JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

PLS-Optimal: A Stepwise D-Optimal Design Based on Latent Variables

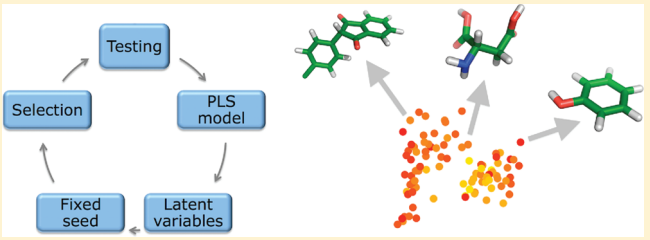
Stefan Brandmaier,*[,†,‡] Ullrika Sahlin,[†] Igor V. Tetko,[‡,§] and Tomas Öberg[†]

[†]School of Natural Sciences, Linnaeus University, 391 82 Kalmar, Sweden

[‡]Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstaedter Landstrasse 1, Neuherberg D-85764, Germany

[§]eADMET GmbH, Ingolstaedter Landstrasse 1, Neuherberg D-85764, Germany

ABSTRACT: Several applications, such as risk assessment within REACH or drug discovery, require reliable methods for the design of experiments and efficient testing strategies. Keeping the number of experiments as low as possible is important from both a financial and an ethical point of view, as exhaustive testing of compounds requires significant financial resources and animal lives. With a large initial set of compounds, experimental design techniques can be used to select a representative subset for testing. Once measured, these compounds can be used to develop quantitative structure−activity relationship models to predict properties of the remaining compounds. This reduces the required resources and time. D-Optimal design is frequently used to select an optimal set of compounds by analyzing data variance. We developed a new sequential approach to apply a D-Optimal design to latent variables derived from a partial least squares (PLS) model instead of principal components. The stepwise procedure selects a new set of molecules to be measured after each previous measurement cycle. We show that application of the D-Optimal selection generates models with a significantly improved performance on four different data sets with end points relevant for REACH. Compared to those derived from principal components, PLS models derived from the selection on latent variables had a lower root-mean-square error and a higher Q2 and R2. This improvement is statistically significant, especially for the small number of compounds selected.



1. INTRODUCTION

The REACH legislation[1] includes the requirement that every chemical compound produced in or imported to the European Union in an amount of more than one ton has to be registered regarding a number of end points. Experimental determination of these properties for all compounds would require high-throughput testing. According to Rovida and Hartung, the financial requirements for such testing are about €9.5 billion.[2] For potentially hazardous, dangerous, or hardly degradable substances, registration also requires information about their bioaccumulation and toxicity. Apart from cost and time efficiency, a sample, for example, bioconcentration, requires around two months and can cost more than €200—this also leads to ethical problems, as experimental determination of end points associated with toxicity and bioaccumulation is achieved by animal tests.

The necessity to keep the overhead of (animal) testing as low as possible is also important in many other research areas, for example, the chemical or pharmaceutical industries. One common strategy to address this problem is to use structure−activity modeling[3] and to predict the required properties rather than performing experimental measurements. This strategy entails testing only a small subset of all the compounds of interest and constructing a predictive model using the experimentally determined values. This basic task can be reduced to the problem of drawing a representative subsample of a larger set. This method is important in other fields of research, e.g., quantitative structure−activity relationship (QSAR) development,[4] large-scale database scanning,[5] in silico drug design,[6] and compound prioritization,[7] as well as in experimental design for risk assessment within REACH.[8]

There are several commonly accepted approaches[9−13] for choosing a representative subset of compounds to deliver the most reliable model. These approaches select the subset according to various criteria. Partition-based approaches, like full or factorial design, attempt to select a sample that is representative of the whole chemical space of interest, separating the descriptor space into subspaces and finding a representative compound for each of these subspaces.[14] Other approaches aim to find the subset that is most descriptive for the remaining compounds by ranking the representativity of compounds according to their pairwise distance in descriptor space.[15,16]

D-Optimal design, which has been recommended as the favorable alternative for linear models in several publications,[17,18] selects the most representative combination of compounds for linear models.[19] In this method, each possible

Received: January 11, 2012
Published: March 30, 2012

dx.doi.org/10.1021/ci3000198 | J. Chem. Inf. Model. 2012, 52, 975−983

subset of a given size is evaluated to derive the information matrix. The most distinct and thereby most optimal of all possible subsets is the one with the maximum determinant of the information matrix. This is equivalent to the set with the maximum entropy.[20] An advantage of the D-Optimal selection criterion is that in our design problem, the training set is selected from a limited candidate set. Pronzato[21] has shown that when the data space is limited, a sequential D-optimal design, given that some conditions are met, is asymptotically optimal.

All the aforementioned approaches select compounds using descriptors only. Usually a principal component analysis (PCA) is applied to these descriptors to extract the so-called principal properties, which are used to select compounds. Although the statistics literature also provides a large variety of sequential approaches,[22,23] their application in QSAR is very limited. Moreover, we are not aware of any available implementations of these approaches. Sequential approaches are arranged in a stepwise manner and adapt to the gathered information about the response. Including the target property with a sequential design strategy might provide a better selection of compounds.

In this study, we investigate an adaptive, stepwise experimental design strategy that is based on the D-Optimal approach. The method combines D-Optimal design with partial least squares (PLS) techniques to iteratively refine the descriptor space for the compound selection. This refinement is realized by using PLS latent variables instead of the principal components. In contrast to the static principal components, the PLS latent variables, which are correlated to the target property, can be recalculated after each measurement cycle. As the number of measurements increases from cycle to cycle, each new model is an improvement of the previous one. Based on these iteratively refined latent variables, an initially selected set of compounds is extended in a stepwise manner. A similar idea was proposed by Lundstedt and Thelin.[24] The authors used a two-step process consisting of a synthesis step and a purification step in which they alternated between PCA and PLS. However, their aim was to select the most important variables for a model, while the aim of our method is to find the most informative compounds for model development.

