Robust Cross-Validation of Linear Regression QSAR Models

Dmitry A. Konovalov,*,† Lyndon E. Llewellyn,‡ Yvan Vander Heyden,§ and Danny Coomans†

School of Mathematics, Physics & Information Technology, James Cook University, Townsville, Queensland 4811, Australia, Australian Institute of Marine Science, PMB 3, Townsville, Queensland 4810, Australia, Department of Analytical Chemistry and Pharmaceutical Technology, Pharmaceutical Institute, Vrije Universiteit Brussel, B-1050 Brussels, Belgium

Received June 24, 2008

A quantitative structure−activity relationship (QSAR) model is typically developed to predict the biochemical activity of untested compounds from the compounds' molecular structures. "The gold standard" of model validation is the blindfold prediction when the model's predictive power is assessed from how well the model predicts the activity values of compounds that were not considered in any way during the model development/calibration. However, during the development of a QSAR model, it is necessary to obtain some indication of the model's predictive power. This is often done by some form of cross-validation (CV). In this study, the concepts of the predictive power and fitting ability of a multiple linear regression (MLR) QSAR model were examined in the CV context allowing for the presence of outliers. Commonly used predictive power and fitting ability statistics were assessed via Monte Carlo cross-validation when applied to percent human intestinal absorption, blood-brain partition coefficient, and toxicity values of saxitoxin QSAR data sets, as well as three known benchmark data sets with known outlier contamination. It was found that (1) a robust version of MLR should always be preferred over the ordinary-least-squares MLR, regardless of the degree of outlier contamination and that (2) the model's predictive power should only be assessed via robust statistics. The Matlab and java source code used in this study is freely available from the QSAR-BENCH section of www.dmitrykonovalov.org for academic use. The Web site also contains the java-based QSAR-BENCH program, which could be run online via java's Web Start technology (supporting Windows, Mac OSX, Linux/Unix) to reproduce most of the reported results or apply the reported procedures to other data sets.

INTRODUCTION

The primary objective of a typical quantitative structure−activity/property relationship (QSAR) study is to develop a mathematical model that can be used for the prediction of considered biological activity, chemical reactivity, or physical properties of new, untested compounds from the compounds' molecular structures.[1−4] The *predictive power*[5] of a QSAR model is often estimated by some form of *cross-validation* (CV).[6−8] A chosen predictive power statistic (i.e., discrepancy measure,[7] loss function,[6] or generalization error) is then used to verify a newly devised set of molecular descriptors or to select a small number of descriptors from the pool of more than 3000 currently known descriptors.[3,9] Quantitative measures/statistics or assessment criteria of the predictive power of a QSAR model in the CV context is the focus of this study.

Note that there exists one other well-known nonparametric approach to estimating the predictive power of a model. This is the bootstrap approach,[10] which, however, was left outside the scope of this work because the CV approach is used more often in QSAR studies.

Most existing variations of the CV technique could be reduced to some form of the leave-group-out cross-validation (LGO-CV), where a sample of n observations is partitioned (i.e., split) into calibration (i.e., training) and validation (i.e., test) subsets. As implied by their names, the calibration subset (with n_c data points) is used to train a model, while the validation subset (with $n_v = n - n_c$ data points) is used to test how well the model predicts the new data, that is, the data points not used in the calibration procedure. Let, without loss of generality for any particular LGO-CV, the calibration and validation data points be numbered from 1 to n_c and ($n_c + 1$) to n, respectively. Then, for example, $\mathbf{y}_c = (y_1, y_2,..., y_{n_c})^T$ is the $n_c \times 1$ calibration vector (i.e., subset) of the activity values, while $\mathbf{y}_v = (y_{n_c+1}, y_{n_c+2},..., y_n)^T$ is the $n_v \times 1$ validation vector, where the superscript T denotes the transpose. After the model is calibrated, the model is used to estimate the activity values obtaining $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2,...,\hat{y}_n)^T$ with the corresponding residuals

$$e_i = y_i - \hat{y}_i \tag{1}$$

where e_i is the calibration (i.e., fitting) residual error for $1 \le i \le n_c$ and the validation (i.e., prediction or generalization) error for $n_c < i \le n$.

As far as the measures of the model's *fitting ability* and predictive power are concerned, the following statistics are typically considered: The mean squared error of calibration (MSE) and prediction (MSEP)

* Corresponding author e-mail: dmitry.konovalov@jcu.edu.au.
† James Cook University.
‡ Australian Institute of Marine Science.
§ Vrije Universiteit Brussel.

10.1021/ci800209k CCC: $40.75
Published on Web 10/01/2008

$$\mathrm{MSE} = \sum_{i=1}^{n_c} e_i^2/n_c \text{ and } \mathrm{MSEP} = \sum_{i=n_c+1}^{n_v} e_i^2/n_v \quad (2)$$

where their square roots are denoted by RMSE and RMSEP, respectively. The mean absolute error of calibration (MAE) and prediction (MAEP)[11]

$$\mathrm{MAE} = \sum_{i=1}^{n_c} |e_i|/n_c \text{ and } \mathrm{MAEP} = \sum_{i=n_c+1}^{n_v} |e_i|/n_v \quad (3)$$

The median absolute error of calibration (MedAE) and prediction (MedAEP)

$$\mathrm{MedAE} = \underset{1\le i\le n_c}{\mathrm{MED}}(|e_i|) \text{ and } \mathrm{MedAEP} = \underset{n_c<i\le n_v}{\mathrm{MED}}(|e_i|) \quad (4)$$

The coefficient of determination for calibration (R_c^2) and prediction (R_v^2), respectively

$$R_c^2 = 1 - \mathrm{SSE}/\mathrm{SST} \text{ and } R_v^2 = 1 - \mathrm{SSEP}/\mathrm{SSTP} \quad (5)$$

where

$$\mathrm{SSE} = \sum_{i=1}^{n_c} e_i^2 \text{ and } \mathrm{SSEP} = \sum_{i=n_c+1}^{n} e_i^2 \quad (6)$$

$$\mathrm{SST} = \sum_{i=1}^{n_c} (y_i - \bar{y}_c)^2 \text{ and } \mathrm{SSTP} = \sum_{i=n_c+1}^{n} (y_i - \bar{y}_v)^2 \quad (7)$$

$$\bar{y}_c = \sum_{i=1}^{n_c} y_i/n_c \text{ and } \bar{y}_v = \sum_{i=n_c+1}^{n} y_i/n_v \quad (8)$$

By definition, the fitting ability statistics such as MSE, MAE, MedAE, and R_c^2 have nothing to do with the measurement of the model's predictive power. It is trivial to devise a model, such as an artificial neural network with a sufficient number of layers and neurons, exhibiting almost arbitrarily "good" fitting ability, while possessing no predictive power (or generalization) at all. This phenomenon is known as overfitting[12] and is well understood in the field of machine learning, for example, via the Vapnik−Chervonenkis (VC) dimension,[13−15] where a more "flexible" model exhibits larger VC dimension corresponding to greater predictive uncertainty. Unfortunately, the definition of the VC dimension is highly mathematical and the dimension is difficult to calculate for a nonexpert.[13] In mathematically less exact terms, the results of VC theory could be interpreted as the more flexible a model, the less correlated the fitting ability and predictive power statistics become.

The main historical justification for reporting various fitting ability statistics was, and still is, because of simple linear regression (SLR) or multiple linear regression (MLR) being widely used in QSAR models. Since a typical MLR is performed with a relatively small number of predictor variables, it is considered to be safe from overfitting, and therefore, a chosen fitting statistic (e.g., MSE) is assumed to be a good approximation of the corresponding predictive statistic (e.g., MSEP). That is, in the case of SLR/MLR, it is generally assumed that there is a positive correlation between the corresponding fitting and predictive statistics: MSE and MSEP; MAE and MAEP; MedAE and MedAEP; R_c^2 and R_v^2. While this assumption is known to be false for more flexible (e.g., nonlinear) models,[16] the assumption is rarely questioned for the MLR models. The examination of the assumption was one of the objectives of this study, in which we found that the assumption could be false even for the relatively "rigid" SLR/MLR-based QSAR models, that is, with only one or few predictors.

In this study, the fitting ability and predictive power statistics are assessed in the context of cross-validation, which could be performed in a number of ways. The *"most primitive but nevertheless useful"*[6] variation of CV is called *hold-out*,[16] when the available sample is divided into a single pair of calibration and validation (i.e., hold-out) subsets or essentially performing a single instance of LGO-CV. The main problem with the hold-out CV is that the QSAR studies often deal with relatively small sample sizes, where n could be less than 100 and often is not larger than a few hundred. Partly as the consequence of the initially small n, small hold-out subsets, between 10 and 100 observations, are chosen, which makes such CV results highly dependent on fortunate or misfortunate selection of each individual hold-out observation and hence is statistically questionable.[17]

In an attempt to improve upon the hold-out cross-validation, the *leave-one-out* (LOO) cross-validation was developed.[6] The LOO cross-validation (LOO-CV) consists of running the LGO-CV n times using each of the observations as a validation subset of size $n_v = 1$. LOO-CV remains very popular with the QSAR community because the computational cost of the validation increases only linearly with n. However,more and more theoretical and practical evidence that LOO-CV must not be used for assessing the predictive power of models nor for model selection has been appearing.[2,5,18,19] In particular, "the high value of LOO q^2 appears to be the necessary but not sufficient condition for the model to have a high predictive power",[5] where

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (9)$$

is defined by "mimicking" the expression for the coefficient of determination (eq 5) between $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ observation and $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)^T$ predicted response (i.e., activity or property) values. In the above context, the activity value $\hat{y}_i$ is predicted for the ith compound by a QSAR model trained on the remaining $n_c = n - 1$ compounds of the calibration subset after the ith compound was *excluded* from the calibration subset. Furthermore, Shao[18] demonstrated that the LOO-CV method is asymptotically inconsistent, that is, the probability of the LOO-CV method to select the MLR model with the best predictive power does not converge to one as the total number of observations $n \to \infty$.

