



Automated QSAR Modelling

David E Leahy
Newcastle University, UK

&

Damjan Krstajic
Research Centre for Cheminformatics, Serbia

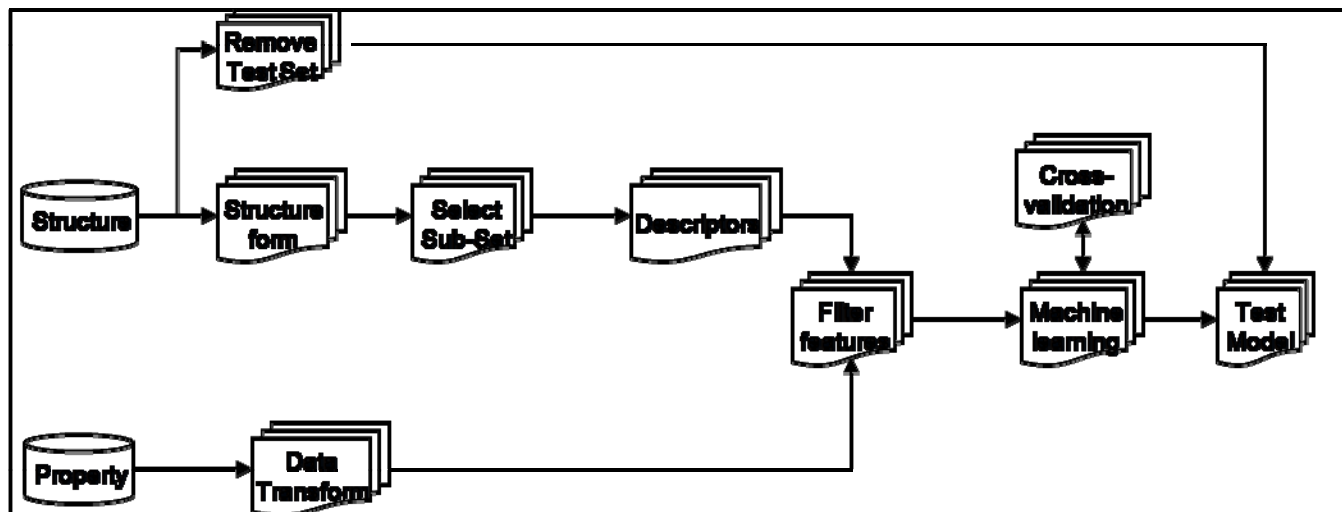
Discovery Bus

- Automate QSAR modelling by Experts
 - Including experience and learning
- Agnostic
 - Let all methods compete
- Auto-Reactive
 - New data, methods, strategies

“The Discovery Bus is not a tool for users. It is a system for deriving QSAR models independent of any user”

