From Descriptors to Predicted Properties: Experimental Design by Using Applicability Domain Estimation

Stefan Brandmaier,[1] Sergii Novotarskyi,[2] Iurii Sushko[2] and Igor V. Tetko[1,2]

[1]Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Munich, Germany; [2]eADMET GmbH, Neuherberg, Munich, Germany

Summary — The importance of reliable methods for representative sub-sampling in terms of experimental design and risk assessment within the European Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system is crucial. We developed experimental design approaches, by utilising predicted properties and the 'distance to model' parameter, to estimate the benefits of certain compounds to the quality of a resulting model. A statistical evaluation of four regression data sets and one classification data set showed that the adaptive concept of iteratively refining the representation of the chemical space contributes to a more efficient and more reliable selection in comparison to traditional approaches. The evaluation of compounds with regard to the uncertainty and the correlation of prediction is beneficial, and in particular, for regression data sets of sufficient size, whereas the use of predicted properties to define the chemical space is beneficial for classification models.

Key words: *bagging, distance to model, representative subset selection*.

Address for correspondence: *Stefan Brandmaier, Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, IngolstädterLandstrasse 1, Neuherberg D-85764, Munich, Germany.*
E-mail: *stefan.brandmaier@gmail.com*

Introduction

The sampling of representative compounds delivers a sensitive subset of a defined chemical space. When used for experimental design, it contributes to financial and time efficiency, as it permits the calculation of highly predictive models at a low experimental cost (1). Furthermore, it is relevant for several other tasks in chemo-informatics and computational chemistry, such as drug design (2) and risk assessment (3).

Most applications in chemistry work in the descriptor space (4). After applying an orthogonal transformation by employing a principal component analysis (PCA) to increase the dimensionality of the chemical space, selection algorithms, such as the D-Optimal criterion (5), the Kennard–Stone algorithm (6) or the Most Descriptive Compound (MDC) selection (7), are applied. Bayesian approaches (8–10) that take preliminary information into account, or adaptive approaches that refine the representation of the chemical space in a step-wise manner, are rarely used. Our recent studies investigate iteratively optimised variables to span the search space for experimental design, by combining a partial least squares technique with a dissimilarity selection (11), and a simple descriptor selection with a similarity selection, to increase the efficiency of the experimental design. Having said that, both approaches still work on a representation of the chemical space that uses descriptors as the basis for the selection of compounds.

In this study, we propose a novel adaptive projection that moves from a descriptor-based construction of the chemical space toward a predicted property-based view. We investigate three strategies for experimental design — one of them uses the predicted properties to define the chemical search space, while the others use the concept of the 'distance to model' (DM) parameter, suggested by Tetko *et al*. (12–14), to estimate the uncertainty of prediction for each compound within a data collection. We are not aware of other studies in computational chemistry that use this parameter for compound selection.

The DM-based approaches decide whether to test a compound either exclusively on the basis of the prediction uncertainty, or they combine this parameter with a compound's hypothetical contribution in decreasing the prediction uncertainty of other relevant compounds. The estimation of this contribution is based on the correlation in ensemble predictions, a concept that is the basis of Associative Neural Networks (ASNN; 15), and which is also used in the ASNN Library mode to make local corrections (16).

We statistically evaluated the new approaches on four regression data sets, each with a different endpoint, and on one classification data set. Furthermore, the performances of the newly developed approaches were compared to the performances

of several non-adaptive strategies. It is shown that the use of predicted properties can significantly improve the efficiency of experimental design and decrease the number of experiments required. An investigation on the usability of the approaches and their limitations was also carried out.

Material and Methods

Data sets

Regression data sets

Four regression data sets were used to evaluate the performance of our approaches, and to compare the results with those of other approaches. The endpoint for the smallest of the data sets was the log-scaled bioconcentration factor (logBCF) in fish. This set contained 238 different compounds, and was taken from a study by Gramatica *et al.* (17). The authors originally split the measurements into a training set of 179 compounds and a validation set of 59 compounds. These data sets are freely accessible in the QMRF database of the European Commission (18), as well as in the Online Chemical Modelling Environment Project (OCHEM; 19) database. The data set, as was used in our study, was merged from the original split.

The second data set contained 648 measurements for the adsorption coefficient between the organic partition of soil and water, $\log K_{OC}$. The original source was a collection by Meylan *et al.* (20).

The third data set contained 1093 measurements of toxicity against the protozoan, *Tetrahymena pyriformis*. The endpoint was the negative, log-scaled inhibition of growth concentration (–logIGC50). All the measurements in this data set were taken from our previous study (12), and originated from the Tetratox database (21) and from several studies by Schulz *et al.* (22–24). The fact that all these measurements were obtained by the same laboratory ensured consistency and helped to avoid problems resulting from different experimental procedures or laboratory conditions.

The last regression data set contained 1198 measurements for the boiling point, and was extracted from the Estimation Program Interface (EPI) Suite data (25). It was collected only for halogenated compounds containing bromine, chlorine and/or fluorine.

Inorganic compounds, radicals, charged molecules and salts were removed from all the data sets. In addition, we excluded compounds without an explicit value, where only an interval, or a minimum or maximum value, was given. Except for the boiling point data set, no structural filters were applied, so the sets for logBCF, $\log K_{OC}$ and –logIGC50 represent a wide chemical diversity. All the collections had already been used in our previous studies on representative compound selection (11, 26).

OCHEM (19) was used to arrange the data sets and calculate the descriptors. We used an aggregation of ALOGPS descriptors (27) and E-State indices (28, 29) to chemically represent the compounds in this study. ALOGPS descriptors estimate the lipophilicity and solubility of a compound, whereas the E-State indices are electro-topological descriptors calculated for each atom in a compound. All the descriptors were normalised to [0,1] interval. The rationale to normalise, rather than standardise, the descriptors was that standardisation estimates a normal distribution of a descriptor, which is not the case for E-State indices.

The classification data set

In addition to the regression data sets, we evaluated the performance of our approaches on a binary classification data set, as the requirements for a meaningful sample of a classification data set are most likely to be different to those for a regression set. The set we used contained 7481 measurements of the human CYP 1A2-inhibition activity of small molecules, which were taken from the bioassay AID410 in the PubChem database. The assay data were deposited in October 2007, and the data set was used in a previous study on comparative modelling of cytochrome inhibition (30). The original data set obtained from this bioassay contained 8348 compounds.

Compounds that were labelled 'inconclusive' were excluded from the data set. Furthermore, if the same molecule was present in both the 'active' and the 'inactive' set, it was removed from all the sets. The final distribution of the remaining 7481 compounds was almost balanced, as 4016 were labelled 'active' and 3465 were labelled 'inactive'. As for the regression data set, normalised ALOGPS descriptors and E-State indices were used to represent the compounds.

Predicted properties

Prediction error versus *standard deviation*

The underlying theory on the ensemble-based applicability domain (AD) estimation is that compounds which are predicted with a high reliability are less prone to small variations in the training data set, and that there is a correlation between the uncertainty in prediction and the prediction error. The simulation of the variations in the data set could be done by using a bagging approach to generate a predefined number of subsets by re-sampling with replacement. For this study, we fixed the number of

bags to 64. Each of these subsets is then used to build a prediction model for either an endpoint of continuous values, or a classification model. The resulting collection of models can then be used to predict the target property for new compounds. By receiving not only one prediction but a whole set of them, not only can the average value, which is used as the prediction value, be calculated, but in addition, the variance in the predictions as a measurement of uncertainty can also be determined.

Previous studies (12–15) have shown the correlation between the uncertainty of the prediction and the prediction error. Figure 1 illustrates this correlation for the –logICG50 data set. Each light grey dot represents a compound, the x-axis represents the standard deviation (SD) of the ensemble predictions and the y-axis represents the prediction error. The black line depicts the cumulative error of all the compounds predicted within a certain standard deviation, and the dark grey line depicts the average error of a sliding window.

