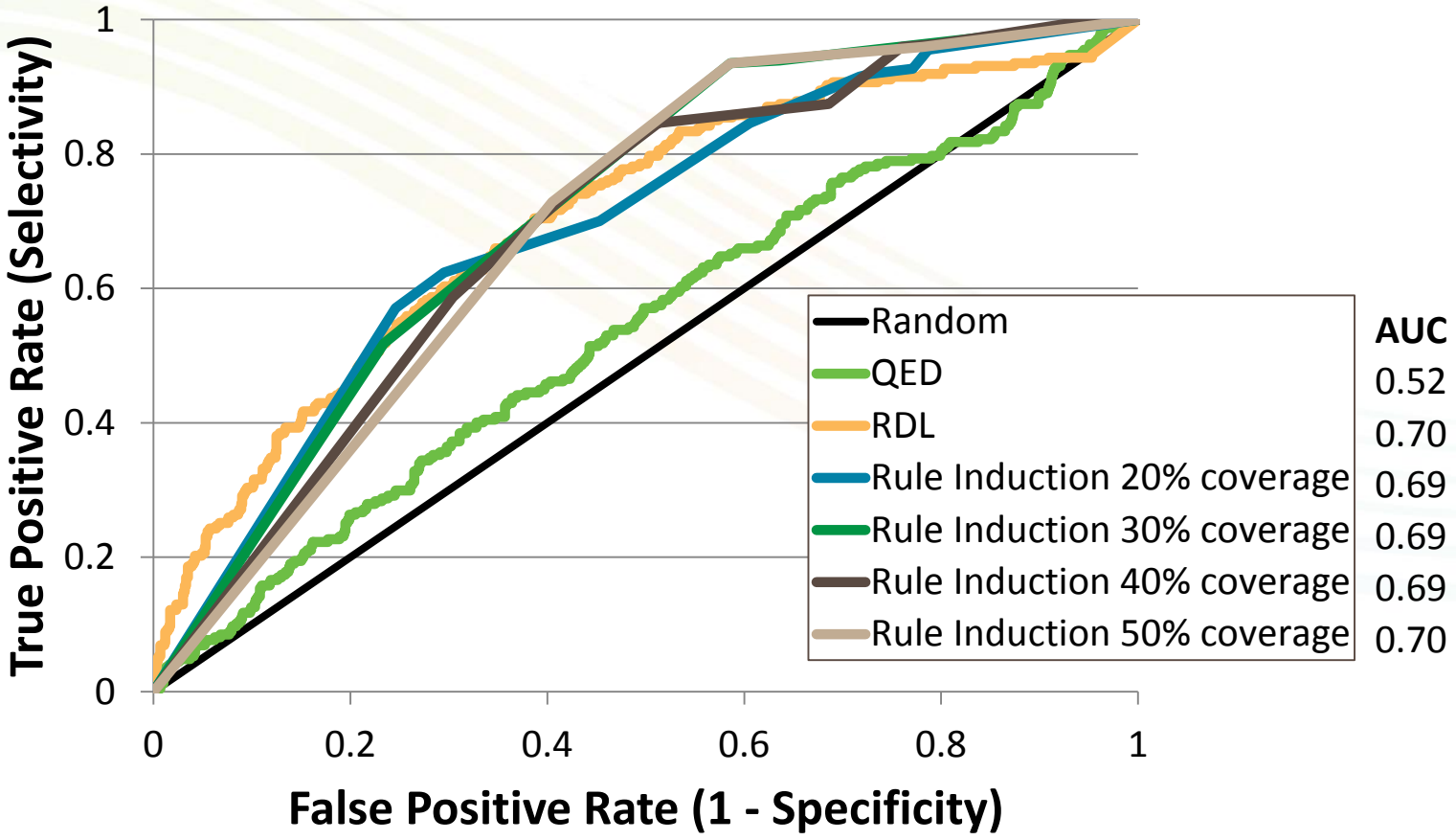
- Minimum coverage 50%



Set	Mean Improvement (%)	Coverage (%)
Train	35	57
Val	35	58

Example: Drug-Like Properties
Results

- Applied to **independent** test set of 247 oral drugs and 1000 compounds randomly selected from ChEMBL



