EDITORIAL

On Outliers and Activity Cliffs—Why QSAR Often Disappoints

Quantitative structure−activity relationships (QSAR) have been around for many years and have been employed in numerous fields from drug design to environmental toxicology. Countless papers have been written employing a wide variety of descriptors and computational methods in order to determine them. Nevertheless, while the jury is still out, it is safe to say that QSAR have not generally lived up to expectations, especially in cases where they are applied to data sets determined after the QSAR models were constructed. But this is true even in many "typical cases" where all of the data are known beforehand and are divided into training and test sets in order to construct and validate a model. Certainly the number of parameters available for use in QSAR models is sufficiently large and diverse to ensure reasonable predictions of bioactivity. In fact, the number of potential parameters is so large that a significant amount of effort in QSAR modeling is centered on development and application of various dimensionality-reduction procedures. In addition, a variety of new and more effective machine learning techniques such as neural networks, decision trees, and support vector machines have become available for implementing QSAR models. Nevertheless, significant mispredictions of activity still arise among *similar* molecules even in cases where overall predictivity is high.

All of this begs the question as to why this is so. Basically, the answer is related to the nature of the *activity landscape* associated with a given assay, which is related to the chemical-space representation used to characterize the set of compounds assayed. Activity landscapes are generally of high dimension (>3) and depend on the nature of the assay (e.g., enzyme-based, cell-based, etc.), on the region(s) of chemical space from which the compounds are drawn, on the density distribution of the compounds in these regions, and, most importantly, on the nature of the molecular representation used. A typical N-dimensional activity landscape is composed of an (N-1)-dimensional chemical space; each dimension is described by a coordinate, which is generally defined by a single molecular descriptor or combination of descriptors. The Nth dimension is defined by the activity space that is derived from the measured activity of each of the assayed compounds. In three dimensions activity landscapes are closely akin to Nature's landscapes.

For many years it has been assumed that similar molecules tend to have similar activities, leading to activity landscapes comparable to the gently rolling hills found on the Kansas prairie. Mounting evidence suggests, however, that this picture is not as universal as once thought but is in many cases rather more like the rugged landscapes of Utah's Bryce Canyon. This new topographical metaphor clearly implies that very similar molecules may in some cases possess very different activities leading to what can be called *activity cliffs*—an activity cliff is defined by the ratio of the difference in activity of two compounds to their "distance" of separation in a given chemical space. The existence of such activity cliffs is not entirely surprising since molecular recognition plays a crucial role in determining activity. For example, a change as "small" as that obtained by replacing an ether oxygen by a secondary amine can have a significant effect on activity.

The greater prevalence of activity cliffs than was earlier suspected has several important and related implications for QSAR modeling. First, purely linear models, even very local ones, in which neither the parameters nor the variables are nonlinear, are unlikely to satisfactorily account for activity landscapes with significant numbers of cliffs. Second, outliers in the data may not be due to statistical fluctuations or to measurement errors but rather may reflect the presence of activity cliffs. Thus, perfectly valid data points located in cliff regions may *appear* to be outliers. Third, the presence of activity cliffs requires the assay of additional compounds in the neighborhoods around these cliffs to ensure that activity landscapes are adequately represented in these rapidly varying regions and, thus, that QSAR models can faithfully represent the SAR data.

Another crucial issue that arises here is the *lack of invariance of chemical space* to changes in the set of descriptors used to represent the molecular information in the model. Such a lack of invariance can have serious consequences, one of which is that neighborhood relationships may be significantly altered—compounds that are nearest neighbors in one chemical-space representation may not be nearest neighbors in another. Since most QSAR models are approximately local this can lead to serious problems, regardless of the power and sophistication of the methodology used to implement the model. The lack of neighborhood invariance also implies that distance relationships will not, in general, be preserved, potentially altering the magnitude and location of activity cliffs.

Addressing all of these problems is a daunting task at best, and it may not be possible to treat some of them in any substantive way. Thus, *all QSAR models are flawed to some degree* due to the limitations that derive from one or more of the problems described above. Of particular importance is the detection of *true outliers*, an inherently difficult problem that is confounded by the presence of cliffs in activity landscapes that are a function of molecular representation, as different representations can lead to dramatic changes in the nature of activity cliffs. Consequently, identifying and removing outliers may not necessarily always be a statistical problem as some outliers may only be *apparent* and may, in fact, arise from activity cliffs in the data. Moreover, while new computational methodologies such as support vector machines may ameliorate some difficulties and produce better, more predictive QSAR models, they are ultimately constrained by the quality of the data, the number and nature of the compounds in the sample, and, importantly, by the underlying characteristics of the molecular representation that define the chemical space in which the compounds lie.

Gerald M. Maggiora
University of Arizona
Tucson, Arizona

CI060117S

10.1021/ci060117s CCC: $33.50
Published on Web 07/24/2006