StarDrop Solubility Models

Olga Obrezanova, PhD

Transnational Computational Chemistry Meeting

1-2 April 2009

BioFocus DPI
A Galápagos Company

StarDrop solubility models

Current - version 4.2 and previous

- Intrinsic aqueous solubility logS (S in µM) (solubility of neutral form)
- Apparent solubility at pH 7.4 logS@7.4 (S in µM)
  - If logS@7.4 is called for a neutral compound logS value is given

Future - version 4.2 beta

- New model for aqueous intrinsic solubility logS

How the new logS model was built

- Automatic model generation algorithm was created and implemented in Auto-Modeler
- To test the algorithm compare 'automatic' models versus 'manual' ones
  - Considered examples of blood-brain barrier penetration and aqueous solubility
  - Automatic solubility model turned out to be better -> new logS model
  - Published in J. Comp. Aided Mol. Design, Feb. 2008

Talk Outline

- Automatic Model Generation process (Auto-Modeler)
  - Stages of the process
  - Gaussian Processes modelling techniques
- 'Manual' model versus 'automatic'
  - Old 'manual' aqueous solubility model
  - New 'automatic' aqueous solubility model
- Comparative evaluation of solubility models
- Solubility at pH 7.4 model (time permitting)

Automatic Model Generation Process

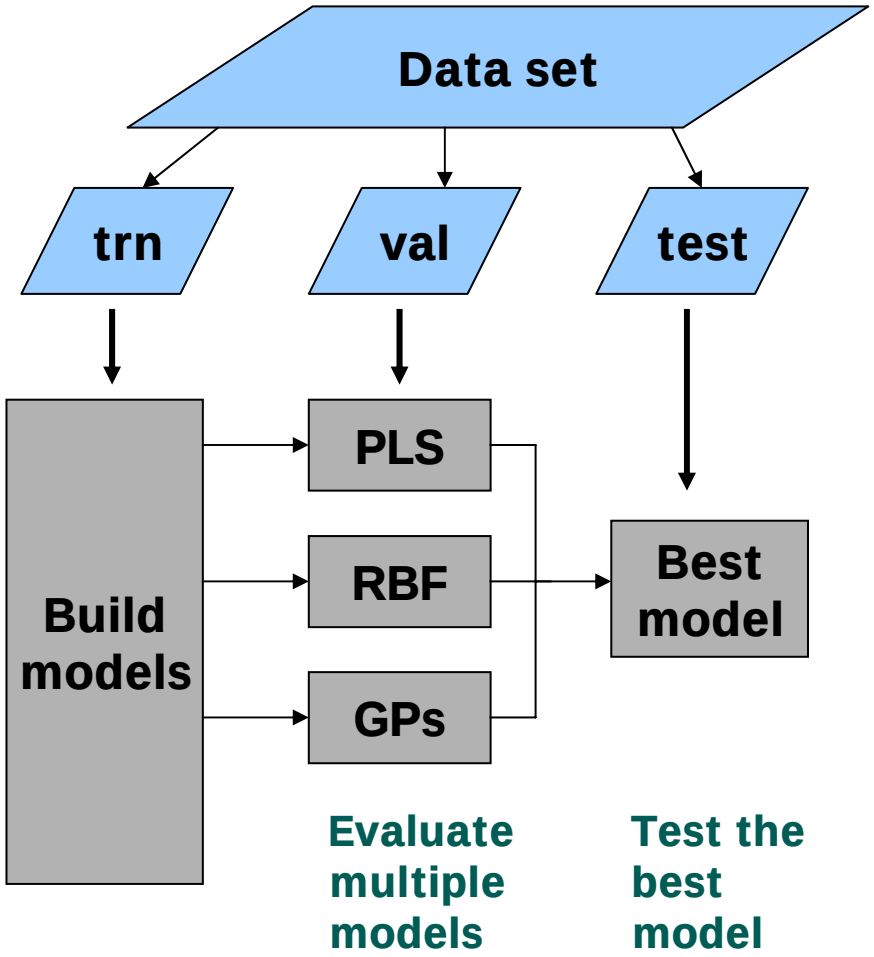
Auto-Modeler

BioFocus DPI
A Galápagos Company

Automatic model generation

- The rapid design-test-redesign cycle of modern drug discovery demands fast model building
- Automatic modelling processes allow
  - exploring large numbers of modelling approaches efficiently
  - making QSAR model building accessible to non-experts
- **Auto-Modeler** is an automatic model generation algorithm implemented in the **StarDrop**. Works at two levels
  - Non-experts, minimal input from the user
  - Expert user can influence each stage of the process

