Struct Chem (2011) 22:795–804
DOI 10.1007/s11224-011-9757-4

ORIGINAL RESEARCH

Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models

Tomasz Puzyn · Aleksandra Mostrag-Szlichtyng · Agnieszka Gajewicz · Michał Skrzyński · Andrew P. Worth

Received: 13 January 2011 / Accepted: 7 February 2011 / Published online: 20 February 2011

Abstract The study was aimed at investigating how the method of splitting data into a training set and a test set influences the external predictivity of quantitative structure–activity and/or structure–property relationships (QSAR/QSPR) models. Six models of good quality were collected from the literature and then redeveloped and validated on the basis of five alternative splitting algorithms, namely: (i) a commonly used algorithm ('*Z*:1'), in which every *z*th (e.g. third) from the compounds sorted ascending (according to the response values, **y**) is selected into the test set; (ii–iv) three variations of the Kennard–Stone algorithm and (v) the duplex algorithm. The external validation statistics reported for each model served as a basis for the final comparison. We demonstrated that the splitting techniques utilizing the values of molecular descriptors alone (**X**) or in combination with the model response (**y**) always lead to the development of the models yielding better external predictivity in comparison with the models designed with methodologies based on the **y** values only. Moreover, we showed that the external validation coefficient (Q_{EXT}^2) is more sensitive to the splitting technique than the root-mean-square error of prediction ($RMSE_P$). This difference becomes especially important when the test set is relatively small (between 5 and 10 compounds). In the case of the models trained/validated with a small number of compounds, it is strongly recommended that both statistics (Q_{EXT}^2 and $RMSE_P$) be taken into account for the external predictivity evaluation.

Keywords Data splitting · External validation · QSAR · QSPR · Predictivity · Kennard–Stone · Duplex · Model reproducibility

Introduction

The practical usefulness of every quantitative structure–activity and/or structure–property relationships (QSAR/QSPR) model depends on its realistic predictivity (i.e. the ability to accurately predict certain activity/property for the chemical compounds that have not contributed to the model's development). Data splitting can be considered as a validation technique, based on the division of the input data into a training set and a test set. The model is developed and internally validated employing the training set, while its predictive power is assessed on the basis of differences between the predicted and experimental values (residuals) determined for a sufficient number of representative test set compounds. The latter procedure is called 'external validation'. Only properly trained and validated models are able to provide reliable predictions for novel compounds [1–5].

Data splitting performed at the initial stage of the QSAR/QSPR development is particularly significant, as it determines, which data are utilized to train (fit) the model, and which are employed for its external validation. The quest to

Electronic supplementary material The online version of this article (doi:10.1007/s11224-011-9757-4) contains supplementary material, which is available to authorized users.

T. Puzyn (✉) · A. Mostrag-Szlichtyng · A. Gajewicz · M. Skrzyński
Faculty of Chemistry, Laboratory of Environmental Chemometrics, University of Gdansk, ul. Sobieskiego 18/19, 80-952 Gdansk, Poland
e-mail: t.puzyn@qsar.eu.org

A. P. Worth
European Commission Joint Research Centre, Institute for Health and Consumer Protection, Via E. Fermi, 21027 Ispra, VA, Italy

find the most appropriate methodology for selecting training and test set compounds has led to active investigations in this area. A vast range of recently published contributions focused on the importance of data splitting, for example [6–9], highlight two major conditions that should be met: (i) representivity of both training and test sets and (ii) sufficient diversity of the training set. However, no model, even when properly validated and yielding "good" values of validation statistics, is able to provide reliable predictions for the entire universe of chemicals. The model usually works much better for the compounds falling inside its applicability domain (typically defined by structural/mechanistic similarity) and the range of activity/property values within the training set [10]. Hence, in the ideal modelling case, chemical structures and the predicted response values for training and test sets should be possibly similar—the representative objects in the training set should be close to the objects in the test set and vice versa [11]. In other words, the training and test sets should scatter over the whole range of the considered space, defined by the descriptors of molecular structure (**X**) and the response (**y**) values [12].

In practice, several algorithms are employed to split the input data. The most common ones are based on the endpoint (**y**) values only (e.g. the repeated test set technique, random selection or activity sampling) [13–16], while more sophisticated techniques take into account also the values of molecular descriptors (**X**) (e.g. maximum dissimilarity method, the Kennard–Stone algorithm, the duplex algorithm, Kohonen's self-organising maps, D-optimal design or sphere exclusion) [3–5, 17–25]. Endpoint-value-based methods of data splitting generate even distributions of compounds along with the endpoint values in both created sets. However, there is a danger that the application of such algorithms may be associated with significant loss of information, as the resulting training sets do not necessarily represent the entire descriptor space of the input data. Consequently, the test set compounds may be distant from those included in the training set. In contrast, algorithms in which **X** values contribute to the data splitting are more likely to generate representative sets consisting of compounds evenly distributed within the chemical space ranged by values of both **y** vector and **X** matrix. Such an approach should ensure the closeness between test and training set compounds [26]. Although opinions have been expressed in the academic literature, no firm and practical recommendations related to dataset splitting have been available so far in any of the official guidelines for QSAR/QSPR modellers.

