

Lehrstuhl für Genomorientierte Bioinformatik
Der Technischen Universität München



Dissertation

*Experimental design methods to increase
the accuracy of in silico models*

Stefan Brandmaier

Supervisor: Prof. Hans-Werner Mewes
Advisor: Dr. Igor V. Tetko

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Experimental design methods to increase the accuracy of in silico models

Stefan Brandmaier

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzende: Univ.-Prof. Dr. I. Antes

Prüfer der Dissertation:

1. Univ.-Prof. Dr. H.-W. Mewes
2. apl. Prof. Dr. Dr. K.-W. Schramm

Die Dissertation wurde am 19.12.2013 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 30.04.2014 angenommen.

Abstract

“Several applications, such as risk assessment within REACH or drug discovery, require reliable methods for the design of experiments and efficient testing strategies. Keeping the number of experiments as low as possible is important both from a financial and an ethical point of view, as exhaustive testing of compounds requires significant financial resources and animal lives. With a large initial set of compounds, experimental design techniques can be used to select a representative subset for testing. Once measured, these compounds can be used to develop QSAR models to predict properties of the remaining compounds. This reduces the required resources and time.”[a]

Most of the commonly used experimental design methods are developed to select all new samples at once (i.e., static approaches). However, due to restricted capacities, the practical applications of these methods mostly perform the experimental testing of the suggested compounds in a sequential manner. Moreover, some measured samples are usually available before the experimental design and they should be incorporated by the selection procedure. Therefore, I developed several new sequential approaches (also called adaptive or stepwise) to apply established selection approaches, such as the D-Optimal criterion, the Kennard-Stone algorithm and similarity-based sampling to larger collections of chemical compounds. The stepwise approaches iteratively refine the representation of the chemical space after each measurement cycle. This is realized by the use of a property-oriented depiction of the chemical space, utilizing techniques, such as PLS latent variables, selected descriptors, predicted properties and the ensemble based applicability domain estimation. Furthermore, I investigated the usability of a static (not stepwise) experimental design approach based on the k-Medoid clustering.

A comparison of the proposed stepwise and classical static approaches was based on statistical performance of models derived from selections of samples using the respective approaches. I validated the performance on five regression datasets with different endpoints, representing toxicity, physicochemical properties and bioconcentration and on two classification datasets. To estimate the quality of the approaches, I evaluated them on criteria, such as the error performance, reliability, consistency, stability and the robustness against structurally diverse compounds.

I show that application of commonly used approaches in a stepwise procedure, which is taking the correlation to the target property into consideration, contributes to the quality of the experimental design. Compared to models derived from static approaches on principal components, models derived from the selection on property-oriented variables had a lower RMSE and a higher Q² and R². Our results indicate that a property-oriented representation of the chemical space enables a more flexible and purposive selection of compounds.

Furthermore, of all the models derived with static approaches, only those derived with the k-Medoid approach showed a significantly improved performance compared to a random selection for all analyzed datasets.

Zusammenfassung

In zahlreichen chemischen Anwendungsgebieten, beispielsweise der Risikobewertung im Rahmen der REACH Gesetzgebung oder bei der Entwicklung neuer Medikamente werden zuverlässige Methoden zur statistischen Versuchsplanung und effizienten Stichprobenanalyse benötigt. Die Anzahl erforderlicher experimenteller Versuche hierbei möglichst gering zu halten ist sowohl aus wirtschaftlicher, als auch als ethischer Sicht geboten, da zahlreiche Standardtests nicht nur hohe finanzielle Kosten verursachen können, sondern auch Tierversuche beinhalten. Ausgehend von einer größeren Sammlung relevanter Molekülen wählen die Methoden zur statistischen Versuchsplanung repräsentative Stichproben. Sobald diese Chemikalien experimentell untersucht wurden, kann die erhaltene Information benutzt werden um QSAR Modelle zu entwickeln, die relevanten Eigenschaften auch für die verbleibenden Chemikalien vorhersagen.

Die meisten Verfahren zur statistischen Versuchsplanung sind darauf ausgerichtet alle Stichproben in nur einem Auswahlschritt zu bestimmen. In Anbetracht eingeschränkter Laborkapazitäten werden die anschließenden Versuche jedoch in einer Vielzahl der Fälle der Reihe nach durchgeführt. Abgesehen davon liegen für die meisten Endpunkte bereits Messwerte aus veröffentlichten Studien vor und sollten bei der statistischen Versuchsplanung berücksichtigt werden. Um diesen Umständen Rechnung zu tragen habe ich mehrere sequenzielle Verfahren (auch adaptive oder schrittweise genannt) entwickelt um bewährte Selektionsverfahren, wie das D-optimale Kriterium oder den Kennard-Stone Algorithmus auf größere chemische Datensätze anzuwenden. Die schrittweisen Verfahren verfeinern hierbei die Darstellung des chemischen Suchraumes nach jedem experimentellen Arbeitsgang. Dies wird durch eine, auf die Zielvariable ausgerichtete Neuberechnung des chemischen Raumes bewerkstelligt. Zugrundeliegende Analyseverfahren sind hierbei Deskriptoren-Selektion, PLS Regression und die statistische Verfahren zur Bewertung des Vorhersagebereiches. Weiterhin wurde die Verwendbarkeit eines statischen (nicht schrittweisen) Verfahrens, basierend auf dem k-Medoid Clustering untersucht.

Um einen Vergleich der entwickelten schrittweisen Verfahren mit klassischen, statischen Verfahren zu ermöglichen, wurde die statistische Performanz von Modellen die aus die Stichproben, die mit den betreffenden Ansätzen gezogen wurden, ausgewertet. Fünf Regressionsdatensätze mit unterschiedlichen Endpunkten (unter anderem aquatische Toxizität, ein Adsorptionskoeffizienten und physikochemische Eigenschaften) und zwei Klassifizierungsdatensätze wurden zur Validierung benutzt. Um die Leistungsfähigkeit der verschiedenen Ansätze zu bewerten wurden mehrere Kriterien, wie der Vorhersagefehler der resultierenden Modelle, ihre Konsistenz, ihre Stabilität und die Widerstandsfähigkeit gegen Moleküle mit abweichender Grundstruktur untersucht.

Die Ergebnisse zeigen, dass eine schrittweise, adaptive und auf die Zielvariable ausgerichtete Anwendung gebräuchlicher Verfahren die statistische Versuchsplanung signifikant verbessern kann. Verglichen mit Vorhersagemodellen die durch statische Ansätze entwickelt wurden, weisen diejenigen, die mit einer schrittweisen Versuchsplanung entwickelt wurden geringere Fehlerwerte und höhere Korrelationswerte auf. Eine, auf die Zielvariable ausgerichtete Darstellung des

chemischen Raumes ermöglicht eine flexiblere und zugleich zweckmäßigere Auswahl ermöglichen.

Weiterhin zeigt diese Arbeit, dass von allen Modellen, die auf statischen Selektionsverfahren beruhen, nur diejenigen die Ergebnisse einer Zufallsauswahl verbessern konnten, die auf dem k-Medoid Ansatz beruhen.

Acknowledgement

Mein besonderer Dank für die Unterstützung beim Anfertigen dieser Doktorarbeit gilt folgenden Personen:

Meinem Doktorvater Professor Dr. Hans-Werner Mewes, dafür, dass er mir die Möglichkeit gegeben hat, diese Dissertation am Helmholtz-Zentrum München zu verfassen, für seine Bemühungen die Doktoranden zu ermutigen über den Tellerrand zu blicken und für seine Vorschläge und Anregungen, die Qualität dieser Arbeit zu verbessern.

Professor Dr. Dr. Karl-Werner Schramm für die Chance, die Methoden, die im Rahmen dieser Arbeit entwickelt wurden in der Praxis zu erproben, für die konstruktive Zusammenarbeit und für seine Bereitschaft als Zweitprüfer dieser Arbeit zu wirken.

Professor Dr. Iris Antes, dafür dass sie sich bereit erklärt hat, den Vorsitz meiner Prüfungskommission zu übernehmen.

Dr. Igor V. Tetko für die Betreuung meiner Doktorarbeit, die damit verbundenen wissenschaftlichen Diskussionen und Ratschläge, ebenso wie für seine Bereitschaft mir bei der Planung meiner wissenschaftlichen Arbeit freie Hand zu lassen, für seine andauernde Erreichbarkeit, die weit über das selbstverständliche hinausging und für sein Verständnis.

Meinem Institutsdirektor Professor Dr. Michael Sattler, den Verantwortlichen der FP7 Projekte CADASTER (grant agreement number 212668) und ECO (Marie Curie Initial Training Networks) (grant agreement number 238701) für die Finanzierung meiner Studien.

Professor Dr. Tomas Öberg, meinem Betreuer an der Linnaeus Universität in Kalmar für die Möglichkeit am Marie Curie Project 'ECO' teilzunehmen, seine fachliche Unterstützung, die wissenschaftliche Anregungen und die Zusammenarbeit.

Dr. Ullrika Sahlin für ihre Kollegialität und die wissenschaftliche Betreuung meiner Arbeit in Schweden.

Professor Dr. Gerrit Schüürmann und Ralf Uwe Ebert vom Helmholtz Zentrum für Umweltforschung in Leipzig für die respektvolle Zusammenarbeit und ihre Beiträge zur Studie über aquatische Toxizität.

Meinen Kollegen Dr. Eva Schlosser, Wolfram Teetz und Ahmed Abdelaziz für die Zusammenarbeit, die Unterstützung und die fruchtbaren wissenschaftlichen Diskussionen.

Dem Entwickler-Team von OCHEM, Dr. Yurii Sushko, Dr. Sergii Novotarsky, Robert Körner und Anil Kumar Pandey, deren Software die Berechnungen, die dieser

Arbeit zugrunde liegen, außerordentlich erleichtert hat, für die gute Zusammenarbeit.

Dr. Frank Westad, für die Vorschläge und Anregungen, meine Studien zu erweitern.

Meinen lieben Kollegen Evanthia Giagloglou, Dr. Ioana Oprisiu, Dr. Pantelis Sopasakis und Aleksandra Rybacka für die wissenschaftlichen Diskussionen.

Meinen Freunden, Aimee C. Sutherland für ihre Hilfe, meine Veröffentlichungen und meine Doktorarbeit sprachlich zu verbessern, Vijyant Srivastava für die Zusammenarbeit am Helmholtz Zentrum München und Sandra Tremml für ihre Beiträge zu einer prägnanten Beschreibung von Regressionsmethoden.

Dr. Faizan Sahigara, Alessandra Pirovano, Jacques Ehret und Elena Salmina für ihre Mitarbeit an Studien, die dieser Arbeit zugrunde liegen und für ihre Kollegialität.

Dandan Shen und Swapnil Chavan, zwei besonders lieben Freunden, für Ihre Unterstützung während meines Aufenthaltes in Kalmar.

Katharina Frank für ihre Anregungen zum Layout und zur graphischen Umsetzung dieser Arbeit, sowie für ihre Geduld und ihr Verständnis während meiner Doktorandenzeit.

Meiner Mutter Barbara Brandmaier und meinem viel zu früh verstorbenen Vater Josef Brandmaier, dafür, dass sie mich zu einem interessierten und neugierigen Menschen erzogen haben, für die Ermöglichung meiner Ausbildung, ihre jahrzehntelange, immerwährende Unterstützung auf meinem Lebensweg, für die Freiheit, die sie mir immer zugestanden haben und ihren Respekt vor jeder meiner Entscheidungen. Ihr Beitrag zu dieser Arbeit war, obwohl nicht mittelbar, so doch fundamental.

Table of Contents

1	Introduction	13
1.1	Motivation	13
1.2	Aims.....	15
1.3	Thesis roadmap.....	15
1.4	State of the art.....	17
1.4.1	QSAR modeling.....	17
1.4.2	Experimental design approaches.....	37
2	Materials.....	47
2.1	Regression datasets.....	49
2.1.1	Boiling point.....	49
2.1.2	$\log K_{OC}$	52
2.1.3	$\log BCF$	54
2.1.4	$\log LC_{50}$	56
2.1.5	$-\log IGC_{50}$	58
2.2	Classification datasets.....	61
2.2.1	AMES mutagenicity.....	61
2.2.2	Cytochrome P450 inhibition	63
2.3	Structural outliers.....	65
3	Methods.....	67
3.1	QSAR modeling with respect to the REACH legislation	68
3.1.1	Constraints in QSAR modeling	68
3.1.2	Modeling aquatic toxicity.....	74
3.2	Customized implementation of experimental design approaches	88
3.2.1	D-Optimal design	88
3.2.2	Kennard-Stone algorithm	88
3.2.3	MDC selection	88
3.2.4	Space filling design.....	88
3.3	Stepwise selection approaches	90
3.3.1	Ordering the chemical space.....	91
3.3.2	Applicability domain approaches.....	95
3.4	Cluster-based selection approach.....	101
3.4.1	Implementation.....	101
3.4.2	Initial cluster centers	102
3.5	Validation pipeline.....	104
3.5.1	Procedure	104
3.5.2	Criteria.....	106
4	Results and Discussion.....	109
4.1	PLS-Optimal: A proof of concept.....	110
4.1.1	Comparison of the performance.....	110
4.1.2	Interpretation.....	116
4.2	DescRep: A generalization of the adaptive concept	119
4.2.1	Comparison of the results.....	119
4.2.2	Interpretation.....	125
4.3	k-Medoid: An improved static approach.....	127
4.3.1	Comparison of the performance.....	127
4.3.2	Interpretation.....	132
4.4	Ensemble based approaches: Using the applicability domain	136
4.4.1	Comparison of the performance.....	136
4.4.2	Interpretation.....	140

4.5	Exhaustive comparison	146
4.5.1	<i>Method based view</i>	146
4.5.2	<i>Property based view</i>	148
4.5.3	<i>General remarks</i>	150
4.6	Practical application	152
4.6.1	<i>Evaluation of previous results</i>	152
4.6.2	<i>Descriptor representation</i>	154
4.6.3	<i>Data set cleaning</i>	154
4.6.4	<i>Experimental design</i>	156
4.6.5	<i>Summary</i>	160
5	Conclusion and outlook	163
5.1	Summary of presented work	163
5.2	Significant findings	163
5.3	Interpretation of the observations	164
5.4	Outlook	165
5.5	Final remarks	165
	References	167
	List of cited publications and studies	181
	Software used	183
	List of publications	185
	Eidesstattliche Erklärung	187
	Curriculum Vitae	189

1 Introduction

1.1 Motivation

“The REACH legislation¹ includes the requirement that every chemical compound produced in or imported to the European Union in an amount of more than one ton has to be registered regarding a number of endpoints. Experimental determination of these properties for all compounds would require high-throughput testing. According to Rovida and Hartung, the financial requirements for such testing are about €9.5 billion.² For potentially hazardous, dangerous, or hardly degradable substances, registration also requires information about their bioaccumulation and toxicity. Apart from cost and time efficiency – a sample for e.g. bioconcentration requires around two months and can cost more than €200 - this also leads to ethical problems, as experimental determination of endpoints associated with toxicity and bioaccumulation is achieved by animal tests.

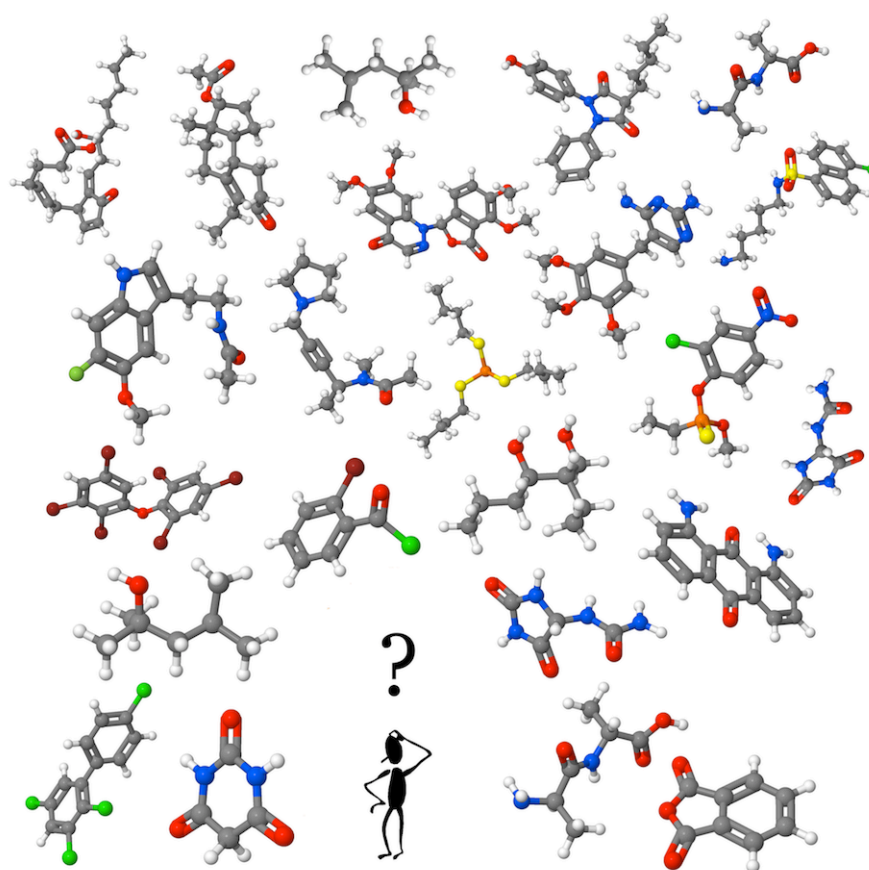


Figure 1. The problem of selecting a representative subset.

The necessity to keep the overhead of (animal) testing as low as possible is also important in many other research areas, for example the chemical or pharmaceutical industries. One common strategy to address this problem is to use structure-activity modeling³ and to predict the required properties rather than performing experimental measurements. This strategy entails testing only a small subset of all the compounds of interest and constructing a predictive model using the experimentally determined values. This basic task can be reduced to the

problem of drawing a representative subsample of a larger set. This method is important in other fields of research - e.g., Quantitative Structure Activity Relationship (QSAR) development,⁴ large-scale database scanning,⁵ in-silico drug design,⁶ and compound prioritization⁷ - as well as in experimental design for risk assessment within REACH.^{8,9}[a]

Experimental design in computational chemistry consists of the sampling of representative compounds and aims to deliver a sensitive subset of a predefined chemical space and for a certain, predefined endpoint. The insight gained by the knowledge of this subset is intended to be the basis of a prediction model or QSAR, which can be applied to new (not yet tested) compounds.

Such techniques are crucial in terms of time and cost efficiency as their goal is to enable the calculation of highly predictive models at a low experimental extent.¹⁰ The challenge in this context, which is indicated in Fig. 1, is to select a combination of compounds for experimental testing, which later on facilitates the calculation of a highly reliable model to predict a given property also for compounds, which are lacking of experimental measurements.

There are numerous approaches^{11,12} for the selection of a representative subset of compounds that are supposed to deliver the most reliable model. These approaches aim to select the subsets by various criteria. They all have in common that their selection is based on the variance of structural characteristics. The selection is done in a one-step procedure, and does not take into account correlations between the structural features and the target property. Unfortunately, an extended variation within a certain structural feature does not necessarily condition that the structural feature is significant for the considered property. In the cases of numerous chemical properties, the contribution of certain structural features is minor. Therefore, the compounds selected with respect to these features may not be optimal for the modeling of a given property.

Furthermore, the exclusive description of the chemical space by structural characteristics with high variance, as it is common for the classical approaches, produces a remarkable fact: For a given dataset, given descriptors and a given number of compounds to be selected, the molecules chosen for experimental testing is identical for all endpoints, irrespective if this endpoint is physicochemical, chemical, biological or toxicological.

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. Given, for example, a set of 600 compounds of interest and the limitations of being able to test only 100 of these compounds, almost no laboratory will test all these 100 compounds in parallel, due to restricted capacities and will rather do this in a stepwise procedure.

Such a procedure implicates that information about the measured property is gathered from testing cycle to testing cycle. In the standard approaches, that select all compounds in one step, this growing amount of information is not taken into

consideration, although exactly this information could be used to refine and thereby significantly increase the quality of the experimental design approach.

The question is whether there are strategies that could provide a better selection of compounds by taking into consideration the correlation to the target property and available data.

1.2 Aims

This thesis aims to provide novel approaches for experimental design, which significantly decrease the number of required experiments. The main focus of attention is hereby on adaptive procedures. Adaptive procedures are executed in a stepwise manner and after each step the representation and description of the chemical space and the compounds populating it is refined. For this rearrangement the accumulating knowledge about the target property, which is growing with each (hypothetical) measurement cycle, is taken into consideration.

In this study three different basic ideas of how to iteratively refine the depiction of the chemical space are suggested and implemented. The applied techniques to correlate the alignment of the chemical space with the target property are partial least squares, supervised descriptor selection and ensemble based applicability domain estimation.

These techniques are combined with generally accepted methods to select representative subsamples. The newly developed approaches are compared to commonly used experimental design approaches in a validation using seven datasets of different characteristics, all of them relevant for recent QSAR research. Additionally, as stepwise procedures are not always feasible, a novel non-adaptive approach using the k-Medoid clustering is introduced and evaluated similarly to the adaptive approaches.

The developed approaches are analyzed regarding the resulting error in prediction, stability, reliability and robustness against structurally diverse compounds and examined with respect to a mechanistic interpretation. Investigations on limitations and applicability restrictions are enclosed.

It is shown that adaptive approaches, which take the correlation to a certain target property into consideration, can significantly contribute to the quality of the QSAR models resulting from the selection of an experimental design approach. Furthermore, it is indicated that adaptability and variability is the underlying principle, which enables this improvement in performance.

1.3 Thesis roadmap

The subsequent part of this introduction provides a short overview of the basic idea of QSAR modeling and the common techniques and methods used in this field of research. The basic ideas of experimental design and commonly used experimental design approaches are also explained. The contents within this

section only recapitulate the state of the art and prerequisites for the understanding of this thesis.

The next chapter (Materials) extensively introduces the datasets used in this study. The detailed discussion of those datasets aims to illustrate the workforce diversity within current QSAR research. Concerning this matter and to provide an example of the practical application of QSAR, as well as its limitations, the subsequent section (Methods) starts with the exemplification of the development of a statistical model predicting aquatic toxicity. The rest of the section focuses on the development of new experimental design approaches.

The second part of the method section firstly elaborates the underlying theory, which is basic for the developed adaptive concepts. A detailed description of the implementation, as well as the introduction of a non-adaptive, cluster-based approach (to handle problems that make a stepwise testing procedure impossible) and the description of the validation pipeline will be presented.

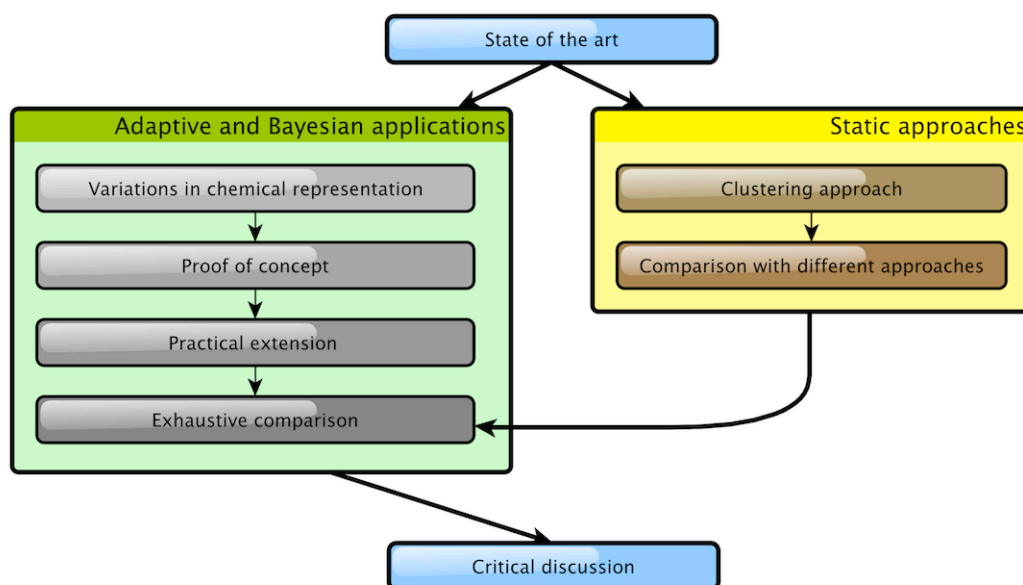


Figure 2. The thesis roadmap.

The following chapters (Results and Discussion, Conclusion) investigate the ability of these new approaches to improve the reliability and performance within experimental design. Fig. 2 visualizes the proceedings starting at this point.

The results and discussion arranges the scientific findings and observations retained from the evaluation on the adaptive approaches in a gradual structure. Starting with a proof of concept, the conceptual principle is successively extended.

In parallel, the performance of the cluster-based approach is examined. The final and exhaustive comparison combines all examined selection approaches with the developed adaptive concepts to represent the property space. The section is closed with a critical disputation of experimental design in general and the adaptive concept, referring to the observations, findings and insights gathered in this study.

1.4 State of the art

1.4.1 QSAR modeling

The abbreviation QSAR stands for 'quantitative structure activity relationship' and describes a field of research that aims to systematically map structural features of chemical compounds to certain properties. These properties can be biological^{13,14,15} (e.g. enzyme inhibition, biodegradation), chemical^{16,17,18} (e.g. lipophilicity, solubility), or physical^{19,20,21} (e.g. melting point, boiling point). The underlying assumption is the frequently cited presumption that similar structures condition similar qualities and QSAR attempts to find which specific structural features are correlated to a certain property.

The first accordant observations supporting this assumption were reported in the end of the 19th century, focusing mostly on the correlation of the molecular size or weight to properties, such as the boiling point of alkenes²² or the narcotic effect of primary alcohols.²³ These days QSAR research has reached a high level of complexity, incorporating knowledge of different fields of science, such as quantum chemistry, toxicology, proteomics, informatics and machine learning, and data mining.

Modern QSAR^{24,25,26} combines the mathematical description of chemical compounds with advanced machine learning techniques and statistical evaluation procedures. The motivation behind can be reduced to two basic requirements:

- **Knowledge amplification**
QSAR can provide a deeper insight into the mechanisms of chemistry, physics and biology, as it describes correlations that have previously not been observed or recognized.
- **Predictive ability**
QSAR can be used to predict properties for new compounds which otherwise have to be experimentally measured. Depending on the property, the measurement procedure can be time consuming or financially expensive.

These two aims are generating an area of conflict, as the models with the highest predictive power are usually calculated with complex approaches. The resulting models are complex as well, which prevents from an easily accessible mechanistic interpretation of the retained correlations. On the other hand, simple approaches can provide a good insight in the underlying mode of action, but their predictive quality is often less than optimal. Finding an appropriate balance between these antipodal targets is one of the most important aims in modern QSAR.

1.4.1.1 Descriptors

A basic requirement in QSPR and QSAR modeling is the description of chemical compounds and their structure in a way which can be processed by statistical methods. This description should be informative and encode chemical features of the molecules, which are relevant for the analyzed property. The common

approach to reach this is the calculation of the so called molecular descriptors to numerically represent molecules.^{27,28,29} Generally spoken, each numerical value derived from a chemical compound can be used as a molecular descriptor. An example thereof is the number of atoms in a molecule. The number of possible descriptors is – as a matter of principle – not limited, but they can be categorized into four main classes:

- **1D descriptors**

1D descriptors are the simplest descriptors in QSAR modeling. They consist of those descriptors, which do not take the connectivity of the atoms in a molecule into account, but just their number, presence or absence. 1D descriptors can be derived from the chemical formula. Examples of this are the molecular weight and the number of occurrences of a certain atom or atom type (e.g., halogens, metals) within the compound. The yellow circles in Fig. 3a) indicate all substructures, which contribute to a descriptor counting the occurrence of nitrogen atoms.

- **2D descriptors**

2D descriptors are those that comprise a molecule as a two dimensional graph, with the atoms as nodes and the bonds as edges (or in rare cases the other way round). They are the most frequently used descriptors in QSAR modeling. Typical examples are fragment-, group- or substructure counts and 2D autocorrelations.^{30,31} Fig. 3b) shows such a 2D descriptor, analogously to Fig. 3a). The decisive element hereby is an O-CH₃ group.

- **3D descriptors**

3D descriptors are those derived from a three dimensional depiction of the molecule. Commonly before the descriptor calculation, a structural optimization of the underlying compound is performed. The intention of this optimization is to find a minimum energy conformation, as this is more likely to coincide with the natural structure. The optimization can be accomplished using a simple force field approach, semi-empirical tools such as MOPAC,³² or *ab initio* calculations. Examples for such descriptors are the surface area, geometric and topological characteristics regarding the molecule's latitude or topological distances between certain atoms and/or chemical groups, as indicated in Fig. 3c).

- **4D descriptors**

4D Descriptors take the flexibility of chemical compounds into consideration. They are calculated on an ensemble of supposable conformations of a compound and their use in QSAR modeling is limited.

Apart from these main classes, there are two more classes of descriptors, which can be interpreted as independent groups, although they are frequently referred to as subgroups of the main groups.

- **Quantum chemistry descriptors**

Quantum chemistry descriptors can be seen a subset of the 3D descriptors, as they are derived from a three dimensional representations of the

molecule. Typical examples are the highest occupied molecular orbital (HOMO) energy and the lowest unoccupied molecular orbital (LUMO) energy,³³ as well as the dipole momentum or the total energy of a structure.

- **String descriptors**

String descriptors can be seen as a special subset of 2D descriptors. They are derived from any string representation of a compound. This representation can be a standardized nomenclature, such as the IUPAC name of a substance, or a notation like the chemical SMILES.³⁴ The standardized names contain information about the presence of certain chemical groups and their connectivity. Examples for this kind of descriptors are substring counts within the string representation.

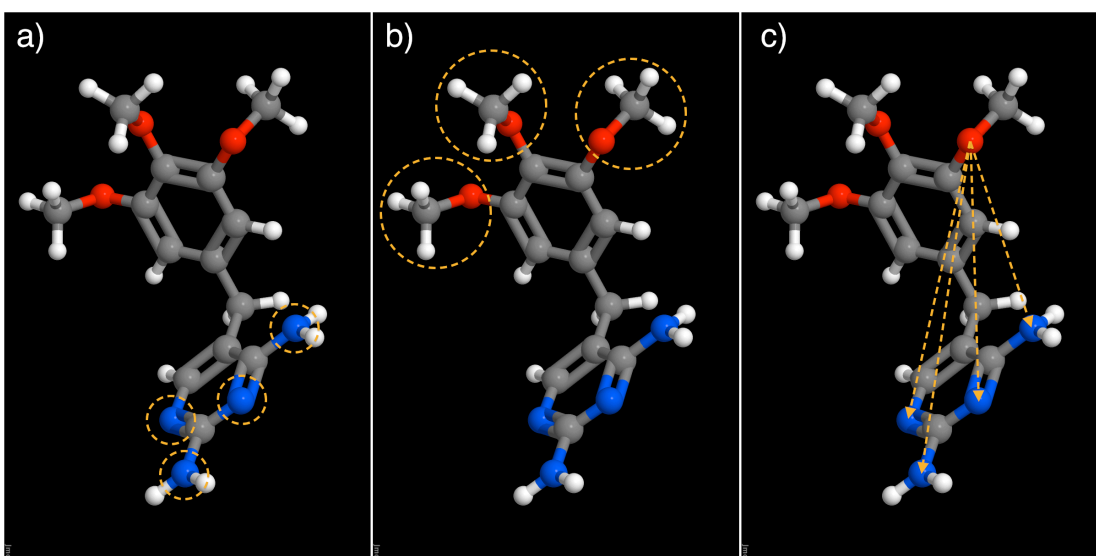


Figure 3. A depiction of different types of commonly used descriptors in chemoinformatics. The yellow circles and lines indicate the decisive structural element. This can be a certain atom (a), a chemical group (b) or a topological distance (c).

Each descriptor has a specific numeric range. If different descriptors types are combined, their respective scales can vary by orders of magnitude. With this said, it is recommended to scale them to a unique order of magnitude. The two most common treatments are thereby the normalization and the standardization. Normalization rescales the values to a predefined range, which usually is [0,1]. Standardization on the other hand rescales the values to a predefined arithmetical mean (usually 0) with a predefined standard deviation (usually 1).

1.4.1.2 Multivariate data analysis

As mentioned in the previous paragraph, there is almost no limitation to the number of hypothetical descriptors. The number of fragment-based descriptors derived on a dataset of several hundred compounds, for instance, can easily exceed a number of 2000 descriptors. Although a high number of descriptors contribute to a unique and precise representation of each chemical compound, there are also disadvantages linked to such an exhaustive representation of the chemical substances. Complications that usually arise are, first of all, that a certain

percentage of the descriptors are not statistically relevant. The information content of a fragment that is present in only one or few compounds within the dataset is limited. Furthermore, the computational performance has to be taken into consideration. Every descriptor can be interpreted as a dimension in the chemical space and the computational efforts required for the calculation of a QSAR model increase with the dimensionality of the descriptor space.

Last but not least, the higher the number of used descriptors is, the higher is the probability of inter-correlations and redundancies within the descriptors. An example thereof is that a descriptor indicating the presence of a minimum of nine carbon atoms in a compound is highly correlated with a descriptor indicating the presence of eight carbon atoms, just as the presence of nine or more carbons implies also the presence of eight carbons. Such redundancies are surely unwanted, as they can overweight certain latent properties. Latent properties are characteristics of a compound that are not explicitly defined, but which are implicitly specified by several descriptors. The size of a molecule is such a latent property and qualities, like the molecular weight, the number of atoms, the number of bonds, etc. are descriptors that cipher this property. Latent properties are of crucial importance, as they describe the basic characteristics of a chemical compound, which are fundamental for predictive modeling of certain endpoints.

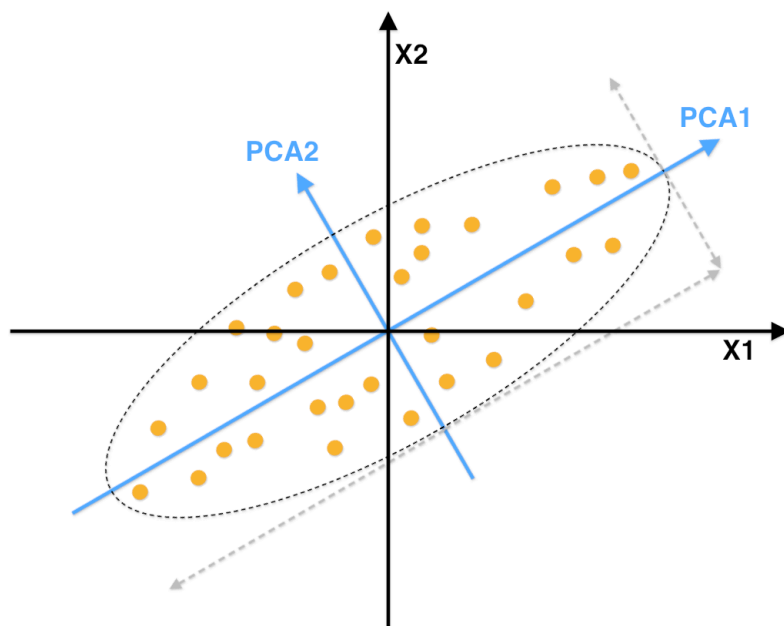


Figure 4. A schematic depiction of the orthogonal transformation resulting from a principal component analysis. The black arrows represent the original coordinate system, the yellow dots represent instances in the space spanned by these axes and the blue axes indicate the newly calculated coordinate system.

The principal component analysis (PCA) allows addressing these problems.³⁵ PCA is an orthogonal transformation of the descriptor space with a rearrangement of the axes according to their variance in the descriptor space. A schematic visualization of the way PCA works can be seen in Fig. 4.

The effects of this transformation is that the new set of orthogonal axes with the highest variance and thereby putatively with the most information content are selected for representation of the chemical space. Furthermore, it helps to drastically reduce the complexity of the chemical space. In practice, the first five to ten axes of the PCA often cover more than 80% percent of the variance in the original data. Thereby, descriptors without statistical significance get eliminated. Apart from that, the axes of the PCA are pairwise orthogonal, which means that they are not inter-correlated.

There are several variants of the principal component analysis, but they all work on the same principle: In the first step, given N instances and M available descriptors, the descriptor matrix D is built. It contains a row for each compound and a column for each descriptor. Then the data columns are centered, so that the mean of the data is zero. In the next step the covariance C matrix is calculated.

$$C = \frac{1}{N - 1} D^T D$$

The covariance matrix is a quadratic matrix and each field contains the pairwise covariance of two descriptors. The covariance matrix is used to calculate its Eigenvectors $x_{1,2,\dots,m}$ and the according Eigenvalues $\lambda_{1,2,\dots,m}$.

$$C * x_i = \lambda_i * x_i$$

The Eigenvectors are orthogonal and they represent the new coordinate system for the data. The ranking of the Eigenvalues indicates the ranking of the principal components, or expressed in other words, the first principle component is the Eigenvector with the highest Eigenvalue, the second principle component is the Eigenvector with the second highest Eigenvalue and so on. After the selection of an appropriate number of principal components to significantly describe the original data, the data points are transformed referring to the new principal components. This is the final step of the PCA and the retained data is reduced in dimensionality and free of redundancies.

Two expressions frequently used in conjunction with the PCA, as well as with any other orthogonal transformation are:

- **Scores**

PCA scores are assigned to compounds. They describe which characteristic a compound has regarding the new coordinate system and thereby also regarding the newly derived principal properties. Or expressed in other words, the score of a compound regarding the first principal component is its x-coordinate referring to the orthogonally transformed descriptor space.