Auto-Modeler

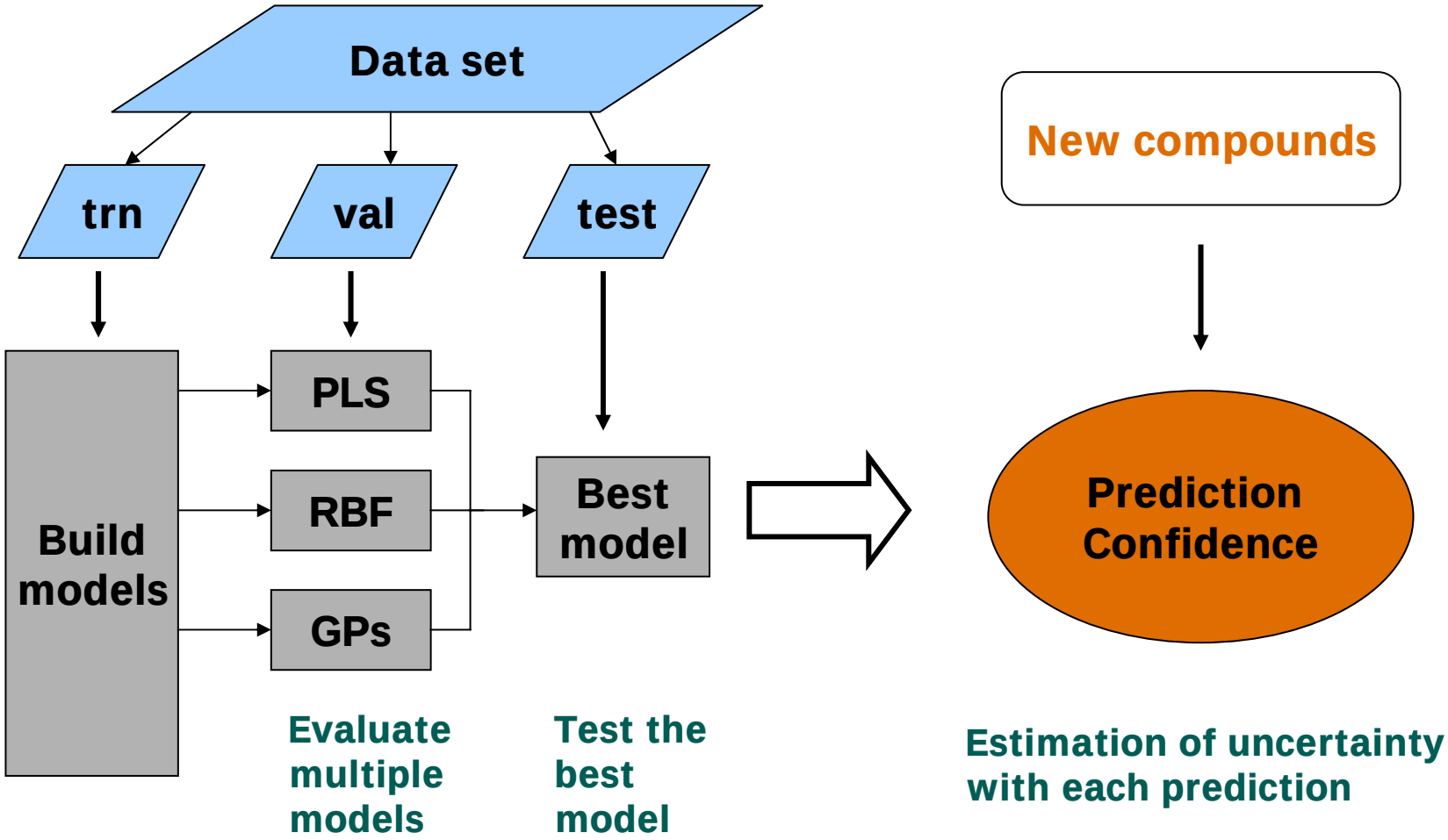


- Splitting data into training, validation and test sets (by cluster analysis)
- Descriptor calculation and filtering (2D SMARTS, logP, TPSA, MW, charge etc.)
- Modelling techniques (PLS, Radial Basis Functions with genetic algorithm, Gaussian Processes, Decision Trees)
- Selection of the best model by performance on the validation set
- Test set is an **independent** set

