Linnæus University

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Final project report /2011

2 February 2012

Experimental design in QSAR modeling

Duration of Short Term fellowship:
18.09.2011-17.12.2011

Early stage researcher:
Stefan Brandmaier

Project supervisor:
Professor Tomas Öberg

Research Institution:
Linnaeus University Kalmar

1. Introduction

The REACH legislation contains the demand, that each chemical compound produced in or imported into the EU in an amount of more than one ton has to be registered respective to a number of endpoints. Experimental determination of these properties for all compounds would require a high throughput testing. According to Rovida and Hartung the financial requirements for such testing are about 9.5 billion Euro. For potentially hazardous, dangerous or hardly degradable substances, the registration requires also information about their bioaccumulation and toxicity. Apart from cost and time efficiency - a sample for e.g. bioconcentration requires around two months of time and can bring up costs of more than 200 Euro, - this leads also to ethical problems, as experimental determination of endpoints associated with toxicity and bioaccumulation are realized, utilizing animal tests.

The necessity to keep the overhead of (animal) tests as small as possible is also important in many other research areas, e.g., in chemical or pharmaceutical industry. One commonly used strategy to address this problem is to use structure-activity modeling 3 and to predict the required properties rather than to perform experimental measurements. This strategy is implemented by testing only a small subset of all compounds of interest and building a predictive model using the experimentally determined values.

This basic task can be reduced to the problem of drawing a representative subsample of a bigger set, which is crucial for many tasks in chemo-informatics and QSAR modeling. This basic step was reported as important for experimental design and risk assessment within REACH, large scale in silico scanning for drug target evaluation and QSAR development. Therefore a variety of "implementations" to solve that problem was developed . All these approaches can be linked to three different basic ideas. Firstly the concept of trying to draw a selection that covers the whole range of the descriptor space, e.g. full or fractional design or space filling design, secondly the idea of selecting the most diverse subset of compounds, e.g. Kennard-Stone or D-Optimal design and thirdly the aim to select the most representative subset, e.g. most descriptive compound selection (MDC).

All these methods deliver advantages to certain problems but also disadvantages were reported for most of them. MDC has a bias towards central data points and disregards the periphery. Approaches aiming to select a maximum diversity subset have a tendency towards selecting outliers, especially in high dimensional spaces and for selection of only a low number of compounds. For space filling designs that try to cover the whole descriptor space, the number of compounds to be selected cannot be fixed, as there are cells in the chemical space, that are unoccupied, as due to the laws of chemistry no compound with the required qualities exist. Space filling designs work well for a selection of equally distributed compounds, but reveal problems for inhomogeneous distributions. Apart from that, the number of subspaces, for which a representative has to be selected, is exponentially increasing with the number of dimensions in the search space. It enables only the use of a low dimensionality search space.

Furthermore, all aforementioned approaches provide selection of 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 statistical literature also provides a large variety of sequential approaches, their application in QSAR is very limited. Sequential approaches are arranged in a stepwise procedure and adapt to the gathered information about the response. In theory, the sequential testing strategy could provide a better selection of compounds by taking into consideration the correlation of molecular descriptors to the target property.

Apart from that, in the recent past, approaches using density based or hierarchical clustering were suggested. Also partition-based approaches, utilizing the k-Means clustering were introduced, but these approaches use the derived clusters to apply other selection algorithms to them. Although pharmaceutical publications mention the possibility, to use the clusters derived by k-Means, to select exactly one representative from each cluster, we are not aware of a study evaluating this technique in QSAR experimental design and comparing its performance to other experimental design techniques. The idea to assign the compounds to different clusters and choose a representative form each cluster seems to be appropriate for chemical compound selection, as the implied separation of the chemical space into clusters is adaptive to the real distribution of compounds and not to a hypothetical distribution.

2. Aims within the fellowship

The aims within my short term fellowship contained two parts:

- **To investigate an adaptive, stepwise experimental design strategy that is based on the D-Optimal approach.**

  We developed a method that combines D-Optimal design with partial least squares techniques to iteratively refine the descriptor space for the compound selection. This refinement is realized by the usage of 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 procedure. We evaluated the performance of the new approach on four datasets and compared it to the original D-optimal design.

- **To investigate the usage of the k-Medoid clustering algorithm to do a representative compound selection in terms of experimental design.**

We compared this approach to several other approaches that aim to give a representative selection, like space filling design, the D-Optimal criterion, the Kennard-Stone algorithm and the most descriptive compound approach, to evaluate its usefulness for experimental design. We applied these approaches to four datasets with different specifications, all of them with relevance for REACH and risk assessment, and compared the performance of models resulting of the selected compounds.

