JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

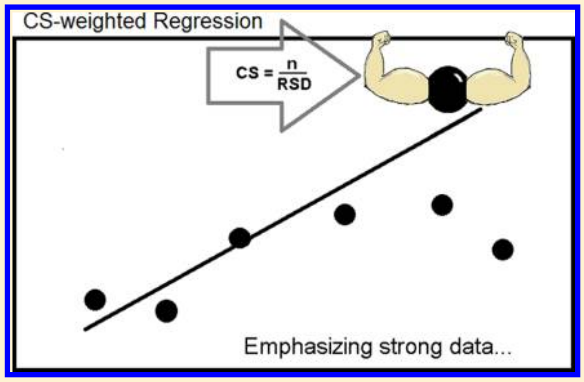
Data Quality in the Human and Environmental Health Sciences: Using Statistical Confidence Scoring to Improve QSAR/QSPR Modeling

Fabian P. Steinmetz, Judith C. Madden, and Mark T. D. Cronin*

School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, England

Supporting Information

ABSTRACT: A greater number of toxicity data are becoming publicly available allowing for in silico modeling. However, questions often arise as to how to incorporate data quality and how to deal with contradicting data if more than a single datum point is available for the same compound. In this study, two well-known and studied QSAR/QSPR models for skin permeability and aquatic toxicology have been investigated in the context of statistical data quality. In particular, the potential benefits of the incorporation of the statistical Confidence Scoring (CS) approach within modeling and validation. As a result, robust QSAR/QSPR models for the skin permeability coefficient and the toxicity of nonpolar narcotics to *Aliivibrio fischeri* assay were created. CS-weighted linear regression for training and CS-weighted root-mean-square error (RMSE) for validation were statistically superior compared to standard linear regression and standard RMSE. Strategies are proposed as to how to interpret data with high and low CS, as well as how to deal with large data sets containing multiple entries.



1. INTRODUCTION

The assessment of biological, and more specifically toxicological, data quality is crucial for many disciplines. Although the quality of data has no absolute definition, it is strongly associated with attributes such as validity, adequacy (i.e., fitness for purpose), reproducibility and reliability.[1] Confidence in the toxicological data, which may be derived in part at least from an assessment of data quality, is of great importance for regulatory bodies which have to make decisions on acceptable limits of chemicals relating to human and environmental exposure. Low, or poor, data quality may also affect the quality of computational models, such as quantitative structure−activity relationships (QSARs), grouping, and read-across, which are relevant both for risk assessment and regulatory decisions.[2−4]

In principle there are two general approaches to assess the quality of biological and toxicological data. The first is based on the assessment of the reported testing information alone. That means data quality is assessed by considering external factors, e.g. data and experimental reliability, completeness of documentation and adoption of protocols such as good laboratory practice (GLP). Schemes such as that developed by Klimisch[1] and its formalization into the ToxRTool (Toxicological data Reliability Assessment Tool) are well-known, established, and relatively accepted within the scientific community.[1,3] A second approach, where there are multiple and comparable data for the same compound in the same test, is to apply a statistical method. In this case, confidence scores (CS) can be calculated to emphasize data with a high weight of evidence, i.e. concordance between two or more independently conducted tests. The CS is the ratio of number of test values (n) and relative standard deviation (RSD) of test results, as defined in eq 1. Thus, if the same compound was tested independently with the same assay and the results were comparable, there will be a high CS for this compound and the associated experimental values.[5]

$$\mathrm{CS} = \frac{n}{\mathrm{RSD}} \tag{1}$$

Examples of calculations of CS are provided in Table 1 representing illustrative scenarios of increasing CS. Compound A is the default (and most common occurrence for a compound with a single experimental value), the CS is 1. Compound B has two relatively divergent data values, differing by an order of magnitude. Clearly there will be greater confidence for the toxicity value than for compound B, but the significant difference in the values introduces some uncertainty, raising CS marginally to 1.73—in this way there is slightly greater confidence associated with two relatively different values than a single value. More data points are considered for compounds C and D, with increasing precision of the data values. While compound C ($n = 4$) has more data than compound D ($n = 3$), the values are more divergent for C (represented by a higher RSD), thus the highest CS is calculated for compound D for

Received: May 19, 2015
Published: July 17, 2015

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746

Table 1. Four Examples of Compounds with Multiple Data in the Same Toxicity Test (EC_{50}), along with Statistical Criteria and CS[a]

compound	EC_{50} (mmol/L)	$\bar{x} \pm$ SD[b]	RSD[c]	n[d]	CS[e]
A	10	10 ± n/a	n/a	1	1[f]
B	1	5.50 ± 6.36	1.16	2	1.73
	10				
C	1	57.75 ± 43.05	0.75	4	5.37
	80				
	50				
	100				
D	1	1.47 ± 0.50	0.34	3	8.74
	2				
	1.4				

[a]See the Supporting Information for data. [b]Mean and standard deviation. [c]Relative standard deviation. [d]Number of data. [e]Confidence Score. [f]CS of a compound with n = 1 is defined as having 1 as the minimum value.

which there are three data points; all relatively consistent in the light of the experimental error that might be associated with an experimental test. As such, compound D has the highest CS value.

As there is growing interest in techniques such as read-across to fill data gaps for regulatory purposes, and there is increasing accessibility to toxicity data through resources such as the OECD QSAR Toolbox to perform read-across, there are more possibilities to apply approaches such as the confidence scoring to improve the robustness of modeling. In this study the relevance of the statistical CS approach has been assessed with regard to established QSARs for two end points, namely skin permeability coefficients and cytotoxicity for which large compilations of historical data are available.

