Journal of
CHEMICAL INFORMATION
AND MODELING

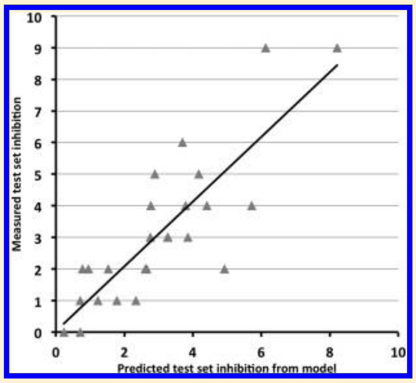
Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models

D. L. J. Alexander,[†] A. Tropsha,[‡] and David A. Winkler*[,§,∥,⊥,#]

[†]CSIRO Digital Productivity Flagship, Private Bag 10, Clayton South, VIC 3169, Australia
[‡]UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, United States
[§]CSIRO Manufacturing Flagship, Clayton, VIC 3168, Australia
[∥]Monash Institute of Pharmaceutical Sciences, Parkville, VIC 3052, Australia
[⊥]Latrobe Institute for Molecular Science, Bundoora, VIC 3046, Australia
[#]School of Chemical and Physical Sciences, Flinders University, Bedford Park, SA 5042, Australia

Ⓢ *Supporting Information*

ABSTRACT: The statistical metrics used to characterize the external predictivity of a model, i.e., how well it predicts the properties of an independent test set, have proliferated over the past decade. This paper clarifies some apparent confusion over the use of the coefficient of determination, R^2, as a measure of model fit and predictive power in QSAR and QSPR modeling. R^2 (or r^2) has been used in various contexts in the literature in conjunction with training and test data for both ordinary linear regression and regression through the origin as well as with linear and nonlinear regression models. We analyze the widely adopted model fit criteria suggested by Golbraikh and Tropsha (*J. Mol. Graphics Modell.* **2002**, *20*, 269−276) in a strict statistical manner. Shortcomings in these criteria are identified, and a clearer and simpler alternative method to characterize model predictivity is provided. The intent is not to repeat the well-documented arguments for model validation using test data but rather to guide the application of R^2 as a model fit statistic. Examples are used to illustrate both correct and incorrect uses of R^2. Reporting the root-mean-square error or equivalent measures of dispersion, which are typically of more practical importance than R^2, is also encouraged, and important challenges in addressing the needs of different categories of users such as computational chemists, experimental scientists, and regulatory decision support specialists are outlined.



1. INTRODUCTION

Although quantitative structure−activity/property relationship (QSAR/QSPR) modeling methods have been used for more than 50 years, there is still, surprisingly, considerable confusion about the best way to characterize the quality of such models. These types of models are typically generated using data-driven statistical or machine learning methods and aim to find a quantitative relationship, often quite complex, between the molecular properties (descriptors) of a series of molecules or materials and a target property such as aqueous solubility, toxicity, drug action, cell adhesion, etc.[1]

To make our terminology clear, we distinguish between the following (mutually exclusive) data sets:

- *Training set.* Data used to generate models. Ideally, a large training set will be available with a high degree of molecular diversity that spans a large range of the property being modeled.
- *Validation set.* Data used to estimate the prediction error in order to compare models. These data are not directly used to fit the models and thus give an independent measure of the model predictive power. However, since models are compared using the validation set, it affects the choice of model, particularly when one model is selected from a large number of candidate models.
- *Test set.* Data used to estimate prediction error of the final chosen model. These data are used neither to fit the models nor to select among them.

All models require training data. Where sufficient data are available, it is preferable to keep some aside as a test set in order to generate a demonstrably unbiased estimate of the magnitude of the prediction error. Ideally, such test sets should be truly independent and drawn from a different data source than the one providing the training data, e.g., by synthesizing molecules predicted by a model to have specific properties and testing them or using a "time-split" cross-validation as proposed recently by Sheridan.[2] However, this type of ideal test set is rarely encountered in practice, and partitioning of a data set into training and test sets (once) is a practical approximation to the ideal case. Additionally, some methods use all of the data in the training set (particularly when the amount of data available is insufficient to allow separate validation and test sets). The

Received: April 14, 2015
Published: June 22, 2015

DOI: 10.1021/acs.jcim.5b00206

two most commonly used methods to estimate the model predictivity in this case are cross-validation and bootstrapping.

Bootstrapping is a resampling method that involves sampling from the original training set of M data points with replacement to form a new set (a bootstrap sample) that is also of size M. QSAR models are generated from the bootstrap sample. This process is repeated a large number of times. As the bootstrap sample is taken from the original data set using sampling with replacement, it is not identical to the original data set. The distribution of statistics such as R^2 is then estimated by the distribution of these statistics in the bootstrap samples. Bagging (bootstrap aggregating) involves averaging of predicted values from a set of bootstrap models.