www.discoverybus.com

Discovery Bus QSAR

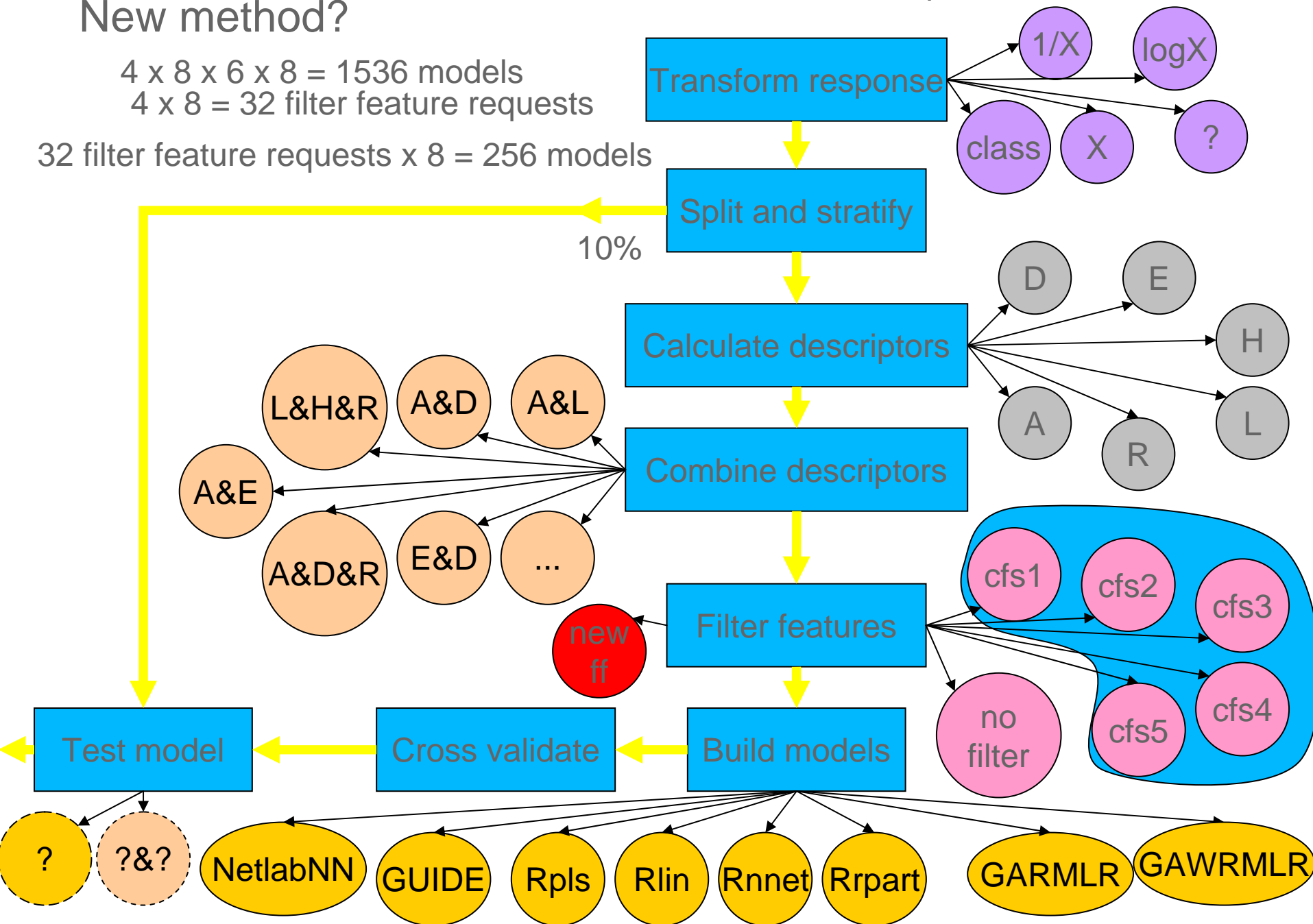