1.1. Skin Permeability. There have been many efforts to develop quantitative structure−permeability relationship (QSPR) models to predict various measures of dermal absorption.[6−11] The most recognized and applied QSPR to predict the skin permeability coefficient (k_p) is that developed by Potts and Guy in 1992 eq 2.[9] They identified the molecular weight (MW), to account for the size of a permeant, and the logarithm of the octanol−water partition coefficient (log K_{OW}), as a descriptor for lipophilicity, as parameters to model k_p following an analysis based on the Flynn data compilation.[12] The mechanistic explanation is that small, lipophilic compounds pass through the *stratum corneum*, the most outer layer of the skin, more easily than larger, more hydrophilic compounds.[9,13]

$$\log k_p(\mathrm{cm/h}) = -2.7 + 0.71 \log K_{OW} - 0.0061\mathrm{MW} \quad (2)$$

Despite the significance of this model, the quality of data compiled by Flynn from the literature, and hence the robustness of the Potts and Guy QSPR, has been the subject of considerable debate.[14,15] More human in vitro k_p data have inevitably become available in the two and half decades since Flynn's seminal publication;[14,16−18] thus the QSPR can be reassessed and rebuilt with a greater consideration and understanding of data quality.

1.2. Aquatic Toxicology. There are thousands of publically available acute and chronic eco-toxicological data, and a significant proportion are compiled within the US Environmental Protection Agency's (US EPA's) ECOTOX database.[19] Of the ecotoxicological data, those for aquatic species are the most prevalent. Of these, the Microtox assay represents a commonly used and standardized acute aquatic toxicity test, based on the marine bacterium *Aliivibrio fischeri*, with a multitude of published data. When the photoluminescent bacteria are exposed to toxicants, the concentration is proportional to the inhibition of light intensity. The negative logarithm of the effective concentration causing 50% light reduction (EC_{50}) is expressed as the pT.[20] Extending the original compilation of Kaiser and Palabrica,[20] Steinmetz et al.[5] collected a large meta-data set with 1813 different values for Microtox toxicity. In order to create meaningful QSAR models in aquatic toxicology, there is an application of the well-established relationship between acute toxicity and hydrophobicity for compounds acting by the nonpolar narcosis mechanism of action.[21−24] Narcosis mechanisms of action, and nonpolar narcosis in particular, are considered to be as a result of membrane perturbation and that specific mechanisms toward endogenous proteins, receptor mediated effects, are not relevant.[25,26] This implies that the toxicity of compounds that are identified as being nonpolar narcotic can be well modeled by descriptors for hydrophobicity, e.g. log K_{OW}. Steinmetz et al.[5] identified a significant proportion of the Microtox toxicity compilation as being capable of acting by the nonpolar narcosis mechanism. In addition Steinmetz et al.[5] confirmed the findings of Cronin and Schultz[27] that for these compounds the standard exposure times (5, 15, and 30 min) had no significant effect on pT, thus enabling global log K_{OW}-derived models (including these three exposure times) to be developed for nonpolar narcotics. Consideration of data quality relating to the confidence associated with multiple data for the same chemical, showed that that toxicity data with certain CS thresholds led to more robust QSAR models.[5]

These two examples of historical data compilations are illustrative of the possibilities of applying confidence scoring metrics to historical compilations of toxicity information. There are many open-access resources such as ChEMBL,[28] PDSP,[29] ACToR,[30] eChemPortal,[31] TOXNET,[32] so the life sciences, and in particular toxicology, has to deal increasingly with large and complex data sets.[33] However, the task of assessing the toxicity data for quality, particularly when contradicting data are present, has not yet been accomplished. Any indication of the quality of data would be very helpful for purposes such as risk assessment, but more crucially for modeling including QSARs and read-across prediction.[3,5]

Therefore, the aim of this study was to investigate how using approaches for statistical data quality, i.e. CS, improve the development of QSAR/QSPR models. Specifically, the effect of directly incorporating the CS into the training and testing of the models was considered. To achieve this, the two end points described above were chosen for analysis, namely human in vitro skin permeability coefficients and the acute toxicity of compounds acting by a nonpolar narcotic mechanism of action to *A. fischeri*. The reasons for choosing these end points included the fact that there were many historical data of variable and unknown quality, many compounds had been tested multiple times (a prerequisite of applying the CS) and that there were simple, robust, and mechanistically interpretable QSAR models for them. Thus, for both data sets, QSARs were constructed with and without reference to the CS.

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746

2. METHODS

2.1. Data Harvest. In vitro skin permeability coefficients (k_p) were collected from the literature by compiling and subsequently merging four of the most comprehensive data sets of human skin k_p values.[14,16−18] All k_p values were converted to a standard unit (cm/h). Duplicate log k_p values (and those within ±0.01 cm/h) were removed as they are most likely to be derived from the same source. SMILES and InChIKey strings were obtained for each compound from the ChemSpider[34] database. The Flynn data set contained k_p values for 94 compounds; however, 11 compounds (all substituted steroids) could not be identified with ChemSpider[34] or ChemIDplus[35] and hence no SMILES were available to calculate descriptors. Since the structure of these compounds could not be completely verified they were excluded from subsequent analysis.