It has been known[18,20] since at least 1993 that in the case of MLR model selection, the probability of the LGO-CV method to select the best predictive MLR model converges to one only when $n_v/n \to 1$ and $n \to \infty$,[18] where the LGO-CV was repeatedly performed a sufficiently large number of times. That is, the larger n_v, the more accurate one can assess the model's predictive power, provided the resampling with replacement is done many times. Let N_{max} denote the total maximum number of distinct LGO cross-validation instances, where $N_{max} = n!/(n_c!n_v!)$. Since N_{max} increases very rapidly with n_v, exhaustive evaluation of all unique LGO-CV instances becomes computationally impossible. If

a computationally feasible number of LGO-CV instances (N) is randomly selected, the resulted procedure is known as the Monte Carlo cross-validation (MCCV).[18]

While the results of Shoa[18] are theoretically very appealing, they were difficult to apply to the QSAR practice since the $n_v/n \rightarrow 1$ condition was considered to be too "radical" to fulfill and hence was essentially ignored. In particular, Shao recommended $n_c = n^{3/4}$ and $N = 2n$, which, for example, would translate to the unacceptably small calibration subsets with $n_c = 32$, 53, and 178 for sample sizes $n = 100$, 200, and 1000, respectively. Contrasting examples were the recommendations to use just $n_v = 7$ by Wold and Eriksson[21] in 1995 and $n_v = n/10$ by Benigni and Bossa in 2008. The main reason for the relatively small n_c/n or large n_v/n recommended by Shoa[18] was an attempt to replace the inconsistent LOO-CV with the asymptotically consistent LGO-CV, while retaining the LOO $O(n)$ computational complexity, that is, the computational cost increases linearly with n. However, the choice of the cross-validation parameters, could be made "more practical" by increasing N.[20] For example, Xu et al.[22] verified that $N = n^2$ could be used to increase the required n_c/n as per the results of Zhang.[20] Note that Shao[18] and Zhang[20] used MSE and MSEP as the fitting and predictive statistics, which are not necessarily the most optimal statistics as we demonstrated in this study.

A practical compromise of

$$n_v = n_c = \frac{1}{2}n \tag{10}$$

was proposed as being arguably more acceptable for QSAR cross-validation testing.[23,24] The choice has an important statistical advantage of providing the maximum possible number of distinct calibration−validation partitions or LGO-CV instances.[1,2] Methodologically, the choice also makes sense because it treats the model building and model testing with equal importance by providing both with an equal share of the limited resource (data points). Then, to be consistent with the results of Shao[18] and Zhang,[20] the total number of MCCV iterations may need to be increased to its theoretical maximum of just a bit more than n^2, that is, essentially the $O(n^2)$ computational complexity.[20] Fortunately, in practice, less than $N = 100n$ iterations are often sufficient.[1]

Since the cross-validation process is often used as a variable selection method, it is important to perform a very large number of distinct LGO cross-validations. Essentially, to guard against overfitting, the number of visited distinct LGO cross-validations should be much greater than the number of adjustable parameters of a particular model. The LOO-CV is the perfect example of misusing the cross-validation process by failing the above requirement, where a sufficiently flexible model could easily fit the n LOO-CV instances. Then on the basis of such LOO "cross-validation", the overfitted model appears to have very good predictive power, when in fact the predictive power has not even been tested, and it is likely that the model possesses very poor predictive/generalization ability.[5] By using the largest possible set of distinct LGO-CV instances (via $n_v = n/2$) and performing MCCV many thousand times, it is arguably much more difficult (if not impossible) to achieve a mere fit of the data points for any QSAR model with the number of adjustable parameters much less than $N_{max} = n!/[(n/2)!]^2$. To illustrate the point, even in the case of a relatively small QSAR data set with just $n = 30$, N_{max} becomes $N_{max} = 1.55 \times 10^8$, which is much greater than the number of "free" parameters in most QSAR models.

Note that the use of the hold-out subset with n_h data points is also known as *external* validation if the remaining subset was used for calibration, as well as the variable selection process via *internal* LGO cross-validation with small n_v ($n_v \ll n$),[5,21] where n_h is typically between $n/10$ and $n/2$. While this external/internal distinction is meaningful for $n_v \ll n$ (and $n_v < n_h$), the difference between the external and internal validation vanishes once a reasonably large $n_v \approx n_h$ is used because each individual LGO-CV instance could be viewed as the external validation. Moreover, such so-called "external" validation still can not compete with the "gold standard" of validation, which could only be performed on the chemicals not considered in any way by the QSAR modelers, that is, chemicals that were not available or not considered at the time when the particular QSAR model was developed.[4]

Finally, one more variation of LGO is worth mentioning. That is, the m-fold or multifold cross-validation,[25] where the sample is partitioned into m mutuality exclusive subsets of similar size. Then each of the subsets is sequentially used as the validation group, while being excluded from the calibration. Multifold cross-validation offers no advantage over the LOO-CV, except for reducing the computation involved.[20]

In summary, the MCCV[1,2,4,18,22] method is emerging as a asymptotically consistent, as well as practical method, of assessing predictive power of MLR-based QSAR models, where the method could be safely used for the model selection without the risk of overfitting the model if used with $n_v = n/2$. Generally speaking, the method should be equally applicable to nonlinear QSAR models, where the $N = O(n^2)$ MLR-result of Zhang[20] should be replaced by the model's specific requirements on N.

The MCCV method was recently combined with nonparametric (bootstrap-like) hypothesis testing of variable selection resulting in the so-called Monte Carlo variable-selection (MCVS) method,[1] in which no assumption is made about the distribution of the residuals. The main idea behind MCVS is that a particular MLR-based QSAR model (i.e., a particular subset of descriptors) is selected if and only if the model consistently exhibits a superior predictive power, as measured by averaged MSEP,[1] during the MCCV procedure. Given the promising results obtained with MCVS, another objective of this study was to investigate which parts of the MCVS procedure could be improved by making them robust to the presence of outliers, which are very common in QSAR data sets. This study accomplished the first step toward fulfilling this objective by identifying more suitable predictive statistics comparing to MSEP.

Typically, when extensive cross-validation is not performed, *residual plots* are examined for a single calibration subset, a single validation subset, or the whole set. Actual or perceived outliers are detected and sometimes removed to improve the apparent performance of the final QSAR model(s). In general, such a handcrafting approach of dealing with outliers is not suitable[26,27] for multivariate models, and in particular, it is not feasible with MCCV/MCVS, when many thousand LGO cross-validations are performed. Moreover, one or more outliers may be large enough to become

the so-called *leverage points* tilting the *least-squares* SLR line or MLR hyperplane to the extent when the outliers become undetectable (e.g., via the residual plots), see the Introduction in the book of Rousseeuw and Leroy[27] for detailed examples and comprehensive discussion.

While the development of a procedure for the assessment of predictive power is mainly a statistical and algorithmic problem,[27] we see the main challenge in defining the new CV procedure in a way that is immediately applicable to the majority of published QSAR-MLR studies. The QSAR field of research is highly multidisciplinary, where it is of great interest to medicinal chemists, computer scientists, and statisticians/mathematicians. Unfortunately, it is too common that theoretical advances are largely ignored in the drug-development/discovery QSAR related practice, where the under-utilization of robust statistical methods[27,28] and the Shao's results[18] are such examples. While the usual scientific inertia is always present, it appears that for the results to have a chance of being adopted, they must be distilled to a clearly stated set of rules or recommendations, which are easy to explain, understand, and implement and which yield immediate measurable improvement for mainstream cases.[29] Therefore the main methodological focus of this study was to simplify and summarize the recent advances in robust statistical methods so that they could be immediately applied to the majority of the MLR-QSAR models.

MATERIALS AND METHODS

Let a QSAR data set consist of n compounds and p descriptors (i.e., predictor variables) stored as a $n \times (p + 1)$ matrix $\mathbf{Z}$,

$$\mathbf{Z} = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{11}$$

where y_i is the activity value of the ith compound and x_{ij} is the jth descriptor value of the ith compound. Generally speaking, $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ is the n observations of the response variable y, while $\mathbf{d}_j = (x_{1j}, x_{2j}, ..., x_{nj})^T$ is the n corresponding observations of the jth predictor variable x_j. The format of the $\mathbf{Z}$ matrix is chosen for its convenience when working with large p in the variable selection problems,[1] that is, the $\mathbf{y}$ vector is stored in the first column.

A MLR model estimates the activity values via

$$\hat{\mathbf{y}}_c = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_{n_c} \end{pmatrix} = \begin{pmatrix} 1, & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1, & x_{n_{c1}} & \cdots & x_{n_c p} \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix} = \mathbf{X}_c \mathbf{b} \tag{12}$$

$$\hat{\mathbf{y}}_v = \begin{pmatrix} \hat{y}_{n_c+1} \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1, & x_{n_c+1,1} & \cdots & x_{n_c+1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1, & x_{n1} & \cdots\cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix} = \mathbf{X}_v \mathbf{b} \tag{13}$$

where $\mathbf{b} = (b_0, b_1, ..., b_p)^T$ is the $(p + 1) \times 1$ column vector of regression coefficients including the intercept term. Depending on the minimization criterion, different estimators of $\mathbf{b}$ are possible.

The ordinary least-squares (OLS) L_2 or ℓ_2-norm MLR estimator of the $\mathbf{b}$ coefficients minimizes MSE and is given by

$$\mathbf{b} = (\mathbf{X}_c^T \mathbf{X}_c)^{-1} \mathbf{X}_c^T \mathbf{y}_c \tag{14}$$

If the $\mathbf{b}$ coefficients are calculated by minimizing MAE,[11] the corresponding estimator is known as the least absolute error (LAE) L_1 or ℓ_1-norm MLR estimator.[11] The OLS and LAE estimators are known to be nonrobust to the presence of outliers.[27]

More robust statistics, compared to MSE or MAE, are the median squared error of calibration (MedSE) and prediction (MedSEP)

$$\text{MedSE} = \underset{1 \le i \le n_c}{\text{MED}}(e_i^2) \text{ and MedSEP} = \underset{n_c < i \le n}{\text{MED}}(e_i^2) \tag{15}$$

the trimmed mean squared error of calibration (TMSE) and prediction (TMSEP)

$$\text{TMSE} = \sum_{i=1}^{h_c} |e_c(\mathbf{b})|_{(i)}^2 / h_c \text{ and TMSEP} = \sum_{i=1}^{h_v} |e_v(\mathbf{b})|_{(i)}^2 / h_v \tag{16}$$

the trimmed mean absolute error of calibration (TMAE) and prediction (TMAEP)

$$\text{TMAE} = \sum_{i=1}^{h_c} |e_c(\mathbf{b})|_{(i)} / h_c \text{ and TMAEP} = \sum_{i=1}^{h_v} |e_v(\mathbf{b})|_{(i)} / h_v \tag{17}$$

where h_c and h_v are the coverage parameters, respectively, given by

$$h_c = [(n_c + p + 1)/2], \; h_v = [(n_v + p + 1)/2] \tag{18}$$

and where $|e_c(\mathbf{b})|_{(i)}$ and $|e_v(\mathbf{b})|_{(i)}$ are the ith smallest absolute calibration and validation residual from fit $\mathbf{b}$, see eqs 12 and 13, respectivly.[30,31]

Perhaps the best known robust or high breakdown (HB)[32,33] MLR estimator is the least median of squares (LMS) estimator,[27,34] which minimizes MedSE. The LMS estimator achieves the theoretical limit of robustness, where up to 50% of the calibration data points could be replaced by outliers.[27] Other known HB estimators are, the least trimmed squares (LTS) estimator,[30] which minimizes TMSE, and the least trimmed absolute errors (LTA) estimator,[30] which minimizes TMAE.