In the present research, we focused on the influence of data splitting on the external predictivity of QSAR/QSPR models. By comparing a series of models redeveloped with use of different splitting schemes (**y**-based, **X**-based, or **y**- and **X**-based) and particular splitting techniques, we have tried to define some general recommendations for QSAR/QSPR practitioners based on the trends observed.

Materials and methods

Six case study models considered to be of high quality were selected from the available literature, and then they were redeveloped and validated on the basis of five alternative training/test sets splitting algorithms, namely: (i) a commonly used **y**-based algorithm we call '*Z*:1', in which the compounds are sorted in ascending order, according to the values of the response (**y**), and then every *Z*th (e.g. third) object is selected into the test set, while the remaining compounds form the training set; (ii–iv) three variations of the Kennard–Stone algorithm (v) the duplex algorithm. The external validation statistics reported for each model served as a basis for the final comparison of the investigated methodologies.

Case study models selection

Six QSAR/QSPR models, published in peer-reviewed journals, were chosen as the case studies. From a large number of published models, we selected only those of "good quality", developed and documented according to the Organisation for Economic Co-operation and Development (OECD) principles for the validation of (Q)SAR/(Q)SPR models [26]. Thus, we considered only models that were internally/externally validated, yielding good statistics for goodness-of-fit, robustness and predictivity and having a well-defined applicability domain. It is emphasised that the purpose of this exercise was not to correct or criticise any of the existing models, but only to use them as illustrative examples for expressing the relationships between their predictive performance and methodology of the training/test sets design.

Reproducibility of the original modelling procedures, appropriately documented by the models' developers, was a crucial criterion for the case study selection. The reproducibility of a model itself is a very general concept. In practice, it depends on two main factors. The first concerns the availability of original data used for the model development and validation. Neither the model nor the training/test set should be proprietary, which means that the values of the dependent (**y**, response) and all independent (**X**, molecular descriptors) variables for each compound used in the model development should be disclosed. The second factor concerns the mathematical approach to the modelling itself. For the sake of reproducibility, models based on linear relationships, developed with more transparent techniques, such as (multiple) linear regression ((M)LR),

would be more desirable. The MLR methodology is extensively described in numerous papers, e.g. [16, 27, 28]. In general, QSAR/QSPR equations developed with MLR consist of a relatively small number of independent variables and, as such, can be more readily interpreted. Moreover, the MLR modelling technique can be relatively easily repeated by other authors, also with software tools other than those originally used. Hence, in this study, we focused on MLR-derived QSAR/QSPR models.

Availability of the data and transparency of the mathematical algorithm are two necessary conditions, but they are not always sufficient to ensure the reproducibility of a QSAR/QSPR model. Another important factor is the adequate and transparent documentation of the applied modelling procedure (i.e. a step-by-step protocol). In order to find as well-documented models as possible, we screened the QSAR Model Reporting Format (QMRF) Database developed by the European Commission's Joint Research Centre which is freely accessible online at http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=QRF [29]. The QMRF Database is an inventory gathering information on several published QSAR/QSPR models, harmonised and structured according to the OECD (Q)SAR/(Q)SPR validation principles [26]. Many of the QMRF reports are supplemented with attachments (e.g. xls files or structure data (.sd) files) providing complete information on the training and test sets compounds (their structures, **y** and **X** values, etc.). The QMRFs provide transparent descriptions of subsequent steps of the modelling procedures used, as well as information on the statistical performance of the models and their applicability domains. For the purpose of the present investigation we have screened 56 documents published in the QMRF Database according to the modelling algorithm (MLR).

Our intention was to compare the impact of different data splitting algorithms on the predictive abilities of the models in the broadest possible sense. As such, we considered either global or local MLR models of various sizes, covering diverse toxicological/environmental endpoints, as well as predicting physical/chemical properties. The only limitation was the practical possibility of reproducing the original model development.

Initially, six QSAR/QSPR models were selected as the case studies (Table 1). Two QSPR models (model **1** and model **2**) originated from our previous work [16]. Four QSARs were selected from the JRC QMRF Database. They were related to toxicokinetic (model **3**), toxicological (model **4**) and eco-toxicological (model **5** and model **6**) endpoints [29–35]. These models were well documented and providing all the necessary information on the training/test set compounds (we extracted the endpoint and descriptors values from .sd files attached to the individual QMRF reports) [32–35].