To apply cross-validation, N members of the training set of size M are removed, a model is generated from the remaining $M - N$ data points, and the properties of the omitted molecules or materials are predicted from the model. This is done M/N times so that each member of the training set is omitted once (in leave-one-out (LOO) cross-validation, the most commonly employed implementation, $N = 1$). The predictions of the properties of the omitted points $\hat{y}_{\mathrm{CV}}$ are used to generate statistics such as R^2.

Both bootstrapping and cross-validation tend to provide overly optimistic estimates of the predictive power of the model, as the data are typically not a truly random sample of molecules. The model may fit the training set data well, but whether it would accurately or precisely predict the properties of test set molecules external to the model generation process remains unproven.[3,4]

The use of an independent test set is considered the "gold standard" for assessing the predictive power of models and is the most stringent approach. The model may be fitted using bootstrapping or cross-validation on training set data, but its performance is measured by predictions of test set data external to the model generation process. The test set can be chosen randomly from the data set (best when the data set is large) or by some other method such as cluster analysis, which chooses test set points that are representative of the data set and property range and thus fall within the domain of applicability of the model (best for small data sets where the statistical fluctuations inherent in random selection are much larger).[5−7] When clustering is used, it can also be argued that this test set is not completely independent of the model generation process, but even so, the test set compounds and their observed property values are not used for either model development or model selection.

Regardless of how the model was fitted, a fixture of QSAR and QSPR model validation is a graph of observed versus predicted target property values derived from the training, cross-validation, and/or test sets.[8] A standard measure of model quality is the coefficient of determination,[9] R^2, but the meaning and appropriateness of this statistic depends on the context. Does the graph refer to training or test data (the model being formulated with reference only to the training data)? Is the best-fit line constrained to pass through the origin? Should the predicted or measured variable be assigned to the vertical axis? Can R^2 ever be negative? Confusion or inconsistency in how R^2 is derived in the various scenarios appears to be fairly common, leading to errors and possible misrepresentation of the predictive strength of a model. This has led to a recent escalation in the number of metrics used to describe the quality of prediction of the test set, adding to the confusion, especially for those relatively new to the QSAR field. For example, Golbraikh and Tropsha[3] define R_o^2 and $R_o'^2$, which represent the quality of test set prediction through the origin, the two values depending on whether the predicted values are plotted on the ordinate or abscissa. More recently, Roy et al.[10] attempted to overcome this ambiguity in R_o^2 values by introducing a number of "r_m" parameters, including r_m^2, $r_m'^2$, r_m^2, and Δr_m^2. This approach has been challenged by Shayanfar and Shayanfar,[11] who further introduced a different r_o^2 value (also see the comments by Roy et al.[12]).

This paper details in practical terms the meaning and usage of R^2 in the various contexts mentioned above and suggests how to use it appropriately as a measure of model fit. In spirit, this paper is a sequel to the well-known study by Golbraikh and Tropsha[3] that identified the *inadequacy* of the LOO cross-validation R^2 (denoted as q^2 in this case) calculated on training set data as a reliable characteristic of the model predictivity. The properties of q^2 were examined, and more rigorous criteria for evaluating the predictive power of QSAR models were proposed. Herein these previously suggested criteria are examined and improved as applied to the external predictivity of QSAR/QSPR models for independent test sets. It should be noted that this paper does not repeat the well-documented arguments in favor of using a test data set;[3,13] instead, we focus on how to properly use R^2 as a measure of how well a model can predict the properties of new data.

The following sections define R^2 both in general and in the specific case of plots of observed and predicted values in QSAR or QSPR modeling. Specific comments on regression through the origin are then followed by an analysis of the widely used model fit criteria of Golbraikh and Tropsha.[3] Particular discussion of the distinct needs of the R^2 metric for optimization purposes, as opposed to pure prediction, precedes the concluding section.

2. WHAT IS R^2?

R^2 is usually defined as the square of the correlation coefficient (also called Pearson's r) between the observed and predicted values in a regression (and for this reason is the same whether the observed values are regressed on predictions or the other way around). However, it may not be widely known that other definitions of R^2 are used, the most common of which agrees with this formula *only* if the predicted values come from an ordinary regression (or similar method) *on the observed values.* Importantly, this is not the case when applying a model to test data or even for training data in the case of some nonlinear models.[14] Sections 5 and 6 demonstrate this point by examples. Care is thus required in choosing an appropriate definition of R^2.