Using the coverage parameter from eq 18, the LTS and LTA estimators become as robust as the LMS, that is, the breakdown values of LTS, LTA, and LMS are equal to the maximum possible of $[(n_c - p)/2 + 1]/n_c \approx 50\%$.[30]

Note that *exact* evaluation of the considered HB estimators is computationally prohibitive.[30,31,35] Therefore, hereafter, while the same abbreviations for HB estimators are used (e.g., LTA, LMS, and LTS) for simplicity, they denote the corresponding *approximate* algorithms. The LTS and LTA estimators are statistically more efficient than LMS if the true residuals are normally distributed.[36] Until very recently the approximate LTS and LTA estimators were harder to compute than the approximate LMS.[30] This has just been changed with the development of reasonably efficient LTS and LTA algorithms.[30,31,35] Thus the approximate LTS or LTA estimator should always now be preferred over LMS, while the choice between the LTS and LTA estimators is a tradeoff between computational speed and statistical efficiency. The statistical efficiency of LTA is not much below

that of LTS, but LTA's computational complexity is of a lower order than that of LTS.[31]

Robust Goodness of Fit. Out of all the statistics considered so far, only the coefficient of determination R^2 (eq 5) could be used directly to assess the goodness of model's calibration or prediction fit. Temporarily ignoring the issue of cross-validation to simplify notation, let $\mathbf{y} = (y_1, y_2,..., y_n)^T$ be the observed activity values and $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2,..., \hat{y}_n)^T$ be the activity values estimated by the model. Then the coefficient of determination R^2 could be defined as[37]

$$R_{\text{var}}^2 = 1 - \frac{\text{var}(e)}{\text{var}(y)} \tag{19}$$

which is interpreted as the explained variance in the response variable since the second term is the fraction of variance in the activity values that is not explained by the model, where

$$\text{var}(e) = \sum_{i=1}^{n} (e_i - \bar{e})^2/(n-1) \quad \text{and}$$

$$\text{var}(y) = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$$

are the unbiased estimators of variance for the residual (see eq 1) and response values, respectively. Note that the standard interpretation relies on the assumption that the mean residual error (see eq 1) is zero, $\bar{e} = \sum_{i=1}^{n} e_i/n = 0$, obtaining the most commonly used expression for R^2 (e.g., eq 5)

$$R_{\text{SSE}}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{20}$$

Note that only in the case of OLS-MLR most commonly used definitions of the coefficient are equivalent,[37] for example, $R_{\text{var}}^2 = R_{\text{SSE}}^2$. Also, while not the focus of this study, it is important to highlight that R_{SSE}^2 could be adjusted for the available degrees of freedom via

$$R_{\text{SSE-MLR}}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-p-1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)} \tag{21}$$

to prevent artificially inflating R_{SSE}^2 values by just adding extra descriptors. This adjustment is rarely used when the results are reported for the nonlinear models with many adjustable parameters, such as artificial neural networks, and therefore not useful for cross-model comparison.

A number of robust generalization for R^2 estimators were proposed over the years.[27,37−40] For example, Rousseeuw and Leroy[27] (R_{RL}^2) and Kvalseth[37] (R_{K}^2) made eq 20 more robust via

$$R_{\text{RL}}^2 = 1 - (\text{MED}(|e_i|)/\text{MAD}(y))^2 \tag{22}$$

$$R_{\text{K}}^2 = 1 - (\text{MED}(|e_i|)/\text{MED}(|y_i - \bar{y}|))^2 \tag{23}$$

where MED is the median and MAD is the median absolute deviation given by

$$\text{MAD}(y) = \text{MED}(|y_i - \text{MED}(y)|) \tag{24}$$

and where $\text{MAD}^2(y)$ is a robust version of variance, var(y).[27]

The main conceptual problem with either considered robust or classical definitions of R^2 is that they could become meaningless by failing to remain in the $0 \le R^2 \le 1$ range when non-OLS MLR estimators are used or when assessing nonlinear models.[37] This certainly presents a major problem when OLS-MLR models are needed to be compared consistently against other types of models, which is a common situation in QSAR studies.[2]

It is known that in the case of the OLS regression with a single predictor $R_{\text{var}}^2 = R_{\text{SSE}}^2 = R_{\text{corr}}^2$,[37] where

$$R_{\text{corr}}^2 = r^2 \tag{25}$$

and where r is the Pearson product-moment correlation coefficient (or simply the sample correlation coefficient) between y and $\hat{y}$ variables

$$r = \frac{\text{cov}(\hat{y}, y)}{\sqrt{\text{var}(\hat{y})\text{var}(y)}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}\hat{})(y_i - \bar{y})}{\left(\sum_{i=1}^{n} (\hat{y}_i - \bar{y}\hat{})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2\right)^{1/2}} \tag{26}$$

Strictly speaking, in all other cases, such as non-OLS MLR and nonlinear models, R_{corr}^2 measures how well y could be OLS fitted by $\hat{y}$ via $y = a + b\hat{y}$. If the $R_{\text{corr}}^2 = r^2$ definition is retained as the generalization of the goodness of fit for any model, the generalization step appears to be very minor compared to all the benefits gained from the use of R_{corr}^2 over R_{SSE}^2. In particular,[37] (1) $0 \le R_{\text{corr}}^2 \le 1$, where $R_{\text{corr}}^2 = 1$ corresponds to perfect fit; (2) R_{corr}^2 is applicable to any type of models, regardless of the statistical properties of the model variables (including residual e); (3) R_{corr}^2 values for different models fitted to the same data set are directly compatible, where larger R_{corr}^2 indicates a better OLS fit $y = a + b\hat{y}$. Note that for any types of models, the $(y_i, \hat{y}_i)$ scatter plot could easily be checked to verify that a model does not exhibit nonlinear fitting behavior, and hence, the $y = a + b\hat{y}$ fit is comparable across different models. A robust version of R_{corr}^2 could then be obtained from a robust correlation coefficient ρ as

$$R_{\text{corr}}^2 = \rho^2 \tag{27}$$

The ρ coefficient could be derived from the alternative (but exactly equivalent to eq 26) expression for r[38]

$$r = \frac{\text{var}(u) - \text{var}(v)}{\text{var}(u) + \text{var}(v)} \tag{28}$$

$$u_i = \tilde{x}_i + \tilde{y}_i, \quad v_i = \tilde{x}_i - \tilde{y}_i \tag{29}$$

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{\text{var}(x)}}, \quad \tilde{y}_i = \frac{y_i - \bar{y}}{\sqrt{\text{var}(y)}} \tag{30}$$

where, generally speaking, $\{(x_1, y_1),..., (x_n, y_n)\}$ are observed values of an arbitrary bivariate random variable $\{(x, y)\}$. A robust version of r is then given by

$$\rho = \frac{\text{var}_{\text{rob}}(u) - \text{var}_{\text{rob}}(v)}{\text{var}_{\text{rob}}(u) + \text{var}_{\text{rob}}(v)} \tag{31}$$

$$u_i = \tilde{x}_i + \tilde{y}_i, \quad v_i = \tilde{x}_i - \tilde{y}_i \tag{32}$$

$$\tilde{x}_i = \frac{x_i - \text{MED}(x)}{\sqrt{\text{var}_{\text{rob}}(x)}}, \quad \tilde{y}_i = \frac{y_i - \text{MED}(y)}{\sqrt{\text{var}_{\text{rob}}(y)}} \tag{33}$$

where the variance (var) and mean values in the original expressions are replaced by the robust variance (var_{rob}) and

median values. While median (MED) is the standard robust replacement of mean, a few choices are available for $\mathrm{var}_{\mathrm{rob}}$. For example, one popular choice is[27,40]

$$\mathrm{var}_{\mathrm{rob}}(y) = \mathrm{MAD}^2(y) \tag{34}$$

where MAD is defined as per eq 24. Another choice is in our opinion a more direct conversion of the sample variance expression

$$\mathrm{var}(y) = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$$

obtaining

$$\mathrm{var}_{\mathrm{rob}}(y) = \mathrm{MED}[(y_i - \mathrm{MED}(y))^2] \tag{35}$$

LTS Estimator. The approximate concentration[30,41] algorithm for LTS[32] (CLTS) was implemented. Note that the calibration subset of the original sample is assumed throughout this subsection, in which the "c" subscript is omitted to simplify the notation. For its first step, the CLTS algorithm generates K so-called "elemental sets",[33] ($\mathbf{X}_1$, $\mathbf{X}_2$,..., $\mathbf{X}_K$), by randomly selecting $p + 1$ sample cases for each elemental set, where $\mathbf{X}_j$ is the $(p + 1) \times (p + 1)$ submatrix of the given $n \times (p + 1)$ descriptor matrix and where the first column contains 1's to allow for intercept as per eqs 12 and 13. The algorithm utilizes a major advance in the theory of HB estimators, where it was shown that restricting trial subsets to only elemental sets introduces asymptotically negligible error.[33] It was tentatively reported[35] that the K parameter could set to $K = 3 \times 2^{p+1}$. Until a peer-reviewed publication confirms that suggestion, we are setting $K = 500$.[30] The OLS estimator is then used to obtain the initial set of K "attractors" or "trial fits", ($\mathbf{b}_{0,1}$, $\mathbf{b}_{0,2}$,..., $\mathbf{b}_{0,K}$), from K elemental sets, where $\mathbf{b}$ denotes a $(p + 1) \times 1$ vector of the MLR coefficients. The OLS estimator from the whole sample is also added and is stored in $\mathbf{b}_{0,K+1}$.[35] This could be viewed as a safety measure to ensure that the obtained solution is never worse than the standard OLS.

The concentration part of the algorithm starts from $K + 1$ attractors, ($\mathbf{b}_{0,1}$, $\mathbf{b}_{0,2}$,..., $\mathbf{b}_{0,K+1}$). The mth iteration of the concentration is performed by computing all n residuals, $\mathbf{e}(\mathbf{b}_{m-1,j}) = \mathbf{y} - \mathbf{X}\mathbf{b}_{m-1,j}$, for $1 \le j \le K + 1$. The OLS $\mathbf{b}_{m,j}$ estimates are computed from h cases corresponding to the smallest squared residuals, for $1 \le j \le K + 1$. The concentration step is performed k times, which is set to between $k = 10$ and $k = 20$ because the $k = 10$ recomendation[30] was sometimes not sufficient to achieve convergence on the considered data sets. The concentration step was considered convergent for a particular $\mathbf{b}_{m,j}$ if the corresponding TMSE value did not change from the previous iteration, that is, $\mathrm{TMSE}_{m-1,j} = \mathrm{TMSE}_{m,j}$.

LTA Estimator. The following approximate LTA algorithm has been implemented for this study as per Hawkins and Olive.[31] As per the CLTS algorithm, the LTA algorithm randomly selects K elemental sets[33] plus the whole calibration subset, ($\mathbf{X}_1$, $\mathbf{X}_2$,..., $\mathbf{X}_{K+1}$). The OLS estimator is then used to obtain K fits, ($\mathbf{b}_1$, $\mathbf{b}_2$,..., $\mathbf{b}_{K+1}$), and the corresponding n residuals are calculated, $\mathbf{e}(\mathbf{b}_j) = \mathbf{y} - \mathbf{X}\mathbf{b}_j$. The smallest h absolute values among the residuals for each fit are found, obtaining $\mathrm{TMAE}(\mathbf{b}_j)$ for the jth fit. The LTA is given by the fit yielding the lowest TMAE.