Chemical structure & response data

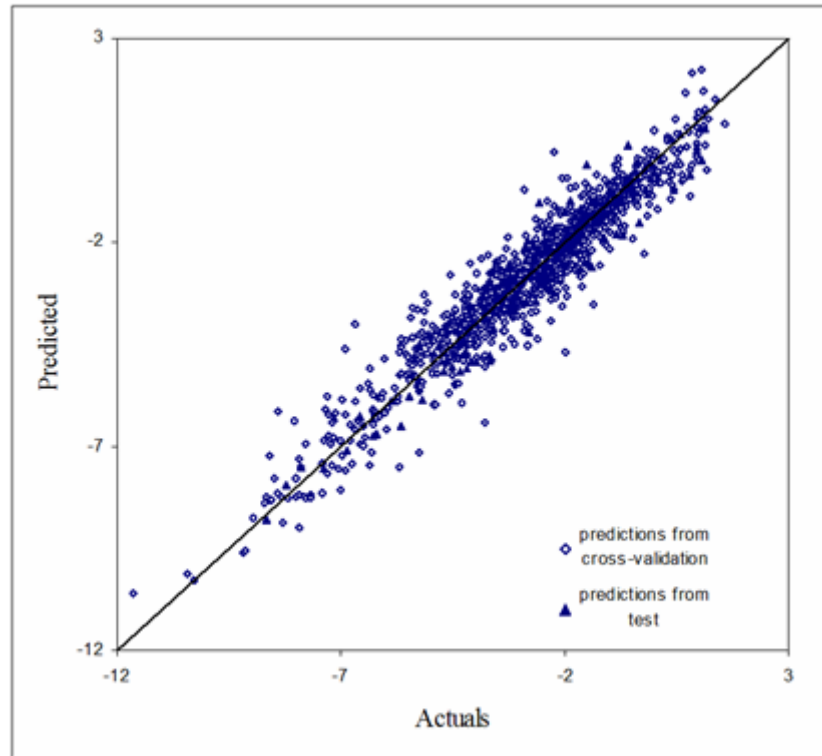
New method?