- **Loadings**

PCA loadings are assigned to the original descriptors. They describe the contribution of a certain descriptor to each of the resulting principal components.

Based on the appropriate preprocessing of the descriptors (normalization, standardization, centering, etc.) PCA (as any other orthogonal transformation) can also handle variables that are not normally distributed. In fact, most fragment descriptors (which will be extensively used in the following) show a distribution, which is similar to a chi-squared distribution. Furthermore, depending on the implementation of the method, PCA can also deal with discrete descriptors. Such variables can be deconstructed into several variables, one of them for each discrete value, displaying the absence or presence of the certain value.

1.4.1.3 Machine learning algorithms

The statistical techniques used to find and extract the correlations and dependencies between a property and the descriptor space are machine learning methods, developed in the field of data mining and artificial intelligence. Machine learning methods can be divided into two groups: classification methods; which make a decision if an instance has to be assigned to a certain class or not and regression methods which assign a specific value to an instance.

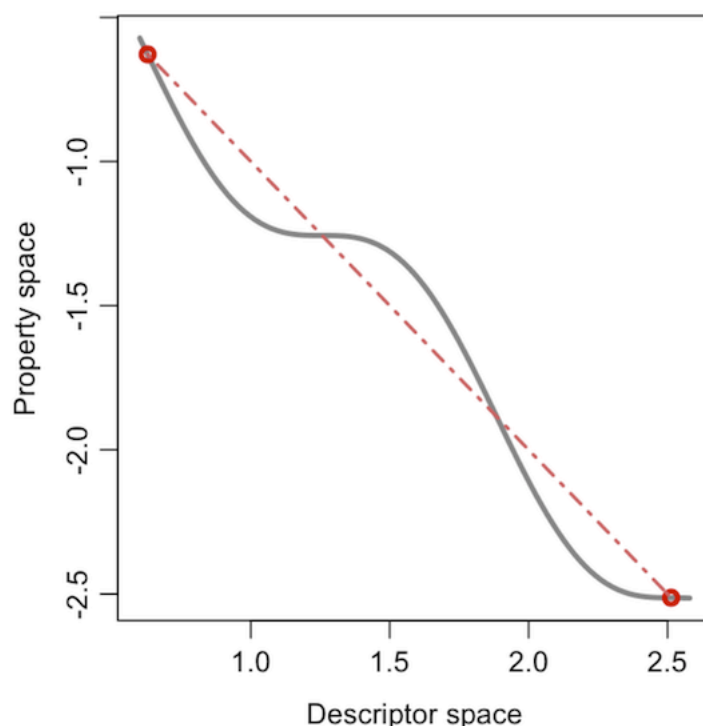


Figure 5. The principle within the regression idea. The solid grey curve shows the correlation between the descriptor space and the property space, the red dots indicate available measurements that are used for the approximation of a linear function, which is indicated by the dashed red line.

Literature provides a variety of implementations and variants for each of the approaches based on Bayesian theory,^{36,37} graph theory,^{38,39} rule-based interpretation^{40,41} and brain science.^{42,43} The following chapter will only focus on those that will be used in this thesis. Further, this chapter will just provide a basic overview of the principles and most common implementations of each approach.

1.4.1.3.1 Regression methods

Contrary to the classification methods that work with a discrete separation of the descriptor space, the regression methods detect continuous correlations between the descriptor space and a target property. A typical regression problem in QSAR modeling is the prediction of the value of a certain property, e.g. the boiling point of a compound.

1.4.1.3.1.1 Linear regression

The most commonly used regression technique in chemoinformatics and computational chemistry is the multiple linear regression (MLR). The underlying principle of this approach is the description of a target property y as a linear combination of input variables. Given a set of M descriptors, x_1, x_2, \dots, x_m , the aim is to express the target property y as an equation in the form of:

$$y = c + \sum_{i=1}^M w_i * x_i$$

where w_i is the specific weight of a descriptor and c is a constant factor, called the bias. The optimization process hereby aims to find the weights and bias that minimize the sum of squared errors over all instances and most commonly uses the ordinary least squares (OLS) estimator:

$$\begin{pmatrix} c \\ w_1 \\ w_2 \\ \dots \\ w_m \end{pmatrix} = (X^T X)^{-1} * X^T Y$$

Hereby the data matrix X and the property vector Y for N observations are defined as:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}; Y = \begin{pmatrix} y_1 \\ y_n \\ \dots \\ y_n \end{pmatrix}$$

A special case of the linear regression is the simple regression, which describes the target property as a function of only one descriptor. The underlying dependencies between descriptor and property space are often of a complex order, but can be approximated with a linear function. Fig. 5 describes such an approximation process.

Often a linear regression is built only on a small set of preselected descriptors, but there are also implementations that aim to filter a large number of input variables in an optimal way.

1.4.1.3.1.2 PLS regression

Several studies show that multiple linear regression is to some extent prone to misinterpretations for a high number of input variables, especially if those are inter-correlated.^{35,44} Furthermore, the computational time required to calculate the weights, drastically increases because of the need to calculate $(X^T X)^{-1}$. A step ahead to automatically solve this problem was the development of PCA regression. The PCA regression combines the principal component analysis with linear regression.

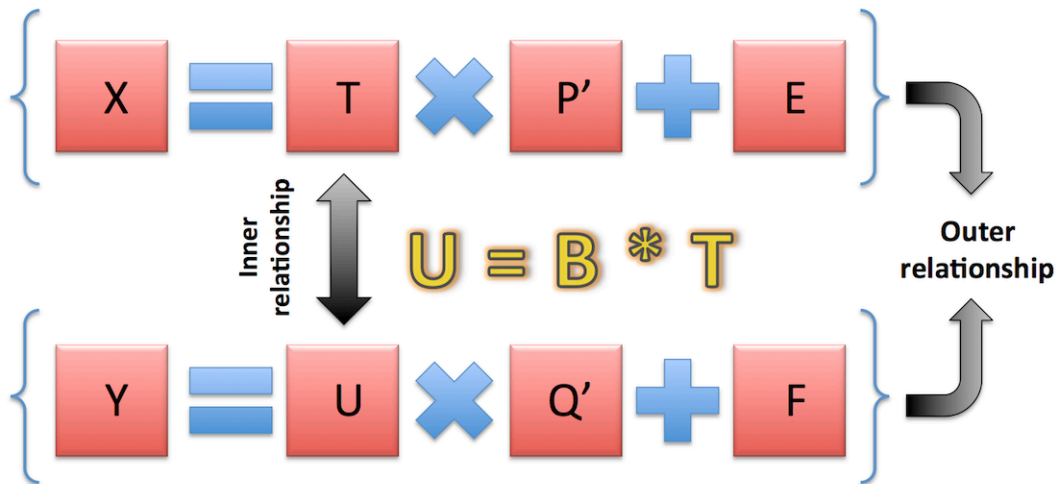


Figure 6. The PLS analysis works with decomposition of both the target properties and the descriptors. The derived score matrices T and U are related to retain the orthogonal transformation.

The advantages arising from the reduction in dimensionality and the elimination of redundant information are obvious. Nevertheless, the decrease of information content involved with the decrease in dimensions is not only beneficial. PCA ranks the components (sometimes also called principal properties) by their variance in descriptor space, but a high variance within such a component does not necessarily imply a high relevance for a certain target property. Quite the contrary, the elimination of dimensions can lead to a loss of information that is crucial for the target property. The partial least squares regression can solve this problem.

Partial least squares (PLS)⁴⁵ contains an orthogonal transformation of the descriptor space, but contrary to the PCA transformation, the components are not just ranked by their variance in descriptor space, but by their correlation to the target property. Additionally, PLS can also be used to predict several properties simultaneously.

Technically, PLS separately calculates the principal components both for the descriptor space and the space spanned by the properties. This decomposition is called the outer relationship. Fig. 6 shows a breakdown of the descriptor space X into the score matrix T , the loading matrix P and the error term E .

$$X = T * P + E$$

Analogously the property space Y is broken down into into the score matrix Z , the loading matrix Q and the error term F .

$$Y = U * Q + F$$

In the following step, the scoring matrices are related to each other by the matrix B , which becomes the new transformation matrix.

$$U = B * T$$

This is called the inner relationship and can be seen as a regression model between the two score matrices.

PLS regression is the most frequently used method within this thesis. Reasons therefore are amongst others the efficient calculation, the conservation of linear dependencies, the mechanistic interpretability and the high acceptance the method is given.

1.4.1.3.2 Classification methods

A typical classification problem in QSAR modeling is, for instance, if a compound is an inhibitor to a certain protein or if it is not. Such binary classification problems, which can be seen as Yes-or-No decisions, are the most common problems in QSAR modeling. Although most classification methods can handle also multi-class problems, these problems will not be considered in the following. The reason for this is that within this thesis, such problems were not analyzed. Furthermore, each multi-class problem can be expressed as an aggregation of binary classification problems.

1.4.1.3.2.1 *k*-Nearest-Neighbors

The simplest and most intuitive approach for classification is the *k*-Nearest-Neighbor (*k*NN) approach. It examines the neighborhood around each instance and assigns new instances referring to these compounds. The neighborhood is usually defined as a distance function in descriptor space. Commonly the Euclidean distance is used, but also other metrics, e.g. the Manhattan distance can be used.

In the original implementation, the decision, which class an instance should be assigned to, is a majority decision. Derived from the given parameter *k*, which defines the number of compounds that should be taken into consideration, each instance is categorized to majority of the neighborhood. The influence of the parameter *k* on this decision can be seen in Fig. 7. Each dot represents a chemical compound in the chemical space. The red and blue coloring symbolizes the affiliation to a certain classification category. The black dot should be classified regarding the nearest neighbor criterion. If five nearest neighbors are taken into consideration, the instance is assigned to the red class, but taking 15 nearest neighbors into consideration, the instance is assigned to the blue class.

Naturally, the majority decision works well only for balanced sets.^{46,47} A set is called balanced, if the number of instances assigned to the first class is approximately similar to the number of compounds in the second set.

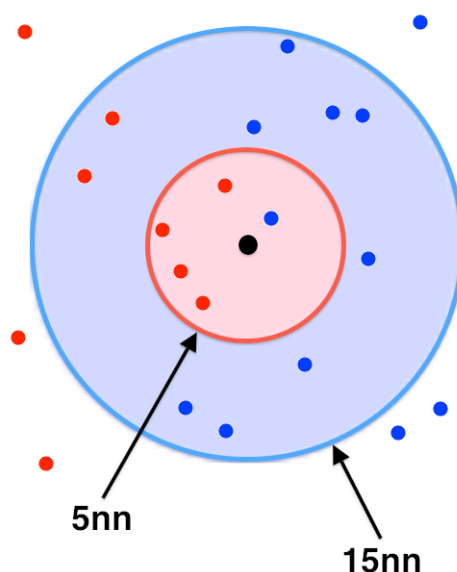


Figure 7. The k-Nearest-Neighbor approach. The new compound is assigned referring to the majority class within a certain range. The assignment value can change for a different number of neighbors taken into consideration.

There are numerous suggestions how to extend and improve the kNN approach, e.g. the weighting of instances according to the overall partition of their classes, the weighting of the k nearest instance referring to their distances,^{48,49} or the incorporation of fuzzy logic.⁵⁰ The use of kNN in computational chemistry is rather limited. The reason for this is two-fold: firstly, there are more advanced approaches, and secondly, the prediction of new instances is relatively time expensive, when compared to other approaches.

1.4.1.3.2.2 Decision trees

A decision tree is the representation of a classification procedure as a connected graph without cycles.⁵¹ Most decision trees are built hierarchically in a top-down-manner by splitting the underlying data referring to the most informative attributes.⁵² The attributes are represented by the graph's nodes, whereas the edges represent particular values of the sourcing node. In classical decision trees, these attributes are binary (or discrete). But if the decision tree is extended by a directed discretization, it can process also continuous or numeric attributes. Fig. 8 shows such a tree, which is based on a young man's decision if he should wash his car at a certain day.

The commonly used criteria to evaluate the most informative attribute are the 'GINI impurity' (or 'GINI index') or the 'Information gain'. The 'GINI index' works on statistical dispersion, whereas the 'Information gain' focuses on the level of entropy. Both criteria produce decision trees with similar performances. If no stop

criterion is given, the branching within a decision trees is carried forward until all cases in a node fall into the same classification category. Furthermore, it is worth mentioning that two branches can use the same attribute on different levels. Every decision tree can be translated into a set of rules, and the other way round.



Figure 8. A decision tree for a binary decision problem. Although the decision problem is binary, the tree does not have to be binary too.

1.4.1.3.2.3 Support vector machines

Support vector machines (SVM) work with a linear separation of the descriptor space. Thereby, the so-called hyper-plane is used to define a border that separates the two classes.⁵³ The adjustment of the hyper-plane is dependent only on the nearest training instances, on the support vectors and it is determined to maximize the margin. The margin is defined as the orthogonal Euclidean distance of the support vectors to the hyper-plane. Fig. 9 depicts the mode of operation of an SVM.

For many classification problems, a clear linear separation is not possible. This can result from measurement errors or dependencies, which are not covered by the used descriptors. To enable the accomplishment of such problems, the concept of a 'soft margin' allows misclassifications. Thereby the cost-factor is used to control the tradeoff between the acceptance of misclassification and the rigidity of the margin.⁵⁴ The higher the cost-factor is set, the higher the investment of computational expenses will be to find an optimal separation. The tradeoff value for misclassifications is usually selected according to cross-validation procedures.

Apart from isolated instances that cannot be fit into a linear separation, there are data distributions that make such a disjunction counterproductive, for example if the underlying classification problem is non-linear. The solution for such problems

is the transformation to space of higher dimensionality. A simple example of such a higher dimensional space is the extension of the original space by the cross product of the original descriptors. Fig. 10 visualizes this transformation for a dataset that is not separable in the original space.

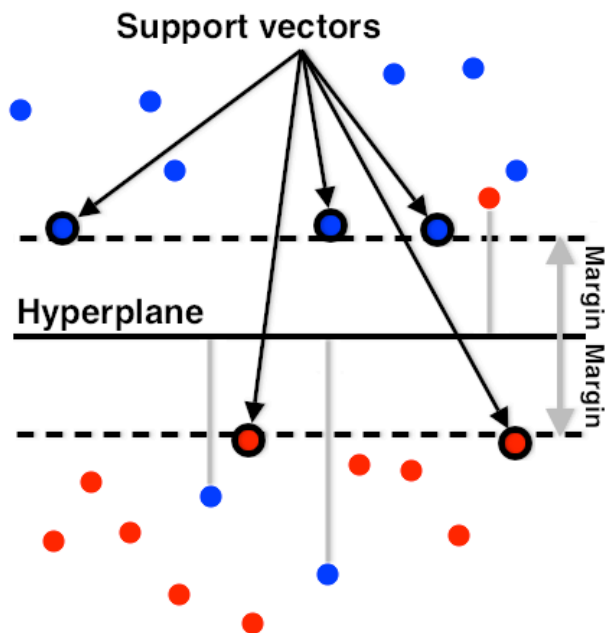


Figure 9. The separation of the chemical space with a SVM. The support vectors that define the hyperplane are indicated with black circles.

Obviously this procedure has two disadvantages: Firstly the computational requirements to calculate the additional descriptors; and secondly the increased computational requirements due to the higher dimensionality of the feature space. In some cases the dimensionality of the new space can be an infinite one. A radial basis function, for example is equivalent to mapping the data into an infinite dimensional Hilbert space.

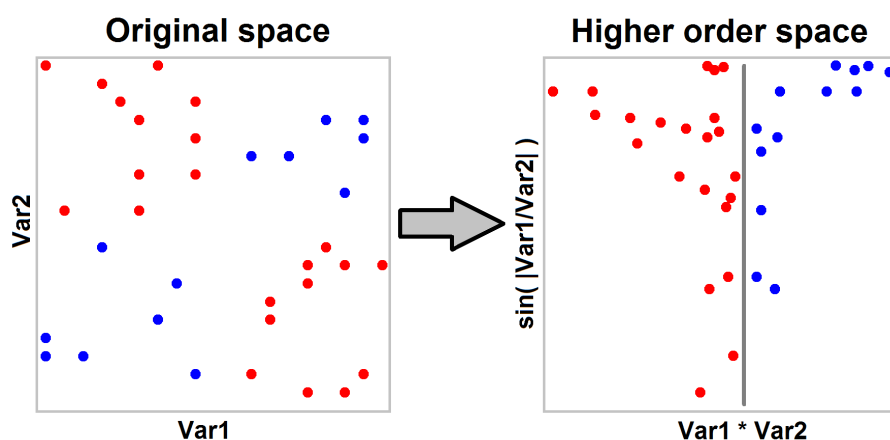


Figure 10. The transformation of a classification problem that is not separable in the original space (left) to the higher order space (right). The transformation to the higher order space on the right shows that the classes can be discriminated with only one dimension.

To avoid these problems, the transformation to a higher dimensional space is implemented using a Kernel function.^{55,56} The kernel function works in the original space, but ‘behaves’ like a dot product in a higher order space. Therefore it eliminates the explicit calculation of additional dimension.

1.4.1.3.2.4 Artificial neural networks

Artificial neural networks (ANN) originate from the effort of implementing a simplified replication of the human brain. The basic component within such networks is the artificial neuron. The best-known artificial neuron is the so-called perceptron.^{57,58,59} Fig. 11 shows the graphical representation of such a perceptron. Given a vector x of n input variables, the artificial neuron i is defined by n weights $w_{i1}, w_{i2}, \dots, w_{in}$, a transfer function Σ , an activation function f and a threshold θ . The transfer function sums the weighted input $\sum_{j=1}^n w_{ij} * x_j$ and the activation function is applied to the resulting sum. If the output of f exceeds θ , the neuron ‘fires’, otherwise not. Thereby the artificial neuron imitates the biological model and its underlying ‘all-or-nothing’-principle.

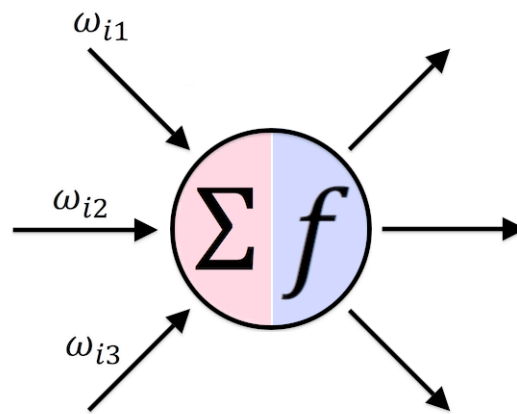


Figure 11. A schematic artificial neuron with its transfer function Σ and an activation function f .

In artificial neural networks, the neurons are arranged in layers and the neurons of a layer are interconnected with the neurons of the subsequent layer. Fig. 12 shows the scheme of such a network. It consists of an input layer, which processes a vector of descriptors, followed by a customizable number of hidden layers. The input of the neurons in a hidden layer is the output of the neurons of the preceding layer. The final layer of the neural network is the output layer. It usually consists only of one neuron, which emits the result of the classification procedure.

The training procedure of an ANN consists of the optimization of the weights of the neurons and aims to minimize the errors in prediction. The other parameters (Σ, f, θ) are generally predefined. There are numerous suggestions for the optimization procedure, using amongst other techniques such as the back-propagation algorithm^{60,61,62} or genetic algorithms.⁶³

Classically, the ANN is a directed, acyclic graph and only neurons in neighboring layers are connected, but there are also implementations allowing cycles or edges

overlapping one or more layers. The artificial neural networks used in this thesis are associative neural networks (ASNN).⁶⁴ ASNNs benefit from the supposedly LIBRARY mode,⁶⁵ which uses local corrections. These local corrections are calculated using a nearest-neighbor method which uses the predictions of an ensemble of models as new descriptors.

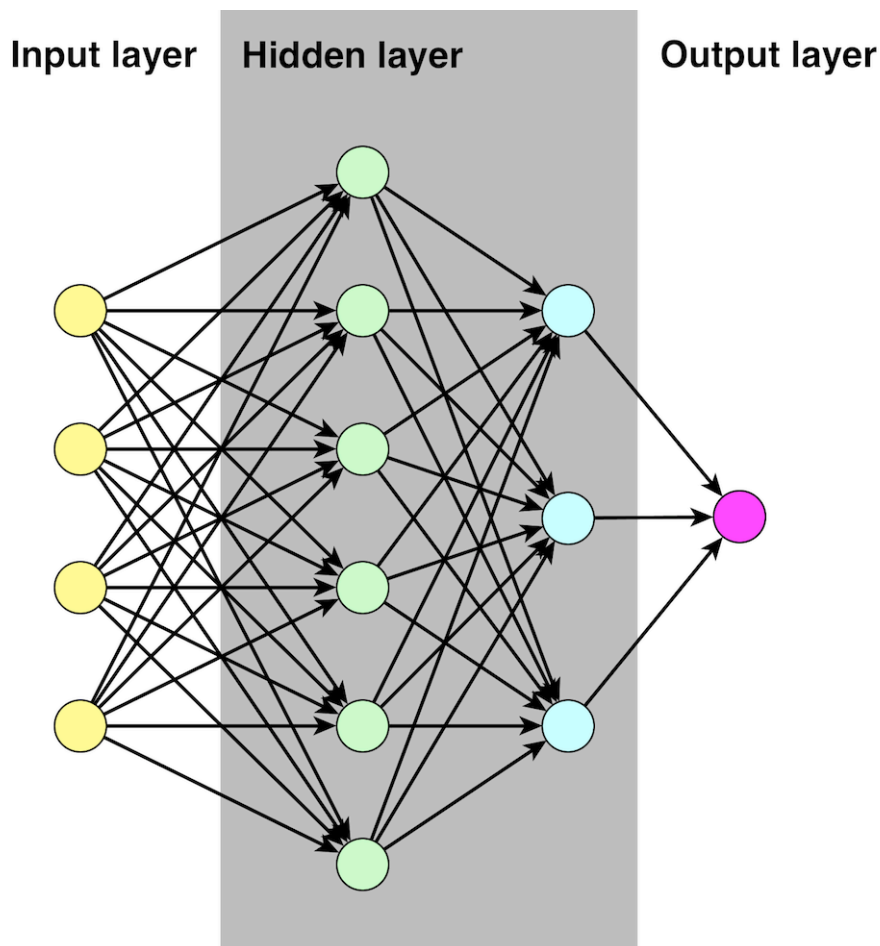


Figure 12. An artificial neural network. The neurons are arranged in layers. The connections between the neurons are directed towards the output layer with only one neuron.

1.4.1.4 Model validation

To obtain a reliable estimation of the predictive quality of a statistical machine learning model it is incorrect to rely on the performance reached only using data that were used for model training. This can be clearly seen when considering a hypothetical kNN approach using only one neighbor. In the chemical space the distance of a compound to itself is naturally zero, the prediction for a compound will surely be exactly the value (regardless if the model is for a discrete or continuous endpoint), which is already provided in the training set. Therefore all instances will be predicted correct and precise, although the values have just been learned by heart.

To avoid this problem, several approaches have been developed, all of them with approximately similar reliability and significance. The four most commonly used techniques are:

- **n-fold cross validation**

For the n-fold cross validation, the dataset is split into n folds of similar size. Usually this splitting is done randomly, but also stratified implementations, which distribute the compounds in a way that the instances in each split have a similar distribution regarding the target property, are used. In the next step a model is trained on each possible combination of n-1 splits, which leads to a final number of n models. All of these models are trained with the same initial parameterization and they are used to predict each instance in the excluded split. The number of folds used for this validation is usually in the range from five to twenty.

- **Leave one out cross validation**

The leave one out cross validation (CV_{Loo}) is a special case of the n-fold cross validation, where the number of splits is exactly the number of compounds in the dataset. This means that for the validation of each instance in a dataset, a model is trained on all other compounds within the dataset. Because of the high computational requirements, the application of this evaluation procedure is mostly used for small datasets with less than 200 instances.

- **External test set validation**

In case of the external validation, the initial dataset is split into a training set, which is used to develop and train a model, and a validation (or test) set, which is then used to evaluate the predictive quality of a developed model. If the number of instances in the validation set is smaller than the number of compounds in the training set, this procedure is similar to the n-fold cross-validation but just using one out of n possible folds. It can be recommended for large sets or for situations, when the computational costs of performing all analyses are too high.

The procedure of splitting the dataset is hereby of high importance. For sufficiently big datasets, a random procedure is the favorable one, but especially in case of smaller datasets (less than 50 instances), the probability to obtain an inappropriate distribution of the compounds grows. In case of a classification dataset this can for instance be that almost all instances of a certain class are assigned to either the test or the validation set. This can lead to an under-estimation of the predictive quality of the resulting model. Stratified approaches, as described for the cross-validation can help to avoid such side effects.

It is of high importance that the split into training and validation set is done before the model development, and that it is completely independent of the model evaluation. Unfortunately, this basic requirement is often violated.

The percentage of compounds used in each of the two splits is thereby highly dependent of the dataset size. In case of datasets with less than 100 instances, a validation set size between 20 and 35 percent is reasonable, but

in case of huge datasets, containing thousands of instances, a validation set size of 90-95% of instances can be reasonable as well.

- **Bagging validation**

The bagging validation works with an ensemble of models. In contrast to the cross fold validation, which produces exactly one prediction per instance, the bagging validation produces several predictions per compound and returns the average prediction value. A predefined number of splits of the data into a validation and a training partition are executed. The usual procedure to divide the dataset is the random sampling with replacement. In case of a dataset of n instances, the training set is formed by n times drawing an instance from the whole data pool. Taking into consideration the probability to draw the same instance multiple times, the training set contains $n * (1 - e^{-1})$ compounds ($\approx 63.2\%$) and the validation set $n * e^{-1}$ compounds. To ensure that each compound of the dataset is at least once present in any of the validation sets, the number of folds for the bagging validation is usually set to a minimum of 32 folds, but already for a number of 10 folds the probability that a compound is not present in any of the validation folds is approximately one percent.

All of these procedures have advantages, as well as disadvantages. The external validation makes it easiest to keep the strict separation between training and validation set and it can be convincing to the final reader. However it uses only a partition of the available data and thereby it can be biased. The leave-one-out cross validation (as well as the cross validation in general) can provide too optimistic predictions if the dataset contains duplicate measurements while the bagging procedure is computationally expensive.

Table 1. The four categories to distinguish binary classifications, depending on the observed and the predicted class.

		Measured / observed class	
		True positives fp	False positives fp
Predicted class	True positives fn		
	False negatives tn		

To quantify and numerically express the predictions quality obtained within the model validation procedures several statistical coefficients have been established.

1.4.1.4.1 Classification models

In case of classification models, each predicted instance gets assigned one out of four labels. These labels contain information if the instance is predicted to be within a certain classification category (positive) or not (negative) and if this

prediction is correct (true) or incorrect (false). Table 1 shows this differentiation, regarding the observed and predicted class. The decision which class is assigned the positive or negative label is done before the modeling and it depends on the nature of the analyzed property.

All statistical coefficients used to express the prediction quality of a classification model are based on the quantitative partition into four labels. The most commonly used ones are:

- **Accuracy**

The accuracy reports the proportion of correctly predicted instances.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

It is the simplest measure for the prediction quality of a model. However, this measure should not be used if the analyzed dataset is strongly unbalanced, as its expressiveness is limited. A model that classifies all instances as member of the majority class has an accuracy identical to the proportion of the majority class. If the majority class contains 90% of instances the accuracy of such a model is also 90%, although the model does not have any predictive power.

- **Recall**

The recall (or sensitivity) describes the proportion of correctly predicted instances of the positive class.

$$Recall = \frac{tp}{tp + fn}$$

- **Precision**

The precision reports the proportion of true predictions within all instances predicted to belong to the positive class.

$$Precision = \frac{tp}{tp + fp}$$

- **Specificity**

The specificity corresponds to the recall of the negative class.

$$Specificity = \frac{tn}{tn + fp}$$

- **F-Measure**

Recall and precision are usually used simultaneously, as the combination of these two measurements is appropriate to identify models with low predictive power. Referring to a classification model on a balanced dataset, that assigns all instances to the majority class, either recall or precision is

very low. A measurement to express the combination of these two values is the F-Measure. The F-Measure is the harmonic mean of recall and precision and it provides a reliable estimation of the overall quality of a prediction model, especially as it takes also the difference between recall and precision values into account.

$$FMeasure = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * tp}{2 * tp + fp + fn}$$

- **Balanced accuracy**

The most frequently used measurement to express the predictive quality of a QSAR classification model is the balanced accuracy. The balanced accuracy is the arithmetic mean of sensitivity and specificity.

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} = 0.5 * \left(\frac{tp}{tp + fn} + \frac{tn}{tn + fp} \right)$$

1.4.1.4.2 Regression models

The statistical coefficients for regression models basically estimate the differences between the measured values y_i and the predicted values \tilde{y}_i for all compounds i out of a dataset of size N .

- **Root mean square error**

The root mean square error (RMSE) quantifies the mean of the squared prediction errors of a model. As a value on its own, the RMSE retained from a specific model gives no information about the predictive quality, as it does not take the range and distribution of the target property into consideration. However, RMSE can be easily used if the performance of models is compared using exactly the same dataset.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \tilde{y}_i)^2}{N}}$$

- **Mean absolute error**

The mean absolute error (MAE) is complementary to the RMSE. It is less sensitive to large outlying predictions, which can heavily affect RMSE. Its use in QSAR modeling is less common.

$$MAE = \frac{\sum_{i=1}^N |y_i - \tilde{y}_i|}{N}$$

- **Coefficient of determination**

The coefficient of determination (Q^2) is a standardized value that enables the estimation of the model quality without further knowledge about the underlying data. Q^2 describes the variance in data explained by the model. The highest value of Q^2 is one (100% explained variance). There are

numerous implementations of the Q^2 , whereas the most commonly used version is calculated as follows:

$$Q^2 = 1 - \left(\frac{RMSE}{\sigma(y)} \right)^2 = 1 - \frac{\sum_{i=1}^N (y_i - \tilde{y}_i)^2}{\sum_{i=1}^N (y_i - E(y))^2}$$

- **Correlation coefficient**

The linear Pearson correlation coefficient (CC) describes the linear correlation between the predicted and the real values. The correlation coefficient gives only an overview of a similar development of measured and predicted value, but does not take the bias into account.

$$CC = \frac{\sum_{i=1}^N (\tilde{y}_i - E(\tilde{y})) * (y_i - E(y))}{\sigma(\tilde{y}) * \sigma(y)}$$

The square correlation coefficient is often denoted R^2 . It is worth mentioning that literature provides several inconsistencies with the usage of Q^2 and R^2 . R^2 is often used to describe the Q^2 for the training set and sometimes both denotations are interchanged. To avoid any misinterpretations, in this thesis the coefficients are strictly used as described above.

1.4.1.4.3 Statistical significance

Given a specific dataset, the application of the aforementioned criteria can detect the superiority of a certain prediction approach (e.g. the combination of a certain machine learning approach with a certain descriptor set) towards others. However, the numerical values retained by these criteria allow just the interpretation of this very narrow excerpt, but without further ado they cannot be used to deduce a general superiority. Differences in performance derived with one single comparison are not necessarily generalizable, as they might be caused by chance. The exclusion or consideration of only one chemical compound can influence the performance of a certain approach. It is therefore necessary to evaluate statistical methods, such as QSAR modeling on a wider scope.

A commonly used method that is used to enable such a generalization is the examination of the statistical significance. The underlying idea of the statistical significance is the refutation of the so-called null hypothesis (H_0). When comparing two methods, H_0 usually claims that both methods perform equally well. The validity of H_0 is estimated from an ensemble of observations. These observations are used to estimate the statistical probability (p-Value) that H_0 is correct. There are numerous statistical tests to detect the level of significance, depending on, which statistical coefficients are compared, e.g. the mean values or variances of the examined distributions. The most reliable and most comprehensible method is the binomial test, which is based on the Bernoulli distribution.

In this thesis, the pairwise comparison of results using the binomial test is the only method to detect statistical significance. Furthermore, we follow the most common recommendation to set the level of significance to refuse H_0 to a p-Value of < 0.05 .

Additionally, results that are obtained with p-Values < 0.001 are called *highly significant*.

In QSAR, as well as in many other areas of statistical applications, the so-called confidence intervals are frequently used as an indication of statistical significance. The confidence intervals usually display the standard deviation or the standard error. Overlapping confidence intervals are thereby used to show that the observed difference is non-significant, whereas non-overlapping confidence intervals are used to show that the observed difference is significant. This has been shown to be basically wrong in several publications.^{66,67}

Therefore, in this thesis the standard error and the standard deviation are only used to estimate the uncertainty in prediction, as well as predictive confidence of a model.

1.4.2 Experimental design approaches

The problem of optimizing scientific experiments, to obtain the maximum information with a minimum number of tests, is well known since more than 200 years. First efforts to systematically design efficient experiments are reported from health studies in the middle of 18th century. The development towards an established area of scientific research and the attempt to generate general solutions can be traced back to the beginning of 19th century.

The optimal design of experiments for regression models with known depending variables was subject to exhaustive research in the beginning of the 20th century. Although QSAR mainly deals with regression problems, these approaches are not applicable to current chemoinformatics problems, basically because in most cases the depending variables are not known and their detection is part of the experimental design problem.

In general, the problem of experimental design in computational chemistry can be seen as the problem to select the most representative subset of compounds from a larger collection. The selected subset should describe the original set of interest to a high extent and moreover, it should represent it respective to a certain property of interest. The subset should also enable both the detection of relevant structural features to describe the property of interest, as well as the quantification of these features. According to this, the basic task is to find a combination of compounds that fulfills two criteria:

1. Each selected compound should be representative for a preferably high number of other compounds in the dataset.
2. The information contained in the selected subset should not be redundant.

With these requirements given, it is obvious that, with regards to the first prerequisite, the selection of structurally diverse compounds should be avoided, as their chemical features might not be applicable to other compounds within the dataset of interest. On the other hand, respective to the second prerequisite, it should be avoided that the final selection contains highly similar compounds.

A variety of different approaches^{11,12,68,69,70,71} have been developed, all of them aiming to select a representative subset of compounds to deliver the most reliable model. All these approaches work with a depiction of the chemical space, which is exclusively based on descriptors. Usually a PCA is applied to these descriptors to extract the principal properties, which are used to span the chemical space to select compounds. Although the statistics literature also provides a large variety of sequential approaches⁷² that refine the representation of the chemical space in a stepwise manner, as well as Bayesian approaches⁷³ that take preliminary information into account, their application in QSAR is very limited.

The difference between the selection approaches is established by the way they evaluate the representativeness of a certain compound or the quality of a combination of compounds. All these approaches can be reduced to three general ideas: firstly the concept of selecting the most descriptive compounds, which is

referred to as similarity selection; secondly the concept of selecting the combination of compounds, which provide the most diverse information, which is referred to as dissimilarity selection; and thirdly, the concept of representing the whole chemical space of interest, referred to as partition based approaches.

In the next chapters these approaches will be explained in more detail as well as their basic features and the advantages and disadvantages linked to them.

1.4.2.1 Dissimilarity selections

The underlying idea of all dissimilarity selection approaches is that the most diverse compounds provide the most widespread information and that they therefore reach the highest coverage of relevant knowledge. Fig. 13 exemplifies their mode of action in a two dimensional space, with the main principle components representing the axes and each black dot representing a compound within the dataset. Compounds surrounded by a red circle indicate those compounds that were selected.

Typically for dissimilarity approaches, the selected compounds are mainly located at the periphery of the data cloud. This bias towards compounds beyond the center of the distribution results in a high sensitivity against the underlying data distribution.

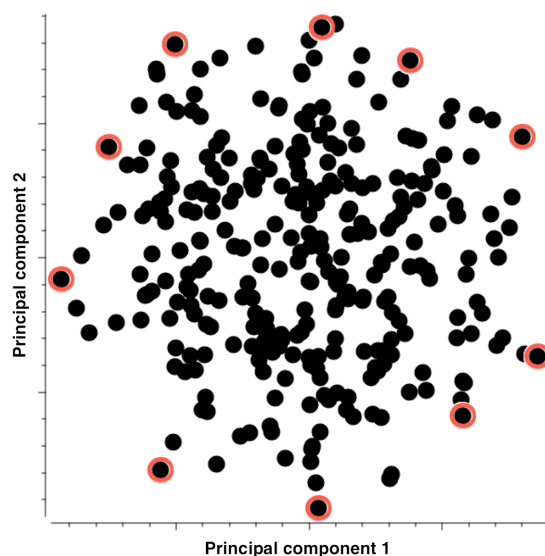


Figure 13. Compound selection with a dissimilarity approach. The selected instances (indicated with red circles) are all at the periphery of the data cloud.