Kvalseth[14] reviewed no fewer than eight such definitions, recommending the following simple and informative formula:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \tag{1}$$

where y is the observed response variable, $\bar{y}$ is its mean, and $\hat{y}$ is the corresponding predicted value. This definition applies to any prediction method, including linear regression, neural nets, etc. Equation 1 therefore measures the size of the residuals from the model compared with the size of the residuals for a null model where all of predictions are the same, i.e., the mean value $\bar{y}$.

The numerator of the fraction in eq 1 is the sum of squared residuals (SSR). The importance of this term is sometimes

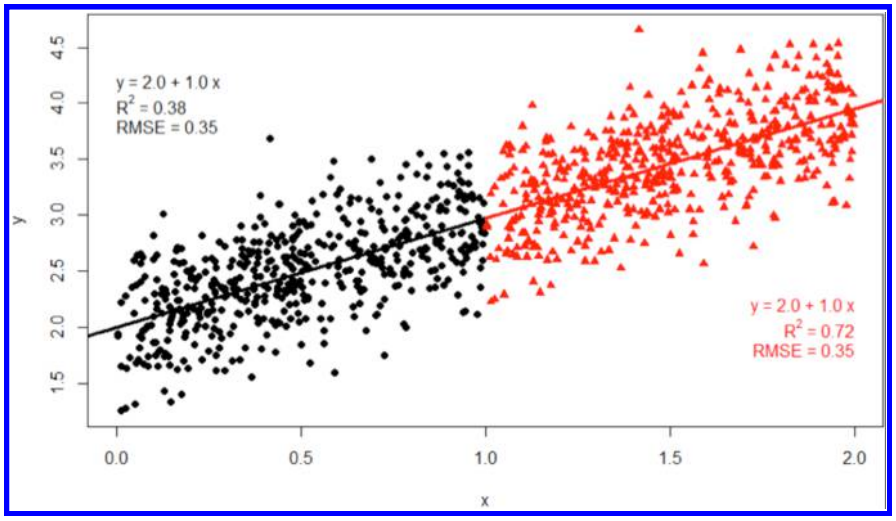


Figure 1. For the black circle data alone, R^2 = 0.38, while for the combined black circle and red triangle data the regression line is the same but R^2 = 0.72. The RMSE is the same in each case, as the red and black residuals are identical, but increasing the range of activity values in the data increases R^2.

forgotten, but it is at the heart of the meaning of R^2: **good models give small residuals**. Squaring the residuals before summing ensures that positive and negative residuals are counted equally rather than canceling each other out. (Less common alternatives to R^2 use the median instead of the sum[14,15] or absolute values of the residuals instead of their squares.[9,14]) For a good model, the SSR is low. High R^2 values are thus preferable; a perfect model has R^2 = 1.

Maximizing R^2 for a particular data set is equivalent to minimizing SSR. Ordinary linear regression finds the best *linear* model by this criterion (hence its common name "least-squares"), but alternative methods may help when variable selection is necessary (e.g., if there are more variables than observations, so not all of them can be used), as is commonly the case in QSAR, or if the relationship between the response and predictor variables is nonlinear, which is also common.

The average squared residual, or mean square error (MSE), which is obtained by dividing the SSR by the number of observations n, is a meaningful measure of model fit. Its square root, the root-mean-square error (RMSE), gives a standard deviation of the residuals. (For the training set, the SSR should be divided by $n - p$, where p is the number of parameters in the model, to obtain an unbiased estimate of the variance of the residuals.[9] Calculating p is straightforward for regression models but less so for more complicated models such as neural nets, etc.[4]) The RMSE, or an estimate of the standard deviation of residuals from the model, should usually be reported. For example, whether a method predicts melting points with a standard deviation of 0.1 or 100 °C is often more relevant to potential users than various other statistics of model fit. An approximate 95% confidence interval for predicting future data is $\hat{y} \pm 2 \cdot \text{RMSE}$ if the model is correct and errors are normally distributed. The use of measures of dispersion such as RMSE has been strongly advocated by Burden and Winkler.[7,16]

R^2 is a measure of how well the model fits *a particular data set*; the fraction in eq 1 compares the variation in the residuals to that of the observations themselves. Again referring to the example of predicting melting points of compounds, achieving an RMSE of 1 °C requires a much better model if the observed data span a range of hundreds of degrees than if they cover only a few degrees. A common verbal definition of R^2 is thus the proportion of variation in the response that can be explained by the model. However, it should be noted that the value of a model is generally in its overall accuracy and precision and not how successfully it explains the variation in a particular data set. RMSE, or an equivalent measure of dispersion, is often a more helpful indicator of a model's usefulness than is R^2. On the other hand, RMSE only indirectly indicates the success of the data modeling process itself, a characteristic directly measured by R^2; the two metrics thus fulfill distinct roles.