$4 \times 8 \times 6 \times 8 = 1536$ models
 $4 \times 8 = 32$ filter feature requests

$32 \text{ filter feature requests} \times 8 = 256$ models



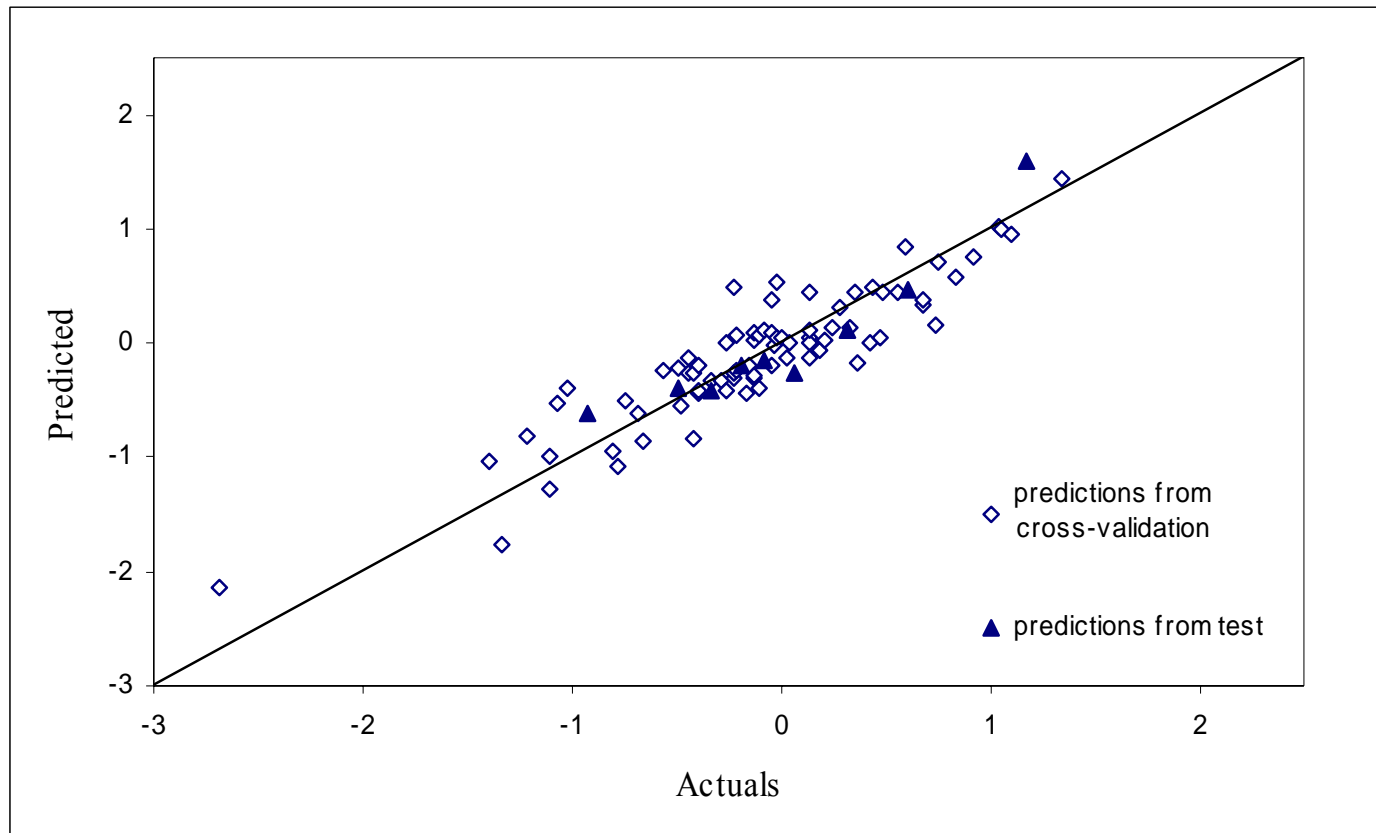
Solubility



Solubility Results

Learner	Filter	Reduction		Types	Linear Fit	Training (1167)		Test (130)	
		Filter	Learner			Rel.MSE	r ²	Rel.MSE	r ²
GUIDE	H	1990 → 558	→ 54	R,D	1.46	0.11	0.89	0.12	0.89
	H	170 → 26	→ 14	A,E,H,D	0.13	0.11	0.89	0.13	0.88
	H	80 → 16	→ 12	A,H,D	0.14	0.11	0.88	0.12	0.87
	C	250 → 2	→ 2	A,R	0.18	0.13	0.87	0.16	0.84
	C	8 → 2	→ 2	A,L	0.16	0.13	0.87	0.16	0.86
GA1	H	80 → 16	→ 16	A,H,D	0.14	0.14	0.86	0.18	0.83
	C	8 → 2	→ 2	A,L	0.16	0.17	0.84	0.17	0.83
NN1	H	250 → 54	→ 54	A,R	0.12	0.09	0.91	0.08	0.92
	H	80 → 16	→ 16	A,H,D	0.14	0.10	0.90	0.12	0.88
	H	326 → 46	→ 46	H,R,D	0.18	0.10	0.90	0.12	0.89

