HelmholtzZentrum münchen
German Research Center for Environmental Health

CADASTER

Performance, Reliability and Robustness - A comparison of several experimental design strategies

Stefan Brandmaier[1], Igor V. Tetko[1,2]

[1] Institute of Structural Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany and [2] eADMET GmbH, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

Motivation

- REACH legislation: Each chemical compound produced in or imported into the EU in an amount of more than one ton has to be registered according to a number of endpoints
- In case of hazardous, dangerous or toxic compounds, these endpoints contain toxicity and bio-accumulation
- Experimental determination of all these values is not possible, as experiments consume a lot of time, money (estimated to €9.5 billion)

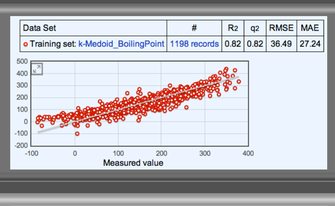
Required specifications

- **Low prediction error**
  - Minimize average prediction error
  - Minimize the error of the worst sample
- **Stability**
  - Low standard deviation in performance
  - Consistent development of performance
- **Flexibility**
  - Adaptability to small variations
- **Robustness**
  - Against small modifications in the dataset
  - Against structural outliers
- **Reliability**
  - Correlation between the number of selected compounds and the resulting performance

Datasets

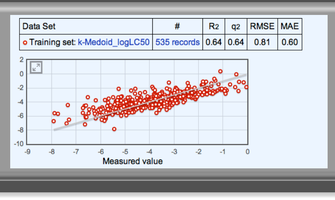
Boiling point

- 1198 compounds
- muted restrictions
- low complexity



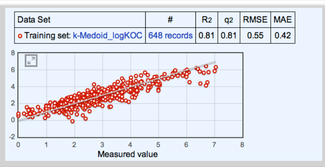
LC_{50}

- 535 compounds
- no restrictions
- high complexity



$logK_{OC}$

- 648 compounds
- no restrictions
- average complexity



Validation

Bagging
- 250 fold random selection

Characterizing
- Using multivariate techniques

Compound selection
- Established approaches

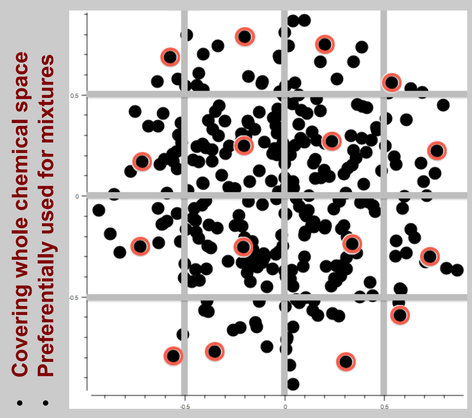
Model building
- Linear kernels

Evaluation
- On multiple criteria

Selection approaches

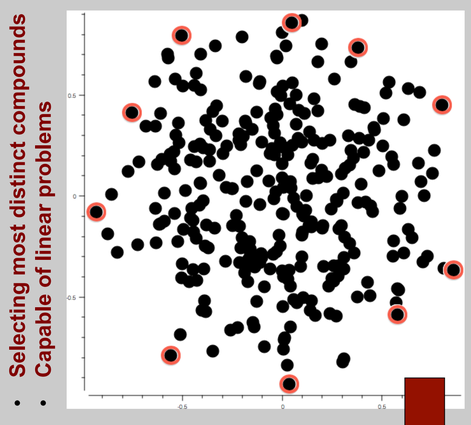
Space filling designs

- Usable only for few dimensions
- Chemical compounds are not equally distributed



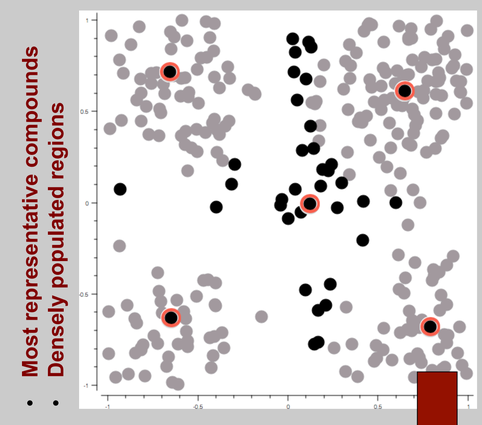
Dissimilarity selections

- Outlier detector in higher dimensional spaces
- Disregarding the center



Similarity search

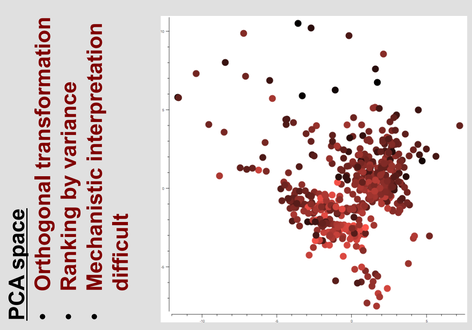
- Bias towards the central region
- Disregarding the periphery