Example: Oral CNS
Rule Induction

- Data set of 1191 drugs labelled as orally administered and CNS active or not
  - By approved route of administration and therapeutic indication (noisy)
- Divided into training (667), validation (286) and test (238) sets
- Calculated ADME properties from StarDrop™ used as input:

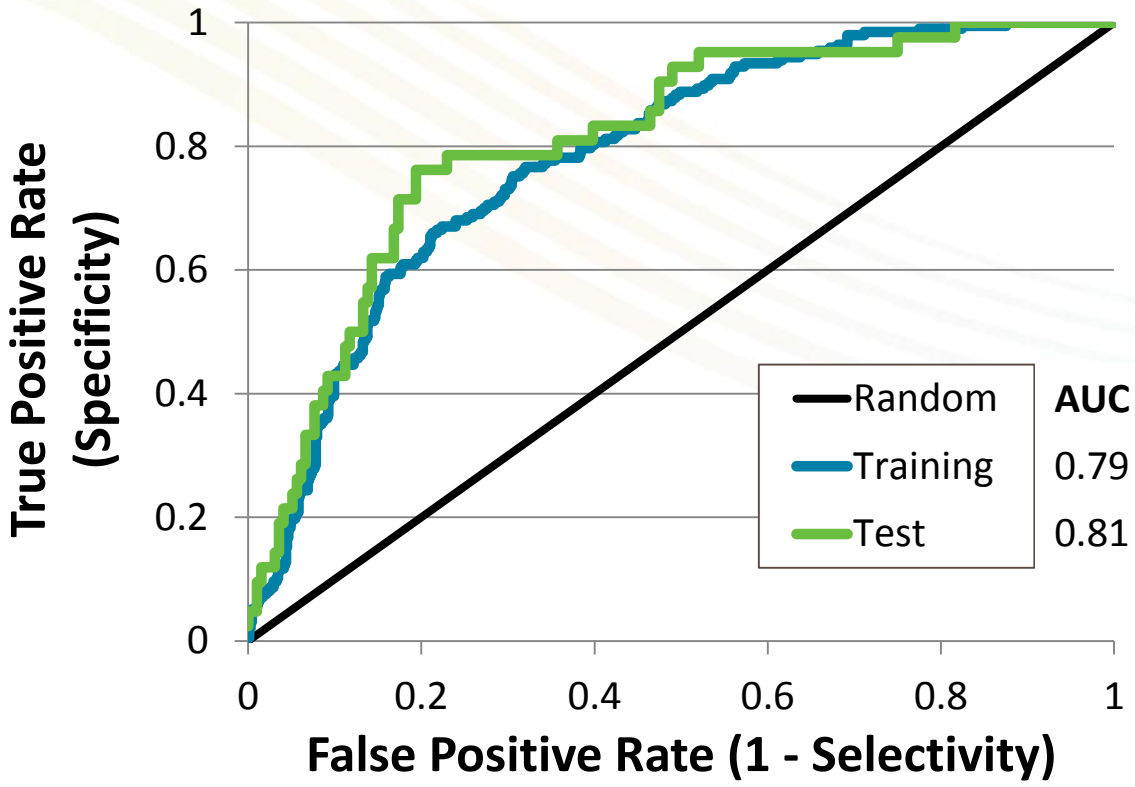
logP	Solubility (logS)	Human Intestinal Absorption category (HIA)
Blood-brain barrier penetration (BBB log)	Plasma protein binding category	P-gp substrate category