Independence of descriptors

An additional feature of the predictions derived with the bagging approach is that they define a compound in terms of the property space, instead of the descriptor space. This enables the representation of a data set to a higher extent of independence of a certain descriptor set. To visualise this we calculated DRAGON 6.0 descriptors (31), E-State indices and quantitative name property relationship (QNPR) descriptors (derived from a SMILES representation of the compounds; 32) for the –logIGC50 data set, and reduced all three representations of the data set to two principal components. The results are shown in Figure 2. The colouring of the compound dots is just a topological indication to enable the identification of changes in the data set depiction — it has no functional meaning.

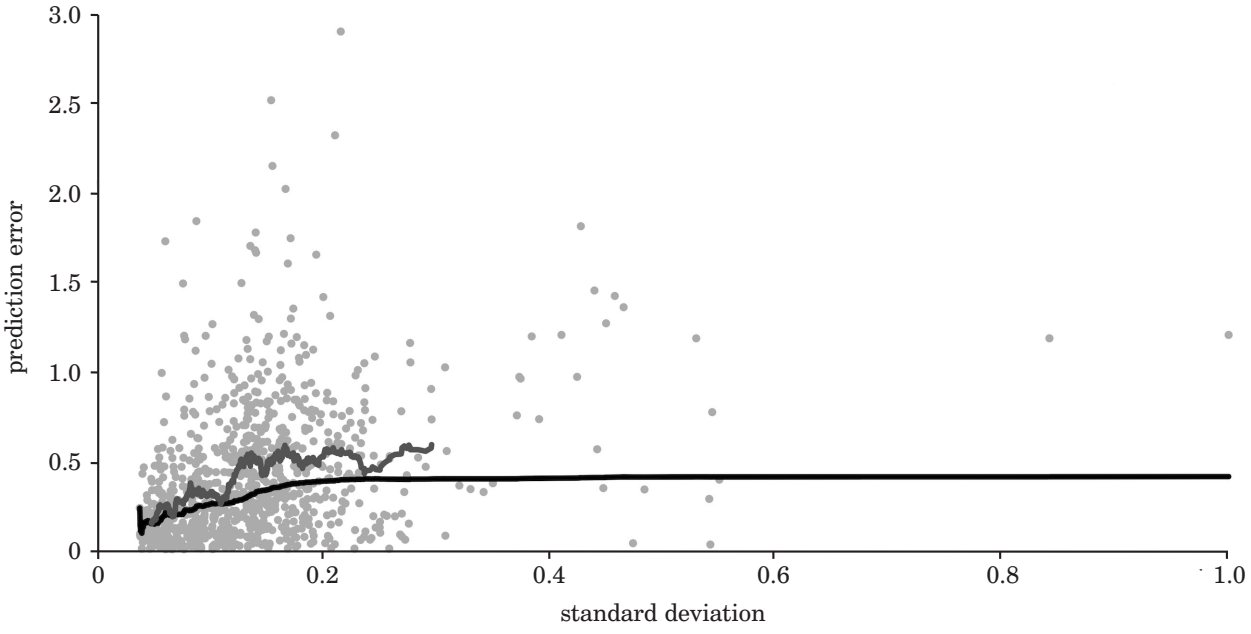
In addition, we used the bagging approach to calculate 64 partial least squares (PLS) regression models (33) on each descriptor set, and reduced the derived predictions for each molecule to two principal components. The results are shown in Figure 2. Again, the colouring has only topological meaning. It is obvious that both the distribution of single compounds (as the overall shape of the distribution), and the variance within the principal components, are harmonised to a greater extent in the predicted properties view.

Selection approaches

Kennard–Stone

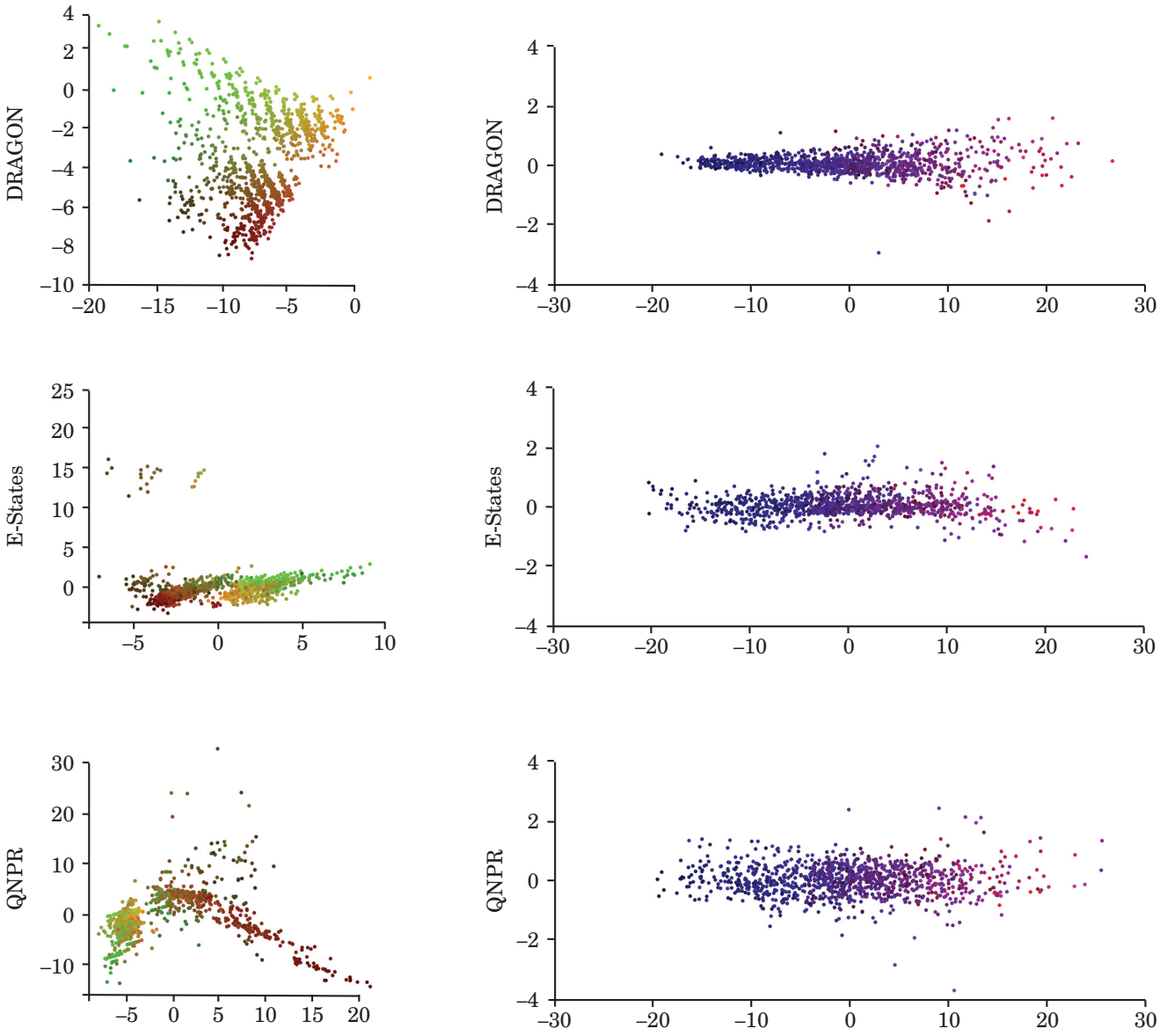
Our implementation of the Kennard–Stone algorithm (6) starts with the most representative com-

Figure 1: The correlation between prediction uncertainty (variation of prediction) and prediction error for a model on environmental toxicity of 1093 molecules against *T. pyriformis*



● = *compounds;* ● = *moving average;* ● =*cumulative error.*

Figure 2: Principal components of a data set on descriptors (left) and predicted properties (right)



Different to a representation by using the principal components derived directly from the descriptors, the representation of the compounds by using predicted properties results in a highly uniform description, independent of the underlying descriptor set.

pound. The most representative compound was defined as the one with the lowest sum of pairwise distances to all other compounds in the data set. From this initially selected compound seed, the Kennard–Stone algorithm works sequentially and selects the compounds in a fixed order. At each step, the next instance to be selected is the one with the greatest distance to its nearest neighbour and which is thereby furthest from the existent selection.

k-Medoid selection

k-Medoid clustering partitions a set of data points into a given number of subsets. Each data point in the underlying data set is thereby assigned to exactly one cluster. The problem of finding the optimal partitioning belongs to the computationally expensive class of NP-hard problems (34). Therefore, iterative heuristics are used to find a local minimum.