We evaluate the performance of the new approach on four different data sets and compare it to the original D-optimal design. D-optimal design based on latent variables can be performed with or without higher order interaction terms. Comparison of the suggested and the original approach was made on experimental designs, with and without higher order terms of the latent variables, since the performance of either method is dependent on the characteristics of the data set.

2. MATERIAL AND METHODS

2.1. QSAR Data Sets. To validate the performance of the stepwise method, four data sets with different end points were collected from the literature. All of the selected end points are relevant for REACH and risk assessment. To cover a broad spectrum of possible applications and to better examine the performance of the new method, the sets collected varied in several criteria: size, modeling, and measurement complexities.

The selected end points included two toxicity end points, namely the log-scaled lethal concentration for fathead minnow (log LC_{50}) and the inhibition growth concentration for *Tetrahymena pyriformis* ($-\log IGC_{50}$), an adsorption coefficient ($\log K_{OC}$), and the boiling point. The number of compounds in these data sets ranged from 96 ($-\log IGC_{50}$) to 1198 (boiling point). The log LC_{50} data set contained 535 compounds and the log K_{OC} data set 648 compounds.

To ensure consistency of the data sets and to avoid problems resulting from different experimental methods, we applied several filters to all collected measurements. As the measurements for toxicity are sensitive to laboratory conditions and experimental procedures, we limited the data points within one data set to one source only. This means that the measurements either had to be from only one lab or had to be taken from a previously reviewed collection.

The Tetratox database[25] and the EPA's fathead minnow acute toxicity database[26] were selected for log IGC_{50} and log LC_{50}, respectively. The log K_{OC} data set was based on the reviewed collection of Meylan et al.[27] As the precision and consistency of boiling point measurements are higher, we did not preselect any data for this end point and used the whole EPI suite data.[28] For all four sets we excluded inorganic compounds, radicals, charged molecules, and salts. Further, we removed compounds for which no exact values, rather an interval or only minimum or maximum values, were given.

For the compounds in the log LC_{50} and log K_{OC} data sets, no structural filters were applied. Therefore, the data sets contained a wide variety of different compound classes and had wide structural diversity, and the resulting models can be designated as "global." For boiling points, a filter was applied to the structures, limiting the compounds in the final data set to halogenated ones, containing bromine, fluorine, and/or chlorine. The initial $-\log IGC_{50}$ data set contained more than 1000 compounds. However, to evaluate the performance of the developed approach on a relatively small data set, a subset of 96 compounds was randomly selected.

The descriptors for model development were ALogPS[29] lipophilicity and solubility and E-state indices.[30] The E-state indices have been shown to provide a high accuracy of predictions for similar end points in our previous publications.[31,32] ALogPS descriptors were added to account for physicochemical parameters, e.g., solubility and distribution, which could be important for the considered end points. For this study, all descriptors were normalized to [0,1] range. The descriptors were calculated using the Online Chemical (OCHEM) database,[33] which is publicly accessible at http://ochem.eu.

2.2. Methods. *2.2.1. D-Optimal Design.* D-Optimal design selects the most distinct subset of molecules of a given size n from a larger initial set containing m compounds. Figure 1a shows the result of a D-Optimal selection. The x- and y-axes represent two first principal components, while each dot represents a chemical compound. Dots marked red are the compounds that were selected using the D-Optimal criterion.

The D-Optimal selection criterion was implemented as suggested in the literature.[34] Fedorov's heuristic approach[35] was used to optimize the selection speed. Further, the implementation was extended by the option to add a fixed seed to the selection. In all the following steps, the fixed seed is a set of compounds for which the target property is considered to be already measured. This additional feature enables us to perform a compound selection that depends on a preselected set of compounds. Newly selected compounds are therefore not only most distinct to one another but also to the preselected compounds. This enhancement was made by adding the preselected compounds to the model matrix.

The application of the D-Optimal criterion to only linear meta descriptors is particularly capable of problems with linear

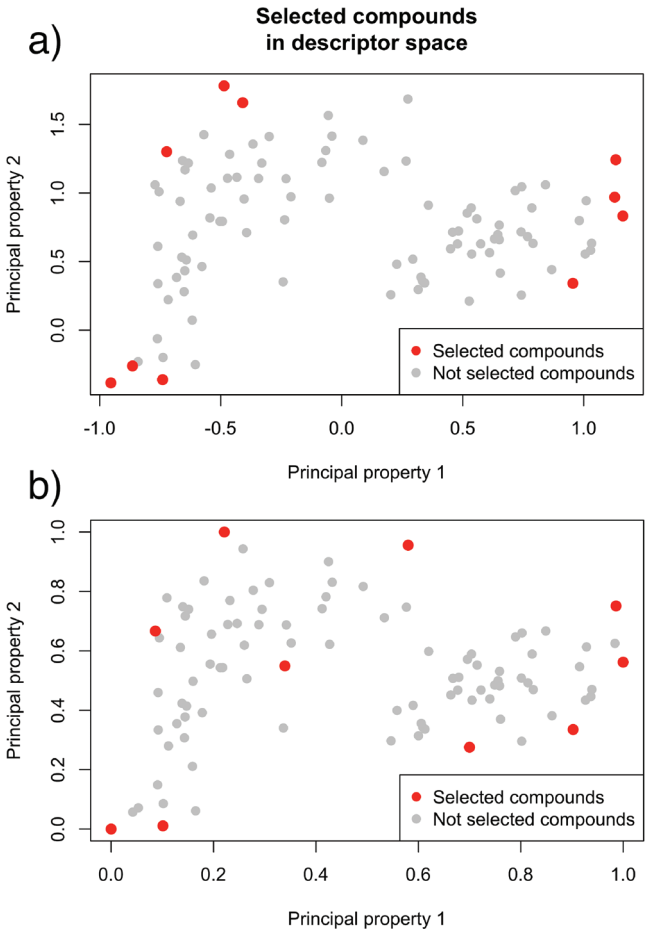


Figure 1. The results of the D-Optimal selection using (a) linear terms only and (b) linear, cross, and square terms.

dependencies but reveals problems for dependencies of other order. Therefore, the D-Optimal criterion was applied not only to the principal components or PLS latent variables but also, in an additional examination, to a set of meta descriptors.