3. Preliminary requirements

3.1. Dataset collections

To validate the performance of the stepwise method, four datasets with different endpoints were collected from literature. All of the selected endpoints have relevance 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 were collected to vary in several criteria, i.e. size, modeling and measurement complexities.
The selected endpoints included two toxicity measurements, namely the log scaled lethal concentration for fathead minnow (logLC50), inhibition growth concentration for T. pyriformis (-logIGC50), an adsorbtion coefficient (logKOC) and the boiling point. The number of compounds in these datasets ranged from 96 (-logIGC50) to 1198 (boiling point). The logLC50 dataset contained 535 compounds and the logKOC dataset had 648 compounds.

For all four sets we excluded inorganic compounds, radicals, charged molecules and salts. Moreover, we removed compounds for which no exact values, but an interval or only a minimum or maximum values were given. For the compounds in the logLC50 and in the logKOC datasets, no structural filters were applied. Therefore the datasets contained a wide variation of different compound classes, had a wide structural diversity and the resulting models can be identified as ‘global’ ones. For boiling point, a diversified filter was applied to the structures, limiting the compounds in the final dataset to halogenated ones, containing bromine, fluorine and / or chlorine. The initial -logIGC50 dataset contained more than one thousand compounds. However, to evaluate the performance of the developed approach on a relatively small dataset, a subset containing 96 compounds was randomly selected.

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

These four sets were collections of literature values and not necessarily intended for model building. Furthermore, they were without an explicit indication regarding the dependency between the general descriptors we decided to use and the endpoint. As we wanted to extend our studies on the k-Medoid approach also to evaluate the performance of the selection approaches on a collection of compounds, for which a clear linear dependency between the endpoint and a selection of descriptors was reported, we additionally selected a dataset that was the fundament for a published model. The dataset we decided for was taken from the OMRF database of the European commission and used for the reviewed bioconcentration factor model in fish, published by Gramatica et al.

The model to predict the logBCF value was trained on 179 compounds and validated on 59 compounds. The model was using a set of five descriptors derived by the DRAGON software 25, reporting a Q2 of 86.4, an R2 of 90.5 and an RMSE of 0.57 on the external dataset. For our study, we merged the trainings and validation set and used the five descriptors, reported in the publication, to represent the compounds.

3.2. Implementation of selection approaches

3.2.1. Partition based selection

This approach is based on the ideas of full factorial design, space-filling design and partition-based approaches. It works by partitioning the chemical space of relevance into subspaces and selecting the most representative compound from each subspace. The implementation of this idea is realized by dividing the axes of the chemical space into bins. Thereby the number of bins within each axis is equal. As the number of subspaces is exponentially increasing with each additional dimension in the chemical space, we fixed the approach to work on three dimensions. And as compounds in the chemical space are usually not equally distributed, the number of bins, each axis is separated into, is not a fixed one, but automatically detected, regarding the number of compounds to be selected. Therefore the number of bins is increased as long, as the number of subspaces, occupied by at least one compound, is not higher than the number of compounds to be selected. Finally from each occupied subspace the compound with the lowest Euclidean distance to the center of the subspace is selected. As this approach is focusing on the separation of the chemical space and not on the distribution of the compounds, the number of selected compounds can be smaller than desired.

3.2.2. Most descriptive compounds

The most descriptive compound selection (MDC) aims to select compounds that are located in the dense regions of the chemical space and therefore highly representative for the other compounds of interest. The algorithm is based on the pairwise distances of the compounds and the deduced information content for all other compounds. The compounds are selected sequentially and after each newly selected compound, the contribution of that compound is eliminated. The implementation also provides a stop criterion, which limits the number of compounds to be selected. As this study concentrates on a comparison for a fixed number of compounds selected, the compounds are used, regarding their selection order.

3.2.3. Kennard-Stone algorithm

Similar to the MDC algorithm, also the Kennard-Stone algorithm selects compounds in a fixed order. Derived from an initial selection, the compounds are selected sequentially. To find the next compound to be selected, the pairwise Euclidean distance of each candidate compound to its nearest already selected neighbor is calculated. The compound with the highest distance to its nearest neighbor and which is thereby furthest from the existent selection, is selected. In this study, the initial selection was a randomly chosen data point.