Before the final selection of the pre-selected case study models we verified that reproducing the original calculations would lead to the same equation coefficients and validation statistics as provided by the original authors. Each of the tentatively selected models was re-developed and re-evaluated in MATLAB v. R2010b [36], employing the original training and test sets. Data from the training sets were used to determine appropriate statistics describing goodness-of-fit, robustness and internal predictivity, namely: the squared correlation coefficient (R^2); the root-mean-square error of calibration ($\mathrm{RMSE_C}$); the leave-one-out cross-validation coefficient ($Q_{\mathrm{CV}}{}^2$) and the root-mean-square error of the leave-one-out cross-validation ($\mathrm{RMSE_{CV}}$). Commonly used mathematical formulations of these statistics can be found elsewhere [16, 27]. The statistics obtained by using the test sets (the external validation coefficient, $Q_{\mathrm{EXT}}{}^2$ and the root-mean-square error of prediction, $\mathrm{RMSE_P}$) were utilized to verify the external predictivity of the models and had a crucial meaning in our comparisons. These parameters were calculated as follows:

$$\mathrm{RMSE_P} = \sqrt{\frac{\sum_{i=1}^{n_\mathrm{v}} (y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}})^2}{n_\mathrm{v}}} \quad (1)$$

$$Q_{\mathrm{EXT}}{}^2 = 1 - \frac{\sum_{i=1}^{n_\mathrm{v}} (y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}})^2}{\sum_{i=1}^{n_\mathrm{v}} (y_i^{\mathrm{obs}} - y_{\mathrm{obs}}^{\mathrm{mean}})^2} \quad (2)$$

where n_v is the number of the test set compounds, y_i^{obs} is the experimental response value for *i*th compound from the test set, y_i^{pred} is the model's response value for *i*th compound from the test set and $y_{\mathrm{obs}}^{\mathrm{mean}}$ is the mean value of the endpoint (**y**).

Re-development of the case study models with various data splitting methods

The essential step of the present investigation was a multiple re-development of each selected case study model with different training sets, designed by employing various data splitting algorithms, namely: (i) *Z*:1 algorithm; (ii) Kennard–Stone algorithm performed on the matrix of molecular descriptors (**X** matrix); (iii) Kennard–Stone algorithm performed on a matrix, in which the molecular descriptors (**X**) were augmented by an additional column including the response values (**y**); (iv) Kennard–Stone algorithm performed on a similar matrix to that in (iii), but this time the additional **y** vector (column) has been replicated *k* times to enhance the influence of the response on the splitting results; (v) Duplex algorithm performed on the descriptor matrix (**X**) only.

Z:1 is the most commonly applied algorithm in QSAR/QSPR studies, mainly due to its simplicity. It does not utilize the values of molecular descriptors—the splitting

Table 1 QSAR/QSPR models selected for the study

Model no.	Type	Description of the endpoint	Equation	Reference
1	QSPR	QSPR for log K_{OW}	Log $K_{OW} = -0.3587$	[16]
			−0.122 [Dipole moment]	
			+0.025 [Solvent accessible surface]	
2	QSPR	QSPR for log K_{OA}	Log $K_{OA} = 7.311$	[16]
			+0.741 [Energy of HOMO]	
			+0.286 [Mean polarizability]	
3	QSAR	QSAR for blood/brain barrier partitioning	Log P (Blood–brain barrier) = −3.03	[29]
			+0.398 [Number of halogenide groups]	[30]
			−25.7 [HA dependent HDCA-2/SQRT(TMSA) (Zefirov) (all)]	[31]
			+ 0.324 [HOMO-1 energy (AM1)]	[32]
			−0.00625 [WFOSA atomic charge (AM1) weighted FOSA]	
			−9.990 [Max net atomic charge (Zefirov) for N atoms]	
4	QSAR	QSAR for relative binding affinity to oestrogen receptor	Log(ER − RBA) = −19.12	[29]
			+2.11 [Average information content (order 1)]	[30]
			+0.80 [Number of rings]	[31]
			+7.33 [Relative ALFA polarizability (DIP) (AM1)]	[35]
			−13.83 [Max net atomic charge (Zefirov) for O atoms]	
			+0.84 [Log P]	
5	QSAR	QSAR for acute toxicity to algae	Log(1/EC50) = 1.656	[29]
			−9.940 [Relative number of rings]	[30]
			−8.465E−002 [WPSA3 weighted PPSA (PPSA3 * TMSA/1000) (AM1)]	[31]
			+1.111E−003 [Gravitation index (all atom pairs) (AM1)]	[33]
			−2.543 [Polarity parameter (AM1)/square distance]	
6	QSAR	QSAR for acute toxicity to fathead minnow	Log(LC50) = 0.97	[29]
			−3.48 [Average bond order (AM1)]	[30]
			−0.32 [Highest total interaction (AM1)]	[31]
			−2.21E−003 [LPSA low polarity (AM1) part of SASA]	[34]
			−0.16 [Count of H-acceptor sites (AM1) (all)]	
			−0.64 [Log P]	

procedure involves the **y** (response) values only. As mentioned above, test compounds are selected in a systematic way based on their sorted response values. Such an approach produces two sets that accurately represent the data [16, 30, 31].