These meta descriptors contain the normalized components from PCA or PLS and their square and pairwise cross terms.[36] For a set of v input variables, d_1, d_2, ..., d_v, additionally the square terms $(d_1)^2$, $(d_2)^2$, ..., $(d_v)^2$ and the cross terms $d_i d_j$ with $i = 1, 2, ..., v$, $j = 1, 2, ..., v$ and $i \neq j$. This extension increases the dimensionality of the search space by a quadratic factor from v input variables to $v * (1.5 + 0.5 * v)$ meta descriptors.

This contributes to the quality of the outcoming sample, as the selected compounds are not located exclusively on the periphery of the data cloud in the chemical space but also in the center. Figure 1b shows the resulting selection on the same data set as Figure 1a.

2.2.2. The Stepwise Approach. The stepwise approach has two phases: First, the application of the extended D-Optimal design, which takes preselected compounds into consideration, and second, an implementation of PLS regression to calculate the so-called latent variables for all compounds. The calculation of these latent variables is based on a PLS model, which is built on the preselected compounds.

Latent variables from PLS are comparable to the principal components of a PCA. However, in contrast to PCA components, which are selected to maximize the variance of the data set (i.e., to cover as much of the data variability as possible), the PLS latent variables are selected to maximize the covariance (i.e., to provide maximum correlation) with the target variable. Therefore, in addition to PCA components, the latent variables contain information about the target variable. In our approach, instead of the uncorrelated PCA components, we use the PLS components as descriptors for the D-Optimal design. With this modification, the representation of the compounds of interest is adjusted to the considered end point and no longer depends only on the uncorrelated structural information.

In the first phase of the stepwise approach, a traditional D-Optimal design is used to select an initial subset, containing a fixed number of compounds. Therefore, a D-Optimal selection is applied to a fixed number of principal components derived from a PCA on a set of descriptors for all compounds within the set of relevant compounds. For all further steps, the compounds selected in the previous steps are considered to be already tested, and a PLS model is built on them. The developed PLS model is then used to calculate the latent variables for all compounds, and the D-Optimal selection is performed utilizing these latent variables instead of the principal components. Further, this selection is taking the fixed seed into consideration, and all preliminarily tested compounds are members of the resulting set of the D-Optimal design.

The most important differences between the stepwise approach based on latent variables and the traditional D-Optimal selection are shown in Figure 2. Whereas the traditional method (left side of the figure, in pink) selects all compounds at the same time, the stepwise approach (right side of the figure, in blue) constantly increases the number of compounds cyclically. Further, the chemical space to represent the compounds is refined with each cycle.

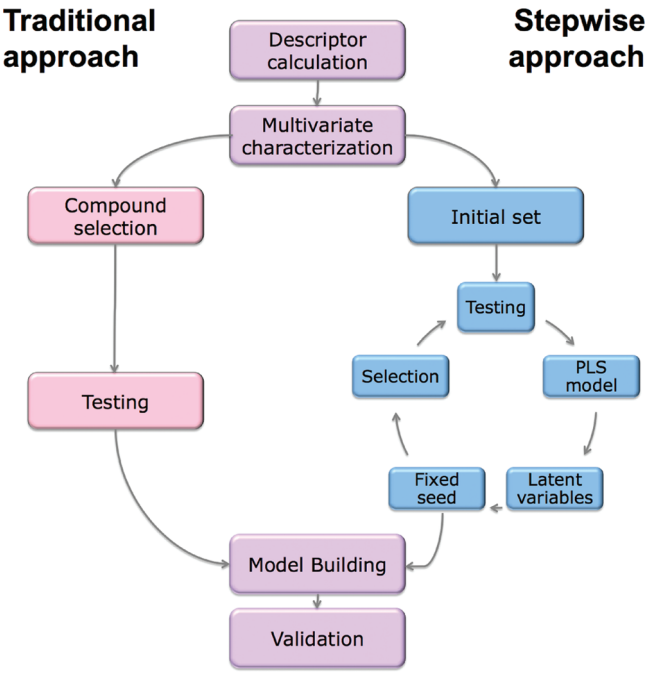


Figure 2. Comparison of the traditional workflow (left) and the suggested stepwise selection (right).

2.2.3. Validation. To obtain a meaningful statistical basis to compare the performance of the sequential approach with the traditional D-Optimal approach, we generated 100 subsets

