Accuracy of Prediction

The non-specialist might have assumed that the main objective of a QSAR study is to predict whether an untested compound will be active or inactive (or to do virtual screening, i.e., predictions about a whole virtual library of compounds). In practice, much work has been devoted to "explanatory" QSAR, relating changes in molecular structure to changes in activity, and only recently has there been considerable interest in predictivity; QSAR is now being used for virtual screening, to find biologically active molecules. There are many reasons why models fail [1,2]: bad data, bad methodology, inappropriate descriptors, domain inapplicability [3], etc. In this article we can address only a few of the issues. Vendors are supplying models that may or may not be applicable to a corporate virtual library [2] and many (in-house approved) models are now available to non-experts on corporate intranets. How are these users to judge applicability?

Significant issues concerning accuracy of prediction are extrapolation (whether the model can be applied to molecules unlike those in the training set) and overfitting. Overfitting has been considered for a long time [4] but extrapolation has received too little attention. Running cross-validation studies on the data to get an overall rms error for prediction is a reasonable check for overfitting but it is inadequate as a measure of extrapolation [5].

The outcome of a leave one out (LOO), or leave-many-out, cross-validation procedure is cross-validated R^2 (LOO q^2). The inadequacy of q^2 as a measure of predictivity was realized more than ten years ago, in the case of 3D QSAR, in what John van Drie refers to as "the Kubinyi paradox": models that give the best retrospective fit give the worst prospective results [6]. To get good values for R^2 you should not choose the highest values of q^2. The "best fit" models are not the best ones in external prediction because internal predictivity tries to fit compounds in the training set as well as possible and does not take new compounds into account [7].

Thus, it is not fair to assume that internally cross-validated models will automatically be externally predictive. Although a low value of q^2 for the training set may well indicate low predictivity in a model, high q^2 does not necessarily imply high predictivity. While a high value of q^2 is a necessary condition for high predictive power, it is not a sufficient condition. Tropsha and his colleagues argue that a reliable model should be characterized by both high q^2 and a high correlation coefficient (or R^2) between the predicted and observed activities of compounds from a test set [8,9]. They have proposed several approaches to the division of experimental data sets into training and test sets and have formulated a set of general criteria for the evaluation of the predictive power of QSAR models.

Kubinyi has another simple explanation of the prediction paradox [1]. Even in the absence of real outliers, external prediction will be worse than fit because the model tries to "fit the errors" and attempts to explain them. Accordingly, external predictions contain the model error *and* the experimental error. When variable selection is carried out, no independent variable selection is performed in the cross-validation runs; correspondingly, variables that were included to "explain the error" remain in the model and cause wrong predictions [1]. The higher the number of descriptors relative to the number of compounds, the higher is the chance to select those of them that give high q^2 values [8]. Other reasons for overestimating q^2 are redundancy in the training set, or, in the case of non-linear methods, the existence of multiple minima [8].

Arthur Doweyko has also published on the elusive nature of 3D QSAR predictions [10], but concludes: "Predictions can be enhanced when the test set is bounded by the descriptor space represented in the training set. Interpretation of significant interaction regions becomes more meaningful when alignment is constrained by a binding site".

At a workshop held in Setubal, Portugal in 2002, a set of principles was proposed to define the validity and applicability domain of QSAR models. These then evolved into the OECD principles in 2004 [11]. Paola Gramatica discusses three of these principles in a recent publication [12], and

in particular, emphasizes the need for external validation using at least 20% of the data. Gramatica, Tropsha and others believe that validation is the absolute essential for successful application and interpretation of QSAR models [3,13].

The necessity for validation has been accepted by leading journals. The policy of *J. Chem. Inf. Model.* on QSAR manuscripts [14] has been adopted by other journals such as *J. Med. Chem.* [15] and *ChemMedChem*. In part, it states: "If a new method/theory is being reported in the paper, it should be compared and "validated" against at least one other common data set for which a published study exists, using at least one other method/approach and preferably a method/approach that has been widely used in the field. The data set should not be small… Evidence that any reported QSAR/QSPR model has been properly validated using data not in the training set must be provided".

Researchers at Merck [5] have proposed a way to estimate the reliability of the prediction for an arbitrary chemical structure, using a given QSAR model, given the training set from which the model was derived. Based on a set of retrospective cross-validation experiments using 20 diverse in-house activity sets, they found two useful measures: the similarity of the molecule to be predicted to the nearest molecule in the training set and the number of neighbors in the training set, where neighbors are those more similar than a user-chosen cut-off. The molecules with the highest similarity and/or the most neighbors are the best-predicted, even for many diverse training sets (though to a lesser degree). The result does not depend on which QSAR method or descriptor is used. Three years later, workers at Strand Life Sciences, unaware of the Merck publication, drew similar conclusions [16].