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Auto-Modeler



StarDrop™
Balanced compounds from better decisions

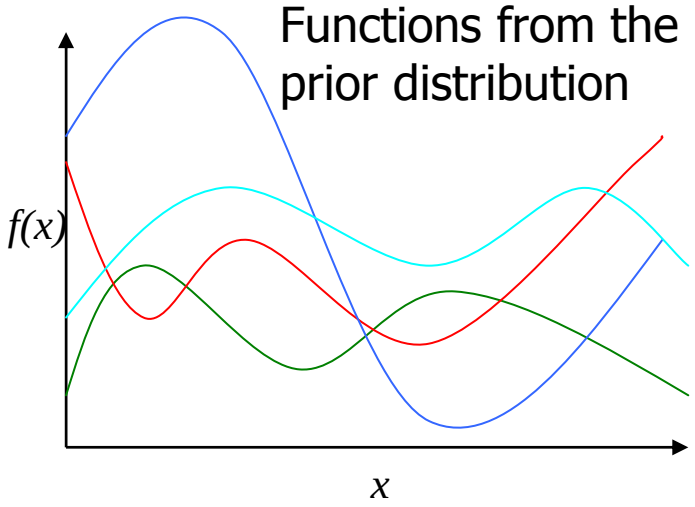
BioFocus DPI
A Galápagos Company

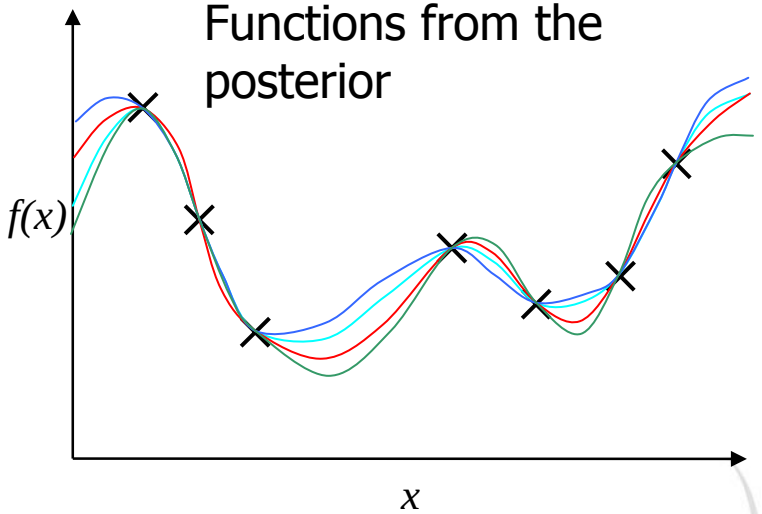
Modelling techniques: Gaussian Processes

- A machine learning method based on Bayesian approach
- Advantages:
  - Does not require a priori determination of model parameters
  - Nonlinear relationship modelling
  - Built-in tool to prevent overtraining - no need for cross-validation
  - Inherent ability to select important descriptors
  - Provides uncertainty estimate for each prediction
- Sufficiently robust to enable automatic model generation

Modelling techniques: Gaussian Processes

- Define **prior distribution** over functions (controlled by hyperparameters, covariance function – ARD function)
- **Posterior distribution**: retain functions which fit experimental data
- **Prediction** is the mean of posterior distribution.
- Standard deviation of the distribution provides estimate of the **uncertainty in prediction**





StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

‘Automatic’ model versus ‘manual’

Experiment

BioFocus DPI
A Galápagos Company

‘Automatic’ model versus ‘manual’

- Data set – experimental values for aqueous solubility
- ‘Manually’ built model – old logS built by a computational chemist (Joelle Gola) in 2003, used since in StarDrop
  - Different modelling techniques, subsets of descriptors, set splitting etc. were investigated
  - Variety of tools – variety of data formats
- Automatic model – apply Auto-Modeler to whole data set
- Compare ‘automatic’ and ‘manual’ models by testing on external data (subset of Huuskonen aqueous solubility set)