HSA Binding



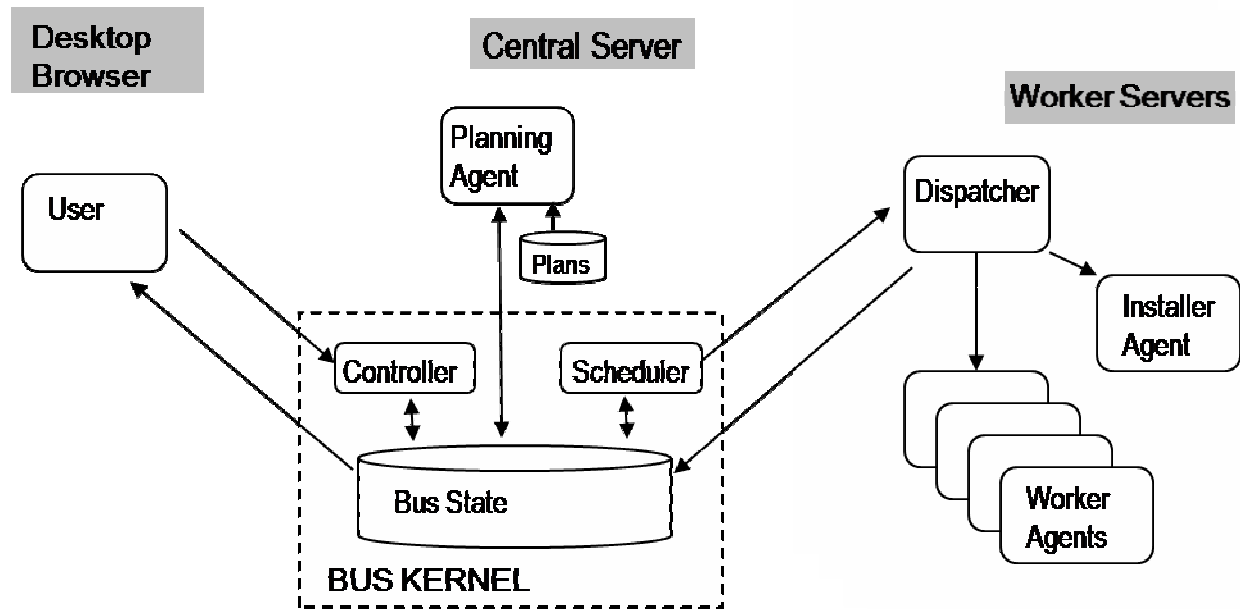
HSA Binding

Learner	Filter	Reduction		Types	Linear Fit	Training (82)		Test (9)	
		Filter	Learner			Rel.MS E	r ²	Rel.MS E	r ²
Guide	Hh2	332 → 39	→ 8	A,E,R	0.92	0.40	0.62	0.25	0.81
	H	250 → 59	→ 12	A,R	1.62	0.47	0.56	0.30	0.76
	Hh4	382 → 20	→ 1	A	0.25	0.50	0.50	0.57	0.49
GA1	Hh2	1998 → 39	→ 26	A,R,D	0.42	0.23	0.77	0.20	0.85
	Hh4	344 → 20	→ 19	H,R,D	0.42	0.26	0.74	0.28	0.78
	Hh10	302 → 9	→ 9	H,R	0.27	0.27	0.73	0.40	0.64
NN1	H	8 → 5	→ 5	A,L	0.37	0.17	0.83	0.15	0.87
	Hh10	346 → 8	→ 8	A,R,D	0.30	0.30	0.70	0.16	0.84
	H	302 → 19	→ 19	H,R	0.27	0.32	0.70	0.39	0.71

P-Glycoprotein

Technique	% Correctly Classified	% Correctly Classified
	Training Set	Test Set
Neural Net Classifier	95.6	69.7
R Part	90.4	81.0

Discovery Bus Architecture



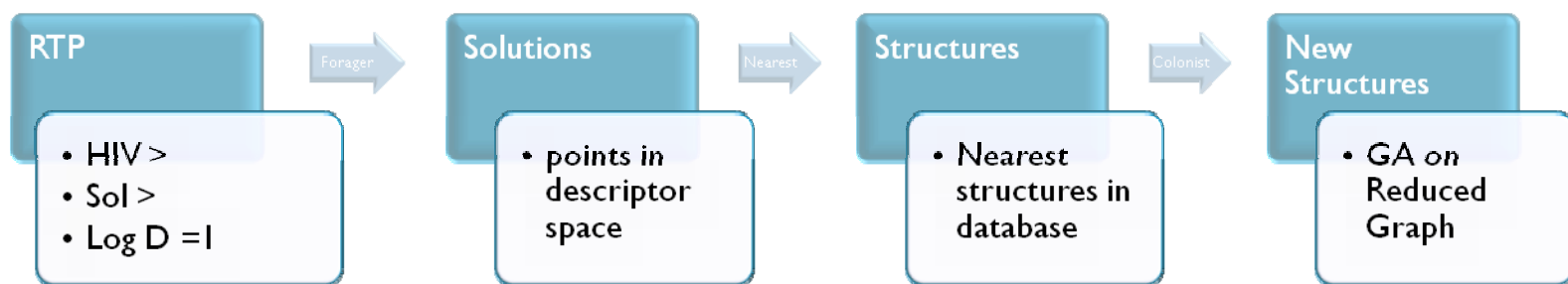
Summary

- The Discovery Bus
 - Automatically tries all possible models
 - Each model according to best practice
 - Automatically applied to new data and methods
 - Easily extended
 - $(N \times M \times P \times Q)$ *versus* $(N + M + P + Q)$
 - Any technology (C++, Java, Perl, Python, ...)
- Efficient
 - Discovery Bus models as good as those by experts
 - 20 Server Engine = 300 QSAR scientists