In contrast, the Kennard–Stone algorithm takes into account only the values of the molecular descriptors (**X**) [20, 28]. Initially, the most representative, 'central' compound is selected into the training set. The algorithm searches for a single compound having the values of all descriptors closest to their mean values calculated for the whole group of compounds. Then, a defined number (sufficiently large and determined by the developer) of the most dissimilar objects (chemicals) is also introduced into the training set. The similarity measure, in this case, is the squared Euclidean distance between particular objects in the multidimensional space in which each descriptor defines a single dimension. Thus, the most dissimilar compounds are the most distant ones (i.e. characterized by the maximal values of the squared Euclidean distance). The remaining compounds are incorporated into the test set.

Usually, the Kennard–Stone algorithm is performed only on the **X** matrix. However, in our contribution we tested also two variations of this methodology. In the first one (**Xy**), we added the response vector (**y**) as an additional column to the matrix of k descriptors (**X**). In the second modification (**X**k**y**), we added the response vector k times (k was equal to the number of descriptors in the **X** matrix), in order to enhance the impact of the response values on the data splitting results.

These ways of data splitting according to the Kennard–Stone algorithm and its modifications should lead to the

formation of two representative sets including all types of chemical structures. The two modifications of the Kennard–Stone method, in principle, should ensure that the training set compounds are distributed evenly within not only in the space defined by the descriptors (**X**), but also by the response values (**y**). As such, the condition of closeness between test and training set compounds in both aspects (**X** and **y**) should be satisfied [26].

The duplex algorithm utilizes **X** values only. Its sequential methodology is based on maximizing the Euclidean distances between the newly selected compounds and the compounds already selected. In the first step, the two most distant (i.e. most dissimilar) objects are picked up and incorporated into the training set. From the remaining compounds, the two most dissimilar ones are included in the test set. Then, from the remaining objects, the one which is furthest away from those previously selected for the training set is labelled as a test set compound. Analogously, subsequent training set compound is selected. The two procedures are repeated alternately, until a sufficient number (indicated by the developer) of training set compounds is chosen. Such a procedure leads to the formation of two balanced sets, consisting of objects uniformly distributed within the whole descriptors (**X**) space [26, 37].

All calculations within this step of the study were performed in MATLAB v. R2010b [36] with external codes (m-files) for Kennard–Stone-based and duplex-based data splitting [12]. Individual models were developed by means of the MLR method [16]. Since the impact of training/test set size on the predictivity of models was not investigated here, when re-splitting and re-developing the models, we kept the ratio of training-to-test compounds proposed by authors of the original contributions. Each newly designed training set was used for the QSAR/QSPR model development, while each test set, for its external validation. The complete set of statistical parameters was calculated for each model (i.e. R^2, $Q_{CV}{}^2$, $RMSE_C$, $RMSE_{CV}$). However, for the purposes of this study, we focused mainly on the ones related to the external predictivity, namely: $RMSE_P$ and $Q_{EXT}{}^2$ (Eqs. 1, 2).

Results and discussion

A positive outcome of the "reproducibility check" confirmed the consistency between the original and repeated calculations. An overview of the selected case study models, as well as the original (if available) and calculated (by the present authors) values of external validation statistics are provided in Table 2.

Each of the six selected QSARs/QSPRs was originally developed and validated with data sets split with the classical *Z*:1 algorithm. Since we re-developed the original models with four additional splitting algorithms, this yielded a set of 30 models in total to be compared. As mentioned above, we applied additional splitting algorithms while keeping the original ratio of training-to-test set compounds. It should be highlighted that models **1**, **2** and **6** had relatively large test sets (Table 2), whereas test sets of models **3**–**5** were very small. This allowed us to observe additional, data splitting-related trends.

Interestingly, the external validation statistics of every original model could be improved by applying alternative data splitting methodologies, which are based not only on the response (**y**) values, but also on the molecular descriptors (**X**) (Table 3). We observed that such algorithms contribute to the formation of more balanced and homogeneous training and test sets. As such, the training/test set compounds were situated close to each other within the considered chemical space, and the condition for both sets to be representative was fulfilled. In the majority of cases, the best results (lowest $RMSE_P$ values) were observed for algorithms that regard the information of both **y** and **X** values, namely the Kennard–Stone **Xy** and the Kennard–Stone **X***k***y**. However, when considering $Q_{EXT}{}^2$ as the measure of external predictivity, the best results (highest values) were obtained for those methods that take into account only information on the structural variance of the compounds (**X**) (duplex-based or Kennard–Stone-**X**-based data splitting). Indeed, the information on **X** seems to have more influence on the appropriate splitting than the response values (**y**).

Some additional observations can also be made. The external validation statistics of the models, when analyzed individually, exhibit different sensitivities to the on the replacement of the **y**-based data splitting methodology with the alternative ones (those taking into consideration the values of **X**). Differences in sensitivity can be observed, when analyzing the values of $\Delta RMSE_P$ and $\Delta Q_{EXT}{}^2$ (Table 3) that quantitatively describe the improvement of external predictivity of the models. Moreover, in general, the sensitivity of $Q_{EXT}{}^2$ is much more dependent on size of the test set than the sensitivity of $RMSE_P$. This becomes evident, when comparing variances for models **1**, **2** and **6** (having large test sets) with models **3**–**5** (having small test sets) (Table 4).