dx.doi.org/10.1021/ci3000198 | J. Chem. Inf. Model. 2012, 52, 975−983

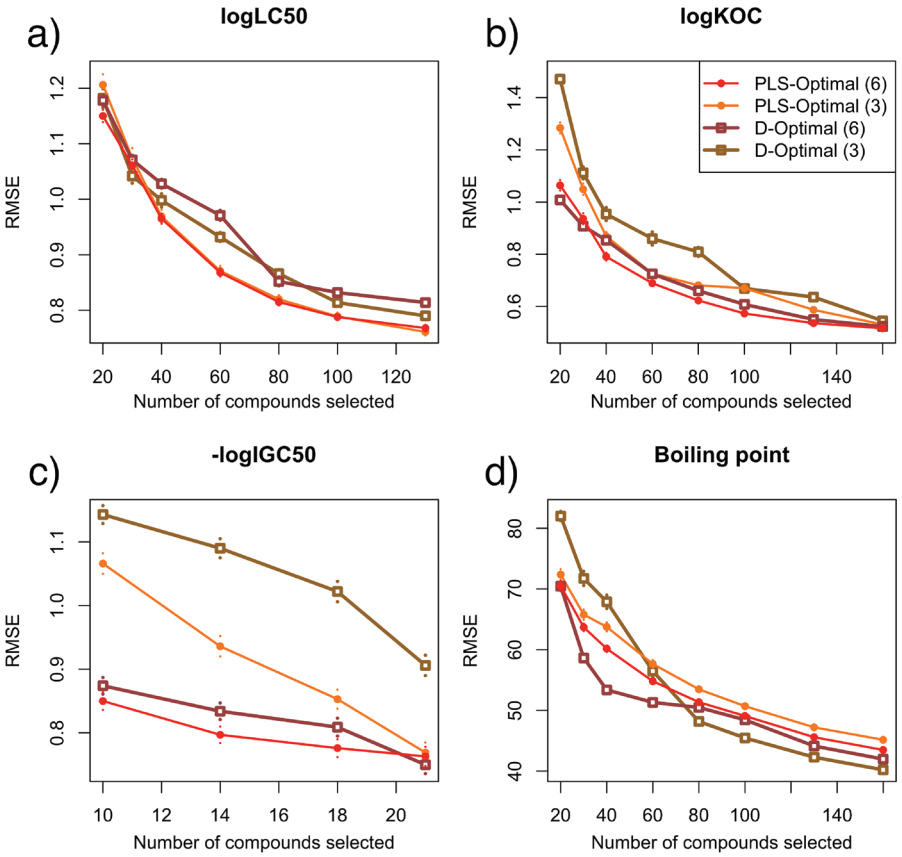


Figure 3. The average error on the (a) log LC_{50}, (b) log K_{OC}, (c) −log IGC_{50}, and (d) boiling point data sets using a linear search space. The performance of the PLS-Optimal approach is shown in red (for six latent variables) and yellow (three latent variables), while the performance of the traditional approach is shown in brown (six principal components) and green (three principal components). On the *x*-axis, the number of compounds used to build the according 100 models is displayed. For the traditional D-optimal design (using principal components), all compounds were selected simultaneously. For the stepwise approach (using latent variables), the preliminary selected compounds were extended with new ones each cycle. The *y*-axis shows the average performance of RMSE.

(design sets) from each data set. The compounds in the subsets were chosen randomly, and the size of each subset was 75% of the whole data set. The remaining 25% of compounds were used as respective external validation sets.

Each of the design sets was used for the experimental design. Both the classical and stepwise approach were used to select a fixed number of compounds, which included 10, 20, 30, 40, 60, 80, 100, 130, and 160 compounds for the three large data sets (log K_{OC}, boiling point, and log LC_{50}) and 6, 10, 14, 18, and 21 compounds for the small data set (−log IGC_{50}).

In the case of the principal components, the number of variables used to describe the search space was always fixed (respectively, cross and square terms). The number of PLS components used could be either fixed or automatically optimized, minimizing the coefficient of determination in cross-validation.

To obtain comparable information about the quality of the compound selection, we used PLS to train a linear regression model on the selected compounds. The number of latent variables for the final model was determined in a five-fold cross-validation on all selected compounds using the coefficient of determination as criterion for the optimal number.[37] The reason why we chose PLS for evaluation of the final selection is the robustness of the method. As it uses a projection of the descriptors, it reliably finds linear correlations of the target property in the descriptor space. Furthermore, by taking the target property into account, PLS removes noise in the descriptor space. The cross and square terms we used to span the search space for the D-Optimal criterion were not used for development of the PLS models.

Although this is unlikely, the use of PLS regression might favor the PLS-based design, whereas PCA-based design might provide better results for the PCA regression. To address this question, we also developed PCA regression models for compounds selected using principal components.

The performance of the developed model was then calculated for two different splits of the data sets. The first split was the external validation set, and the second split was the selection set without the compounds that were suggested for testing. The validation was performed on these two splits to represent different targets or intentions for the compound selection. The performance on the external validation set gives a measurement of a global validity, as it contains only compounds excluded from the selection. It is thus an independent measurement that enables estimation of the model quality for new compounds. Another point of relevance is the performance of the model for compounds of interest that were not selected for testing. In most cases, it is the performance for precisely these compounds that is the underlying motivation for the experimental design.

For both splits, root-mean-square error (RMSE) was calculated as a measurement of error. The mean value of RMSE for the 100 models calculated for each data set was then

dx.doi.org/10.1021/ci3000198 | *J. Chem. Inf. Model.* 2012, 52, 975−983

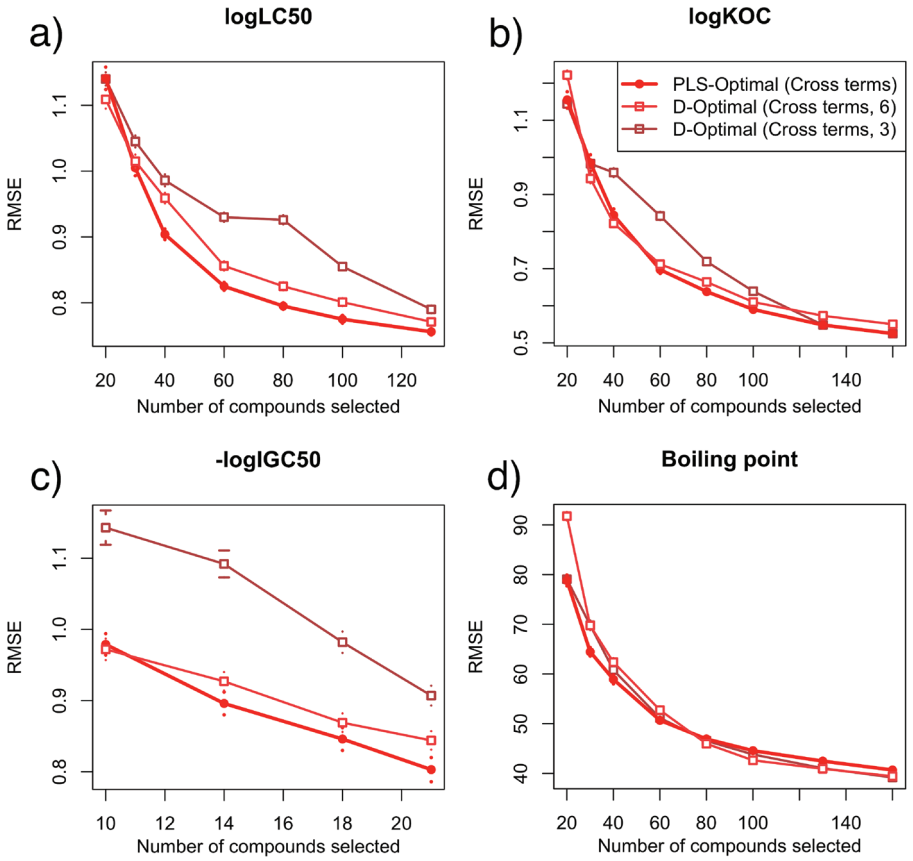


Figure 4. Development of the average error on the (a) log LC_{50}, (b) log K_{OC}, (c) −log IGC_{50}, and (d) boiling point data sets using a search space extended by cross and square terms. For all end points, the bold red line represents the development of the stepwise approach; the development of the traditional approach on six principal components is represented by a dark-red line and on three principal components by a red-brown line.