Chemical space representation

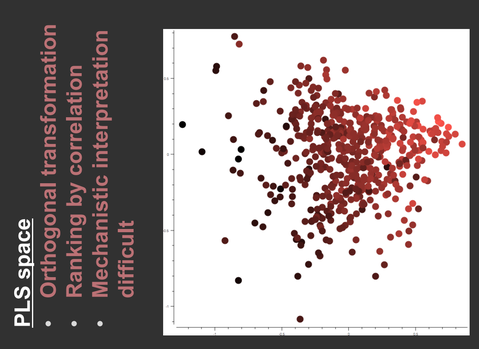
k-Medoid approach

- Non-adaptive
- Using principal components
- Based on space filling idea



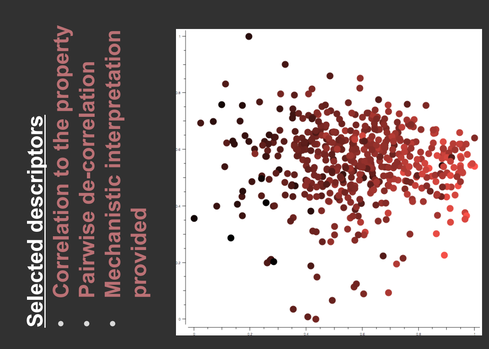
PLS-Optimal

- Stepwise execution
- Using PLS latent variables
- Based on dissimilarity

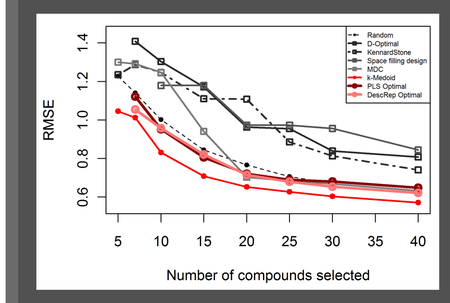


DescRep

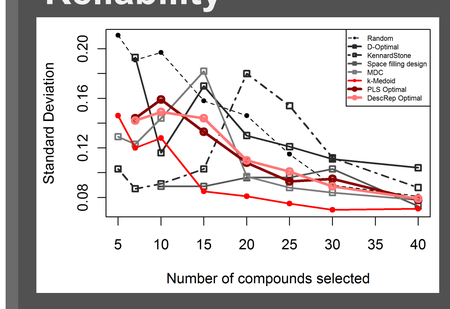
- Stepwise execution
- Using selected descriptors
- Based on similarity



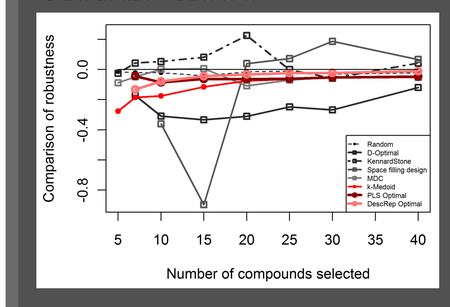
Performance



Reliability



Robustness



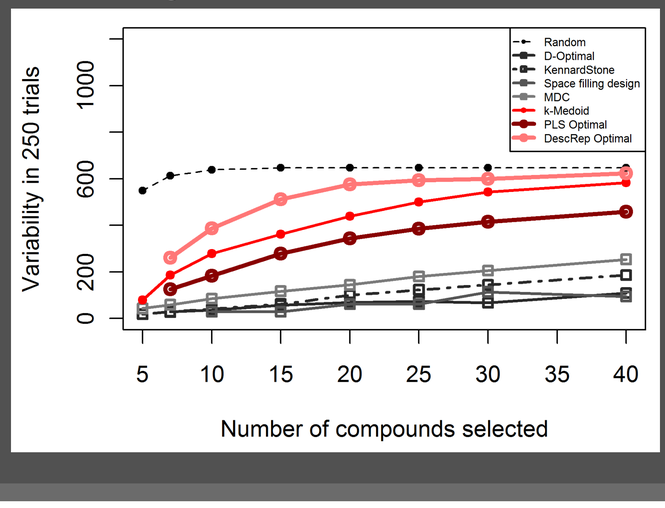
Results

Referring to a binomial test, models resulting from a selection based on
- adaptive approaches
- clustering approaches

perform significantly better concerning
- RMSE
- Q^2
- correlation coefficient

In terms of
- all tested endpoints
- both external and internal validation sets
- each examined size of the dataset (250-5000 compounds)
- the full range from 5% to 25% selected points
- regression and classification datasets

Flexibility



Conclusion

- k-Medoid provides the best performance for all examined datasets
- DescRep is robust against structural outliers
- Adaptive approaches help to stabilize the performance and to increases the reliability
- The major influence regarding the quality of resulting models is the informational basis
- Flexibility and adaptability are the key criteria for a stable and reliable performance
- Stepwise approaches and k-Medoid are the only ones to significantly improve the results of a random selection

HELMHOLTZ ASSOCIATION