3.2.4. D-Optimal criterion

D-Optimal design selects the most distinct combination of compounds. Therefore each possible subset of a given size is evaluated according to the D-Optimal criterion. The model matrix of a subset is used to derive, the information matrix. The most distinct and thereby optimal of all possible subsets is the one with the maximum determinant of the information matrix. The implementation of the D-Optimal selection criterion was according to literature specifications and utilizing the Fedorov heuristic to optimize the speed of the selection.

As the application of the D-Optimal criterion is known to work well for linear dependencies between the principal components and the target property, but revealing problems for dependencies of higher order, it was also applied to a set of meta-descriptors, that contained the normalized principal components, their cross and square terms. This additional feature increases the search space by a quadratic factor. The compounds selected with this enhancement are not anymore exclusively located at the periphery of the dataset, but also in the central regions.

3.2.5. K-Medoid approach

The fundament for the selection approach, we investigated within the fellowship was an implementation of the k-Medoid clustering. The k-Medoid clustering, which is a type of k-Means clustering, partitions a set of data points into a given number of subsets, the clusters. Each data point is assigned to only one cluster and each cluster must contain at least one data point.

For the k-Medoid approach, a number of k randomly selected data points gets initially assigned as cluster centers. In the second step, each data point, that is not assigned to be a cluster center, gets assigned to the nearest cluster center (regarding the Euclidean distance). In the third step, the cluster centers are reassigned to that data point within a cluster, which has the lowest sum of pairwise distances to all other data points in the cluster. Now steps two and three are alternatingly executed, until convergence is reached, which means, that the clusters, and thereby also the cluster centers do not change anymore. For each assigned cluster a representative is returned, which is in our case the cluster center. As the cluster center is the point with the lowest sum of pairwise distances to all other points within a cluster, this point can be also seen as the most representative point within the cluster.

3.3. Implementation of a stepwise selection procedure

The utilization of the stepwise approach has two phases. Firstly the application of the extended D-Optimal design, that takes preselected compounds into consideration and, secondly, an implementation of Partial Least Squares regression (PLS) 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. But contrary to PCA components, which are selected to maximize the variance of the dataset (i.e., to cover as much as possible variability of data), the PLS latent variables are selected to maximize the covariance (i.e., 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. By this modification, the representation of the compounds of interest

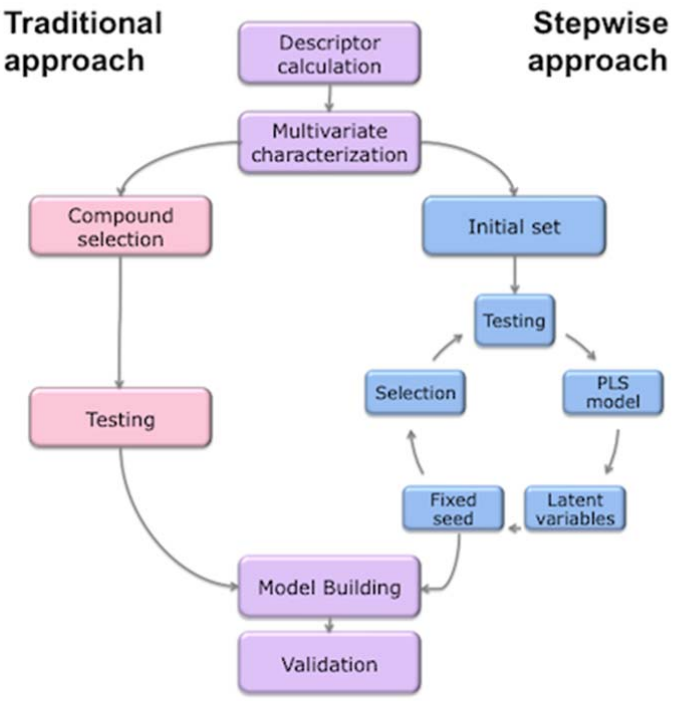

Figure 1. Comparison of the traditional workflow (left) and the suggested stepwise selection (right)

is adjusted to the considered endpoint, and it is not anymore depending 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 seed, 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. During all further steps, the compounds selected in the previous steps are considered as being 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. Furthermore, all preliminary tested compounds are fixed 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 is shown in Fig. 1. Whereas the traditional method (left side of the figure, colored pink) selects all compounds at the same time, the stepwise approach (right side of the figure, depicted blue) constantly increases the number of compounds in cyclic way. Furthermore the chemical space to represent the compounds is refined with each cycle.