used to compare the quality of experimental designs for PLS-Optimal and the traditional method.

3. RESULTS

3.1. Linear Search Space. To compare the methods, we used three and six components alternatively for the three large data sets and two and four components alternatively for the −log IGC_{50} data set. The performance of the developed models is shown in Figure 3a−d. The performance of the models built with PCA regression was significantly worse than that of PLS regression for all analyzed data sets. Therefore PCA regression results are no further provided.

The selection of these numbers of components for comparison is a reasonable one. The lower number of PLS or PCA components (3 or 2) shows the performance for a low dimensionality search space, which is particularly interesting regarding runtime requirements. The higher number of PLS or PCA components (6 or 4) adopts the Organization for Economic Co-operation and Development (OECD) principles[38] regarding the number of descriptors to be used for a linear model.

There are several important observations: First, with an increasing number of selected compounds, the model performance also improves. Second, with an increasing number of latent variables (or principal components for the traditional method), the performance of the resulting models also increases. This observation is particularly clear for the stepwise approach, with the exception of the log LC_{50} data set.

This is an expected result. A larger number of molecules allows the development of better models, while higher dimensionality in the search space provides a more diversified representation of the compounds and thereby increases the information content of the search space.

Let us take a closer look at the performance of the methods on the external validation split. It is clear that for all data sets, except for boiling point, error decreases faster with the stepwise method. Further, using the stepwise approach, a point of convergence, where the performance of the outcoming model no longer changes, is reached with a lower number of compounds. For the log LC_{50}, the log K_{OC}, and the −log IGC_{50} (Figure 3a−c, respectively) data sets, the performance of the stepwise approach is better than that of the traditional approach using the same number of latent variables or principal components and the same number of compounds selected.

This improvement is statistically significant with a p-value <0.05 for 40 compounds and a p-value <0.001 for the range of 60−130 selected compounds for the log LC_{50} data set according to the binomial test (the binomial distribution with N = 100 trials corresponding to the number of models used in our study). The sequential approach using 40 selected compounds and 6 latent variables provides the same accuracy of prediction as the traditional approach using 60 selected compounds.

The results for the validation on the log K_{OC} data set are similar. The increase in performance derived with the sequential approach using 6 components in the range of 40−130 compounds selected is significant and on average is 0.037

dx.doi.org/10.1021/ci3000198 | J. Chem. Inf. Model. 2012, 52, 975−983

log units. For the same range of compounds, the increase of RMSE for three components is 0.079 log units.

For a search space of two dimensions, the performance of PLS-Optimal on the −log IGC_{50} data set is better with statistical significance ($p < 0.001$) for the whole range from 10 to 21 compounds (14−30%). The greatest difference in the performance of the methods is found for 18 selected compounds: in 90 of 100 cases, the models built with the stepwise selection delivered a better result than those built on the traditional selection.

Comparison of the performance of both approaches on the boiling point data set (Figure 3d) reveals results that differ from those of the other data sets. The performance of the traditional approach using PCA components is better. In the case of 6 principal components used to define the search space, the incline in the error is steep for the first 40 compounds selected. Beyond that, until 100 selected compounds, there is almost no improvement in performance.

The results for the compounds in the design set that was not used to train the model were very similar and are therefore not explicitly discussed in this or the following sections.

3.2. Square and Cross Terms. The same calculations as for the linear search space were also performed for the search space using square and cross terms of the PCA or PLS components. For the traditional approach, the number of principal components was fixed to the same values as for the linear approach. The number of resulting meta descriptors was thus also fixed to 27 and 9 for the three large data sets and to 14 and 5 for the small data set. In contrast, the number of PLS latent variables for the stepwise approach was automatically optimized for this calculation. The procedure for estimating the optimal number was the same as for estimating the number of components to evaluate the resulting mode. We also tried to optimize the number of principal components in a similar way, regarding the error on reconstruction.[39,40] However, our examinations indicated that the performance of the resulting models improves with any further principal component.

The results for the validation on the square terms are shown in Figure 4a−d. The axes are similar to Figure 3. Similar to the linear search space, the performance of the resulting models improves with an increasing number of compounds selected. A further observance on all end points, except for the boiling point (Figure 4d), is that the performance of the traditional approach improves with the number of principal components used. The selection performance on six principal components (or four for −log IGC_{50}) is better than the selection performance on three (or two) principal components for the whole examined range. Additionally, for six or four principal components and the use of cross terms and square terms, the development of the error describes a constant curve with a continuously increasing incline, without the inconstancies observed for the linear search space.

Although the performance for six or four principal components converges with that of the stepwise approach, the models built on the compounds selected by PLS-Optimal are still better for most of the examined ranges on the log LC_{50}, the log K_{OC}, and the −log IGC_{50} (Figure 4a−c, respectively) data set. For log LC_{50}, this improvement is significant from 40 to 130 compounds selected. The average error for 40 selected compounds is 6% lower for the selection derived using the stepwise approach. A model with better performance than that, derived from 100 selected compounds using the traditional approach, could be achieved with the 80 compounds selected with the stepwise approach.