LMedA Estimator. The above LTA estimator could be generalized for other statistics. In particular, the least median absolute error (LMedA) estimator minimizes the MedAE statistic (eq 4), where the LTA algorithm was modified to select the elemental set with the lowest MedAE.

Huber and LAE Estimators. Note that the following two estimators were fully implemented (disabled in QSAR-BENCH) but did not yield sufficient improvement over the OLS estimator in this study. The estimators are reported here merely to demonstrate that they were considered and possibly save unproductive effort in future studies.

Huber[42] suggested to minimize objective (cost or loss) function

$$L = \sum_{i=1}^{n_c} L(e_i) \tag{36}$$

with

$$L_{\mathrm{H}}(e) = \begin{cases} \frac{1}{2}e^2, & |e| < t \\ t|e| - \frac{1}{2}t^2, & |e| \ge t \end{cases} \tag{37}$$

thus combining the LAE's more "robust" treatment of outliers and OLS's Gaussian treatment of small residuals,[43] where t is the threshold parameter controlling the linear and quadratic treatment of the residual errors or a *tuning constant*. The corresponding estimator of $\mathbf{b}$ coefficients is known as the Huber estimator. The estimator is currently quite popular because of the discovery of the relatively efficient *iteratively reweighted least-squares* (IRLS) method.[44]

In general, the IRLS method minimizes any objective (or cost) function L, where the OLS, LAE, and Huber estimators are given by $L_{\mathrm{OLS}}(e) = e^2$, $L_{\mathrm{LAE}}(e) = |e|$, and $L_{\mathrm{H}}(e)$, respectively. The solution is iterative starting from the least-squares estimates ($B^{(m=0)} = B$)[45,46]

$$B^{(m)} = (\mathbf{X}_c^{\mathrm{T}}\mathbf{W}^{(m-1)}\mathbf{X}_c)^{-1}\mathbf{X}_c^{\mathrm{T}}\mathbf{W}^{(m-1)}Y_c \tag{38}$$

where $\mathbf{W}$ is a diagonal matrix of weights

$$\mathbf{W}^{(m-1)} = \begin{pmatrix} w\left(e_1^{(m-1)}\right) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w\left(e_{n_c}^{(m-1)}\right) \end{pmatrix} \tag{39}$$

$\varepsilon_i^{(m-1)}$ is the residual at iteration t, and the Huber weight function is given by

$$w_{\mathrm{H}}(e) = \begin{cases} 1, & |e| < t \\ t/|e|, & |e| \ge t \end{cases} \tag{40}$$

Then, the Huber parameter t controls the tradeoff between the number of iterations and robustness: decreasing t makes the Huber estimator more "robust" to outliers, but a greater number of iterations is typically required to achieve the same numerical accuracy of the solution. In this study, the IRLS method was stopped once the relative change of MAE was less than 0.01%, that is, $2|\mathrm{MAE}_t - \mathrm{MAE}_{t-1}|/(\mathrm{MAE}_t + \mathrm{MAE}_{t-1}) < 10^{-4}$. It was verified that less than five iterations was often sufficient to achieve the 0.01% accuracy of MAE with $t_m = 1.345 \times s_{m-1}$, where s_{m-1} is the $(m - 1)$th iteration estimate of the standard deviation σ of the residuals. A popular robust estimator of σ is $s = 1.483 \times \mathrm{MAD}$.[28,47] We used the MAE-based estimator of σ given by

$$s = \sqrt{\pi/2} \times \mathrm{MAE} = 1.2533 \times \mathrm{MAE} \tag{41}$$

because MAE is recalculated at each iteration and to make each iteration a smooth function of MAE.

For the Huber estimator, the relatively large t is justified by the context of this study, when OLS works well for small and normally distributed residual errors. In particular, the current choice of $t = 1.345 \times \sigma = 1.686 \times$ MAE yields 95% efficiency when the errors are normal,[28,45,46] while still protecting (only to some extent) the Huber estimator against outliers.

The LMA estimator is approximated by setting $t = 0.01 \times s$.

logBB-TPSA Data Set. The blood−brain (BB) partition coefficient values, log BB = log (C_{brain}/C_{blood}), where C_{brain} and C_{blood} are the equilibrium concentrations of the compound in the brain and the blood, respectively.[48] The 289 unique compounds were extracted[1] from the data set of Abraham et al.[24] The first log BB related data set was labeled as the logBB-TPSA data set combining the log BB activity values with the Iv,[24] Ic,[24] and TPSA(NO)[49] indicators/descriptors,[1] see Supporting Information, Table S1. The Iv indicator is the origin-of-data indicator, where Iv = 1 and Iv = 0 are for the data points measured in vitro and in vivo, respectively.[24] The Ic indicator counts the number of carboxylic-acid groups in a compound. The TPSA(NO) descriptor is the topological polar surface area[49] using N and O polar contributions calculated via the freely accessible E-DRAGON Web site.[50]

logHIA-ALOGP Data Set. The logHIA-ALOGP data set was identical to the KS127-logHIA data set from Konovalov et al.,[1] see Supporting Information, Table S2. This data set consisted of the percent human intestinal absorption (%HIA) of 127 compounds, which had neither 0 nor 100% HIA values.[51,52] With MCVS,[1] it was found that the Ghose−Crippen octanol−water partition coefficient[53,54] (ALOGP) was the single best performing predictive descriptor out of about 1500 E-DRAGON[9,50,55−57] descriptors for the log HIA = log {ln [100/(100 − %HIA)]} values.[51]

logTox-X5Av Data Set. The highest reported toxicity values of 30 saxitoxins (STXs) have been recently collated for QSAR analysis and will be referred to as the LL30 data set.[3] Llewellyn[3] also encoded the considered STXs in SMILES[58] notation, which was used in this study to calculate 1795 descriptors[9,59] via the freely available Parameter Client (PCLIENT)[55,57] Web site. Using the freely available QSAR-BENCH[1,2] program, the descriptors were pruned by removing the descriptor columns which were constant, duplicated, or had error codes (i.e., the descriptor could not be calculated) arriving at 1303 descriptors. Note that the majority of the considered descriptors (1264 out of 1303) were from the E-DRAGON program.[9,50]

It was found that the following descriptors were the most correlated to the toxicity values: $r_{X5Av} = -0.716$, $r_{X4A} = -0.685$, $r_{R2p} = -0.638$, and $r_{HATS7e} = -0.635$, where the topological descriptor X5Av is an average valence connectivity index,[60,61] which encodes the presence of double and triple bonds as well as heteroatoms in the molecule.[9,59]

Because it is quite common[1,2] for an activity or property to have exponential dependence on its molecular descriptors, the logarithm of the toxicity (log Tox) was also considered. High toxicity would normally imply high solubility and high absorption of a compound, and since both solubility and absorption are better modeled by the logarithm function,[51,62] the log Tox value would be expected[23] to be better described by the molecular descriptors. This was confirmed by examining first few most correlated to log Tox descriptors, $r_{X5Av} = -0.815$, $r_{X3Av} = -0.747$, $r_{X4Av} = -0.744$, and $r_{nOHs} = -0.741$, which were consistently more correlated (in absolute terms) to the logarithm of toxicity than to the toxicity values themselves.

The logTox-X5Av data set combined the log Tox values and the X5Av descriptor,[1] see Supporting Information, Table S3.

CYGOB1, GESELL, BODY-BRAIN Data Sets. The control (or benchmark) data sets were taken from the book by Rousseeuw and Leroy[27] (Table 3 on page 27, Table 5 on page 47, and Table 7 on page 57). The CYGOB1 data form the Hertzsprung-Russell diagram of the start cluster CYG OB1, which contains 47 starts in the direction of Cygnus.[27] The GESELL data set contains the age (in months) at which a child utters its first word as the explanatory variable, while the response variable is its Gesell adaptive score.[27] The BODY-BRAIN data set presents the brain weight and the body weight of 28 animals.[27] The CYGOB1, BODY-BRAIN, and GESELL data sets were selected as the benchmark data sets with known outlier contaminations having major, moderate and negligible contaminations, respectively.[27]

Simulated Companion Data Sets. While the CYGOB1, BODY-BRAIN, and GESELL data sets are well studied, their underlying residual distributions are not known for certain. Since the OLS estimator is known to be the best available estimator for asymptotically large, homogeneous data sets with normally distributed residual errors, simulated data sets with normal residuals were created while resembling the considered data sets in some other respects. For each of the above-described real data sets, a corresponding simulated data set was created via

$$u_i = a + bv_i + \varepsilon_i, \quad 1 \le i \le n \tag{42}$$

where n is the number of points in the corresponding data set and ε_i is a random variable with mean zero and variance

$$s^2 = \sum_{i=1}^{n} e_i^2/(n-2) \tag{43}$$

The a and b coefficients and the $\{e_1, e_2, ..., e_n\}$ residuals were obtained from the CLTS fit of the original data set. In the case of the logBB-TPSA data set, the CLTS regression coefficient for the TPSA(NO) variable was used as b.

RESULTS AND DISCUSSION

The Matlab and java source code used in this study is freely available from the QSAR-BENCH section of www.dmitrykonovalov.org and from the Supporting Information.