Equation 1 shows that if it were possible to augment the data set so that the observed values have greater variation while maintaining the same model accuracy, then R^2 would increase because $\sum(y - \overline{y})^2$ would be larger. However, such a procedure would improve neither the RMSE nor the practical usefulness of the model at any point in the range. Figure 1 illustrates this point: increasing the range of the data but maintaining an identical distribution of residuals causes R^2 to increase.

It is instructive to view eq 1 from another perspective also: the denominator is the sum of squared residuals for *a model that ignores all of the predictor variables* (the "model" that minimizes the SSR in this case is simply the average response). The denominator thus acts as a scaling factor relating the SSR to the overall variation in the observed data. Ordinary regression applied to a training data set can do no worse than the model $\hat{y} = \overline{y}$, so eq 1 implies that $R^2 \geq 0$ in this case. However, if the model is applied to data for which the response values y were unknown when fitting the model (as is the case in predicting observed activity values of the test set data) or if a method other than ordinary regression is used, eq 1 can sometimes give negative values. This is often highly confusing for novices struggling with a notion that a squared value of a parameter is negative. However, in this case the interpretation is simply that the model fit is *so poor* that the ratio in the right-hand side of the formula achieves values exceeding 1!

3. ASSESSING A MODEL

In QSAR and QSPR studies, the aim is to generate the model that gives the best predictions of a property (the dependent variable) on the basis of other properties of molecules or materials (i.e., their descriptors) in the training set. The quality of a model is assessed by a plot of the observed versus predicted dependent variable property values. This can be done for a training set, where it illustrates how well the model predicts the data used to generate it, or for a test set that contains data never used to train the model. The accuracy of prediction of the dependent variable property value for the test set data is a measure of the ability of the model to generalize to new data it has not seen before. The closer the data in such a plot lie to the line $y = \hat{y}$, the better the model is, as the predicted numerical values are very close to those measured by experiment. Including this line on the graph helps the predictive power of the model to be assessed (and the graph also provides a check for outliers or trends in the data).

Equation 1 provides the formula for R^2 as a measure of model fit. It should be noted that in a plot of test data, y, $\hat{y}$, and $\bar{y}$ in eq 1 should all relate to test data, not training data. Some authors[13] have recommended using y and $\hat{y}$ from the test set but $\bar{y}$ from the training set; however using $\bar{y}$ from the test set is not only simpler and more consistent but also minimizes R^2 and thus is more conservative. This question does not arise in calculating RMSE, another advantage of this metric.

The value of R^2 from a regression on the observed and predicted values themselves (rather than that calculated directly using eq 1 with the residuals $y - \hat{y}$ from the original model) is sometimes reported for test set data. However, this provides a measure of not the absolute degree of accuracy of the model but rather the degree to which its predictions are *correlated* with the observations. Such a correlation should be reported separately, as discussed in section 6 below.

When *training set* observations are regressed on their predicted values obtained by regression or a similar method, the fitted model is simply $y = \hat{y}$ (as $\hat{y}$ is already the linear combination of predictors that minimizes the SSR) and R^2 is the same as it was for the original model; thus a separate regression is unnecessary. The same will usually hold at least approximately for other (e.g., nonlinear) prediction methods.

However, this is *not* the case for test data. The regression of observed versus predicted values in this case will have a value of R^2 that is *larger* than that of the original model. The original model based on the training set data can estimate each test set observation y by a predicted value $\hat{y}$, but the linear regression of observed on predicted values maximizes R^2 for a *secondary* model, $y = a + b\hat{y}$. The fitted model will not be $y = \hat{y}$ in this case, since the test set is not identical to the training set; thus, the secondary model will give a larger value of R^2. A large difference between the two regression R^2 values would indicate a systematic error in the model, since it does not correctly predict the test data with good precision (although it may predict the trend correctly). The larger value of R^2 is not a true test set estimate of the model fit since it comes from a regression model using test set data, while the definition of test set data is that they are not used to fit a model. Instead, the test set value of R^2 must be calculated directly from eq 1.

Finally and for completeness, regressing predicted values on the actual observations has also been suggested.[3] Using observations in this way to predict model estimates, rather than the other way round, is counterintuitive. Besalu et al.[8] showed that for training data of a regression model, the regression line will be $\hat{y} = \bar{y} + R^2(y - \bar{y})$ (though the reason for this is not regression to the mean as claimed). The slope in such a graph is thus (surprisingly) R^2, which may be well less than the ideal of 1, and the intercept is $\bar{y}(1 - R^2)$, which may be far from the ideal of zero if $\bar{y}$ is large—even for relatively good models with R^2 (for example) around 0.8. The same features will appear on graphs of test data, provided that the regression model is unbiased. Again, this will usually hold at least approximately for prediction methods other than regression, provided that the average predicted observation is close to the actual average observation. It is thus unnecessary to perform such a regression and unreasonable to expect it to give a slope of 1 or an intercept of zero.