Referring to the given number of k expected clusters, k compounds are randomly assigned as cluster centres. In the next step, all remaining compounds are assigned to the nearest cluster, and the cluster centres get reassigned within each cluster to the data point with the lowest sum of pairwise distances to all other compounds in the cluster. This step of reassigning compounds to a cluster centre and reassigning the cluster centres, is iteratively executed until convergence is reached. In this context, convergence means that the cluster centres do not change any more between two reassignment cycles.

As the cluster centre is the point with the lowest sum of pairwise distances to all other points within a cluster, this point can also be seen as the most representative point within the cluster. Therefore, in our experimental design approach, it is returned as the selected instance. This approach was shown to be highly efficient for experimental design in our previous study (26).

Adaptive approaches

Based on a predefined selection of compounds, the stepwise approaches extend this seed in a stepwise manner and according to the predicted properties, and thereby, the AD estimation. After each (hypothetical) measurement cycle, a new ensemble of 64 bagging models is calculated by using PLS regression (33, 35).

AD-Fetcher: The first approach we implemented (referred to as 'AD-Fetcher') uses the predictions of the bagging models and the derived standard deviation for the compound selection. Based on the (simplified) assumption that a high variance in prediction implies a high uncertainty, in each step it selects the compound with the highest standard deviation. This can improve the experimental design performance in two ways: firstly, it extends the existing selection (and implicitly also the resulting model) with a compound that is not yet within the AD and therefore with new information; secondly, the selected compound, which was predicted with a high uncertainty, no longer needs to be measured.

AD-Spider: The second approach not only takes the uncertainty of the prediction into account, but also combines it with the representativeness of a compound, which is deduced from the correlation in the predictions. The underlying estimation is that correlated predictions imply a common mode of action, and that the extension of the selected set by a certain compound also leads to an increased uncertainty and prediction error in correlated compounds.

As for the previous approach, we start with the assumption that the variance in prediction is proportional to the prediction error. Furthermore, we assume that the extension of a model with a certain compound decreases the uncertainty in prediction of another compound in proportion to the correlation in the predictions of these two compounds. A similar concept has been successfully applied for the local corrections in the log P prediction by Tetko *et al*. (16) and is called LIBRARY mode. Both assumptions, as used in this study, are simplifications of a more complex context. Nevertheless, these assumptions should be sufficient for the prioritisation of representative compounds in a data set. In addition, one has to take into consideration that experimental design aims toward efficiency, not toward exhaustiveness.

From our two estimations, it can be inferred that the compound which has to be chosen in each measurement cycle, is the compound that decreases the prediction error of all the remaining compounds to the highest extent. Our implementation of this concept works with one matrix and two vectors: first, the standard deviation vector, S, which contains the compound-wise standard deviation of the predictions; then the correlation matrix, C, which contains the pairwise correlation of the prediction of each compound; and finally, the decision vector, D, which is initially derived from a matrix multiplication of C and S and displays the representativeness of each compound. The first compound to be selected in a measurement cycle is the one with the highest representativeness — it is the compound with the highest correlation to those compounds with the highest variances.

After the selection of a compound, the decision matrix has to be updated to remove the estimated contribution of the recently selected compound. The correction factor for a compound is thereby calculated from the correlation between the recently selected compound and the compound itself. The decision score is multiplied by the difference between one and this correlation.

AD-Descriptors: In addition to the approaches for dealing with the variance in predictions, we implemented stepwise procedures for both the Kennard–Stone algorithm and the k-Medoid clustering to work on the predicted properties, instead of the descriptors. Whereas the non-adaptive selection of these approaches was executed on a fixed number of five principal components derived from the descriptors, the stepwise approach was executed on only three latent variables derived from the 64 predicted properties of each compound. The reason to decrease the number of dimensions was that a PCA on predicted properties usually covers most of the variance in the data set within the first principal component, as shown in Figure 2. Therefore, the influence of any further dimension on the selected compounds would have been marginal, but would have extended the computational complexity.

As the Kennard–Stone algorithm selects compounds based on an initial selection, no modifications to its implementation were required. On the contrary, the k-Medoid implementation was adapted to use the pre-selected compounds as fixed cluster centres, which could not be reassigned, and to select new compounds in reference to these cluster centres.

Validation procedure

To obtain a meaningful statistical basis and to compare the performances of the implemented approaches on both regression and classification data sets, all five collections were split into two partitions. In each data set, 16% of the compounds were excluded from the selection process and used as a respective external validation set. The remaining 84%, referred to as the design set, were used to execute the selection approaches. Two hundred and fifty of these random splits were made and each of them was subject to the same experimental design strategies. For the logBCF and the $\log K_{OC}$ data sets, selections of 5, 7, 10, 15, 20, 25, 30 and 40 compounds were drawn in this predefined order. For the –logIGC50 and the boiling point data sets, additional selections of 50 and 60 compounds were drawn. As the number of compounds in the classification data set was at least six times higher than the number of compounds in the largest regression data set, we selected samples of 10, 20, 30, 40, 60, 80, 100 and 120 compounds.

The selection process for the non-adaptive approaches (random sampling, as well as the k-Medoid clustering and the Kennard–Stone algorithm on principal components) was started from scratch for each sample size. Contrary to the stepwise approaches, the selection process was strictly based on the sequence as mentioned above. Thereby, the compounds selected in each previous step are used in the next step as a known seed; the newly selected compounds just extend this seed.

The performance of each approach was determined with a PLS regression model on all normalised descriptors. PLS regression was also used for the classification data set and the retained continuous values were discretised into two bins. The criterion to determine the number of latent variables for the final model was the best coefficient of determination derived in a ten-fold cross validation (35). The measurement of quality was the root mean squared error (RMSE), as well as the correlation coefficient for the regression data sets. The balanced accuracy (the arithmetic mean of recall of both classes) as well as the F-Measure (harmonic mean of recall and precision) was used to determine the quality of the classification models.

The statistical significance of the different performances (derived by different approaches) was estimated according to a binomial test, by using the binomial distribution with $n = 250$ trials corresponding to the number of models used in our study.

Software Used

PLS models to evaluate the performance of the analysed approaches were calculated by using WEKA (39).

