Automated QSAR Modeling to Guide Drug Design

Olga Obrezanova, Joëlle Gola, Ed Champness, Matthew Segall

BioFocus DPI, Chesterford Research Park, Saffron Walden, Essex, CB10 1XL, UK, Matt.Segall@glpg.com

StarDrop™
Balanced compounds from better decisions

Automatic Model Generation Process

The rapid design-test-redesign cycles of modern drug discovery and the demand for fast model (re)building whenever data becomes available have given rise to a trend to develop computational algorithms for automatic model generation. Automatic modelling processes allow computational scientists to explore large numbers of modelling approaches very efficiently and make QSAR/QSPR model building accessible to non-experts.

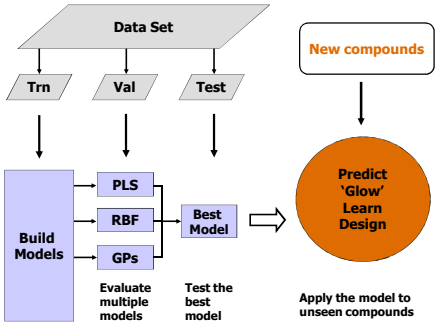


Figure 1. Stages of the Auto-Modeler™.

In this poster we will present an automatic model generation process for building QSAR models. The stages of the process that ensure models are built and validated within a rigorous framework are:

- Splitting data into training, validation and test sets (by cluster analysis)
- Descriptor calculation and filtering (2D SMARTS descriptors, whole molecular properties and user's imported descriptors)
- Application of modelling techniques (PLS, Radial Basis Functions with genetic algorithm, Gaussian Processes (GP) [1])
- Selection of the best model (performance on the validation set is used as criterion) and evaluating it on the test set

This algorithm is implemented in the **StarDrop** environment for decision support within drug discovery and is referred to as the **Auto-Modeler**.

A model can be used to predict values for new compounds and together with the **Glowing Molecule** visualisation tool can help to interpret the SAR for a chemical series and to guide redesign of compounds to overcome liabilities.

'Manual' Models versus 'Automatic'

We applied this automatic process to data sets for blood-brain barrier penetration and aqueous solubility and compared the resulting automatically generated models with models developed 'manually' by computational chemists by test them on new external data. The results demonstrate the effectiveness of the automatic model generation process for two types of data sets commonly encountered in building ADME QSAR models, a small set of *in vivo* data and a large set of physico-chemical data (see [2] for details of the study).

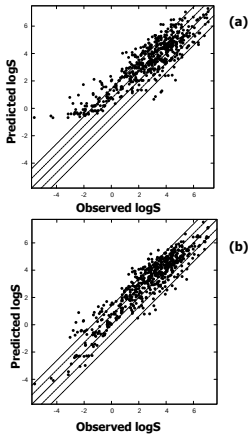


Figure 2. Predicted logS for 'Huuskonen' set by (a) manual model and by (b) automatic model.

Blood-brain barrier penetration (logBB)

- Data set of 151 compounds with logBB values derived from various sources. 'Manual' model achieved R^2=0.73 and RMSE=0.36 log units on internal test set.
- The best automatic model is produced by GP technique with nested sampling. It achieved R^2=0.72 (val), R^2=0.66 (test) and RMSE=0.44 log units on combined test and val sets.
- External test data – 143 compounds from 'Abraham' set [3] not present in the initial set:

Model	Desc	% pred within ±0.4 log unit	% pred within ±0.8 log unit	R^2	r^2_{corr}	RMSE
manual	7	62.9	93.0	0.39	0.44	0.44
automatic	162	63.6	90.9	0.27	0.36	0.49

- The low R^2 for both models on the external test set is due to the small range of experimental values represented therein. However, the RMSEs of both models are good and comparable.