Some further comments on regression through the origin, recommended by Golbraikh and Tropsha,[3] are relevant before improvements to their criteria for a good model are suggested.

4. REGRESSION THROUGH THE ORIGIN

All of the regression equations discussed to this point have had no constraints placed on them. However, it is also possible to constrain regression equations so that the line passes through the origin. Adding this constraint invalidates the observations of section 2; for example, the values given by eq 1 would differ depending on whether observed values are regressed on predictions or the other way around.

Regression through the origin can also lead to negative values of R^2 in eq 1 even for the training data, since the added constraint may make the model even worse than simply estimating each observation by the sample mean. For this reason, a different formula, applicable only to regression through the origin, is used (without warning!) by various software packages,[11,12] including R and at least some modern versions of Microsoft Excel:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - 0)^2} \tag{2}$$

In eq 2, the SSR is again compared to the residuals from a model that ignores all of the predictor variables. The only such model that passes through the origin estimates each observation by the value zero. With this definition, R^2 values for linear regression on training data are again non-negative. Equation 2 gives higher values of R^2 than eq 1 (much higher when the mean observation is high), so values from the different equations are not comparable. R^2 values from eq 2 are also commonly close to 1 even when the model is very poor, greatly reducing their utility in distinguishing good models from less-good ones.[13]

5. CRITERIA FOR A GOOD MODEL FOR PREDICTIVE PURPOSES

Kubinyi et al.[17] and Golbraikh and Tropsha[3] clearly demonstrated that q^2 values on training data often bear no relation to the predictive ability of models on test data (where q^2 is defined as the value of eq 1 obtained using the training data set, where the predicted values are $\hat{y}_{CV}$ from LOO cross-validation[13]). The advice to assess model fit and predictive ability using test data is extremely pertinent.

However, the measures of model fit for test data suggested in the latter paper[3] are not strictly statistically correct. The application of F ratios is invalid[9] since it assumes that there is only one predictor, which is unlikely to hold, and there are other errors of interpretation as well (see the Supporting Information).

The main criteria for validity of a model as assessed by the test set predictions suggested by Golbraikh and Tropsha[3] and endorsed by Tropsha, Gramatica, and Gombar[13] relate to regression through the origin of observed and predicted values. These criteria are the following:

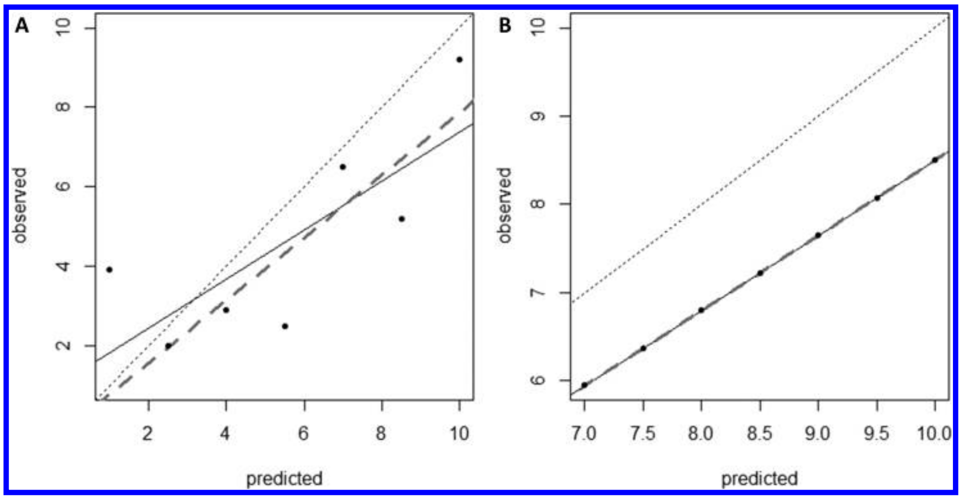


Figure 2. Illustrative plots of observed and predicted data that pass the Golbraikh and Tropsha criteria but are poor models. The dotted line shows the relationship $y = \hat{y}$; data points for good models would lie close to this line. Solid lines for ordinary regression and dashed lines for regression through the origin have also been added, although plotting of such lines on the graphs is not advocated here. (A) Ordinary regression and regression through the origin for a poor model that meets the criteria. (B) Ordinary regression and regression through the origin for an extreme case that predicts the order of the test set perfectly but none of the numerical values and also meets the criteria.