Accessibility of the Data

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

Results

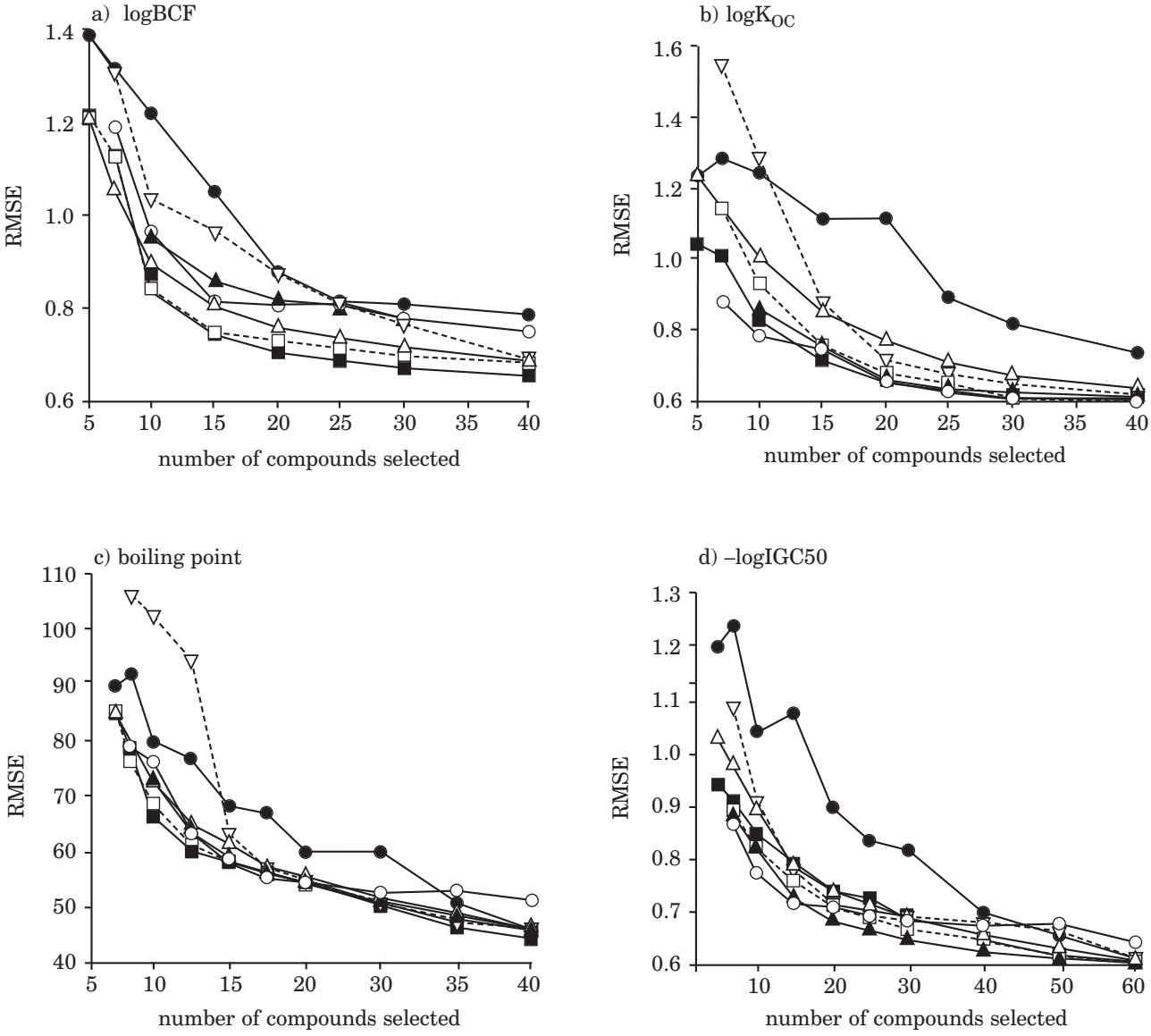
Regression data sets

Prediction error

Figure 3 shows the average error performance on the non-selected compounds for: a) the logBCF data set; b) the $\log K_{OC}$ data set; c) the boiling point data set; and d) the –logIGC50 data set. The x-axis represents the number of selected compounds, and the y-axis represents the average RMSE out of 250 trials. The random selection is illustrated by the open triangles, the k-Medoid approach by the closed squares, and the Kennard–Stone approach by the closed circles. With regard to the underlying data on which the approaches have been executed, the lines are solid for the static approaches on principal components derived from descriptors, or dashed for adaptive approaches on principal components derived from the predicted properties. The AD-Fetcher approach that works only on the distance to model is represented by the open circles, and the AD-Spider approach, which additionally takes the pairwise correlations between the predictions into account, is shown as closed triangles. The same symbols and lines were used in all further figures.

The first observation that can be deduced is that, for all the approaches and for all the data sets, the average error decreased with an increasing number of selected compounds. The second observation is that, for all data sets, the performance of the Kennard–Stone algorithm on principal components derived from descriptors delivered the worst models. Although the use of principal components on predicted properties improved the performance, it was still significantly worse than most other approaches. What is further noteworthy is the high

Figure 3: A comparison of the average RMSE performance on the 250 splits for the regression data sets



△ = *Random;* ● = *Kennard–Stone;* ■ = *k-Medoid;* ▽ = *Kennard–Stone (PP);* □ = *k-Medoid (PP);* ▲ = *AD-Spider* ○ = *AD-Fetcher.*

The k-Medoid clustering performs well on all data sets, the AD-Spider approach performs equally well on the logK$_{OC}$ and the boiling point data set and better on the –logIGC50 dataset.

initial average error for both implementations of the Kennard–Stone approach, as well as the inconsistent development of the error performance for the static implementation. The development of all the other approaches is smoother, approaching a hyperbolic function.

Referring to the logBCF data set, the only approaches that performed significantly better than the random approach were the two k-Medoid approaches (on principal components derived from descriptors and derived from predicted properties). The performances of all the other approaches were worse, at a statistically-significant level ($p < 0.05$). Furthermore, the two approaches that used the AD estimation performed better than the Kennard–Stone approaches. On the logK$_{OC}$ data set, all systematic approaches, except the Kennard–Stone approaches (i.e. AD-Spider, AD-Fetcher, k-Medoid

and k-Medoid on predicted properties), performed equally well and significantly better than the random approach. With regard to its low initial average error, the AD-Fetcher can be seen to be the best-working approach.

The observations on the boiling point data set are similar to those on the $\log K_{OC}$ data set, with the exception that the best initial performance was derived from the clustering approaches and that the improvement to the performance of the random approach was not so significant. In addition, from 40 selected compounds upward, the performance of the AD-Fetcher did not improve any further. Finally, on the –logIGC50 data set, only the AD-Spider approach performed significantly better than the random approach. The k-Medoid approaches showed a similar performance to that of the random approach, whereas the clustering approach on predicted properties consistently performed better than the clustering approach on descriptors. Nevertheless, this difference was not statistically significant. Comparable to the performance on the boiling point data set, the AD-Fetcher had a good initial error performance, but revealed stagnation from 30 selected compounds upward.

The evaluation of performance with regard to the correlation revealed no further insights. The observations were equivalent to those for the RMSE, and for this reason, they are not discussed in detail in this paper. Furthermore, the development of RMSE and correlation on the external validation set was similar to the development on the non-selected data set for all endpoints and methods.

Robustness

To permit a comparison of the stability and reliability of the various approaches, we calculated the SD of the RMSE for all the approaches, for all the data sets. The results can be seen in Figures 4a–d.

We explicitly chose not to show the SD on the same graph as the average RMSE, as this could imply that it is possible to evaluate the significance of an improved performance by the overlapping intervals. In fact, due to the preceding random exclusion of 16% of the compounds from each design set, this is not the case. A sampling of design sets showed that the performance derived from different splits differed by more than two SDs. Nevertheless, the SD of the models derived with the selection approaches is a valid measurement for estimating the uncertainty within one selection approach and for comparing it to that of other approaches.

The observations on the SD are similar for all data sets. Compared to the other approaches, the initial SD of the random approach was higher, but it decreased the fastest. For all other systematic approaches, the standard deviation decreased (and thereby the reliability of the resulting models increased) with a growing number of selected compounds. The SD of the Kennard–Stone algorithm on descriptors revealed quite an inconsistent occurrence: for the logBCF, it increased consistently; for the $\log K_{OC}$ data set, it peaked at 20 selected compounds; for the boiling point data set, the peak was reached at 40 selected compounds; and only for the –logIGC50 data set was the pattern similar to that of the other systematic approaches. Furthermore, the model error development for the boiling point data set was remarkable when considering the reliability of the AD-Spider approach. For the whole range, from 15 to 40 compounds, we compared the models derived for each of the 250 validation splits with the model derived in the previous step. We found that, across the whole range, a minimum of 200 models (80%) improved with any additional selected compound, that is, in 200 of 250 cases the model quality increased when the training set was extended with new compounds, selected by the AD-Spider approach.

Classification data set

With reference to the size of the CYP-inhibition data set, due to the high computational costs of the AD-Fetcher approach (i.e. a total number of 130,000 PLS models is required) and its poor performance, we excluded a full statistical validation of the AD-Fetcher approach with the classification data.

