Automated QSAR Modeling to Guide Drug Design

Olga Obrezanova and Matthew D. Segall

ASC National Meeting

Spring 2009

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
- Hence the requirement for unsupervised computational techniques
  - Bayesian Neural Networks, Associative Neural Networks, Gaussian Processes ...

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

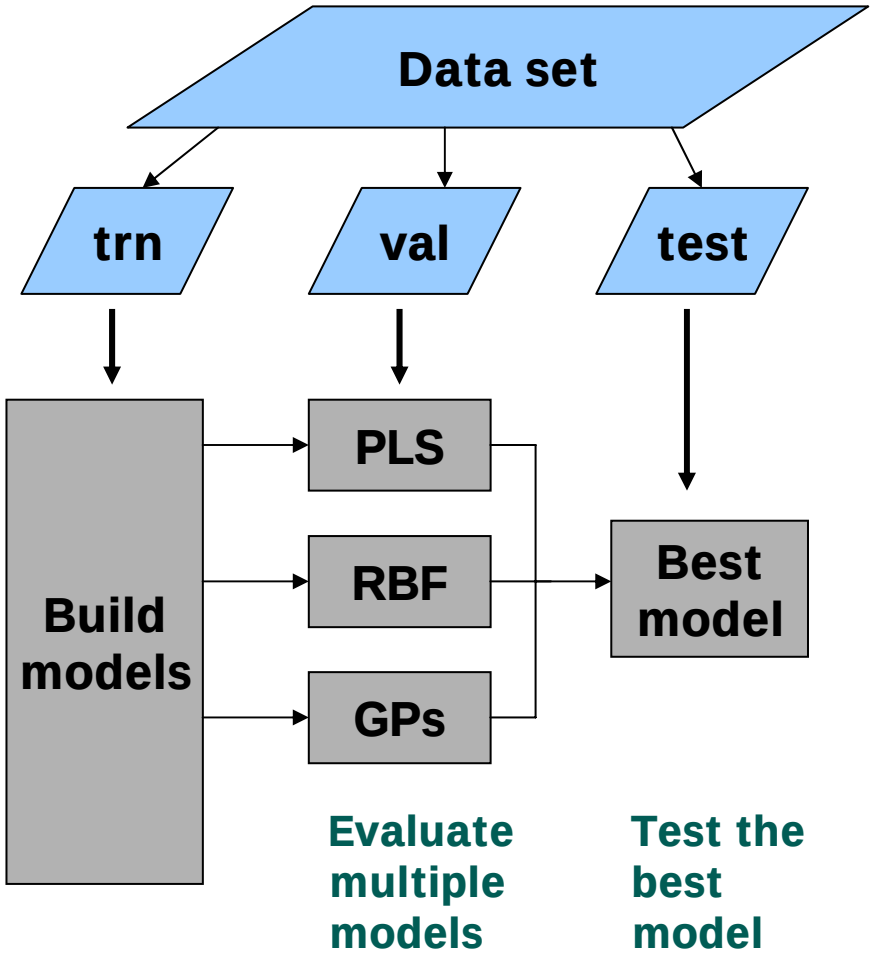
Talk Outline

- Automatic Model Generation process (AMG)
  - Stages of the process
  - Gaussian Processes modelling techniques
- 'Manual' model versus 'automatic'
  - Blood-brain barrier penetration
  - Aqueous solubility
- Building QSAR model to guide drug design