On the one hand, dissimilarity selections work well for datasets, with limited chemical diversity that cover only a compact, not too widespread subspace of the chemical space. In the case of linear problems and known depending variables, dissimilarity selections perform with a high reliability. On the other hand, they reveal a tendency towards the selection of outliers, which results in negative effects, especially for datasets with structurally diverse compounds or in high dimensional spaces.¹¹

The most frequently used variants that belong to this group of selection approaches are the Kennard-Stone algorithm⁷⁴ and the D-Optimal design.^{44,75,76}

1.4.2.1.1 D-Optimal criterion

“D-Optimal design, which has been recommended as the favorable alternative for linear models in several publications,^{44,75} selects the most representative combination of compounds for linear models.⁷⁶ In this method, each possible subset of a given size is evaluated to derive the information matrix.”[a]

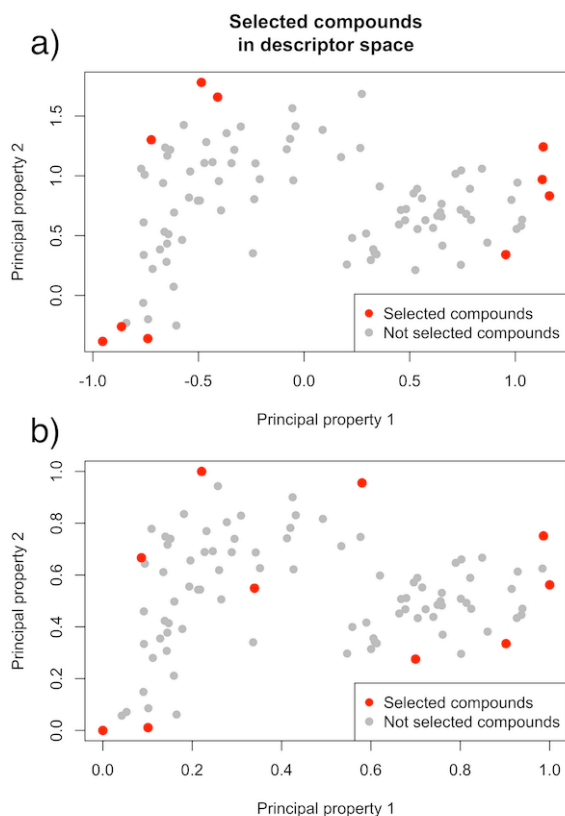


Figure 14. The results of the D-Optimal selection using a) linear terms only, b) linear, cross, and square terms. [a]

The information matrix I is hereby derived from the data matrix X , as it was introduced in 1.4.1.3.1.1.

$$I = (X^T X)^{-1}$$

“The most distinct and thereby most optimal of all possible subsets is the one with the maximum determinant of the information matrix. This is equivalent to the set with the maximum entropy.⁷⁷ An advantage of the D-Optimal selection criterion is that in the design problems that are analyzed in this work, the training set is selected from a limited candidate set. Pronzato has shown that when the data space is limited, a sequential D-optimal design, given that some conditions are met, is asymptotically optimal.⁷⁸

Fig. 14a) shows the result of a D-Optimal selection. The x-axis and the y-axis represent two first principal components, while each dot represents a chemical compound. Dots marked red are the compounds that were selected using the D-Optimal criterion.”[a]

As mentioned in the previous chapter, the application of the D-Optimal criterion to principal components is particularly capable of problems with linear dependencies. In contrast, the performance of models derived from compounds selected with the D-Optimal criterion is clearly decreased for non-linear dependencies, such as quadratic or hyperbolic ones. An explanation therefore can be seen in Fig. 15a-f. The x-axis shows the descriptor space and the y-axis the property space. The grey line shows the real dependency between the structural properties and the endpoint of concern, whereas the red and green dots illustrate available measurement values.

The D-Optimal criterion selects compounds from the periphery of the descriptor space, which causes a high structural diversity. The maximum structural dissimilarity usually causes also a dissimilarity regarding a certain endpoint. Extreme structures often account for extreme values regarding certain endpoints. The compounds selected by the D-Optimal criterion are therefore in most cases the ones with the highest or lowest value regarding the target property. This can be problematic if the dependency between the descriptor space and the property space is non-linear.

If the selection of compounds contains only those with extreme values, as indicated by the red dots in Fig. 15a, the derived regression line (dashed, red), which is added in Fig. 15b enables good predictions in the area of the extreme values, as they were used for the resulting model. But for intermediate values, which are allocated in the center of the descriptor space (and which usually are the most frequent ones), the prediction error is relatively high, which is illustrated by the red arrow in Fig. 15c. Taking additional compounds from the distribution center into consideration can eliminate suchlike effects. The green dots in Fig. 15d represent such compounds. They can be interpreted as a correction factor, to get a better indication of the shift (or bias) that is added to the new regression line, which is shown as a green dashed line in Fig. 15e. Although the prediction error for values at the borders of the descriptor space increases, the overall prediction quality is clearly increased, whereas the maximum error is decreased.

To address this problem, the D-Optimal criterion is frequently applied not only to the principal components but also to a set of meta descriptors. “These meta descriptors contain the normalized components from PCA, their square terms, and their pairwise cross terms.⁷⁹ For a set of v input variables, d_1, d_2, \dots, d_v , additionally the square terms $d_1^2, d_2^2, \dots, d_v^2$ and the cross terms $d_i d_j$ with $i = 1, 2, \dots, v, j = 1, 2, \dots, v$ and $i \neq j$. This extension increases the dimensionality of the search space by a quadratic factor from v input variables to $v * (1.5 + 0,5 * v)$ meta descriptors.

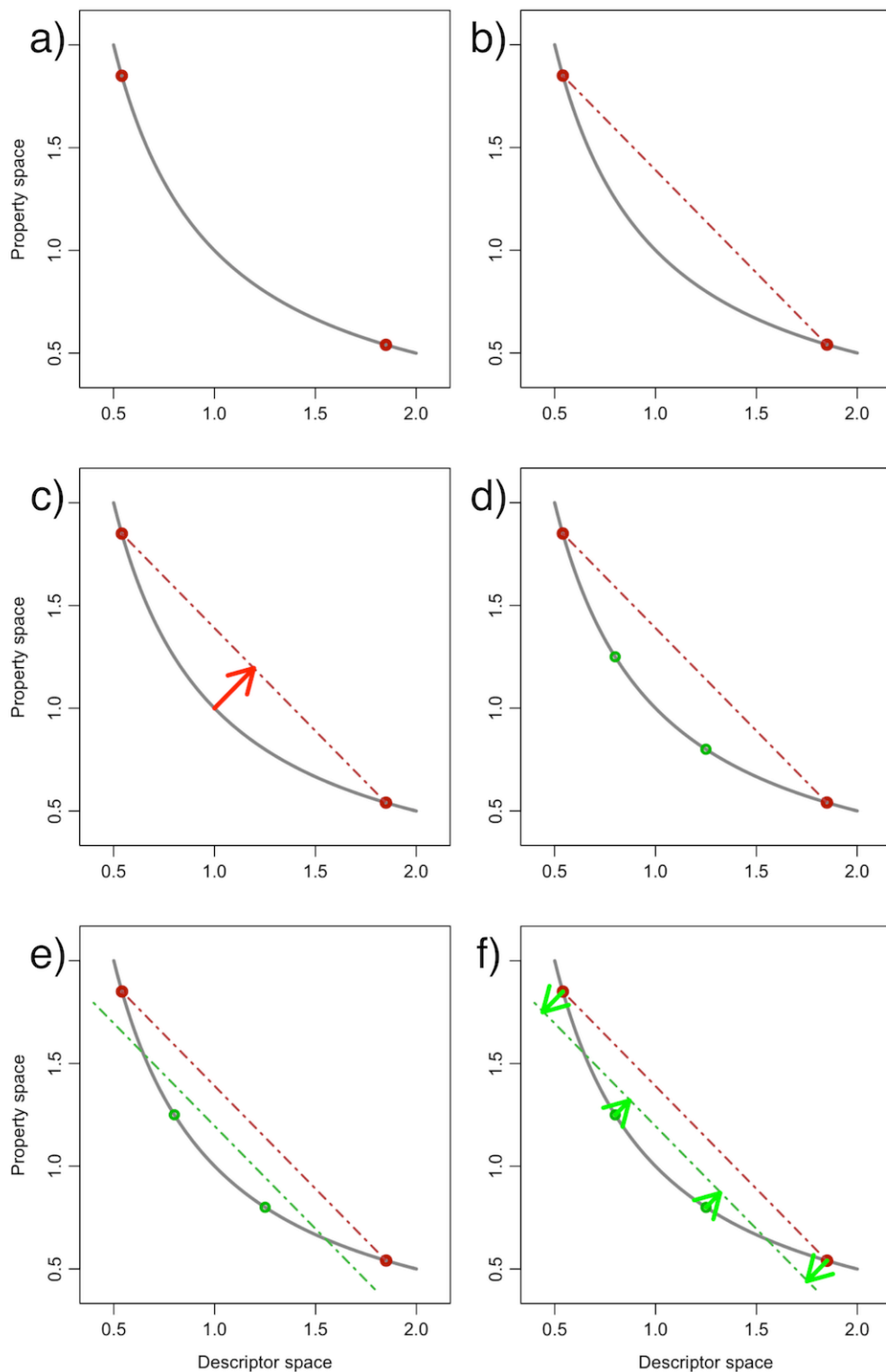


Figure 15. Problems arising from a compound selection that takes only extreme compounds with values of the target property into account (a-c) and the effects of the inclusion of values with intermediate values of the target property (d-f). Whereas the consideration of only extreme values results in a comparably high prediction error at in the center of the data distribution, which is usually the most densely populated area, the additional consideration of intermediate values can clearly improve the prediction quality.

This contributes to the quality of the outcoming sample, as the selected compounds are not located exclusively on the periphery of the data cloud in the chemical space but also in the center. Fig. 14b) shows the resulting selection on the same dataset as Fig. 14a). “[a]

1.4.2.1.2 Kennard-Stone algorithm

The Kennard-Stone algorithm⁷⁴ selects compounds in a fixed order. Derived from an initial selection, the compounds are chosen sequentially. From this initially selected seed, each step in the selection process extends the collection of chosen compounds by that one that has the highest Euclidean distance to its closest neighbor within the previously selected ones.

1.4.2.1.3 Sphere exclusion

A further variation of the dissimilarity selection is the sphere exclusion method.^{80,81} Similar to the Kennard-Stone algorithm, the sphere exclusion method selects compounds in a fixed order. The selection starts with the compound with the largest sum of Euclidean distances to all other compounds in the dataset. Referring to a predefined sphere exclusion radius, all compounds that are closer to the selected compound than this radius, are supposed to be represented by the selected compound and they are removed from the list of selectable molecules. The next compound to be selected is the one most distant from all previously selected compounds. All compounds within the sphere exclusion radius are also removed from the list of selectable compounds. This procedure is iteratively repeated until the list of selectable compounds is empty. Contrary to the D-Optimal criterion and the Kennard-Stone algorithm, the sphere exclusion method cannot select a predefined number of compounds, as the underlying stop criterion depends on the exclusion radius.

1.4.2.2 *Similarity selections*

The general assumption of the similarity selection approaches is diametrically opposed to that of the dissimilarity selections. The basic idea is that QSAR works on the similarity, not on dissimilarity of molecules.^{70,71} Therefore the most informative and descriptive compounds within a dataset might be those with a high similarity to a preferably high number of other compounds within the dataset of interest. Similarity based approaches usually work in a sequential manner and their mode of action is shown in Fig. 16. Starting with an instance in the center of a densely populated area, which is represented by the highlighted dot in the top right corner, they assign the compounds within a defined area around the selected compound as represented and do not take them into consideration in all further steps.

The advantage of similarity selection approaches is their comparably high robustness against structurally diverse compounds in a dataset. These approaches focus on the center of the data distribution, which minimizes the influence of

structurally diverse compounds on the selection. Additionally, as these approaches select compounds with a high structural similarity to a preferably high number of other compounds, they converge fast and deliver reasonable predictions even with a low number of selected compounds.

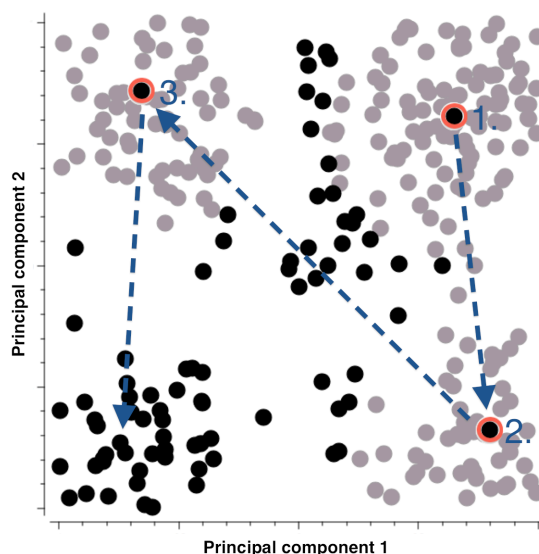


Figure 16. The selection mode of a similarity based approach. Compounds are selected in a stepwise manner and compounds in densely populated areas are preferred.

Similarity based approaches are successful in detecting natural clusters within the data distribution, but their selection approach is arbitrary if such clusters are missing. Similarity approaches can hardly handle equally distributed or centric data and furthermore, they reveal a bias towards disregarding the periphery of the data clod. An additional disadvantage, which makes the application of such approaches uncommon, is that most of them require an elaborate parameterization.

1.4.2.2.1 MDC selection

“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 of 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 criterion hereby is to select the compounds with a low sum of pairwise distances to other compounds, as these compounds are estimated to provide the highest information content for all other compounds. The compounds are selected sequentially, and after each new compound is selected, the contribution of that compound is eliminated.”[b]

1.4.2.2.2 Other similarity based approaches

Apart from the MDC approach, there are other implementations that rank the compounds' representativeness by their pairwise structural similarity to all other compounds, but as mentioned before, all these methods require expert knowledge

for a proper parameterization. All these approaches result in selections comparable to the one resulting from the MDC selection and are therefore not discussed in detail, and will not be used within this thesis.

1.4.2.3 Partition based approaches

The last relevant group of selection approaches are those that aim to cover the whole chemical space of interest, namely the partition based approaches. The full (or fractional) factorial design¹⁰ and space filling designs⁶⁸ are examples thereof. Partition-based approaches attempt to select a sample that is representative for all relevant compound, by separating the descriptor space into subspaces and finding a representative compound for each of these subspaces.¹⁰ Fig. 17 exemplifies the underlying idea to use a 'grid' to partition the two-dimensional chemical space into subspaces of similar size.

One of the disadvantages all the partition based approaches suffer from, is that the number of compounds to be selected cannot be fixed.¹¹ The reason therefore is that in general the molecules' distribution in the chemical space is heterogeneous. The resulting consequence are cells (or subspaces), which are unoccupied as, due to the laws of chemistry, no compound with the required structural qualities exists. Although the partition based approaches work well for equally distributed datasets, they reveal problems with inhomogeneous data distributions.

An additional problem is that such approaches allow only the use of a very limited number of dimensions to span the search space. This is due to an exponential increase in the number of subspaces with each additional dimension which is taken into consideration.⁴⁴

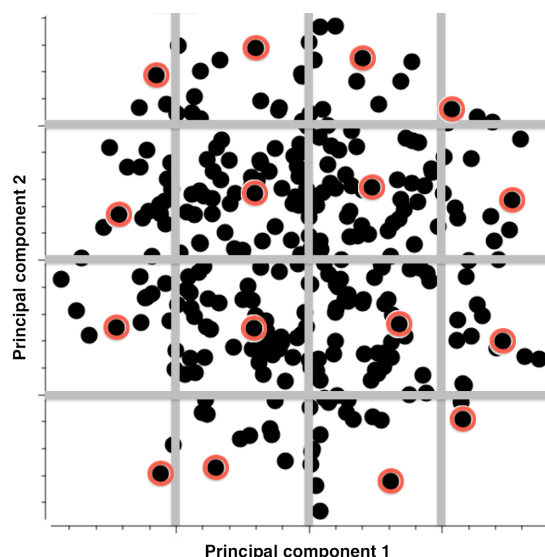


Figure 17. The basic idea behind partition based approaches. The chemical space is divided into subspaces and from each subspace a representative compound is chosen.

The classical factorial design¹⁰ works with a binary representation of each variable, which is used to describe the search space. Instead of a numerical value, the compounds are assigned the labels 'high' or 'low' for each considered variable.

For n relevant variables, this results in 2^n possible combinations. The full factorial design aims to find one representative for each of these combinations plus the most central point in the dataset as a respective calibration point.

The procedure for the fractional factorial design works in a similar manner, with the difference being that only a fraction of all possible combinations is selected. This is usually reached, by the precondition that all instances in the final selection are required to differ in the label-values of at least two variables. Contrary, the compounds in the final selection of the full factorial design have to vary in the label of only one variable. This is equivalent to a 50% decrease in the number of required measurements.

1.4.2.4 Selection of a design approach

All the aforementioned approaches revealed problems for certain preconditions. It is therefore impossible to refer to one of these approaches as 'the best'. Fig. 18 shows that apart from the shape of data distribution, several other parameters have to be taken into consideration.

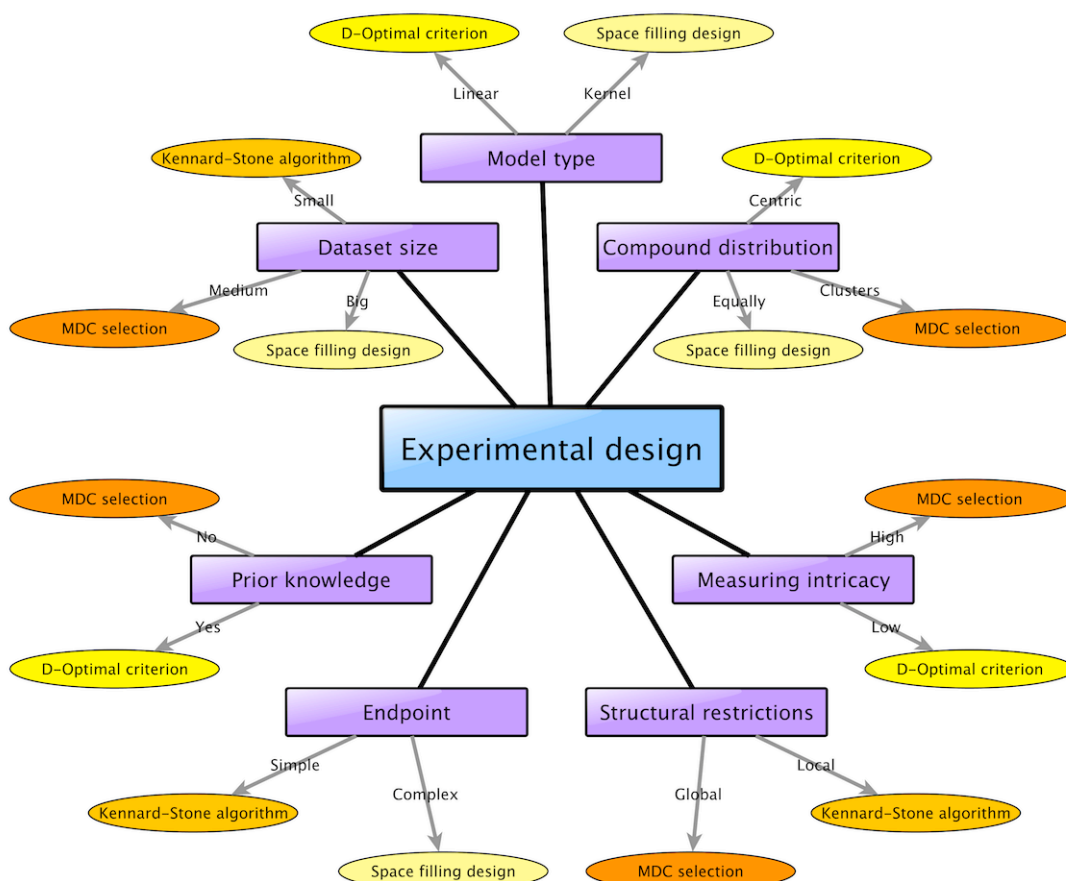


Figure 18. The problem of finding the optimal experimental design approach. Each purple rectangle represents a specific condition that is relevant in terms of the decision for a certain approach.

Prior knowledge about (linear) correlations between the descriptor space and the target property might suggest the D-Optimal criterion as the most reasonable

choice, whereas a structurally unrestricted dataset might result in a preference of a space filling design. Other parameters, such as the intended type of machine learning algorithm to train the model, the dataset size or the endpoint of relevance should have an influence on the final decision. Furthermore, the mentioned parameters for the decision for a certain approach can be ambiguous, or result in conflicting suggestions.

Still, the most relevant predicate for the evaluation of a selection approach is the underlying data. With this in mind, the next chapters will exhaustively introduce and discuss the datasets that were used for the evaluation of the results (presented later in this thesis).

2 Materials

To validate the performance of the stepwise method, five regression datasets and two classification datasets, were collected from the literature. All seven datasets varied regarding the considered endpoint. These endpoints were selected with respect to the REACH legislation and risk assessment in general. Amongst others, the endpoints contained of a physicochemical property, an adsorption coefficient, toxicity measurements against freshwater fish and a protozoa and protein inhibition. Furthermore, to cover a broad spectrum of possible applications and to better examine the performance of newly developed method, the sets collected varied in several other criteria: size, modeling and measurement complexities.

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

In order to avoid problems in descriptor calculation, we excluded inorganic compounds, radicals, charged molecules, and salts from all collected datasets (both for regression and for classification). Further, from all regression datasets we removed compounds for which no exact values, rather an interval or only minimum or maximum values, were given. All datasets were free of duplicate compounds and only of them was structurally restricted thus the sets represent a wide chemical diversity.

“For each dataset, a collection of two types of descriptors was calculated. The first type was calculated using the ALOGPS 2.1 program⁸² and contained two descriptors: solubility and lipophilicity of molecules. ALOGPS was the top-ranked model for prediction of logP.⁸³ The second type included E-State indices^{84,85} These are electrotopological descriptors calculated for each atom and each bond in a compound and then summed according to their types over all atoms. The number of descriptors for the second type is determined by number of different chemical groups and thus it was not a fixed one.

The Online CHEmical database and Modeling environment (OCHEM)⁸⁶ was used for the calculation of the descriptors. To represent the chemical space of each dataset the descriptors were normalized to a [0,1] range. The rationale to use normalization instead of standardization is that standardization works on the underlying assumption that the objects are normally distributed. This assumption is not true for descriptors determined for chemical groups, e.g., in particular for the E-State indices. As they are linked to the presence of certain substructures, for most compounds, their value is just zero.”[c]

In the following chapter, all seven datasets will be highlighted regarding their endpoint and the underlying data source. Furthermore, to allow a more detailed insight, we analyzed the datasets regarding their data distribution in chemical

space and their eligibility in terms of QSAR modeling. Both examinations are of high relevance, to correlate the performance retained from the following experimental design to expectable results and to facilitate the applicability of different selection approaches to certain data distributions and comprise the limitations.

The distribution analysis was performed using PCA to extract the five dataset-characteristics with the highest variance and therefore with the highest descriptive power. The subsequent modeling of the data was realized in a five-fold cross validation using the PLS technique. The descriptors used for both multivariate techniques were the E-State indices, as well as ALog_{PS} descriptors.

2.1 Regression datasets

In this chapter, we will introduce the datasets on regression endpoints that were used for this study. The five regression datasets build the core entity for the evaluation of selection approaches. Detailed statistical specifications, as well as other characteristics are shown in Table 2.

Table 2. Comparison of the five regression datasets. The analysis takes into consideration general characteristics, such as the measurement intricacy of the endpoint or the size of the dataset (shaded green), as well as the chemical representation and distribution of the data (shaded orange) and the results of the evaluation of the developed PLS models (shaded blue).

	Boiling Point	logK_{oc}	logBCF	logLC₅₀	-log(IGC₅₀)
Endpoint	Physico-chemical	Sorption coefficient	Biological	Toxicity	Toxicity
Measurement uncertainty	Low	Medium	High	High	Medium
Compounds	1198	648	238	535	1093
Structural restrictions	Halogenated	No	No	No	No
Minimum	-85.7	0	-0.22	-7.92	-2.67
Maximum	378.9	7.05	5.97	-0.04	3.34
Standard deviation	85.3	1.25	1.37	1.34	1.05
Descriptors	232	232	122	178	182
Distribution	Dense, loose periphery	Triangular, no outliers	Centered, scattered outliers	Consistently, expanded	Clustered, few outliers
PLS latent variables	12	4	3	5	7
Variance covered	41.7%	21.5%	27%	24.8%	37%
RMSE	24.8942	0.4645	0.5899	0.6835	0.4635
Correlation coefficient	0.9569	0.9281	0.9028	0.8595	0.898
Q ²	0.9148	0.8615	0.8157	0.7383	0.8062

2.1.1 Boiling point

The first dataset was collected for the boiling point (BP). The experimental determination of this physicochemical property is straightforward and usually without complications. Correspondingly, the metering precision is high. The only experimental condition that has to be taken into consideration is the ambient pressure.

Measurements for the same compounds, but gathered from different sources, usually vary less than 0.5°C, and only in extreme cases more than 1.0°C. With this in mind, the requirement to limit the data to measurements from only one

laboratory, is needless and we decided to use data extracted from the EPI suite.⁸⁷ The EPI suite contains a collection of physicochemical measurements derived from various sources.

As a consequence of the fast and comparably cheap testing procedure, literature provides large amounts of measurement results, which is reflected in an amount of more than 5.500 boiling point values for different compounds, contained in the EPI suite. In order to decrease the number of considered compounds to a reasonable size, and to enable also the examination of the selection approaches on structurally restricted datasets, we kept only measurements on compounds, containing bromine, chlorine and/or fluorine and disregarded all other measurements.

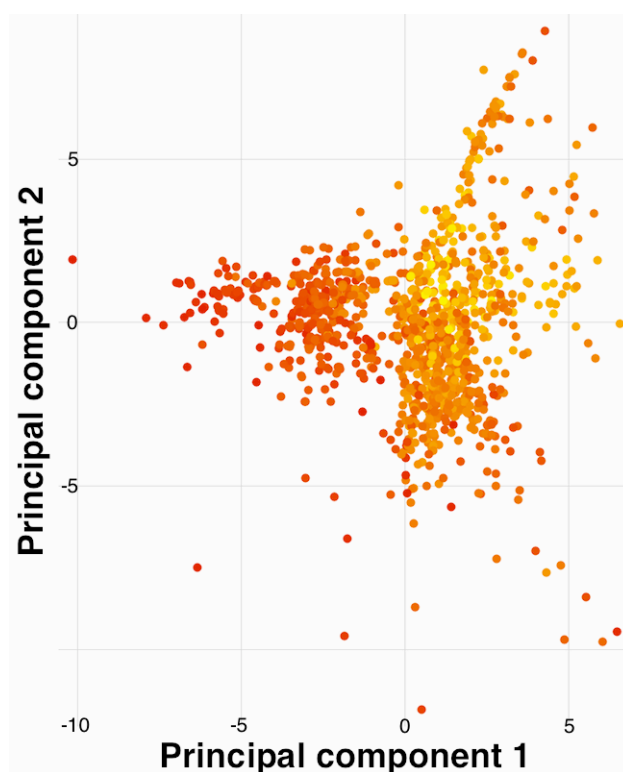


Figure 19. The compound distribution for the boiling point dataset in PCA space. Dark red dots represent compounds with a high boiling point, yellow dots represent those with a low boiling point.

The resulting collection contained 1198 measurements for halogenated compounds, all of them obtained at 1.0 atm ambient pressure. As no further structural filters were applied, and as halogenated compounds do not form a homogeneous chemical class, this set still provided a broad diversity with regards to the molecule size and chemical structures, which can also be seen in the comparably wide range of contained boiling point values (450°C).

An analysis of the five most important principal components revealed that these components encode for: 1) the compounds aromaticity; 2) fluorine substitutions; 3) polarity; 4) hydrophobicity; and 5) the bond saturation.

The data distribution is illustrated in Fig. 19. The x-axis displays the first principal component, whereas the second principal component is displayed by the y-axis. Each compound is represented by a dot and the colors of the dot display the value

of the target property. Those compounds with high measured values are highlighted in red and the compounds with low measured values are indicated in yellow. The distribution shows a densely crowded center of the chemical space, with three comparably crowded outgrowths around. The distribution on the periphery beyond this center is loose. Only few compounds reveal a significant discrimination towards the majority of compounds, hence no real outliers are observable. Furthermore, a slight correlation to the target property can be identified in the PCA distribution. The aromatic compounds (on the left in the figure) are those with a higher BP, whereas the fluorine containing compounds (top) are aligned with a low BP.

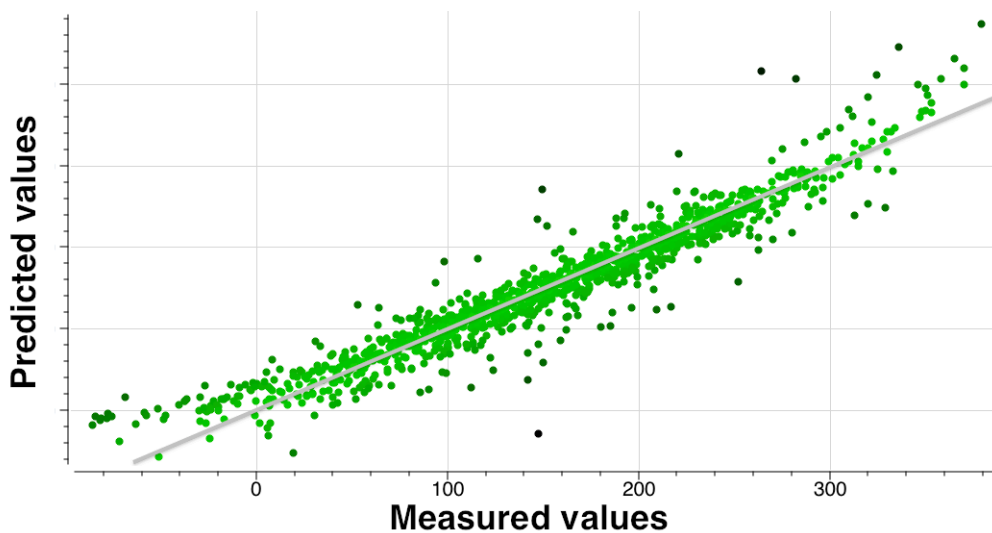


Figure 20. Measured versus predicted values for the boiling point model [°C].

Furthermore, these five components covered 10% of the data variance that is contained within the 232 used descriptors. Other important statistical values are:

- 35 components covered 50% of the data variance
- 87 components covered 80% of the data variance
- 129 components covered 95% of the data variance (56% of the used descriptors)

This small coverage provides an indication how widespread the variability of compounds within the dataset still is, despite the limitation of halogenated molecules.

Fig. 20 shows the results of the PLS regression on the dataset. The x-axis displays the measured values and the y-axis displays the predicted values and the grey diagonal indicates the area of optimal predictions. It is obvious that the majority of compounds are predicted with a high accuracy, which is emphasized by a $Q^2 > 0.9$. The model exhibits a low prediction quality for only a few compounds. Particularly for compounds with measured boiling point values from 20°C to 300°C; the correlation between measured and predicted values is high, whereas the predictions for compounds beyond those borders reveal a tendency towards frazzling.

Correlating to the high measurement precision, also the modeling of the endpoint is relatively simple. Furthermore, the boiling point is one of the properties which have undergone exhaustive research. This is reflected in the variety of available global^{20,88,89} and local models.^{90,91,92} Global models mainly depend on the compound size and polar properties, but topological, topo-chemical and geometrical variables in general were shown to improve global models.⁹³ With local models, which are trained only on a certain class of compounds, the property can be predicted to a precision up to 1-2°C. Although our boiling point dataset is structurally restricted to halogenated compounds, it contains a variety of different chemical and structural classes. The observation that the prediction accuracy in our model is clearly lower (30°C) is a logical consequence.

2.1.2 logK_{OC}

The endpoint of the second dataset is an adsorption coefficient, namely logK_{OC}. The dataset contained 648 measurements and it was based on the reviewed collection of Meylan et al.⁹⁴ Although logK_{OC} is no REACH endpoint, it is of high importance for risk assessment.⁹⁵ logK_{OC} is frequently referred to as a partition coefficient, which is incorrect. In fact, it is a log-scaled adsorption coefficient, and it describes a material's tendency to adsorb to soil particles.

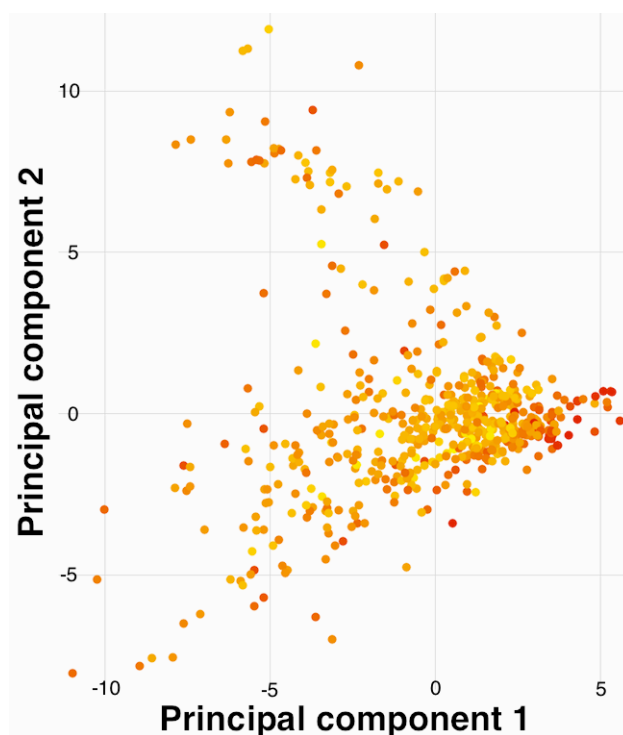


Figure 21. The compound distribution for the logK_{OC} dataset in PCA space. Dark red dots represent compounds with a high logK_{OC}, yellow dots represent those with a low logK_{OC}.

The properties, represented by the five most important principal components are: 1) the length of the skeletal backbone; 2) non-aromatic substitutions with inorganic groups; 3) the compound size; 4) aromaticity; and 5) lipophilicity. Fig. 21 displays the distribution of the measurements referring to the first and second principal components. The basic shape of the data cloud is triangular, which can be

frequently observed for depictions that are dependent on a descriptor or component, which represents the compound size or length.

A dense accumulation of high $\log K_{OC}$ values can be observed for high values of the first PC and intermediate values for the second PC, still in the same area there are also numerous compounds with very low values for $\log K_{OC}$. In general the first two PCs deliver a very unordered depiction which does not allow any conclusion about the target property. Furthermore, the dataset contains no structural outliers.

Five components covered 14% of the data variance contained within the 232 used descriptors. Further specifications regarding the variance are:

- 28 components covered 50% of the data variance
- 72 components covered 80% of the data variance
- 114 components covered 95% of the data variance (49% of the used descriptors)

This indicates the widespread chemical diversity within this intermediate size dataset.

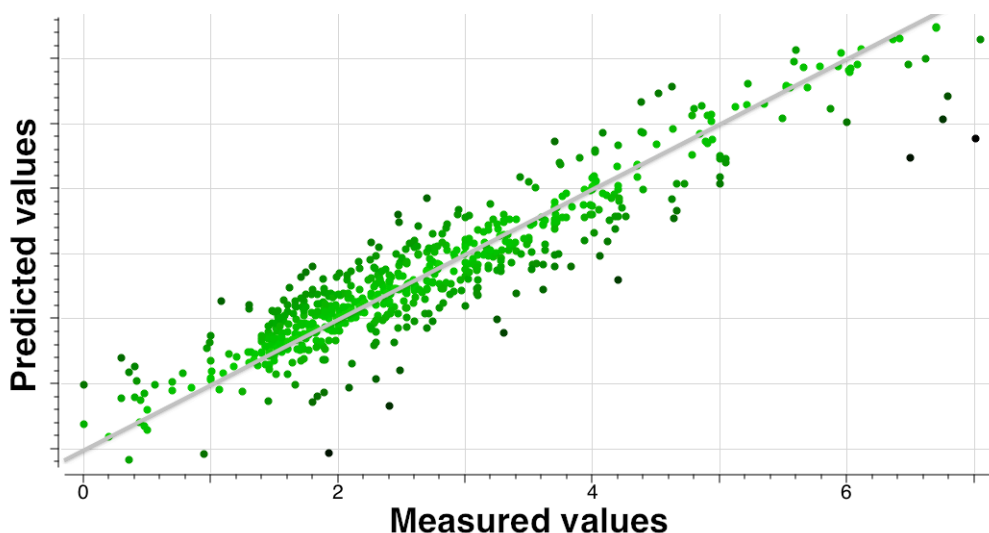


Figure 22. Measured versus predicted values for the $\log K_{OC}$ model [log unit].

Fig. 22 displays the model predictions for the $\log K_{OC}$ dataset. The legend is the same as for Fig. 20. As the measurement intricacy is low (although not comparable to that of the boiling point) the related model provides the second highest Q^2 value. Additionally, the number of latent variables used for the PLS model is low. This is reasonable, as the soil organic carbon-water partitioning coefficient is known to mostly depend on the lipophilicity of a compound.^{96,97,98,99}

Regarding the distribution of property values, the dataset consists of measured $\log K_{OC}$ values with a range of more than 7 log units, but the majority of compounds exhibits values within a range from 1.3 - 4.5. Especially in areas with high $\log K_{OC}$ values, the predictions reveal a bias towards an underestimation.