Old Aqueous Solubility Model
'Manual' model

- Data set of 3313 compounds
  - Intrinsic aqueous solubility (logS, S in μM), measured within 20-30°C
  - PHYSPROP database (Syracuse Research Corporation, SRC)
- Random set split
  - (80% in Trn, 20% in Test=663 comp)
- 108 descriptors (SMARTS based and MW, TPSA etc)
  - Initial set of 157 was reduced by filtering on low variance, correlation
- Final model – Radial Basis Functions (RBF) technique
- On test set R^2=0.82, RMSE=0.79 log units

New ‘automatic’ model

- Auto-Modeler was applied to all data set of 3313 compounds
- Set split by cluster analysis at Tanimoto=0.7 (15% - Val, 15% - Test)
- Best model - Gaussian Processes with 2D search

manual

Test set	
R^2	0.82
RMSE	0.79

automatic

Val+Test set	
R^2 val	0.84
R^2 test	0.85
RMSE	0.69

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

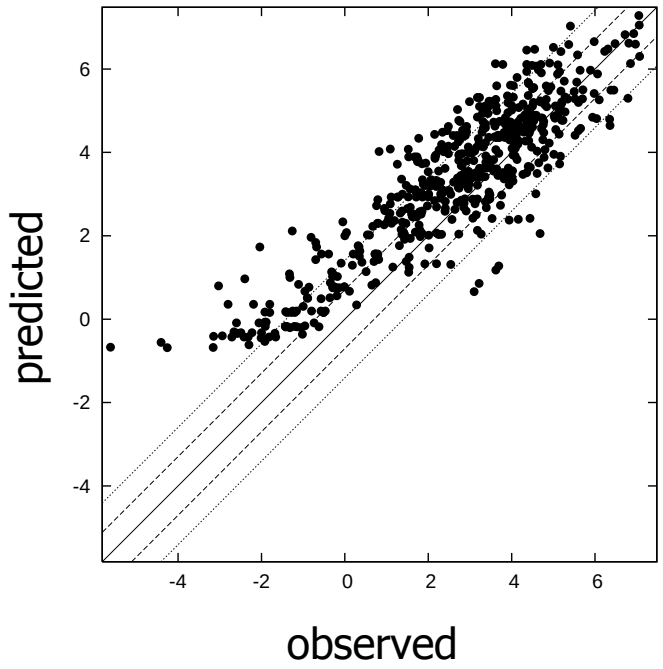
Performance on external test set

- External test data – 564 compounds from 'Huuskonen' set – not used in original modelling set
- Pure water solubility (or intrinsic?), in total 1297 compounds
- Huuskonen J., J. Chem. Inf. Comput. Sci., 2002, 42

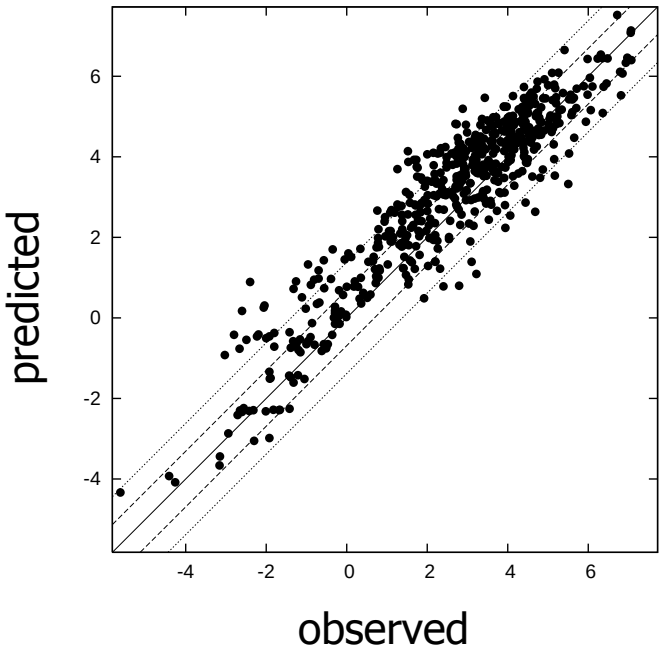
Model	Desc	% pred within ±0.7 log unit	% pred within ±1.4 log unit	R^2	r^2_{corr}	RMSE
manual	108	39.9	70.9	0.68	0.80	1.28
automatic	166	54.1	85.9	0.82	0.86	0.96