Automatic Model Generation Process

BioFocus DPI
A Galápagos Company

Automatic model generation

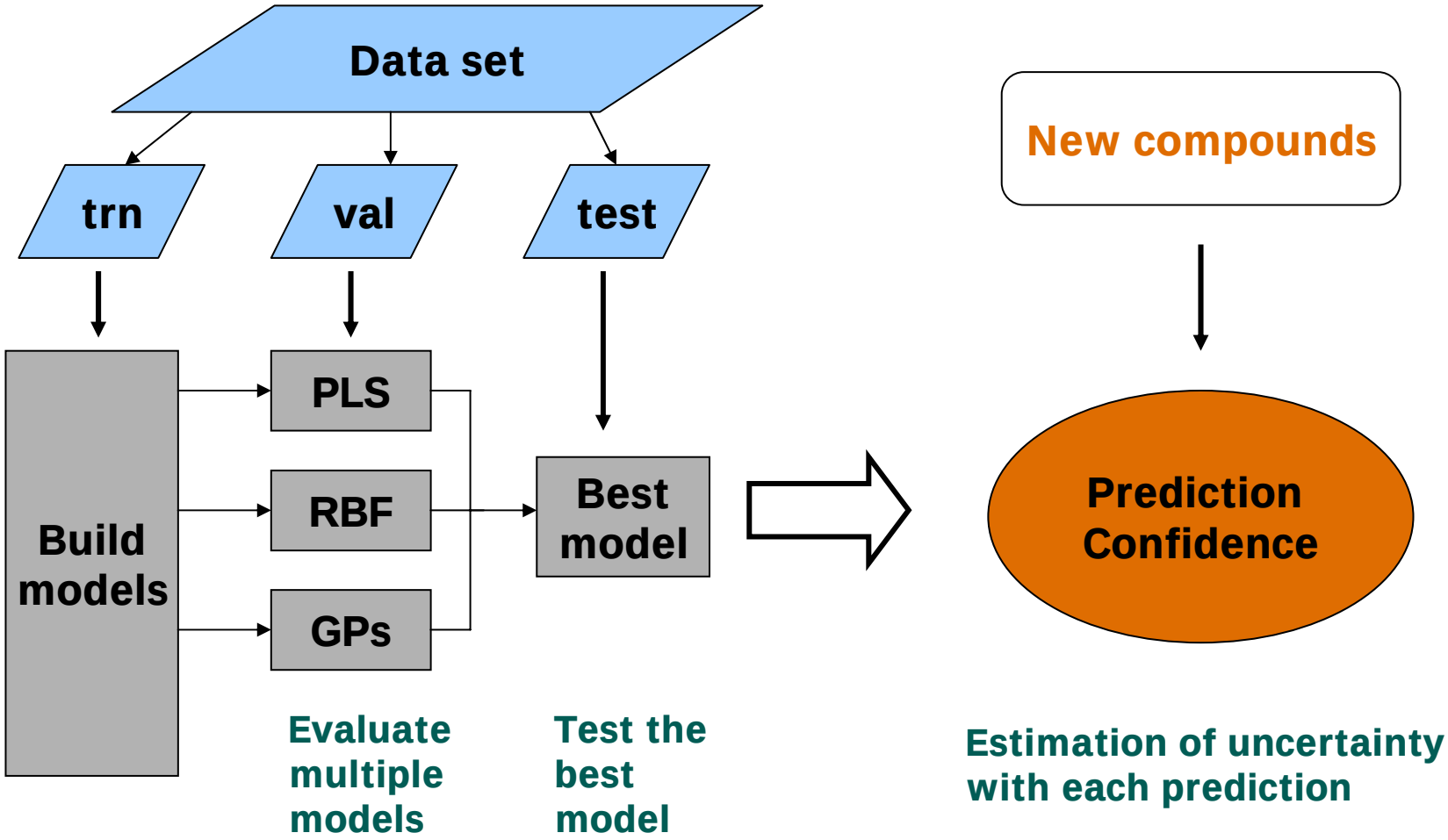


- Splitting data into training, validation and test sets (by cluster analysis)
- Descriptor calculation and filtering (2D SMARTS, logP, TPSA, MW, charge etc.)
- Modelling techniques (PLS, Radial Basis Functions with genetic algorithm, Gaussian Processes, Decision Trees)
- Selection of the best model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