The observations above can be explained by the following reasoning. The predictivity of particular QSAR/QSPR model is strongly driven by the distribution of the training set compounds in the chemical space defined by the **X** values on one hand, and by the **y** values on the other one. Ideally, the training set compounds should be evenly scattered over the whole space. Under such a condition, the model is well trained and the predictions of the response (**y**) are satisfactory. However, the ability to correctly

Table 2 Validation statistics of the models selected for the study

Model no.	Number of compounds		R^2		$RMSE_C$		Q_{CV}^2		$RMSE_{CV}$		Q_{Ext}^2		$RMSE_P$	
	Training (n)	Test (n_t)	Orig.	Rep.	Orig.	Rep.	Orig.	Rep.	Orig.	Rep.	Orig.	Rep.	Orig.	Rep.
1	178	59	0.920	0.920	0.315	0.315	0.918	0.918	0.321	0.321	0.924	0.924	0.302	0.302
2	77	26	0.972	0.972	0.320	0.320	0.970	0.970	0.333	0.333	0.961	0.961	0.376	0.376
3	54	6	0.753	0.752	NA	0.360	0.680	0.679	NA	0.410	NA	0.721	NA	0.377
4	62	6	0.80	0.797	NA	0.808	0.73	0.726	NA	0.939	NA	0.422	NA	0.846
5	40	5	0.924	0.924	NA	0.282	0.881	0.881	NA	0.352	NA	0.629	NA	0.515
6	423	46	0.76	0.763	NA	0.680	0.75	0.754	NA	0.693	NA	0.699	NA	0.696

Orig. value provided by authors of the original contribution, *Rep*. value reproduced within our study, *NA* not applicable (not provided in the original work)

Table 3 The impact of investigated data splitting algorithms on the statistical external validation parameters for selected case study models

Model no.	Splitting algorithm	$RMSE_P$	$\Delta RMSE_P$	Q_{EXT}^2	ΔQ_{EXT}^2	Size of the test set
1	*Z*:1 (**y**)	0.302		0.924		Large
	Kennard–Stone (**X**)	0.296	0.007	0.926	−0.001	
	Kennard–Stone (**Xy**)	0.239	0.065	0.884	0.041	
	Kennard–Stone (**X*k*y**)	0.224	0.079	0.894	0.031	
	Duplex (**X**)	0.344	−0.041	0.900	0.025	
2	*Z*:1 (**y**)	0.376		0.961		Large
	Kennard–Stone (**X**)	0.369	0.007	0.881	0.081	
	Kennard–Stone (**Xy**)	0.303	0.073	0.930	0.031	
	Kennard–Stone (**X*k*y**)	0.303	0.074	0.954	0.007	
	Duplex (**X**)	0.322	0.055	0.979	−0.018	
3	*Z*:1 (**y**)	0.377		0.721		Small
	Kennard–Stone (**X**)	0.360	0.017	0.640	0.081	
	Kennard–Stone (**Xy**)	0.264	0.113	0.619	0.102	
	Kennard–Stone (**X*k*y**)	0.297	0.081	0.468	0.253	
	Duplex (**X**)	0.230	0.147	0.812	−0.091	
4	*Z*:1 (**y**)	0.846		0.422		Small
	Kennard–Stone (**X**)	0.914	−0.068	0.738	−0.316	
	Kennard–Stone (**Xy**)	0.518	0.328	0.544	−0.122	
	Kennard–Stone (**X*k*y**)	0.604	0.242	0.333	0.089	
	Duplex (**X**)	0.865	−0.019	0.805	−0.383	
5	*Z*:1 (**y**)	0.515		0.629		Small
	Kennard–Stone (**X**)	0.368	0.147	−0.114	0.743	
	Kennard–Stone (**Xy**)	0.368	0.147	−0.114	0.743	
	Kennard–Stone (**X*k*y**)	0.368	0.147	−0.114	0.743	
	Duplex (**X**)	0.425	0.090	0.873	−0.244	
6	Z:1 (**y**)	0.696		0.699		Large
	Kennard–Stone (**X**)	0.579	0.117	0.696	0.003	
	Kennard–Stone (**Xy**)	0.498	0.199	0.776	−0.077	
	Kennard–Stone (**X*k*y**)	0.486	0.210	0.767	−0.068	
	Duplex (**X**)	0.894	−0.198	0.728	−0.029	

$\Delta RMSE_P$—the decrease of the $RMSE_P$ value, calculated as the difference between the $RMSE_P$ obtained for *Z*:1 algorithm based model and the $RMSE_P$ obtained for a model designed with an alternative splitting methodology; ΔQ_{EXT}^2—the decrease of the Q_{EXT}^2 value, calculated as the difference between the Q_{EXT}^2 obtained for the model based on *Z*:1 algorithm and the Q_{EXT}^2 obtained for a model designed with an alternative splitting methodology

Table 4 Variances (s^2) of the external validation parameters in comparison with size of the test set

Model no.	$RMSE_P$	Q_{Ext}^2	Size of the test set
1	0.002	0.000	Large
2	0.001	0.001	Large
3	0.004	0.016	Small
4	0.031	0.041	Small
5	0.004	0.232	Small
6	0.029	0.001	Large

predict the response for novel compounds (not used for training the model) must be verified with use of the external test set. To be representative, the test set should also evenly cover the whole chemical space. In practice, this condition can be fulfilled only for sufficiently large test sets. For small test sets there is a very high probability that all their constituents will be unevenly distributed within the considered chemical space. In the extreme situation, the test set compounds form a small cluster situated in only one region of the chemical space covered by the training set. Such a test set is neither representative nor well balanced and contributes to the superfluous results of the external validation.