Nevertheless, says Gerry Maggiora, incorrect predictions of activity still arise among *similar* molecules even in cases where overall predictivity is high, because, in his metaphor, activity landscapes are not always like gently rolling hills, but may be more like the rugged landscape of the Bryce Canyon [17]. Even very local, linear models cannot account satisfactorily for landscapes with lots of "cliffs", and perfectly valid data points located in cliff regions may *appear* to be outliers, even though they are perfectly valid data points. It may also be necessary to assay additional compounds in the neighborhoods around the cliffs, to ensure that activity landscapes are adequately represented in these rapidly varying regions. Maggiora also discusses the consequences of lack of invariance of chemical space to changes in the set of descriptors.

Bob Clark referred to "clumpy" data sets, rather than "activity cliffs" in a recent presentation [18]. A larger data set is not necessarily better than a smaller one in the case of cross-validation: larger data sets in which the observations are unevenly distributed through the descriptor space are particularly susceptible to problematic distortions of the validation statistics. Clark's paper was given in a symposium on evaluation of computational methods at the Fall 2007 ACS Meeting. Papers arising from that symposium, selected by guest editors, should shortly appear in the *Journal of Computer-Aided Molecular Design.*

Only a small number of the oral presentations related to QSAR; most papers concerned measures of the quality of docking results. In his concluding remarks, Terry Stouch said that there was agreement on the need for better test, validation, and decoy sets and we are approaching agreement on what more is necessary [19]. Two significant new data sets are now available for testing docking algorithms: a Directory of Useful Decoys (DUD) [20,21] and WOMBAT Data for Enrichment Studies [22,23]. Is there a need for newer, better QSAR data sets and what should be the criteria for building them? Since QSAR and docking are both being used now for virtual high throughput screening, comparisons of the two methods are likely to be of interest. I can see here topics worthy of further discussion in *QSARWorld*. The Spring 2008 ACS meeting also promises a symposium entitled "Model Applicability Domains: When Can I Use my Model?". Maybe I will be writing more about accuracy of prediction for *QSARWorld* in 2008.

References

1. Kubinyi, H. Why models fail. http://americanchemicalsociety.mediasite.com/acs/viewer/?peid=7a194d47-baa9-4b2d-a823-1fd17bf5301c (accessed 21st September 2007).
2. Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X.-Q.; Doweyko, A.; Li, Y. *In Silico* ADME/Tox: Why Models Fail. *J. Comput.-Aided Mol. Des.* **2003**, 17, 83-92.
3. Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, in press.
4. Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1-12.
5. Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912-1928.
6. Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553-2564.
7. Kubinyi, H. Validation and Predictivity of QSAR Models. In *QSAR & Molecular Modelling in Rational Design of Bioactive Molecules (Proceedings of the 15th European Symposium on QSAR & Molecular Modelling, Istanbul, Turkey, 2004)*; Sener, E. A.; Yalcin, I. Eds.; CADDD Society: Ankara, Turkey, 2006; pp. 30-33.
8. Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269-276.
9. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational Selection of Training and Test Sets for the Development of Validated QSAR Models. *J. Comput.-Aided Mol. Des.* **2003**, 17, 241-253.
10. Doweyko, A. 3D-QSAR Illusions. *J. Comput.-Aided Mol. Des.* **2004**, 18, 587–596.
11. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. http://www.oecd.org/dataoecd/33/37/37849783.pdf (accessed September 19, 2007).
12. Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26*(5) 694-701.
13. Tropsha, A.; Gramatica, P. Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.*, **2003**, *22*, 69-77.
14. Jorgensen, W. QSAR/QSPR and Proprietary Data. *J. Chem. Inf. Model.* **2006**, *46*, 937-937.
15. Editorial. QSAR/QSPR and Proprietary Data. *J. Med. Chem.* **2006**, *49*, 3431-3431.
16. Dogra, S.; Jamois, E. A Composite Predictability Metric that Combines Structure- and Descriptor-based Similarity into a Single Measure. Poster Presented at CHI's Fourteenth Molecular Medicine Tri-Conference, San Francisco, CA, February 27 - March 2, 2007. *QSARWorld*, Poster Gallery. http://www.qsarworld.com/qsar-poster-gallery.php?mm=9 (accessed September 19, 2007).
17. Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535-1535.
18. Clark, R. D. Validation and the Downside of the Law of Large Numbers. In *Abstracts of Papers*, 234th ACS National Meeting, Boston, MA, USA, August 19-23, 2007; American Chemical Society, Washington, DC, 2007; COMP-265.
19. Stouch, T. R. Private communication.
20. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789-6801.
21. DUD: a Directory of Useful Decoys. http://dud.docking.org/ (accessed September 23, 2007).
22. Good, A. C. Virtual Screening Enrichment Studies: a Help or Hindrance in Tool Selection? In *Abstracts of Papers*, 234th ACS National Meeting, Boston, MA, USA, August 19-23, 2007; American Chemical Society, Washington, DC, 2007; COMP-266.
23. WOMBAT data presented in reference 22. http://dud.docking.org/wombat/ (accessed September 23, 2007).

Dr. Wendy A. Warr, Wendy Warr & Associates (wendy@warr.com, http://www.warr.com), September 2007