2.1.3 logBCF

“The endpoint of the smallest of the datasets was the log-scaled bio-concentration factor (logBCF) in fish. The set contained 238 different compounds and was taken from a study done by Gramatica et al.¹⁰⁰ The authors originally split the measurements into a training set of 179 compounds and a validation set of 59 compounds. These datasets are freely accessible in the QMRF database of the European commission,¹⁰¹ as well as in the On-line CHEMical database and modeling environment (OCHEM).⁸⁶ The dataset, as we used it in our study, was merged from the original split.”[e] The underlying measurements were primarily collected by Lu et al.¹⁰² and subject to other publications as well.¹⁰³

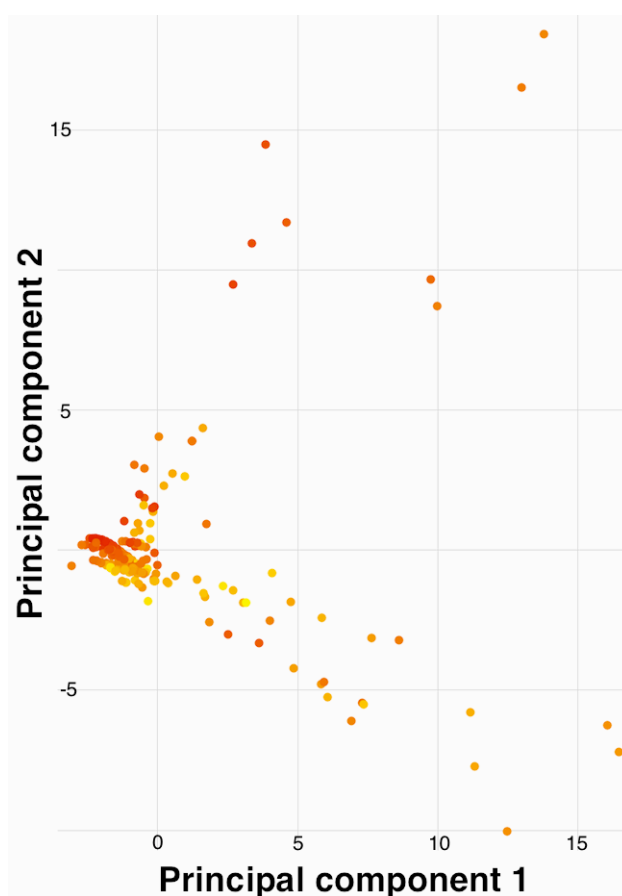


Figure 23. The compound distribution for the logBCF dataset in PCA space. Dark red dots represent compounds with a high bioconcentration factor, yellow dots represent those with a low one.

Although the examination of the bioconcentration factor has been conducted since the late 70s,^{104,105,106} the research on this endpoint gained importance with the upcoming REACH regulation.^{107,108,109} The bioconcentration factor gives an overview of how intensively a compound is accumulated in an organism. It depends on numerous experimental conditions, such as the temperature, the time of exposure, the lipid content of the considered tissue, or the test duration, and experimental testing is expensive. The conditions within the used measurement collection were not consistent, but varied for example regarding the fish species.

Within all datasets used in this study, the one on logBCF is the smallest. It mainly consists of typical environmental contaminants, which originates from agricultural and industrial use. The dataset contains halogenated benzenes, phthalates, anilines and biphenyls amongst others.

The first five principal components encoded were the: 1) size of the organic backbone; 2) degree of branching; 3) aromaticity; 4) lipophilicity; and 5) content of Nitrogen, Sulfur or Phosphor. The effect of the backbone length in a PCA depiction, as mentioned in the previous chapter, is reflected in the compound distribution (Fig. 23). The triangular shape, as it is observable for the logK_{OC} dataset, is appearing for logBCF as well. Furthermore, the majority of compounds are located in a very narrow subspace of the chemical space with the remaining compounds sparsely located as outliers in the periphery.

Several widespread compounds surround a dense cluster of small molecules (x-axis) with an intermediate degree of branching (y-axis). The dots located on the bottom of the left side encode for larger molecules with a sequential backbone, whereas those located on the top encode for larger, branched molecules. The triangular shape is most likely resulting from the preprocessing of the dataset in terms of model building. Additionally, the disordered agglomeration of the majority of compounds indicates that the principal components provide only limited information about the target property.

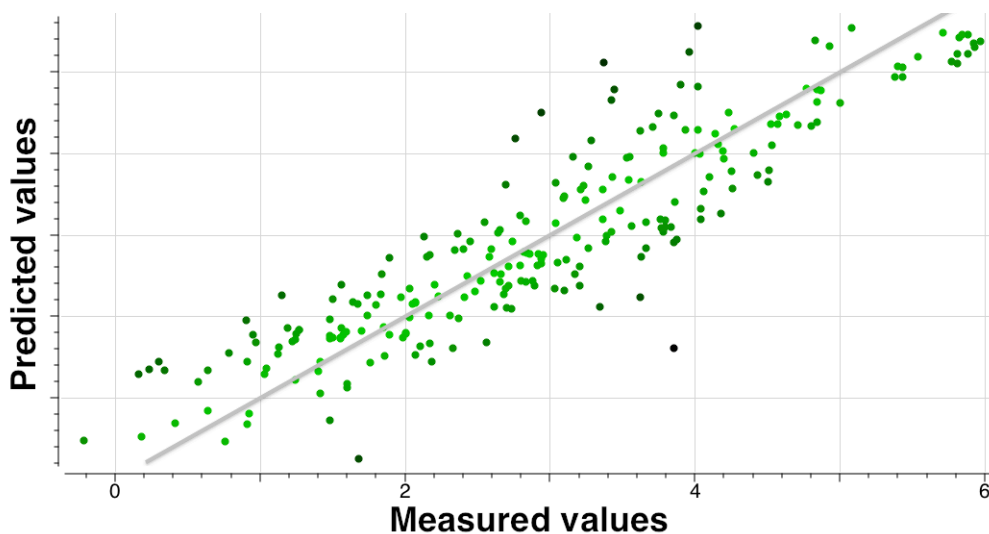


Figure 24. Measured versus predicted values for the logBCF model [log unit].

The first five components covered 28% of the data variance, which is the highest coverage within the datasets used in this study. Furthermore:

- 11 components covered 50% of the data variance
- 25 components covered 80% of the data variance
- 41 components covered 95% of the data variance (34% of the used descriptors)

The performance of the PLS regression shows good results, which was expected, as the dataset was used for a published linear model. Furthermore, the prediction accuracy works equally well for the whole range of considered logBCF values,

which are shown in Fig. 24. A closer look at this figure reveals that the values of the target property are equally distributed over the whole range. This gives an indication that the dataset was undergoing a pre-filtering of compounds. The dispersion of the values of the target property usually follows a normal distribution.

The cross validation performance of $Q^2 > 0.8$ is remarkable as only three latent variables are used for the underlying model. This results from two specifications of the data collection: Firstly the dataset is comparably small, so that only few dependencies can be extracted with statistic reliability; and secondly, which is of even greater importance, the logBCF is mainly correlated with hydrophobicity.

2.1.4 logLC₅₀

“The endpoint for the first toxicity dataset was the log scaled aquatic LC₅₀ value on the fathead minnow. The measurements were taken from the fathead minnow acute toxicity database¹¹⁰ of the Environment Protection Agency (EPA).”[c]

“As the measurements for toxicity are sensitive to laboratory conditions and experimental procedures, we limited the data points within the dataset to this source only.”[b] No structural filters were applied to the collected compounds, implying the dataset contains a widespread structural diversity. All measurements we used in this dataset were produced with test durations of 96 hours.

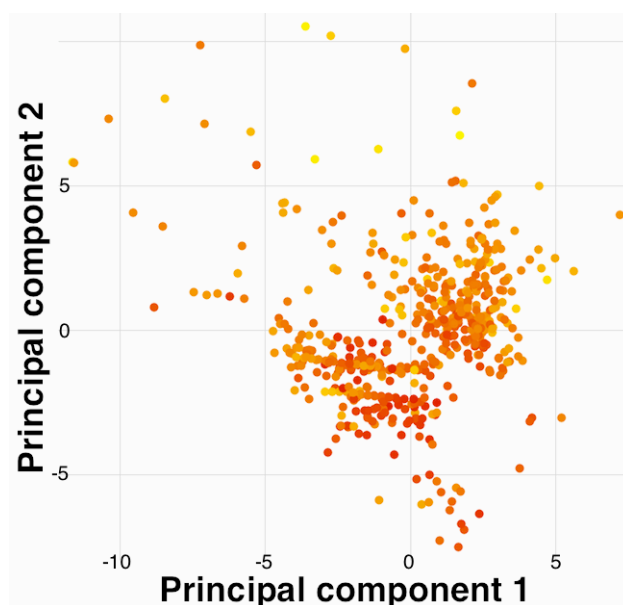


Figure 25. The compound distribution for the logLC₅₀ dataset in PCA space. Yellow dots represent compounds with a high level of toxicity, red dots represent those with a low one.

“The log-scaled lethal concentration against fish is one of the key properties towards aquatic risk assessment and the aquatic toxicity against the fathead minnow as a model organism has been subject to numerous studies^{111,112,113,114,115,116} and published models.^{117,118,119,120} The fathead minnow has been reported as a good model organism by Ankley and Villeneuve,¹²¹

coherently, the variety of models to predict the toxicity against this species available in literature is large. A frequently used source for these models is the EPA's fathead minnow database.¹¹⁰ Measurements of this database are also the informational basis for several QSARs in the JRC's QMRF database.¹⁰¹

Most of the available models work on data collections ranging from approximately 70 to 550 compounds. Their complexity covers a wide range, from a twelve-descriptor model derived with an artificial neural network, developed by the JRC, to a one-descriptor model relying exclusively on logP, developed by Pavan et al.”[d]

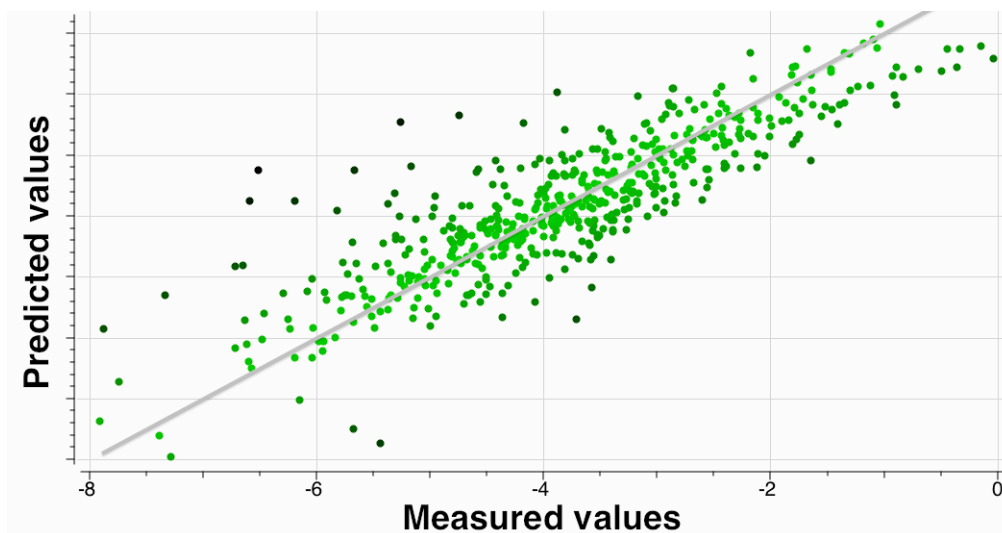


Figure 26. Measured versus predicted values for the logLC₅₀ model [log(mole/L)].

The measurements on this endpoint are characterized by a high variance and uncertainty. This is a consequence of multiple factors: 1) the estimation of the precise concentration that was reached when the fish died is difficult; 2) the endpoint is sensitive to laboratory conditions, in spite of keeping the same measurement protocol; 3) the resistance against a certain toxic compound can vary for more than two log units between two individuals of the same species. With respect to this, the modeling usually concentrates on basic properties, such as molecular weight, chemical groups of known toxic effects and lipophilicity.

An analysis of the five most important principal components revealed that these components encode for: 1) the backbone length; 2) the number of benzene rings; 3) the compound saturation; 4) the atom count; and 5) sulfonic substitutions. The resulting data distribution is shown in Fig. 25. Almost the whole chemical space spanned by the first two principal components is populated. The backbone length shows no influence on the endpoint, but a tendency towards higher toxicity for compounds with benzene rings is observable.

Five components covered 15% of the data variance, and:

- 24 components covered 50% of the data variance
- 60 components covered 80% of the data variance
- 92 components covered 95% of the data variance (52% of the used descriptors)

Regarding the Q^2 from the cross-validation, the performance of the $\log LC_{50}$ model (Fig. 26) shows the worst performance of all regression models. This is expected, due to the measurement uncertainty and as $\log LC_{50}$ is an unspecific endpoint, which is dependent on numerous modes of action. In areas of low toxicity (high $\log LC_{50}$ values) the prediction is quite precise, whereas the number of compounds with high prediction error increases for highly toxic compounds.

2.1.5- $-\log IC_{50}$

“The last regression dataset contained 1093 measurements of toxicity against *T. pyriformis*. The endpoint was the negative, log-scaled inhibition of growth concentration ($-\log IC_{50}$). All measurements in this dataset were taken from our previous study¹²² and originated from Tetratox database¹²³ and several studies of Schulz et al.^{124,125,126} 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.”[e]

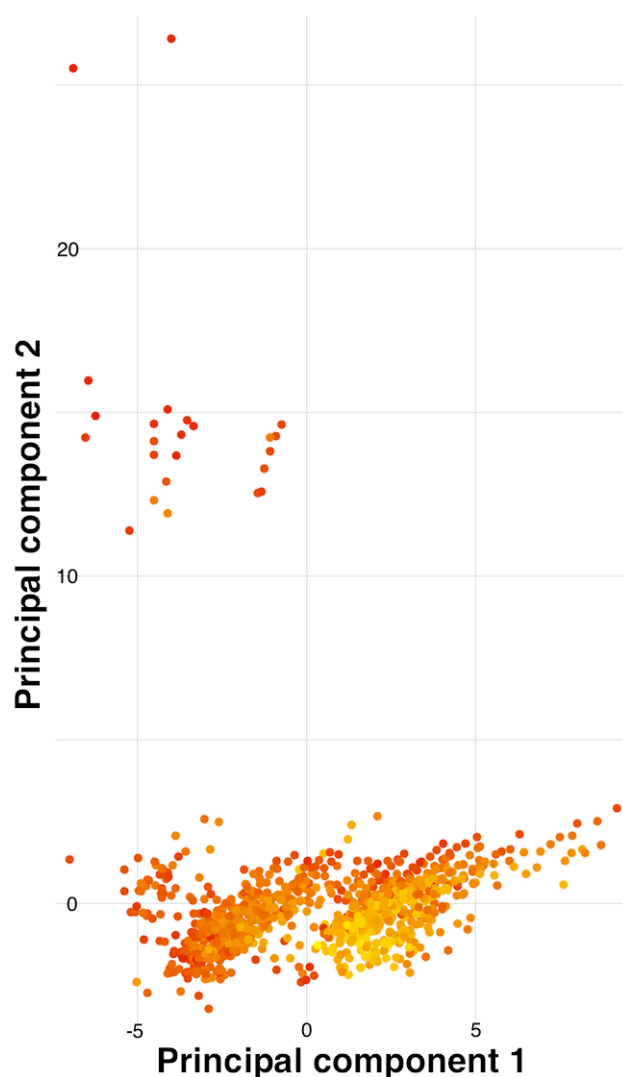


Figure 27. The compound distribution for the $-\log IC_{50}$ dataset in PCA space. Red dots represent compounds with a high level of toxicity, yellow dots represent those with a low one.

The dataset was subjected to the CADASTER toxicity prediction challenge¹²⁷ with more than 100 contributions from 90 participants all over the world. Further worth mentioning, a prediction model submitted by Fabian Buchwald and Stefan Brandmaier reached the seventh best performance and is listed within the first pass winners of the challenge. Although $-\log\text{IGC}_{50}$ is an in-vivo toxicity measurement, such as $\log\text{LC}_{50}$, the measurement uncertainty is clearly lower.

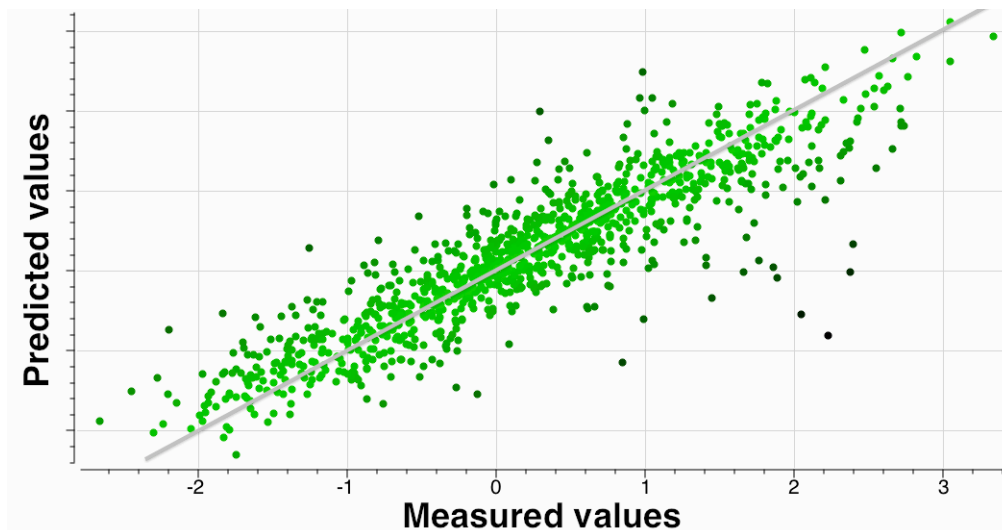


Figure 28. Measured versus predicted values for the $-\log\text{IGC}_{50}$ model $[-\log(\text{mmol/L})]$.

The PCA resulted in the finding that the five most important principal components encode for: 1) the chain length; 2) heterogeneity in the skeletal chain; 3) oxygen content; 4) the substitutions with nitro groups; and 5) the content of amino groups. The distribution of the dataset referring to the first two components is shown in Fig. 27.

The first observation from this depiction is that the compounds are arranged in three layers lying upon one another. The lowermost layer is a large cluster, containing the majority of compounds. These compounds in this cluster are those with a distinct heterogeneity in the skeletal backbone, whereas the loose and clearly smaller cluster above consists of a backbone containing not only carbon. Finally, the two outliers on the top of the data distribution are highly heterogeneous regarding their skeletal backbone. The second observation is a non-linear gradient in the lowermost layer regarding the value of target property. The arrangement is circular with low $-\log\text{IGC}_{50}$ measurements in the center of the cluster and increasing values towards the periphery.

The five most important components covered 15% of the data variance and:

- 26 components covered 50% of the data variance
- 67 components covered 80% of the data variance
- 99 components covered 95% of the data variance (54% of the used descriptors)

Fig. 28 shows the measured values on the x-axis and the predicted values, derived in a cross validation on the y-axis. Although the observed Q^2 is lower than the one for the boiling point dataset, the $\log\text{K}_{\text{OC}}$ dataset and even the $\log\text{BCF}$ dataset, the

model performance is decent, with most compounds predicted well, but some clearly under predicted molecules for high $-\log IC_{50}$ values.

2.2 Classification datasets

As the requirements for a meaningful sample of a classification dataset are most likely to be not identical to those for a regression set, we decided to evaluate the experimental design approaches also on binary classification datasets. The endpoints we use in this thesis consist of an in-vivo test for mutagenicity test and the in-vitro inhibition of an enzyme. Detailed information is given in Table 3.

Table 3. Comparison of the two classification datasets. The analysis is similar to the one presented on the regression sets, but the statistical parameters to evaluate the prediction models are adapted to the needs of classification problems.

	AMES	CYP inhibition
Endpoint	Mutagenicity	Protein inhibition
Measurement uncertainty	Medium	Medium
Compounds	4359	7481
Structural restrictions	No	No
Active/Inactive	2343/2016	3465/4016
Ratio	1.16	0.863
Descriptors	364	428
Distribution	Scattered, multiple outliers	Centered loose periphery
PLS latent variables	7	8
Variance covered	27.4%	35.3%
Accuracy	0.7857	0.8245
Balanced accuracy	0.7855	0.8248
F-Measure	0.7983	0.8137

Still, as the REACH legislation mostly requires information on continuous endpoints and as this thesis therefore focuses on regression datasets, the analysis of the classification datasets will not be as exhaustive. Hereby it is also important to keep in consideration that classification assignments in computational chemistry modeling are mostly derived by the discretization of continuous values.

2.2.1 AMES mutagenicity

The AMES test, as it was applied to produce the measurements used in this study, detects the mutagenic effect of a small chemical compound to a histidine-dependent strain of *Salmonella typhimurium*, which is a gram-negative bacteria. The exposure to a mutagen is expected to restore the original ability to synthesize histidine. Such a mutagenic effect would be observable in a growth of bacterial colonies on a medium deficient in histidine. The measurable mutagenic ability of a compound provides an indication about its potential carcinogenicity.¹²⁸

The dataset we use contained 4359 compounds, whereby 2343 of them were determined to be active, 2016 to be inactive. This dataset was derived from a bigger dataset described in one of our previous studies.¹²⁹ The compounds we used in this study have been limited to those used in the training set of Sushko et al.'s study on the applicability domain of classification models.¹³⁰

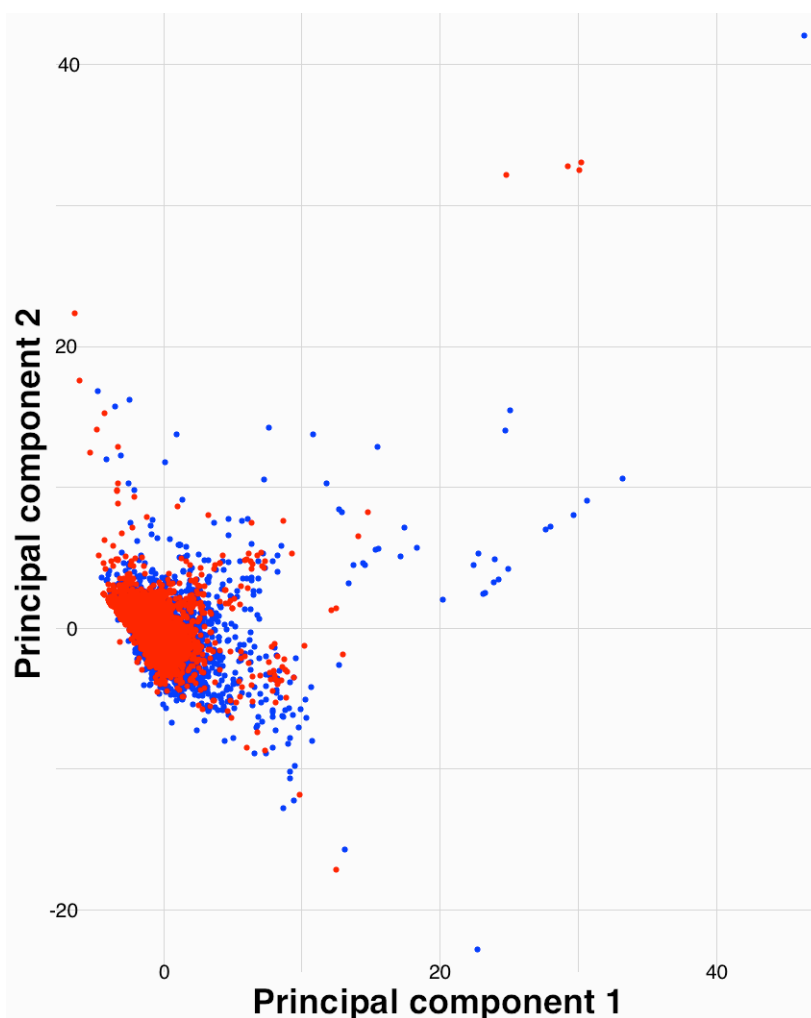


Figure 29. The compound distribution for the AMES dataset. The red dots represent compounds with mutagenic ability, whereas the blue dots represent all other compounds.

The distribution of the compounds in the dataset is shown in Fig. 29. The two axes represent the two main principle components and each dot represents a compound. The colors of the dots discriminate mutagenic active (red) from inactive (blue) compounds. For both classes the majority of compounds are located in a very narrow subspace of the chemical space. The two principle components with the highest variance seem inappropriate for the differentiation of the two classes. Beyond the center of data cloud, there is a small partition of sparsely distributed compounds, which also do not show any correlations to the principal components.

Similar to the regression datasets, the method to build a classification model for the AMES dataset was PLS. The cross validation statistics showed an accuracy value as well as a balanced accuracy of 79%. The results are illustrated in Fig. 30.

The color of the dots corresponds to the measured class, whereas the color of the background corresponds to the predicted class. The majority of active, as well as of inactive compounds, is predicted correctly. The number of false negatives (red dots on blue background) and false positives (blue dots on red background) is minor.

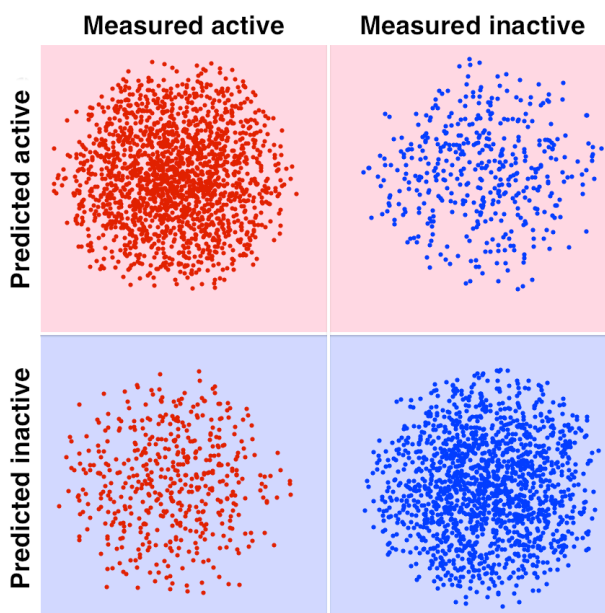


Figure 30. Results of the classification model for the AMES dataset. The background color refers to the predicted class, the dot color to the measured class. Each dot represents a compound in the dataset. The majority of compounds are predicted correctly.

2.2.2 Cytochrome P450 inhibition

“The second classification set we collected contained 7481 measurements of human cytochrome 1A2 inhibition activity of small molecules, which were taken from the bioassay AID410 in the PubChem database. The assay was deposited in October 2007 and the dataset was used in a previous study on comparative modeling of cytochrome inhibition.¹³¹ The original dataset obtained from this bioassay contained 8348 compounds.”[e]

“Compounds that were labeled ‘inconclusive’ were excluded from the dataset. Further, if the same molecule was present in both the ‘active’ and the ‘inactive’ set, it was removed from all sets. The final distribution of the remaining 7481 compounds was almost balanced, as 4016 were labeled ‘active’ and 3465 ‘inactive’.”[e]

The distribution of the majority of compounds is more widespread, compared to the distribution of the compounds in the AMES dataset. Furthermore the structurally diverse compounds beyond the center of distribution are actively measured active in a large part. Although the graphical representation, shown in Fig. 31, indicates a slight shift of inactive compounds towards lower values of the

first principal components, the majority of compounds within the two classes are overlapping.

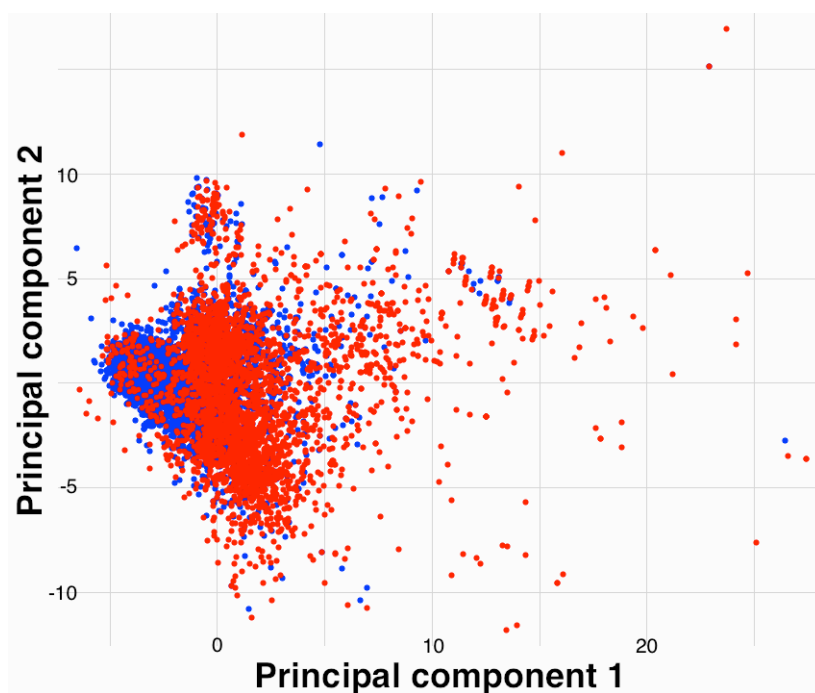


Figure 31. The compound distribution for the CYP inhibition dataset. The red dots represent inhibiting compounds, whereas the blue dots represent non-inhibiting compounds.

The classification results (shown in Fig. 32) derived on the CYP inhibition set are comparable to those derived on the AMES dataset. The statistical parameters for accuracy and balanced accuracy are increased by three percent.

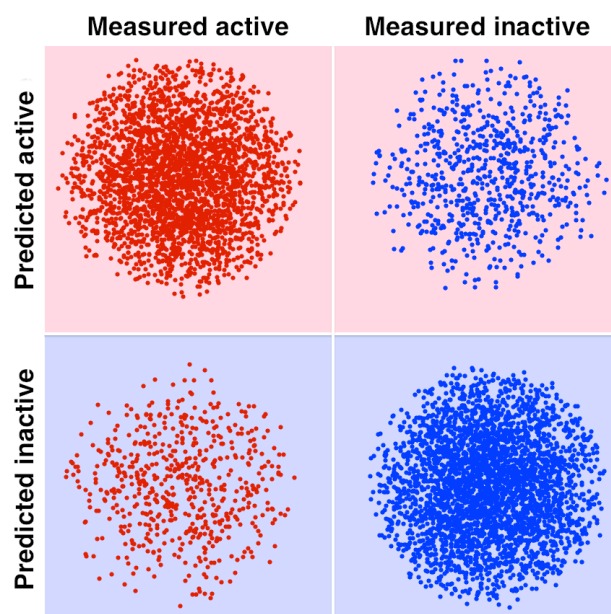


Figure 32. Results of the classification model for the CYP inhibition dataset. The background color refers to the predicted class, the dot color to the measured class. Each dot represents a compound in the dataset. The majority of compounds are predicted correctly.

2.3 Structural outliers

“One of the aims of this study was to investigate the influence of structurally diverse compounds on the selection and accuracy of the resulting models. Therefore each of the three datasets was extended by the inclusion of a compound, which was characterized as a structural disrupter. We defined a structural disrupter as a data point that (a) influences the recalculated loadings of the first or the second principal component in such a manner that the principal properties represented by these components are changed and (b) results in one or more instances in the data set that are – according to the distribution of the instances in that principal component – at least five standard deviations from 97% of all other compounds.

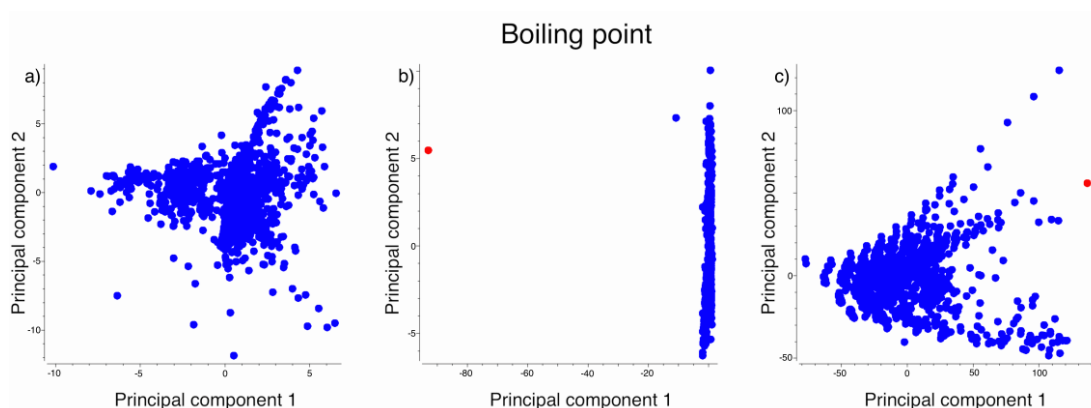


Figure 33. The change in the principal components view due to one structural outlier in the dataset. The principal components were calculated for the dataset with (b, c) and without (a) structural outlier. ALOGPS and E-State indices were used (a, b), as well as DRAGON descriptors (c). The protocol to calculate the principal components was always the same. [c]

Structural outliers like the ones used in this study are not artificial, but can result from several reasons, e.g. (a) from few compounds within the dataset, which have a specific chemical group that is different from other compounds and functionally is not relevant, (b) from the choice of a specific descriptor set, or (c) from a certain procedure within the multivariate analysis (centering or not the data, usage of raw, normalized or standardized data).

The structural outliers in our study were (a) ethyl 2-chloro-3-[2-chloro-5-[4-(difluoromethyl)-3-methyl-5-oxo-1,2,4-triazol-1-yl]-4-fluorophenyl]propanoate (carfentrazone-ethyl) for the boiling point dataset, (b) (1R,4aR,4bS,7S,10aR)-7-ethenyl-1,4a,7-trimethyl-3,4,4b,5,6,8,10,10a-octahydro-2H-phenanthrene-1-carboxylic acid (isopimaric acid) for the logLC₅₀ dataset and (c) (1,2-dimethyl-3,5-diphenyl-pyrazol-1-yl) methyl sulfate for the logK_{OC} dataset. All these three compounds were retrieved from the same source as the rest of the respective dataset. Fig. 33a shows the first two principal components of the boiling point dataset without outliers whereas Fig. 33b shows the first principal components of the same dataset with the structural disrupter. The structural disrupter has a red color. The principal components were derived from the whole set of normalized ALOGPS descriptors and E-State indices and thus no variable selection was performed. Furthermore, the data were not centered before the orthogonal transformation.

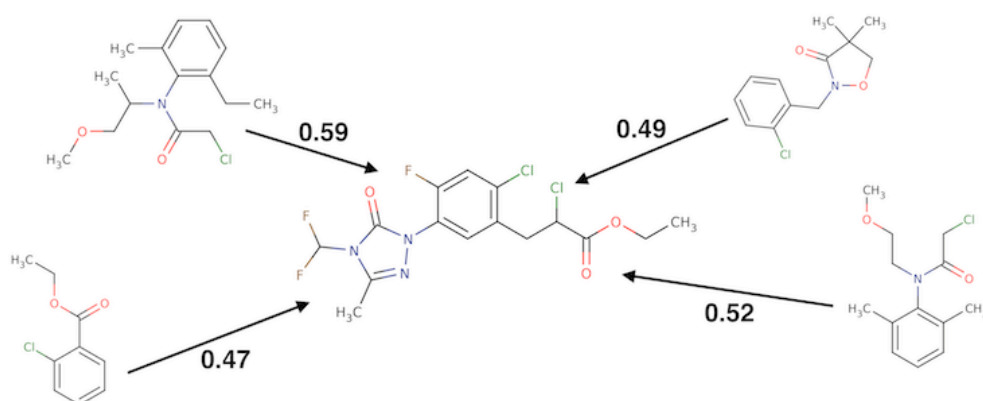


Figure 34. The structural outlier and similar compounds in the dataset. The most significant unique structural feature of the outlier is a triazole group. [c]

To show how the concerted outlier for boiling point structurally fits into the dataset, we calculated its Tanimoto distance to all other compounds. ISIDA fragments³¹ were used therefore. Fig. 34 shows the outlier in the center and the four most similar compounds around. The value assigned to the edges indicates the similarity score. It is obvious that the outlier is a larger molecule and contains a triazole group, which is absent in other compounds. Such types of outliers could naturally happen to be present in the datasets. The appearance of such outliers depends on the used descriptors. Fig. 33c shows, if DRAGON descriptors¹³² are used, this compound is not anymore an outlier (although it is located at the periphery of the data cloud). Indeed, Dragon software calculates many more descriptors and in their space the analyzed molecule does not have descriptors, which make it to be the outlying point in the PCA space. Thus, a property of a molecule to be a structural outlier depends on the used set of descriptors, i.e. on the representation of the molecule.”[c]

3 Methods

The first part of this chapter is intended to enable an estimation of the applicability of QSAR in terms of model building with respect to the REACH regulation and to deepen the ability to assess the peculiarities and pitfalls attached to statistical modeling in general and to computational chemistry in particular. A critical discussion about the current status of and conflicts within QSAR modeling is followed by a case study, in which I step wisely demonstrate the line of action that is required to build a valid QSAR model, which is reliable in terms of its predictive ability and explanatory in terms of the underlying mechanistic.

The subsequent parts of this chapter provide an overview about the techniques that are examined regarding their contribution to experimental design. Starting with a short overview of established methods that were implemented for this thesis and special customizations and parameterizations they have been subject to, I proceed with a detailed description of the conceptual innovations, this work focuses on.

Different to the approaches presented in the introduction of this thesis, all following methods were newly developed for experimental design in QSAR modeling. Furthermore, apart from a detailed description of all methods, a deeper insight into the underlying ideas of the newly developed concepts is given. I will explain the principle of stepwise, adaptive approaches and the expected effects on the resulting selection and I will outline the idea of a cluster based experimental design.

Finally, the last section of this chapter describes the validation procedure we adapted to enable a representative, statistically valid estimation of the performance derived with the used approaches.

3.1 QSAR modeling with respect to the REACH legislation

This chapter is intended to discuss the difficulties attached to statistical modeling in general and the proposed solutions referring to QSAR modeling in particular. A reliable estimation of the benefits of an approach requires information about the respective limitations and peculiarities as well. As a matter of fact, the predictions of a statistical model on large amounts of data work well to provide a representative overview of the general predictive ability of the model, but still the quality and reliability of predictions for single instances may vary to a large extent.