Model **5** can serve as an illustrative example of such a situation. Unexpected values of its statistical validation parameters ($Q_{EXT}^2 < 0$! for the Kennard–Stone-based data splitting) reflect the unusual localisation of test set compounds in the corresponding chemical space. The test set covers only the lowest values of **y**, thus the whole space of the response/descriptors is not appropriately represented. The statistical external validation, when performed only on the basis of Q_{EXT}^2, reveals that such a model is completely externally unpredictive. Clearly, this is not entirely true. When the $RMSE_P$ value is considered, the lowest values of this statistic are observed after applying the Kennard–Stone-based splitting techniques. These contradictory results can be explained, when looking at the mathematical formulas of Q_{EXT}^2 and $RMSE_P$ (Eqs. 1 and 2). Both statistics are calculated from the sum of squared residual values (i.e. differences between the observed and predicted values of **y**). $RMSE_P$ is simply the root of the average squared residual in the test set. The calculation of the external validation coefficient, however, is more complicated. The value of Q_{EXT}^2 is the difference between 1 and the ratio of the sum of squared residuals (PRESS) to the sum of squared deviations of particular observed values of **y** from the average y (TSS). In consequence, the influence of one or more unexpected predictions (unusually high residuals) on Q_{EXT}^2 is stronger than on $RMSE_P$, since in the second case the (squared) residuals are averaged and the root value is calculated at the end. In the case of Q_{EXT}^2, we are operating on the sum of squared values and there is no averaging. Thus, when one or two residuals are extremely high, and the test set is small, it is possible that the ratio PRESS/TSS is higher than 1. In such a case, the calculated external validation coefficient would be negative. This also explains why we have observed a strong influence of the test set size on the Q_{EXT}^2 values.

This case study reflects that, particularly for the models evaluated on the basis of very small test sets, the conclusions on the final external predictivity should not be drawn on the basis of one statistical parameter alone, but should be related also to the other relevant measures. The small-size test set models are much more sensitive to the choice of data splitting methodology which means that the results obtained might be less robust and meaningful than those for the large-size test set ones. Consequently the decision concerning the data splitting algorithm must be made with particular care.

When discussing the most appropriate choice of splitting algorithm, a significant comment concerning the reliability of our results related to **X**-based techniques must be added. Actually, the truly external validation could be performed only for the models **1**–**2**, since both were developed on the basis of the molecular descriptors selected a priori, on the mechanistic basis only. In case of the remaining models, the reasonable amount of independent variables were selected by the authors of the original contributions from broad "pools" of more than 1000 tentatively calculated descriptors. The selection of descriptors was performed on a statistical basis, for instance by using a genetic algorithm. This leads to a lack of reproducibility in the modelling procedure. In the majority of cases the complete information on the descriptors forming the large "pool" was not available in the original publication. Available data sets contained only the values of the final variables selected on a statistical basis and incorporated into the model equation. Therefore, in the case of models **3**–**6** we were only able to perform the alternative data splitting and calculate the validation statistics on the basis of the pre-selected independent (**X**) variables. As a consequence, since the compounds labelled as 'test' in our study had been previously involved in the variable selection, the validation procedures with such test sets were not strictly 'external'. In a real situation, when **X** variables need to be selected from a large pool of calculated descriptors, test compounds should never be involved in the variable selection process. In contrast, it is highly probable that, when the splitting with **X**-based algorithms (i.e. Kennard–Stone or duplex) is performed on the whole pool of 1000 or more descriptors (before the final selection of variables), neither the training nor test set would be sufficiently representative (not evenly distributed in the space of the finally selected variables). This is a serious limitation of such splitting algorithms.

Our results are highly consistent with previous contributions to the area and supplement some of the findings. Leonard and Roy [9] demonstrated that the application of the K-means clustering technique (utilizing the descriptor **X** values) for input data splitting leads to much better external validation statistics of the resulting models than the random splitting and/or splitting methods that are based only on the response (**y**) values (i.e. 'activity ranges algorithm'). Moreover, they highlighted that the splitting procedure should take into account the proximity of training and test compounds; both training and test sets should consist of the molecules representing the whole multidimensional descriptor space.