The Microtox data compilation from Steinmetz et al.[5] was used as the resource for the aquatic toxicology data set. This comprised 1227 compounds for which there were 1813 data points for 5, 15, and 30 min exposure. Where there were data for different time end points, the longest was taken. For modeling all exposure times were combined, since it has been demonstrated that this has no significant effect on the toxicity of nonpolar narcotics.[5,27] The EC_{50} values were considered in millimoles per liter and converted to pT. The SMILES and InChIKeys were obtained from ChemSpider.[34] The structures of all compounds were run through IDEAconsult's Toxtree v2.6.6[36] (mod. Verhaar) and nonpolar narcotics were identified as being Class 1 according to the Verhaar scheme.[21]

2.2. Descriptor Generation. Values of log K_{OW} and molecular weight (MW) were calculated for compounds in both data sets. The SMILES strings were used as the input format for all calculations. Values of log K_{OW} were calculated with KOWWIN v1.68 within EPI Suite 4.11 (estimated values exclusively).[37] MW data was calculated with the CDK node "molecular properties" within KNIME 2.9.[38]

2.3. Calculation of Confidence Scores (CS). Confidence scores were calculated for the compounds in both data sets with regard to their k_p and EC_{50} values, respectively. For compounds with more than a single experimental value, the arithmetic mean ($\overline{x}$), number (n), standard deviation (SD), and relative standard deviation (RSD) were calculated with reference to data in the units stated in section 2.1 and before logarithmic transformation. A confidence score (CS) was assigned to the arithmetic mean of the experimental values for each compound. Compounds with a single entry (n = 1) were assigned a confidence score of one (CS = 1). For $n > 1$, the CS was calculated as in eq 1.

2.4. Development of QSARs. Uni- and multivariate linear regression was performed on the data sets using R Studio 0.98.501.19.[39] Linear equations were generated and the following statistical, and other, criteria, were recorded: n (number of data points), S (standard error), R_{adj}^2 (coefficient of determination, adjusted for the number of degrees of freedom), t statistics for the descriptors, and F statistics for the equation. The regression analysis was performed to develop the QSARs for both data sets with and without weighting. Nonweighted regression analysis and weighted regression analysis was performed by applying CS values as weights in R using lm{stats}. Weighting in linear regression means that each datum point is associated with a weight. A high weight strengthens, and a low value weakens, the impact of the data point toward the linear regression. In this manner, data for compounds associated with a high confidence score would be more heavily weighted in the regression analysis than compounds with a lower confidence score. Comparison of the statistics of the weighted and unweighted regression analysis provides an indication of whether CS is able to improve the robustness of models.

2.5. Evaluation of the Predictivity of the QSARs/QSPRs. Statistical evaluation of the predictive capability of the CS-weighted QSAR and the CS-weighted QSPR was performed using 10-fold cross-validation, i.e. the compounds were ordered by k_p and pT, respectively, and every 10th compound was removed in turn leading to 10 training and validation sets. After applying the CS-weighted linear regression, the 10 data sets were investigated by the root-mean-square error (RMSE); predicted (f_i) versus experimental (y_i) values. Additionally the root-mean-square error adjusted for CS ($RMSE_{CS}$) was calculated eq 3. It is expected that during the validation process, the $RMSE_{CS}$, which incorporates CS-weighting, will be lower than the standard RMSE. As the residuals ($f_i - y_i$) of the compounds with low CS values are weakened and the residuals of high CS compounds are strengthened, the sum of (squared) errors of the $RMSE_{CS}$ should be reduced in comparison to the conventional RMSE. The R script for $RMSE_{CS}$ cross-validation and the equations are available in the Supporting Information.

$$RMSE_{CS} = \sqrt{\frac{\sum_i CS_i(f_i - y_i)^2}{\sum_i CS_i}} \qquad (3)$$

3. RESULTS

Names of compounds, their InChIKeys, their SMILES strings, and all k_p and pT values including references are available for the two data sets in the Supplorting Information. In addition the R script for $RMSE_{CS}$ cross-validation and a glossary of relevant statistical equations are also available in the Supplorting Information.

3.1. Data Harvest. The compilation of human in vitro k_p data resulted in 342 values for 226 different compounds. 55 of these compounds have more than a single k_p value. The log k_p values covered a broad range from −6.10 to 0.16. The structures included in the data set were diverse in terms of physicochemical properties and structure, e.g. solvents, alkaloids, steroids, sugars, nonsteroidal anti-inflammatory drugs, etc. The solvents, sugars, and steroids in particular had many multiple data points. Water, with 13 different data points, had the most k_p values. The range of CS values is from 1 (for single entries) to 76.8 for chlorphenamine (based on two data points). Illustrating the capability of the CS approach, two compounds have moderately high CS values: the synthetic opioid sufentanyl with a CS value of 9.97 (based on two data points) and the cytostatic drug 5-fluorouracil with a CS value of 5.00 (based on four data points).

From the complete data set of acute toxicity values to *A. fischeri*, comprising 1227 compounds, 203 were identified as potentially acting as nonpolar narcotics according to the Verhaar scheme as implemented in Toxtree v2.6.6.[36] A total of 418 different pT values were available for these compounds, with 71 of the 203 compounds having more than a single experimental value. pT values covered a broad range from −4.00 to 4.12. The structures included in the data set were conservative in their structural diversity as they had been

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746

selected to represent the nonpolar narcosis domain, including mainly solvents and medium- and long-chained alkanes, partly branched and halogenated, with only a few functional groups, such as hydroxyl- and amino-groups. The compounds investigated have a moderate spread of MW and log K_{OW} and can generally be regarded as lipophilic (cf. Table 2). The CS spread shows the diversity between high confidence compounds, such as methyl isobutyl ketone (CS of 205 with 3 data points) and acetone (CS of 43.7 with 14 entries) and the single entry low confidence compounds (defined as CS = 1).