It is therefore of crucial importance to carve out the process of model development with all required steps, such as the collection of sound data, the identification and analysis of observable correlations, the validation procedure, the definition of an applicability domain and the identification of resulting benefits, when compared to prior knowledge referring to the same question.

3.1.1 Constraints in QSAR modeling

Although efforts to correlate molecule structures with chemical properties are carried out since more than 100 years, QSAR and computational chemistry are scientific disciplines with a disunity regarding general questions. Furthermore to enable a proper evaluation of the discussed methods, it is of concern to be aware of these issues. Statistical approaches to prioritize chemical compounds are well established in the field of pharmaceutical research, but they are critically eyeballed in for regulatory purposes. Partially this is based on the fact that statistical methods are prone to systematic errors. It is therefore required to highlight these errors.

3.1.1.1 Area of conflicts

Referring to several basic questions there are opposed directions in research within the scientific community. The most relevant of these questions will be briefly introduced in the following paragraphs.

3.1.1.1.1 Interpretability versus performance

One of the most frequently discussed issues in QSAR and computational chemistry is whether the crucial requirement towards a computational model is its predictive ability or its descriptive character. Whereas statistical scientists tend to prefer the version of a reliable statistical tool with the major task to predict well, chemists tend to distrust models without an explicative mechanistic interpretation.

Naturally both qualities are of high concern and the optimal model should deliver the best possible predictions lined with a proper chemical reasoning of the observed effects. But such models are exceptional. In most instances modeling approaches, which are of a higher complexity, result in a higher predictive accuracy. But modeling methods of higher complexity, such as for example an SVM

with a non-linear kernel, or neural networks elude from a simple and straightforward interpretation, as the resulting regression formula usually does not allow a simple weighting of the contribution of different descriptors.

Whereas linear regression approaches enable a simple access to the underlying mechanistic principles, the predictive performance of resulting models is often not optimal. Nature is more complex and in particular biological properties often cannot be expressed with only three or four variables, as they are preferred in linear models. To compensate for this, the models are often trained to match available observations, which results in good predictions on the training set and a decrease in statistical reliability.

On the contrary, machine-learning approaches of higher complexity and with comparably better predictive performance, such as SVMs or ANNs deliver mostly blackbox models, which prevent an interpretation. They produce either incomprehensible, highly complex equations or none at all.

3.1.1.1.2 Local versus global models

A further critical issue is the question whether to prefer the development of global models, which are trained on a variety of chemical classes and intended to deliver reliable predictions for widespread areas of the chemical space, or to rely on local models, which are restricted to a comparably narrow sector of the chemical space.

On the one hand, every constraint to the use of available data increases the risk to rely on chance correlations, on the other hand, due to a lack of experimental data, the calculation of a statistical significance is often impossible. In most cases chemical dependencies can be linearly approximated, at least on a local model. Especially for a constricted variation in the underlying structures the convergence is sufficient, which supports the use of local models in terms of the interpretability. But these models often lose track of the global context. Nevertheless, a study by Puzyn et al.¹³³ on aquatic solubility resulted in the conclusion that the improvement reached with local models is negligible small and statistically not significant.

3.1.1.1.3 Statistical modeling in general

The arising question, why not to switch from a statistical modeling approach with a subsequent mechanistic interpretation to a direct mechanistic modeling, can be answered by the complexity of the chemical space. Contrary to many biological macromolecules, such as amino acids or DNA, small chemical compounds are not sequential, but branched. Therefore the structural variability is clearly higher so that neither current methodologies, nor the available experimental data enable the application of such approaches.

Furthermore, the urgent problem to identify persistent organic pollutants (POPs) and curtail their dispersion requires efficient solutions.¹³⁴ Pesticides are detectable in human milk,^{135,136} per- and polyfluorinated compounds are highly

persistent^{137,138} and their presence can be observed all over the Atlantic Ocean.¹³⁹ The in-utero exposure to low concentrations of polybrominated diphenyl ethers have been shown to cause severe effects on motor activity and male fertility.¹⁴⁰ The precocious recognition of the dispersion, persistence and toxicity of such compounds is highly important.

To address the problem of individual unreliable predictions, numerous studies focus on the question, how to distinguish reliable predictions from those, which are unreliable.^{141,142,143,144} This problem is generally referred to as the definition of an applicability domain.

The suggested approaches to estimate the domain of applicability, or in other words, if a model provides a reliable prediction for a certain compounds, are of different assumptions. There are similarity based approaches,^{145,146} as well as prediction based ones,^{130,147} but as a general rule of thumb, the more a compound differs to the molecules the model was trained on, the higher is the uncertainty in prediction.¹⁴⁸

3.1.1.2 Regulatory restrictions

To bring these areas of conflict to a balanced consensus and to establish general guidelines for model development, the member countries of the Organisation for Economic Co-operation and Development (OECD) released a list of five principles for QSAR modeling. These principles are referred to as the 'OECD' principles.¹⁴⁹

3.1.1.2.1 OECD principles and QMRF

The main motivation behind the OECD principles is to ensure a basic quality standard for predictive models, which are intended to support regulatory purposes. The defined requirements are:

1. a defined endpoint
2. an unambiguous algorithm
3. a defined domain of applicability
4. appropriate measures of goodness-of-fit, robustness and predictivity
5. a mechanistic interpretation, if possible

To help scientist to respect these principles and to simplify the evaluation QSAR models in terms of regulatory purposes, the QSAR modeling report format (QMRF) was established. The QMRF is xml based and enforces information about used descriptors, the underlying selection process, a model equation, the description of a validation set, as well as the conformation to a predefined validation protocol.

All this required information should furthermore contribute to enable a fast and simple reproduction of QSAR models. However, the increased use of 3D descriptors in recent years has shown a new difficulty arose, regarding the model transparency. The structural optimization of molecules is in general not

deterministic, so that also the descriptor values resulting from these optimized structures vary.

Further issues arising from the five principles are, e.g. the question if a defined endpoint can comprise more than one species, or if the validation of a model requires an external test set. Although the OECD principles do not state a specification for acceptable or not acceptable machine learning algorithms, for the most part, models that are used in the practical regulatory work are linear, limited to one specific species and validated on an external test set.

3.1.1.2.2 Consequences

The studies presented in this thesis were initiated within and financed by the CADASTER project. CADASTER aims to provide a risk assessment framework that is exemplary within REACH.

With respect to this, we only use linear models (PLS). PLS models enable interpretability, which is not directly derivable from descriptors, but principal properties. Furthermore, with respect to the complexity of the biological endpoints, relevant for REACH, robust modeling approaches seem favorable. It has to be taken into consideration, that the measurements of most biological endpoints are derived with in-vivo experiments, which causes a high uncertainty. Furthermore, properties, such as the lethal concentration are unspecific. The underlying mechanism, why a certain compound is toxic can be versatile. Hundreds, if not thousands of modes of action are possible.

3.1.1.3 Pitfalls in QSAR modeling

Finally a critical look on QSAR additionally requires the consideration of pitfalls attached to any kind of statistical modeling.¹⁵⁰ Numerous published models lack of statistical reliability or of an established explanation of the underlying modes of action. Although we are aware of examples for each of the following pitfalls, for obvious reasons, we disclaim references in this paragraph.

3.1.1.3.1 Overfitting

The most frequent problem in statistical modeling is overfitting. A model, which is mainly trained to fit the underlying measurements, but without predictive ability for new instances is referred to as overfitted. The sources of overfitting can be numerous.

- **Descriptor selection on the whole dataset**

The statistically questionable *modus operandi*, which is most frequently not even recognized as a severe source of error, is an improper validation procedure for the descriptor selection. A supervised descriptor selection is often interpreted as a part of the preprocessing of the data. Actually, it is a step within the modeling procedure. Therefore it is inappropriate to use the

validation results, retained within the optimization process as a reference for the goodness of fit. An independent external validation or a double cross validation has to be performed.

- **Use of too many descriptors**

The variety of available descriptors is huge. The DRAGON 6.0 descriptor package offers a selection of almost 5.000 variables to describe a chemical compound. Furthermore, packages, such as ISIDA, which are mining the compounds for available substructures can easily exceed the number of 10.000 variables, depending on the concerned length of the substructures, even for small datasets. The smaller the dataset is and the larger the number of available descriptors is, the more likely it is to find a correlation that is just caused by chance. Although several established criteria, such as the Akaike information criterion,^{151,152} the Bayesian information criterion,¹⁵³ or the Hannan-Quinn information criterion¹⁵⁴ can help to prevent from using an exceeding number of variables, there is no guarantee that the correlation of a certain variables to the target property is mechanistically reasoned or just the side effect of a non representative or too small sample of the chemical space.

- **Measurement uncertainty is not taken into consideration**

Especially in case of biological endpoints, with measurements derived from in-vivo experiments, the measurement uncertainty has to be taken into consideration. For example measurements on bioaccumulation, for the same compound and for the same species in average vary at least one log unit. Furthermore, variations by two log units are frequent. Therefore, a model predicting bioaccumulation with a reported average error of 0.5 log units is most likely trained too much to match the available observations.

3.1.1.3.2 Nonsense interpretations

Apart from statistically unreliable models, even for properly validated models the interpretation in terms of underlying mechanisms is a frequent source of errors or sloppiness. To be sure, model analyses lacking from a well-founded interpretation do not influence the quality of the underlying prediction approach or the model itself. Still, they might cause systematical error in the generation process of new models as such interpretations might be used to define a precondition for a descriptor selection.

- **Correlation is not causality**

The most common error in the interpretation of statistical models is the confusion of correlation and causality. A simple example for such a kind of nonsense interpretation is the development of the mosquito population around Finnish lakes. Whereas the number of mosquitos drastically increases from June to September, their number drops to zero from December to February. The same development is observable for tourists visiting the Finnish lakes.

This is a correlation, but not a dependency. The idea to increase the appearance of tourist by importing mosquitos is therefore not too promising. Neither do the tourists visit the lakes because of the mosquitos, nor do the

mosquitos come to watch the tourists. The underlying reason is the weather, or more precisely, the temperature.

- **Face-value interpretations**

Another way to shoot a serious interpretation down, is accepting descriptors at face value, without questioning the principle property they represent. If a model for the boiling point of organic compounds mostly relies on a descriptor, which displays the number of carbon atoms in a compound, the effective reference to the concerned endpoint is not the one to the descriptor itself. The descriptor represents a latent variable, which is most likely to be the molecule size. Such over-interpretations could be easily identified by an analysis of further descriptors with a comparable correlation to the target property and a high correlation to the used descriptor.

3.1.1.3.3 Intentional cheating

Finally, apart from the impure use of statistical methods, published QSAR models may lack of trustworthiness, as they use oblique procedures to 'prettify' the statistical results derived with a certain model.

- **Dataset pruning**

A frequently observable habit is the exclusion of available data. The usual practice is to extract i.e. 50 compounds of a certain chemical class from a dataset, whereas the underlying dataset provides 65 of such chemicals. The disregarded 15 compounds are not even mentioned in the publication, but the application of the model to those compounds reveals, that the observed correlation does not fit to those compounds.

To be sure, the exclusion of available measurements from a dataset is not axiomatically wrong. But there should be good reasons for disregarding individual compounds and. Such reasons might be a high structural diversity to all other compounds within a dataset or a highly deviant molecular weight. Furthermore, putatively erroneous measurements have to be disregarded within the modeling process, but the reasoning to exclude a compound from a dataset should not be that the compound did not fit into the model.

- **Biased splits between validation set and test set**

Another way to manipulate statistical values is to use a non-representative split between a training and validation set. By a directed selection of only well fitted compounds for the test set, the statistics derived from the external validation appear better than they really are. A way to detect such biased splits is by a repeated random splitting of the whole underlying data and to recalculate the statistics to get a representative overview about expectable statistics.

3.1.2 Modeling aquatic toxicity

In order to constitute and critically assess the facilities and restrictions connected to statistical methods and in order to exemplify the application of techniques to implement the requirements of REACH and the OECD principles, I will present a feasible procedure for the development of QSAR models with a demonstration of all required steps.

The following case study presents “an easily interpretable mechanistic and generalized model for aquatic toxicity against fish, which is not limited to only one species. We collected 1358 measurements in total for LC₅₀ against different fish, e.g. *Pimephales promelas*, *Oncorhynchus mykiss*, *Danio rerio* and others from a broad literature research.

We calculated a collection of 7595 descriptors for the underlying compounds and applied a genetic algorithm to select a fixed number of three descriptors. Those descriptors were subject to a linear regression to derive a model, which is explaining aquatic toxicity as a combination of hydrophobicity, mass autocorrelation and the presence of alkyl or cyanide groups.

We compared the performance of our model to the results derived with more complex approaches on the same dataset and show that the three-descriptor solution, we propose, is sufficiently good. Further, to assess the performance of our model in terms of regulatory purposes, we compared it to the quantitative read-across model published by Schüürmann et al.¹⁵⁵[d]

3.1.2.1 Data collection

“From a literature review on numerous publications,^{110,114,156,157} online databases such as the PPDB¹⁵⁸ and the JRCs QMRF database¹⁰¹ we collected 2391 toxicity measurements in total on several species of fish. The endpoint of all these measurements was the lethal concentration for 50% of the population and the test duration was 96 hours. We excluded measurements on inorganic compounds, radicals, charged molecules and removed compounds for which no exact values, rather an interval or only minimum or maximum values, were given. Additionally, all measurements from the PPDB, which were assigned with the quality code ‘1’ (lowest rating), were refused. All remaining measured values were then transformed to log-scaled concentration values (mole/L).

In the next step, we used the *Chemaxon standardizer* to standardize and neutralize the remaining compounds, as well as to remove salts. The Hamiltonian AM1 algorithm,¹⁵⁹ implemented in MOPAC 7.1³² was used to structurally optimize the compounds. The ration to choose the AM1 algorithm for optimization was that the derived atomic charges and LUMO energies were found to correlate well with several properties.¹⁶⁰ Measurements on compounds for which the structural optimization or the following descriptor calculation failed were excluded.

Finally, to manage duplicate measurements on the same compounds, we implemented a strict decision pipeline. It allowed to better understand, which measurements should be retained and which ones to be discarded. We ranked the priority of a measurement thereby as follows:

1. Measurements from the fathead minnow database
2. Measurements on the fathead minnow
3. Measurements from the PPDB
4. Other measurements

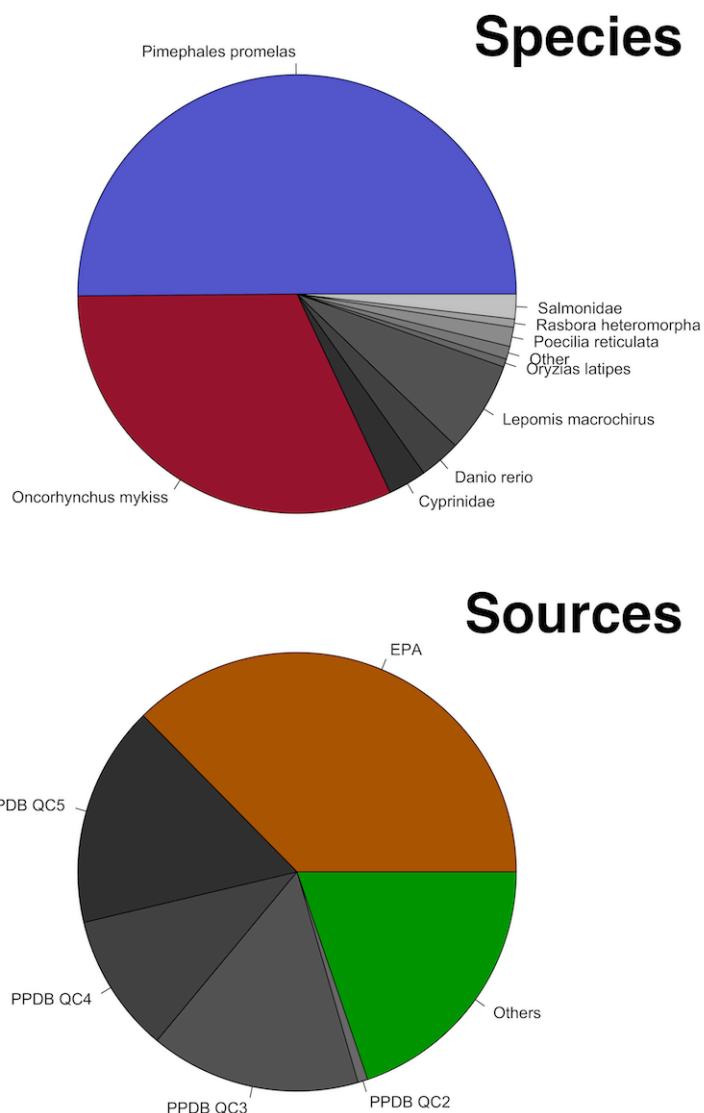


Figure 35. Data sources and species in the resulting dataset. Regarding the underlying species, half of the measurements were derived on the fathead minnow and one third on the rainbow trout.

If the pool of duplicates consisted of several measurements with the same priority, the decision was as follows: if three or more measurements on the same compound were available, we decided for the median; if only two duplicates were available and if their values did not differ by more than log one unit, one of the

measurements was arbitrarily refused; in all other cases, both measurements were removed from the set.

After all these filtering steps, we retrieved a dataset containing 1358 measurements for unique compounds. Approximately half of these measurements (660) were taken on *Pimephales promelas*, one third on *Oncorhynchus mykiss* (450) and the rest on different species (*Danio rerio*, *Poecilia reticulata*, *Cyprinidae*, *Percoidei*, etc.). Regarding the source of data, most values were taken from the PPDB (617) and from the fathead minnow toxicity database (553). The remaining compounds were collected from a variety of publications and QMRFs. Fig. 35 visualizes these distributions.”[d]

3.1.2.2 Descriptors selection

“For the storage, management, organization and filtering of the measurements as well as the calculation of an initial pool of descriptors to work with, we used the Online CHEMical database and Modeling environment (OCHEM).⁸⁶ Following sets of descriptors were calculated for all the compounds:

- ALogPS⁸²
- ISIDA fragments³¹
- MERA, MERSY^{161,162}
- E-State indices^{84,85}
- DRAGON 6.0¹³²
- Chemaxon¹⁶³
- CDK¹⁶⁴
- MOPAC¹⁶⁵
- Inductive¹⁶⁶
- Spectrophores^{167,168}

The resulting number of descriptors was 7595. This collection was intended to cover a broad spectrum of chemical representation, as it contains fragment-based descriptors (ISIDA, E-States), topological 2D and 3D descriptors (DRAGON), quantum chemical descriptors (MOPAC, MERA, MERSY), hydrophobicity (ALogPS) and others.

To prevent from over-fitting, and to remove descriptors with only a low or by chance correlation, we used a leave-one-out cross-validation to calculate the correlation coefficient (R) of each descriptor to the target property and excluded those with $R < 0.1$. A genetic algorithm was customized to select a prefixed number of descriptors in a double cross validation. The aim of this study was to deliver a model with a high degree of simplicity in terms of interpretability and to avoid over-fitting our model to the measurement uncertainty, resulting from various factors (differences in laboratory techniques, individuals, species, etc.). Therefore we selected a fixed number of three descriptors.

In each generation of the genetic algorithm 250 offspring were produced, 30 of them assigned to be the survivors for the next generation. The offspring inherited one descriptor from the first parent and the second descriptor from the other

parent. The third descriptor was chosen randomly from the remaining pool and the number of generations was 25. The criterion to select the ‘fittest’ combination of descriptors was the correlation coefficient derived by a multiple linear regression in a ten-fold internal cross validation.”[d]

3.1.2.3 Regression and validation

“To build the final model, we used a MLR analysis on the descriptors selected by the genetic algorithm. The validation - using the root mean square error (RMSE), the correlation coefficient and the Q^2 referring to Schüürmann et al.¹⁶⁹ as measurements of quality - consisted of two parts: firstly, a leave one out cross validation on the whole dataset (CV_{Loo}); and secondly a statistical evaluation by random splitting. We therefore generated 1000 splits on the dataset, each containing 62% of the compounds in the training set and 38% for the validation set, which is corresponding to a theoretical split of $(1-e^{-1})$. The result of the external validation is therefore an average Q^2 value as well as the corresponding standard deviation, which can be interpreted as a confidence interval. We explicitly refuse to select one particular split as ‘the’ validation set, as we believe, that the use of an average value instead gives a better insight to the reliability of the model.”[d]

3.1.2.4 Model calculation

“The three descriptors derived applying the genetic algorithm were Se1C2C3ts, ALogPS_logP and ATSm2. The first descriptor, Se1C2C3ts, is an E-State index for single bonds between two carbon atoms, one of them with two skeletal bonds (one of these bonds is a triple bond), the other one with three skeletal bonds. ALogPS_logP corresponds to the hydrophobicity of a compound calculated with the ALOGPS 2.1 program. ATSm2 is derived from the CDK package and it refers to the Moreau-Broto autocorrelation of a topological structure. It specifically describes the distribution of the mass for topological distances of length 2.

Table 4. Pairwise correlation between the selected descriptors and the target property.

	Se1C2C3ts	ALogPS_logP	ATSm2	logLC ₅₀
Se1C2C3ts	-	11.0%	7.9%	22.4%
ALogPS_logP	11.0%	-	50.8%	74.0%
ATSm2	7.9%	50.8%	-	63.5%
logLC ₅₀	22.4%	74.0%	63.5%	-

Tab. 4 shows the (absolute) pairwise correlation between these three descriptors and the target property. The correlation coefficient of hydrophobicity to logLC₅₀ is thereby the highest. It covers almost three fourth of the variance within the dataset. The correlation of the target property to ATSm2 is slightly lower, but still covering approximately two third of the variance. The inter-correlation of these two descriptors is around 50%, which is indicating that – referring to the target property - they are at least partially independent. Finally, the E-State index has a comparably lower correlation to the target property, which is logical, as it

describes only one specific bond type but shows high orthogonality with the other descriptors.

The formula resulting from the linear regression on these descriptors is:

$$\log LC_{50} = -(2.21 + 0.945 * Se1C2C3ts + 0.473 * ALogPS_logP + 0.0465 * ATSm2)$$

The corresponding model validation statistics are shown in Tab. 5. The last column contains the average values from the 1000-fold external validation on random splits. The according standard deviation is indicated in brackets.

Table 5. Statistics derived from the model validation.

	Internal	CV_{Loo}	External bagging
Correlation coefficient	0.809	0.808	0.81 (0.01)
RMSE	0.847	0.850	0.85 (0.02)
Q²	-	0.653	0.65 (0.02)

The internal validation (applying the model trained on all instances) results in a correlation of 80.9% with an RMSE of 0.847. The statistics retrieved from the leave one out validation and the external validation, using a 1000-fold split do not significantly vary from the results derived with the internal validation.”[d]

3.1.2.5 Evaluation of the model performance

“The achieved performance of $Q^2 = 0.65$ is surely not comparable to that of local models or models for an endpoint of lower complexity; however, to achieve a performance like that was not the intention of this study. Several aspects have to be taken into consideration: first, we did not apply any structural restriction to the compounds in the dataset; second, the data was collected from various sources; and finally, we used a variety of species in this study.

Table 6. Q₂ values derived with other methods.

	E-States + ALogPS	MERA + MerSy	CDK	Chemaxon	ISIDA
kNN	0.61	0.57	0.64	0.61	0.43
FSMLR	0.63	0.57	0.0	0.62	0.67
PLS	0.66	0.52	0.64	0.63	0.64
ASNN	0.70	0.66	0.71	0.70	0.70
SMOreg *	0.70	0.65	0.71	0.69	0.62
M5P *	0.68	0.59	0.67	0.66	0.65

To estimate the explanatory power of our three-descriptor model, we used OCHEM to apply a variety of combinations of different descriptor sets and machine learning methods to the collected dataset. The results derived from a leave-one-out cross-validation are shown in Tab. 6. Models marked with (*) were calculated with WEKA.¹⁷⁰ The descriptors were filtering to eliminate redundant ones with a pairwise correlation coefficient larger than 0.95. The parameterization of the machine

learning methods were default ones, except for the support vector regression (SMOreg), which used a normalized, quadratic poly-kernel.

The only linear model that reached a better performance than our models was a fast stage-wise multiple linear regression (FSMLR) model built on 18 out of 660 ISIDA fragments and a PLS regression model on ALogPS descriptors and E-State indices. The PLS model was calculated with nine latent variables derived from 212 descriptors. In any other case, the linear approaches could not improve the results of our regression, neither for PLS regression, nor for the FSMLR. Overall, the k-Nearest-Neighbor approach (kNN) showed the weakest performance, while the associative neural network (ASNN) delivered the best results. The performance of a support vector machine and a tree learner (M5P) were comparable.

In summary, all the models that delivered a better performance than our regression on three selected descriptors had a clearly higher complexity: firstly, the ASNN, M5P and SMOreg models used correlations of higher order, whereas our model only used the linear dependencies; secondly, the PLS model and the FSMLR model, as well as the ASNN, M5P and SMOreg models, were built on a multiple number of descriptors, compared to our model with only three descriptors.

Table 7. Model performance for different number of selected descriptors.

Number of variables	RMSE	AIC
1	0.94	0.89
2	0.87	0.77
3	0.84	0.71
4	0.83	0.70
5	0.81	0.66
6	0.80	0.65

Additionally, we used the genetic algorithm to select a fixed number of 1, 2, 4, 5 and 6 descriptors and applied a linear regression to them. Tab. 7 contains the number of selected descriptors, the RMSE derived in the double cross-validation and the corresponding Akaike Information Criterion (AIC). Our results indicated that the improvement in the RMSE derived with more than two descriptors was minimal. The decrease in the average error was 0.07 log units for two instead of one descriptor and additional 0.03 log units for using a third descriptor. In contrast, the improvement derived with a number of four or more descriptors, is small. Compared to the performance with three descriptors, a six-descriptor model decreased the error only 0.04 log units.”[d]

3.1.2.6 Comparison with an established model of regulatory relevance

“To investigate the performance of our regression model in comparison to a model of regulatory relevance, we applied the read-across model for aquatic toxicity against fish, published by Schüürmann et al.,¹⁵⁵ to our dataset. This model is based on a k-NN approach taking into consideration the three nearest neighbors.

The similarity between the compounds is thereby defined by atom centered fragments (ACF).^{145,171,172} The underlying source of data was the EPA's fathead minnow toxicity database, which results in a model that is trained on only one species. The assessment of the reliability of the derived predictions is ACF-based as well and works with similarity thresholds. We compared the prediction quality of both models, taking into consideration the suggested similarity levels (0.8 and 0.9), as well as the different species in our data set. The retrieved results can be seen in detail in Tab. 8.

Table 8. Results derived from a comparison of the developed model (3Desc) with those of the read-across approach (R-A). The results are discriminated with reference to the similarity thresholds (Simi) and the underlying fish species. In addition to RMSE, Q² and R², also the model bias, the maximum positive error (MPE) and the maximum negative error (MNE) are shown.

Species	Simi	Comp.	Model	RMSE	R ²	Q ²	MNE	MPE	Bias
<i>All</i>	n/a	1358	3Desc	0.85	0.65	0.65	2.26	2.63	0
			R-A	0.89	0.68	0.62	3.23	4.40	0.21
	0.8	769	3Desc	0.73	0.73	0.72	2.11	2.60	0
			R-A	0.53	0.86	0.85	2.21	3.00	0.05
	0.9	656	3Desc	0.71	0.71	0.70	2.11	2.45	-0.03
			R-A	0.43	0.90	0.89	2.23	2.07	0.02
<i>Fathead minnow</i>	n/a	660	3Desc	0.76	0.68	0.67	2.22	2.51	-0.06
			R-A	0.53	0.85	0.84	3.29	4.27	0.01
	0.8	604	3Desc	0.71	0.70	0.69	1.90	2.45	-0.05
			R-A	0.37	0.92	0.92	1.94	2.25	0
	0.9	581	3Desc	0.71	0.69	0.68	1.90	2.45	-0.05
			R-A	0.33	0.94	0.93	1.38	1.46	0
<i>Rainbow trout</i>	n/a	450	3Desc	0.94	0.55	0.53	2.14	2.63	0.12
			R-A	1.16	0.53	0.28	2.40	4.40	0.49
	0.8	91	3Desc	0.78	0.79	0.76	1.42	2.36	0.27
			R-A	0.85	0.78	0.71	2.05	2.36	0.36
	0.9	43	3Desc	0.66	0.77	0.75	1.22	1.74	0.15
			R-A	0.78	0.70	0.65	1.50	2.07	0.24
<i>Others</i>	n/a	248	3Desc	0.91	0.56	0.55	2.26	2.60	-0.03
			R-A	1.07	0.52	0.37	2.69	4.09	0.22
	0.8	74	3Desc	0.81	0.44	0.38	2.11	2.60	0.05
			R-A	0.95	0.48	0.14	2.21	3.00	0.02
	0.9	32	3Desc	0.76	0.41	0.29	2.11	1.92	0.08
			R-A	0.99	0.38	-0.20	2.23	1.67	-0.10

The first observation is that the read-across model shows an excellent performance for measurements on the fathead minnow in general and for measurements on chemicals with high similarity templates in the training set in particular. This was expected, as the EPA database is the basic source of measurements also for our model and given a similarity threshold of 0.8, the read-across model covers 45% of measurements with a Q²=0.92. Concerning the whole of these compounds, the read-across model is clearly preferable to ours, which performs with a Q²<0.7.

The evaluation of the results derived on other species is more difficult, as most measurements were affecting compounds beyond the applicability domain of the read-across model. Only 20% of the measurements for the rainbow trout and 30% of the measurements for other fish passed the medium similarity threshold of 0.8. In general, the three-descriptor model reached a more reliable performance on species different to the fathead minnow and for compounds failing the medium similarity threshold.

Remarkably, the three descriptor model resulted in a lower maximum prediction error for all examined partitions of the dataset, except for compounds on the fathead minnow with a similarity level > 0.9. This means that it has a lower decline to under-estimate the toxic level of a compound.”[d]

3.1.2.7 Mechanistic interpretation

“The dependency between logP and (chronic) toxicity has been frequently reported and is reasonable.¹⁷³ An increased hydrophobicity (which is in general equivalent with lipophilicity) causes an increased bioconcentration factor. Therefore, the duration until a chemical is excreted increases and its toxic effect swells. Further, a recent study reports a correlation between the protein binding affinity and logP.¹⁷⁴ This relationship also increases the effects of toxic metabolites of a compound.

ATSm2 describes how the mass is distributed along the topological structure. It has very low values for small molecules and linear, especially aliphatic compounds, while it reaches high values for large, branched molecules containing heavier atoms at a topological distance of two. Fig. 36 shows this mechanism for two compounds of similar molecular mass (350) and comparable predicted logP (4.7).

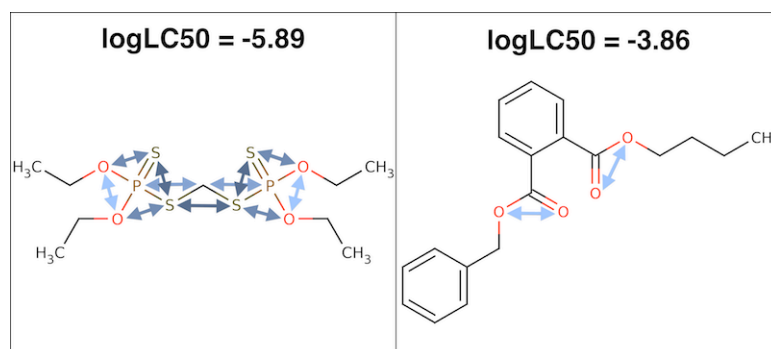


Figure 36. Visualization of the mass autocorrelation (ATSM2) for diethion (left) and butyl benzyl phthalate (right). Both compounds have a comparable molecular weight and comparable logP value. Their difference in toxicity can be explained by the mass autocorrelation. [d]

The molecule on the left is diethion and the right one is butyl benzyl phthalate. The blue arrows indicate the autocorrelation according to the ATSM2 descriptor. The darker an array is, the higher also the mass autocorrelation is. The difference in the aquatic toxicity (more than two log units) of these two compounds can be attributed to the different mass autocorrelation. Whereas the only noteworthy of these autocorrelations for the phthalate are between two pairs of oxygen (256

each), the diethion contains numerous autocorrelations of at least the same intensity. The highest impact thereby results from three pairs of sulfur (1024 each).

As an articulative example, dithiophosphates fall into this category with dioxathion (78-34-2), an insecticide, having the largest descriptor value of all followed by besultap (17606-31-4), an ester of thiosulfonic acid (also an insecticide). Other thiophosphates such as diethion (563-12-2) follow closely. Ionophores such as Monensin (similar to 17090-79-8) and Lasalocid, widely used polyether antibiotics, which act by its ability to transport metal cations through cellular and subcellular membranes, also have high values. Both Monensin and Lasalocid have some degree of activity on mammalian cells and thus toxicity is common, especially horses and dogs are very susceptible to the toxic effects and it comes to no surprise that fish are affected as well. Perhalogenated compounds (4234-79-1, 173584-44-6, 4151-50-2), PCBs and sulfuramides (86209-51-0) also achieve high ATSm2 descriptor values, especially when coupled with nitro groups like Bromethalin (63333-35-7), a rodenticide.

Table 9. Differentiation of compounds containing triple bonds.

	Non-zero	Zero	Σ
Cyanides	10	42	52
Alkynes	3	21	24
Σ	13	63	

Although the value of the E-State descriptor (Se1C2C3ts) is non-zero for only 13 out of 1358 compounds (1 % of the dataset), it has an overall correlation of 22.4% to the target property. It is highly specific and affects mostly cyanic compounds (10 out of the 13). This is reasonable, as cyanides form highly toxic prussic acid upon metabolization, which halts cellular respiration. The other 3 compounds are highly reactive alkynes, which affect protein functions. Tab. 9 shows this distribution in detail.

Out of a total of 76 structures containing triple bonds, 52 are cyanides. Out of these, 10 are relevant for Se1C2C3ts. In general, cyanides are already classified as toxic by logP and ATSm2. Se1C2C3ts was non-zero only for those cyanides that are easily accessible and thus rapidly metabolized. This makes them very strong poisons such as Cyfluthrin (68359-37-5) - and separates them from the less toxic such as adiponitrile (629-40-3), a common nylon precursor with delayed metabolization, neonicotinoids such as thiacloprid, or benzyl cyanides.

The other 3 compounds, for which Se1C2C3ts was non-zero, are secondary alkynols containing a terminal alkyne unit next to a secondary alcohol group. Se1C2C3ts distinguishes these from the bulky and less reactive 21 tertiary alkynols, e.g. 3,6-dimethylhept-1-yn-3-ol (19549-98-5). Secondary alkynols are known GABA receptor blockers and strong poisons. Well-known members of this class are cicutoxin and aethusin.

Although the E-State descriptor is relevant only for a small partition of the dataset (< 1%) its contribution to the prediction quality, especially in regions of high

toxicity is disproportionate. Fig. 37 shows the effect of the descriptor for all 13 compounds that are affected. The x-axis shows the measured values, whereas the y-axis displays the predicted values. The red dots show the values predicted by a regression model derived only from ATSm2 and AlogPS_logP. Apparently the toxicity of all compounds is underestimated. The green dots show the prediction values derived with the three-descriptor model. The prediction error is clearly increased for twelve out of the 13 compounds. The influence on the remaining 1345 compounds was negligible.

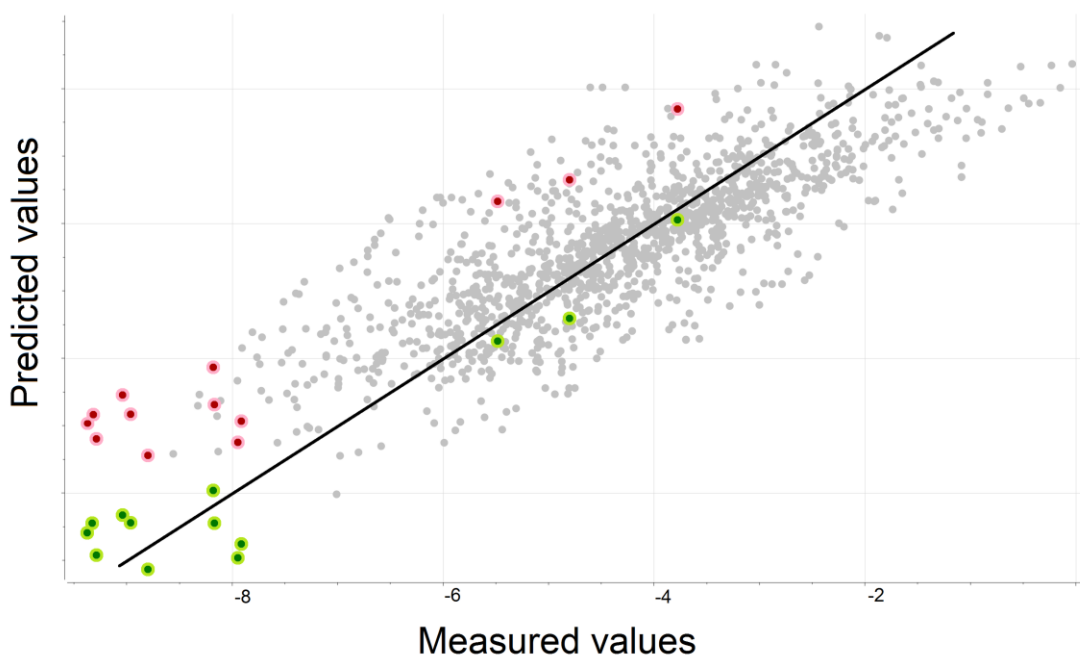


Figure 37. The influence of the E-State descriptor on the resulting model. The red dots show the values predicted by a model limited to two descriptors (ATSm2 and AlogPS_logP), whereas the green dots show the prediction values for the same compounds derived with the developed three descriptor model. [d]

Se1C2C3ts is therefore the only descriptor in our regression model that is directly associated with a toxic group (cyanides), whereas the other two descriptors mainly depict general qualities of a compound (lipophilicity and a combination of size, ramification and content of heavier atoms in a topological distance of 2 bonds like dithiophosphates).”[d]

3.1.2.8 Applicability domain estimation

“Naturally, the applicability domain of this model is predefined by the scope of the databases that sourced the measurements used for the models. As a solid foundation (profound basis), the fathead minnow database was used to ensure a wide coverage. It is based on an U.S. industrial chemical inventory of discrete organic chemicals.¹⁷⁵ We have specifically extended the applicability domain to pesticides by adding data from the pesticide properties database with a wide coverage of the field. These compounds were optimized towards a balance of strong toxic effects that are typically controlled in mammals by specific adsorption, distribution, metabolism, and or binding effects.

