Trends and Plot Methods in MLR Studies

Emili Besalú*

Institute of Computational Chemistry, Universitat de Girona, Facultat de Ciències, Avda. Montilivi s/n, 17071 Girona, Spain

Jesus V. de Julián-Ortiz

Insituto de Tecnología Química, Av. de los Naranjos s/n, 46022 Valencia, Spain

Lionello Pogliani

Dipartimento di Chimica, Università della Calabria, via P. Bucci 14/C, 87036 Rende (CS), Italy

Received November 6, 2006

Regression toward the mean effects are presented within the field of quantitative structure−activity relationship modeling and in situations in which multilinear regression techniques are considered for model building. The concept is related to the graphical aspect of some scatter plots (experimental vs fitted and fitted vs experimental values). These graphs demonstrate how the point cloud is not always symmetrically distributed along the so-called "ideal" or "desired" line, that is, the bisector of the first and third quadrants. The deviation from the ideal line is fixed, and it is also related to the coefficient of determination. An extrapolation of these regression effects is also discussed within the context of property predictions obtained via the leave-one-out cross-validation technique.

INTRODUCTION

During these past few years, there has been a new and justified interest in statistical methods applied to model quantitative structure−activity relationship/quantitative structure−property relationship (QSAR/QSPR) studies.[1−5] Some of these studies focus on plot methods, which should be widely used in model studies, as exemplified by the discovery of some poor plots of QSAR studies which showed quite good statistical values.[6,7] Furthermore, two recent studies have shown that some plots are far from reliable.[8,9] The present article focuses on multilinear regression (MLR) plot methods, which in the literature continue to be treated rather superficially, a superficiality which seems to be present and practiced by many chemists and QSAR investigators. The plot problem is not only a QSAR/QSPR problem; misconceptions on the subject can also be detected in other chemical fields (see ref 8 and references therein), as the bulk of many experimental studies continue to be based on linear regressions and on their graphical representation. Normally, fitted (calculated) or predicted data are graphically depicted against experimental values, and very often these kinds of representations are accompanied by a "desired" or "ideal" line bisecting the first and third quadrants. It is believed that calculated values correlate with experimental data across this line and bear a symmetric distribution around it. In fact, the data do not necessarily have to be symmetrically placed around this bisector. At least for the least-squares case, the experimental versus fitted and the fitted versus experimental plots are not equivalent. Regarding these bidimensional representations, the graphs exhibit predictable behavior. For instance, in the fitted versus experimental values plot, the slope of the line adjusted by the least-squares technique is lesser than or equal to 1. Even more, it is exactly equal to r^2, the coefficient of determination of the original MLR fitting calculation. Visually, the aspect of the plot is compatible with a regression toward the mean effect.

In 1886, Sir Francis Galton published the results of an anthropological study establishing a relationship between the heights of parents and their respective grown children.[10] He noticed that the heights of children of parents who are both tall (or both short) "revert" or "regress" toward the mean of the group. That is, children of tall parents are usually tall, but not as tall as expected. Conversely, children of short parents are usually short, but not as short as expected. Globally, the mean statistical results are clear: if one depicts the children's heights against the heights of the corresponding parents, a regression line with a slope lesser than 1 is obtained. Initially, Galton referred to this effect as a "regression toward the mediocrity" and later renamed it "regression toward the mean". Together with the work of other scientists such as Adrien-Marie Legendre, Karl Friedrich Gauss, and Auguste Bravais, this view constituted the historical origin of the regression technique.[11]

The regression toward the mean effect accounts for systematic deviations which are well-known in pure and applied statistics and in other fields, especially in those involved with experimental research. For instance, in the medical sciences, psychology, and biostatistics,[12−27] the effect is taken into account because it reveals how some treatments are not as effective as believed a priori. Within this panorama, correction methods are sometimes proposed.[27,28] Other areas in which the effect is known include econometrics,[29−33] the social sciences,[34−36] and discussions on sport performance.[37−40]

* Corresponding author e-mail: emili.besalu@udg.es.

10.1021/ci6004959 CCC: $37.00
Published on Web 04/25/2007

This work comprises three parts. In the first one, a practical numerical example illustrates the behavior of data fitted by the ordinary MLR technique, leading to nonequivalent experimental versus fitted and fitted versus experimental plots. The second part illustrates how the problem can be presented phenomenologically and visually within the context of the regression toward the mean effect. Finally, it is demonstrated how a very common cross-validation technique, namely, the leave-one-out (LOO) procedure, leads to more widely spread data, as expected with lesser values of correlation, and how this is also connected to the regression toward the mean effects.

METHOD AND DISCUSSION

1. A Graphical Example. To underline the importance of the intimate relationship between statistical parameters and linear regression and contemporarily of the importance of a graphical representation of the regressions, let us go back to Anscombe's famous regression examples.[41] With these examples, Anscombe wanted to raise the issue of whether or not it is possible for a regression analysis to be misleading, especially if no visualization, in the form of plots, is offered of the results of the regression. Anscombe's examples share not only the same correlation but the values of any normally used statistics (number of observations, means of x and y, variance, and regression line, among others). Let the vector of the descriptors be

$$d_{A,B,C} = d_1 = (10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)$$

and the following vectors of the experimental actual properties be

$$y_A = (8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68)$$

$$y_B = (9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74)$$

$$y_C = (7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73)$$

Also, consider the following second vector of descriptors, which is the vector of the described experimental property, y_D,

$$d_D = d_2 = (8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8)$$

$$y_D = (6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89)$$

Anscombe's study is a bit changed here to conform more with the QSAR/QSPR problem: here, the y_I versus $y_{I,calcd}$ plots are completed with a plot of the residuals (Anscombe shows, instead, the y vs d_i plots), but changes are only minor at the plot level. The statistics of the four regressions are, surprisingly, exactly the same:

$$F = 18,\ r = 0.816,\ r^2 = 0.667,\ s = 1.2,\ N = 11\ \langle d_I \rangle = 9.0,\ \langle y_I \rangle = 7.5,\ \text{Regr. SS} = 27.5,\ \text{Res. SS} = 13.8\ (9\ \text{df})$$

The regression equation is the same for the four cases: $y_{A,B,C,D} = 3.00\ (\pm 1.12) + 0.50 d_{1,2}\ (\pm 0.12)$.