Table 2. Ranges of Properties and CS for the Two Datasets Considered in the Analysis

	human in vitro skin permeability coefficients	pT of nonpolar narcotics to *A. fischeri*
MW (Da)	18.01 to 764.4	32.04 to 342.4
log K_{OW}	−6.76 to 8.39	−1.34 to 6.43
CS	1 to 76.8	1 to 205

3.2. Development of QSARs/QSPRs. QSAR/QSPR models were developed using linear regression with the experimental log k_p and pT as the dependent variables and log K_{OW} and MW (for k_p only) as descriptors. Linear regression analysis was performed on both data sets, the resultant QSPRs for skin permeability coefficients based on the Potts and Guy approach eq 4 (unweighted) and eq 5 (weighted), Figure 1, and the log K_{OW}-based QSARs for the acute toxicity of nonpolar narcotics to *A. fischeri* eq 6 (unweighted) and eq 7 (weighted), Figure 2, are reported below.

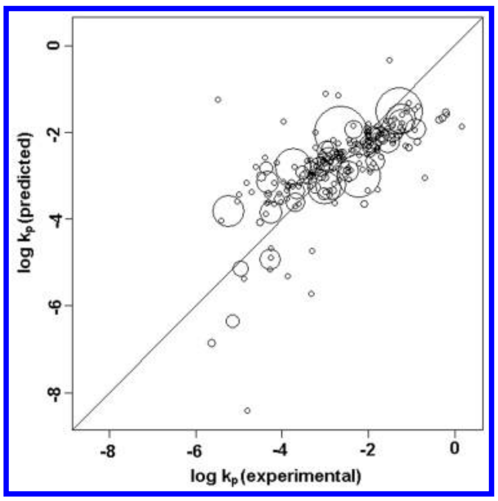


Figure 1. Experimental log k_p versus predicted log k_p from eq 5. The area of circles correspond to the CS value; the larger the CS, the greater the area of the circle. The solid line indicates a slope of unity and an intercept of zero.

3.2.1. QSPR: Modeling of Skin Permeability Coefficients. The unweighted QSPR for the data set of skin permeability coefficients, using the Potts and Guy approach, was:

$$\log k_p = -2.45 + 0.40 \log K_{OW} - 0.0045\text{MW} \quad (4)$$

$n = 226$, $S = 0.82$, $R_{adj}^2 = 0.48$, $t_{\log K_{OW}} = 13.3$, $t_{MW} = -8.97$, $F = 105$.

The reanalysis using CS-weighted k_p provided the following, similar, equation with improved statistical fit:

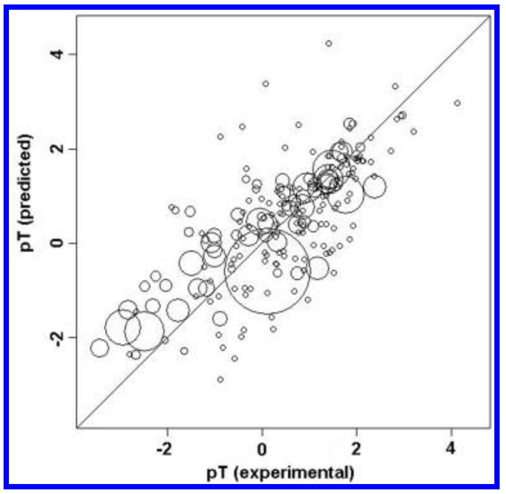


Figure 2. Measured pT versus pT predicted from eq 7. The area of circles corresponds to CS value; the larger the CS, the greater the area of the circle. The solid line indicates a slope of unity and an intercept of zero.

$$\log k_p = -2.51 + 0.50 \log K_{OW} - 0.0051\text{MW} \quad (5)$$

$n = 226$, $S = 1.39$, $R_{adj}^2 = 0.61$, $t_{\log K_{OW}} = 18.7$, $t_{MW} = -9.25$, $F = 177$.

Experimental k_p values are plotted against predicted values from eq 5 in Figure 1, demonstrating good overall predictivity. In particular, there is a good fit about the line of unity, with a significant trend for compounds with the highest CS (represented by larger circles) to be well predicted, and the significant outliers tending to be compounds with low CS, i.e. single data points.

The QSPR model represented by eq 5 was tested using 10-fold cross-validation. The statistical summary is presented in Table 3. Notably the RMSE_{CS} is lower than the RMSE.

3.2.2. QSAR: A. fischeri Nonpolar Narcosis. The unweighted QSAR for the nonpolar narcotics in the Microtox data set, using a log K_{OW}-based linear regression was

$$\text{pT} = -1.14 + 0.68 \log K_{OW} \quad (6)$$

$n = 203$, $S = 0.95$, $R_{adj}^2 = 0.50$, $t_{\log K_{OW}} = 14.3$, $F = 204$.

The reanalysis using CS-weighted pT provided the following equation with improved statistical fit:

$$\text{pT} = -1.67 + 0.92 \log K_{OW} \quad (7)$$

$n = 203$, $S = 1.77$, $R_{adj}^2 = 0.68$, $t_{\log K_{OW}} = 20.9$, $F = 478$.

Figure 2 demonstrates the relative predictivity of eq 7. There is a good fit about the line of unity, with a significant trend for compounds with the highest CS (represented by larger circles) to be well predicted, and the significant outliers tending to be compounds with low CS, i.e. single values.

The QSAR model eq 7 was assessed with 10-fold cross-validation. The summary of the statistics for eq 7 is presented in Table 4. The RMSE_{CS} is lower than the RMSE.