Out of the 46 compounds with an absolute prediction error larger than 2 log units, 33 were underestimated regarding their toxicity potential. The 20 compounds with the largest absolute error were all ‘underpredicted’. A careful analysis revealed that this is partly due to the systematic errors of defined compound classes. In addition, several compound classes could be identified as out of the applicability domain of this model.

One of the major systematically ‘underpredicted’ compounds are carbamates and dithiocarbamates. The complete dataset consisted of more than 70 compounds belonging to these classes. These compounds were ‘underpredicted’ by an average of 0.6 log units, and generally showed no quality prediction results ($Q^2=0.0$). Activated alkenes and alkynes (those next to $-C=O$, CN, C(O)N, NO_2 or SO_2 groups), were also not so well predicted ($Q^2=0.2$) and should be considered out of the AD of this model.

Toxicological intermediates can create a variety of hazardous effects in vivo due to their special behavior and should also be considered out of the applicability domain. They especially include thiocarbonyls and quinones accounting for 5 out of the 20 most underestimated predictions.

To analyze if our model is predicting especially well for a certain group of chemicals, we used the structural alerts implemented in OCHEM¹⁷⁶ to assign 248 functional groups to the compounds in the dataset. In the next step, we dropped all those groups, which did not apply to a minimum of 20 compounds and calculated the Q^2 for the compounds assigned to the remaining 66 groups. Tab. 10 shows the result for the ten best predicted structural groups in detail.

Table 10. The ten best predicted structural groups within the dataset and their according Q^2 values.

Structural group	Number of compounds	Q^2
Nitriles	50	0.88
Secondary amines	63	0.85
Tertiary mixed amines (aryl alkyl)	28	0.83
Secondary mixed amines (aryl alkyl)	29	0.83
Tertiary alcohols	40	0.81
Diarylethers	54	0.81
Primary aliphatic amines	40	0.81
Aryl bromides	33	0.79
Secondary aliphatic amines	29	0.78
Primary amines	117	0.77
Overall	354	0.84

It is obvious that the prediction of both nitriles and amines in general is reliable. Furthermore, tertiary alcohols, diarylethers and aryl bromides belong to the well predicted structural compound classes. Due to the overlap within different structural groups, the aggregation of all compounds in these ten groups resulted in

a dataset of 354 different compounds. The Q^2 we retained from the model to predict them was 0.84, which is of regulatory relevance.

Apart from those well predicted groups, we also investigated the worst performing ones and found five structural groups predicted with an R^2 lower than 0.35. Those were carboxylic acid tertiary amides (47 compounds), ureas (45), hydrocarbons (34), thioethers (34) and thiocarbamic acid derivatives (20). They are therefore clearly out of the applicability domain of the model. A further observation was that 37 sulfonic acid derivatives and 25 sulfonamides revealed an acceptable R^2 (> 0.5), while the according Q^2 was 0 in both cases. This resulted from a systematic error in the model, which estimated the toxicity of those compounds in average 0.7 log units too low.

Finally, we decided to assess the justified application of our three-descriptor model by comparing it to a linear regression model, exclusively derived from the estimated logP. We therefore used the ALogPS_logP descriptor. Within all available logP estimates, it showed the highest correlation to the measured LC_{50} values. A CV_{LOO} of the derived model delivered an RMSE=0.96, a correlation coefficient of 74%, a $Q^2=0.54$ and an MPE=3.2 log units. Referring to the three-descriptor model, the RMSE is increased by more than 0.1 log units and the Q^2 decreased by 0.11.

Table 11. Comparison of the difference in prediction error of the logP model and the three descriptor model.

Δ_{Err}	Our model	logP model
0.0 - 0.1	150 (11.0%)	140 (10.3%)
0.1 - 0.2	143 (10.5%)	108 (8.0%)
0.2 - 0.3	136 (10.0%)	95 (7.0%)
0.3 - 0.4	151 (11.1%)	78 (5.7%)
0.4 - 0.5	113 (8.3%)	36 (2.7%)
0.5 - 0.6	54 (4.0%)	16 (1.2%)
0.6 - 0.7	37 (2.7%)	7 (0.5%)
0.7 - 0.8	24 (1.8%)	14 (1.0%)
0.8 - 0.9	11 (0.8%)	8 (0.6%)
0.9 - 1.0	5 (0.4%)	2 (0.1%)
>0.1	19 (1.4%)	11 (0.8%)
Σ	843 (62.1%)	515 (37.9%)

To gain a deeper insight we calculated the difference in the absolute prediction error (Δ_{Err}) between the logP model and our three-descriptor model for all considered measurements. For 843 out of the 1358 compounds, the three-descriptor model performed better. Tab. 11 shows this distribution depending on the value of Δ_{Err} . For more than 770 compounds, the prediction error was inconsiderable ($\Delta_{Err} < 0.3$). In the range of an intermediate difference in the prediction quality ($0.3 < \Delta_{Err} < 0.6$) for 318 out of the affected 448 compounds (71%) the three-descriptor model delivered a higher accuracy. Similarly, in the range of a significant difference in the prediction error ($0.6 < \Delta_{Err}$) for 71% of affected compounds the three-descriptor model was preferable.

The attempt to correlate the Δ_{Err} with the ISIDA fragments (length 2-5) in order to define a structurally based applicability domain to decide in which cases the predictions derived from the three descriptor model are favorable (or disadvantageous) compared to those of the logP model, revealed two major findings: 1) the three descriptor model performs significantly better on cyanides and alkynes. This is logical, as these are the compounds, which are covered by the E-State descriptor; 2) the three descriptor model reveals worse predictions in case of per- or polyfluorinated compounds. A possible reason therefore is that substitutions with fluorine increase the value of the mass autocorrelation and thereby the predicted toxicity, which is equivalent to an overestimation. The acute toxicity of such compounds is known to be relatively low and the major problem resulting from them is their persistence and low degradability.¹⁷⁷

Apart from that, the examination was unspecific. The improvement in prediction that we derived cannot be linked to certain chemical groups, but is a mostly global observation without explicit reasoning in small substructures.”[d]

3.1.2.9 Closing remarks

“The endpoint $\log\text{LC}_{50}$ is highly complex, as it is not limited to one or few modes of action, for instance, the inhibition of a certain protein. The possibility to cover this complexity with current QSAR approaches is ambitious, if not impossible. Therefore, we believe that a general solution is preferable. Our model uses three descriptors and delivers a reasonable mechanistic explanation of aquatic toxicity. It is not limited to a certain species, but takes into account a variety of them. Furthermore, the model was trained and validated on measurements from numerous sources. This is important in terms of reliability and robustness.

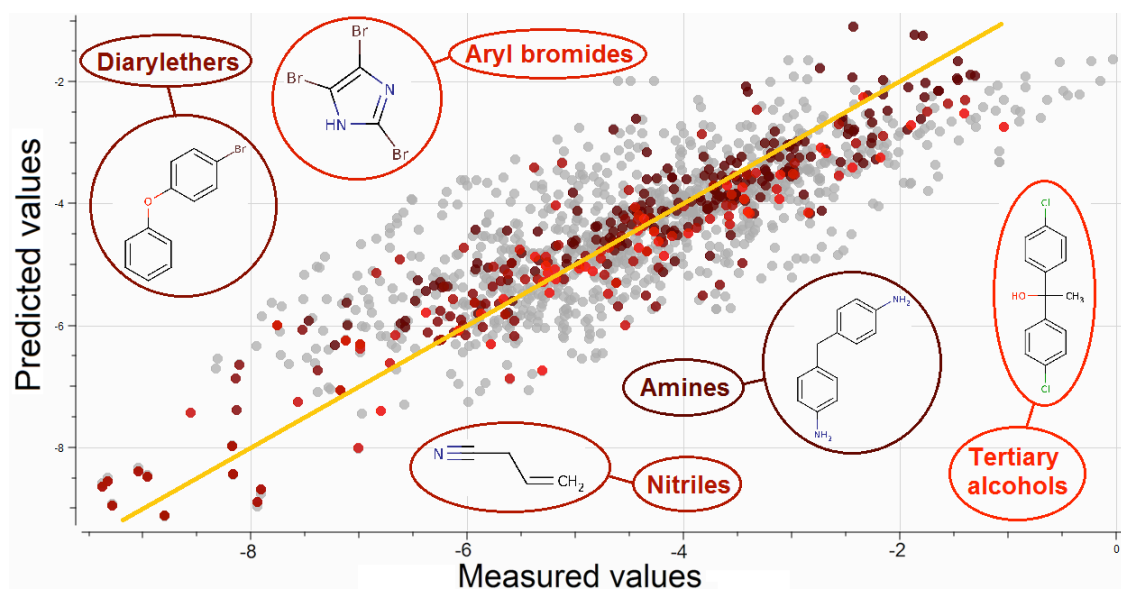


Figure 38. Measured versus predicted values for five chemical groups with high prediction accuracy in our model. Remarkably, the model reveals a prediction error exceeding one log unit for only few compounds that belong to these groups. The prediction accuracy thereby approximately concurs with the measurement accuracy.

The improvement of the predictive quality that we reached by applying higher complex models was negligible. This shows that the limitation to three variables is not so much of an over-simplification, rather than the reduction to the dependencies we can reliably confirm.

Our model clearly improves the global estimation of aquatic toxicity compared to a simply logP based estimation. And although the performance of read-across models is clearly better in those areas of the chemical space, which are experimentally covered with a high density, our models shows a comparably robust performance in the areas that are no more applicable with such read across models. Referring to the chemical groups we determined to be predicted with a high reliability (amines, nitriles, etc.), our model can be of regulatory relevance for compounds belonging to those classes.”[d] Fig. 38 shows representative structures for each of the classes, as well as the comparably low discrepancy between the measured and predicted values for the majority of affected compounds.

“Furthermore, our model provides an intuitive access to the estimation of the toxicity of a compound. Except the hydrophobicity of a compound, all other criteria used in the model can be deduced from a simple look at the compound’s structure, that is the existence of a cyan group, molecular size, or the number of heavy atom pairs with two bonds in between.”[d]

3.2 Customized implementation of experimental design approaches

In order to enable an overview of the performance of frequently used experimental design approach and to enable a comparison with more complex approaches, I implemented at least one example procedure for each of the conceptual ideas (dissimilarity-based, similarity-based and partition-based). Hereby, the focus was to enable the application of these approaches with the precondition of prior knowledge, amongst other criteria.

3.2.1 D-Optimal design

“The D-Optimal selection criterion was implemented as suggested in the literature.¹⁷⁸ Fedorov’s heuristic approach¹⁷⁹ was used to optimize the selection speed. Further, the implementation was extended by the option to add a fixed seed to the selection. This additional feature enables us to perform a compound selection that depends on a preselected set of compounds. Newly selected compounds are therefore not only most distinct to one another but also to the preselected compounds. This enhancement was implemented by adding the preselected compounds to the model matrix.”[a]

3.2.2 Kennard-Stone algorithm

As the Kennard-Stone algorithm was developed, to start from a predefined seed of instances, no efforts were required to customize it for the use of a fixed seed. The initially selected compound is the central point within the dataset. The central point is defined to be the compound with the minimum sum of Euclidean distances to all other compounds within the dataset.

3.2.3 MDC selection

Due to the concept of using reciprocal ranks, instead of directly using the pairwise distances of compounds, a customization of the MDC selection to work with an initial seed is not feasible. Furthermore, the original implementation, by Hudson et al.⁷¹ provides a stop criterion, which limits the number of compounds to be selected. “As this study concentrates on a comparison of a fixed number of selected compounds, the compounds are used regarding their selection order.”[b]

3.2.4 Space filling design

The space filling design is an extension of the full factorial design. Instead of a simple binary representation of each variable, it works by partitioning the chemical space into subspaces of equal size. This fragmentation is derived by the division of each axis into the same number of bins. Consequently, for a number of b bins and a number of a axes, the number of resulting subspaces is b^a . To avoid the problem of exceeding the number of relevant subspaces, we limited the number of considered dimensions to a maximum of three. “From each of the resulting

subspaces a compound is selected, but as the compounds are not equally distributed in the chemical space, subspaces can be completely without a representative compound. It is therefore difficult to fix the number of finally selected compounds.”[c]

“To address this problem in our implementation, the number of bins, the axes are separated into, is not preliminary fixed, but automatically detected. Therefore, the number of bins is iteratively 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 focuses 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.”[b]

3.3 Stepwise selection approaches

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. Given, for example, a set of 600 compounds of interest and the limitations of being able to test only 100 of these compounds, almost no lab will test all these 100 compounds in parallel, but due to restricted capacities, in a stepwise procedure.

A stepwise procedure like that implicates, that information about the measured property is gathered from testing cycle to testing cycle. In the standard approaches, that select all compounds in one step, this growing amount of information is not taken into consideration, although exactly this information could be used to refine and thereby significantly increase the quality of the experimental design approach.

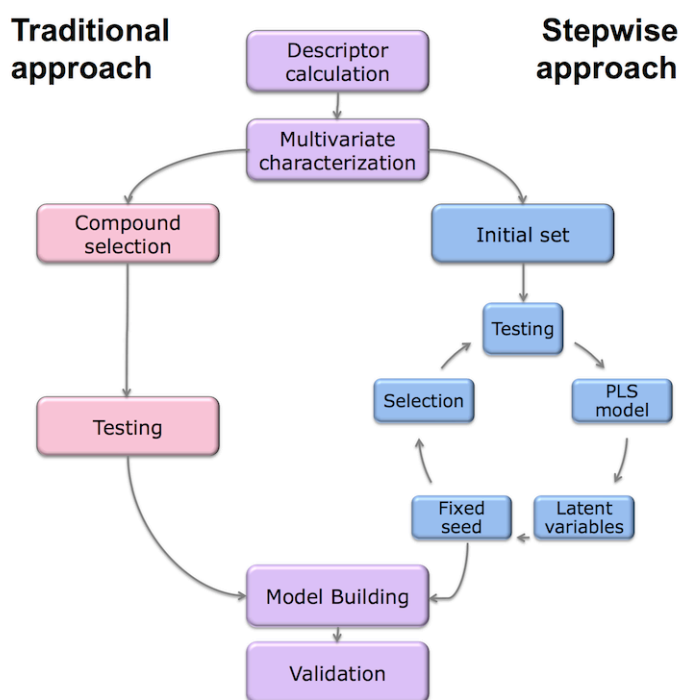


Figure 39. Comparison of the traditional workflow (left) and the suggested stepwise selection (right). Instead of measuring all compounds at the same time, the stepwise approach works with an iterative refinement of the depiction of the chemical space.[a]

Another difficulty is that the selection of compounds is performed in the principal property space. Although the application of PCA to extract the principal components is a powerful tool, there is no guarantee, that a principal component is correlated with the target property. Principal components are selected just by ranking the variation in descriptor space. But in exceptional cases, for a certain endpoint, principal components can represent nothing else than noise without any relevant information.

Principal components are maintained from just the descriptors, irrespective of the target property. This leads to the fact, that an experimental design and thereby the selected compounds, derived from the classical approaches, is identical for all endpoints, Irrespective if, this endpoint is physicochemical, chemical, biological or toxicological. The question is whether there is a strategy that could provide a better selection of compounds by taking into consideration the correlation to the target property and available data.

“The adaptive experimental design approaches we use, work in a stepwise manner, where each step consists of two phases. In the first phase the representation of the chemical space is refined. This is done by using the preliminary gathered information from the target property and analyzing its correlation to the chemical space. In the second phase a selection algorithm is executed on the newly arranged chemical space. The selection is hereby taking all previously selected compounds into consideration. These phases are executed in an alternating way until a prefixed number of compounds are reached.”[c]

“The most important differences between the stepwise approaches and the traditional selection approaches are shown in Fig. 39. Whereas the traditional method (left side of the figure, in pink) selects all compounds at the same time, the stepwise approach (right side of the figure, in blue) constantly increases the number of compounds cyclically. Further, the chemical space to represent the compounds is refined with each cycle.”[a]

3.3.1 Ordering the chemical space

“The idea behind the iterative rearrangement of the chemical space is to adjust the design of an experiment to a certain endpoint and consequently to reach a faster increase in the resulting model performance. Experimental designs derived from PCA space are not aligned to the target property, but are identical for the same selection algorithm and executed on the same compound collection, regardless of the endpoint. For this reason they are unspecific and most probably not optimal.”[c]

3.3.1.1 Underlying theory

A further disadvantage of the PCA-based representation of the chemical space is shown in Fig. 40a. The axes represent the first principal components and the dots represent the compounds in the logLC₅₀ dataset. The coloring of the compounds corresponds to the target property. Compounds with high logLC₅₀ values are indicated red, whereas compounds with low values are indicated yellow.

It is obvious that there is no or only low correlation between the target property and the principal components. This results in a distribution of similar logLC₅₀ values over the whole chemical space. Given a partition based experimental design approach or a dissimilarity selection, there is an increased probability that the collection of selected compounds contains multiple molecules with similar endpoint intensity. For obvious reasons, this is adverse in terms of modeling. A

representative selection should contain a representative overview of the value facet.

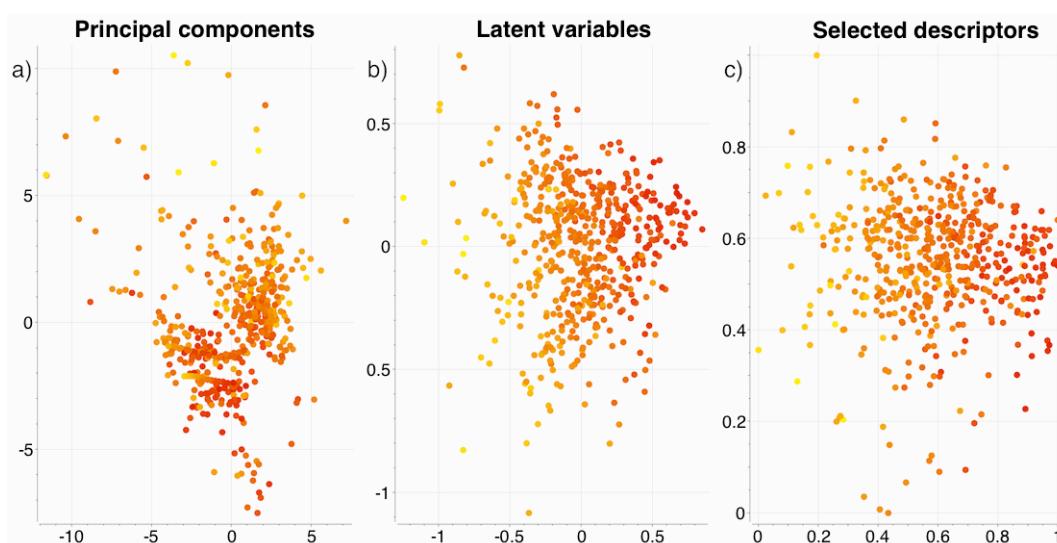


Figure 40. The chemical space of the logLC₅₀ dataset represented by a) the first principle components; b) PLS latent variables; and c) selected descriptors. The coloring of the compounds represents the value of the target property. The PCA representation results in an unordered distribution.

3.3.1.1.1 PLS latent variables

“Latent variables from PLS are comparable to the principal components of a PCA. However, in contrast to PCA components, which are selected to maximize the variance of the dataset (i.e., to cover as much of the data variability as possible), the PLS latent variables are selected to maximize the covariance (i.e., to provide maximum correlation) with the target variable. Therefore, in addition to PCA components, the latent variables contain information about the target variable.”[a]

Fig. 40b shows the logLC₅₀ dataset as well, but the axes represent the first PLS latent variables. The depiction of the chemical space with latent variables contributes to a clearly higher degree of order in the chemical space, compared to a PCA depiction. Thereby it minimizes the probability of selecting samples, which are not optimal in terms of the representation of the target property.

3.3.1.1.2 Supervised descriptor selection

The concept of a supervised descriptor selection is similar to the concept of PLS. Variables are ranked by their correlation to the target property and their contribution to explain the variance in the data. But contrary to PLS, the descriptor selection does not work with an orthogonal transformation of the chemical space, but directly on selecting relevant descriptors. Literature provides a variety of sophisticated approaches, which are based on neural networks,^{180,181} simulated annealing,¹⁸² or other ideas.^{183,184}

Fig. 40c shows the logLC₅₀ dataset represented by selected descriptors. The first observation is the similarity to the representation based on PLS latent variables. The underlying selection procedure is explained in chapter 5.1.1.3.2.

3.3.1.2 PLS-Optimal

“PLS-Optimal is an adaptive approach that combines the D-Optimal criterion with the partial least squares technique (PLS). The representation of the chemical space within this approach is realized with PLS components instead of PCA components.”[c]

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

As the number of measurements increases from cycle to cycle, each new model is an improvement of the previous one. Based on these iteratively refined latent variables, an initially selected set of compounds is extended in a stepwise manner. A similar idea was proposed by Lundstedt and Thelin.¹⁸⁵ The authors used a two-step process consisting of a synthesis step and a purification step in which they alternated between PCA and PLS. However, their aim was to select the most important variables for a model, while the aim of our method is to find the most informative compounds for model development.”[a]

3.3.1.3 DescRep

“The next approach, DescRep works in a similar way as PLS-Optimal. However, it combines a similarity-based approach (instead of a dissimilarity based one) with a representation of the chemical space using selected descriptors (instead of PLS components). As for PLS-Optimal, the preselected compounds are used as reference information to evaluate the most important descriptors.”[c]

3.3.1.3.1 Similarity based selection approach

“As a stepwise experimental design procedure requires a selection method, which is able to take a preselected initial seed of compounds into consideration, it is not possible without further ado to use the MDC selection, as its concept of ranking distances cannot be adapted to this precondition. Therefore we developed a

selection method based on the idea that structural similarity of compounds also conditions similar values regarding a certain endpoint.

We select an initial seed of compounds starting from a k-Means-based partition of the chemical space (represented by the principal components) into a predefined number of clusters. The initial seed contains the most representative compound of each cluster. The most representative one is hereby defined as that compound with the lowest sum of pairwise distances to all other compounds within the same cell. The k-Means clustering was initialized 15 times with randomly assigned starting compounds. The finally picked clustering was the one with the lowest sum of pairwise differences within each cluster.

In each further step the chemical space is represented by a selection of descriptors based on the preselected compounds. The preselected compounds are extended by new ones, which are assigned to be the most informative ones for all other compounds based on a priority score (*PS*) calculated for each compound. *PS* estimates how well a compound is represented by all previously selected compounds.

Initially, all compounds are assigned a *PS* of 1.0 and the distance matrix *DM*, containing the pairwise distances between all compounds, is calculated. The distance matrix (normalized to [0,1] range) is used as *PS* to select the first compound.

For all following compounds, the normalized pairwise distances of *N* preselected compounds to the remaining compounds in the dataset are used to determine how well each compound is already represented and select the least represented ones. Each compound *x* within the set gets assigned a correction factor CF_{xi} for each preselected compound *i*. The correction factor refers to a hyperbolic distance function and it is used to adjust the *PS* to the preselected compounds.

$$PS_{x_new} = PS_x * \prod_{i=1}^N CF_{xi}$$

The correction factor is calculated as $CF_{xi} = (1 - (1 - DM_{xi})^{exp})$.

The exponent *exp* is not fixed, but depends on the distribution of the data in the descriptor space and the number of compounds to be selected. It is recalculated in each selection cycle. Referring to the most central point within the dataset (which is again defined as the one with the lowest sum of pairwise distances to all other compounds) *exp* is determined as the value for which the number of preselected compounds and compounds to be selected in the present cycle has to have a value of $(1 - DM_{xi})^{exp}$, which is higher than a given threshold. The threshold value we used in this study was $\lambda=0.75$. For the calibration datasets on density, bioconcentration, lipophilicity and solubility were used. We experimentally determined that this is an appropriate value.

This additional feature of the recalculated exponent enables one to also handle exceptional data distributions. The method is not sensitive to the parameterization. We tried different versions of λ within the range of 0.5-0.9 and did not observe significant changes in the method performance. Moreover, it should be mentioned that parameterization and an appropriate distance function are general issues for similarity-based selections.⁷¹

Based on these prior conditions, the correction factors for all combinations of not yet selected compounds are calculated. The collection of compounds finally selected for testing is the one that minimize the sum of priority scores over all compounds.”[c]

3.3.1.3.2 Supervised descriptor selection

“The search space for DescRep is spanned by a fixed number of selected descriptors. The selection process follows a simple idea and is therefore straight and efficient.

In the first step a scoring list S , containing the correlation coefficient of each descriptor to the target property, is built. This correlation coefficient is derived only from the preselected compounds, which are already ‘measured’. Additionally, the correlation matrix M , containing the absolute pairwise correlation of any combination of two descriptors, is built.

According to the scoring list, the descriptor that should be selected is the one with the highest score, which is initially equivalent to the one with the highest correlation to the target property. After the selection of a descriptor x the scores are updated to avoid pairwise correlations in the final selection. Regarding the compound i , its score S_i gets updated to $S_i * (1 - M_{ix})^3$. Thereby the scores of descriptors, which are highly correlated to the preselected ones are decreased, which helps to avoid the selection of redundant information. The underlying idea of selecting variables with a high correlation to the target property and the elimination of inter-correlated variables is similar to partial least squares. This procedure is repeated until a predefined number of descriptors are selected.”[c]

3.3.2 Applicability domain approaches

The use of a descriptor-based representation of chemical compounds in terms of experimental design is established and well founded. In the classical meaning, also the latent variable representation we use for the PLS-Optimal approach, is member of such a representation, as the PLS components are an orthogonal transformation of the descriptor space. Still, the relation and correlation of chemical compounds can be represented in other ways.

We developed a workflow to apply “a novel adaptive projection that moves from a descriptor-based construction of the chemical space towards 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, the others

utilize the concept of the ‘distance to model’-parameter (DM) suggested by Tetko et al.^{122,130,147} to estimate the uncertainty of prediction for each compound within a data collection. We are not aware of other studies in computational chemistry, using this parameter for compound selection.”[e]

3.3.2.1 Underlying theory

The following paragraphs are intended to give a short overview of the ensemble based applicability domain approaches and their potential use in terms of experimental design.

3.3.2.1.1 Uncertainty and error

“The underlying theory on the ensemble based applicability domain 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 dataset could be done using a bagging approach to generate a predefined number of subsets by resampling 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 the average value, which is used as the prediction value, can be calculated, but additionally the variance in the predictions as a measurement of uncertainty could also be determined.

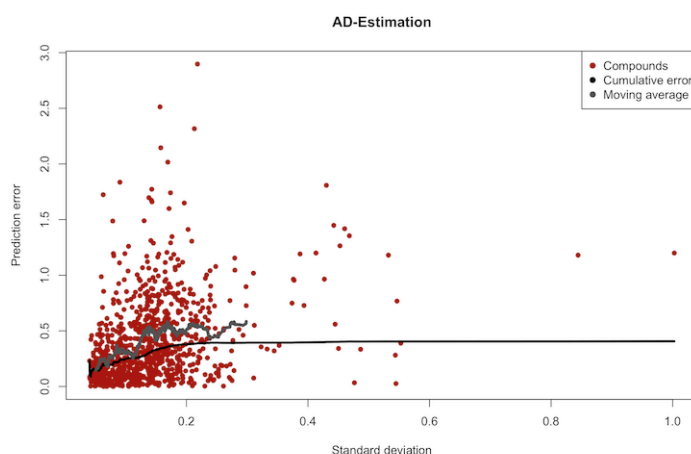


Figure 41. The correlation between prediction uncertainty (variation of prediction) and prediction error for a model on environmental toxicity of 1093 molecules against *T. pyriformis*. [e]

Previous studies^{65,122,130,147} have shown the correlation between the uncertainty of the prediction and the prediction error. Fig. 41 illustrates this correlation for the -logIC₅₀ dataset. Each red dot represents a compound, the x-axis represents the standard deviation of the ensemble predictions and the y-axis represents the prediction error. The black line depicts the cumulative error of all compounds

predicted within a certain standard deviation and the grey line depicts the average error of a sliding window.”[e]

3.3.2.1.2 Unification of the chemical space

“An additional feature of the predictions derived with the bagging approach is that they define a compound referring to the property space, instead of the descriptor space. This enables the representation of a dataset to a higher extent of independence of a certain descriptor set.

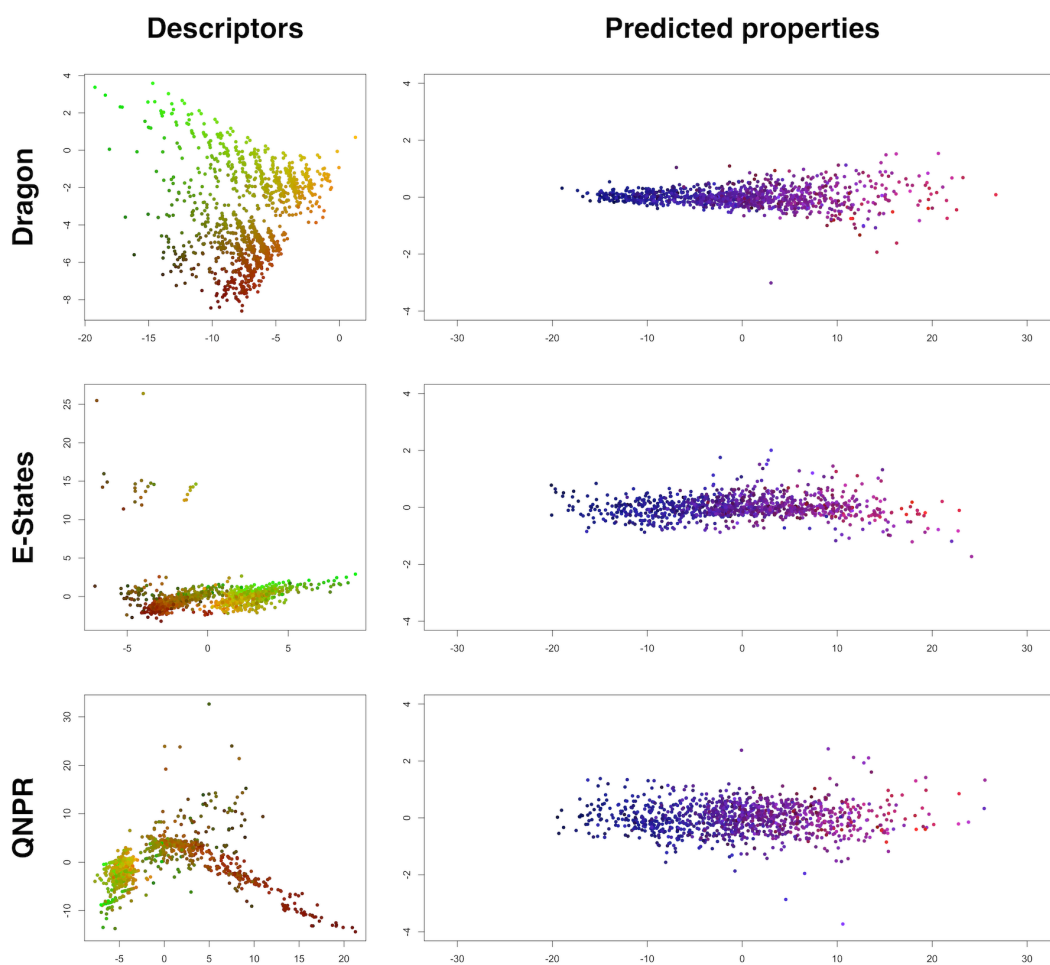


Figure 42. Principal components of a dataset on descriptors (left) and predicted properties (right). [e]

To visualize this, we calculated DRAGON 6.0 descriptors,¹³² E-State indices and quantitative name property relationship (QNPR) descriptors³⁴ (derived from a SMILES representation of the compounds) for the $-\log\text{IC}_{50}$ dataset and reduced all three representations of the dataset to two principal components. The left column of Fig. 42 shows the results. The coloring of the compounds is just a topological one to enable the identification of changes in the dataset depiction, it has no functional meaning.

Additionally, we used the bagging approach to calculate 64 partial least squares (PLS) regression models¹⁸⁶ on each descriptor set and reduced the derived predictions for each molecule to two principal components. The results are shown in the right column of Fig. 42. Also in this depiction, the coloring 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 is harmonized to a higher extent in the predicted properties view.”[e]

3.3.2.2 Implementation

“The DM-based approaches decide on testing a compound either exclusively on 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 the ASociative Neural Networks (ASNN)⁶⁵ and which is also used in the ASNN LIBRARY mode to make local corrections.⁸³

Based on a predefined selection of compounds, the stepwise approaches extend this seed in a stepwise manner and based on the predicted properties and thereby the applicability domain estimation. After each (hypothetical) measurement cycle, a new ensemble of 64 bagging models is calculated, using partial least squares (PLS) regression.^{76,186”[e]}

3.3.2.2.1 Distance-based selection

“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 applicability domain and therefore with new information; secondly, the selected compound, which was predicted with a high uncertainty, does not have to be measured anymore.

To prove if this concept works, we executed this approach in a one-by-one selection on the logK_{OC} dataset. We compared the performance of a PLS model derived on the selected compounds and applied it to the remaining ones with that of the k-Medoid approach, the Kennard-Stone algorithm and a random selection. The results are shown in Fig. 43.

The x-axis shows the number of selected compounds and the y-axis the RMSE on the non-selected compounds. As the RMSE for the Stepwise approach is initially low, and as the performance is constantly better than that of the Kennard-Stone algorithm and of higher stability than that of the random approach, the concept appears to work.

On a larger scale, and for the practical application, a selection and testing of compounds one by one is inappropriate. Therefore in all further applications, the n compounds, which should be selected in a measurement cycle, are the n compounds with the highest variance in prediction. Further, the initial seed of compounds is selected with a clustering approach.”[e]

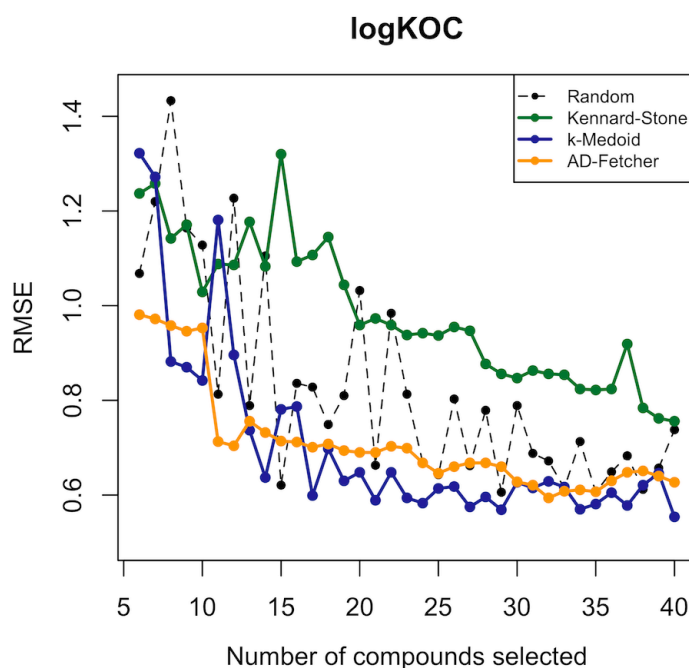


Figure 43. Proof of concept for AD-Fetcher. [e]

3.3.2.2.2 Representativeness-based selection

“The second approach does not take only the uncertainty of prediction into account, but 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 in the previous approach, we start with the assumption that the variance in prediction is proportional to the prediction error. Further, we assume that the extension of a model with a certain compound decreases the uncertainty in prediction of another compound proportionally to the correlation in the predictions of these two compounds. A similar concept has been successfully applied for the local corrections in the logP prediction by Tetko et al.⁸² and is called LIBRARY mode. Both assumptions, as we use them in this study, are simplifications of a more complex context. Nevertheless, these assumptions should be sufficient for the prioritization of representative compounds in a dataset. Additionally, one has to take into consideration that experimental design aims towards efficiency, not towards exhaustiveness.

Reasoning from our two estimations, the compound, which has to be chosen in each measurement cycle is the compound which decreases the prediction error of all 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; secondly the correlation matrix C , which contains the pairwise correlation of the prediction of each compound; and thirdly, the decision vector D , which is initially derived from a matrix multiplication of C and S and which 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 with the difference between 1 and this correlation.”[e]

3.3.2.2.3 Predicted property representation

Additionally, to investigate the direct use of predicted properties in terms of experimental design, we applied principle component analysis to them and used the obtained variables to span the search space for selection approaches.