- Minimum coverage 20%

Example: Oral CNS
Results

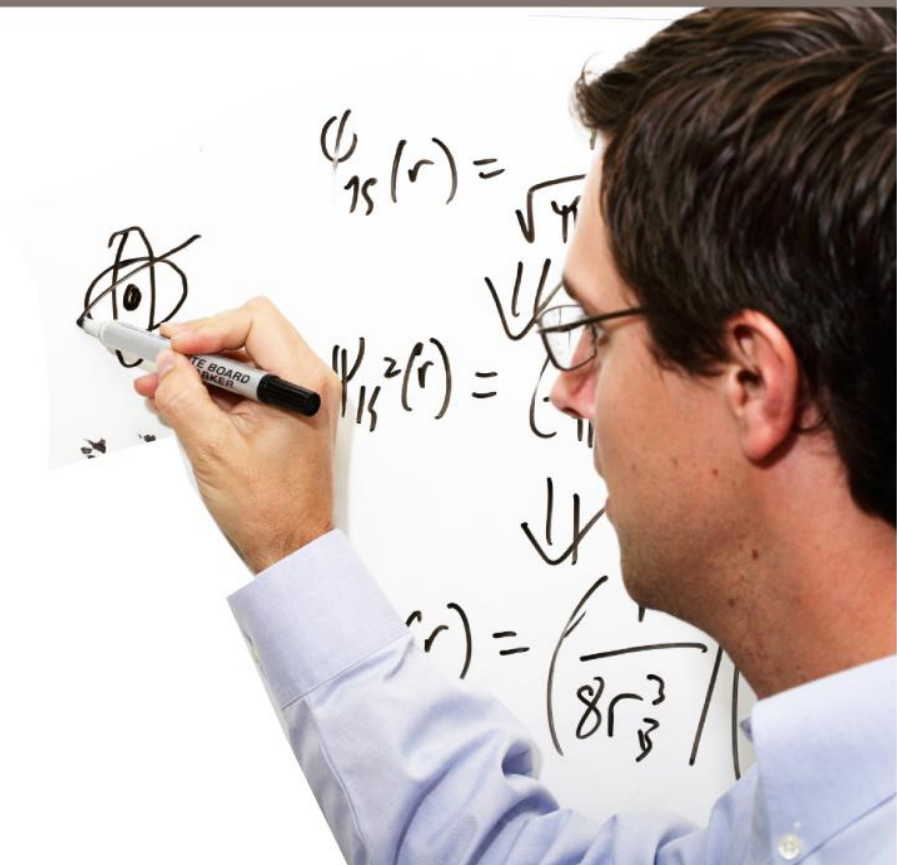
Profile	Desired Value	Importance
HIA category	+	
logS	> 1.00601	
BBB log([brain]:[blood])	> -0.424961	
logP	> 1.12432	
P-gp category	no	