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Automatic model generation



StarDrop™
Balanced compounds from better decisions

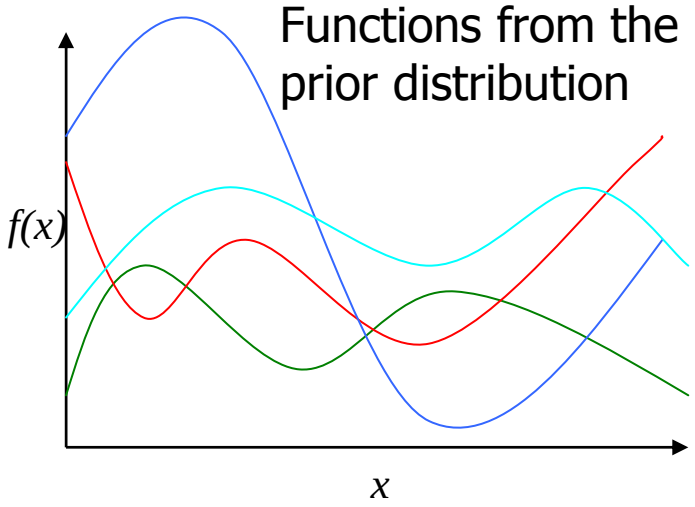
BioFocus DPI
A Galápagos Company

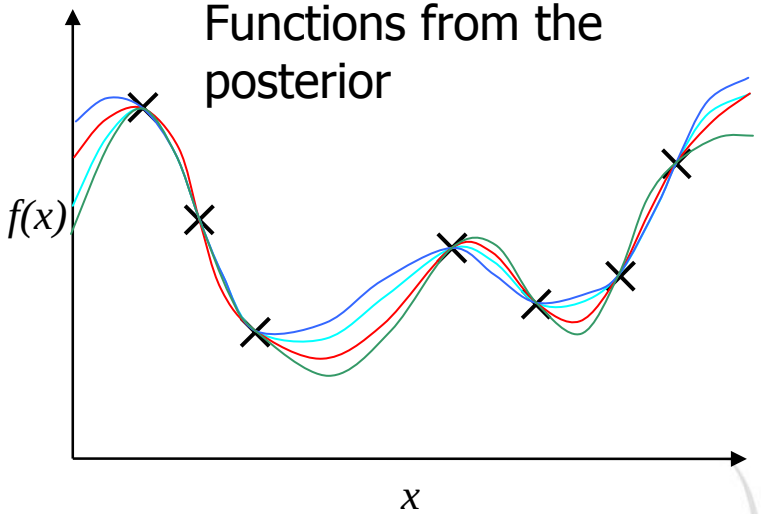
Modelling techniques: Gaussian Processes

- 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

Gaussian Processes: Hyperparameters

- Learning the Gaussian Process ~ finding hyperparameters
  - Optimize the marginal log-likelihood (prevents overtraning, no need for validation set)
- Techniques for finding hyperparameters
  - "Fixed" values for length scales. Search for noise parameter
  - Forward variable selection provides feature selection
  - Optimization by conjugate gradient methods
    - Length scales show which descriptors are most relevant
  - Nested sampling
    - Search in the full hyperparameter space
    - Search does not get trapped in local maxima

‘Automatic’ models versus ‘manual’

Applications:

blood-brain barrier penetration

and aqueous solubility

BioFocus DPI
A Galápagos Company

Blood-brain barrier penetration (logBB)

- Data set of 151 compounds with logBB values (collected from literature)
- 'Manually' built model (random set split TRN=108, TEST=43)
- Build a model by the automatic model generation (AMG) process (apply to all 151 compounds)
- Compare 'automatic' and 'manual' models by testing on external data
  - 143 compounds from 'Abraham' set not present in the initial set (Abraham et al. J.Pharm. Sci., 2006, 95)

Blood-brain barrier penetration (logBB)

- 'Manual' model
  - 2D SMARTS descriptors reduced by FVS, various modelling techniques (PLS, RBF, MLR) – performance supervised on test set
  - Final model is built by Radial Basis Functions on 7 descriptors (logP, flexibility, charge, hydrogen bonding…)
- 'Automatic' model
  - cluster at t=0.7, val=23 comp, test=22 comp
  - Best model by GP with nested sampling
  - 162 descriptors