The performance of the other approaches is shown in Figure 5a, and the corresponding SDs are shown in Figure 5b. The y-axis shows the development of the balanced accuracy.

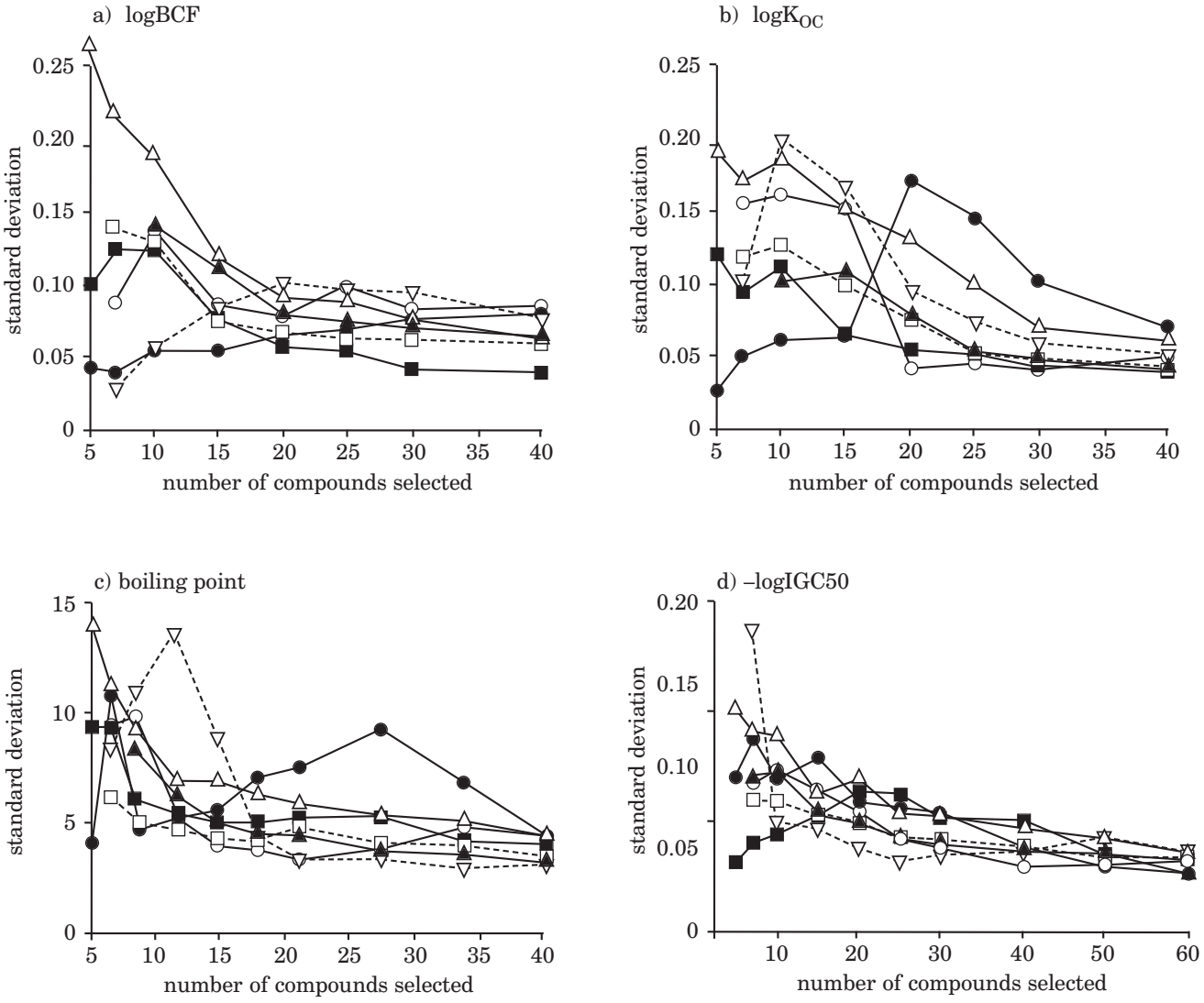
Similarly to the results derived on the regression data sets, the performance of the Kennard–Stone approaches was significantly worse than that of the other approaches. Furthermore, the k-Medoid approaches were within the best methods for compound selection. The performance of the AD-Spider approach was significantly worse than that of the clustering approach, and was also worse than the results derived from a random selection. A comparison of the F-measure as a criterion of prediction quality resulted in the same observations.

Discussion

AD-Spider

The AD-Spider approach, which takes the variance and the correlation of predictions into account, per-

Figure 4: A comparison of the standard deviation derived from the RMSE performance



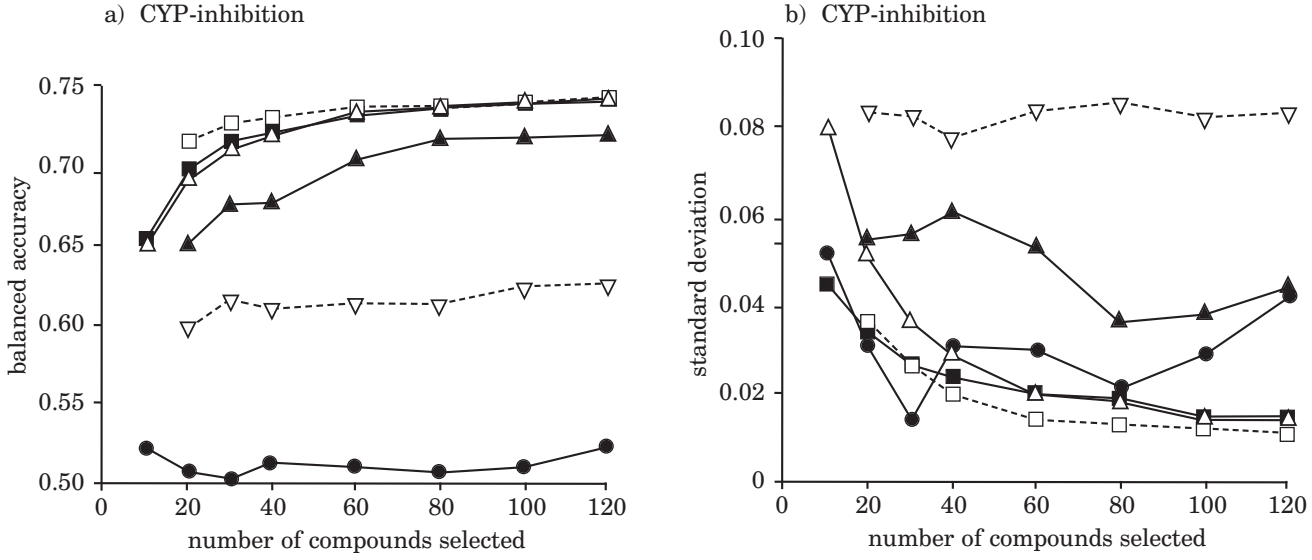
△ = *Random;* ● = *Kennard–Stone;* ■ = *k-Medoid;* ▽ = *Kennard–Stone (PP);* □ = *k-Medoid (PP);* ▲ = *AD-Spider* ○ = *AD-Fetcher.*

formed significantly better than a random selection of the data sets for logK$_{OC}$, the boiling point and –logIGC50. Furthermore, with the logK$_{OC}$ data set and the boiling point, its performance was as good as that of the k-Medoid approach. When compared on the –logIGC50 data set, it performed even better (i.e. with statistical significance). In contrast, the average AD-Spider performance with the logBCF data set was significantly worse than that of a random selection. The reason for this difference can be found in a depiction of the principal components derived from the E-State indices for the data set. Figure A1.1 in Appendix 1 shows that most compounds are within a small subspace, with the remaining compounds widely scattered and sparsely filling the rest of the chemical space.