Aqueous solubility (logS, S in μM)

- Data set of 3313 compounds with experimental solubility values from Syracuse database. 'Manual' model achieved R^2=0.82 and RMSE=0.79 log units on the test set.
- The best automatic model is produced by GP-2DSearch technique. It achieved R^2=0.85 (val), R^2=0.84 (test) and RMSE=0.69 log units on combined test and val sets.
- External test data – 564 compounds from 'Huuskonen' set [4] not present in the initial 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

Building QSAR Model to Guide Compound Design

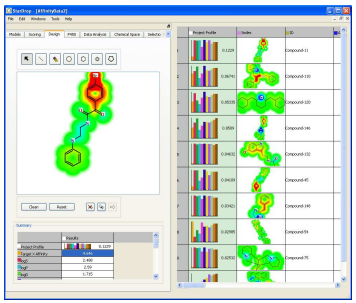
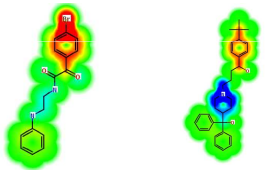


Figure 3. Scoring compounds and using the Glowing Molecule and QSAR model predictions to guide compound design.



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

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

Figure 4. Glowing Molecule examples suggest that a para-substituted phenyl has positive influence to the high affinity.

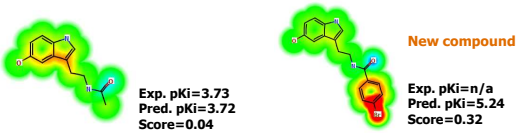
QSAR model for Target X affinity. We applied the Auto-Modeler to a set of 138 compounds with pKi data from screening against 'Target X', the therapeutic target for a drug discovery project. The best QSAR model of Target X affinity achieved R^2=0.96 and RMSE=0.23 log units on validation set and R^2=0.95 and RMSE=0.29 log units on the test set.

Predicting affinity. Additional experimental affinity data was subsequently gathered for 10 new compounds. Affinity values predicted by the model correlate very well with the experimental values for these new compounds (R^2=0.98, RMSE=0.22).

Scoring against project profile. We also used proprietary ADME QSAR models from StarDrop to predict a range of ADME properties for the set of 10 compounds. The Probabilistic scoring functionality from StarDrop allows all the compound data to be rapidly integrated to prioritise compounds. A scoring profile, incorporating all the project criteria and their relative importance, has been defined for an orally bioavailable, potent molecule for a non-CNS target. This includes the predicted potency against Target X. The resulting scores estimate each compound's likelihood of success against the project profile. As seen from Fig. 3, the top scoring compound has a relatively low potency (light orange bar in histogram), while some compounds have high affinity for Target X but poor balance of ADME properties.

Glowing Molecule. This visualisation tool highlights regions of a molecule that most strongly influence a predicted property or activity. Examples in Fig. 4 show that a para-substituted phenyl contributes positively to the predicted affinity for Target X (red 'glow').

Design. Compound in row 5 has a best balance of ADME properties but low affinity for Target X. Addition of a para-substituted phenyl improves predicted potency. The new compound is predicted to have a better balance of potency and ADME properties.



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

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

Figure 5. Interactive design of a new compound. Adding a para-substituted phenyl improved affinity to Target X and increased the total score.

Conclusions

- We have described an automatic model generation process for QSAR modelling implemented in the Auto-Modeler functionality of StarDrop.
- We have applied the Auto-Modeler to build blood-brain barrier penetration and aqueous solubility models. In the case of blood-brain barrier penetration, it can be seen that the automatically built model reports a slightly higher but comparable RMSE to the original manual model. For the aqueous solubility, the automatically built model reports a lower RMSE, i.e. higher accuracy, than the manual model. We have demonstrated that the performance of the automatic model generation process is robust and comparable to manual model building. Additionally, it is much quicker than manual modelling 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 to redesign compounds to overcome liabilities. We have built a QSAR model of target activity and used this, combined with Glowing Molecule visualisation, to guide the design of a new compound that is predicted to have a better balance of potency and ADME properties.

References

[1] Obrezanova et al. J. Chem. Inf. Model., 2006, **47**, pp. 1847-1857
[2] Obrezanova et al. J. Comput. Aided Mol. Des., accepted for publication, 2008.
[3] Abraham et al. J.Pharm. Sci., 2006, **95**, p. 2091.
[4] Huuskonen J., J. Chem. Inf. Comput. Sci., 2002, **42**, p. 651.

BioFocus DPI
A Galápagos Company