manual

Test set	
R^2	0.73
RMSE	0.36

automatic

Val+Test set	
R^2 val	0.72
R^2 test	0.66
RMSE	0.44

StarDrop™ Balanced compounds from better decisions

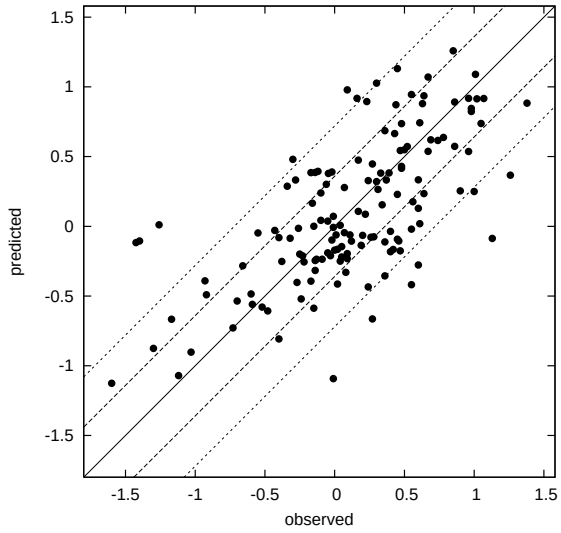
BioFocus DPI A Galápagos Company

Blood-brain barrier penetration

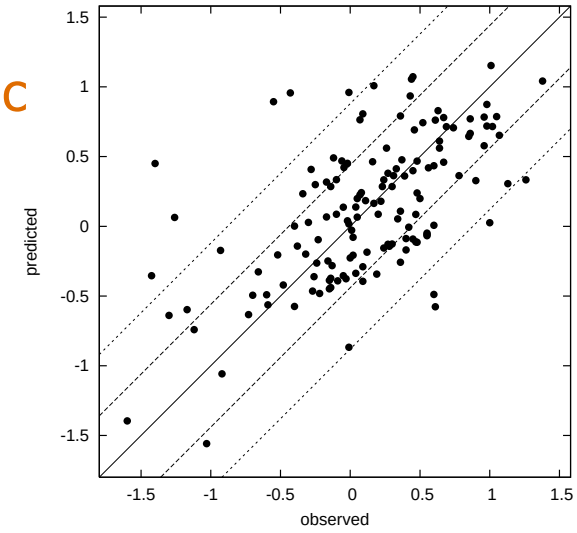
Performance on external 'Abraham' test set

Model	RMSE pred	% pred within ±0.4 log unit	% pred within ±0.8 log unit	R^2	r^2_{corr}	RMSE
manual	0.36	62.9	93.0	0.39	0.44	0.44
automatic	0.44	63.6	90.9	0.27	0.36	0.49

manual



automatic



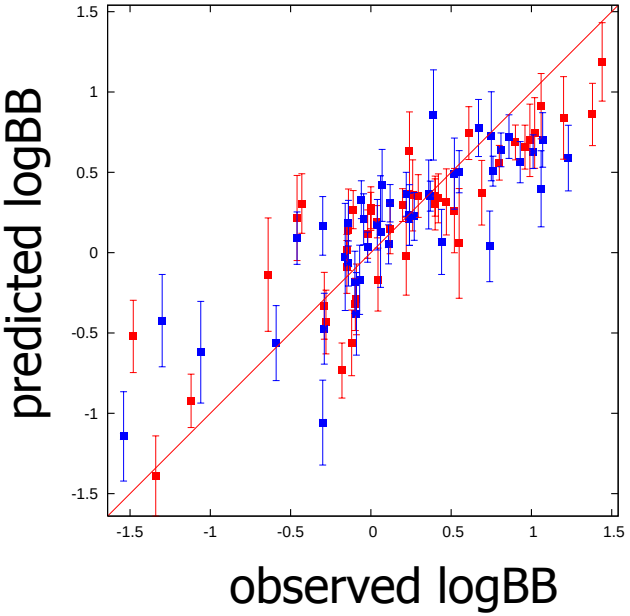
StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Automatic rebuilding logBB model to include new data

- Original 151 compounds and 143 compounds from 'Abraham' set
- Best model – GP with 2Dsearch on 167 descriptors:

