# The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy)

Stephen R. Johnson*

Bristol-Myers Squibb, Co., Princeton, New Jersey 08543

Received September 7, 2007

**Abstract:** A general feeling of disillusionment with QSAR has settled across the modeling community in recent years. Most practitioners seem to agree that QSAR has not fulfilled the expectations set for its ability to predict biological activity. Among the possible reasons that have been proposed recently for this disappointment are chance correlation, rough response surfaces, incorrect functional forms, and overtraining. Undoubtedly, each of these plays an important role in the lack of predictivity seen in most QSAR models. Likely to be just as important is the role of the fallacy *cum hoc ergo propter hoc* in the poor prediction seen with many QSAR models. By embracing fallacy along with an over reliance on statistical inference, it may well be that the manner in which QSAR is practiced is more responsible for its lack of success than any other innate cause.

In recent years, we have witnessed an increased alertness to the pitfalls of QSAR. There has been an increased awareness of the impotence of $q^2$,[1] the importance of careful statistical validation, much posturing about extrapolation monitoring,[2,3] and a renewed focus on training data quality.[4,5] However, not much has truly changed, and most in the field continue to be frustrated and disappointed with the inability of QSAR to reach its potential. All of this leads to the question—why do QSAR models continue to yield significant prediction errors for molecules similar to the training data.

Professor Maggiora[6] posits an interesting argument for why this is so—that QSAR has disappointed because the structure−activity optimization surface is not as smooth as originally anticipated. Certainly, this rationale has substantial merit as a sudden change in activity as a result of a seemingly conservative molecular change is not particularly unusual in the course of a medicinal chemistry project. Considering that an order of magnitude change in IC50 arises from only a 1.2 kcal/mol change in $\Delta G$, it is quite reasonable to imagine activity cliffs around the formation or loss of hydrogen bonds. The presence of activity cliffs would create data points with significant statistical leverage making measures like $q^2$ unreliable. Indeed, the mere existence of such cliffs would complicate outlier detection as it would be impossible to separate measurement errors from observations that simply do not obey the physical assumptions of the model.

The net effect of activity cliffs on model development and reliability is dependent on how frequently these activity cliffs occur in the particular chemical space that is relevant to the activity being modeled. One would expect that with the application of a modeling algorithm that is robust to outliers, such as least median squares or support vector machines, that poor predictions would be largely limited to a few examples of a chemical series that are over the activity cliff. The other members of the series should be predicted reasonably well. In practice, however, poor prediction seems to be the rule for most compounds rather than the exception. This implies one of two possibilities—that activity cliffs are more common than the statistical breakdown point of most modeling paradigms (the Bryce Canyon analogy) or that the model itself is an incorrect representation of reality.

While the presence of some activity cliffs is almost certainly true, it is this latter possibility that seems the more significant confounding problem. How could it be that we consistently arrive at wrong models? With the near infinite number of molecular descriptors coupled with incredibly flexible machine learning algorithms, perhaps the question really should be why do we expect anything else. QSAR has devolved into a perfectly practiced art of logical fallacy. *Cum hoc ergo propter hoc* (with this, therefore because of this) is the logical fallacy in which we assign causality to correlated variables. To be sure, such correlations should serve as a starting point for a hypothesis regarding activity modulation. Indeed, many reports have appeared in the literature with statements akin to “correlation does not imply causation, but it is useful to consider the descriptors meaning...” followed by a series of unsubstantiated statements regarding the descriptors in the model. Rarely, if ever, are any designed experiments presented to test or challenge the interpretation of the descriptors. Occasionally, the model will be tested against a set of compounds unmeasured during the development of the model. It appears unusual, however, that they are targeted compounds specifically chosen to test a particular feature of the model. Rather, they are used as a general yes/no screen for the viability of a model. In this respect, they are fairly uninformative about how we improve the model. In short, QSAR disappoints because we have largely exchanged the tools of the scientific method in favor of a statistical sledgehammer. Statistical methodologies should be a tool of QSAR but instead have often replaced the craftsman tools of our trade—rational thought, controlled experiments, and personal observation.

Still this does not explain how we *arrive* at the wrong model, just how we *accept* the wrong model. Maggiora highlights a bigger problem than the response landscape—that of the lack of invariance of chemical space. Feature selection seeks to exploit the lack of invariance of chemical space in order to find a molecular representation that places compounds in proximity to (or, depending on the statistical method, collinear to) compounds with similar activity by iteratively searching combinations of molecular descriptors to find a set that best forces compounds onto an acceptable response landscape. In practice, acceptability is evaluated

* Corresponding author e-mail: stephen.johnson@bms.com.

10.1021/ci700332k CCC: $40.75 © xxxx American Chemical Society
Published on Web 12/28/2007 PAGE EST: 1.5

by a fitness function that selects for descriptors that predict the available activity data with as low an rms error (or greatest $R^2$, $Q^2$, etc.) as possible.

The problem with this approach is that there are typically many possible solutions that yield approximately equal statistical measures of quality. Indeed, the very trait that makes QSAR appealing—that we could identify a few molecular properties critical for activity from a nearly infinite pool of detailed possibilities (the exploitation of the lack of invariance of chemical space)—in fact makes the method nearly intractable. With such an infinite array of descriptions possible, each of which can be coupled with any of a myriad of statistical methods, the number of equivalent solutions is typically fairly substantial. Each of these equivalent solutions, however, represents a hypothesis regarding the underlying physical or biological phenomenon. It may be that each largely encodes the same basic hypothesis but only in subtly different ways. Alternatively, it may be that many of the hypotheses are distinctly different from one another in a meaningful, perhaps unclear, physical way. The common practice has been to select the model with the best fitness function score and predict a small group of observations that were withheld at the beginning. All too often, the model development process stops here, or, worse, the validation set is poorly predicted and models are iteratively tested until one predicts this set of compounds well.

So why do the resulting models perform so poorly on new compounds? The predictions on future compounds are disappointing because we chose the wrong model from our selection of near equivalent models. Reliable prediction of future compounds requires that the model have some basis in physical reality. Short of this, it is implausible that a model can be “local enough” for reliable prediction because of the sheer number of possible variations of the chemical space that are evaluated during feature selection.

QSAR suffers from the number and complexity of hypotheses that modern computing can generate. The lack of interpretability of many descriptors only further confounds QSAR. We can generate so many hypotheses, relating convoluted molecular factors to activity in such complicated ways, that the process of careful hypothesis testing so critical to scientific understanding has been circumvented in favor of blind validation tests with low resulting information content. QSAR disappoints so often, not only because the response surface is not smooth but because we have embraced the fallacy that correlation begets causation. By not following through with careful, designed, hypothesis testing we have allowed scientific thinking to be co-opted by statistics and arbitrarily defined fitness functions. Statistics must *serve* science as a tool; statistics cannot replace scientific rationality, experimental design, and personal observation.

## REFERENCES AND NOTES

(1) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(2) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G. et al. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, *45*, 839−849.

(3) Guha, R.; Jurs, P. C. Determining the validity of a QSAR model - A classification approach. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 65−73.

(4) Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X.-Q.; Doweyko, A. et al. In silico ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 83−92.

(5) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *THEOCHEM* **2003**, *622*, 39−51.

(6) Maggiora, G. M. On Outliers and Activity Cliffs—Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

CI700332K