Now, let us look at the corresponding plots, shown in Figure 1. In these plots, the corresponding residuals (▲) are also shown. When these four plots (residuals included) are looked at, it is evident that the first picture is the plot visually imagined when a simple linear regression equation is reported. Here, residuals are really random, as confirmed by the linear regression (solid lower line), which is coincident with the abscissa axis. Nevertheless, the same statistics apply to the other three plots of Figure 1, all of which show either a curvilinear trend or a "strange" linear trend, due to outliers. The last plot, the y_D plot, is the most troublesome, as here a straight vertical line seems more appropriate. In fact, leaving out the strong outlier, which determines the regression entirely, we obtain a perfect vertical line which, among other poor statistics, has $F = 0$, $r^2 = 0$, $s = 1.2$, and $N = 10$. Here, the importance of the plot becomes evident, and this should have told us, from the start, that something is completely wrong here. Concerning the y_C (third) plot erasing the outlier, we obtain an optimal linear regression, with $F = 1\ 160\ 688$, $r^2 = 1$, $s = 0.003$, and $N = 10$. This tells us the importance of checking results with plots and is further strengthened by the recent discovery of some anomalous plots, which had not been shown in the respective QSAR/QSPR studies, that harbored very good statistics.[6,7]

A problem which is also related to the problem of wrong plots is the problem of the "reversed plot", that is, the y_{calcd} versus the y (experimental) plot, which is exemplified here by the two Anscombe plots shown in Figure 2. These are the reversed plots of the first two plots of Figure 1 and are also given here with their regression equations. The least that can be said about these two plots is that they are symmetric, which many researchers continue to accept uncritically. This topic can be thus rephrased: (*i*) the plot of experimental versus calculated values is centered around the bisectors of the first and third quadrants; (*ii*) the plot of calculated versus experimental values is a mirror image of the former with respect to the bisector which, in an ideal case, must act as the bidimensional data fitting line, and (*iii*) the residuals should always show Gaussian distribution with respect to the zero line. In the following, it will be stressed that assumptions *ii* and *iii* are not always fulfilled, as shown by our plots and by an accurate analysis of any plot, and as has already been demonstrated elsewhere.[8,9]

To check the importance of such a description, let us now consider the case of the introduction of a second descriptor to allow for a multilinear description of the given properties. The two second descriptors are

$$^2d_{A,B,C} = (8, 7, 15, 12, 16, 18, 3, 6, 12.5, 5.5, 3.2)$$

and

$$^2d_D = (7, 6, 5, 6, 7, 8, 6, 12, 6, 6, 6)$$

Can a second descriptor really improve the bad quality of the first descriptor? Leaving aside some subjective preferences, the answer of the plots, shown in Figure 3, is that it cannot! In fact, let us look at the main statistics of each

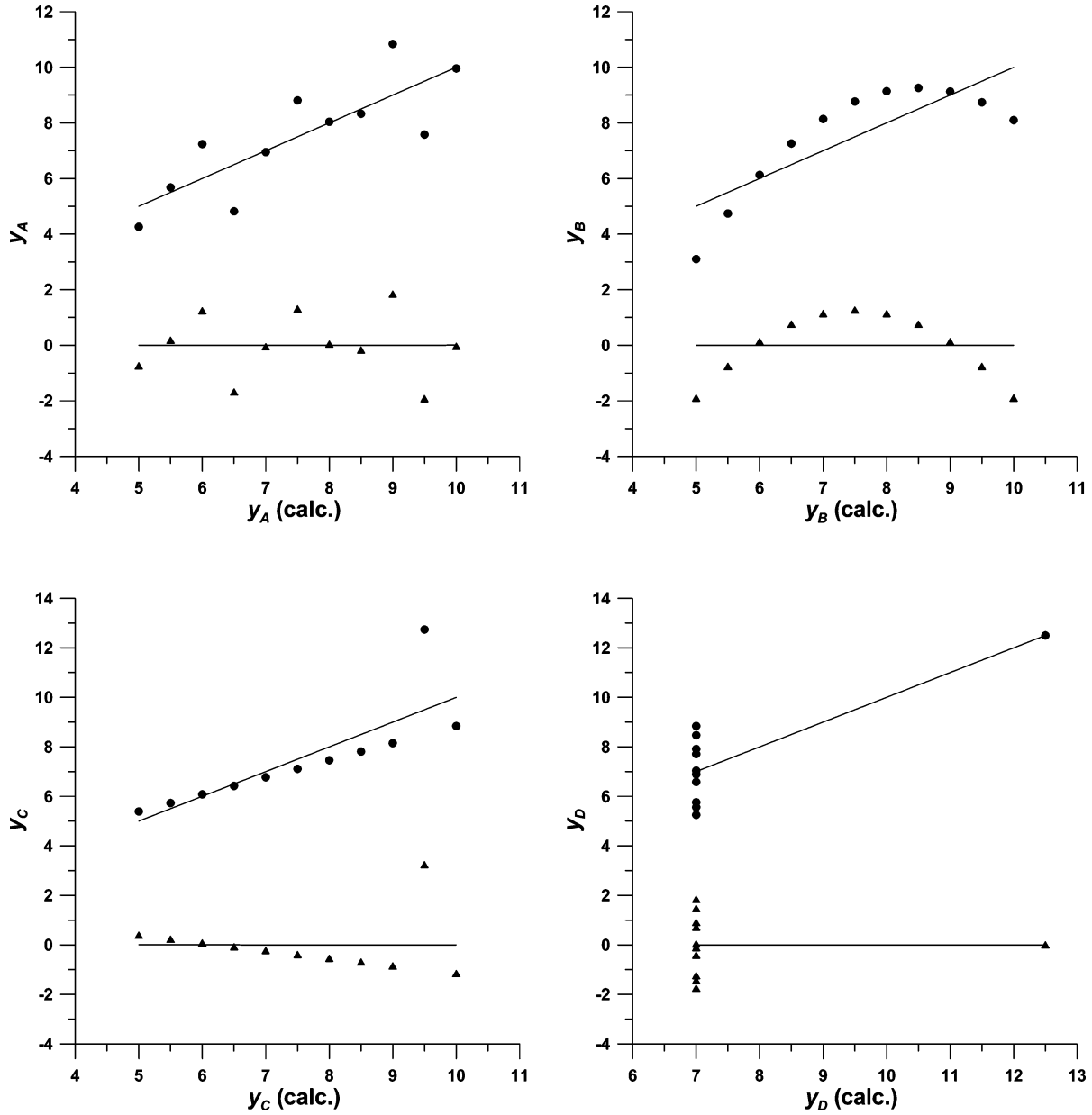

Figure 1. Plot of the experimental vs calculated properties, y_A/y_{Acalc}, y_B/y_{Bcalc}, y_C/y_{Ccalc}, and y_D/y_{Dcalc} (●), together with the plot of their residuals (▲).

description (u = |coefficient/its error| are the utilities of each coefficient and are indicated in parenthesis)

y_A: $F = 8.2$, $r = 0.819$, $r^2 = 0.671$, $s = 1.3$, $N =$ 11; regr.coeff. = [0.58(2), 0.06(0.3), 2.81(2.1)]

y_B: $F = 9.4$, $r = 0.838$, $r^2 = 0.702$, $s = 1.2$, $N =$ 11; regr.coeff. = [0.75(2.8), −0.17(1.0), 2.45(2.0)]

y_C: $F = 8.0$, $r = 0.817$, $r^2 = 0.668$, $s = 1.3$, $N =$ 11; regr.coeff. = [0.54(1.9), −0.03(0.17), 2.90(2.2)]

y_D: $F = 8.0$, $r = 0.817$, $r^2 = 0.667$, $s = 1.3$, $N =$ 11; regr.coeff. = [0.46(1.5), 0.07(0.13), 2.85(1.7)]

The consistent negative variation of F clearly shows that the new descriptor is "fishy" and should be thrown away, while the other statistics are practically constant. The only "poor" novelty is that now the statistics for the four cases are no longer completely equal to each other.