Set	N	R^2	RMSE
TRN	205	0.80	0.29
VAL	44	0.73	0.33
TEST	43	0.67	0.35



- Improvement in prediction of 30 compounds from 'Abraham' set, now in val and test sets:
  - current model – RMSE=0.27
  - previous automatic model – RMSE=0.44

val set in red
test set in blue

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Aqueous solubility

- 3313 compounds with intrinsic aqueous solubility (logS, S in µM)
  - from PHYSPROP database (Syracuse Research Corporation, SRC)
- Automatic model produced by Gaussian Processes with 2D search
- External test data – 564 compounds from 'Huuskonen' set
  - Huuskonen J., J. Chem. Inf. Comput. Sci., 2002, 42

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

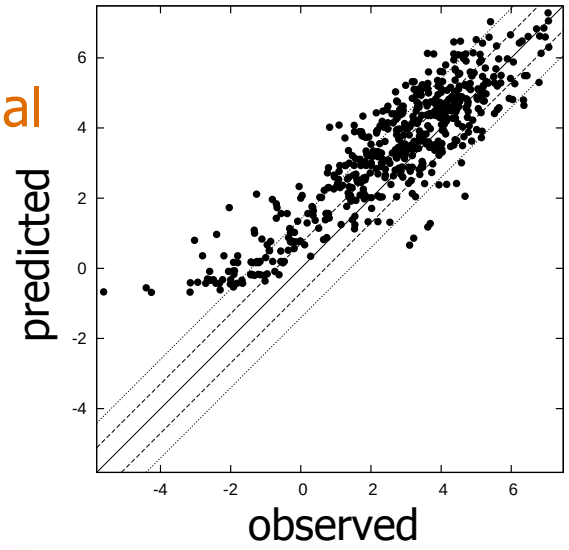
BioFocus DPI
A Galápagos Company

Aqueous solubility

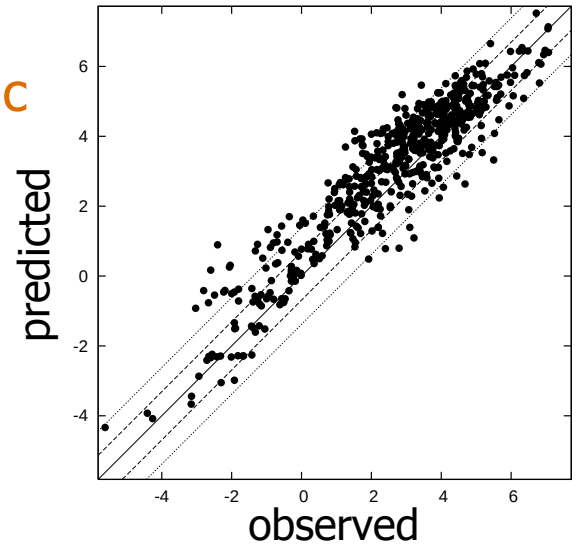
Performance on external 'Huuskonen' test set

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

manual



automatic



StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

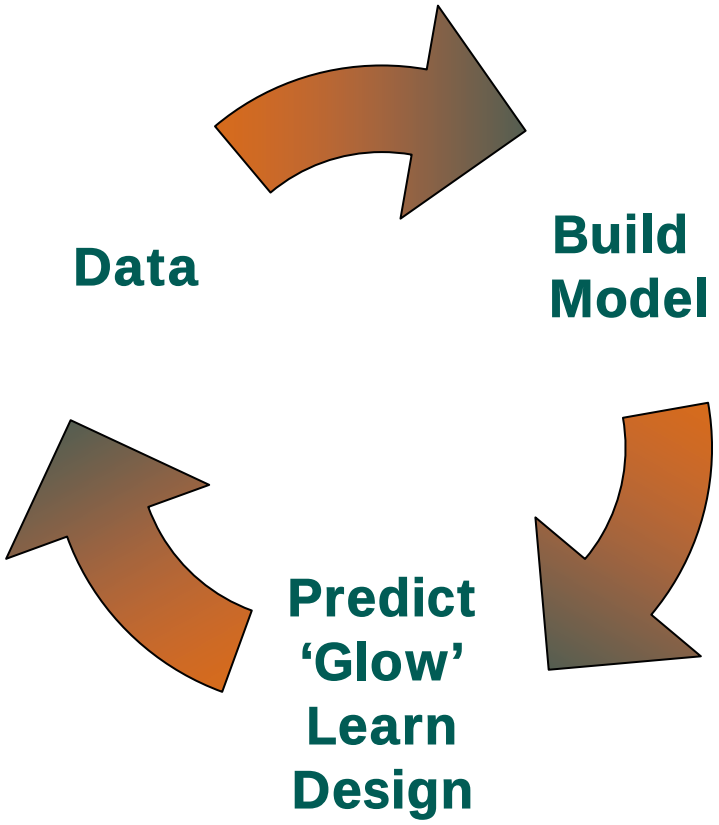
Building QSAR model to guide drug design