Performance on 'Huuskonen' test set

manual



automatic

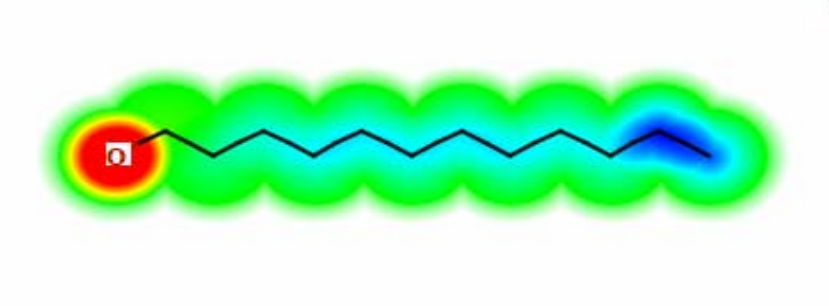


StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Glowing molecule visualization

- Makes a link between predicted property and compound's structure
- Interpret SAR and guide redesign of compounds to overcome liabilities



Obs logS = 1.33

Pred logS = 2.09

Obs logS = 3.04

Pred logS = 2.33

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Comparative evaluation of solubility models

Galapagos study 2008

Dearden study 2006

Solubility Challenge 2008

BioFocus DPI
A Galápagos Company

Galapagos Study

by Pieter Stouten and David Sys

- Solubility models from Pipeline Pilot/Cerius, Pipeline Pilot/Tetko, ACDlabs (4 models), Q-pharm, StarDrop
- Evaluation data sets
  - Training set of old (manual) logS model – 2650 compounds
  - Test set of old logS model – 663 compounds
  - Subset of Huuskonen set – 564 compounds
- Possible performance bias
  - Tetko model is built on Huuskonen set (biased on all 3 sets)
  - StarDrop model will be biased on the first set (on the second set as well)
  - Cerius used compounds from PHYSPROP database (biased on all sets?)

Galapagos Study: Results

Pearson's correlation coefficient (r)

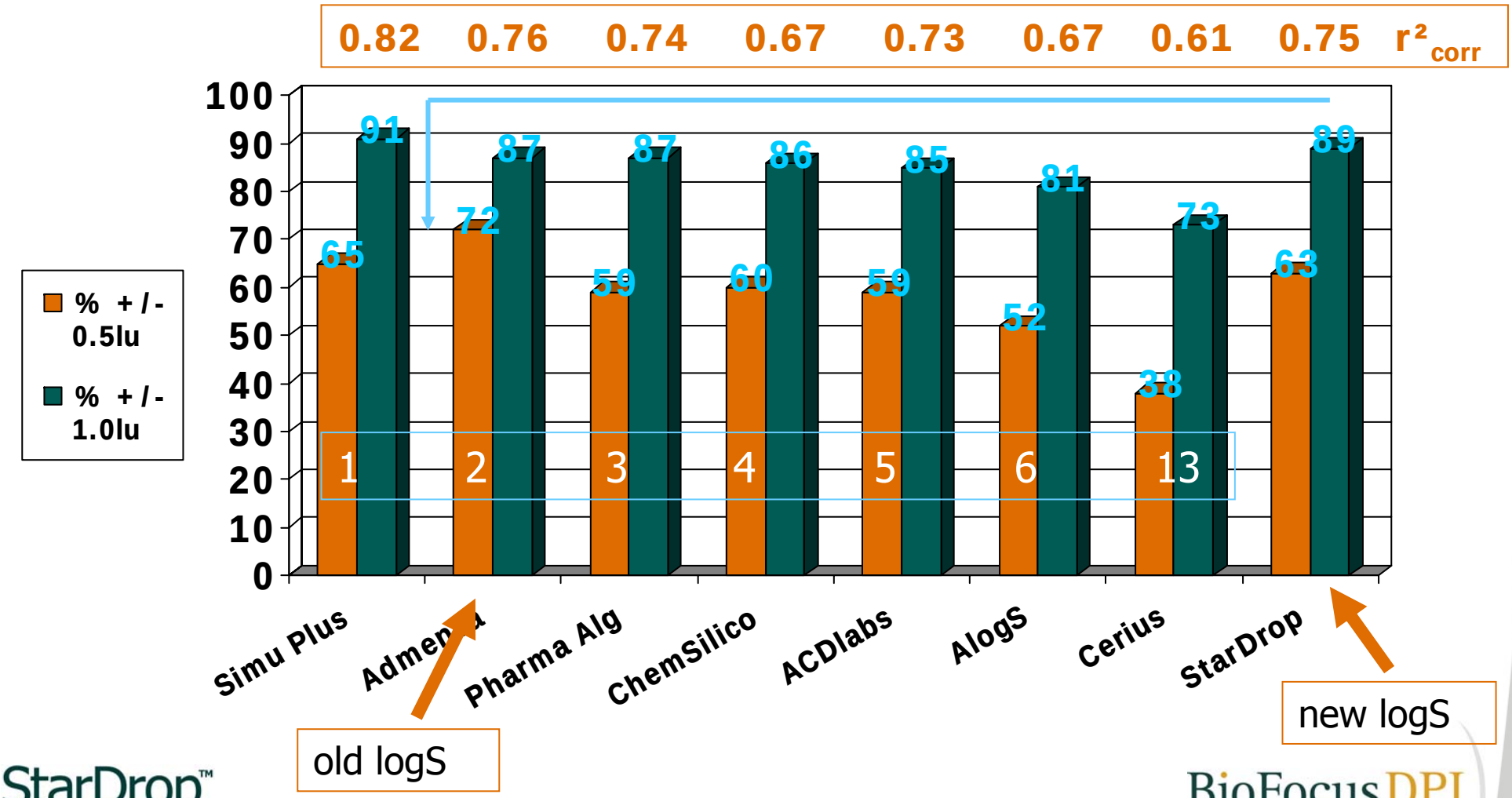
	ACDlabs				StarDrop	
	intrinsic	pure water	Tetko	Cerius²	logS old manual	logS new auto
Trn set 2650 cpds	0.91	0.91	0.80	0.81	0.99*	0.93
Test set 663 cpds	0.91	0.92	0.80	0.84	0.92	0.94
Huuskonen 564 cpds	n/a	n/a	0.93	0.92	0.89	0.93