3.4 Cluster-based selection approach

“In the recent past, approaches using density-based¹² or hierarchical clustering¹⁸⁷ were suggested. Partition-based approaches, utilizing the k-Means clustering, were also introduced; however, these approaches use the derived clusters to apply other selection algorithms to them.¹⁸⁸ Although pharmaceutical publications mention the possibility of using the clusters derived by k-Means to select exactly one representative from each cluster,¹² we are not aware of a study evaluating this technique in QSAR experimental design, comparing its performance to other experimental design techniques. The idea of assigning the compounds to different clusters and choosing a representative from 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 rather than to a hypothetical distribution.”[b]

Furthermore, in terms of a meaningful partition of the chemical space, an approach that divides the chemical space into hyper-dimensional spheres seems to be more appropriate in terms of experimental design, than density based approaches, such as DBSCAN¹⁸⁹ or OPTICS.¹⁹⁰

3.4.1 Implementation

“The basis for the selection approach was an implementation of the k-Medoid clustering. The k-Medoid clustering, partitions a set of data points into a given number of subsets, the clusters. Each data point is assigned to one cluster only and each cluster must contain at least one data point. The problem of finding the optimal partitioning belongs to the class of NP-hard problems,¹⁹¹ which are computationally very expensive. Heuristics are therefore used to find a local minimum by iteratively reassigning data points to certain clusters until convergence is reached.

For the k-Medoid approach, a number of k randomly selected data points are initially assigned as cluster centers. In the second step, each data point that is not assigned as a cluster center is assigned to the nearest cluster center (according to the Euclidean distance). In the third step, the cluster centers are reassigned to the data point within a cluster that has the lowest sum of pairwise distances to all other data points in the cluster. In case of two data points with identical values for the sum of pairwise distances, the decision, which of them to assign as cluster center is arbitrary. Following that, steps two and three are executed alternately, 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 in our case is 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.”[b]

An example of the data partitioning derived with the k-Medoid clustering can be seen in Fig. 44. The underlying dataset is the boiling point collection. The red dots indicate the cluster centers and according to this the compounds selected for

experimental testing. The distribution of these compounds is covering most of the occupied chemical space, with an adequate structural differentiation between them.

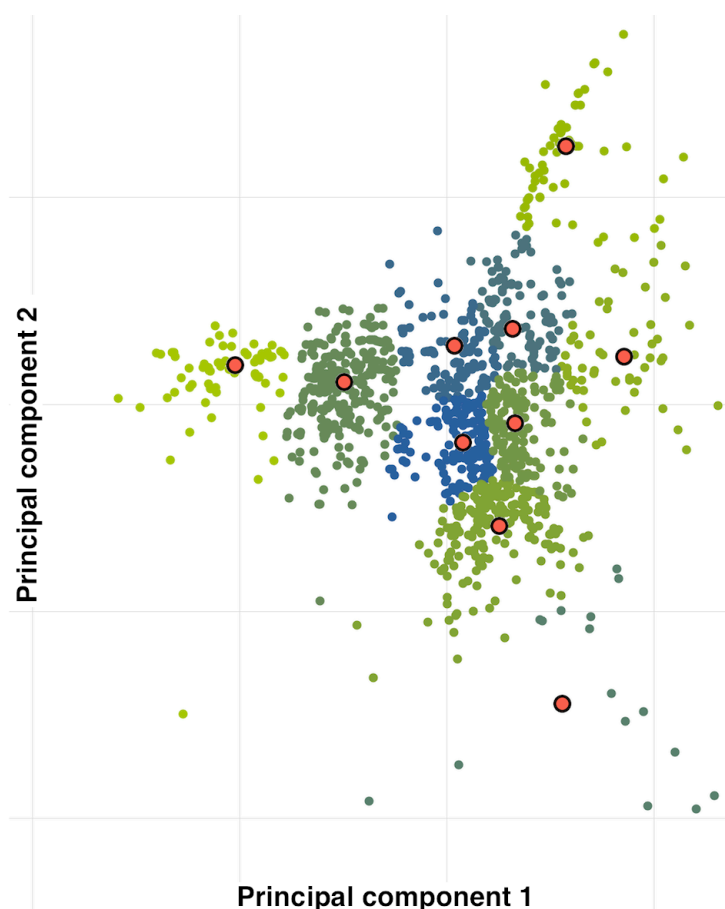


Figure 44. A k-Medoid data partitioning on a two-dimensional scale. Each compound is represented as a dot in a space spanned by the principal components. The coloring of the compounds indicates their affiliation to a certain cluster. The cluster centers are indicated red.

“The standard implementation of k-Medoid checks the membership of a data point to a certain cluster in an undefined order and reassigns the cluster centers each time a data point is assigned to a new cluster. Therefore, the final clustering is not deterministic, but depends on the order of the data points to be assigned to a new cluster.

This means that the clustering on a dataset, starting with a fixed set of initial cluster centers, can deliver different results if the order in which the data points are reassigned to new clusters is changed. To address this problem, our implementation does not reassign the cluster center before each data point has been assigned to a cluster. This modification results in a higher runtime requirement but contributes into the stability of the outcome.”[b]

3.4.2 Initial cluster centers

“A second specification of the k-Medoid clustering approach that prevents a deterministic result is the randomized selection of the initial cluster centers. To

address this problem and to investigate the influence of the initial centers, we implemented two different approaches to assign the initial cluster centers. The first utilizes the selection derived with using another approach - in our study the D-Optimal criterion and the Kennard-Stone algorithm - and initializes the k-Medoid clustering with these data points as initial cluster centers. The second approach follows the standard implementation and works with a random selection of the initial centers. To avoid getting stuck in a disadvantageous local minimum because of an unfortunate selection of initial cluster centers, 15 of these initial assignments were made, with the finally accepted clustering the one with the highest density.”[b]

3.5 Validation pipeline

To enable a reliable comparison of the selection approaches in terms of experimental design and model building, we implemented a robust validation pipeline. A schematic overview of the steps that are included in this procedure is given in Fig. 45. It starts with a random sampling, in order to simulate small variations within the dataset. In the next step the underlying data is characterized, using multivariate techniques or a supervised descriptor selection, followed by the application of the selection approaches. Subsequently, the collected measurements are used for model building and finally, these models are evaluated referring to distinct criteria, such as reliability, performance and robustness.

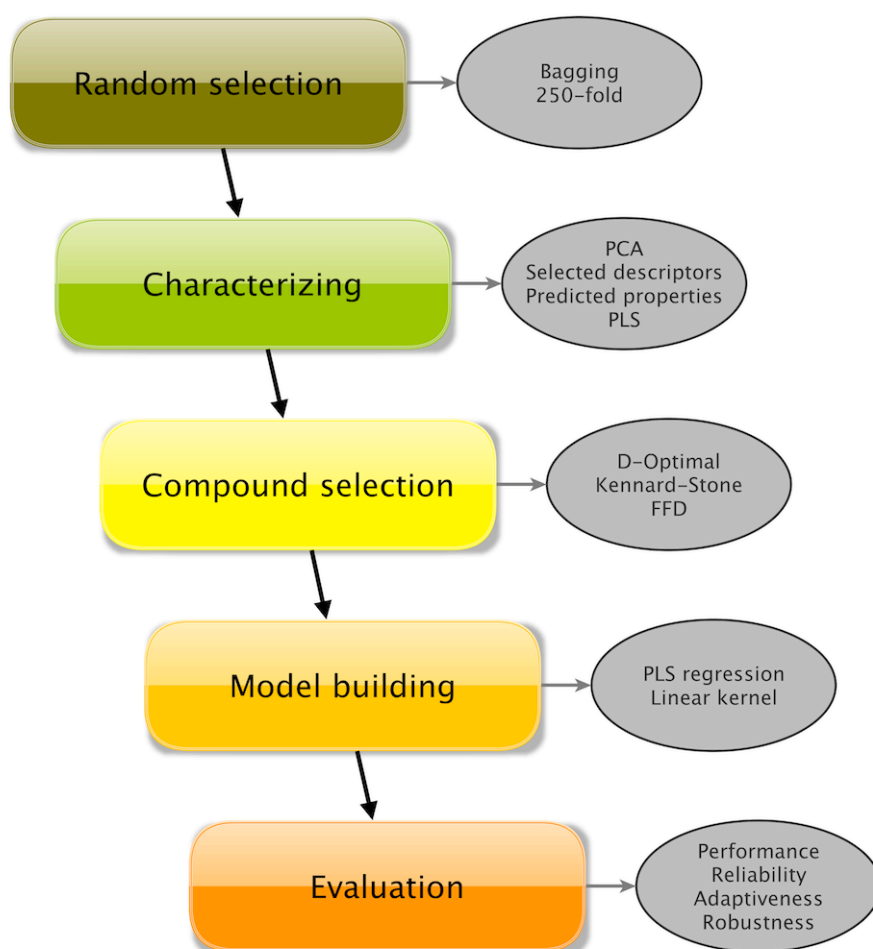


Figure 45. The validation pipeline contains of a random selection to enable a statistical evaluation of the observed results, a predefined representation of the compounds, an unambiguous modeling algorithm and several quality criteria.

3.5.1 Procedure

“All datasets (*both regression and classification*) were split into two partitions. The first partition (design set) was used to execute the selection approaches and the second partition was used as a respective validation set. To retrieve a statistically meaningful foundation to evaluate and compare the approaches, an ensemble of

these splits was generated. Each of these splits was used for the evaluation of each of the selection approaches.”[c]

If not otherwise stated the design set contained 84% of the compounds in the original collection and the validation set contained the remaining 16%. A split of that size was chosen as default value, as it guarantees a minimum overlap of two thirds in two arbitrarily chosen design sets. Furthermore, the default number of validation splits used in this thesis was 250.

“As a standard, for the logLC₅₀, the logBCF and the logK_{OC} dataset a number of 5, 7, 10, 15, 20, 25, 30 and 40 compounds were drawn in the predefined order. For the -logIGC₅₀ and the boiling point dataset, an additional number of 50 and 60 compounds were drawn. As the classification datasets contained a number of compounds, which was at least four times higher than the number of compounds in the largest regression dataset, 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 approaches on principal components) was started from scratch for each sample size. Contrary for the stepwise approaches, the selection process was strictly based on the sequence as mentioned above. Thereby the compounds selected in each previous step are in the next step used as a known seed, and the newly selected compounds, just extend this seed.”[e]

“To obtain comparable information about the quality of the compound selection, we used PLS to train a linear regression model on the selected compounds. The number of latent variables for the final model was determined in a five-fold cross-validation on all selected compounds using the coefficient of determination as criterion for the optimal number.⁴⁵ The reason why we chose PLS for evaluation of the final selection is the robustness of the method. As it uses a projection of the descriptors, it reliably finds linear correlations of the target property in the descriptor space. Furthermore, by taking the target property into account, PLS removes noise in the descriptor space. The cross and square terms we used to span the search space for the D-Optimal criterion were not used for development of the PLS models.

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

“PLS regression was also used for the classification dataset set and the retained continuous values were discretized into two bins.” [e]

3.5.2 Criteria

“In QSAR modeling and chemoinformatics the focus within the evaluation of a novel approach is often exemplified with a spot check on a particular dataset. Statistical evaluations, taking performance measures such as reliability and robustness of an approach into consideration are rare. Due to chance correlations, this can result in misleading conclusions about the applicability of an approach.”[c]

“The measurement of quality was the root mean squared error (RMSE), as well as the correlation coefficient and the Q^2 for the regression datasets. The balanced accuracy as well as the F-Measure was used to estimate the quality of the classification models.”[e]

3.5.2.1 Average performance

We use the mean values of RMSE, Q^2 and the correlation coefficient for all validations splits as a basic reference to assess the applicability of a certain approach. The mean values are an accessible reference for a first evaluation and enable a fast direct comparison of different approaches.

“The statistical significance of the different performances (derived by different approaches) is estimated according to a binomial test, using the binomial distribution corresponding to the number of models used in our study.”[e]

3.5.2.2 Reliability in performance

Given a first collection of ten samples, five of them labeled with the value ‘1’, the other five labeled with the value ‘0’ and given a second collection of ten samples, all of them labeled with ‘0.5’, the average value of both collections is 0.5. Nevertheless, the explanatory power of the average value is restricted. Whereas the assumption that the majority of values within a data collection is close to its average value is true for the second dataset, it is basically wrong for the first dataset. This is referred to as the level of uncertainty.

The level of uncertainty can be expressed by the standard deviation, which is 0.5 for the first dataset and 0.0 for the second dataset. It expresses the variance within the underlying data.

The same effect that the average value on its own has only limited expressiveness is observable for the observation of the performance on the n validation samples, we use. Therefore, we use the belonging standard deviation and standard error as measurements of reliability. The standard deviation gives information about the adaptability of an approach to the small variations in the dataset, resulting from the random sampling that was used to generate the validation splits.

Once again to clarify this, the standard deviation is only used as an indicator for the reliability, but we do not use it as an indicator for statistical significance.

3.5.2.3 *Steadiness in prediction*

Another important criterion is a steady decrease in the prediction error with the selection of additional molecules. The general assumption hereby is that the performance quality of a model increases with a higher number of compounds in the training set.

3.5.2.4 *Robustness against structurally diverse compounds*

The adaptability to small variations within a dataset is not the only indicator for a powerful experimental design approach, but also the ability to handle structurally diverse compounds. In order to investigate if this capability is ensured in the selection approaches, we compare the average model quality derived with the selection approaches applied to the datasets with and without the structural outliers, specified in chapter 3.3.

4 Results and Discussion

This chapter contains the results of five studies that were made on the newly developed experimental design approaches, as well as a practical application of these selection approaches.

Hereby, the first study, which focuses on the combination of the partial least squares technique with the D-Optimal criterion, acts as proof of the concept, that an iterative refinement of the representation of the chemical space can be beneficiary in terms of a representative compound selection.

In the second study, the paradigm of the adaptive approaches is generalized and shown to work for the combination of a similarity-based approach with selected descriptors as well. Furthermore, the approaches are also compared with various frequently used experimental design approaches.

The third study focuses on a cluster based selection approach, which is mainly examined as a static approach (non-adaptive), which selects all compounds at the same time.

The fourth study, on the other hand, emancipates of the descriptor-based interpretation of the chemical space and shows the benefits of approaches using predicted properties and to define the chemical space. Hereby the compounds similarity is not anymore defined by their pairwise Euclidean distance in the descriptor space, but by their correlation in the property space and the applicability domain estimation is used to prioritize them.

In the fifth study, all previously used approaches are extensively compared on the whole collection of regression datasets, which were used in this thesis. Hereby the comparison refers to the underlying data representation, as well as to the selection paradigm. Finally, the gained findings and insights are used to execute a purposive prioritization of compounds that are commonly used in chemical laboratories, respective to the required voltage to dissociate a chemical compound. As the phase of experimental testing is still in progress, the final results of this study are open.

4.1 PLS-Optimal: A proof of concept

The first study we carried out was intended to be a proof of concept, that property based experimental design approaches are appropriate to improve the performance of the resulting models. Thereby the focus was to show that a certain selection algorithm performs more efficient, if it is applied to an iteratively refined, property-correlated chemical space, instead of principal components.

We decided to select PLS-Optimal for this study and investigate its advantage over the classical D-Optimal criterion. The reasons to decide for the D-Optimal criterion as the selection criterion were: firstly its high acceptance within the scientific community; and secondly, the comparably simple adaption of the approach, to deal with preselected instances.

The study was limited to four datasets with regression, endpoints. Those endpoints were the boiling point, the $\log K_{OC}$, $\log LC_{50}$ and $-\log IGC_{50}$. "The initial $-\log IGC_{50}$ dataset contained more than one thousand compounds. However, to evaluate the performance of the developed approach on a relatively small dataset, a subset of 96 compounds was randomly selected.

Both the classical and stepwise approach were used to select a fixed number of compounds, which included 10, 20, 30, 40, 60, 80, 100, 130, and 160 compounds for the three large datasets ($\log K_{OC}$, boiling point, and $\log LC_{50}$) and 6, 10, 14, 18, and 21 compounds for the small dataset ($-\log IGC_{50}$)."[a]

As the main purpose of this study was to display the applicability of concept, we decided to use only 100 validation splits. Furthermore, to ensure a higher variability within those splits, we increased the size of the validation sets to 25%.

4.1.1 Comparison of the performance

4.1.1.1 Linear search space

"To compare the methods, we used three and six components alternatively for the three large datasets and two and four components alternatively for the $-\log IGC_{50}$ dataset." [a] The order of variables used to define the search space was assigned by the variance in the descriptor space for the PCA components and by the correlation to the target property for the PLS latent variables. "The performance of the developed models is shown in Fig. 46(a-d). The performance of the models built with PCA regression was significantly worse than that of PLS regression for all analyzed datasets. Therefore PCA regression results are no further provided.

The selection of these numbers of components for comparison is a reasonable one. The lower number of PLS or PCA components (3 or 2) shows the performance for a low dimensionality search space, which is particularly interesting regarding runtime requirements. The higher number of PLS or PCA components (6 or 4) adopts the OECD principles¹⁴⁹ regarding the number of descriptors to be used for a linear model.

There are several important observations. First, with an increasing number of selected compounds the model performance also improves. Second, with an increasing number of latent variables (or principal components for the traditional method) the performance of the resulting models also increases. This observation is particularly clear for the stepwise approach, with the exception of the logLC₅₀ dataset.

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

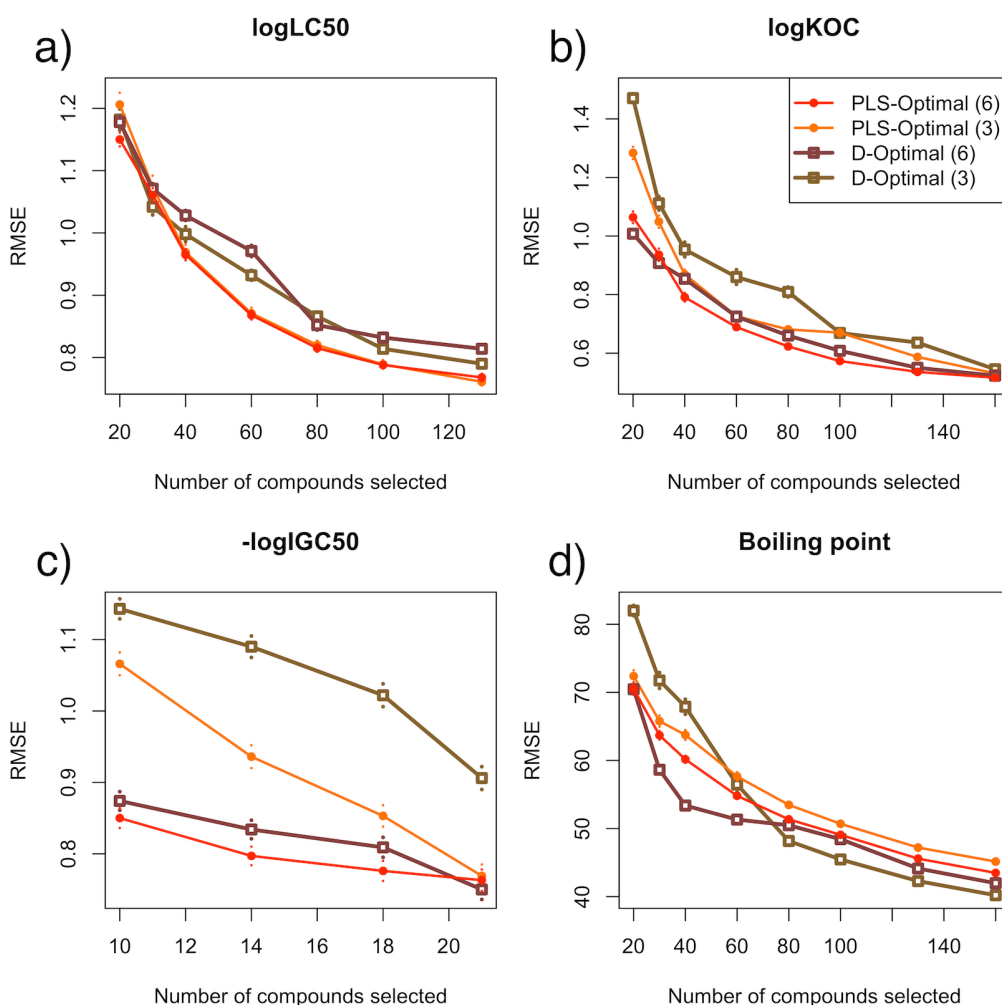


Figure 46. The average error on the a) logLC₅₀, b) logK_{OC}, c) -logIC₅₀, and d) boiling point datasets using a linear search space. The performance of the PLS-Optimal approach is shown in red (for six latent variables) and yellow (three latent variables); the performance of the traditional approach is shown in brown (six principal components) and green (three principal components). On the x-axis the number of compounds used to build the according 100 models is displayed. For the traditional D-optimal design (using principal components) all compounds were selected simultaneously. For the stepwise approach (using latent variables) the preliminary selected compounds were extended with new ones each cycle. The y-axis shows the average performance of RMSE. [a]

Let us take a closer look at the performance of the methods on the external validation split. It is clear that for all datasets, except for boiling point, error

decreases faster with the stepwise method. Further, using the stepwise approach a point of convergence, where the performance of the outcoming model no longer changes, is reached with a lower number of compounds. For the logLC₅₀ dataset (Fig. 46a), the logK_{OC} dataset (Fig. 46b), and the -logIGC₅₀ dataset (Fig. 46c), the performance of the stepwise approach is better than that of the traditional approach using the same number of latent variables or principal components and the same number of compounds selected.

This improvement is statistically significant with a p-value < 0.05 for 40 compounds and a p-value < 0.001 for the range 60 to 130 selected compounds for the logLC₅₀ dataset according to the binomial test (the binomial distribution with N=100 trials corresponding to the number of models used in our study). The sequential approach using 40 selected compounds and six latent variables provides the same accuracy of prediction as the traditional approach using 60 selected compounds.

The results for the validation on the logK_{OC} dataset are similar. The increase in performance derived with the sequential approach using six components in the range of 40 to 130 compounds selected is significant and on average 0.037 log units. For the same range of compounds, the increase of RMSE for three components is 0.079 log units.

For a search space of two dimensions, the performance of PLS-Optimal on the -logIGC₅₀ dataset is better with statistical significance (p < 0.001) for the whole range from 10 to 21 compounds (14%-30%). The greatest difference in the performance of the methods is found for 18 selected compounds: in 90 of 100 cases, the models built with the stepwise selection delivered a better result than those built on the traditional selection.

Comparison of the performance of both approaches on the boiling point dataset (Fig. 46d) reveals results that differ from those of the other datasets. The performance of the traditional approach using PCA components is better. In the case of six principal components used to define the search space, the incline in the error is steep for the first 40 compounds selected. Beyond that, until 100 selected compounds, there is almost no improvement in performance.

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

4.1.1.2 *Cross and square terms*

“The same calculations as for the linear search space were also performed for the search space using square and cross terms of the PCA or PLS components. For the traditional approach, the number of principal components was fixed to the same values as for the linear approach. The number of resulting meta descriptors was thus also fixed to 27 and 9 for the three large datasets and to 14 and 5 for the small dataset. In contrast, the number of PLS latent variables for the stepwise approach was automatically optimized for this calculation. The procedure for estimating the optimal number was the same as for estimating the number of components to

evaluate the resulting mode. We also tried to optimize the number of principal components in a similar way, regarding the error on reconstruction.^{192,193} However, our examinations indicated that the performance of the resulting models improves with any further principal component.

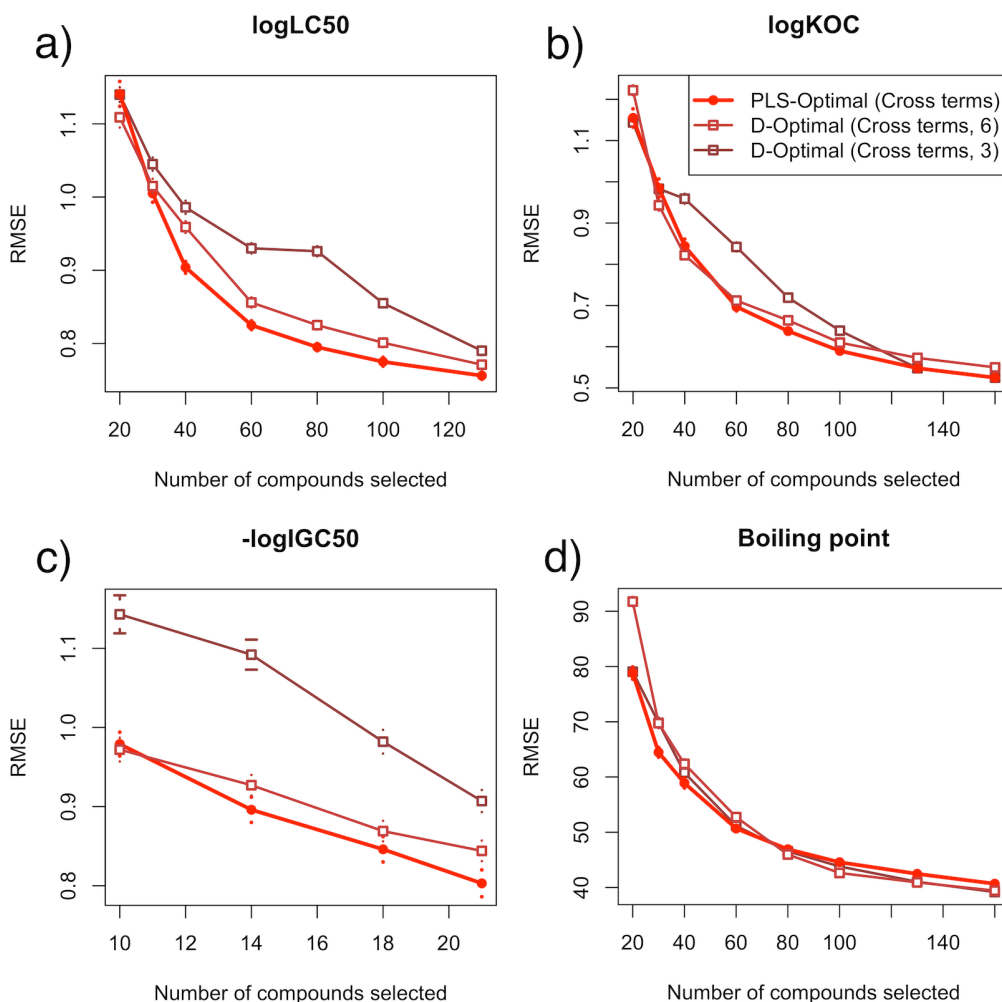


Figure 47. Development of the average error on the a) logLC₅₀, b) logK_{OC}, c) -logIC₅₀, and d) boiling point datasets using a search space extended by cross and square terms. For all endpoints, the bold red line represents the development of the stepwise approach; the development of the traditional approach on six principal components is represented by a dark red line and on three principal components by a red-brown line. [a]

The results for the validation on the square terms are shown in Fig. 47a-d. The axes are similar to Fig. 46. Similar to the linear search space, the performance of the resulting models improves with an increasing number of compounds selected. A further observation on all endpoints, except for the boiling point (Fig. 47d), is that the performance of the traditional approach improves with the number of principal components used. The selection performance on six principal components (or 4 for -logIC₅₀) is better than the selection performance on three (or 2) principal components for the whole examined range. Additionally, for six or four principal components and the use of cross terms and square terms, the development of the error describes a constant curve with a continuously increasing incline, without the inconsistencies observed for the linear search space.

Although the performance for six or four principal components converges with that of the stepwise approach, the models built on the compounds selected by PLS-Optimal are still better for most of the examined ranges on the logLC₅₀ dataset (Fig. 47a), the logK_{OC} dataset (Fig. 47b), and the -logIGC₅₀ dataset (Fig. 47c). For logLC₅₀, this improvement is significant from 40 to 130 compounds selected. The average error for 40 selected compounds is 6% lower for the selection derived using the stepwise approach. A model with better performance than that, derived from 100 selected compounds using the traditional approach, could be achieved with the 80 compounds selected with the stepwise approach.

For logK_{OC} the development is similar. After an almost similar performance on the first 40 compounds selected, the stepwise approach performs significantly better in the range from 60 to 160 compounds (12% - 33%). The average error within that range is 0.022 log units (3.5%) lower than for the traditional approach. The development on the -logIGC₅₀ dataset is almost analogous. After a similar performance for the first 10 selected compounds, the average error of the stepwise approach decreases more quickly than for the traditional approach on four principal components. We also evaluated the models built on the selected compounds on the 1000 compounds excluded from this dataset for this study and found the results to be similar.

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

4.1.1.3 *Decreased step-size*

“As the quality of the crossed traditional approach seems to increase with any additional principal component, it is interesting to take a look at the selection of only very few compounds. As it is a requirement for the D-Optimal criterion to work that the model matrix has more observations than variables, the number of components to be used is strictly limited. Therefore, on the three large datasets another examination within the range from 5 to 35 selected compounds was initiated. We used the meta descriptors containing the normalized components, their square products 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 one component for less than six selected compounds, two for less than 10, three for less than 15, four for less than 21, five for less than 28, and six components for less than 30 selected compounds.

The results in Fig. 48a –c show that the stepwise approach clearly achieves better performance for all three endpoints. This improvement is significant ($p < 0.001$) for the whole range from 10 to 35 selected compounds. In the case of the logK_{OC}

dataset (Fig. 48b) and for the range from 13 to 24 selected compounds, the stepwise approach performed better for more than 90 out of 100 splits.

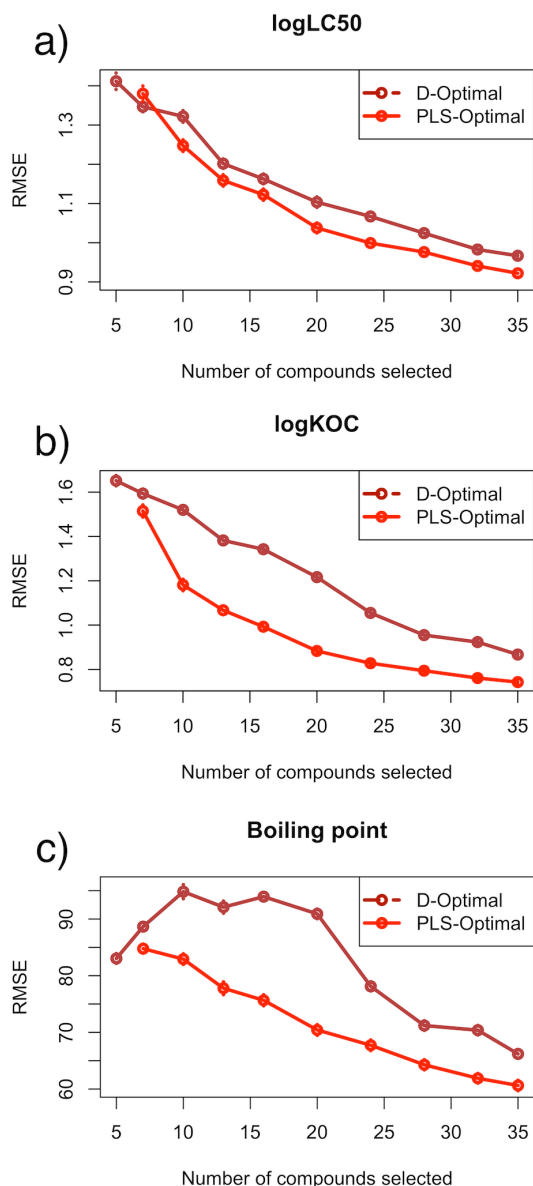


Figure 48. Results of the error validation for cross and square terms with a low number of compounds. [a]

Regarding the boiling point (Fig. 48c), the average RMSE performance for 24 compounds selected by the traditional approach could be achieved with only 13 compounds selected in a stepwise procedure. Furthermore, in the range from 13 to 32 selected compounds, the improvement of the average RMSE for the same number of selected compounds is better by at least 9 degrees. For logLC₅₀ (Fig. 48a), the average performance with 24 compounds selected in a stepwise manner could not be achieved with less than 32 compounds selected based on principal components. In the case of the logK_{OC} dataset, the stepwise approach delivers an average performance for 13 selected compounds that cannot be achieved with less than 24 compounds utilizing the traditional method. The RMSE for that dataset was on average 21% less in the range from 10 to 35 selected compounds.

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

4.1.2 Interpretation

“Our results, derived from examination of the PLS-Optimal performance on the logLC₅₀, logK_{OC}, and -logIGC₅₀ datasets within a range of 5% to 35% of compounds selected, show that the stepwise approach utilizing PLS latent variables can significantly increase the quality of the resulting model and thereby help to save resources. Compared to a model based on selection of compounds by the traditional D-Optimal design approach, the model derived from the same number of compounds selected using the stepwise approach delivered a decreased RMSE and an increased R² and Q² for both the linear search space and a search space extended by cross and square terms. The convergence of the error to a minimum was clearly faster and the improved performance can be observed in the whole range from approximately 10% to 30% of compounds selected.

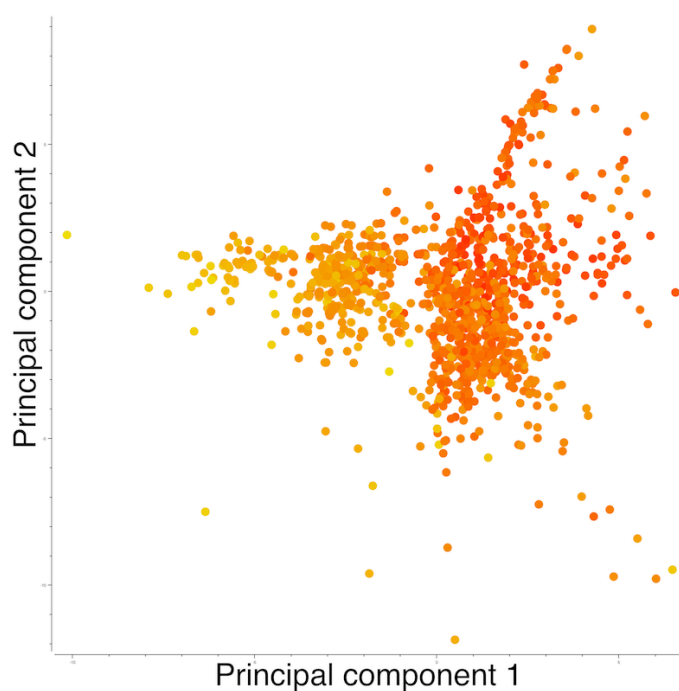


Figure 49. Compounds of the logLC₅₀ dataset in PCA space. The color of the data points represents the measured value. [a]

The performance on the boiling point dataset can be explained by the depiction, shown in Fig. 49, of the chemical space using the principal components. The x-axis represents the first principal component, the y-axis the third principal component, and the color of the data points displays the endpoint values. We can clearly see that not only the first principal component but also the third is strongly correlated

to the endpoint. Furthermore, the principal components are not just correlated with the endpoint; they are almost similar to the PLS latent variables, derived on the whole dataset. Tab.12 shows this correlation for the first PLS and PCA component. The correlation in the second and third dimension is comparable. While the PLS-Optimal approach tries in a stepwise manner to find a stable depiction and correlation, the PCA used for the traditional approach provides them exactly. As the boiling point is a very simple endpoint and widely cleared up, this effect was a foreseeable one. Nevertheless, it is a good depiction of the limitations of the developed approach and we suggest using the stepwise approach particularly for experimental designs for complex endpoints.

Table 12. The loadings and rank of five descriptors for the first PLS latent variable and the first principal component

Descriptor	PLS loading	Rank PLS	PCA loading	Rank PCA
SeaC2C3aa	0.506	1	-0.33	1
SaaCH	0.471	2	-0.321	2
SeaC2C2aa	0.332	3	-0.272	3
SsF	-0.273	4	0.167	8
Se1C3Cl1a	0.273	5	-0.249	4

The models built on PLS-Optimal design deliver a more stable performance regarding the error development for all four examined endpoints. Whereas with the classic approach the performance shows some variability and deviations with an increasing number of selected compounds, the performance development of the PLS-Optimal design is much smoother and approximates a hyperbolic function. This is observable even for a search space of only three variables.

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

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

We repeated the whole study with raw (non-normalized) and standardized descriptors, which resulted in a worse performance of the resulting models. The average error performance was worse and the development of the error was less stable for both analyzed approaches. We also compared the stepwise approach with the traditional one on other descriptor sets, i.e. ISIDA fragments³¹ and QNPR descriptors.³⁴ The results were similar and did not influence our conclusions.”[a]

4.2 DescRep: A generalization of the adaptive concept

The results of our first study support the theory that adaptive, property-oriented experimental design approaches result in a compound selection, which enables the building of significantly improved models, still, results were obtained for only one selection approach. Therefore, our second study aimed towards the generalization of the stepwise adaptive approach we introduced with the PLS-Optimal design. Furthermore, it was intended to give a more widespread overview of the performance of adaptive approaches in comparison to the variety of available static approaches.

We therefore decided to use a similarity based selection algorithm instead of a dissimilarity based one and combine it with a representation of the chemical space by selected descriptors. This approach is referred to as DescRep. Additionally we wanted to enable the estimation of the applicability of such approaches to classification datasets. Therefore we examined the performance of the selection approaches not only on the three regression datasets for logLC50, logKOC and the boiling point, but also on the AMES classification dataset. The dimensionality of the search space was fixed to five components for all datasets and approaches.