Case Study

BioFocus DPI
A Galápagos Company

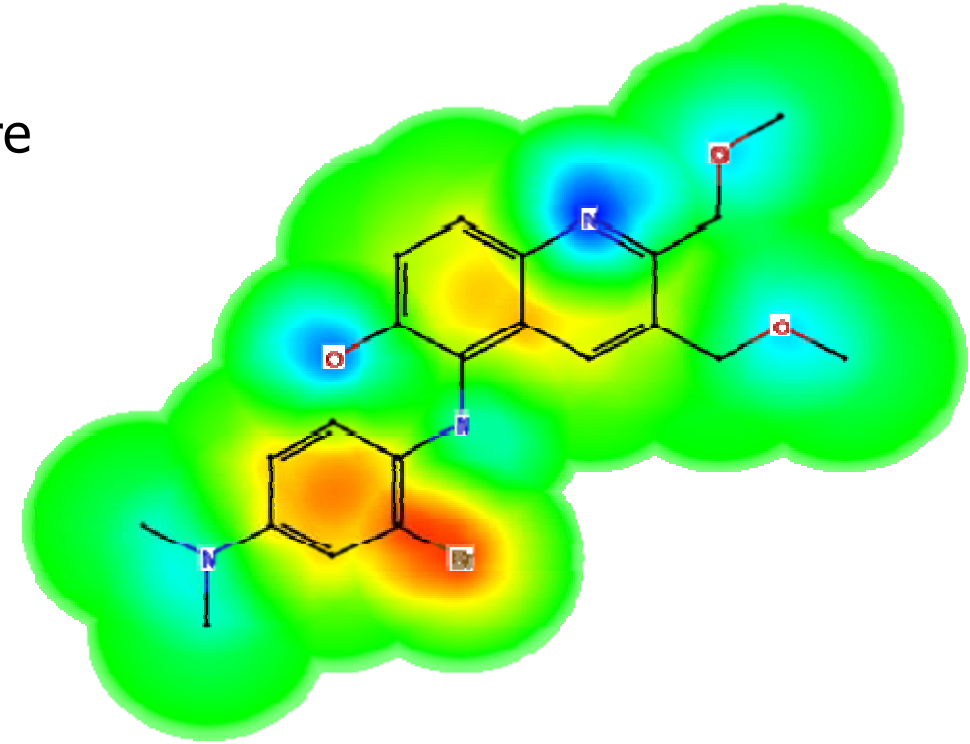
Building QSAR model to guide drug design

- The automatic model generation algorithm is implemented in the **StarDrop** environment for decision support in drug discovery and is referred to as the **Auto-Modeler**
- QSAR models can be used to predict new compounds together with the **Glowing Molecule** visualisation tool
- Interpret SAR and guide redesign of compounds to overcome liabilities



The 'Glowing Molecule': visualisation tool

- Makes a link between predicted property and compound's structure
  - "Why is a property value predicted?"
  - "Where can I change this property?"
  - Interpret SAR
  - Guide efficient redesign of molecules
- No-more 'black box' models!



logP property

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Building QSAR model

- QSAR model for Target X affinity
  - 138 compounds with pKi data from screening against 'Target X'
  - Apply Auto-Modeler