We therefore attempted a comparison of the approaches, with the same data set, but with different (not fragment-based) descriptors. It was decided to use Inductive descriptors (36) and MERA descriptors (37, 38) for the representation of the compounds. A repeat of the examination of the logBCF data set with those descriptors resulted in a clearly better performance for the AD-Spider approach. Obviously, the AD-Spider approach is not appropriate for scattered compound distributions (see Appendix 1, Figure A1.2).

Furthermore, taking that into consideration, in comparison to the k-Medoid approach, the AD-Spider performs significantly worse for a data set of 238 compounds, equally well for a data set of 648

Figure 5: Balanced accuracy and according standard deviation for the classification data set



△ = *Random;* ● = *Kennard–Stone;* ■ = *k-Medoid;* ▽ = *Kennard–Stone (PP);* □ = *k-Medoid (PP);* ▲ = *AD-Spider.*
The best performing approach is the k-Medoid clustering executed on predicted properties.

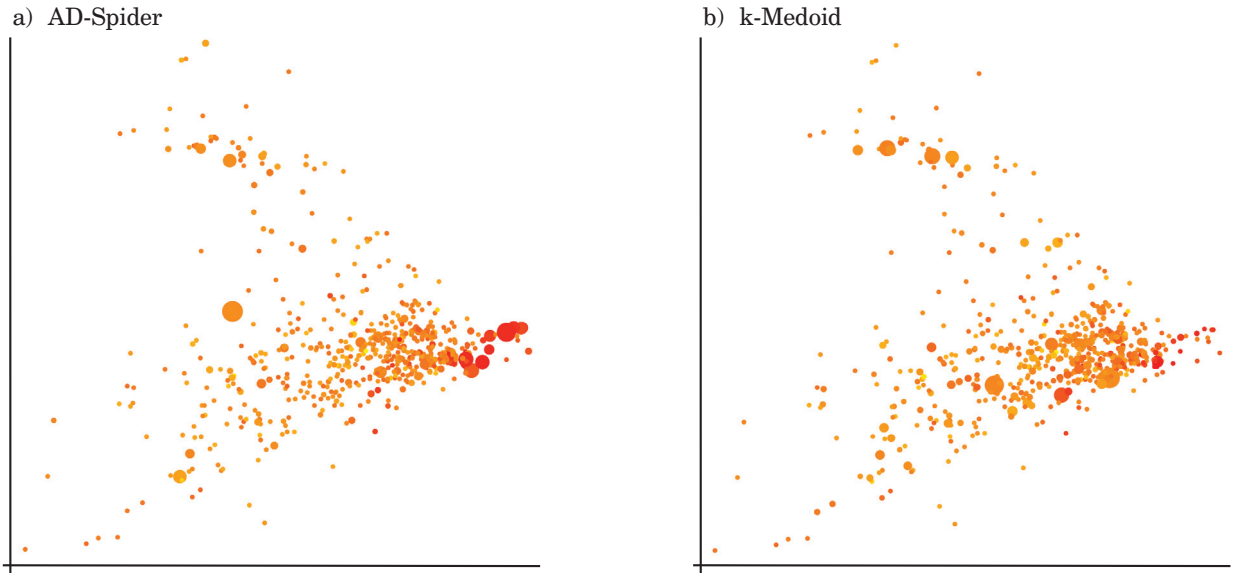
compounds, and significantly better for a set of 1093 compounds. This implies that there is a correlation between the size of the data set and the performance of AD-Spider with it. Such a dependency seems logical as a chemical space that is defined by a lower number of compounds is less densely populated. Therefore, the probability of finding pairwise correlations in predictions between the compounds is decreased or is just arbitrary. As the approach relies on these correlations, small data sets affect its performance.

Referring to the classification data set, the performance of the AD-Spider approach was not able to reach the performance of the best approaches, and in particular, the clustering approaches or the random selection. This can be justified with the use of discretised PLS regression predictions to define the predicted property space. This discretisation can lead to a loss of information, as the resulting variance in prediction differs from the one calculated by the continuous PLS predictions.

To gain a deeper insight into the mechanisms in the approach, we investigated the compounds within the $\log K_{OC}$ data set, because they are highly significant with regard to the quality of the resulting model. Therefore, we built 7000 models based on the data set, each with 20 randomly selected compounds. We used these models to predict the remaining compounds, which had not been used for model building, and calculated the RMSE. For each of the 648 compounds in the data set, we calculated the average RMSE of all the models to which it contributed, and used it as a measurement of representativeness. Finally, we used the selected compounds from the 250 validation trials, and applied AD-Spider to draw 20 compounds, and counted the number of cases each molecule had been selected.

The result of this analysis is shown in Figure 6a. The axes depict the principal components, and each data point represents one compound. Highly representative compounds, which contributed to good models, are coloured red; those contributing to poor models are coloured yellow. The size of the data points indicates how often a compound was selected with the AD-Spider approach. Remarkably, almost all the frequently selected compounds have a high, or very high, representative quality. Figure 6b shows the same correlation for the k-Medoid selection. Although this approach also favours the selection of compounds with good representativeness, compared to the AD-Spider approach, it is not as successful for the highly representative compounds, nor as specific to certain compounds. This leads to the conclusion that the good performance of the k-Medoid approach is, largely, a result of its good statistical coverage of the chemical space, while the good performance of the AD-Spider approach results from its ability to recognise highly representative compounds. We repeated this comparison for the boiling point and the –logIGC50 data set, and the same correlations were observed.

Figure 6: A selection of highly representative compounds



Compounds with high explanatory power are coloured red, others yellow; the size of each dot represents how often a compounds is chosen. The selection of the AD-Spider approach recognises these highly representative compounds better than the k-Medoid clustering.

AD-Fetcher

The performance with the logBCF data set was comparable to that for AD-Spider and, over the whole range, significantly worse than the random approach. The data set on $\log K_{OC}$ is the only one where the AD-Fetcher approach could perform similarly to the AD-Spider and the k-Medoid approaches, and where it performed significantly better than the random approach. For the boiling point data set, it performed well initially. However, from 40 selected compounds, the performance of the approach did not significantly improve any further, which was contrary to all the other approaches tested. The same observation can be made for the –logIGC50 data set. This approach works for fewer than 30 selected compounds, but from this point, its performance is significantly worse than that of the AD-Spider or the k-Medoid approach.

This finding can be explained by the parameterisation employed. The decision for selecting a compound is made exclusively by its variance in prediction. By selecting not only one compound per measurement cycle, but five or ten, we do not ensure that the selected compounds are not correlated. This means that we do not have a mechanism to avoid drawing redundant information within a cycle.

AD-Descriptors

With regard to the regression data sets, the use of a three-dimensional PCA space derived from predicted properties, instead of a search space defined by descriptors or their orthogonal transformation, has to be interpreted in two ways. First, in the case of the k-Medoid clustering, the performance did not significantly change. A clear tendency toward descriptor space or predicted property space was not observable for the error performance, the SD or the correlation coefficient. Only when the reliability in terms of improvement was regarded, was there a slight (but not significant) bias toward favouring the predicted properties. The switch in the search space representation neither improved nor diminished the performance of the selection approach. The robustness of the k-Medoid approach against the dimensionality of the search space had already been shown in our previous study (26). Furthermore, the results of this study indicate that, if the search space takes information about the target property into account, it has no influence on the performance of the approach.

Second, in the case of the Kennard–Stone approach, switching the search space significantly improved the performance with all the regression sets with regard to error and correlation. In the case of the boiling point and the $\log K_{OC}$ data sets, the initial performance with fewer than 20 and 15 selected compounds, respectively, was not improved by the use of a stepwise approach on predicted properties. However, starting from this point significantly improved it. For the other two data sets, the performance improved when predicted properties were used instead of principal components starting with

Table 1: Reference models on the whole data set

Data set	Reference RMSE	Reference balanced accuracy	k-Medoid	AD-k-Medoid	AD-Spider
logBCF	0.65 ± 0.06	—	20	25	N/A
$logK_{OC}$	0.65 ± 0.05	—	20	25	20
Boiling point	45.00 ± 2.20	—	50	60	60
–logIGC50	0.62 ± 0.04	—	40	40	30
CYP inhibition	—	75.4 ± 1.0	150	120	

— = *no data. Shaded boxes indicate the best approaches.*

only seven selected compounds, i.e. the performance improved immediately.