Robust LMS results in Figures 4, 5, and 7 of Rousseeuw and Leroy' book,[27] corresponding to the CYGOB1, GESELL, and BODY-BRAIN data sets, were closely reproduced using the CLTS and LTA estimators (first column, last three rows of Figure 1) verifying our implementation of the algorithms. In all three benchmark data sets the differences between the CLTS and LTA fitting lines were qualitatively negligible. For Figure 1, the CLTS algorithm was run with $K_{CLTS} = 100$ and $k = 10$, while the LTA algorithm was run with $K_{LTA} = 1000$, consuming similar amount of computational time compared to CLTS. The algorithms were run a number of times with the above

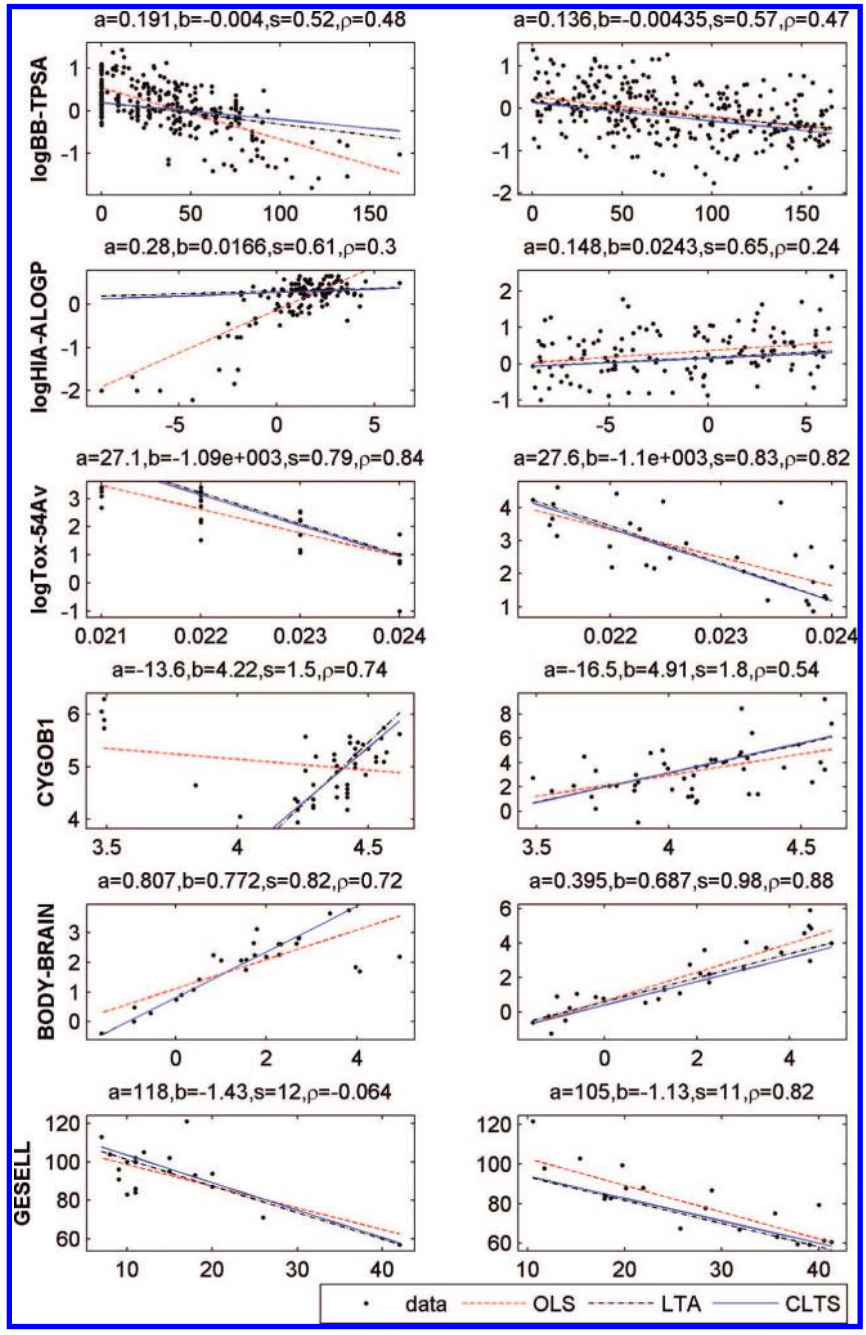

Figure 1. OLS (dashed line), LTA (dotted-dashed line), and CLTS (solid line) estimators. Data sets are in rows, where the corresponding simulated companion data sets are in the second column. The a and b coefficients (in the titles) are from the $y = a + bx$ CLTS regression of the data sets, s is the standard deviation estimate (eq 43), and ρ is the robust correlation coefficient (eqs 31 and 35) between $\mathbf{y}$ and CLTS estimate $\hat{\mathbf{y}}$.

parameters to verify that the obtained solutions were stable (i.e., convergent). Note that the algorithms with much smaller parameter values, $K_{CLTS} = 20$ and $K_{LTA} = 200$, arrived at similar solutions most of the time, confirming that $K_{CLTS} = 100$ and $K_{LTA} = 1000$ were sufficiently large for testing the stability of the final results in Figure 1.

Figure 1 was generated using only the Matlab implementation of the CLTS and LTA algorithms, where each robust (CLTS or LTA) regression took under one minute on a Pentium 4 desktop PC. This verified that the robust linear regression could be routinely and easily performed even using the relatively slow Matlab program.

The CLTS, LTA, and LMedA results in Figures 4 and 5 were calculated via the QSAR-BENCH program with $K_{CLTS} = 500$, $k = 20$ and $K_{LTA} = K_{LMedA} = 10000$, respectively.

QSAR Data Sets. While the CLTS and LTA algorithms where visually very similar in the graphical plot of the results in Figure 1, the algorithms exhibited very noticeable difference in the convergence rate when the MLR expressions were required. The CLTS-MLR expressions were extremely stable with the $K_{CLTS} = 500$ and $k = 20$ parameters, confirming the full convergence of the CLTS estimator for all considered data sets. On the other hand, the LTA estimator produced MLR expressions where the MLR coefficients kept varying in the first two significant digits while consuming similar or greater computational time comparing to CLTS (with $K_{LTA} = 10000$). Hence only the convergent CLTS-MLR models of the considered QSAR data sets were presented here (obtained via QSAR-BENCH with $K_{CLTS} = 500$ and $k = 20$)

$$\log BB_{CLTS} = 1.076 - 0.0191 \times TPSA(NO) - 0.760 \times Iv - 1.134 \times Ic,\ R_{SSE}{}^2 = 0.518,\ \rho_c{}^2 = 0.39,\ MedAE = 0.255,\ F = 158 \quad (44)$$

$$\log HIA_{CLTS} = 0.28 + 0.0166 \times ALOGP,\ R_{SSE}{}^2 = -0.0245,\ \rho_c{}^2 = 0.092,\ MedAE = 0.163,\ F = 15 \quad (45)$$

$$\log Tox_{CLTS} = 27.27 - 1098 \times X5Av,\ R_{SSE}{}^2 = 0.497,\ \rho_c{}^2 = 0.708,\ MedAE = 0.335,\ F = 68 \quad (46)$$

where $\rho_c{}^2$ is the robust coefficient of determination (eqs 27, 31, and 35) with the subscript "c" as a reminder that the value was calculated from the calibration subset (whole data set in this case), $R_{SSE}{}^2$ is the classical coefficient of determination (eq 20), and F is the conventional F statistic adjusted for MLR

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / (n - p - 1)} \quad (47)$$

Note that logBB-TPSA subfigure in Figure 1 presents results for SLR models of log BB, log BB = $a - b \times$ TPSA(NO).

In the above MLR expressions, $R_{SSE}{}^2$ values were reported to illustrate that $R_{SSE}{}^2$ is not necessarily always less or always greater than the corresponding robust equivalent, ρ^2. In addition, meaningless negative $R_{SSE}{}^2$ values were easily obtained for non-OLS-MLR models, for example, $R_{SSE}{}^2 = -0.0245$ in the CLTS-MLR model of log HIA.

The robust (CLTS or LTA) regression of the QSAR data sets revealed that the ALOGP descriptor may be quite irrelevant in the %HIA prediction problem (log HIA − ALOGP row, first column in Figure 1) beyond the simple observation that a compound is absorbed better if it is soluble (well approximated by ALOGP). However, the considered %HIA data set was originally constructed to remove the solubility effect by including compounds, which had neither 0 nor 100 %HIA.[51,52] Some residual solubility effect still remained but only less than 10% of variance in log HIA is explained by ALOGP (eq 45). In fact, the strongly tilted OLS line is essentially controlled by less than 10% of 127 log HIA data points while the bulk of the data (about 50% of it) is best fitted by a near horizontal line (CLTS and LTA lines). This is quite typical of OLS, where the OLS regression is essentially controlled by the outliers or extreme data points, while the bulk of the data could have disproportionally small influence. While general justification of the robust regression is outside the scope of this study, we could only assume that for some class of problems a regression line should approximate the bulk of the data points and hence the robust regression must be always preferred over the OLS line in such instances. This is, arguably, the case for many QSAR

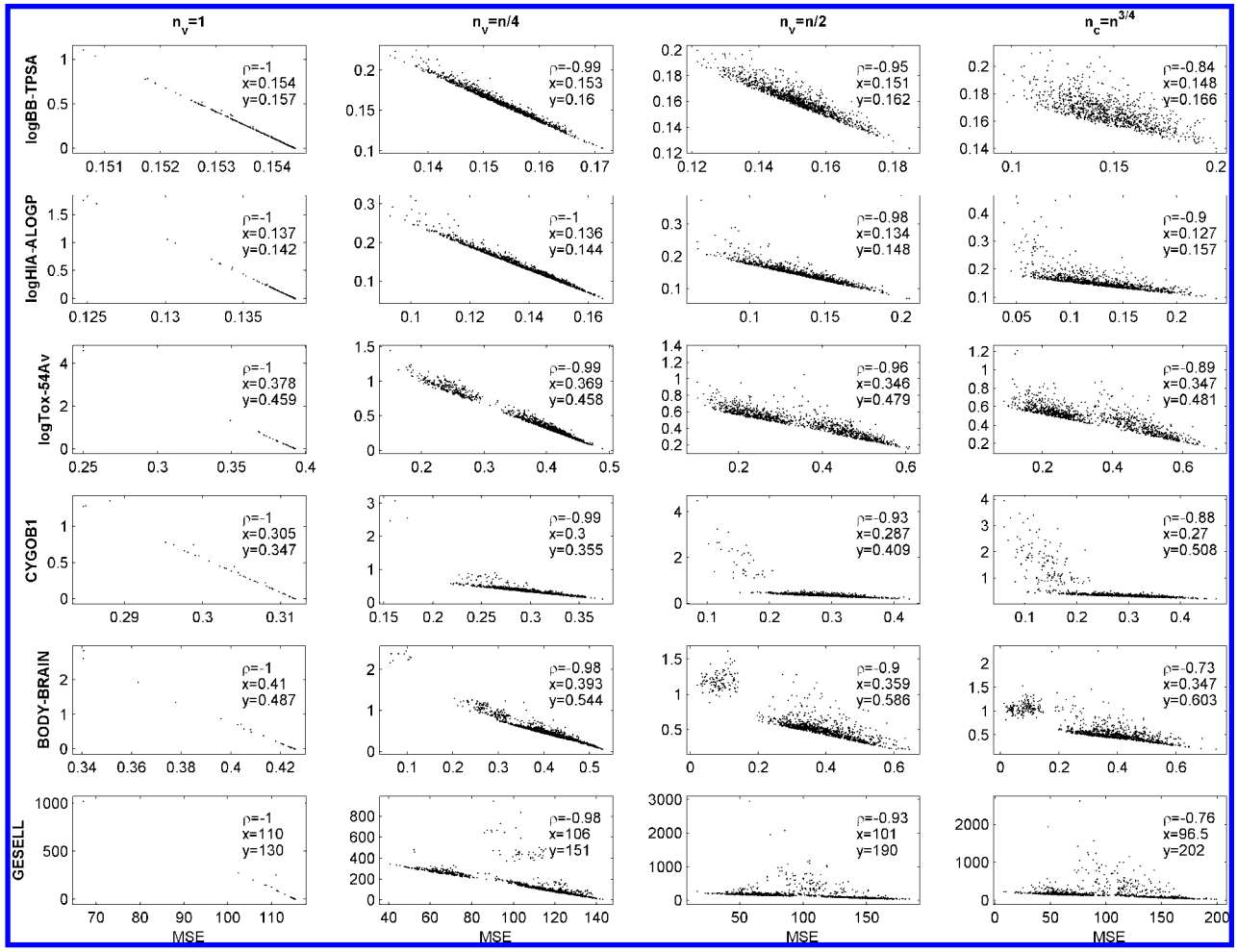

Figure 2. MSEP as a function of MSE obtained from MCCV($N = 1000$) with the OLS estimator, where ρ is the robust correlation (eqs 31 and 35) between the MSE and MSEP values, and x and y are the averaged MSE and MSEP values, respectively. Data sets are in rows; cross-validations are in columns.