Residual plots have another advantage, which is independent of the type of representation: they can show nonlinearity and also the phenomenon called heteroscedasticity[6] (also heteroskedasticity); that is, residuals diverge for growing values of the property, a case many tests have been developed to cope with.[42] Residual analysis is, thus, a useful tool to detect deficiencies of the model as it is a direct way to detect patterns, nonrandomness, and, finally, heteroscedasticity. But heteroscedasticity is not the worst of a model's problems, and sometimes it can hardly be avoided. The origin of the word is: hetero + skedastikos (related to scattering). Heteroscedasticity means nonconstancy of the variance of a measure over the levels of the property under study, that is, the dependence of the variance on the value of the measurand. The opposite phenomenon is called homoscedasticity. An

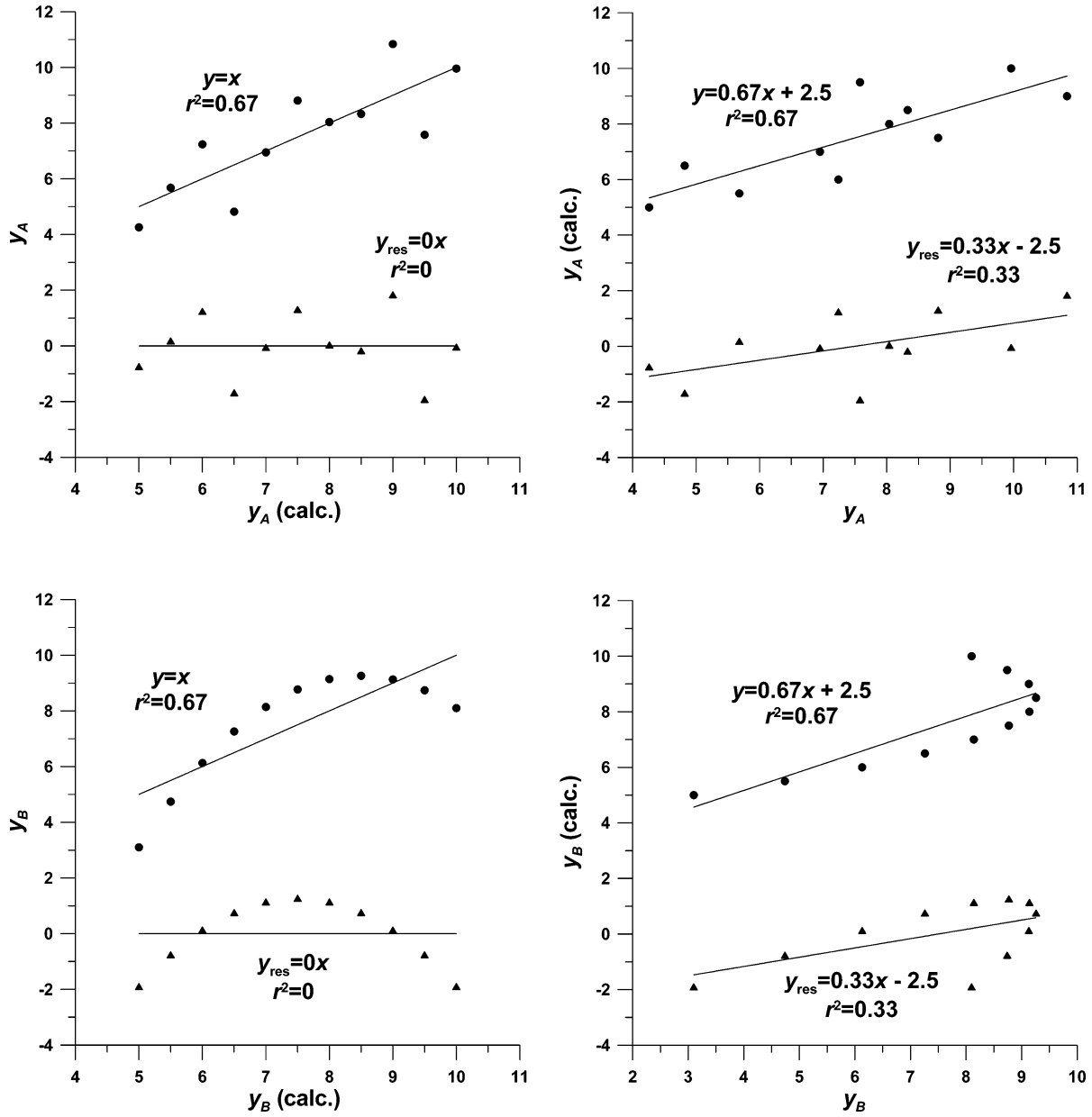

Figure 2. Original, y_A/y_{Acalc} and y_B/y_{Bcalc} (left-side), and reversed plots, y_{Acalc}/y_A, and y_{Bcalc}/y_B (right-side), together with their residuals (▲).

easy to remember the rule of thumb states that, statistically, residuals are better homo than hetero.

2. Regression toward the Mean Effects. In this section, a phenomenological presentation is given to illustrate, in a graphical way, the concept of regression toward the mean effect. Figure 4 shows a simulated case where 5000 training points have been depicted. Every point represents an item (a molecule, if the QSAR field is being considered), and for each one, four random descriptors have been generated following a normal distribution (mean zero and standard deviation unity). In order to demonstrate the general effects of regression toward the mean, we have generated the property value by adding a normally distributed random number to the sum of the four descriptors. For graphical presentation purposes, this random value which is added to construct the property was expanded by a factor of 1.6 in order to force the coefficient of determination among the data (fitted and actual) to be $r^2 = 0.60$. In this way, we reproduce an ideal situation in QSAR, that is, to have descriptors of equal importance and a property value, all normally distributed and related by a subjacent MLR model. These numerical manipulations are similar to those described by Wang and Wang.[43] The Box−Muller transformation algorithm[44] has been employed to generate random numbers, which follow a normal distribution. Finally, a shrinkage factor of 0.5 and a shift of five units have been applied to the obtained property in order to fit the points of Figure 4 inside a 10 × 10 square. All in all, the equation which generates the property from the four normally distributed random descriptor variables z_1-z_4 is

$$y_i = 5 + 0.5(1.6z_0 + \sum_{i=1}^{4} z_i)$$

where z_0 is another normally distributed random variable. The artificial property variable y_i has a mean of 5 and a standard deviation of about 1.28 units.