For log K_{OC}, the development is similar. After an almost similar performance on the first 40 compounds selected, the stepwise approach performs significantly better in the range from 60 to 160 compounds (12−33%). The average error within that range is 0.022 log units (3.5%) lower than for the traditional approach. The development on the −log IGC_{50} data set is almost analogous. After a similar performance for the first 10 selected compounds, the average error of the stepwise approach decreases more quickly than for the traditional approach on four principal components. We also evaluated the models built on the selected compounds on the 1000 compounds excluded from this data set for this study and found the results to be similar.

For the cross and square-term usage, too, the development on the boiling point data set differs from the other data sets. Both the stepwise and the traditional approaches on six or three latent variables derived from PLS gave almost the same performance. The error for the PLS-Optimal approach converges faster in the range from 20 to 40 selected compounds; however, the performance of the traditional approach is better in the range from 100 to 160 selected compounds.

3.3. Comparison with a Small Number of Selected Compounds. As the quality of the crossed traditional approach seems to increase with any additional principal component, it is interesting to take a look at the selection of only very few compounds. As it is a requirement for the D-Optimal criterion to work that the model matrix has more observations than variables, the number of components to be used is strictly limited. Therefore, on the three large data sets another examination within the range from 5 to 35 selected compounds was initiated. We used the meta descriptors containing the normalized components and their square and cross products. The number of PLS latent variables used in the stepwise approach was automatically determined, whereas the number of principal components used for the traditional approach was fixed to the maximum that could be used, respective of the number of compounds to select. This means 1 component for less than 6 selected compounds, 2 for less than 10, 3 for less than 15, 4 for less than 21, 5 for less than 28, and 6 components for less than 30 selected compounds.

The results in Figure 5a−c show that the stepwise approach clearly achieves better performance for all three end points. This improvement is significant ($p < 0.001$) for the whole range from 10 to 35 selected compounds. In the case of the log K_{OC} data set (Figure 5b) and for the range from 13 to 24 selected compounds, the stepwise approach performed better for more than 90 out of 100 splits.

Regarding the boiling point (Figure 5c), the average RMSE performance for 24 compounds selected by the traditional approach could be achieved with only 13 compounds selected in a stepwise procedure. Furthermore, in the range from 13 to 32 selected compounds, the improvement of the average RMSE for the same number of selected compounds is better by at least 9 degrees. For log LC_{50} (Figure 5a), the average performance with 24 compounds selected in a stepwise manner could not be achieved with less than 32 compounds selected based on principal components. In the case of the log K_{OC} data set, the stepwise approach delivers an average performance for 13 selected compounds that cannot be achieved with less than 24 compounds utilizing the traditional method. The RMSE for

dx.doi.org/10.1021/ci3000198 | J. Chem. Inf. Model. 2012, 52, 975−983

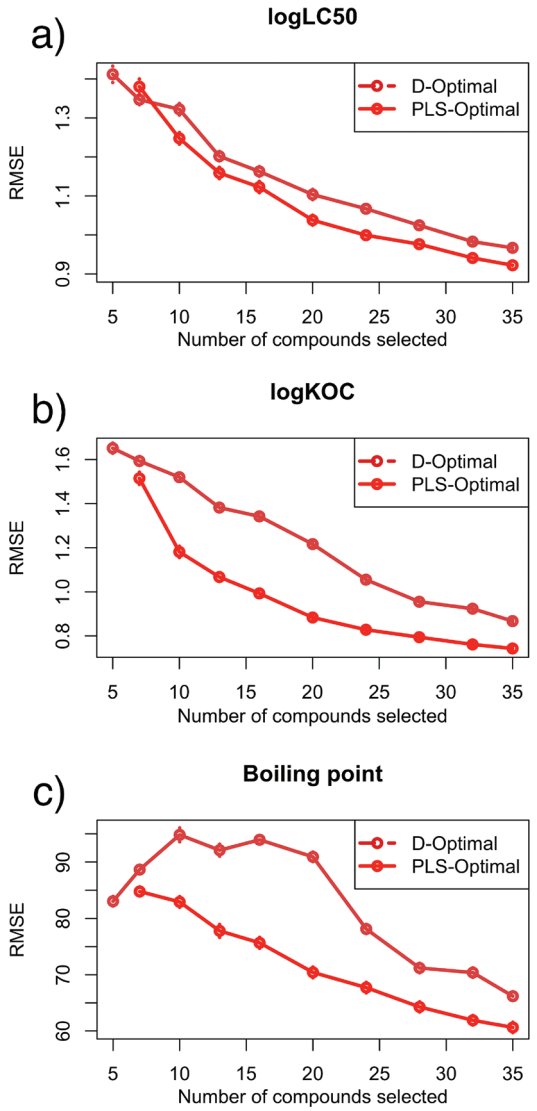


Figure 5. Results of the error validation for cross and square terms with a low number of compounds.

that data set was on average 21% less in the range from 10 to 35 selected compounds.

Finally, comparing the results of the stepwise approach applied to a sequence of 10, 20, and 30 selected compounds with the results of the stepwise approach applied to the increased step size, the latter delivers better model quality for the same number of compounds selected. The average RMSE for 28 selected compounds using the smaller step size is 0.19 log units better for the log K_{OC} data set and 0.03 log units better for the log LC_{50} data set.

4. DISCUSSION

Our results, derived from examination of the PLS-Optimal performance on the log LC_{50}, log K_{OC}, and −log IGC_{50} data sets within a range of 5−35% of compounds selected, show that the stepwise approach utilizing PLS latent variables can significantly increase the quality of the resulting model and thereby help to save resources. Compared to a model based on selection of compounds by the traditional D-Optimal design approach, the model derived from the same number of compounds selected using the stepwise approach delivered a decreased RMSE and an increased R^2 and Q^2 for both the linear search space and a search space extended by cross and square terms. The convergence of the error to a minimum was clearly faster, and the improved performance can be observed in the whole range from approximately 10−30% of compounds selected.