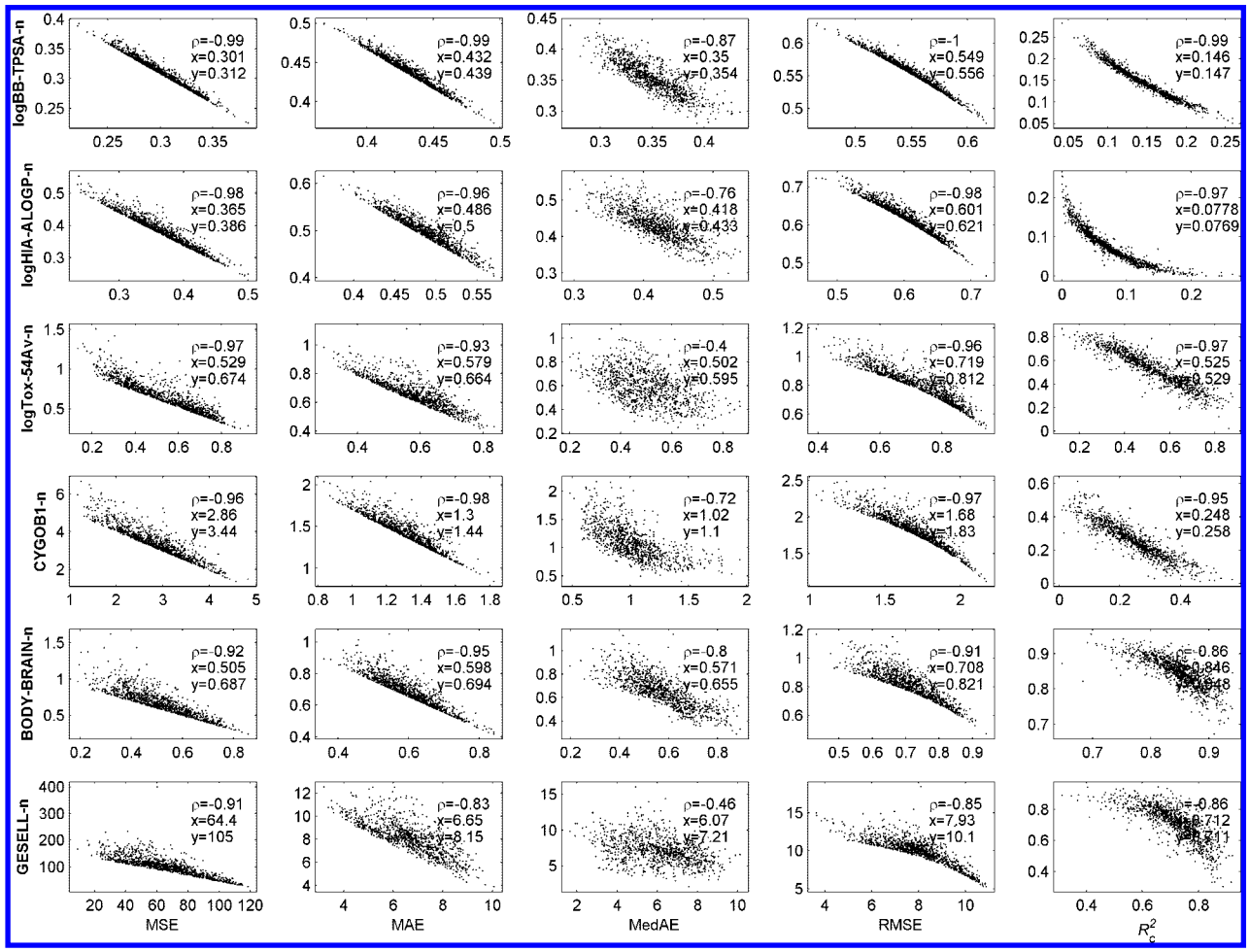

Figure 3. Same as in Figure 2 but for MSEP, MAEP, MedAEP, RMSEP, and R_v^2 as functions of MSE, MAE, MedAE, RMSE, and R_c^2 obtained with the OLS estimator, respectively, where the MCCV was performed with $n_v = n/2$ and $N = 1000$. Simulated data sets are in rows.

studies, where extreme points and outliers are quite common but could be quite irrelevant to the problem under consideration, for example, active transport processes may be at play for the outliers while passive absorption is studied.

While the above argument in favor of the robust regression is mainly methodological rather than purely statistical, in the following subsections, we attempted to demonstrate that the robust regression should also be preferred over OLS from the cross-validation point view.

The ALOGP descriptor was selected by "data mining" the E-DRAGON descriptors using the MCCV/MCVS method,[1] that is, MSE was minimized via OLS-MLR, while MSEP was minimized via MCCV/MCVS to select ALOGP as the best predictive descriptor. The absence of robust correlation between %HIA and ALOGP is an interesting result as it indicates that the variable selection method (e.g., MCVS)[1] may be less important than the choice of the fitting and predictive statistics.

The OLS regressions of the logBB-TPSA and logTox-X5Av data sets were confirmed to a large extent by the robust CLTS and LTA fits.

Simulated Data Sets. The close statistical resemblance of the real data sets and their simulated companions were verified by observing similarity in the corresponding *a*, *b*, and *s* parameters from their CLTS fits, compared in the two subfigures in each row in Figure 1. The simulated companion data sets were generated many times, and Figure 1 displays a typical example, which was not optimized in any way and could be trivially regenerated using the supplied Matlab source code.

The second column of Figure 1 highlights the fact that the OLS estimator has no "practical" advantage over the robust CLTS and LTA estimators in the presence of perfectly normal residuals, all three estimators produce very similar lines.

MCCV of the OLS Regression. Each subfigure in Figure 2 is a scatter plot of 1000 MCCV instances of MSE (i.e., calibration error) and MSEP (i.e., validation error) values obtained with the OLS estimator. Four cases were considered using the same MCCV computer (in Matlab) code: $n_v = 1$ (first column of subfigures), $n_v = n/4$ (second column), $n_v = n/2$ (third column), and $n_v = n - n^{3/4}$ (forth column). The generic MCCV processing of all n_v ensured consistency in the comparison across different n_v for each data set. Note that the exhaustive treatment of $n_v = 1$ (LOO-CV), where only *n* distinct cross-validation instances existed, produced essentially identical results (not shown) because the large *N* made sure that most if not all distinct LOO instances were visited.

Without exception, Figure 2 clearly demonstrates that the common "unquestioned" assumption about the existence of positive correlation between MSE and MSEP could be in fact false. Moreover, the robust correlation coefficient between the MSE and MSEP values $\rho_{MSE,OLS}$ could be

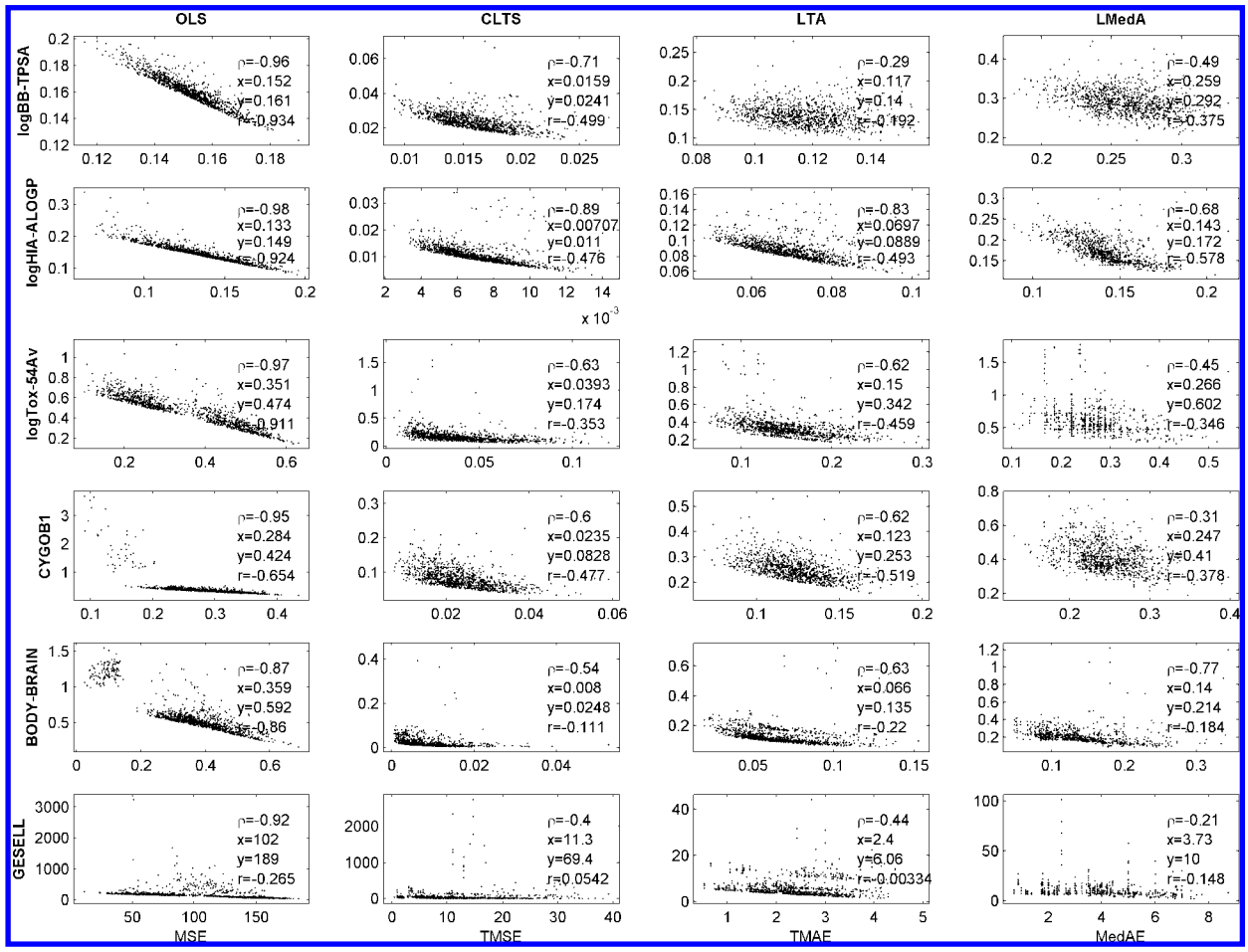

Figure 4. Same as in Figure 3 but for MSEP, TMSEP, TMAEP, and MedAEP as functions of MSE, TMSE, TMAE, and MedAE obtained with the OLS, CLTS, LTA, and LMedA estimators, respectively, where r is the classical correlation coefficient (eq 26).

negative. That is, the better the OLS regression, the worse the predictive power of the corresponding OLS-MLR model. Interestingly, the correlation coefficient $\rho_{MSE,OLS}$ becomes slightly less negative as n_v increases from 1 to $(n-n^{3/4})$ for all considered data sets, regardless of the degree of outlier contamination (see the last three rows of subfigures). This is arguably consistent with the results of Shao[18] on the OLS estimator, who showed that the predictive power of MSEP improves as n_v/n increases.