4. DISCUSSION

There are many future challenges in human and environmental health sciences which require the use of adequate and reliable data, these include toxicological risk assessment for occupational health and consumer goods. As the quality of toxicological data is variable and often not stated, practical

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746

Table 3. Intercept, Coefficients of Descriptors and Statistical Summary of 10-fold Cross-Validation Based on Equation 5 (Skin Permeability)

training				test	
intercept	log K_{OW}	MW	R_{adj}^2	RMSE	$RMSE_{CS}$
−2.51 ± 0.09	0.497 ± 0.026	−0.0051 ± 0.0004	0.61 ± 0.02	0.83 ± 0.21	0.79 ± 0.21

Table 4. Intercept, Coefficients of Descriptors and Statistical Summary of 10-fold Cross-Validation Based on Equation 7 (Aquatic Toxicology)

training			test	
intercept	log K_{OW}	R_{adj}^2	RMSE	$RMSE_{CS}$
−1.67 ± 0.14	0.92 ± 0.04	0.68 ± 0.03	0.99 ± 0.12	0.87 ± 0.13

and feasible methods to overcome this issue are crucial to many scientific and regulatory fields. Beside approaches such as Klimisch scoring,[1] we suggest a purely statistics-based method to support modeling approaches. It is difficult to determine the extent to which such a statistically driven approach could be used for regulatory purposes, but neglecting the information multiple data hold for the same substance is not recommended if such data are available.

The aim of this work was not to build new QSAR/QSPR models, but to make two existing models more robust using independent, heterogeneous data sets. The two QSARs and associated data sets chosen are well established. In this study the data sets have been extended by further data harvesting and collection. As part of the data collection activity, multiple data were compiled for the same chemical, thus allowing for the application of the CS approach to determine the reliability of the data. This approach has not been applied formally in the development of QSARs and there are no clear guidelines on how to develop QSARs when multiple data are available for the same chemicals (i.e., use of the mean, most conservative value, etc.). In addition, there appear to be few, if any, attempts to include information such as data quality as a metric or criterion for QSAR development, this being despite it being logical and acknowledged that data quality will affect the robustness of a QSAR.[40] It should also be noted that current means of documenting QSARs provide little opportunity for assessing the quality of data. Therefore, approaches that allow us to identify data quality quantitatively and without subjective bias are of value to develop in silico models.

Skin permeability is often assessed in vitro, but also some in vivo work is undertaken. In silico models are increasingly desirable in areas such as risk assessment where there is a dermal exposure (e.g., for cosmetics) and for assessing adverse effects to the skin, e.g. skin sensitization. Since the publication of the Flynn data,[12] there have been a number of QSAR analyses of skin permeability coefficients including refinements and extensions to the database.[13] The Potts and Guy approach,[9] based on fundamental and mechanistically comprehensible descriptors is one of the more commonly utilized QSAR modeling methodologies. This study has derived a Potts and Guy equation for a larger data set not only increasing the coverage of the model (i.e., greater chemical space) but also incorporating multiple data points for the same chemical and allowing for an assessment of quality through the application of CS. It is noted that published skin permeability coefficients are highly variable, due in no small part to high experimental error arising from the variable nature of the (human) skin utilized and test protocols, e.g. use of solvents, enhancers, finite doses, vehicles, solvents, etc.[14,15] As such, it is to be expected that models will not have a very significant statistical fit (i.e., a high R^2) and this is borne out by many of the published models,[9,14] indeed models with significant fit should be treated with some caution as they may be overfitted.

While high statistical fit was not achieved for the skin permeability QSPRs, the results show a significant relationship with log k_p and log K_{OW} and MW with both variables demonstrating high t-values. The new QSPR have moderately improved statistical fit as compared to that of Potts and Guy. It should be noted that some values within the Flynn data set were proven to be incorrect and would have increased the error in the Potts and Guy QSPR.[15] The novel QSPR model (cf. eq 5 and Figure 1) derived from the skin permeability data has some advantages over the original Potts and Guy[9] model: first of all robustness, due to model development incorporating statistical data quality (cf. Table 3); second a greater applicability domain due to implementing a data set with greater chemical diversity (in terms of properties and structure) than Flynn;[12] and third due to the usage of calculated log K_{OW} (whereas the original model used measured values which are more difficult to obtain consistently). Nevertheless the differences between Potts and Guy's eq 2 and eq 5 are only marginal. It is recognized that there are many limitations to this use of this model. For example it does not predict the effects of mixtures and formulations on the penetration of single compounds, which could be of great importance for risk assessment of products and dermal drug delivery.[41] However, the QSPR approach allows for a "relative" estimation of skin permeability which may be useful to rank compounds, or identify compounds with a high probability of dermal absorption and hence prioritise such compounds in the risk assessment process (e.g., for skin sensitization).

The assessment of effects of chemicals to the bacterium *A. fischeri* (or the Microtox test) is one of the more rapid, cheaper and fundamental measurements of cytotoxicity. Data from the Microtox test show good correlation with higher species, especially for compounds acting by nonspecific mechanisms of action such as nonpolar narcosis.[42] Thus, if a compound can be identified as being a nonpolar narcotic, Microtox data may, if used appropriately and with caution, add further to the weight of evidence associated with a prediction. It is very well established that there is a strong relationship between hydrophobicity, as described by log K_{OW}, and nonpolar narcosis for many species.[43,44] This study has expanded the number of chemicals with data within nonpolar narcosis domain for *A. fischeri*, hence expanding the chemical space and extended a previous study.[5] It is of no surprise that the nonpolar narcosis data for *A. fischeri* (Microtox) are significantly correlated with log K_{OW}, even if some historical data are obviously of quite

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746

poor quality.[5] The QSAR eq 7 is similar to earlier published aquatic toxicology QSAR models, i.e. toxicity is increasing linearly with lipophilicity.[5,22−24,43]