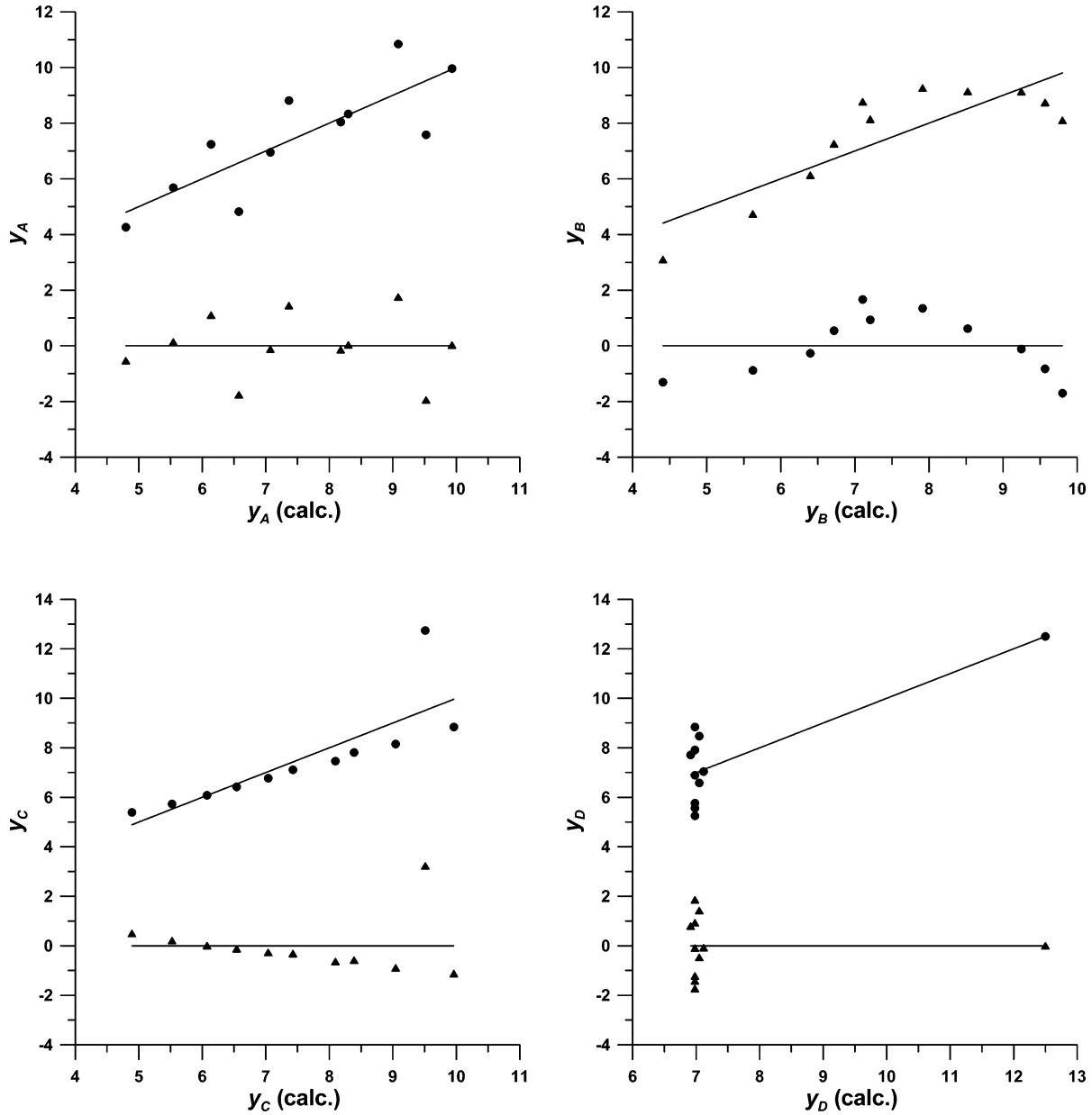

Figure 3. Plot of the experimental vs calculated properties, y_A/y_{Acalc}, y_B/y_{Bcalc}, y_C/y_{Ccalc}, and y_D/y_{Dcalc} (●), together with the plot of their residuals (▲). This time the calculated values have been obtained with two descriptors.

In order to depict Figure 4, the MLR model involving all 5000 cases has been adjusted and the fitted property values are depicted against the experimental ones. The representation in Figure 4 of a big number of cases helps to appreciate the density distribution of the points cloud and to see how this cloud is rotated with respect to the "ideal" bisector. Figure 4 shows how the density of the spread points also bears the structure of a bidimensional Gaussian distribution. The most remarkable feature is that the point cloud is slightly rotated with respect to the bisector of the first and third quadrants (solid line). Clearly, the points depicted in Figure 4 are not symmetrically spread (balanced) around this bisecting line, but rather around the line having the equation $y = 1.40 + 0.72x$ (not depicted). This line coincides with the first principal component of the plotted bidimensional data and also corresponds to the fitted line by the orthogonal distance method.[45] If the points shown in Figure 4 are fitted to a straight line by the usual least-squares technique, the obtained equation is (see dashed line in Figure 4)

$$y = 2.00 + 0.60x \quad (1)$$

The MLR coefficient of determination of the data is $r^2 = 0.60$ and coincides with the slope of the previous equation (eq 1). This is not casual, as a theorem demonstrates that both values must coincide exactly[8] (the same kind of coincidences can be seen in the right-most graphs in Figure 2). Moreover, the ordinate at the origin of eq 1 is not zero. The same theorem states that this value must be $a = (1 - r^2)\bar{y}$, which here is $a = (1 - 0.60)5.00 = 2.00$.

The higher the coefficient of determination, the more the cloud of points in Figure 4 will be symmetrically spread around the bisector and, simultaneously, the lower the degree of spreading. The limiting case is the perfect fit along the bisector, and the data having $r^2 = 1$ (and $a = 0$). This constitutes a simple example showing how the "desired" line

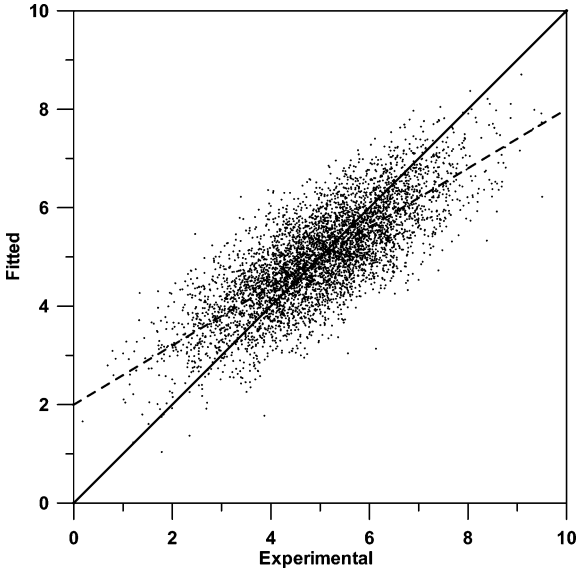

Figure 4. Representation of fitted values against the experimental property value. The data concern the simulation of 5000 molecules, with four descriptors each.