The performance on the boiling point data set can be explained by the depiction, shown in Figure 6, of the chemical

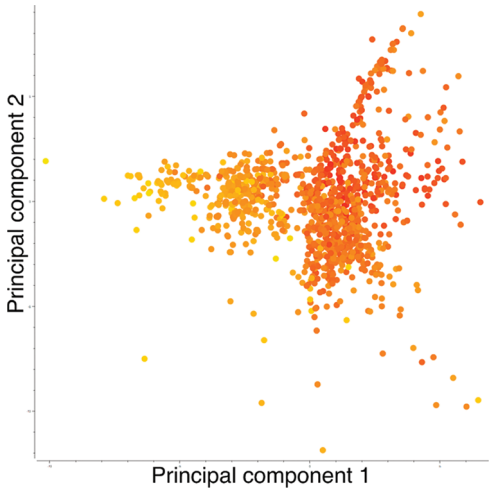


Figure 6. Compounds of the log LC_{50} data set in PCA space. The color of the data points represents the measured value.

space using the principal components. The *x*-axis represents the first principal component and the *y*-axis the third principal component, and the color of the data points displays the end point values. We can clearly see that not only the first principal component but also the third is strongly correlated to the end point. Furthermore, the principal components are not just correlated with the end point; they are almost similar to the PLS latent variables, derived on the whole data set. Table 1

Table 1. Loadings and Rank of Five Descriptors for the First PLS Latent Variable and the First Principal Component

descriptor	PLS loading	rank PLS	PCA loading	rank PCA
SeaC2C3aa	0.506	1	−0.33	1
SaaCH	0.471	2	−0.321	2
SeaC2C2aa	0.332	3	−0.272	3
SsF	−0.273	4	0.167	8
Se1C3Cl1a	0.273	5	−0.249	4

shows this correlation for the first PLS and PCA component. The correlation in the second and third dimension is comparable. While the PLS-Optimal approach tries in a stepwise manner to find a stable depiction and correlation, the PCA used for the traditional approach provides them exactly. As the boiling point is a very simple end point and widely cleared up, this effect was a foreseeable one. Nevertheless, it is a good depiction of the limitations of the developed approach, and we suggest using the stepwise approach particularly for experimental designs for complex end points.

The models built on PLS-Optimal design deliver a more stable performance regarding the error development for all four examined end points. Whereas with the classic approach the performance shows some variability and deviations with an increasing number of selected compounds, the performance

dx.doi.org/10.1021/ci3000198 | J. Chem. Inf. Model. 2012, 52, 975−983

development of the PLS-Optimal design is much smoother and approximates a hyperbolic function. This is observable even for a search space of only three variables.

Whereas a principal component can be completely uncorrelated to the target property and thereby lead to an accumulation of noise, the PLS components contain only correlated information. Furthermore, they are ranked by their importance for the specific end point, whereas the principal components are ranked solely by their variance. This leads to an accumulation of irrelevant information in the principal components. Therefore, the number of principal components required to capture the same amount of information for an end point is usually higher than the required number of PLS latent variables. This is important, both in terms of stability and efficiency, in order to keep the dimensionality of the search space as low as possible.

The effect that PLS components are less prone to noise can be observed for the selection of only a small number of compounds, in particular when using cross terms. In the range from 5 to 35 selected compounds, PLS-Optimal delivers significantly improved performance compared to the traditional D-optimal design.

We repeated the whole study with raw (non-normalized) and standardized descriptors, which resulted in a worse performance of the resulting models. The average error performance was worse, and the development of the error was less stable for both analyzed approaches. We also compared the stepwise approach with the traditional one on other descriptor sets, i.e., ISIDA fragments[41] and QNPR descriptors.[42] The results were similar and did not influence our conclusions.

5. CONCLUSION

Our results show that the performance of D-optimal experimental design in QSAR model development can be significantly improved by taking the correlation between descriptors and property into consideration. The PLS-optimal design operates in the property-based space; therefore, the selection of compounds is not only based on their structural properties but also tuned for a specific end point. Similar advantages of property-based space were demonstrated in assessing the accuracy of predictions for quantitative and qualitative models.[31,43]

The results presented in this study are limited to the application of the D-Optimal criterion to PLS latent variables. The concept of taking the correlation or covariance with the target property into account could be realized with any other selection criterion. Furthermore, the use of sequentially refined latent variables is a powerful tool, but an integrative process of descriptor selection, based on the preselected compounds, could also realize the stepwise optimization of the chemical space.

However, the concept of substituting the PCA representation of the descriptor space with PLS latent variables is also relevant in terms of efficiency. Although the performance of models derived from the traditional selection converged with the performance of compounds selected on PLS components, the search space required a higher dimensionality. This carries weight especially in terms of use of cross and square terms and for large-scale scans on databases containing more than 100 000 compounds. The runtime requirements for such operations can be reduced to a fraction with the approach presented here.

The sequential approach could, we suggest, also be extended to a Bayesian one, simply by performing the initial selection on the latent variables derived from a model, built on measurements, collected by a literature search. The use of nonlinear methods, e.g., kernel PLS, could be an interesting work to further extend the method we have developed in this article.

6. SOFTWARE USED

PLS models to evaluate the performance of the analyzed approaches were calculated using WEKA.[44] The PLS latent variables were calculated using PLS package[45] in the statistical language [R].[46]

7. IMPLEMENTATION AND ACCESSIBILITY OF DATA

An implementation that enables users to make use of the stepwise approach for experimental design can be publically accessed at http://qspr-thesaurus.eu. The web interface enables users to apply the D-Optimal criterion to principal components or latent variables and visually compare and explore the selection.

The data sets used in this article and the models built on them are available at http://ochem.eu/article/9423.

AUTHOR INFORMATION

Corresponding Author

*stefan.brandmaier@gmail.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This study was partially financed by FP7 project “Case studies on the development and application of in-silico techniques for environmental hazard and risk assessment” (CADASTER), grant agreement number 212668, and by Marie Curie Initial Training Network “Environmental ChemOinformatics” (ECO) project, grant agreement number 238701.

REFERENCES

(1) EC, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Official Journal of the European Union, L: Legislation (English Edition) 2006, L 396/1 of 30.12.2006, 3−280.