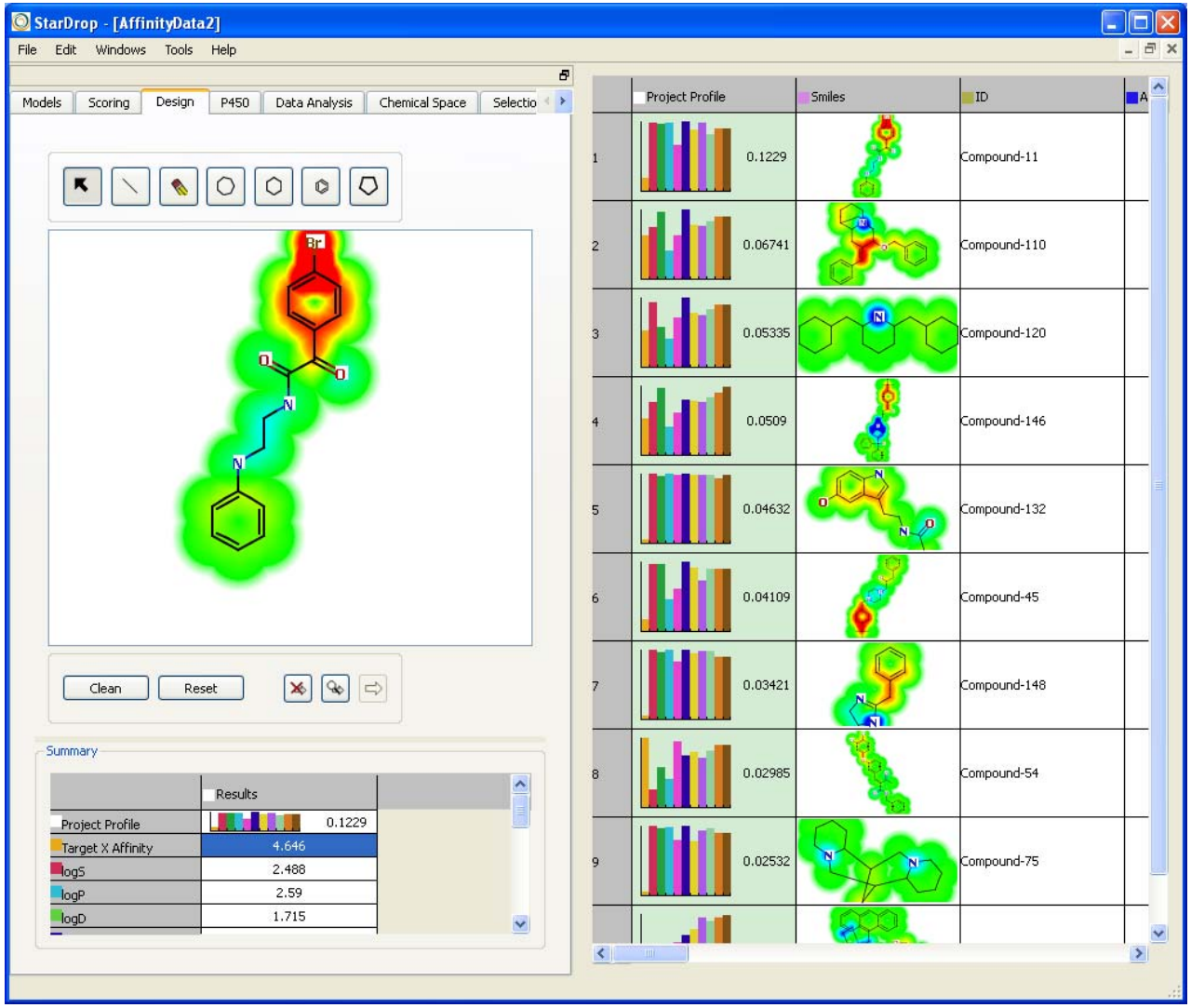
    The best QSAR model of affinity:

Set	R^2	RMSE
VAL	0.96	0.23
TEST	0.95	0.29

- Predicting affinity
  - Additional experimental affinity data for 10 new compounds
  - Model predictions correlate very well with the experimental data R^2=0.98, RMSE=0.22