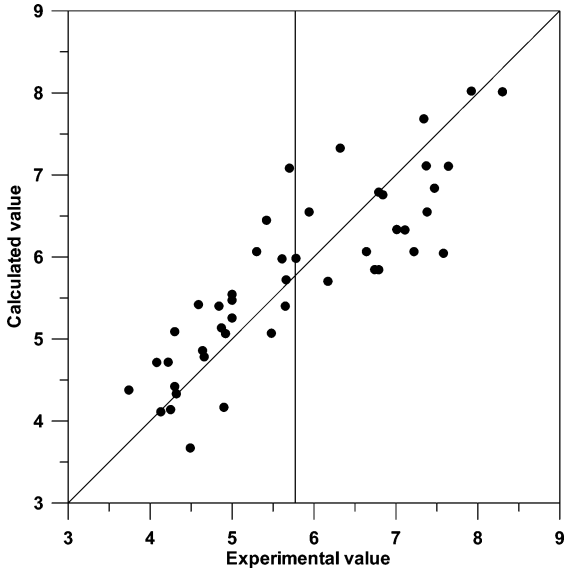

Figure 5. Reproduction of the fitted vs experimental plot from the data of Huuskonen's article[50] (activities of TIBO molecules expressed as minus the logarithm of the IC50 dose protection of MT-4 cell against the cytopathic effect of HIV-1). The vertical line signals the experimental mean value.

(bisector) is only attained for a perfect correlation and is not reproduced for intermediate situations. In general, the cloud is not symmetrically distributed along the ideal or desired line, as some authors claim.[1]

Additionally, and as a consequence of the previous situation, when reversing the representation axis in Figure 4, that is, when experimental values are plotted against the fitted ones, the slope of the one-dimensional least-squares fitting line will be exactly 1.[8] This juncture opens the possibility to make bidimensional representations of experimental versus fitted values where the depicted points always appear to be well-aligned around the "desired" line (see, for instance, some graphs depicted in refs 46−48, or the leftmost graphs in Figure 2). In these cases, the goodness of fit of the data has to be graphically inferred from the dispersion of the point cloud and not from the arrangement relative to the bisector, as this apparent "symmetrical" arrangement is justified (forced) by a theorem.

In order to illustrate the regression toward the mean effects in real cases, two bibliographic examples have been selected. The first one is a result due to Wang and co-workers[49] concerning the fit and prediction of binding p*K* values with models involving four parameters. Wang et al.'s Figure 4, in ref 49, shows, in the training fitted versus observed plot, how a 200-point cloud visually presents the effect of regression toward the mean. That is, a nonuniform cloud is observed spread along the bisector. This example is particularly clear and illustrative because the coefficient of determination is moderate ($r^2 = 0.59$) and mainly because quite a large number of points have been fitted, a situation which is not very common when constructing QSAR models. As a second example, it is here reproduced a figure obtained from the numerical data of ref 50. This result involves 46 molecules and shows a value of $r^2 = 0.74$ for the training set. The reported experimental data (activities of TIBO molecules expressed as minus the logarithm of the IC_{50} dose protection of the MT-4 cell against the cytopathic effect of HIV-1) and electrotopological descriptors allowed to reproduce the linear model and to obtain fitted values. Figure 5 represents the calculated (fitted) versus experimental values plot (an equivalent plot is depicted in ref 50), and the effect of the regression can also be seen. In the figure, the vertical line signals the experimental mean value (5.77 logarithmic units). The reader can check for a distinct tendency behavior on the left and on the right of this separator line.

The effect in the first example is more evident than in the second one, and this is due to two reasons mainly: (i) because the number of points in the training set differs considerably (200 against 46), leading to distinct point cloud density aspects, and (ii) because the r^2 values are also distinct and greater (less favorable for visualization purposes) in the second example. In general, the effect of regression in fitting can be reproduced for any set, being more and more evident as more points are available and even more if lower r^2 values are reproduced. Thus, in general, linear regression techniques yield models whose predictions systematically tend too high for weakly active structures and too low for highly active structures. The aspects we are revisiting here apply not only to ordinary MLR calculations but to other more sophisticated techniques based on models ultimately obtained by an ordinary MLR model. This includes very common methods such as, for instance, principal component regression,[51] or post-treatment methods for data obtained by comparative molecular field analysis,[52,53] comparative molecular similarity indices analysis,[54] or holographic QSPR[55] calculations, among others.

In this work, the phenomenon of regression toward the mean is being treated in a purely descriptive way and the underlying laws governing it are not explored. The global situation is neither clear nor apparent in some cases, and nowadays there are some epistemological currents debating Galton's approach or interpretation. For instance, Los[30] speaks in terms of "a monumental scientific error". In fact, a relaxed posture consists of identifying the phenomenon with its own explanation. In the medical or social fields outlined above, another source accounting for the effect deals with the particular study of and the characteristics attached to extreme data, especially those catalogued in this way a

posteriori. But this assumption is not always valid in QSAR approaches, as special deviations or ill-conditioned treatments cannot generally be blamed on extreme experimental values. Another class of general explanations for the lack of perfect correlation and the associated regression has to do with identifying casual factors relating the original data (trained points) and the regressed data (test predictions). Again, this can be justified in some cases and in some of the fields outlined above, but not in the QSAR field. The QSAR effects explored here are purely mathematical and related to the intrinsic nature of the multilinear model.

3. Regression toward the Mean Effects in MLR−LOO Cross-Validation. Regarding cross-validation procedures, the LOO approach is a widely used technique, especially when MLR models are being constructed, as cross-validated properties can be obtained very fast without explicitly reproducing all the LOO steps.[56] Here, it will be shown how MLR−LOO predictions present systematic deviations which magnify the regression toward the mean effects. The same general ideas which will be reviewed also apply to leave-many-out procedures, but this aspect will be not explored here.