* RBF model, complete fit on training set

John C. Dearden study

- "In silico prediction of aqueous solubility"
  - Expert Opinion Drug Discovery (2006), 1: 31-52
- Comparison of software for aqueous solubility prediction
  - Tested 17 software programs
- Test set - 122 drugs, with experimental pure water solubility
  - 58 /14 cpds from this set are present in the training/test set of StarDrop models
  - Some experimental values in pairs of duplicates are very different, squared correlation coefficient $r^2 = 0.88$

John C. Dearden study: Results



Solubility Challenge

- Organized by University of Cambridge in summer 2008
- Competition
  - Trn set - 105 compounds with accurate measurements of intrinsic solubility
  - Build model on that set or use existing solubility model
  - Predict on test set of 32 compounds
- 50 cpds are present in StarDrop modelling set, 3 pairs of experimental values are very different, $r^2=0.53$
- We did not participate

Solubility Challenge: Results

Model	Full 32 cpds % ±0.5 lu	28 cpds * % ±0.5 lu	28 cpds r^2_{corr}	24 cpds ** % ±0.5 lu	24 cpds r^2_{corr}
logS old	46.9	50	0.31 **# 62**	58.3	0.84 **# 1**
logS new	40.1	42.9	0.32 **# 58**	50	0.82 **# 3**
Ranges	15.6 - 62.5	10.7 – 60.7	0.02 – 0.65	12.5 – 70.8	0 – 0.835