The authors [9] noticed high values of the external validation coefficient (Q_{EXT}^2) irrespective of the size of the data set. It is worth noting, however, that three models studied by Leonard and Roy [9] were externally validated with test sets of moderate or even large size ($n_t = 9$, 14 and 22), in comparison with the relatively small test sets of models **3**, **4** and **5** investigated in our study. Thus, Leonard and Roy did not have a chance to observe the strong influence of test set size on the Q_{EXT}^2 values, as elaborated here. Their results are in agreement with our suggestion that the influence becomes important when the test set is very small (contains fewer than 10 test compounds).

Leonard and Roy [9] also state that "The size of the test set is an important factor in identifying the predictive potential of the data set, so one may intend to explore the optimum size of the test set in relation to the size of the training set". The same research group investigated also the impact of the training set size on the predictive ability of QSAR models [38]. They concluded that the optimum size of the training set depends on many factors: the particular data set, number and types of descriptors, as well as statistical analysis being used; no general rule can be formulated. In the context of our results, this leads to the recommendation that the experimental input data set should be large enough to ensure an appropriate number of training compounds (dependent on the factors mentioned above), and at least 10 test compounds. When the number of test compounds is much less than 10, an external validation is still possible, but one should expect the Q_{EXT}^2 to be strongly dependent on the splitting technique.

Gramatica [6] performed a broad evaluation of other statistical approaches for the validation of QSAR/QSPR models. As the part of this project, she compared three data splitting algorithms, namely: (i) a D-optimal experimental design, (ii) the Kohonen Artificial Neural Network (K-ANN) and (iii) a random splitting. The conclusions were similar to those from our study: models with small test sets were found to be more sensitive to the data splitting methodology than models validated with the test sets containing many compounds. This confirms the importance of selecting the most appropriate splitting technique, whenever a QSAR/QSPR model is developed and validated with a small set of data, for instance in case of local models for particular congeneric groups of Persistent Organic Pollutants [39, 40].

Conclusions

In the present study we illustrated the impact of data splitting methodologies on the external predictivity of QSAR/QSPR models. We demonstrated that, although the results varied slightly for the selected models, it was possible to make some generalizations and identify several common trends.

The results of external validation are strongly dependent on the composition of the training and test sets. The application of splitting techniques that utilize the values of molecular descriptors alone (**X**) or in combination with the model response (**y**) always lead to the development of the models yielding better external predictivity in comparison with the models designed with methodologies based on the **y** values only.

In case of the models trained and validated with a very small number of compounds, the splitting methodology might influence the external validation results. We recommend that, since Q_{EXT}^2 seems to be more sensitive to the splitting technique than $RMSE_P$ when the test set is small (contains between 5 and 10 compounds), both statistics should be taken into account when evaluating the external predictivity of such models.

Whenever the model input variables are selected by a statistical approach (e.g. with a genetic algorithm) from the large pool of calculated descriptors, **y**-based splitting techniques should be preferred to ensure the possibility of performing external validation and the best predictive ability of the final QSAR/QSPR.

In our contribution we selected the most commonly used methodologies of data splitting, other than those previously evaluated by Leonard and Roy [9] and Gramatica [6]. However, taking into account the strong research needs for developing practical guidance of QSAR/QSPR, further investigations should also include more sophisticated resampling methods, i.e. bootstrapping [41, 42].

Acknowledgments The authors thank the editors for rapidly considering our submission and the anonymous reviewer for valuable comments, which helped to improve scientific quality of this contribution. T.P. thanks the Foundation for Polish Science for granting him with a fellowship and a research grant in frame of the HOMING Program supported by Norwegian Financial Mechanism and EEA Financial Mechanism in Poland. This work was supported by the Polish Ministry of Science and Higher Education Grant No. DS/8430-4-0171-11.