As is well-known, values predicted by the MLR−LOO technique are obtained from the following equation:[56,57]

$$y_i' = \frac{h_{ii} y_i - \hat{y}_i}{h_{ii} - 1} \qquad i = 1, 2, ..., n. \tag{2}$$

where y_i are experimental values, $\hat{y}_i$ are the values adjusted by the overall MLR data fitting, and each h_{ii} term is a diagonal element of the so-called hat matrix, $\mathbf{H} = \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T$. In present notation, $\mathbf{D}$ is the matrix of descriptors including a fixed column of ones. From eq 2, it can be seen that the numerical differences between experimental, fitted, and cross-validated values are related:

$$\hat{y}_i - y_i = (1 - h_{ii})(y_i' - y_i) \tag{3}$$

On the other hand, h_{ii} terms are bounded:[4,57]

$$\frac{m}{n} \le h_{ii} \le 1$$

m being the number of descriptors and n the number of equations. This condition implies the following inequalities:

$$0 \le 1 - h_{ii} < 1 \tag{4}$$

because $n > m$. Then, as condition 4 involves a positive term, the differences $\hat{y}_i - y_i$ and $y_i' - y_i$ appearing in eq 3 bear the same sign, and additionally, the second difference is magnified, in absolute value, with respect to the first one. These restrictions admit a graphical interpretation, depicted in Figure 6, where filled circles represent MLR-fitted values (global calculation involving all the data) and the empty circles represent the corresponding LOO predictions. As a consequence of relation 3 and conditions 4, fitted points lying above the bisector (diagonal solid line) are attached to LOO cross-validated values which in turn are more overestimated (for these points: $y_i' > \hat{y}_i > y_i$). Conversely, fitted points originally underestimated and placed below the bisector will give even more underestimated LOO predictions once they are cross-validated (for these cases: $y_i' < \hat{y}_i < y_i$). Up to here, systematic deviations have been described, and as expected, the point cloud spreads even more because the new predicted values obtained by cross-validation will always be more overestimated or underestimated than the corresponding fitted ones. Then, the correlation coefficient diminishes with respect to the one which would be obtained when fitting all the data in a single calculation. Going further, and due to the regression effects reviewed above, it is assumed that a majority of fitted points lie above the bisector if the corresponding experimental value is lower than its mean (vertical solid line in Figure 6). This will generate many cross-validated points with a regression toward the mean effect (raising predicted values) on the left of the bidimensional graph. For the same reason, there is a tendency for a majority of fitted points having experimental values greater than the mean to be preferably placed below the bisector. Hence, the corresponding cross-validated values will be even lower, and regression toward the mean effect (this time, decreasing predicted values) is expected to also be found on the right part of the graph. These two kinds of tendencies are represented for a couple of cases in Figure 6 (see arrows). Then, when comparing cross-validated values by LOO against actual values, the global effect is an overall clockwise rotation of the cross-validated point cloud around the experimental mean value. In this way, the fitting line relating cross-validated and experimental values (point-dash line) exhibits a regression effect with respect to the training fitted line (dashed line), bearing an even lesser slope than the original r^2 value. This effect also contributes to decreasing the statistical q^2 value, as the data globally tend to regress toward the mean value.

Considering the linear regression of n points with a single independent explanatory variable, x, the diagonal elements in the hat matrix, also called "leverage" terms, are[4,57]

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

This expression constitutes a relevant clue because it establishes a relation between the numerical value of h_{ii} and the magnitude of the difference between the independent descriptor value and its mean: the greater the distance of a descriptor from its mean, the greater the leverage. In this way, one infers that extrapolated points from the experimental mean may be attached to a greater value of the term (eq 4), magnifying the scaling (eq 3). In other words, systematic overestimations or underestimations are magnified, and possibly, the effects of regression toward the mean may be greater for points lying far away from the mean trained value. In order to illustrate this, consideration has been given to a particular QSAR model (eq 4 and Figure 1 of ref 58) for which data is related to molecular binding affinities (ΔG measured in kilocalories per mole) and the linear model involves two independent descriptors. For this case, Figure 7 is the bidimensional plot where the differences between the values cross-validated by LOO and the adjusted ones ($y_i' - \hat{y}_i$ terms) are represented against the difference between the experimental property value and its mean ($y_i - \bar{y}$). As can be seen, there is a clear tendency reflecting that the difference between cross-validated and fitted values is

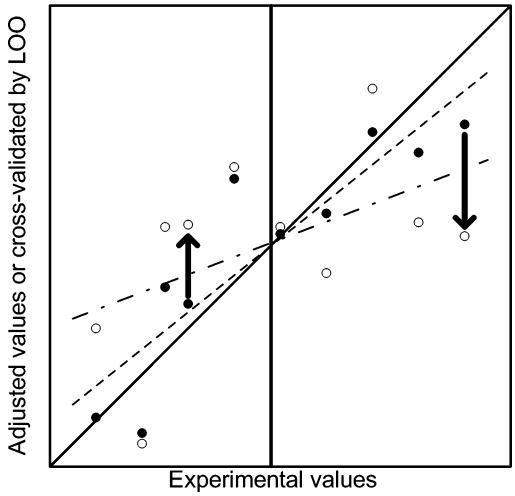

Figure 6. Comparison between the fitted training points in the overall MLR calculation (filled circles) and the corresponding LOO predictions (empty circles). The arrows signal a couple of shifts. The diagonal solid line is the bisector of the first and third quadrants. The dashed line fits the original data. The point-and-dash line fits the LOO cross-validated data.

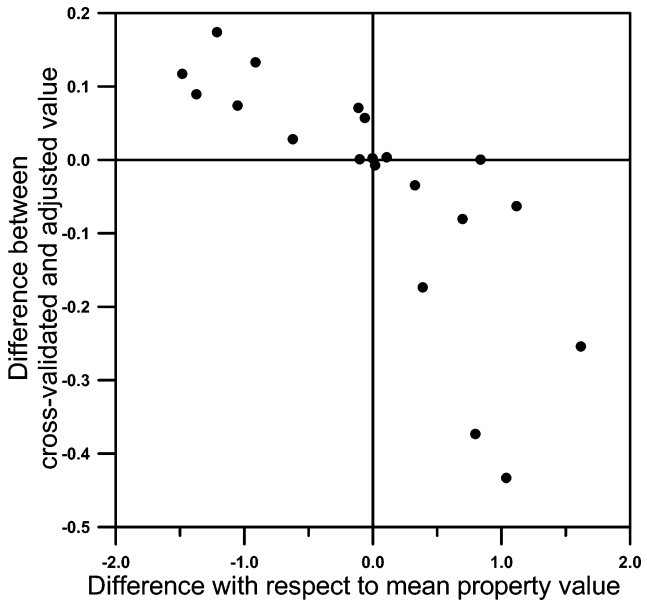

Figure 7. Illustrative case taken from the literature for which the difference between the cross-validated and the fitted values is clearly correlated with the proximity of the experimental value to its mean. The measured units for the ΔG values are kilocalories per mole.

correlated with the proximity of the experimental point to the mean value. This qualitative result has been found in other explored cases. The conclusion is that, for this kind of representation, very extreme points will presumably exhibit greater deviations from the corresponding fitted value. In many cases, the deviation will bear a direction according to the regression toward the mean effect.