1. $q^2 > 0.5$ on training data
2. in regressing observed on predicted test data (or vice versa):
   a. $R^2 > 0.6$
   b. R^2 through the origin close to the unconstrained R^2
   c. slope of the test set regression through the origin close to 1

Criterion 1 is a valid condition for training data where cross-validation has been applied, though as Golbraikh and Tropsha[3] demonstrate, this gives little indication of the model's predictive power on test data. Criterion 2a, which applies to test data, ensures that no good model capturing the correlation between the actual and predicted data irrespective of the absolute error will be rejected; good models have small residuals, and thus, the graph of observed versus predicted values will have data points clustered around the line $y = \hat{y}$. Yet even a model with low R^2 may still be practically useful; again, the practical usefulness of a model is better determined by its RMSE than by the value of R^2.

As discussed in section 3, it is more meaningful to report simply both the RMSE and the value of R^2 calculated with the residuals $y - \hat{y}$ from the original model and additionally the squared test set correlation between observed and predicted values, as discussed in section 6. Calculating regressions through the origin is unnecessary.

Furthermore, the criteria of Golbraikh and Tropsha[3] identify as good some models that actually have poor fit, as the following examples demonstrate. It should also be noted that the definitions of R^2 through the origin in Golbraikh and Tropsha[3] are incorrect: not only do they use eq 1 rather than eq 2 (which is merely a matter of convention), but also, some of their equations contain typographical errors (see the Supporting Information for details). Corrected formulas are applied in the following discussion.

Figure 2 shows two hypothetical plots of observed and predicted test data values that pass the above criteria despite representing poor model fits. In Figure 2A, the regression of observed on predicted data has an R^2 of 0.60, and regressing predicted on observed data through the origin gives an R^2 of 0.58 and a slope of 1.1499, satisfying the conditions above. (Criterion 1 relates to training data, not shown here.) However, as described in section 2, the proper measures of fit are RMSE = 2.1 and $R^2 = 0.22$ calculated from eq 1. The data are far from the dotted line representing $y = \hat{y}$, and the model fit is in fact poor.

Figure 2B shows a more extreme example, in which the ordinary regression and regression through the origin are identical with $R^2 = 1$ and a slope of 0.85. The criteria are thus again satisfied, even though the data are far from the dotted line representing $y = \hat{y}$. In this case the RMSE is 1.3, which is very high given that the entire observed range is only 2.5. The correct R^2 value, −1.28, is negative in this case since the residuals are larger than the variation of the observations around their mean. This again indicates that the model fits extremely poorly, despite passing the criteria of Golbraikh and Tropsha.[3] (Reducing the range of the predicted values but maintaining the form of this graph, with predicted values 85% of the observed values, would decrease the value of R^2 (correctly calculated) even further.)

These examples amply demonstrate a weakness in the criteria described by Golbraikh and Tropsha.[3] There is no value in calculating R^2 for regression through the origin of observed on predicted values or vice versa; instead, the RMSE and R^2 calculated from eq 1 using $\hat{y}$ from the original model are of much greater practical use. Our recommended criteria for a useful model for predictive purposes are thus simply the following:

- High R^2, e.g., $R^2 > 0.6$, calculated for test set data using eq 1. This ensures that the model fits the data well.
- Low RMSE of test set predictions. How low the RMSE must be depends on the intended use of the model. Even a model with low R^2 can be practically useful if the RMSE is low.

These criteria are simpler and clearer than those proposed above and more recently by Tropsha,[18] because of this sharper understanding of how to calculate the statistics for the test set predictions.