Current & Future Work in QSAR

- “Load up the Bus”
 - More Data
 - Descriptors
 - Learning methods
- Workflow Extensions
 - Structure normalisations
 - Chemotype identification & Fragment methods
 - Model Selection
- Meta-QSAR
 - Performance comparisons
 - Model analysis
- Combinatorial Problem
 - Tree pruning
 - Reinforcement

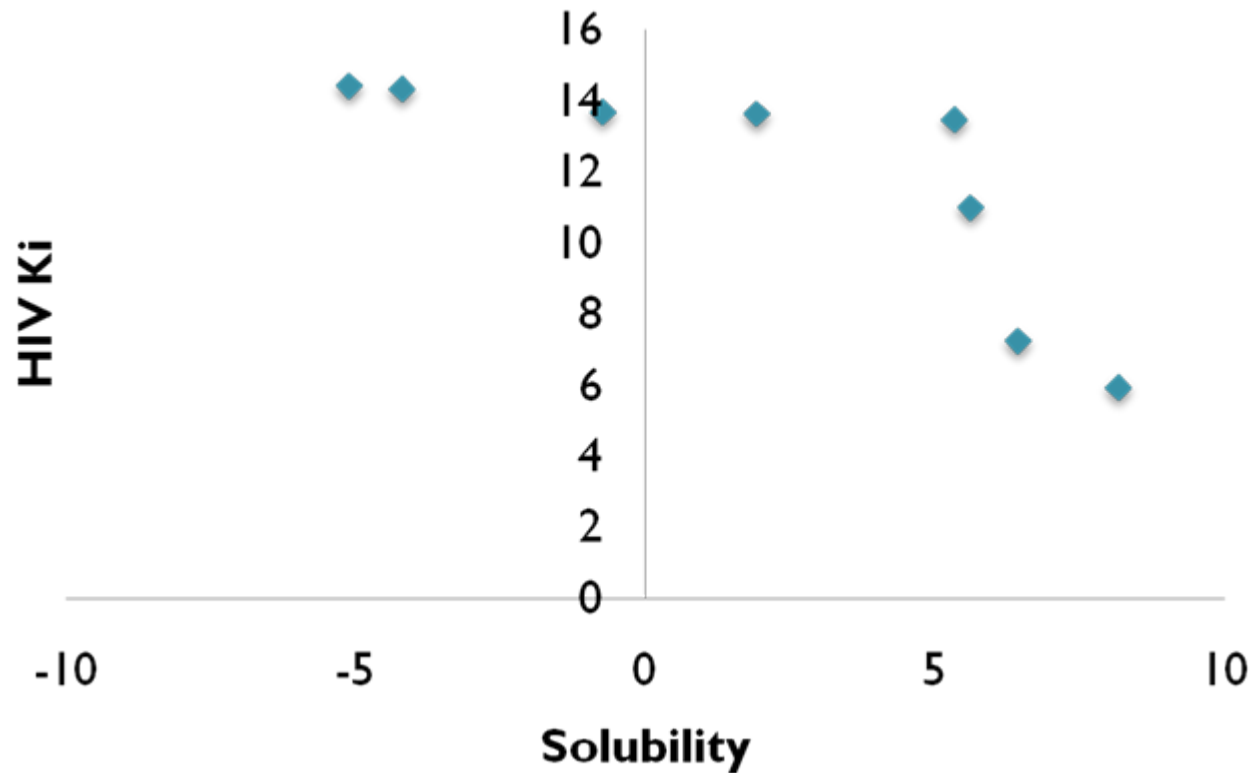
Reverse QSAR Engineering



Forager: A PSO for Reverse QSAR

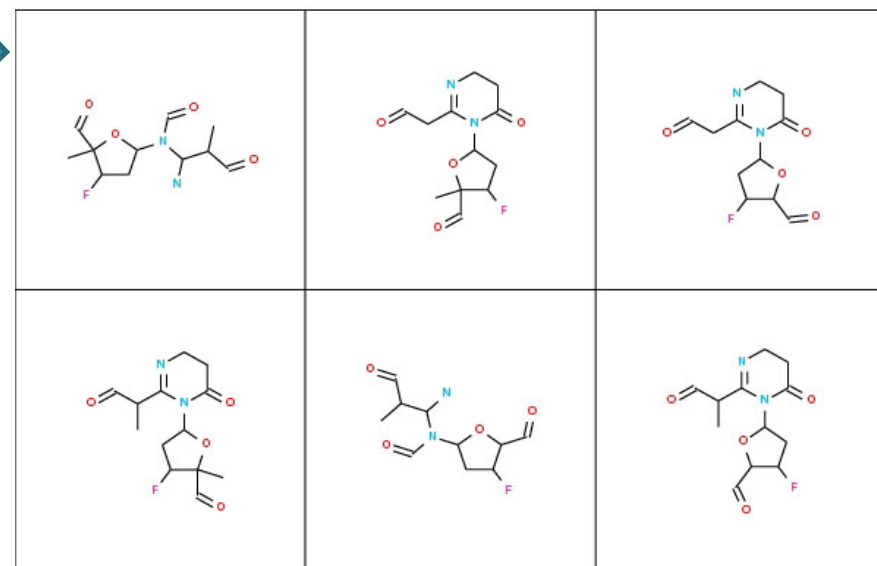
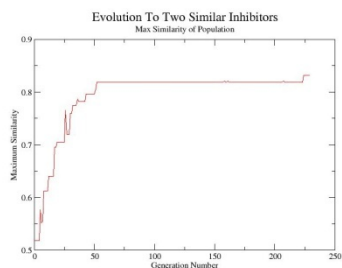
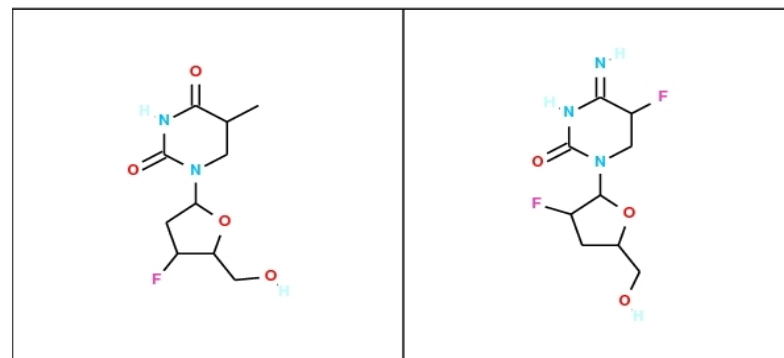
- Particle Swarm Optimisation of multiple properties
 - Swarm moves in union descriptor Space
 - Swarm use QSAR models as optimisation criteria
 - Memory of best position
 - Memory of nearest neighbours best positions
- Herding
 - Separation (avoid crowding)
 - Alignment (steer towards common direction)
 - Cohesion (steer towards mean position)
- Varied Speed
 - Speeds up if no solutions
 - Slows down when solutions found

Forager Optimisation



Thanks to Tudor Oprea for a copy of Wombat

Colonist



Acknowledgements

RCC
Belgrade

- Damjan Krstajic

Newcastle
University

- Vladimir Sykora (Colonist)
- Robert Leahy (Forager)

Cyprotex PLC

- John Cartmell, Jim Bowen, Steve Enoch