

QSAR modeller seeks meaningful relationship

C.L. Bruce¹

S.D. Pickett², J.D. Hirst¹

¹ *School of Chemistry, University of Nottingham, Nottingham, U.K.*

² *GlaxoSmithKline, Stevenage, U.K.*

Disappointment with QSAR has been articulated recently¹ and although the technique is an important tool in the drug discovery process, improvements perhaps have not been as forthcoming as in other areas. A good model comprises several components. Predictive accuracy is paramount, but it is not the only important aspect. In addition, one should apply robust and appropriate statistical tests to the models to assess their significance or the significance of any apparent improvements. The real impact of a QSAR, however, perhaps lies in its chemical insight and interpretation, an aspect which is often overlooked.

Any insight into the relationship between descriptors and structure can be used to further our understanding, but obtaining this insight is not always as straightforward as calculating predictive accuracy. Interpretation is dependent on the classifier. For example, a decision tree is simple to interpret, but does not produce the most predictive models. Similarly, support vector machines offer excellent predictive capability, but generate a model that is difficult to interpret.

Previously, we have shown random forests predict with accuracies comparable to support vector machines.² A decision tree is easier to interpret than a random forest; Breiman gave them an 'A+' and 'F' for interpretability, respectively.³ It is the different tree construction and number of trees present in a forest that makes their interpretation complicated. One cannot simply glance through the forest and readily see the model, whereas one can with a decision tree.

Therefore, to obtain useful interpretation from a random forest we have employed a selection of tools. This includes alternative representations of the trees using SMILES and SMARTS. Using existing methods we can compare and cluster the trees in this representation. Descriptor analysis and importance can be measured at the tree and forest level. Pathways in the trees can be compared and frequently occurring sub graphs identified. The ability to distinguish multiple modes of action in a data set is tested. In terms of model assessment, all test data can be assigned a level of confidence, reflecting the extent to which the prediction is an extrapolation from the model. These tools have been built around the Weka machine learning workbench⁴ and are designed to allow further additions of new functionality.

References:

1. Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008**, 48, 25-26.
2. Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, 47, 219-227.
3. Breiman, L. Statistical Modeling: The Two Cultures. *Statist. Sci.* **2001**, 16, 199-231.
4. Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.