With regard to the classification data set for CYP-inhibition, the use of a predicted property space could improve the performance of the Kennard–Stone algorithm. In addition, the balanced accuracy of the k-Medoid approach was also significantly improved for 20 to 60 selected compounds.

Comparison with models on the whole data set

To permit an overview of the approaches examined, we used OCHEM to calculate reference models for each data set. The reference models were built on the same descriptors as the validation models for the selection, by using PLS regression on a fixed number of three latent variables. For the evaluation, a ten-fold cross validation was used, and one SD was used as a measurement of uncertainty.

For the k-Medoid approach on descriptors as well as on predicted properties and the AD-Spider approach, we investigated the number of compounds that were required to reach a model of the same accuracy. The results can be seen in Table 1. The first column indicates the data set, the second and third columns contain information on the average performance and associated uncertainty, and the following columns display the number of compounds that are required to build a model within one SD of the reference model. The best approaches, referring to the number of required compounds, are indicated in shaded boxes.

The AD-Spider approach delivers the best performance for the $logK_{OC}$ and the –logIGC50 data set, and delivers models with similar performance for only 20 out of 648 (3.1%) and for 30 out of 1093 (2.7%) compounds. The k-Medoid approach on predicted properties is the best performing approach on the CYP-inhibition data set, with 120 out of 7481 (1.6%) compounds.

Conclusion

We showed in this study that the variance in predictions can not only be used to estimate the AD of a model, but also to make an intelligent and purposive selection of representative compounds. A stepwise solution that iteratively refines the depiction of the chemical space depending on prior knowledge is target-oriented and can improve the results.

The attempt to select compounds exclusively by their variation in predictions appears to be inappropriate if not executed in a one-by-one manner. Therefore, the number of suitable applications is limited. On the contrary, the combination of this variance with correlated development in predictions, which putatively indicates a common mode of action, produced very good results. We could show the efficiency of this approach, especially for sufficiently large regression data sets (more than 500 compounds) with a non-scattered distribution of compounds.

The observation that the stepwise use of predicted properties, instead of the static use of principal components, improves the performance of selection approaches is not limited to the Kennard–Stone algorithm. In this study, we also examined its influence on other selection approaches, such as the D-Optimal criterion or the full factorial design. Although these studies were not as exhaustive as the ones presented in this paper, they indicated the same effect.

Except for the logBCF data set, the models derived with the AD-Spider and the k-Medoid approach on predicted properties required less than 5% of compounds of the whole data set to create models for which the performance did not significantly differ from models derived on the whole data set.

Therefore, stepwise, adaptive experimental design approaches that make use of predicted properties for a representative compound selection, are efficient and can be recommended.

Acknowledgement

This study was funded by the EU FP7 project "**CA**se studies on the **D**evelopment and **A**pplication of *In Silico* **T**echniques for **E**nvironmental Hazard and **R**isk Assessment" (CADASTER), grant agreement No. 212668.

References

1. Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, Å., Pettersen, J. & Bergman, R. (1998). Experimental design and optimization. *Chemometrics & Intelligent Laboratory Systems* **42**, 3–40.
2. Taylor, R. (1995). Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *Journal of Chemical Information & Computer Sciences* **35**, 59–67.
3. Lahl, U. & Gundert-Remy, U. (2008). The use of (Q)SAR methods in the context of REACH. *Toxicology Mechanisms & Methods* **18**, 149–158.
4. Eichler, U., Ertl, P., Gobbi, A. & Rohde, B. (1999). Definition of an optimal subset of organic substituents. Interactive visual comparison of various selection algorithms. *Internet Journal of Chemistry* **2**, 1–10.
5. de Aguiar, P.F., Bourguignon, B., Khots, M.S., Massart, D.L. & Phan-Than-Luu, R. (1995). D-optimal designs. *Chemometrics & Intelligent Laboratory Systems* **30**, 199–210.
6. Kennard, R.W. & Stone, L.A. (1969). Computer aided design of experiments. *Technometrics* **11**, 137–148.
7. Hudson, B.D., Hyde, R.M., Rahr, E., Wood, J. & Osman, J. (1996). Parameter based methods for compound selection from chemical databases. *Quantitative Structure–Activity Relationships* **15**, 285–289.
8. Chaloner, K. & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science* **10**, 273–304.
9. Van Den Berg, J., Curtis, A. & Trampert, J. (2003). Optimal nonlinear Bayesian experimental design: An application to amplitude *versus* offset experiments. *Geophysical Journal International* **155**, 411–421.
10. Roy, A., Ghosal, S. & Rosenberger, W.F. (2009). Convergence properties of sequential Bayesian D-optimal designs. *Journal of Statistical Planning & Inference* **139**, 425–440.
11. Brandmaier, S., Sahlin, U., Tetko, I.V. & Öberg, T. (2012). PLS-Optimal: A stepwise D-Optimal design based on latent variables. *Journal of Chemical Information and Modeling* **52**, 975–983.
12. Tetko, I.V., Sushko, I., Pandey, A.K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D. & Varnek, A. (2008). Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *Journal of Chemical Information & Modeling* **48**, 1733–1746.
13. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Kovalishyn, V.V., Prokopenko, V.V. & Tetko, I.V. (2010). Applicability domain for *in silico* models to achieve accuracy of experimental measurements. *Journal of Chemometrics* **24**, 202–208.
14. Sushko, I., Novotarskyi, S., Körner, R., Pandey, A.K., Cherkasov, A., Li, J., Gramatica, P., Hansen, K., Schroeter, T., Müller, K-R., Xi, L., Liu, H., Yao, X., Öberg, T., Hormozdiari, F., Dao, P., Sahinalp, C., Todeschini, R., Polishchuk, P., Artemenko, A., Kuz'min, V., Martin, T.M., Young, D.M., Fourches, D., Muratov, E., Tropsha, A., Baskin, I., Horvath, D., Marcou, G., Muller, C., Varnek, A., Prokopenko, V.V. & Tetko, I.V. (2010). Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *Journal of Chemical Information & Modeling* **50**, 2094–2111.
15. Tetko, I.V. (2009). Associative Neural Network. *Methods in Molecular Biology* **458**, 180–197.
16. Tetko, I.V., Poda, G.I., Ostermann, C. & Mannhold, R. (2009). Large-scale evaluation of log P predictors: Local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chemistry & Biodiversity* **6**, 1837–1844.
17. Gramatica, P. & Papa, E. (2005). An update of the BCF QSAR model based on theoretical molecular descriptors. *QSAR & Combinatorial Science* **24**, 953–960.
18. JRC (2008). *(Q)SAR Model Reporting Format Inventory*. Ispra, Italy: European Commission Joint Research Centre. Available at: http://qsardb.jrc.it/qmrf/ (Accessed 10.02.13).
19. 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. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *Journal of Computer-aided Molecular Design* **25**, 533–554.
20. Meylan, W., Howard, P.H. & Boethling, R.S. (1992). Molecular topology/fragment contribution method for predicting soil sorption coefficients. *Environmental Science & Technology* **26**, 1560–1567.
21. Schultz, T.W. (1997). Tetratox: *Tetrahymena pyriformis* population growth impairment endpoint — a surrogate for fish lethality. *Toxicology Mechanisms & Methods* **7**, 289–309.
22. Schultz, T.W., Netzeva, T.I. & Cronin, M.T.D. (2004). Evaluation of QSARs for ecotoxicity: A method for assigning quality and confidence. *SAR & QSAR in Environmental Research* **15**, 385–397.
23. Aptula, A.O., Roberts, D.W., Cronin, M.T.D. & Schultz, T.W. (2005). Chemistry−toxicity relationships for the effects of di- and trihydroxybenzenes to *Tetrahymena pyriformis*. *Chemical Research in Toxicology* **18**, 844–854.
24. Schultz, T.W., Hewitt, M., Netzeva, T.I. & Cronin, M.T.D. (2007). Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. *QSAR & Combinatorial Science* **26**, 238–254.
25. US EPA (2012). *Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11*. Washington, DC, USA: US Environmental Protection Agency. Available at: http://www.epa.gov/opptintr/exposure/pubs/episuite.htm (Accessed 04.02.13).
26. Brandmaier, S., Tetko, I.V. & Öberg, T. (2012). An

