HelmholtzZentrum münchen
German Research Center for Environmental Health

CADASTER

Stepwise D-Optimal design based on latent variables

Stefan Brandmaier[1], Ahmed Abdelaziz[1], Ullrika Sahlin[2], Tomas Öberg[2], Igor V. Tetko*

[1] Institute of Bioinformatics & Systems Biology, 85764 Neuherberg, Germany and [2] Linneaus University, School of Natural Sciences, SE-391 82 Kalmar, Sweden

Introduction and Motivation

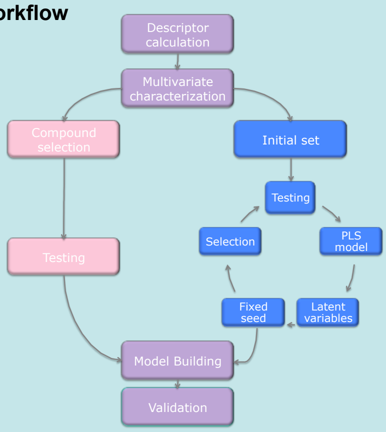
In the course of REACH, each chemical compound produced in or imported into the EU in amount of more than 1 ton has to be registered according to a number of environmental endpoints, including bioaccumulation and toxicity. Experimental determination of these properties requires a high number of animal tests. Apart from ethical reasons, animal experiments are expensive and time consuming. Therefore, the number of these tests should be kept as small as possible. This can be achieved by testing only a small representative subset of compounds, using them to build QSAR models and predict the remaining compounds. The challenge is now, to select exactly that combination of compounds, that delivers the most reliable model. Approaches that are aimed to solve the problem of selecting a representative subset of compounds for testing, are summed up by the expression 'experimental design'.

There are several standard approaches for the selection of diverse sets of compounds for model purposes, such as factorial [1] or D-Optimal [2] design. The later method is frequently considered to be a better choice [2]. The D-optimal design selects compounds using principal component analysis (PCA) of molecular descriptors. The analysis is done in one step and does not take into account the target property. Therefore, the selected compounds may not be optimal for modeling of the given property. Taking a look at the practical course of action in laboratories, the modus operandi of the standard approaches, to select all compounds in one single step, seems to be quite artificial. Most labs, e.g. because of restricted capacities, test compounds not in parallel but in a stepwise procedure. The question is whether there is a better strategy that could provide better selection of compounds by taking into consideration the target property and available data.

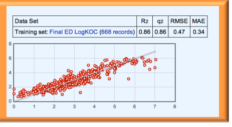
Material and Methods

A stepwise solution for experimental design, that utilizes the D-Optimal approach and combines it with partial least squares techniques to iteratively refine the descriptor space for the compound selection, by the usage of the PLS latent variables. These latent variables were retrieved from a PLS model built on all compounds that were considered as to be already tested. In each step, the compounds selected in the previous steps were used for model development.
The advantage of this approach is, that the PLS latent variables are correlated with the target property, which makes them specific for an endpoint. Thereby both noise and irrelevant information gets filtered out.

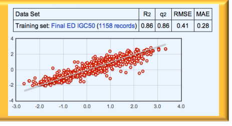
Workflow



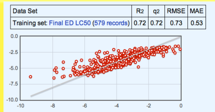
LogKOC
(Partition coefficient in the organic fraction of the soil)
- 668 compounds
- no restrictions
- average complexity



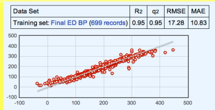
IGC50
(Inhibition growth concentration on T. Pyriformis)
- 1158 compounds
- no restrictions
- high complexity



LC50
(Lethal concentration on fathead minnow)
- 579 compounds
- no restrictions
- high complexity

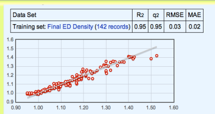


Boiling point
(Compound boiling point at 1 bar)
- 699 compounds
- muted restrictions
- low complexity



Density
(Mass per volume)

- 142 compounds
- restrictions
- low complexity

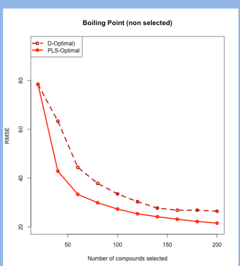


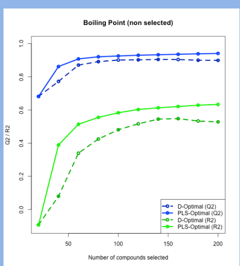
Dataset validation strategy

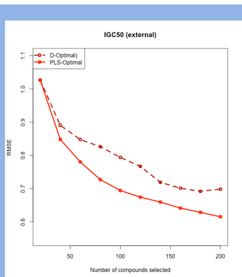
- 100 splits on each dataset
- 70% of compounds as operative set on which design is performed
- 30% of compounds as external validation set

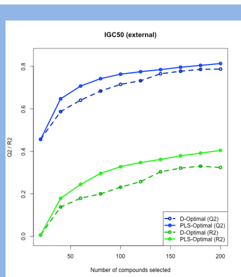
Results

A comparison of D-Optimal versus PLS-Optimal design was done with 40 to 200 compounds. Within the range from 40 to 160 selected compounds, the models based on the stepwise approach provide significantly lower RMSE ($p < 0.05$, direct method, binomial distribution, 100 trials) compared to those developed using molecules selected with D-optimal design.









On average, RMSE calculated using the PLS-Optimal approach was lower for about 10% compared to those developed the traditional method. In a similar way R^2 and Q^2 were significantly higher for models developed using PLS-Optimal design.
These results indicate that sets of molecules selected using proposed method have significantly higher quality compared to those selected with traditional D-Optimal design approach.

Validation facts

- A binomial test showed, that the improvement of performance is highly significant concerning RMSE, Q^2 and R^2.
- This improvement takes place for
  - all tested endpoints
  - both external and internal validations
  - each size of the datasets
  - the full range from 5% to 25% selected points
- Models of equally good performance can be created with only 50% of compounds

Conclusion Our results show, that the performance of D-optimal experimental design can significantly be improved by taking into consideration the correlation between descriptors and property. The PLS-optimal design uses latent variables, which incorporates also information about the target property and descriptors. The models developed using proposed PLS-optimal design provided significantly higher accuracy of prediction compared to the models developed using D-optimal for the whole range that can be particular interesting for practical application.

References

[1] Torbjörn Lundstedt et al. 1998. Experimental design and optimization. Chemometrics and Intelligent Laboratory Systems 42:3-40
[2] Massimo Baroni et al.. 1993. D-Optimal Designs in QSAR. Quantitative Structure-Activity Relationships 12:225–23

HELMHOLTZ | ASSOCIATION