OLS Regression with Normal Residuals. Figure 3 explores an interesting question regarding the OLS estimator: Maybe there is nothing wrong with the OLS estimator as such but rather the MSE and MSEP measures should be replaced by more appropriate predictive and fitting statistics for OLS. To have a definitive answer, the OLS estimator was applied to the simulated companion data sets with the normal residuals. Figure 3 clearly demonstrated for the OLS estimator that even on the easiest possible for the estimator data sets, all but one of the commonly used predictive power and fitting ability statistics yielded very similar results, where the "-n" marker was added to the data set names to highlight their normalized residuals. The only noticeable improvement by reduction in the absolute value of ρ was obtained with MedAE and MedAEP, which are robust statistics. It is reasonable to assume that for any other residual distribution, OLS would perform even worse, because OLS is statistically the most efficient estimator in the presence of the normal residuals.

The CLTS and LTA estimators produced MCCV results (not shown) very similar to the OLS results, since the CLTS, LTA and OLS regressions of data sets with normally distributed residuals are virtually identical (recall the second column of subfigures in Figure 1).

In summary, the OLS regression does not exhibit any MCCV advantages over the robust regression even when applied to the normal residuals.

Robust MCCV. Arguably, the results presented so far indicated that the OLS estimator had no advantage over the robust regression for the data sets, which require the regression line to be fitted through the bulk of the data points. The same conclusion persists when OLS is compared against robust (CLTS, LTA, and LMedA) MLR estimators, see Figure 4, which displays the results of 1000 MCCV instances in each subfigure. Each estimator was assessed using its "native" calibration/fitting-ability statistic (e.g., MSE for OLS and TMSE for CLTS) and the corresponding validation/predictive-power statistic (e.g., MSEP for OLS and TMSEP for CLTS). The performance of the considered robust estimators was mixed: the robust correlation coefficient ρ (first line in each subfigure in Figure 4) between the corresponding fitting ability and predictive power statistics was the best (the least negative) for LTA on logBB-TPSA, LMedA on logHIA-ALOGP, logTox, CYGOB1, and GESELL, and CLTS on BODY-BRAIN. Note that the fourth line in each subfigure of Figure 4 reported the classical correlation coefficient r, which again did not favor a single

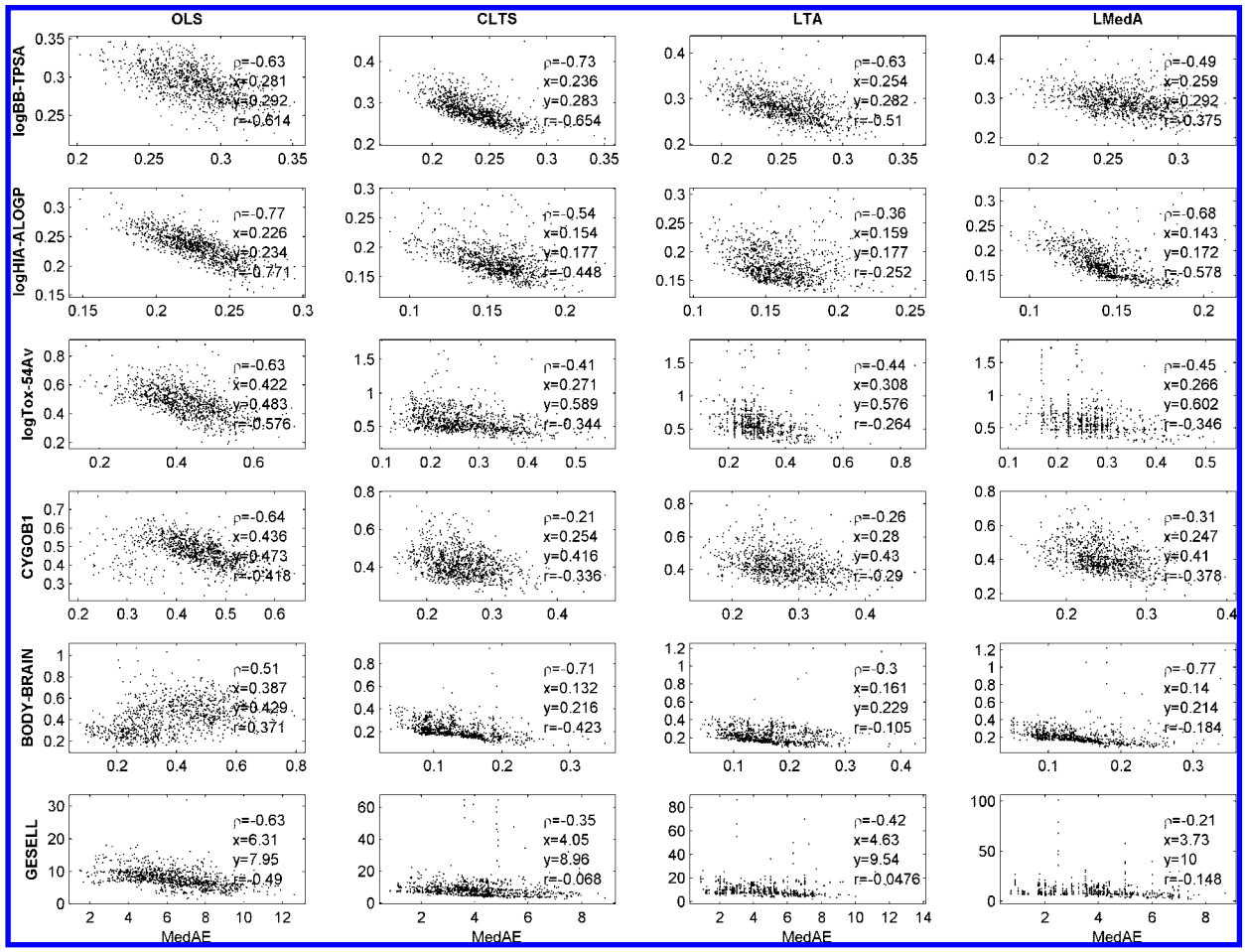

Figure 5. Same as in Figure 4 but for MedAEP as functions of MedAE obtained with the OLS, CLTS, LTA, and LMedA estimators.

robust estimator. However, without exception, every robust estimator performed better (by ρ and r) than the OLS estimator on every considered data set. The new robust CV procedure could be summarized in the following steps, which at present appear to be independent of outlier contamination, the number of predictors, underlying sample distribution, and sample size.

Calibration Step. A robust statistic of the model's fitting ability should be minimized when performing MLR of a calibration subset.

Validation Step. A robust statistic of the model's predictive power should be calculated from a validation subset which was completely excluded from the model's calibration.

Cross-Validation Procedure. The calibration and validation steps should be repeated a large number of times (not exceeding $N = n^2$ for MLR models) within the MCCV framework with $n_v \approx n_c \approx n/2$ to obtain convergent average value of the model's predictive power statistic (denoted CV-PP). The CV-PP statistic should be used as the primary quantitative characteristic for assessing the model's predictive power and for comparison between different models.

In this study, we were unable to select neither a single robust estimator nor the best robust predictive power statistic from the MCCV point of view as the best all-around recommendation. This was understandable because the individual performance of each of the robust estimators and statistics is highly depended on the data set. However the CLTS estimator stood out as a robust estimator that yielded convergent MLR expressions using the smallest number of iterations comparing to the LTA and LMedA estimators. Our recommendation choice for the predictive statistic was based on how easy the statistic could be understood and interpreted. The MedAEP and ρ^2 statistics appear to be the only choices because the trimmed predictive statistics (TMSE and TMAE) are likely to be confusing for most researchers who have not been actively working with the robust estimators, while MedSEP is essentially just the square of MedAEP. MedAEP is then averaged within MCCV obtaining the MedAEP_{cv} value, which could be used to assess the likely prediction error obtainable with the model. That is, half of the predictions are expected to fall into the $[\hat{y}_j - \text{MedAEP}_{cv}, \hat{y}_j + \text{MedAEP}_{cv}]$ range, where $j > n$ denotes the future untested compounds. By studying the residual distribution, further assessment could be made on the distribution of the predicted value, for example, estimating the percentage of values belonging to the $[\hat{y}_j - 2 \times \text{MedAEP}_{cv}, \hat{y}_j + 2 \times \text{MedAEP}_{cv}]$ range. The standard QSAR convention should always be checked as per the central premise of medicinal chemistry in that only structurally similar molecules are expected to exhibit similar biological activities. Therefore any QSAR model could only expect to predict within the chemical/structural space covered by the initial data set used to develop the model.[63]

The feasibility of our recommendations was verified by Figure 5, where the correlation between MedAE and MedAEP values was studied. The logBB-OLS subfigure (first row and column) in Figure 5 supported the results in Figure 3, where the use of a robust predictive statistic (see Validation

Step) could improve the predictive performance of even a nonrobust estimator such as OLS. However, this is situation should not be relied upon in general. The only reason the MCCV result with OLS is comparable by ρ to the results obtained with the robust estimators is that the logBB-TPSA data set is equally well described by both the robust and OLS estimators (see first row and column in Figure 1). A somewhat peculiar OLS result on the BODY-BRAIN data set (fifth row, first column in Figure 5) needs to be clarified, where the robust correlation between MedAE and MedAEP became positive, $\rho = 0.51$. The OLS estimator yielded the average MedAEP ($\text{MedAEP}_{cv} = 0.429$, the y value in the subfigure) much higher than the corresponding value from CLTS ($\text{MedAEP}_{cv} = 0.216$). Therefore regardless of the MedAE/MedAEP correlation, the CLTS estimator is expected to predict twice more accurately than OLS on the data set.