3.4. Validation

To obtain a meaningful statistical fundament to compare the performance of different approaches, from each dataset 100 subsets (design sets) were generated. The compounds in the subsets were chosen randomly and the size of each subset was 75% of the whole dataset. The detained 25% of the compounds were used as respective external validation sets. Each of the design sets was used for the experimental design.

In order to receive comparable information about the quality of the compound selection, PLS was utilized to train a regression model on the selected compounds. The number of latent variables to be used 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. The reason, to choose PLS for evaluating 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 performance of the developed model was then calculated for the exteral validation Split and RMSE was calculated as a measurement of error. The mean value of RMSE on the 100 models calculated for each dataset was then used to compare the quality of experimental designs for PLS-Optimal and the traditional method.

4. Results

4.1. PLS-Optimal

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 datasets another examination within the range from 5 to 35 selected compounds was initiated. We were using the meta descriptors containing the normalized components, 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 the number of compounds to select. This means 1 component for less than 6 compounds selected, 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 compounds selected.

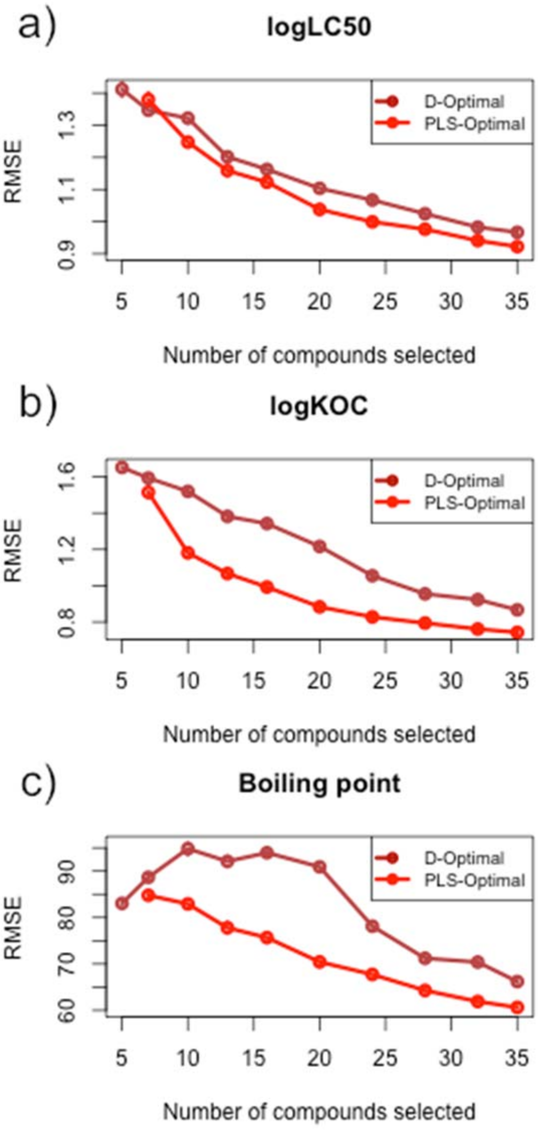

Figure 2. Results of the error validation for cross and square terms on a low number of compounds

The results in Fig. 2a –c show, that the stepwise approach reaches a clearly better performance for all three endpoints. This improvement is significant ($p < 0.001$) for the whole range from 10 to 35 selected compounds. In case of the logKOC dataset (Fig. 2b) 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 (Fig. 2c), the average RMSE performance for 24 compounds selected with the traditional approach could be reached with only 13 compounds selected in a stepwise procedure. Furthermore, in the range from 13 to 32 compounds selected, the improvement of the average RMSE for the same number of compounds selected is better by at least 9 degrees. For logLC50 (Fig. 2a), the average performance with 24 compounds selected in a stepwise procedure could not be reached with less than 32 compounds selected based on principal components and in case of the logKOC dataset, the stepwise approach delivers an average performance for 13 compounds selected, that cannot be reached with less than 24 compounds utilizing the traditional method. The RMSE for that dataset was in average decreased by 21% in the range from 10 to 35 compounds selected.

Finally, comparing the results of the stepwise approach applied to a sequence of 10, 20 and 30 compounds selected with those of the stepwise approach applied to the increased step size, the latter delivers an increased model quality for the same number of compounds selected. The average RMSE for 28 compounds selected using the smaller step size is 0.19 log units better for the logKOC dataset and 0.03 log units for logLC50 dataset.