Many of the above considerations are statistical and probabilistic. The phenomena will be more evident if medium or large sets are considered or if the results of many calculations (eventually involving a few cases each) are collected and processed with a unified approach. Unfortunately, only a reduced number of points is usually considered to depict calculated (adjusted, cross-validated, or predicted) versus experimental representations, making the visual distinction of regression effects hard to detect in the QSAR field. This could be the reason why these systematic deviations have not been evident to many researchers and why the statistical effect has not been previously considered in the QSAR literature.

Another relevant question concerns the molecular property ranking obtained from the calculations: was the regression effect able to swap some molecular positions in this ordering? In this case, correction techniques should be applied, but the task is not so straightforward. For instance, simple statistical tests and corrections such as those proposed by Mee and Chua[27] cannot be applied to the QSAR examples presented here as we are not dealing with repetitive measures, but with a more complex situation: in QSAR, we are comparing the performance found for a cross-validated (or external) set against the performance of trained data. In this field, training involves not only dealing with a population that is distinct from cross-validation (or external validation) but also knowing that the process of linear model building is essentially distinct from the simpler process of model application to a new external data set.

From our results, it can be inferred that, if the MLR−LOO calculations present an even stronger regression effect than in fitting, probably even stronger effects might, in some cases, be found in predictions obtained by linear or related to linear models (this can be visually checked in Figures 5, 3, 3, 2, and 5 of refs 49, 50, 59, 60, and 61, respectively, or in refs 62 and 63). In fact, present results justify that MLR−LOO calculations do not provide pure or true predictions, as there are internal laws (inherited from the MLR technique) which control the global behavior of the cross-validated results. In accordance with other authors,[64] we claim that cross-validation is not as useful as a pure external test validation. In any case, we do not refute cross-validation, but we are pointing toward an inherent characteristic of MLR−LOO cross-validation calculations: to be aware of "natural" regression toward the mean effects. The researchers must know about the advantages, disadvantages, and special characteristics of the numerical tools they are using.

CONCLUSIONS

The importance of using plots to inspect QSAR models has been stressed. It has been shown that regression toward the mean effects is present in QSAR when MLR techniques are considered for model building. The relation of this general effect to the aspect of fitted versus experimental scatter plots has been shown, demonstrating that, in general, the natural "ideal" or "desired" line in these plots is not the bisector of the first and third quadrants. On the contrary, a symmetric pattern is always reproduced in experimental versus fitted representations. The manifestation of the regression effect has also been mathematically justified regarding the MLR−LOO cross-validation technique. This evidences another aspect of the intrinsic nature of this particular cross-validation procedure in the context of linear modeling.

ACKNOWLEDGMENT

This work has been supported by grants number SAF2000-0223-C03-01, BQU2003-07420-C05-01, and CTQ2006-04410/BQU of the Ministerio de Ciencia y Tecnología within the Spanish Plan Nacional I+D. The comments of three

anonymous and attentive reviewers helped to improve the quality of the present manuscript.