References

1. Golbraikh A, Tropsha A (2002) Beware of q2!. J Mol Graph Model 20:269–276
2. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comp Aided Mol Des 16:357–369
3. Gramatica P, Pilutti P, Papa E (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. J Chem Inf Comput Sci 44:1794–1802
4. Gramatica P, Papa E (2005) An update of the BCF QSAR model based on theoretical molecular descriptors. QSAR Comb Sci 24:953–960
5. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs and theoretical descriptors for the modelling of the aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). J Chem Inf Model 45:1256–1266
6. Gramatica (2004) Evaluation of different statistical approaches for the validation of quantitative structure–activity relationships. JRC Contract ECVA-CCR.496576-Z. http://ecb.jrc.ec.europa.eu/qsar/information-sources/
7. Roy PP, Paul S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive QSAR models. Molecules 14:1660–1701
8. Roy PP, Paul S, Mitra I, Roy K (2010) On two novel parameters for validation of predictive QSAR models–correction. Molecules 15:604–605
9. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 25(3):235–251
10. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, van de Sandt JJM, Tong W, Veith G, Yang C (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The Report and Recommendations of ECVAM Workshop 52. ATLA 33:155–173
11. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–77
12. Daszykowski M, Walczak B, Massart DL (2002) Representative subset selection. Anal Chim Acta 468(1):91–103
13. Boggia R, Forina M, Fossa P, Mosti L (1997) Chemometric study and validation strategies in the structure-activity relationship of new cardiotonic agents. QSAR 16:201–213
14. Yasri A, Hartsough D (2001) Toward an optimal procedure for variable selection and QSAR model building. J Chem Inf Comput Sci 41:1218–1227
15. Kauffman GW, Jurs PC (2001) QSAR and k-nearest neigbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. J Chem Inf Comput Sci 41:1553–1560
16. Puzyn T, Suzuki N, Haranczyk M (2008) How do the partitioning properties of polyhalogenated POPs change when chlorine is replaced with bromine? Environ Sci Technol 42(14):5189–5195
17. Potter T, Matter H (1998) Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. Med Chem 41:478–488
18. Taylor R (1995) Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. J Chem Inf Comput Sci 35:59–67
19. Bourguignon, de Aguiar PF, Khots MS, Massart DL (1994) Optimization in irregularly shaped regions: pH and solvent strength in reversed phase high-performance liquid chromatography separations. Anal Chem 66:893–904
20. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11:137–148
21. Hudson BD, Hyde MR, Rahr E, Wood J, Osman J (1996) Parameter based methods for compounds selection from chemical databases. QSAR 15:285–289
22. Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropscha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comp Aided Mol Des 17:241–253
23. Snarey M, Terrett NK, Willett P, Wilton DJ (1997) Comparison of algorithms for dissimilarity-based compound selection. J Mol Graph Model 15:373–385
24. Nilakatan R, Bauman N, Haraki KS (1997) Database diversity assessment: new ideas, concepts and tools. J Comp Aided Mol Des 11:447–452
25. Gobbi A, Lee ML (2003) Database DISE: directed sphere exclusion. J Chem Inf Comput Sci 43:317–323
26. OECD (2007) Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models, Paris
27. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26(5):694–701
28. Mostrąg A, Puzyn T, Haranczyk M (2010) Modeling the overall persistence and environmental mobility of sulfur-containing polychlorinated organic compounds. Environ Sci Pollut Res 17:470–477
29. QSAR Model Reporting Format (QMRF) Database developed by the Joint Research Centre and accessible online at http://ecb.jrc.ec.europa.eu/qsar/qsar-tools/index.php?c=QRF. Accessed Jan 2011
30. Karelson M, Dobchev D, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D, Karelson G (2008) Correlation of blood-brain penetration and human serum albumin binding with theoretical descriptors. ARKIVOC 16:38–60
31. Karelson M, Karelson G, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D, Dobchev D (2009) QSAR study of pharmacological permeabilities. ARKIVOC 2:218–238
32. The The JRC QMRF Q2-10-25-184 "QSAR for blood-brain barrier (BBB) partitioning". http://qsardb.jrc.ec.europa.eu/qmrf/search_catalogs.jsp?id=184&idstructure=. Accessed Jan 2011
33. The JRC QMRF Q8-10-27-209 "QSAR for acute toxicity to algae". http://qsardb.jrc.ec.europa.eu/qmrf/search_catalogs.jsp?id=209&idstructure=. Accessed Jan 2011
34. The JRC QMRF Q2-10-14-174 "QSAR for acute toxicity to fathead minnow". http://qsardb.jrc.ec.europa.eu/qmrf/search_catalogs.jsp?id=174&idstructure=. Accessed Jan 2011
35. The JRC QMRF Q8-10-14-171 "QSAR for Relative Binding Affinity to Estrogen Receptor". http://qsardb.jrc.ec.europa.eu/qmrf/search_catalogs.jsp?id=171&idstructure=. Accessed Jan 2011
36. MATLAB® The Language of Technical Computing v. R2010b (2010) The MathWorks Inc., http://www.mathworks.com. Accessed Jan 2011
37. Chang J, Lei B, Jiazhong L, Lia S, Shen Y, Yao X (2008) Accurate and validated quantitative structure–activity relationship model of caspase-mediated apoptosis-inducing activity of phenolic compounds using density functional theory calculation and genetic algorithm–multiple linear regression. QSAR Comb Sci 27(11–12):1318–1325
38. Roy PP, Leonard JT, Roy K (2008) Exploring the impact of training sets for the development of predictive QSAR models. Chemom Int Lab Syst 90:31–42
39. Puzyn T, Mostrag A, Falandysz J, Kholod Y, Leszczynski J (2009) Predicting water solubility of congeners: chloronaphthalenes—a case study. J Hazard Mater 170(2–3):1014–1022

40. Puzyn T, Gajewicz A, Rybacka A, Haranczyk M (2011) Global vs. local QSPR models for persistent organic pollutants: balancing between predictivity and economy. Struct Chem. doi:10.1007/s11224-011-9764-5
41. Stine R (1989) An introduction to bootstrap methods: examples and ideas. Sociol Methods Res 18(2–3):243–291
42. Wehrens R, Putter H, Buydens LMC (2000) The bootstrap: a tutorial. Chemom Int Lab Syst 54:35–52