The models built on PLS-Optimal design deliver a more stable performance regarding the error development for all four examined endpoints. Whereas for the classic approach the performance shows some variability and deviations for an increasing number of selected compounds, the development of the performance of the PLS-Optimal design is much more smooth and approximates a hyperbola function. This is observable even for a search space of only three components 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 endpoint, whereas the principal components are just ranked by their variance. This leads do an accumulation of irrelevant information in the principal components. Therefore, the required number of principal components to catch up the same amount information for an endpoint is usually higher than the required number of PLS latent variables. This is important, both, in terms of stability and in terms of efficiency, 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 use of traditional D-optimal design.

4.2. K-Medoid approach

Fig.3 a-f) shows the results of the validation on the selection approaches for the logLC50 dataset (a-b), the logKOC dataset (c-d) and the boiling point dataset (e-f). The number of principal components used in each column is 3 (left column) and 7 (right column). Exceptions hereby are the random selection, as it is independent of the number of latent variables and the space filling design, which was performed on a fixed number of 3 latent variables. Furthermore, the performance derived by a selection on the D-Optimal criterion using cross and square terms was only examined for three latent variables. The x-axis in each figure indicates the number of compounds selected and the y-axis shows the average RMSE performance. The initial cluster centers for the k-Means approach in this section were assigned randomly.