Consideration of the QSARs developed in this study shows an improvement in the models when utilizing CS-weighted regression. The improvement is both the statistical fit but also the slope for log K_{OW} which approaches one when employing CS-weighting, i.e. from 0.68 to 0.90 (cf. eq 6 to 7). A slope of one is the theoretical optimum which is commonly associated with models for simple unicellular organisms, i.e. the absorption of the compound alone directly into the cellular membrane is responsible for narcosis; whereas in higher organisms, other factors such as distribution and clearance become important. The improvements following the application of CS are consistent with the notion that some historical data are of poor quality[45] and demonstrates the utility of an approach such as this when generalistic QSARs are being developed for data sets from various sources and of unknown quality. The importance of the compounds with high CS values can be seen in Figure 2, when considering that all large CS-circles are close to the line of best prediction. The quantity of data and the incorporation of statistical data quality make a robust equation with an extensive applicability domain—for nonpolar narcotics. Clearly this approach could be extended to other data compilations for aquatic acute toxicity.[46]

The identification of compounds acting by the nonpolar narcotic mechanism of action is essential to the development of the models. Various approaches have been applied to identify mechanisms of action including analysis of molecular descriptor space,[47] multivariate analysis of mode and mechanism of action space,[48] and definition of molecular fragments[26] as well as the Verhaar classification scheme that was applied in this study due to its ease of use following coding in the ToxTree software. Due to this definition of the nonpolar narcosis domain in the ToxTree software, there appear to be a number of anomalies. For example, aflatoxins are identified by the ToxTree software as being Verhaar Class 1 compounds (nonpolar narcotic) but, in reality, they are potent, specifically acting toxins[5] and therefore do not act as nonpolar narcotics, e.g. aflatoxin B2 has $pT_{experimental}$ = 1.17 (CS = 15.4) whereas eq 7 calculates $pT_{predicted}$ = 0.54. This emphasizes that continual development is required of decision criteria presented in approaches such as the Verhaar scheme as new knowledge and understanding becomes apparent.

Overall for both data sets, applying CS as a weighting tool improves the training and validation of the QSAR/QSPR models. The improvements are demonstrated as increases in R^2 eq 4 to 5 and eq 6 to 7 as a direct result of CS-weighting. Whereas increasing t and F values show improvements in the models as a result of weighting by CS, the S value does not incorporate weights and so only indicates absolute, unweighted error thus it actually increases when the nonweighted regression is compared to the weighted regression. Generally the higher the CS for the data associated with a compound, the greater the evidence is, in terms of similar results for that compound (cf. Figures 1 and 2). In the validation process, the $RMSE_{CS}$, which incorporates CS-weighting, is lower than the standard RMSE. As residues $(f_i - y_i)$ of low CS compounds are weakened and residues of high CS compounds are strengthened, the sum of (squared) errors of the $RMSE_{CS}$ becomes lower than in the conventional RMSE. Therefore, this approach could be used even for the validation of models where any metric could be applied to imply confidence, i.e. without calculating CS. For example a reversed Klimisch score (4 as the most reliable; 1 the least) could be used as a weight similar to the fuzzy logic approach of Yang et al.[49] In the context of validation these weights then determine to what extent residues should have impact on the RMSE.

The CS-weighting approach, whether in model development or validation, is limited by the presence of multiple entries for one compound. Thus, if multiple values are available for the data set, more robust models may potentially be built.[5] This robustness and the associated confidence are helpful in reducing uncertainty and hence increasing acceptance for regulatory decisions. For example in the context of REACH, there is a demand for robust QSAR models to support the toxicological assessment of chemicals. The approach described herein could thus be used to support read-across and QSAR-based predictions.[50,51]

5. CONCLUSIONS

The assessment of data quality is not trivial. This study has shown that CS provides a means of assessing confidence in data when there are more than a single datum point. The CS scores can be applied to develop QSAR models through the use weighted regression, as demonstrated in this study for historical data compilations with known variability in the quality of the data. Additionally cross-validation with $RMSE_{CS}$ provides a measure of the robustness of an equation utilizing metrics (here CS) for weighting.

■ ASSOCIATED CONTENT

Supporting Information

Microtox and skin permeability data, including statistics glossary. R-Script for $RMSE_{CS}$ cross-validation. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00294.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: M.T.Cronin@ljmu.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007−2013) COSMOS Project under grant agreement no. 266835 and Cosmetics Europe.

■ ABBREVIATION

CS = Confidence Score
EC_{50} = concentration (mmol/L) causing 50% of the stated effect
f = predicted value
F = F-value (cf. linear regression)
K_{OW} = octanol−water partition coefficient
k_p = skin permeability coefficient
n = number of data points/test values
InChIKey = international chemical identifier key
MW = molecular weight (Da)
pT = negative decadic logarithm of EC_{50} for toxicity
QSAR = quantitative structure−activity relationship
QSPR = quantitative structure−permeability relationship

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746

R_{adj}^2 = coefficient of determination adjusted for degrees of freedom
RMSE = root mean square error
$RMSE_{CS}$ = CS-adjusted RMSE
RSD = relative standard deviation (also known as coefficient of variation)
S = standard error (cf. linear regression)
SD = standard deviation
SMILES = simplified molecular-input line-entry system
t = t-value (cf. linear regression)
$\bar{x}$ = arithmetic mean
y = experimental value