* 4 cpds did not have meas. logS

** 4 worst outliers removed

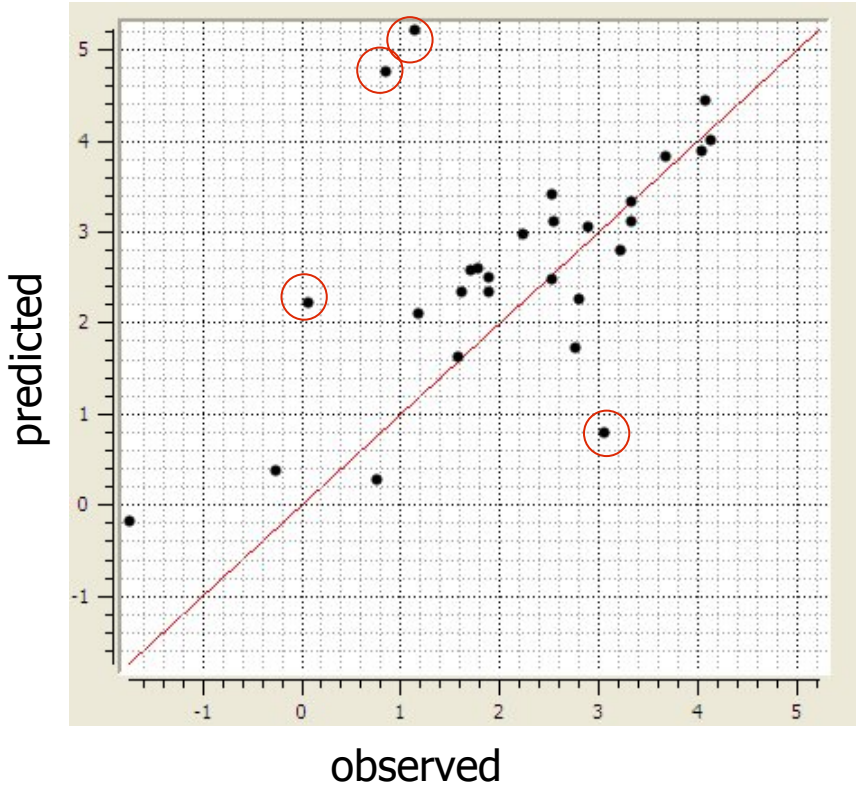
99 participants

Places were not allocated

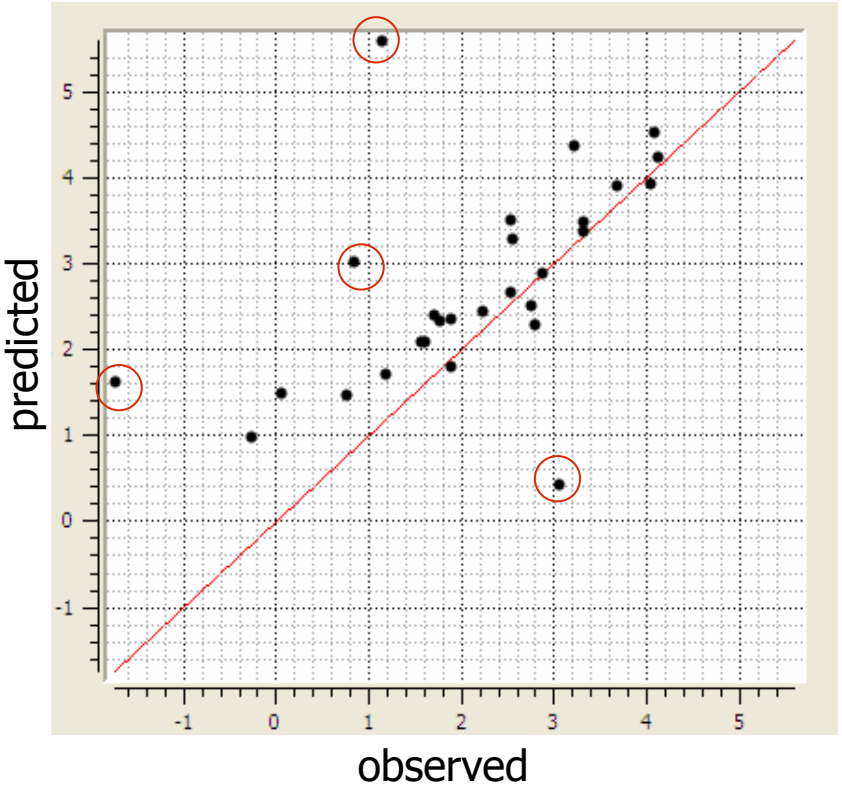
StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Solubility Challenge:
Predicted versus observed on 28 compounds



new logS



old logS

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Conclusions

- Building StarDrop solubility models
  - Described the automatic model generation process for QSAR modelling
  - 'Automatic' aqueous solubility model compares well to one built 'manually', it reports lower RMSE.
  - The automatic process is robust, much quicker than manual building and can be applied by non-experts
- Comparative evaluation of solubility models
  - Need to evaluate on real unseen data, not used in building the model!