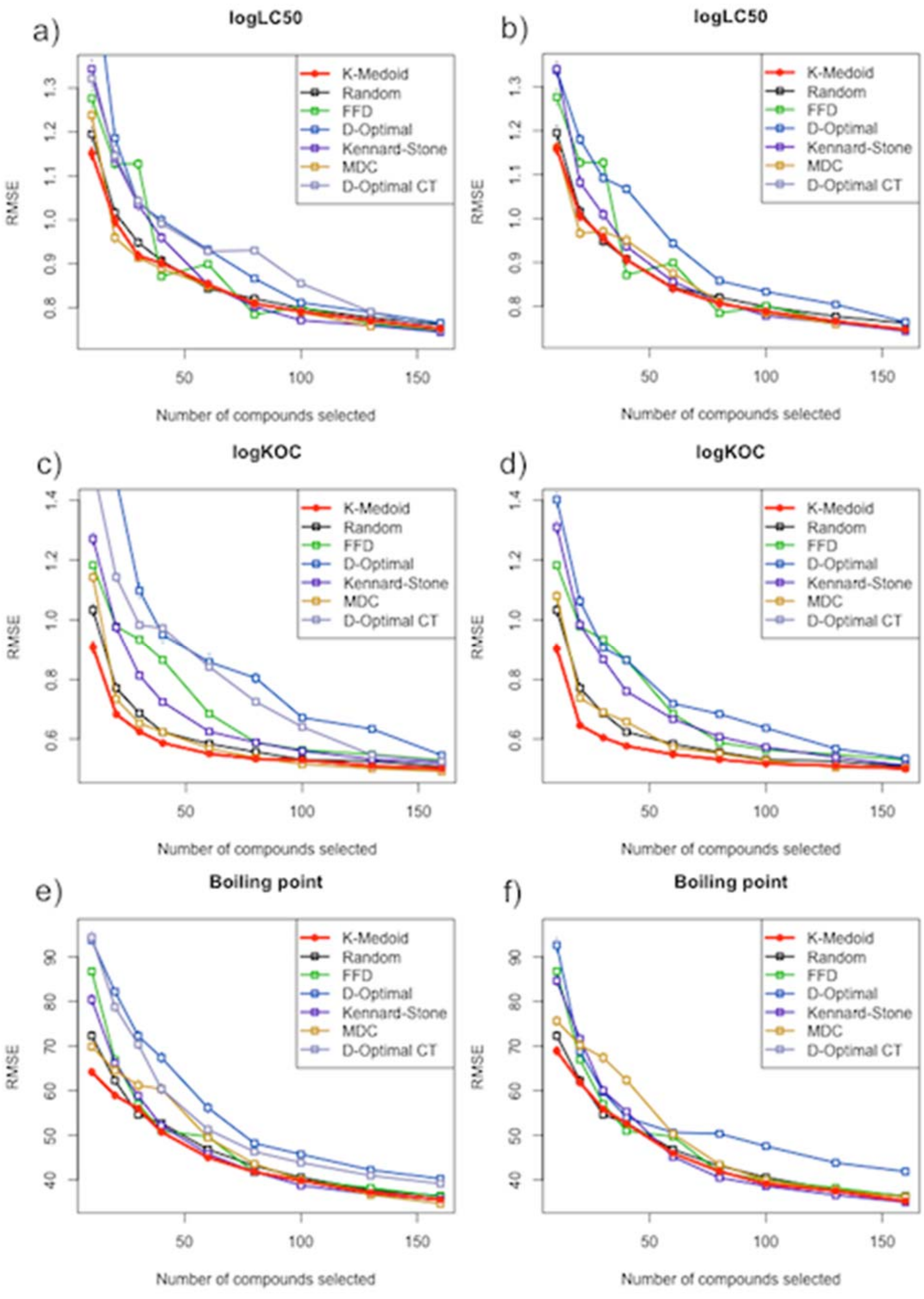

Figure 3. Performance of the approaches on the logLC50 dataset, the logKOC dataset and the boiling point dataset, with 3 and 7 latent variables. k-Medoid is displayed in red, the space filling design in green, Kennard-Stone purple, the D-Optimal criterion is blue or bright blue for linear or a quadratic search space and the MDC approach is displayed yellow. The random selection is displayed by the black curve.

The first observation is, that with an increasing number of selected compounds also the model performance improves. This observation applies for all approaches and it is expected, since a larger number of molecules allows developing better models. A second observation is that for all datasets, the random selection performs well. Also this is expected, as there are publications that report better performance for random approaches, than for elaborated approaches.

Figure 1. Performance of the approaches on the logLC50 dataset, the logKOC dataset and the boiling point dataset, with 3 and 7 latent variables. k-Medoid is displayed in red, the space filling design in green, Kennard-Stone purple, the D-Optimal criterion is blue or bright blue for linear or a quadratic search space and the MDC approach is displayed yellow. The random selection is displayed by the black curve. The x-axis represents the number of selected compounds and the y-axis represents the average RMSE of 100 trials.

Furthermore and of more interest, for all datasets, for the whole range of compounds selected and for each number of latent variables to define the search space, the performance of the k-Medoid approach (displayed by the bold red curve) is within the best. Compared to all other approaches, it has the best initial performance and a fast decrease of the error.

Especially for a small number of compounds selected, in the range from 10 to 30 compounds, the models derived from the k-Medoid selection perform better than for any other approach, except the MDC selection on the logLC50 dataset. This improvement is statistical significant. Regarding a binomial test from the direct method, using the Binomial distribution and 100 trials, the p-Value is lower than 0.01. The best initial performance for 10 compounds selected is in all nine examples derived from the selection of the k-Medoid approach. In the range from 80–160 compounds selected, the performances of the models derived by the k-Medoid approach, the random selection, the MDC approach and the space filling design converge and come up with a comparable performance.

Contrary to other approaches, the development of the error for the k-Medoid approach describes a permanent curve with a constantly increasing incline, without the inconstancies, that can be observed for the MDC approach, the space filling design and also the D-Optimal design. The development of the performance of the k-Medoid approach is more smooth and approximates a hyperbola function, regardless of the number of principal components used or the dataset.

A reason for the good performance of the k-Medoid approach is that it approach unites the advantages of the three basic ideas, whereas it minimizes their disadvantages.

- Like a space filling design, it covers the whole chemical space, but respective to the real distribution.
- Like approaches based on the selection of the most distinct compounds, each point in periphery of the data cloud is represented in a cluster.
- From each cluster the most representative compound is selected, as the criterion of the minimal distance is applied.

Furthermore worth mentioning, the k-Medoid approach is not subject to restrictions like other approaches. Whereas for the space filling design it is impossible, to fix the number of finally selected compounds, with k-Medoid approach the resulting number of compounds can always be precisely defined. k-Medoid has no stop criterion, like MDC and even a small number of compounds can be selected from a high dimensional search space, which is not possible with the D-Optimal criterion, as the number of compounds to be selected, must be higher than the number of principal components.

5. Summary and outlook

The results for the stepwise approach, presented in this report, were limited to the application of the D-Optimal criterion to PLS latent variables. The concept of taking the correlation or covariance to the target property into account could be realized with any other selection criterion. Furthermore, the usage 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.

The sequential approach, we suggest, could also be extended to a Bayesian one, just by performing the initial selection on the latent variables derived from a model, built on measurements, collected by a literature research.

The manuscript about the study on the usage of PLS latent variables, instead of principal components for experimental design is finished and will be submitted to a peer reviewed journal within the next days. The manuscript about the usage of k-Medoid clustering for experimental design in QSAR is reviewed by the authors and will be submitted to a peer reviewed journal, as son as his procedure is finished. A third study, focussing on the robustness of several standard approaches for experimental design and the comparison of these approaches to a newly developed approach, could not be finished within the three months period, but will be continued within the next months.