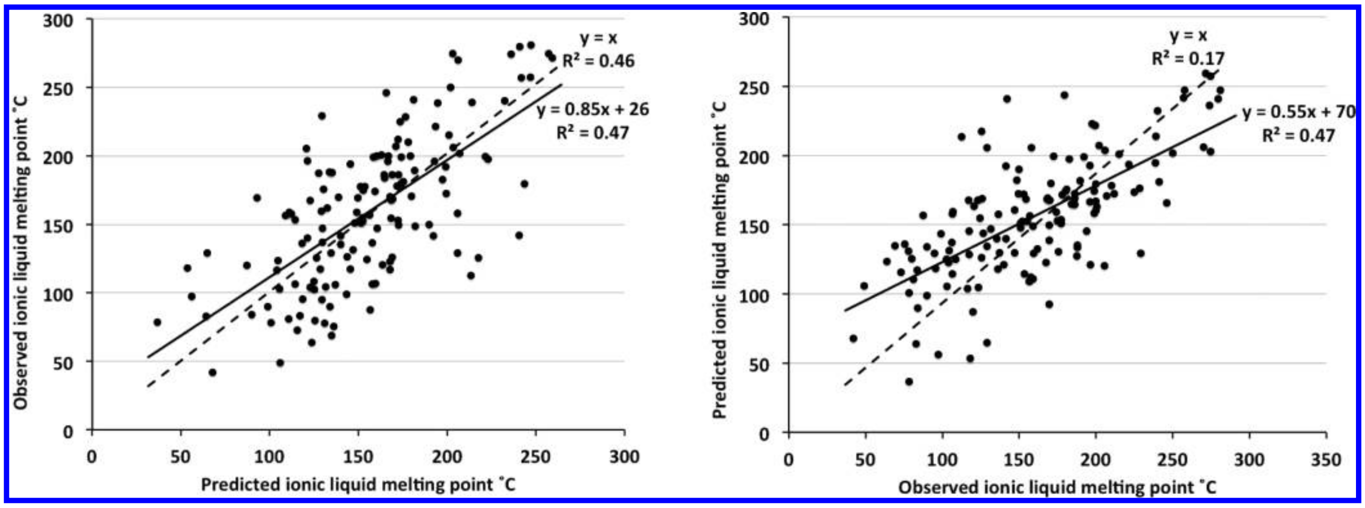


Figure 3. Test set predictions (143 compounds in test set) from a model predicting the melting point of 771 ionic liquids (ionic liquid data from Varnek et al.[20]) A regression line has been drawn through the data (solid line). The dotted line is for y=x and the regression values relate to fitting points to each line. The two graphs represent the plotting of the same test set data but with reversal of the assignment of the observed melting points to the axes.

6. CRITERIA FOR A GOOD MODEL FOR RANKING MOLECULES

The preceding sections establish improved criteria for good models *for predictive purposes*. For these purposes, we recommend plotting the predicted and observed values for the test set but calculating R^2 directly via eq 1 rather than from a line of best fit on this graph. Such a line gives the squared test set correlation between observed and predicted values rather than a test set measure of model fit.

However, the best fit correlation between observed and predicted values in the test set is still important in some situations. As most QSAR/QSPR practitioners are aware, the quantity and diversity of the training data are often relatively low because of the cost and difficulty in generating larger data sets or a limitation on the number of compounds that could show appreciable activity in a biological assay. For this or other reasons, the model may not fully capture the relationship between the molecular properties (descriptors) and the dependent variable property of interest. In this rather more realistic scenario, the numerical values of the target properties of the test set data may not be predicted as accurately, but the model may still allow correct identification of *trends*. It is still of great value to know which molecules or materials are the "best" (highest activity against a drug target, highest water solubility, lowest toxicity, lowest pathogen adhesion, etc.) and which are likely to have poor properties. This allows molecules or materials to be prioritized for synthesis. In this case, the trend-capturing *correlation* between the observed and predicted values for the test set may be more relevant than either the RMSE or accurate numerical prediction of the property by the model.

Preferably, a regression line for the observations fitted to their predicted values should have a slope close to 1 ($y = \hat{y}$) and may pass close to the origin, as would be the case for an ideal model. However, this is not essential when the relative ranking of molecules or materials in the test set is more important than numerically accurate values for the property being modeled. The value of R^2 that this secondary regression model generates is the square of the correlation coefficient between the observed and predicted values of the test set; to avoid confusion with the value of R^2 for the original model, we suggest this statistic be described explicitly as the squared test set correlation between observed and predicted values. This is simple and unambiguous. This squared test set correlation need only be reported if it is significantly higher than the value of R^2 calculated from eq 1 and when relative ranking of molecules or materials suffices rather than accurate prediction of the properties of each molecule or material.

Other measures may be even more appropriate in this case. Both the observed and predicted values may be ranked from 1 to n, where n is the number of data points; the correlation between these rankings, known as Spearman's rank correlation coefficient,[19] assesses how well the predictions *order* the observed values, regardless of how accurate the predictions are. (Squaring the Spearman rank correlation would make it more comparable to an R^2 value.)

These measures are now illustrated by example. Returning to Figure 2a, R^2 is only 0.22; the observed values are far from their predictions. However, if relative rankings of the data are more important than accurate predictions, the squared test set correlation between observed and predicted values, equal to 0.6, may also be mentioned; the model did at least order the test data fairly accurately, which may be all that is required. Spearman's rank correlation coefficient (0.71) is also much higher, indicating the model predictions were ordered approximately correctly even though the absolute prediction accuracy was poor.

In Figure 2b, the test set observations were so far from their predicted values that R^2 was negative. Nevertheless, the squared test set correlation between observed and predicted values is equal to 1: the model ordered the data perfectly, which may suffice in some applications. Spearman's rank correlation coefficient was also 1 in this case.