(2) Rovida, C.; Hartung, T. Re-evaluation of animal numbers and costs for in vivo tests to accomplish REACH legislation requirements for chemicals - a report by the transatlantic think tank for toxicology (t(4)). *ALTEX* **2009**, *26*, 187−208.

(3) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*.

(4) Öberg, T. A QSAR for the hydroxyl radical reaction rate constant: validation, domain of application, and prediction. *Atmos. Environ.* **2005**, *39*, 2189−2200.

(5) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.

(6) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.

(7) Stenberg, M.; Linusson, A.; Tysklind, M.; Andersson, P. L. A multivariate chemical map of industrial chemicals − Assessment of

various protocols for identification of chemicals of potential concern. *Chemosphere* **2009**, *76*, 878−884.
(8) Lahl, U.; Gundert-Remy, U. The Use of (Q)SAR Methods in the Context of REACH. *Toxicol. Mech. Method.* **2008**, *18*, 149−158.
(9) Eichler, U.; Ertl, P.; Gobbi, A.; Rohde, B. Definition of an Optimal Subset of Organic Substituents. Interactive Visual Comparison of Various Selection Algorithms. *Internet J. Chem.* **1999**, *2*.
(10) Daszykowski, M.; Walczak, B.; Massart, D. L. Representative subset selection. *Anal. Chim. Acta* **2002**, *468*, 91−103.
(11) Mason, J.; Pickett, S. Partition-based selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.
(12) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65−84.
(13) Wootton, R.; Cranfield, R.; Sheppey, G. C.; Goodford, P. J. Physicochemical-activity relations in practice. 2. Rational selection of benzenoid substituents. *J. Med. Chem.* **1975**, *18*, 607−613.
(14) Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, Å.; Pettersen, J.; Bergman, R. Experimental design and optimization. *Chemometr. Intell. Lab.* **1998**, *42*, 3−40.
(15) Chaudhuri, B. B. How to choose a representative subset from a set of data in multi-dimensional space. *Pattern Recognit. Lett.* **1994**, *15*, 893−899.
(16) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J.; Osman, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285−289.
(17) Eriksson, L.; Johansson, E. Multivariate design and modeling in QSAR. *Chemometr. Intell. Lab.* **1996**, *34*, 1−19.
(18) Baroni, M.; Clementi, S.; Cruciani, G.; Kettaneh-Wold, N.; Wold, S. D-Optimal Designs in QSAR. *Quant. Struct.-Act. Relat.* **1993**, *12*, 225−231.
(19) Wold, S.; Josefson, M.; Gottfries, J.; Linusson, A. The utility of multivariate design in PLS modeling. *J. Chemometr.* **2004**, *18*, 156−165.
(20) Van Den Berg, J.; Curtis, A.; Trampert, J. Optimal nonlinear Bayesian experimental design: an application to amplitude versus offset experiments. *Geophys. J. Int.* **2003**, *155*, 411−421.
(21) Pronzato, L. One-step ahead adaptive D-optimal design on a finite design space is asymptotically optimal. *Metrika* **2010**, *71*, 219−238.
(22) Roy, A.; Ghosal, S.; Rosenberger, W. F. Convergence properties of sequential Bayesian D-optimal designs. *J. Stat. Plan. Infer.* **2009**, *139*, 425−440.
(23) Chaloner, K.; Verdinelli, I. Bayesian Experimental Design: A Review. *Stat. Sci.* **1995**, *10*, 273−304.
(24) Lundstedt, T.; Thelin, B. A multivariate strategy for optimizing a two-step process. *Chemometr. Intell. Lab.* **1995**, *29*, 255−261.
(25) Schultz, T. W. Tetratox: Tetrahymena pyriformis population growth impairment endpoint − a surrogate for fish lethality. *Toxicol. Mech. Methods* **1997**, *7*, 289−309.
(26) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.
(27) Meylan, W.; Howard, P. H.; Boethling, R. S. Molecular topology/fragment contribution method for predicting soil sorption coefficients. *Environ. Sci. Technol.* **1992**, *26*, 1560−1567.
(28) US EPA. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.10 2011.
(29) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.
(30) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(31) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733−1746.
(32) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P. Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **2007**, *47*, 1111−1122.
(33) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Chem. Inf. Model.* **2011**, *25*, 533−554.
(34) de Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Than-Luu, R. D-optimal designs. *Chemometr. Intell. Lab.* **1995**, *30*, 199−210.
(35) Fedorov, V. *Theory of Optimal Experiments*; Academic Press: Waltham, MA, 1972.
(36) Olsson, I.-M.; Gottfries, J.; Wold, S. D-optimal onion designs in statistical molecular design. *Chemometr. Intell. Lab.* **2004**, *73*, 37−46.
(37) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab.* **2001**, *58*, 109−130.
(38) OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models; OECD: Paris, France, 2004; http://www.oecd.org/dataoecd/33/37/37849783.pdf.
(39) Valle, S.; Li, W.; Qin, S. J. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods†. *Ind. Eng. Chem. Res.* **1999**, *38*, 4389−4401.
(40) Qin, S. J.; Dunia, R. Determining the number of principal components for best reconstruction. *J. Process Control* **2000**, *10*, 245−250.
(41) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solovev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191−198.
(42) Thormann, M.; Vidal, D.; Almstetter, M.; Pons, M. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Name. *Open Appl. Inf. J.* **2007**, *1*, 28−32.
(43) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability domain for in silico models to achieve accuracy of experimental measurements. *J. Chemometr.* **2010**, *24*, 202−208.
(44) Holmes, G.; Donkin, A.; Witten, I. H. In *WEKA: a machine learning workbench*, Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems;Brisbane, QLD, Australia, November 29−December 2, 1994; IEEE: Washington, DC, pp 357−361.
(45) Wehrens, R. Journal of Statistical Software. *J. Stat. Softw.* **2007**, *18*, 1−24.
(46) R Development Core Team *R: A Language and Environment for Statistical Computing*; R Project: Vienna, Austria, 2011.