REFERENCES AND NOTES

(1) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.
(2) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.; 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.
(3) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.
(4) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.
(5) Peterangelo, S. C.; Seybold, P. G. Synergistic Interactions among QSAR Descriptors. *Int. J. Quantum Chem.* **2004**, *96*, 1−9.
(6) Pogliani, L.; de Julián-Ortiz, J. V. Plot Methods in Quantitative Structure−Property Studies. *Chem. Phys. Lett.* **2004**, *393*, 327−330.
(7) Pogliani, L.; de Julián-Ortiz, J. V. Residual Plots and the Quality of a Model. *MATCH* **2005**, *53*, 175−180.
(8) Besalú, E.; de Julián-Ortiz, J.; Iglesias, M.; Pogliani, L. An Overlooked Property of Plot Methods. *J. Math. Chem.* **2006**, *39*, 475−484.
(9) Besalú, E.; de Julián-Ortiz, J.; Pogliani, L. Some Plots Are Not that Equivalent. *MATCH* **2006**, *55*, 281−286.
(10) Galton, F. Regression towards Mediocrity in Hereditary Stature. *J. Anthrop. Inst.* **1886**, *15*, 246−263.
(11) Denis, D. J. The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists? *Hist. Philos. Psychol. Bull.* **2001**, *13*, 36−44.
(12) Landow, L. Another Example of Regression to the Mean. *Aneseth. Analg. (Hagerstown, MD, U.S.)* **2002**, *94* (6), 1673−1673.
(13) Browne, S. M.; Halligan, P. W.; Wade, D. T.; Taggart, D. P. Cognitive Performance after Cardiac Operation: Implications of Regression toward the Mean. *J. Thorac. Cardiovasc. Surg.* **1999**, *117* (3), 481−485.
(14) Bland, J. M.; Altman, D. G. Regression towards the Mean. *BMJ [Br. Med. J.]* **1994**, *308*, 1499−1499.
(15) Bland, J. M.; Altman, D. G. Statistics Notes: Some Examples of Regression towards the Mean. *BMJ [Br. Med. J.]* **1994**, *309* (6957), 780−780.
(16) Fitzmaurice, G. Regression to the Mean. *Nutrition* **2000**, *16* (1), 81−82.
(17) Newell D.; Simpson, J. Regression to the Mean. *Med. J. Aust.* **1990**, *153* (3), 166−168.
(18) Erev, I.; Wallsten, T. S.; Budescu, D. V. Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes. *Psychol. Rev.* **1994**, *101*, 519−527.
(19) Nesselroade, J. R.; Stigler, S. M.; Baltes, P. B. Regression toward the Mean and the Study of Change. *Psychol. Bull.* **1980**, *88* (3), 622−637.
(20) Clarke, D. B.; Clarke, A. M.; Brown, R. I. Regression to the Mean − A Confused Concept. *Br. J. Psychol.* **1960**, *51*, 105−117.
(21) Furby, L. Interpreting Regression toward the Mean in Developmental Research. *Dev. Psychol.* **1973**, *8*, 172−179.
(22) Davis, C. E. The Effect of Regression to the Mean in Epidemiological and Clinical Studies. *Am. J. Epidemiol.* **1976**, *104*, 493−498.
(23) Egelberg, J. The Impact of Regression towards the Mean on Probing Changes in Studies on the Effect of Periodontal Therapy. *J. Clin. Periodontol.* **1989**, *16*, 120−123.
(24) Blomqvist, N. On the Bias Caused by Regression toward the Mean in Studying the Relation between Change and Initial Value. *J. Clin. Periodontol.* **1987**, *14*, 34−37.
(25) Denke, M. A.; Frantz, I. D. Response to a Cholesterol-Lowering Diet: Efficacy Is Greater in Hypercholesterolemic Subjects even after Adjustment for Regression to the Mean. *Am. J. Med.* **1993**, *94*, 626−631.
(26) Chuang-Stein, C.; Tong, D. M. The Impact and Implication of Regression to the Mean on the Design and Analysis of Medical Investigations. *Stat. Methods Med. Res.* **1997**, *6*, 115−128.
(27) Mee, R. W.; Chua, T. C. Regression toward the Mean and Paired Sample t Test. *Am. Stat.* **1991**, *45*, 39−41.
(28) Curnow, R. N. Correcting for Regression in Assessing the Response to Treatment in a Selected Population. *Stat. Med.* **1987**, *6*, 113−117.
(29) Koenker, R. Galton, Edgeworth, Frisch, and Prospects for Quantile Regression in Econometrics. *J. Econometrics* **2000**, *95* (2), 347−374.
(30) Los, C. A. Galton's Error and the Under-Representation of Systematic Risk. *J. Banking Finance* **1999**, *23* (12), 1793−1829.
(31) Friedmand, M. Do Old Fallacies Ever Die? *J. Econ. Lit.* **1992**, *30* (4), 2129−2132.
(32) Quah, D. Galton's Fallacy and Tests of the Convergence Hypothesis. *Scand. J. Econ.* **1993**, *95* (4), 427−443.
(33) Zimmerman, D. J. Regression towards Mediocrity in Economic States. *Am. Econ. Rev.* **1992**, *82* (3), 409−429.
(34) Davis, G. A. Accident Reduction Factors and Causal Inference in Traffic Safety Studies: A Review. *Accid. Anal. Prev.* **2000**, *32* (1), 95−109.
(35) Persaud, B. Relating the Effect of Safety Measures to Expected Number of Accidents. *Accid. Anal. Prev.* **1986**, *18* (1), 63−70.
(36) Anand, D.; Schinnar, A. P. Technical Issues in Measuring Scholastic Improvement due to Compensatory Education Programs. *Socio-Economic Planning Sci.* **1990**, *24* (2), 143−153.
(37) Audas, R.; Dobson, S.; Goddard, J. The impact of managerial change on team performance in professional sports. *J. Econ. Business* **2002**, *54* (6), 633−650.
(38) Smith, G. Do Statistics Test Scores Regress Toward the Mean? *Chance* **1997**, *10* (4), 42−45.
(39) Lee, M.; Smith, G. Regression to the Mean and Football Wagers *J. Behav. Dec. Making* **2002**, *15*, 329−342.
(40) Gamson, W. A.; Scotch, N. A. Scapegoating in Baseball. *Am. J. Sociol.* **1964**, *70*, 69−72.
(41) Anscombe, F. J. Graphs in Statistical Analysis. *Am. Stat.* **1973**, *27*, 17−21.
(42) Verbeek, M. *A Guide to Modern Econometrics*, 2nd ed.; John Wiley & Sons: Chichester, U. K., 2004.
(43) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1422−1426.
(44) Box, G. E. P.; Muller, M. E. A Note on the Generation of Random Normal Deviates. *Ann. Math. Stat.* **1958**, *29*, 610−611.
(45) Adcock, R. J. A Problem in Least Squares. *Analyst* **1878**, *5*, 53−54.
(46) Murcia-Soler, M.; Pérez-Giménez, F.; Nalda-Molina, R.; Salabert-Salvador, M. T.; García-March, F. J.; Cercós-del-Pozo, R. A.; Garrigues, T. M. QSAR Analysis of Hypoglycemic Agents Using the Topological Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1345−1354.
(47) Yin, S.; Shuai, Z.; Wang, Y. A Quantitative Structure−Property Relationship Study of the Glass Transition Temperature of OLED Materials. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 970−977. After inspection of the data in this article, one concludes that, in Figure 2, the "calculated" label must be "experimental" and "predicted" label must read "fitted" or "calculated".
(48) Peterangelo, S. C.; Seybold, P. G. Synergistic Interactions among QSAR Descriptors. *Int. J. Quantum Chem.* **2004**, *96*, 1−9.
(49) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(50) Huuskonen, J. QSAR Modeling with the Electrotopological State: TIBO Derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425−429.
(51) Esbensen, K. H. *Multivariate Data Analysis - In Practice*, 5th ed.; CAMO Process AS: Oslo, 2002.
(52) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110* (18), 5959−5967.
(53) Cramer, R. D.; Depriest, S. A.; Patterson, D. E.; Hecht, P. The Developing Practice of Comparative Molecular Field Analysis. In *3D QSAR in Drug Design*; Kubinyi, H. Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 443−467.
(54) Klebe, G. Comparative Molecular Similarity Indices: CoMSIA. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic Publishers: Great Britain, 1998; pp 3−87.
(55) Tong, W.; Lowis, D. R. Evaluation of Quantitative Structure−Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (4), 669−677.
(56) Besalú, E. Fast Computation of Cross-Validated Properties in Full Linear Leave-Many-Out Procedures. *J. Math. Chem.* **2001**, *29* (3), 191−204.
(57) Weisberg, S. *Applied Linear Regression*, 2nd ed.; John Wiley and Sons: New York, 1985.
(58) Kroeger, M. B.; Hose, B. M.; Hawkins, A.; Lipchock, J.; Farnsworth, D. W.; Rizzo, R. C.; Tirado-Rives, J.; Arnold, E.; Zhang, W.; Hughes, S. H.; Jorgensen, W. L.; Michejda, C. J.; Smith, R. H. Molecular Modeling Calculations of HIV-1 Reverse Transcriptase Nonnucleoside Inhibitors: Correlation of Binding Energy with Biological Activity for Novel 2-Aryl-Substituted Benzimidazole Analogues. *J. Med. Chem.* **2003**, *46*, 1940−1947.
(59) Prathipati, P.; Pandey, G.; Saxena, A. K. CoMFA and Docking Studies on Glycogen Phosphorylase a Inhibitors as Antidiabetic Agents. *J. Chem. Inf. Model.* **2005**, *45* (1), 136−145.

(60) Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries. *J. Med. Chem.* **2005**, *48* (16), 5154−5161.

(61) Boström, J.; Böhm, M.; Gundertofte, K.; Klebe, G. A 3D QSAR Study on a Set of Dopamine D_4 Receptor Antagonists. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 1020−1027.

(62) Ragno, R.; Marshall, G. R.; Di Santo, R.; Costi, R.; Massa, S.; Rompei, R.; Artico, M. Antimycobacterial Pyrroles: Synthesis, Anti-Mycobacterium Tuberculosis. Activity and QSAR Studies. *Bioorg. Med. Chem.* **2000**, *8*, 1423−1432.

(63) Kuo, C.-L.; Assefa, H.; Kamath, S.; Brzozowski, Z.; Slawinski, J.; Saczewski, F.; Buolamwini, J. K.; Neamati, N. Application of CoMFA and CoMSIA 3D-QSAR and Docking Studies in Optimization of Mercaptobenzenesulfonamides as HIV-1 Integrase Inhibitors. *J. Med. Chem.* **2004**, *47* (2), 385−399.

(64) Davies, A. M. C. Cross-Validation: Do We Love It too Much? *Spectrosc. Eur.* **1998**, *10*, 24−25.

CI6004959