Set	Mean Improvement (%)	Coverage (%)
Train	132	23
Val	123	20

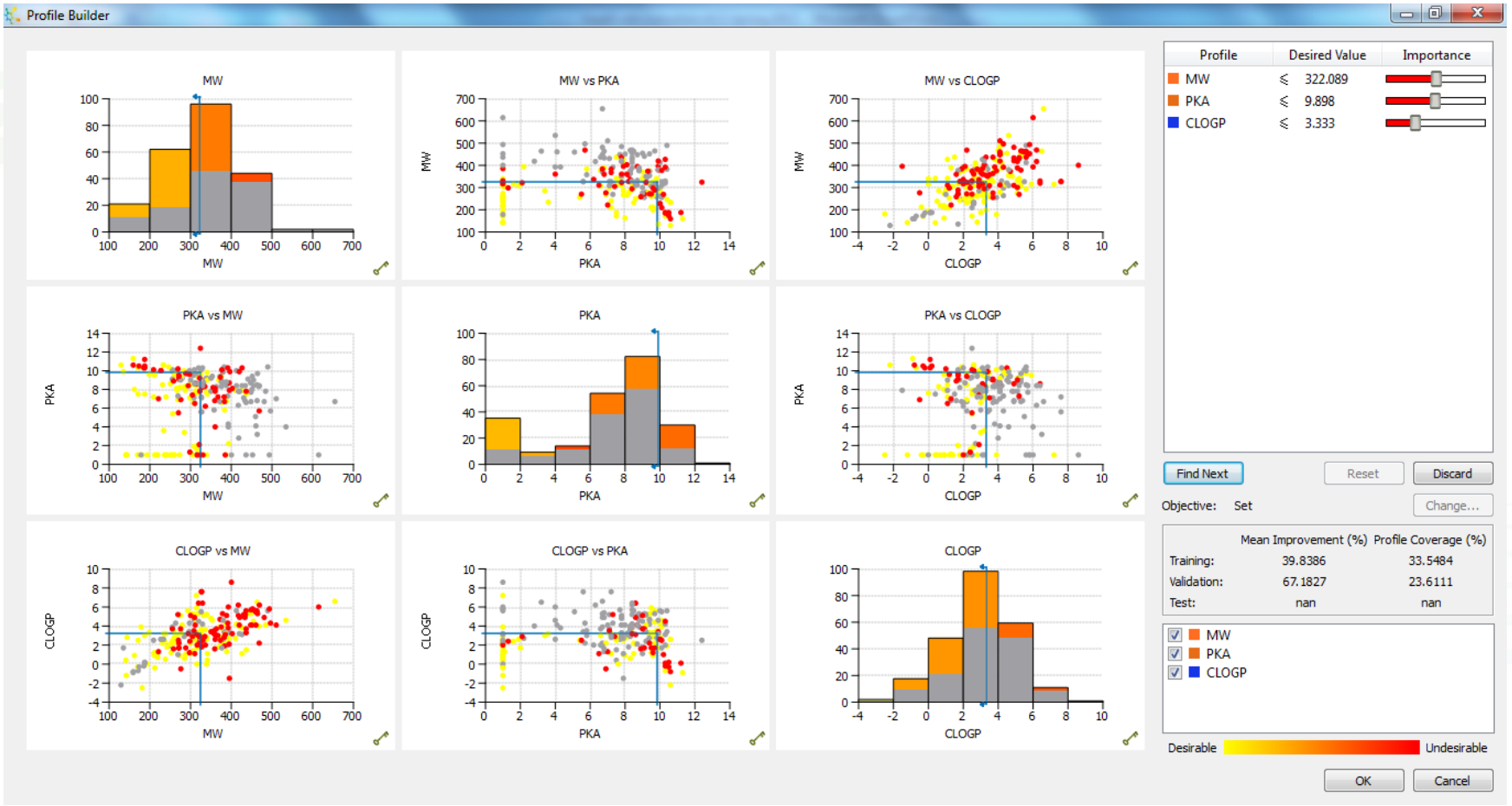


User Interaction

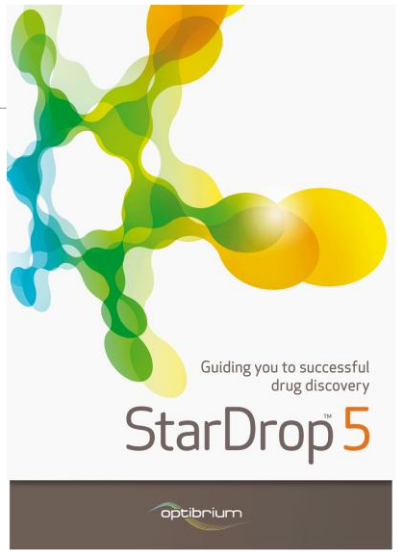
optibrium



Interactively Explore Profile Building



Conclusion



- MPO is a powerful approach to select and design compounds with a high chance of success
- Rule Induction helps to guide the development of scoring profiles to select compounds for a drug discovery objective
  - Apply to any objective
  - Use experimental or calculated data
  - Not black box – synergy between computer and expert
- Identify most important data to guide selection of successful compounds
  - Optimise screening strategy and prioritise experimental resources
- For more information:
  - matt.segall@optibrium.com
  - www.optibrium.com

Acknowledgements

- Tatsu Hashimoto – MIT
- Optibrium team, including:
  - Iskander Yusof
  - Ed Champness
  - Chris Leeding
  - James Chisholm
  - Hector Garcia Martinez