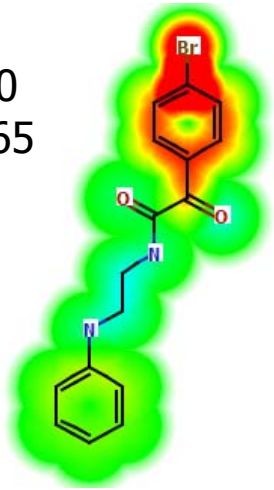
Balance of potency and ADME properties

- Need to have balance of potency and ADME properties, hence incorporate predictions from StarDrop ADME models (logS, hERG, BBB, HIA, PPB, logP, 2C9 affinity, pgp …)
- Apply probabilistic scoring – all compound data integrated to allow prioritization
- Score new molecules against project profile
  - Scoring profile is for an orally bioavailable, potent molecule for a non-CNS target (incorporates desired project criteria and their importance)
- Resulting score estimates each compound's likelihood of success against the project profile

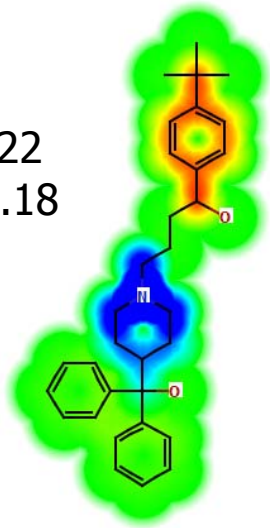

StarDrop - [AffinityData2]
File Edit Windows Tools Help
Models Scoring Design P450 Data Analysis Chemical Space Selectio
Clean Reset
Summary
Results
Project Profile 0.1229
Target X Affinity 4.646
logS 2.488
logP 2.59
logD 1.715
Project Profile Smiles ID
1 0.1229 Compound-11
2 0.06741 Compound-110
3 0.05335 Compound-120
4 0.0509 Compound-146
5 0.04632 Compound-132
6 0.04109 Compound-45
7 0.03421 Compound-148
8 0.02985 Compound-54
9 0.02532 Compound-75

Compound redesign

Exp. pKi=4.60
Pred. pKi=4.65
Score=0.12



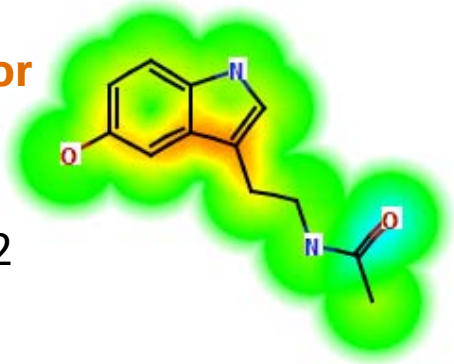
Exp. pKi=6.22
Pred. pKi=6.18
Score=0.05



A para-substituted phenyl has positive influence to the high affinity

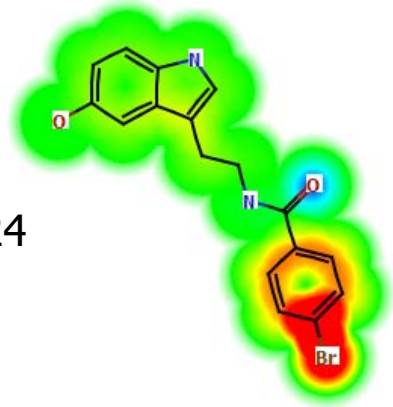
Compound for redesign

Exp. pKi=3.73
Pred. pKi=3.72
Score=0.04



New compound

Exp. pKi=n/a
Pred. pKi=5.24
Score=0.32



Adding a para-substituted phenyl improved affinity and increased the total score

StarDrop™
Balanced compounds from better decisions

BioFocus DPI
A Galápagos Company

Conclusions

- Described the automatic model generation process for QSAR modelling
- The process was applied to modelling blood-brain barrier penetration and aqueous solubility
- Automatic models compare well to ones built manually. The automatic process is robust, much quicker than manual building and can be applied by non-experts
- The case study demonstrates how building a QSAR model can help to understand SAR for a chemical series and redesign compounds to overcome liabilities

Thanks

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

http://www.biofocusdpi.com/stardrop