When the above recommendations were applied to the QSAR data sets, the following MCCV results were obtained:

$$\text{logBB: } \bar{\rho}_v^2 = 0.356(\rho_c^2 = 0.39),\ \text{MedAEP}_{cv} = 0.281(\text{MedAE} = 0.255) \quad (48)$$

$$\text{logHIA: } \bar{\rho}_v^2 = 0.014(0.092),\ \text{MedAEP}_{cv} = 0.178(0.163) \quad (49)$$

$$\text{logTox: } \bar{\rho}_v^2 = 0.596(0.708),\ \text{MedAEP}_{cv} = 0.599(0.335) \quad (50)$$

where the corresponding non-cross-validated values (i.e., from the calibration on the whole data set) are given in parentheses and where the robust coefficient of determination from the validation subsets ρ_v^2 was averaged during the MCCV resampling process obtaining $\bar{\rho}_v^2$. The interpretation of the above MCCV results is straightforward: while the robust MLR expressions should be presented for the whole available data set (eqs 44−46), their predictive power could not be assessed from the calibration-based fitting-ability statistics such as MedAE, but rather, the cross-validated predictive power statistics should be used. For example, we could only claim that the CLTS-MLR expression for log BB (eq 44) is accurate within ±0.281 log unit with 50% confidence (because of the median nature of the statistic).

In the case of the relatively large logBB data set, the difference between cross-validated predictive power ($\text{MedAEP}_{cv} = 0.281$) and non-cross-validated fitting ability (MedAE = 0.255) statistics is minor. The other two QSAR data sets provided two distinctly different examples. The log HIA data set also have quite similar $\text{MedAEP}_{cv} = 0.178$ and MedAE = 0.163. This could be reconciled with very low $\bar{\rho}_v^2 = 0.014$ by observing that the [0.28 − 0.178, 0.28 + 0.178] prediction range (0.28 is the intercept from eq 45) essentially covers the bulk of the available y values (see second row, first column in Figure 1); hence, the predicted log HIA (via eq 45) is basically a completely random function of ALOGP. The log Tox CLTS-MLR expression (eq 46) was also very interesting because it showed that a model with relatively high $\bar{\rho}_v^2 = 0.596$ may still have a large increase of the predictive error range during the MCCV procedure ($\text{MedAEP}_{cv} = 0.599$) comparing to the fitting MedAE = 0.335. This was mainly caused by the small number of data points. Therefore QSAR studies, which previously reported high R_{SSE}^2 on small data sets, very likely underestimated the predictive error range and thus overestimated predictive power.

CONCLUSIONS

In this study, we examined some currently popular statistics of predictive power and fitting ability of MLR models. It was demonstrated that the OLS-MLR estimator should always be replaced by a robust MLR estimator regardless of how contaminated with outliers the data sets were and regardless of the residuals' distributions. The robust CLTS-MLR estimator was identified as a possible candidate for a "standard" robust MLR estimator mainly because of its high numerical stability and fast convergence rate. The steps for the robust cross-validation procedure were summarized so that they could be potentially applied to most available MLR-QSAR models. It appears that most of our results/recommendations are directly applicable to nonlinear models, but this should be verified in future studies.

The open source of the QSAR-BENCH program allows replacement of the currently implemented algorithms with new algorithms with "minimal" effort, which could be valuable for testing of new ideas in the future.

ACKNOWLEDGMENT

We thank David Olive, Bruce Litow, and Nigel Sim for useful discussions, as well as two anonymous reviewers for constructive comments on earlier version of this manuscript.

Supporting Information Available: The logBB-TPSA, logHIA-ALOGP, logTox-X5Av CYGOB1, GESSELL, and BODY-BRAIN data sets, together with the corresponding simulated companion data sets, as well as Matlab code for all figures. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Konovalov, D. A.; Sim, N.; Deconinck, E.; Vander Heyden, Y.; Coomans, D. Statistical confidence for variable selection in QSAR models via Monte Carlo cross-validation. *J. Chem. Inf. Model.* **2008**, *48*, 370–383.
(2) Konovalov, D. A.; Coomans, D.; Deconinck, E.; Vander Heyden, Y. Benchmarking of QSAR models for blood−brain barrier permeation. *J. Chem. Inf. Model.* **2007**, *47*, 1648–1656.
(3) Llewellyn, L. E. Predictive toxinology: An initial foray using calculated molecular descriptors to describe toxicity using saxitoxins as a model. *Toxicon* **2007**, *50*, 901–913.
(4) Benigni, R.; Bossa, C. Predictivity of QSAR. *J. Chem. Inf. Model.* **2008**, *48*, 971−980.
(5) Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
(6) Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. B, Met.* **1974**, *36*, 111–147.
(7) Geisser, S. The predictive sample reuse method with applications. *J. Am. Stat. Assoc.* **1975**, *70*, 320–328.
(8) Mosier, C. I. I. Problems and designs of cross-validation. *Educ. Psychol. Meas.* **1951**, *11*, 5–11.
(9) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(10) Shao, J. Bootstrap model selection. *J. Am. Stat. Assoc.* **1996**, *91*, 655–665.
(11) Bassett, G., Jr.; Koenker, R. Asymptotic theory of least absolute error regression. *J. Am. Stat. Assoc.* **1978**, *73*, 618–622.
(12) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1–12.
(13) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; The MIT Press: Cambridge, MA, 2001.
(14) Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons: New York, 1998.
(15) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(16) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.

(17) Duffy, E. M.; Jorgensen, W. L. Prediction of properties from simulations: Free energies of solvation in hexadecane, octanol, and water. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.
(18) Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.
(19) 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.
(20) Zhang, P. Model selection via multifold cross-validation. *Ann. Stat.* **1993**, *21*, 299–313.
(21) Wold, S.; Eriksson, L. Statistical validation of QSAR results. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 309−318.
(22) Xu, Q. S.; Liang, Y. Z.; Du, Y. P. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J. Chemom.* **2004**, *18*, 112–120.
(23) Toropov, A. A.; Rasulev, B. F.; Leszczynski, J. QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: Comparative analysis by MLRA and optimal descriptors. *QSAR Comb. Sci.* **2007**, *26*, 686–693.
(24) Abraham, M. H.; Ibrahim, A.; Zhao, Y.; Acree, W. E. A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *J. Pharm. Sci.* **2006**, *95*, 2091–2100.
(25) Burman, P. A comparative-study of ordinary cross-validation, nu-fold cross-validation and the repeated learning-testing methods. *Biometrika* **1989**, *76*, 503–514.
(26) Croux, C.; Gallopoulos, E.; Van Aelst, S.; Zha, H. Machine learning and robust data mining. *Comput. Stat. Data Anal.* **2007**, *52*, 151–154.
(27) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression & Outlier Detection*; John Wiley & Sons: New York, 1987.
(28) Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J.; Stahel, W. A. *Robust Statistics: The Approach Based on Influence Functions*; John Wiley and Sons: New York, 1986.
(29) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
(30) Rousseeuw, P. J.; Van Driessen, K. Computing LTS regression for large data sets. *Data Mining Knowledge Discovery* **2006**, *12*, 29–45.
(31) Hawkins, D. M.; Olive, D. Applications and algorithms for least trimmed sum of absolute deviations regression. *Comput. Stat. Data Anal.* **1999**, *32*, 119–134.
(32) Hawkins, D. M.; Olive, D. J. Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm. *J. Am. Stat. Assoc.* **2002**, *97*, 136–148.
(33) Olive, D. J.; Hawkins, D. M. Behavior of elemental sets in regression. *Stat. Probability Lett.* **2007**, *77*, 621–624.
(34) Rousseeuw, P.; Daniels, B.; Leroy, A. Applying robust regression to insurance. *Insur. Math. Econ.* **1984**, *3*, 67–72.
(35) Olive, D. J.; Hawkins, D. M. High breakdown multivariate estimators. http://www.math.siu.edu/olive/preprints.htm (accessed March 31, 2008).
(36) Hossjer, O. Rank-based estimates in the linear-model with high breakdown point. *J. Am. Stat. Assoc.* **1994**, *89*, 149–167.
(37) Kvalseth, T. O. Cautionary note about R2. *Am. Stat.* **1985**, *39*, 279–285.
(38) Devlin, S. J.; Gnanadesikan, R.; Kettenring, J. R. Robust estimation and outlier detection with correlation-coefficients. *Biometrika* **1975**, *62*, 531–545.
(39) Hubert, M.; Verboven, S. A robust PCR method for high-dimensional regressors. *J. Chemom.* **2003**, *17*, 438–452.
(40) Shevlyakov, G. L. On robust estimation of a correlation coefficient. *J. Math. Sci.* **1997**, *83*, 434–438.
(41) Ruppert, D. Computing S estimators for regression and multivariate location/dispersion. *J. Comput. Graphics Stat.* **1992**, *1*, 253–270.
(42) Huber, P. J. Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Stat.* **1973**, *1*, 799–821.
(43) Guitton, A.; Symes, W. W. Robust inversion of seismic data using the Huber norm. *Geophysics* **2003**, *68*, 1310–1319.
(44) Holland, P. W.; Welsch, R. E. Robust regression using iteratively reweighted least-squares. *Commun. Stat. A-Theor.* **1977**, *6*, 813–827.
(45) Fox, J. *An R and S-PLUS Companion to Applied Regression*; Sage Publications, Inc.: Thousand Oaks, CA, 2002.
(46) Fox, J.http://socserv.mcmaster.ca/jfox/Books/Companion/appendix-robust-regression.pdf (accessed November 29, 2007).
(47) Ronchetti, E.; Field, C.; Blanchard, W. Robust linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1997**, *92*, 1017–1023.
(48) Kaznessis, Y. N.; Snow, M. E.; Blankley, C. J. Prediction of blood−brain partitioning using Monte Carlo simulations of molecules in water. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 697–708.
(49) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.
(50) E-DRAGON. *Dragon 5.4*; http://www.vcclab.org/lab/edragon/ (accessed December 4, 2007).
(51) Abraham, M. H.; Zhao, Y. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Reynolds, D. P.; Beck, G.; Sherborne, B.; Cooper, I. On the mechanism of human intestinal absorption. *Eur. J. Med. Chem.* **2002**, *37*, 595–605.
(52) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749–784.
(53) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762–3772.
(54) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Model.* **1989**, *29*, 163–172.
(55) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.; Radchenko, E.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory−Design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453–463.
(56) Tetko, I. V. Computing chemistry on the web. *Drug Discovery Today* **2005**, *10*, 1497–1500.
(57) VCCLAB. Virtual Computational Chemistry Laboratory, www.vcclab.org (accessed November 30, 2007).
(58) Weininger, D. Smiles, a chemical language and information-system 0.1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(59) E-DRAGON User Manual. http://michem.disat.unimib.it/chm/Help/edragon/index.html (accessed December 4, 2007).
(60) Kier, L. B.; Hall, L. H. Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.* **1981**, *70*, 583–589.
(61) Kier, L. B.; Hall, L. H. General definition of valence delta-values for molecular connectivity. *J. Pharm. Sci.* **1983**, *72*, 1170–1173.
(62) Wang, J. M.; Krudy, G.; Hou, T. J.; Zhang, W.; Holland, G.; Xu, X. J. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395–1404.
(63) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.

CI800209K