evaluation of experimental design in QSAR modelling utilizing the k-medoid clustering. *Journal of Chemometrics* **26**, 509–517.
27. Tetko, I.V. & Tanchuk, V.Y. (2002). Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 Program. *Journal of Chemical Information & Computer Sciences* **42**, 1136–1145.
28. Kier, L.B. & Hall, L.H. (1990). An electrotopological-state index for atoms in molecules. *Pharmaceutical Research* **7**, 801–807.
29. Hall, L.H. & Kier, L.B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *Journal of Chemical Information & Computer Sciences* **35**, 1039–1045.
30. Novotarskyi, S., Sushko, I., Körner, R., Pandey, A.K. & Tetko, I.V. (2010). Classification of CYP450 1A2 inhibitors using PubChem data. *Journal of Cheminformatics* **2**, Suppl. 1, 40.
31. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match Communications in Mathematical & in Computer Chemistry* **56**, 237–248.
32. Thormann, M., Vidal, D., Almstetter, M. & Pons, M. (2007). *Nomen Est Omen*: Quantitative prediction of molecular properties directly from IUPAC name. *Open Applied Informatics Journal* **1**, 28–32.
33. Geladi, P. & Kowalski, B.R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta* **185**, 1–17.
34. Mahajan, M., Nimbhorkar, P. & Varadarajan, K. (2009). The planar k-means problem is NP-hard. *WALCOM: Algorithms & Computation* **5431**, 274–285.
35. Wold, S., Sjöström, M. & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics & Intelligent Laboratory Systems* **58**, 109–130.
36. Cherkasov, A., Ban, F., Santos-Filho, O., Thorsteinson, N., Fallahi, M. & Hammond, G.L. (2008). An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *Journal of Medicinal Chemistry* **51**, 2047–2056.
37. Potemkin, V. & Grishina, M. (2008). A new paradigm for pattern recognition of drugs. *Journal of Computer-aided Molecular Design* **22**, 489–505.
38. Potemkin, V.A., Pogrebnoy, A.A. & Grishina, M.A. (2009). Technique for energy decomposition in the study of "receptor–ligand" complexes. *Journal of Chemical Information & Modeling* **49**, 1389–1406.
39. Holmes, G., Donkin, A. & Witten, I.H. (1994). WEKA: A machine learning workbench. In *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, pp. 357–361. New York, NY, USA: Institute of Electrical and Electronics Engineers.

Appendix 1

Figure A1.1: Principal components derived from various descriptors

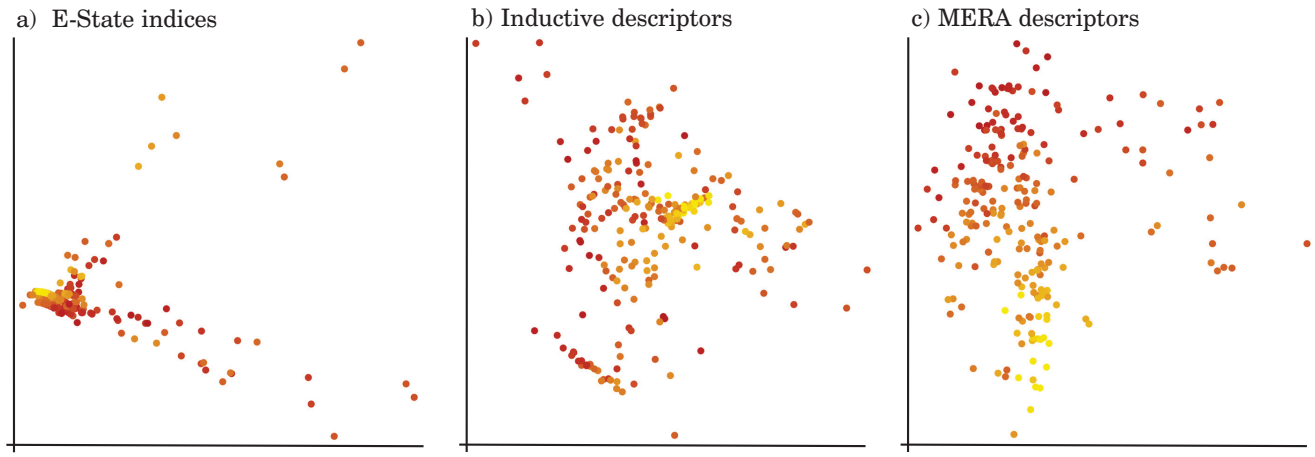
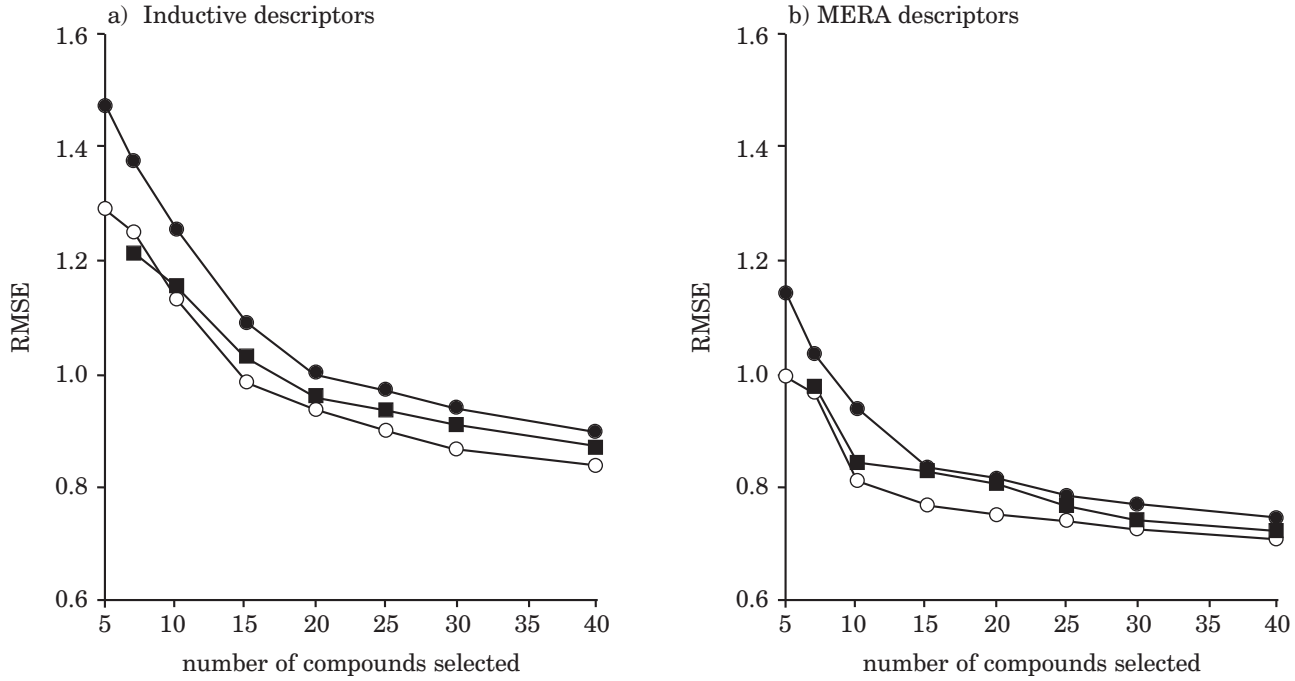


Figure A1.2: Comparison of the performance of a random approach with the k-Medoid clustering and the AD-Spider approach on various descriptors



● = *Random;* ■ = *k-Medoid (PP);* ○ = *AD-Spider*