■ REFERENCES

(1) Klimisch, H.-J.; Andreae, M.; Tillmann, U. A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data. *Regul. Toxicol. Pharmacol.* **1997**, *25*, 1−5.
(2) Péry, A. R. R.; Schüürmann, G.; Ciffroy, P.; Faust, M.; Backhaus, T.; Aicher, L.; Mombelli, E.; Tebby, C.; Cronin, M. T. D.; Tissot, S.; Andres, S.; Brignon, J. M.; Frewer, L.; Georgiou, S.; Mattas, K.; Vergnaud, J. C.; Peijnenburg, W.; Capri, E.; Marchis, A.; Wilks, M. F. Perspectives for Integrating Human and Environmental Risk Assessment and Synergies with Socioeconomic Analysis. *Sci. Total Environ.* **2013**, *456*, 307−316.
(3) Przybylak, K. R.; Madden, J. C.; Cronin, M. T. D.; Hewitt, M. Assessing Toxicological Data Quality: Basic Principles, Existing Schemes and Current Limitations. *SAR QSAR Environ. Res.* **2012**, *23*, 435−459.
(4) Nendza, M.; Aldenberg, T.; Benfenati, E.; Benigni, R.; Cronin, M. T. D.; Escher, S.; Fernández, A.; Gabbert, S.; Giralt, F.; Hewitt, M.; Hrovat, M.; Jeram, S.; Kroese, D.; Madden, J. C.; Mangelsdorf, I.; Rallo, R.; Roncaglioni, A.; Rorije, E.; Segner, H.; Simon-Hettich, B.; Vemeire, T. Data Quality Assessment for *In Silico* Methods: A Survey of Approaches and Needs. In *In silico toxicology: principles and applications*, first ed.; Cronin, M. T. D., Madden, J. C., Eds.; Royal Society of Chemistry Publishing: Cambridge, UK, 2010; pp 59−69.
(5) Steinmetz, F. P.; Enoch, S. J.; Madden, J. C.; Nelms, M. D.; Rodriguez-Sanchez, N.; Rowe, P. H.; Wen, Y.; Cronin, M. T. D. Methods for Assigning Confidence to Toxicity Data with Multiple Values − Identifying Experimental Outliers. *Sci. Total Environ.* **2014**, *482*, 358−365.
(6) Khajeh, A.; Modarress, H. Linear and Nonlinear Quantitative Structure-Property Relationship Modelling of Skin Permeability. *SAR QSAR Environ. Res.* **2014**, *25*, 35−50.
(7) Dancik, Y.; Miller, M. A.; Jaworska, J.; Kasting, G. B. Design and Performance of a Spreadsheet-based Model for Estimating Bioavailability of Chemicals from Dermal Exposure. *Adv. Drug Delivery Rev.* **2013**, *65*, 221−236.
(8) Magnusson, B. M.; Anissimov, Y. G.; Cross, S. E.; Roberts, M. S. Molecular Size as the Main Determinant of Solute Maximum Flux across the Skin. *J. Invest. Dermatol.* **2004**, *122*, 993−999.
(9) Potts, R. O.; Guy, R. H. Predicting Skin Permeability. *Pharm. Res.* **1992**, *9*, 663−669.
(10) Abraham, M. H.; Martins, F.; Mitchell, R. C. Algorithms for Skin Permeability Using Hydrogen Bond Descriptors: The Problem of Steroids. *J. Pharm. Pharmacol.* **1997**, *49*, 858−865.
(11) Scheuplein, R. J.; Blank, I. H. Permeability of the Skin. *Physiol. Rev.* **1971**, *51*, 702−747.
(12) Flynn, G. L. Physicochemical Determinants of Skin Absorption. In *Principles of Route-to-Route Extrapolation for Risk Assessment*, first ed.; Gerrity, T. R., Henry, C. J., Eds.; Elsevier: New York, USA, 1990; pp 93−127.
(13) Mitragotri, S.; Anissimov, Y. G.; Bunge, A. L.; Frasch, H. F.; Guy, R. H.; Hadgraft, J.; et al. Mathematical Models of Skin Permeability: An Overview. *Int. J. Pharm.* **2011**, *418*, 115−129.
(14) Moss, G. P.; Cronin, M. T. D. Quantitative Structure − Permeability Relationships for Percutaneous Absorption: Re-Analysis of Steroid Data. *Int. J. Pharm.* **2002**, *238*, 105−109.
(15) Johnson, M. E.; Blankschtein, D.; Langer, R. Permeation of Steroids through Human Skin. *J. Pharm. Sci.* **1995**, *84*, 1144−1146.
(16) ten Berge, W. Homepage of Wil ten Berge. http://home.wxs.nl/~wtberge/skinperm2013a.zip (accessed March 1, 2014).
(17) Chen, L.; Han, L.; Lian, G. Recent Advances in Predicting Skin Permeability of Hydrophilic Solutes. *Adv. Drug Delivery Rev.* **2013**, *65*, 295−305.
(18) Chauhan, P.; Shakya, M. Role of Physicochemical Properties in the Estimation of Skin Permeability: *In Vitro* Data Assessment by Partial Least-Squares Regression. *SAR QSAR Environ. Res.* **2010**, *21*, 481−494.
(19) US EPA. ECOTOX Database. http://cfpub.epa.gov/ecotox/ (accessed February 20, 2015).
(20) Kaiser, K. L. E.; Palabrica, V. S. Photobacterium phosphoreum Toxicity Data Index. *Water Poll. Res. J. Can.* **1991**, *26*, 361−431.
(21) Verhaar, H. J. M.; van Leeuwen, C. J.; Hermens, J. L. M. Classifying Environmental Pollutants. 1: Structure-Activity Relationships for Prediction of Aquatic Toxicity. *Chemosphere* **1992**, *25*, 471−491.
(22) Cronin, M. T. D.; Schultz, T. W. Structure-Toxicity Relationships for Three Mechanisms of Action of Toxicity to *Vibrio Fischeri*. *Ecotoxicol. Environ. Saf.* **1998**, *39*, 65−69.
(23) Zhao, Y. H.; Ji, G. D.; Cronin, M. T. D.; Dearden, J. C. QSAR Study of the Toxicity of Benzoic Acids to *Vibrio fischeri*, *Daphnia magna* and Carp. *Sci. Total Environ.* **1998**, *216*, 205−215.
(24) Zhao, Y. H.; Cronin, M. T. D.; Dearden, J. C. Quantitative Structure-Activity Relationships of Chemicals Acting by Non-Polar Narcosis − Theoretical Considerations. *Quant. Struct.-Act. Relat.* **1998**, *17*, 131−138.
(25) van Wezel, A. P.; Opperhuizen, A. Narcosis due to Environmental Pollutants in Aquatic Organisms: Residue-based Toxicity, Mechanisms, and Membrane Burdens. *Crit. Rev. Toxicol.* **1995**, *25*, 255−279.
(26) Ellison, C. M.; Cronin, M. T. D.; Madden, J. C.; Schultz, T. W. Definition of the Structural Domain of the Baseline Non-Polar Narcosis Model for *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* **2008**, *19*, 751−783.
(27) Cronin, M. T. D.; Schultz, T. W. Validation of *Vibrio fischeri* Acute Toxicity Data: Mechanism of Action-based QSARs for Non-Polar Narcotics and Polar Narcotic Phenols. *Sci. Total Environ.* **1997**, *204*, 75−88.
(28) ChEMBL. https://www.ebi.ac.uk/chembl/ (accessed February 11, 2015).
(29) PDSP. http://pdsp.med.unc.edu/pdsp.php (accessed February 11, 2015).
(30) US EPA. ACToR. http://www.epa.gov/actor/ (accessed February 20, 2015).
(31) OECD. eChemPortal. http://www.echemportal.org/ (accessed February 20, 2015).
(32) US NIH. TOXNET. http://toxnet.nlm.nih.gov/ (accessed February 20, 2015).
(33) Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays to Identify Potential Toxicants. *Chem. Res. Toxicol.* **2014**, *27*, 1643−1651.
(34) Royal Society of Chemistry. ChemSpider. http://www.chemspider.com/ (accessed September 1, 2014).
(35) NIH. ChemIDplus. http://chem.sis.nlm.nih.gov/chemidplus/ (accessed September 1, 2014).
(36) IDEAconsult. Toxtree v2.6.6. http://toxtree.sourceforge.net/ (accessed December 3, 2014).
(37) US EPA. EPI Suite 4.11. http://www.epa.gov/opptintr/exposure/pubs/episuite.htm (accessed August 20, 2014).
(38) KNIME 2.9. https://www.knime.org/ (accessed August 20, 2014).
(39) The R Project. R Studio 0.98.501.19. http://www.r-project.org/ (accessed August 20, 2014).