Acknowledgements

- Matthew Segall
- Chris Leeding
- Ed Champness
- Joelle Gola

Results of solubility models comparison:

- David Sys
- Pieter Stouten

Spare slides

BioFocus DPI
A Galápagos Company

Solubility at pH 7.4

BioFocus DPI
A Galápagos Company

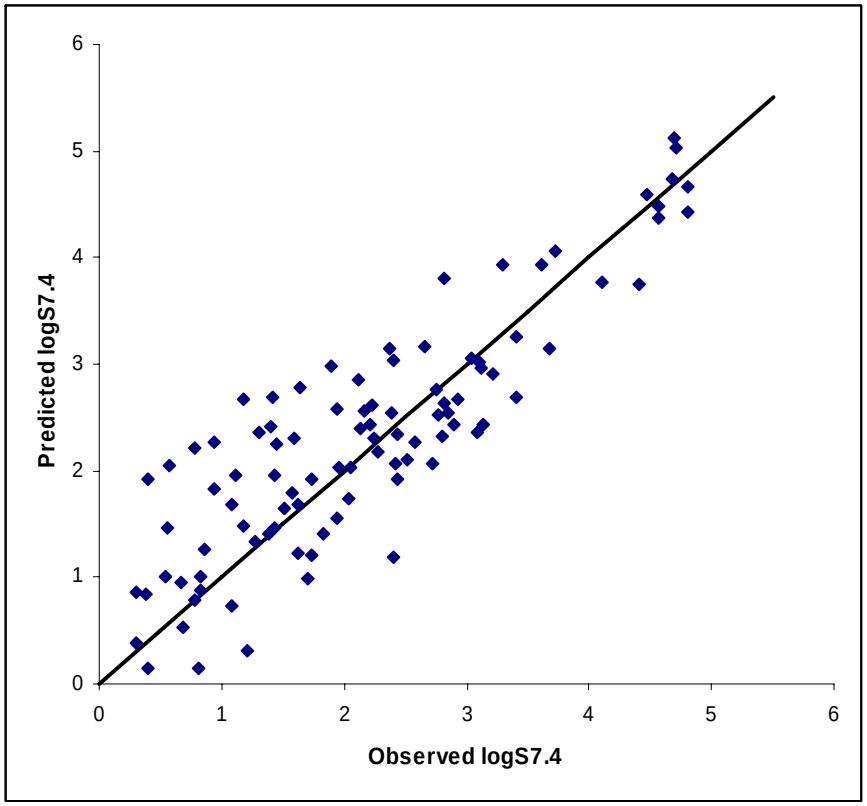
LogS @ pH7.4

- Apparent solubility of ionized compounds at pH7.4 ($logS_{7.4}$ with $S_{7.4}$ in µM)
  - 322 compounds, measured in buffered solution at 25-35°C
  - gathered from StARLITe database
- Built by Auto-Modeler, cluster split t=0.7 (test set - 96 cpds), RBF technique with genetic algorithm
- 28 descriptors (logP, negative charge, counts of groups and fragments …)

Val+Test set	
R^2 val	0.74
R^2 test	0.74
RMSE	0.61

LogS @ pH7.4 model performance on groups of compounds

Group	Val+Test set	
	%	RMSE
Overall		0.61
Acidic	14	0.41
Basic	54	0.60
Zwitterionic	32	0.69



StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

'Automatic' logS model performance on groups of compounds

Group	Trn set		Val set		Test set		Huuskonen set	
	%	RMSE	%	RMSE	%	RMSE	%	RMSE
Overall		0.66		0.68		0.66		0.96
Neutral	66	0.63	66	0.66	68	0.62	80	1.01
Acidic	15	0.71	17	0.76	16	0.69	7	0.74
Basic	15	0.72	13	0.63	13	0.83	10	0.63
Zwitterionic	4	0.75	4	0.44	3	0.56	3	0.9

Galapagos Study: Results

Pearson's correlation coefficient (r)

	ACDlabs						StarDrop		
	intrinsic	pure water	at pH7.4	pure water pH7.4	Tetko	Cerius	logS old man	logS new auto	logS at pH7.4
Trn set 2650 cpds	0.91	0.91	0.80	0.40	0.80	0.81	0.99 *	0.93	0.82
Test set 663 cpds	0.91	0.92	0.78	0.43	0.80	0.84	0.92	0.94	0.77
Huuskonen 564 cpds	n/a	n/a	n/a	n/a	0.93	0.92	0.89	0.93	0.82

* RBF model, complete fit on training set

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company