A further example is provided by the melting points of ionic liquids. The data taken from Varnek et al.[20] relate to the melting points for four chemical classes of ionic liquids. These data were modeled by Varnek et al. as separate chemical classes, but we have used molecular descriptors and a Bayesian regularized neural network to generate a model that predicts the melting points for all of the ionic liquids taken as a single group. The purpose is not to describe this particular model but to show how the metrics for model validation apply to a realworld problem. The QSPR model resulting from the analysis of these data is given in Figure 3. As explained in sections 2 and 4, the regression through the data gives identical R^2 values regardless of the assignment of the observed values to the x or y axis, whereas the regression of $x = y$ (corresponding to numerically accurate predictions, not just ranking) is very dependent on the choice of axes. When observed values are the ordinate, the R^2 value is 0.46, but when the axes are switched the R^2 value falls to 0.17.

The model tended to overestimate low melting points and underestimate high melting points in the test set, which is reflected in a fairly low value of $R^2 = 0.46$ as calculated from eq 1. This may suggest modifications to the model. The Spearman rank correlation for the predictions of the test set was 0.66. Squaring this correlation gives 0.44, which is comparable to the squared test set correlation of 0.47 between observed and predicted melting points. If these figures had been much higher than the value of eq 1, it would suggest that the model may still be effective in finding which ionic liquids have the lowest or the highest melting points, even though exact numerical estimates may not be accurate.[21]

When the purpose of a model is to find the best molecules or materials, not necessarily to predict their properties, the utility of a model is not solely dependent on quantitative accuracy.[20] In such cases, the squared test set correlation between observed and predicted values, or the Spearman rank correlation coefficient, may be more relevant.

7. CONCLUSIONS

The value of Golbraikh and Tropsha's work[3] is in its very effective demonstration that model quality and predictivity should be assessed using test data, not only on the training data. However, the measures of model fit suggested are overly complicated and potentially misleading. Instead, researchers should simply report the R^2 and RMSE or a similar statistic such as the standard error of prediction for the test set, which readers are more likely to be able to interpret. The value of R^2 should be calculated for the test data using eq 1, not from a regression of observed on predicted values. However, if relative rankings of the data suffice, rather than accurate numerical prediction, then it may be relevant to report, in addition, the squared test set correlation between observed and predicted values (or an equivalent metric).

This paper has elucidated some common misunderstandings surrounding the use of R^2 as a measure of model fit. Much confusion could be spared if everyone knew how to use R^2!

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Typographical and other errors in the paper by Golbraikh and Tropsha.[3] The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00206.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dave.winkler@csiro.au.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure−Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112*, 2889−2919.

(2) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.

(3) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(4) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, 2009; p 745.

(5) Burden, F.; Winkler, D. Bayesian Regularization of Neural Networks. *Methods Mol. Biol.* **2008**, *458*, 25−44.

(6) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.

(7) Burden, F. R.; Winkler, D. A. Optimal Sparse Descriptor Selection for QSAR Using Bayesian Methods. *QSAR Comb. Sci.* **2009**, *28*, 645−653.

(8) Besalu, E.; de Julian-Ortiz, J. V.; Pogliani, L. Trends and Plot Methods in MLR Studies. *J. Chem. Inf. Model.* **2007**, *47*, 751−760.

(9) Seber, G. A. F. *Linear Regression Analysis*; John Wiley & Sons: New York, 1977; p 465.

(10) Roy, K.; Chakraborty, P.; Mitra, I.; Ojha, P. K.; Kar, S.; Das, R. N. Some Case Studies on Application of "r_m^2" Metrics for Judging Quality of Quantitative Structure−Activity Relationship Predictions: Emphasis on Scaling of Response Data. *J. Comput. Chem.* **2013**, *34*, 1071−1082.

(11) Shayanfar, A.; Shayanfar, S. Is regression through origin useful in external validation of QSAR models? *Eur. J. Pharm. Sci.* **2014**, *59*, 31−35.

(12) Roy, K.; Kar, S. The r_m^2 metrics and regression through origin approach: Reliable and useful validation tools for predictive QSAR models (Commentary on "Is regression through origin useful in external validation of QSAR models?"). *Eur. J. Pharm. Sci.* **2014**, *62*, 111−114.

(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) Kvalseth, T. O. Cautionary Note about R^2. *Am. Stat.* **1985**, *39*, 279−285.

(15) Rousseeuw, P. J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871−880.

(16) Burden, F. R.; Winkler, D. A. An Optimal Self-Pruning Neural Network and Nonlinear Descriptor Selection in QSAR. *QSAR Comb. Sci.* **2009**, *28*, 1092−1097.

(17) 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.

(18) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.

(19) Kendall, M. G. *Rank Correlation Methods*, 4th ed.; Griffin: London, 1970.

(20) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P. Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **2007**, *47*, 1111−1122.

(21) Pearlman, D. A.; Charifson, P. S. Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System. *J. Med. Chem.* **2001**, *44*, 3417−3423.