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746

(40) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 125−134.
(41) Samaras, E. G.; Riviere, J. E.; Ghafourian, T. The Effect of Formulations and Experimental Conditions on *In Vitro* Human Skin Permeation − Data from Updated EDETOX Database. *Int. J. Pharm.* **2012**, *434*, 280−291.
(42) Cronin, M. T. D.; Dearden, J. C.; Dobbs, A. J. QSAR Studies of Comparative Toxicity in Aquatic Organisms. *Sci. Total Environ.* **1991**, *109-110*, 431−439.
(43) Könemann, H. Quantitative Structure-Activity Relationships in Fish Toxicity Studies. Part 1: Relationship for 50 Industrial Pollutants. *Toxicology* **1981**, *19*, 209−221.
(44) Verhaar, H. J. M.; Urrestarazu Ramos, E.; Hermens, J. L. M. Classifying Environmental Pollutants. 2: Separation of Class 1 (Baseline Toxicity) and Class 2 (Polar Narcosis) Based on Chemical Descriptors. *J. Chemom.* **1996**, *10*, 149−162.
(45) Cronin, M. T. D.; Schultz, T. W. Structure-Toxicity Relationships for Phenols to *Tetrahymena pyriformis*. *Chemosphere* **1996**, *32*, 1453−1468.
(46) Martin, T. M.; Young, D. M.; Lilavois, C. R.; Barron, M. G. Comparison of Global and Mode of Action-based Models for Aquatic Toxicity. *SAR QSAR Environ. Res.* **2015**, *26*, 245−262.
(47) Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Identification of Mechanisms of Toxic Action of Phenols to *Tetrahymena pyriformis* from Molecular Descriptors. In *Quantitative structure-activity relationships in environmental sciences—VII*, first ed.; Chen, F., Schuurmann, G., Eds.; SETAC Press: Pensacola, USA, 1997; pp 329−342.
(48) Aptula, A. O.; Netzeva, T. I.; Valkova, I. V.; Cronin, M. T. D.; Schultz, T. W.; Kühne, R.; Schüürmann, G. Multivariate Discrimination between Modes of Toxic Action of Phenols. *Quant. Struct.-Act. Relat.* **2002**, *21*, 12−22.
(49) Yang, L.; Neagu, D.; Cronin, M. T. D.; Hewitt, M.; Enoch, S. J.; Madden, J. C.; Przybylak, K. Towards a Fuzzy Expert System on Toxicological Data Quality Assessment. *Mol. Inf.* **2013**, *32*, 65−78.
(50) Patlewicz, G.; Ball, N.; Becker, R. A.; Booth, E. D.; Cronin, M. T. D.; Kroese, D.; Steup, D.; van Ravenzwaay, B.; Hartung, T. Read-Across Approaches − Misconceptions, Promises and Challenges ahead. *ALTEX* **2014**, *31*, 387−396.
(51) Cronin, M. T. D. Evaluation of Categories and Read-Across for Toxicity Prediction Allowing for Regulatory Acceptance. In *Chemical Toxicity Prediction: Category Formation and Read-Across*, first ed.; Cronin, M. T. D., Madden, J. C., Enoch, S. J., Roberts, D. W., Eds.; The Royal Society of Chemistry: Cambridge, UK, 2013; pp 155−167.

DOI: 10.1021/acs.jcim.5b00294
J. Chem. Inf. Model. 2015, 55, 1739−1746