In contrast to the evaluation of PLS-Optimal, which mainly focused on the average error performance, we also investigated the quality of the resulting models referring to criteria, such as outlier robustness, stability and reliability. The parameterization of the validation pipeline was according to the default values.

4.2.1 Comparison of the results

4.2.1.1 Regression datasets

4.2.1.1.1 Model performance

“To enable a comparison of the quality of the models resulting from the examined selection approaches, we calculated the average RMSE performance and the average correlation coefficient for each number of compounds selected. Fig. 50a-c) shows the results of this comparison using the prediction error, whereas Fig. 50d-f) shows the comparison using the correlation. The x-axis displays the number of selected compounds and the y-axis the measurement of quality.

The first general observation on all of the datasets and selection approaches is that with an increasing number of selected compounds the average error decreases, whilst the average correlation in the models increases. This is expected as a larger number of molecules provide an increase in the amount of information obtained and thereby enables one to build a better model. Furthermore, for all datasets the stepwise approaches reach a good performance, which is constantly within the range of the best approaches.

PLS-Optimal reveals problems with the BP dataset, these problems were explained in our previous study with the similarity between the loadings of the PLS latent variables and the loadings of the principal components. The average performance

of models derived from compounds selected with DescRep is also the best for the boiling point.

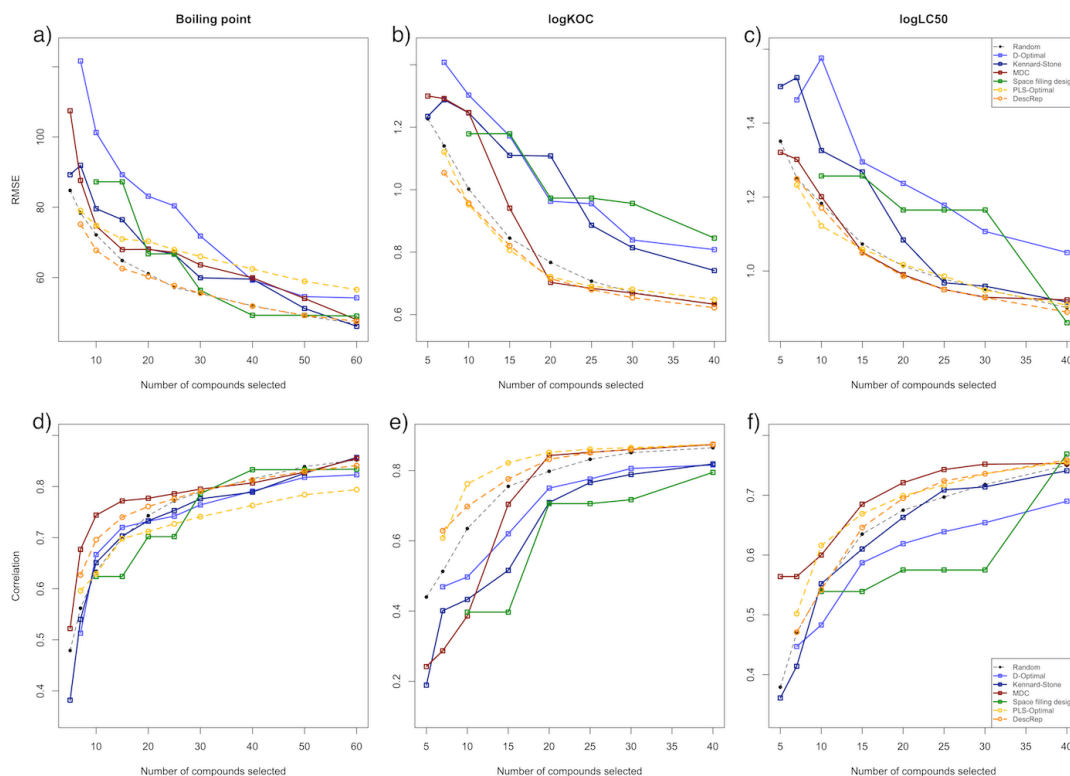


Figure 50. Average performance of the models resulting from the selections of the examined approaches, displayed as a-c) RMSE and d-f) correlation coefficient on the datasets for a, d) boiling point, b, e) logK_{OC} and c, f) logLC₅₀. The stepwise approaches are displayed by the dashed orange lines (DescRep) and the dashed yellow lines (PLS-Optimal). The color assigned to the random selection is black, red for the MDC selection, green for the space filling design and blue for the dissimilarity selections. [c]

A further observation is the smooth hyperbolic development of the average error performance on the 250 splits for each dataset. Whereas the static approaches result in unexpected deviations, there are no irregularities for the stepwise approaches, neither in the error, nor in the correlation development. MDC is the only systematic approach that derives selections resulting in a performance, which is as comparably good, although it reveals similar problems as the other approaches for the boiling point and the logK_{OC} dataset until 20 selected compounds.

The models derived from the selection of both stepwise approaches show a low initial prediction error. The performance of PLS-Optimal for seven selected compounds is better than e.g. than that of the D-Optimal criterion for 25 selected compounds on the boiling point dataset, for 15 compounds on the logK_{OC} dataset and on the logLC₅₀ dataset for 20 compounds. Further worth mentioning, is the good performance of models resulting from the random selection. Like the stepwise approaches, the random selection provides models that reliably decrease in average error and increase in average correlation for a growing number of compounds selected.

Regarding the correlation coefficient, MDC shows the fastest increase of all examined methods for the boiling point and the logLC₅₀ dataset. The models from the MDC selection on the logK_{OC} dataset, clearly show a worse initial correlation for less than 20 selected compounds. Although the convergence in the correlation for the stepwise approaches is not that fast, it works equally well on all datasets and it is still faster in comparison to all other systematic approaches.

Referring to the binomial test, we found that the observed improvements in the resulting models derived with DescRep are of high statistical significance ($p < 0.001$) for the range of 7 to 20 selected compounds for the boiling point dataset, 7 to 25 selected compounds for logK_{OC} and 15 to 40 selected compounds for logLC₅₀, when compared to the random selection. Regarding a comparison of PLS-Optimal with a random approach, we observed this level of statistical significance for the range of 10 to 25 selected compounds for the logK_{OC} dataset and 7 to 15 selected compounds for the logLC₅₀ dataset. Furthermore, DescRep performed better than MDC (the best static approach) with high statistical significance ($p < 0.001$) over the whole examined range for the boiling point and for 5 to 15 selected compounds for the logK_{OC} dataset.”[c]

4.2.1.1.2 Consistency and stability

“In addition to the average error, the reliability and stability in the performance of the resulting models have to be taken into consideration. We therefore calculated the standard deviation within the models of the 250 trials on each dataset, for each number of selected compounds, and for each selection approach. The results are shown in Fig. 51. The colors are identical to that of Fig. 50 and the y-axis displays the standard deviation, whereas the x-axis displays the number of compounds selected.

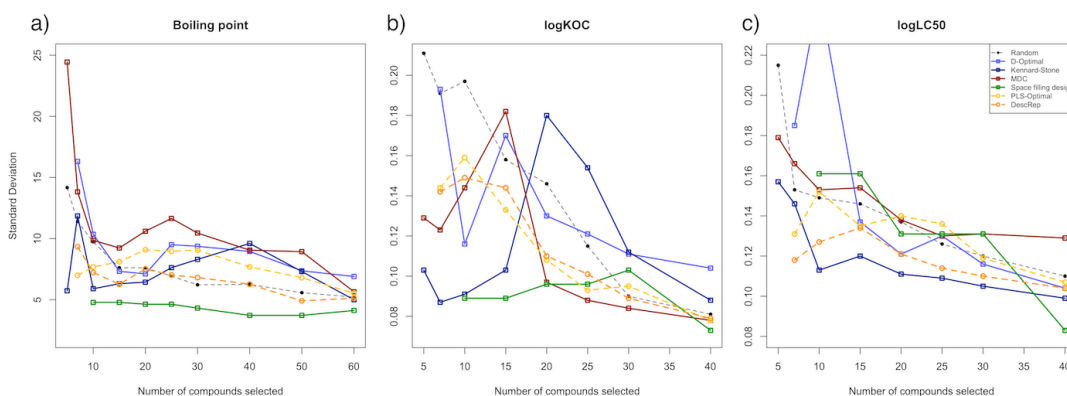


Figure 51. Comparison of the standard deviation of the selection approaches on a) the boiling point, b) the logK_{OC} and c) the logLC₅₀ dataset. [c]

The first general observation is that with an increasing average error the standard deviation also increases for most of the approaches. The exceptions are the models derived with the Kennard-Stone algorithm on the logK_{OC} dataset, as they show an increase in standard deviation by a factor of two for 20 compounds selected in comparison to 10 compounds selected.” [c] As this extraordinary behavior is affecting exclusively the static approaches, whereas the development of the

standard deviation of the stepwise approaches is as expected, it demonstrates the exceeding sensitivity of the all-in-one approaches to small variations in the dataset. This is likely to be reasoned by the distribution of the compounds in the $\log K_{OC}$ dataset, when represented with principal components (Fig. 21). Most compounds with high values for $\log K_{OC}$ are located on the right peak of the data distribution. However, the affected area of the descriptor space contains compounds with low $\log K_{OC}$ values as well. The exclusion of a certain compound with a high property value from the design might therefore consequence in the selection of a locally proximate compound, with a completely different value regarding the target property, which consequences in a decrease in the $\log K_{OC}$ range, taken into account for modeling.

“Regarding the random approach, the variations in the initial performance are high. This high level of uncertainty in the resulting models is why this approach is frequently found inappropriate, in spite of its reasonable average performance.

The space filling design has the lowest standard deviation for the resulting boiling point and $\log K_{OC}$ models, whereas the MDC approach, the only systematic method that could at least partially reach the same performance as the stepwise approaches, has a significantly higher standard deviation than DescRep on all datasets and for the whole range of selected compounds.

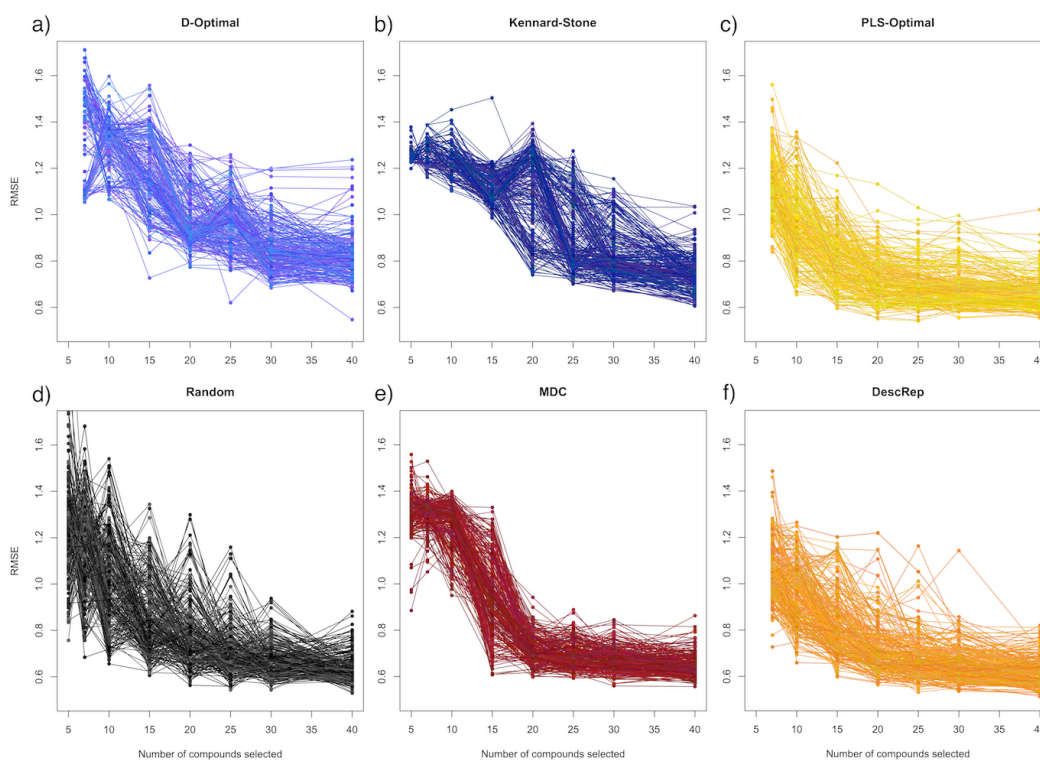


Figure 52. All 250 models on the $\log K_{OC}$ dataset. [c]

Fig. 52 provides a more detailed insight into the distribution of performance of the resulting models and the development of particular validation splits. It shows the RMSE development of all 250 validation splits on the $\log K_{OC}$ dataset for a) the D-Optimal criterion, b) the Kennard-Stone algorithm, c) PLS-Optimal, d) the random selection, e) the MDC selection and f) DescRep.

Both stepwise approaches produce only a small number of low performance outliers, whereas the majority of the validation splits results in models with quite similar performance. Additionally, for almost all splits, the initial performance of the resulting model is lower than for the other approaches and the error performance shows a fast convergence.

Furthermore, the error on the validation splits steadily decreases for a higher number of selected compounds. Especially for the dissimilarity approaches this is not the case, e.g. Kennard Stone selection delivers a worse model for 20 than for 15 selected compounds. And for the D-Optimal criterion these deviations of worse models for a larger training set are widely spread over the whole range of selected compounds.”[c]

4.2.1.1.3 Outlier robustness

“All calculations were repeated with the extended sets, each containing a structural outlier. To compare the effects of such outliers to models derived by the selection approaches, we determined the difference in the average RMSE between the sets without and the sets with outliers. The results are shown in Fig. 53. The colors are in accordance with all previous figures, and the y-axis displays the difference in average performance. Approaches that result in models with a better performance on datasets with structural outliers, have positive values, those performing better on sets without structural outliers, have negative values.

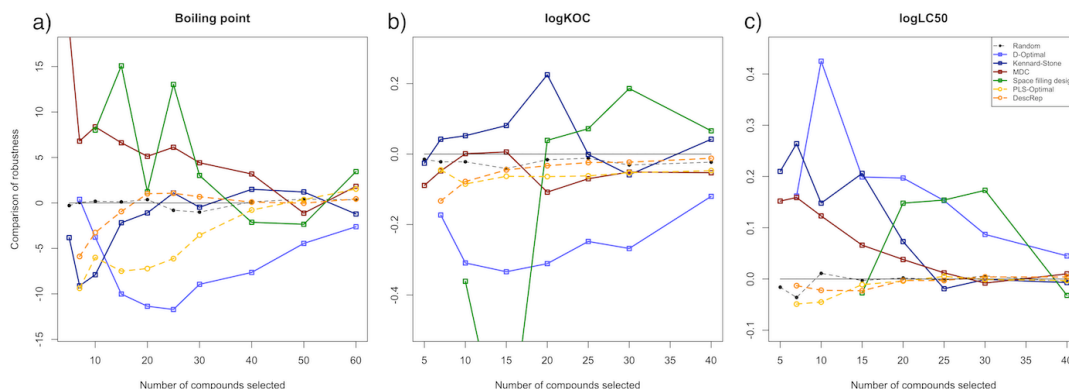


Figure 53. Effects of the structural outliers to the selection approaches to the examined datasets, displayed by the difference in average RMSE performance. [c]

Both stepwise approaches show only small deviations in the resulting models. Apart from an initial better performance of PLS-Optimal on the boiling point dataset without structural outlier, the selections derived with the adaptive approaches perform equally well on the extended datasets. Also the MDC selection is mostly resistant to the outlier, whereupon a tendency to deliver better selections on datasets with outliers is observable.

Contrary, the effect of only one additional compound on the other approaches was incalculably. The models derived with the space filling design, the D-Optimal criterion on principal components and the Kennard-Stone algorithm, have no clear tendency towards the original or the modified dataset. The sign of the difference in

the average error of the resulting models differs from dataset to dataset. This is also the case for the space filling design, even within the logK_{OC} dataset.”[c]

4.2.1.2 Classification dataset

“The examination of the influence of structural outliers in a classification set was not considered in our study as we could not find a structural outlier in the dataset with such a strong influence on the principal components. Furthermore, the size of the dataset made it impossible to investigate the performance of the MDC selection, as the runtime requirements exceeded more than one hour for one sample. We therefore concentrated on the prediction quality of the models resulting from the remaining selection approaches. The measurement of quality therefore is the balanced accuracy and the deduced standard deviation as a measurement of uncertainty.

As shown in Fig. 54, the random selection reveals the best performance, combined with a low level of uncertainty. This is due to the relatively big size of the dataset. The number of selected compounds therefore had to be larger, which resulted in a significantly decreased probability of selecting an adverse and clearly unrepresentative combination of compounds with the random approach.

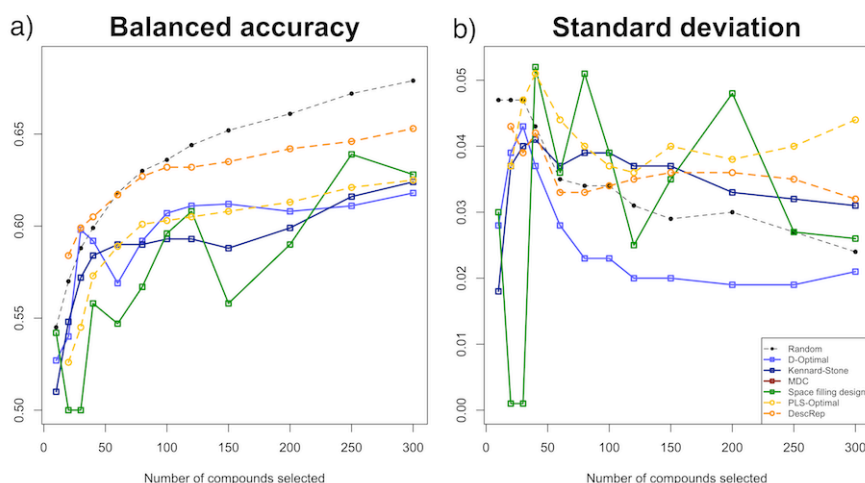


Figure 54. Performance on the classification dataset, showing a) the average F-measure and b) the according standard deviation. [c]

Initially the performance of DescRep is equally good, but starting from 120 selected compounds the balanced accuracy is lower, and starting from 100 selected compounds, the standard deviation is higher, when compared to the random approach. A direct comparison of the performance on the 250 splits resulted in the finding that this difference in performance is not statistically significant until 150 selected compounds. Further worth mentioning is the good initial performance of DescRep for 20 selected compounds, which is better with high significance than the performance of all other approaches for the same number of selected compounds.

The performance of PLS optimal is comparable to that of other systematic approaches and significantly worse than that of the DescRep approach. By example, the models resulting from 300 compounds selected with PLS-Optimal, the space filling design, the Kennard-Stone algorithm or the D-Optimal criterion, had a similar average prediction performance as the models derived with DescRep selecting only 60 compounds. Furthermore, the development of the error curve of all approaches, except the random selection and DescRep, shows numerous inconsistencies and irregularities.”[c]

4.2.2 Interpretation

“Both stepwise approaches: DescRep and PLS-Optimal, performed equally well on the analyzed datasets. The error performance of their resulting models is in general lower than that of the approaches that select all compounds at the same time. The development of the error is smooth and reliable. Both methods reveal a lower standard deviation compared to MDC, which is the best performing non-stepwise approach. The average correlation coefficient develops in a similar way. Neither on the logLC₅₀ dataset, nor on the logK_{OC} dataset any of the classic approaches was performing better than the stepwise approaches and on the boiling point dataset, none of the classic approaches performed better than DescRep.

This good performance can also be observed in the depiction of the specific models in Fig. 52. At large, for both stepwise approaches an increase in the number of selected compounds results in a decrease of the error. This is not the case for the Kennard-Stone algorithm and the D-Optimal criterion where high variations in performances were observed.

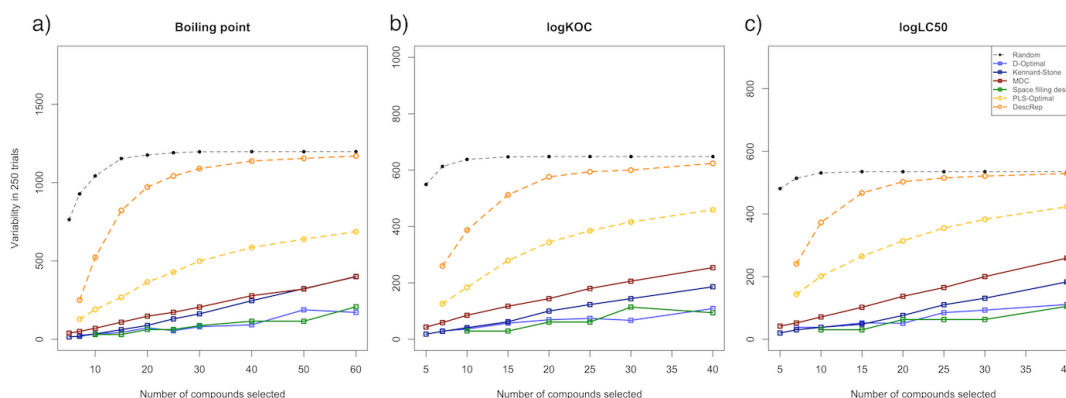


Figure 55. Variability in selection. [c]

Overall, DescRep is superior over the PLS-Optimal approach, as it was able to deliver high quality performance models even on the dataset where the performance of PLS-Optimal was not ideal. Nevertheless, the decrease in the performance accuracy of PLS-Optimal on the boiling point and the classification dataset can be easily explained and is therefore avoidable. The boiling point dataset resulted from a correlation between the PLS components and PCA components. In the case of the classification dataset, the choice of a dissimilarity

selection was quite inappropriate as the D-Optimal criterion selects compounds at the periphery of the chemical space which are most likely furthest away from the separating plain. It is important that DescRep is not affected with such problems.

To investigate the major difference between the stepwise and the non-stepwise (static) approaches, we analyzed the compounds selected by the different methods and compared their distribution in the design sets. To compare the variability in the selections of the methods, we counted the number of different compounds selected in the 250 trials. We found a significant difference between the stepwise and the static approaches. Whereas the systematic approaches, which select all compounds at the same time, have a comparably small pool of compounds that are selected, the stepwise approaches are resulting in a higher variety of selected compounds. This variability is shown in Fig. 55.

The stepwise approaches have a better adaptability to small variations in the datasets. The observance that PLS-Optimal has a lower variability in selection than DescRep is coherent as the D-Optimal criterion also has lower variability than MDC. Still, the variability of DescRep is significantly lower than that of the random approach. This shows that the selection process is still systematic and contributes to better performance of DescRep compared to random selection.

It is interesting to note that despite step-wise approaches have a higher variation in the number of selected compounds, the models developed with these compounds have lower variation compared to those developed using static approaches. The contradiction clearly indicates that the variability in selected compounds in both stepwise approaches is a meaningful adaption to changes in the dataset. Whilst the variation within the selected compounds is clearly increased for the MDC approach compared to the stepwise approaches, the resulting models show a significantly higher standard deviation than the stepwise approaches.

Additionally, not only referring to the adaption of small variations in the dataset, but also in terms of outlier adaption, the stepwise approaches show a convincing performance. The average error of the resulting models is similar with or without an outlier. The influence of structurally diverse compounds is only minor, when compared to the changes in performance for the static approaches.

We repeated all calculations with design sets of different size (66% and 75% of compounds) for all datasets and found no significant difference to the results presented in this study.”[c]

4.3 k-Medoid: An improved static approach

Our previous studies clearly indicate a better performance of models derived with adaptive approaches. Still, there are design problems in applied chemistry, for which the application of a stepwise or sequential measurement procedure is absolutely inappropriate. This applies, for example, for endpoints that require experiments with comparably low financial expenses, but which are highly time-consuming. Accordingly, there is a need for efficient static approaches as well.

In our third study, we investigated the use of a cluster based experimental design, utilizing the k-Medoid partitioning. We compared this approach to other static experimental design approaches and examined the performance of models derived from their selection for the logLC₅₀ dataset, the logK_{OC} dataset and the data collection on boiling point.

Once again, in terms of efficiency, we used only 100 validation splits with validation sets containing 25% of the available data. "All approaches were used to select a fixed number of compounds, namely 10, 20, 30, 40, 60, 80, 100, 130 and 160. For each of the datasets and for each selection approach, the compound selection was achieved on three, five and seven latent variables." [b]

Additionally, we used the logBCF for the validation of the evaluation of the k-Medoid approach, but instead of a representation of the dataset with ALog_{PS} descriptors and E-State indices, we decided to use the five descriptors that were suggested for optimal modeling by Gramatica et al.¹⁰⁰

"For this dataset, two hundred and fifty splits were generated, each containing 90% of the compounds to be used in the selection process and 10% excluded from it. Furthermore, the number of compounds selected was fixed to 5, 8, 10, 15, 20, 30, 40, 50, 75, 100, 125, 150 and 180; the selection of the compounds was performed in a search space containing four principal components, derived from the five descriptors.

The evaluation of the performance on this dataset took not only the predictions on the compounds within the external validation set into consideration, but all compounds that were not selected by the applied approach and which were thereby not used for the training of the model. Further, the model for this dataset was not built with PLS, but from a multiple linear regression on the five descriptors used for the published model." [b]

4.3.1 Comparison of the performance

4.3.1.1 Performance of the collections logLC₅₀, logK_{OC} and boiling point

"Fig. 56a-f) shows the results of the validation on the selection approaches for the logLC₅₀ dataset (a-b), the logK_{OC} dataset (c-d) and the boiling point dataset (e-f), including the standard errors as dots. The number of principal components used in each column is three (left column) and seven (right column). An exception therefore is the random selection, as this is independent of the number of latent

variables and the space-filling design, which was performed on a fixed number of three latent variables. To provide a statistical reference, for each of the datasets, a PLS model was built using a five-fold cross validation. These reference models provided an RMSE of 0.77 for $\log LC_{50}$, 0.52 for $\log K_{OC}$ and 35°C for the boiling point.

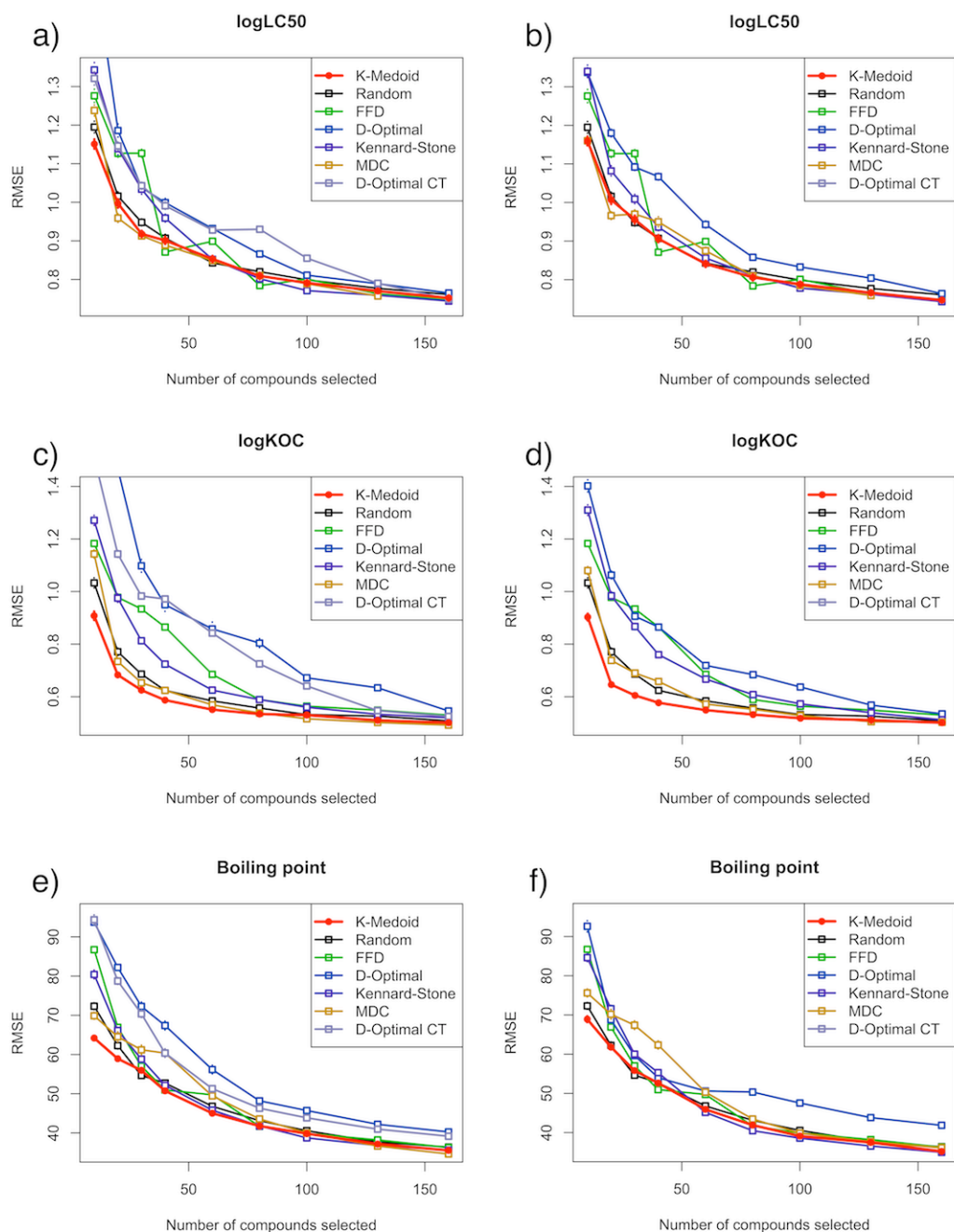


Figure 56. Performance of the approaches on the $\log LC_{50}$ dataset, the $\log K_{OC}$ dataset and the boiling point dataset, with three and seven latent variables. k-Medoid is displayed in red, the space-filling design in green and Kennard-Stone in purple; the D-Optimal criterion is blue or bright blue for a linear or a quadratic search space, and the MDC approach is marked in yellow. The random selection is shown by the black curve. The x-axis represents the number of selected compounds and the y-axis the average RMSE of 100 trials. [b]

To investigate the underlying chemistry, we analyzed, which descriptors spanned the first and the second PLS-component. Tab. 13 contains these descriptors and their loadings for all three datasets. The first component for the logLC₅₀ dataset is highly correlated with lipophilicity whereas the second indicates the molecule size as highly important. A similar mechanism can be seen for the logK_{OC} dataset. Also for the boiling point dataset the second component was mainly correlated with the length of the carbon chain, but for the first component mostly polar properties are crucial.

Furthermore, the performance derived from a selection on the D-Optimal criterion using cross and square terms was examined for three latent variables only. 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-Medoid approach in this section were assigned randomly.

The first observation is that with a greater number of selected compounds the model performance also improves. This observation applies for all approaches and it is expected, since a larger number of molecules allows better model development. A second observation is that for all datasets the random selection performs well. At first glance this may seem surprising but two facts have to be taken into consideration. First, the shown error is an average value and although this average value is quite well, single values can provide clearly worse performances. And second, the three datasets used for the evaluation are collections without chemical review. Their composition is thus not restricted to a specific class of chemical compounds.

Furthermore and of greater interest, for all datasets for the whole range of compounds selected and for each number of latent variables used to define the search space, the performance of the k-Medoid approach (shown by the bold red curve) is among the best. Compared to all other approaches, this approach shows the best initial performance and a rapid decrease of the error.

Especially for a small number of selected compounds, 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 statistically significant. For 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 ten selected compounds is, for all six examples derived from the selection of the k-Medoid approach. In the range of 80–160 selected compounds, the performances of the models derived by the k-Medoid approach, the random selection, the MDC approach and the space-filling design converge and are comparable.

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 the D-Optimal design. The development of the performance of

the k-Medoid approach is smoother and approximates a hyperbolic function, regardless of the number of principal components used or the dataset.”[b]

4.3.1.2 Performance on the logBCF dataset

“For the validation of the performance on the logBCF dataset, we compared the k-Medoid approach to a random selection, the Kennard-Stone algorithm, a space-filling design, MDC and the D-Optimal criterion. As the dependency between the endpoint and the descriptors is known to be linear, use of cross and square terms is not required. Fig. 57 shows the results of the validation.

The first observation is that, compared to the other datasets, the decrease in the error for all approaches is faster. The point of convergence, where the quality of performance no longer increases significantly, is reached between 20 and 30 selected compounds for all approaches except the space-filling design. This results not only from the increased set size, but also from the linear correlation between the descriptors and the endpoint.

The area of interest for experimental design is thereby the range from 5 to 30 compounds selected. Comparison of the models derived using the different selection approaches within that range reveals converse results to those on the other datasets, as all approaches performed significantly better than the D-Optimal criterion for logK_{OC}, logLC₅₀ and boiling point. For the logBCF dataset, the D-Optimal design delivers the best initial performance for five compounds selected and a fast decrease of error. Furthermore, Kennard-Stone and the space-filling design, which delivered good results for the other datasets, are significantly worse than the other approaches in the range from 40 to 120 compounds selected. The random selection, like the MDC approach, shows a high average error for 10 or fewer compounds selected, and the standard error within that range of selected compounds is high, compared to the other approaches.

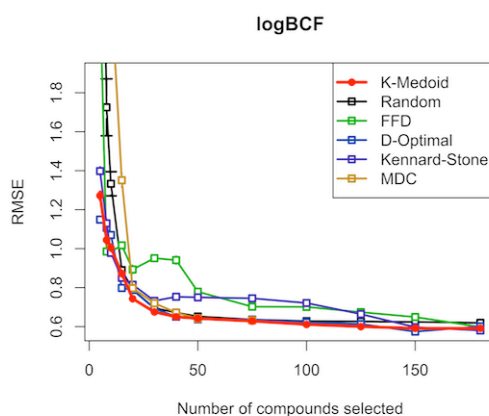


Figure 57. Performance of the approaches on the logBCF dataset. Colors and axes are as in Fig. 56. [b]

Only the k-Medoid approach performs equally as well on the logBCF dataset as on the other datasets. It delivers the second-best initial performance and reaches an average RMSE, which is lower than that derived with the D-Optimal design, for

seven compounds selected. Furthermore, for this dataset, too, the k-Medoid approach shows a very smooth development of the error without any deviations.”[b]

4.3.1.3 Initial cluster centers

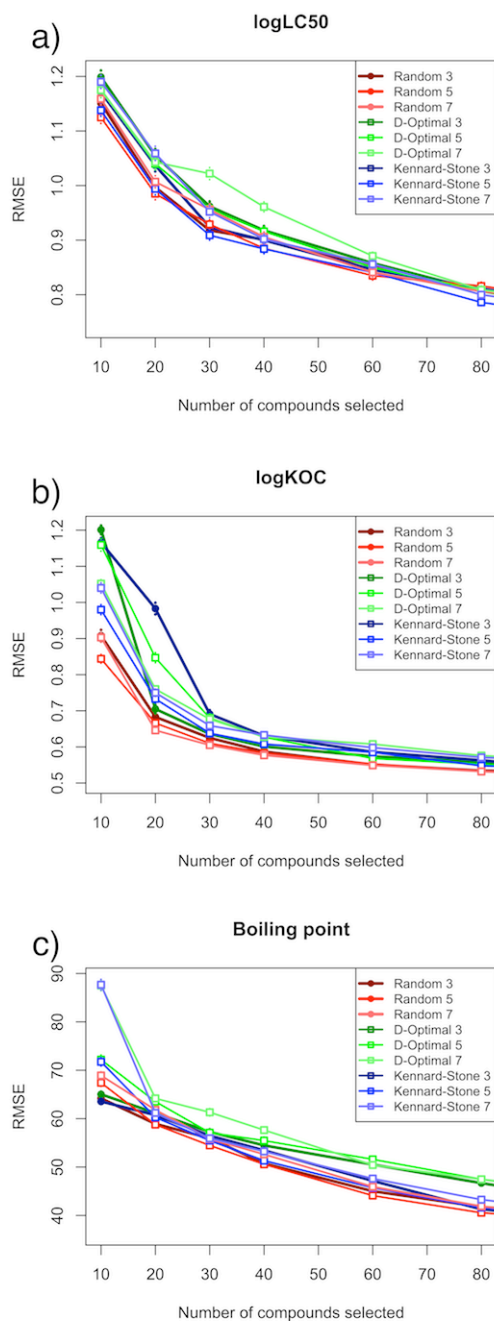


Figure 58. Comparison of the performance of the k-Medoid approach with different initial cluster centers on three, five and seven latent variables. The 15-fold random approach is shown in red, the approach utilizing the D-Optimal criterion for the initial selection is in green and Kennard-Stone is in blue. [b]

“To investigate whether and how the selection of the initial compounds influences the resulting selection and thereby the performance of the derived model, we

compared the results of the standard approach using the 15-fold random approach to results derived with a fixed initial selection of compounds. To select these initial compounds, we used the Kennard-Stone algorithm and the D-Optimal criterion. The performance was calculated on the logLC₅₀ dataset, the logK_{OC} dataset and the boiling point dataset for three, five, and seven latent variables. The results of this examination can be seen in Fig. 58.

Our results show that contrary to other approaches, the number of latent variables used for the search space does not influence the quality of the resulting model. The average RMSE does not show any statistically significant difference regarding the number of principal components used except for the boiling point dataset and the logLC₅₀ dataset, using seven latent variables, and the D-Optimal criterion for the initial selection. Taking the initially selected cluster centers into consideration, the results are not so clear. While the performance seems to be independent of the initial selection for the logLC₅₀ dataset, the best performance for the logK_{OC} dataset is achieved by the random selection. Regarding the boiling point dataset, both the random selection and initial selection derived with the Kennard-Stone algorithm work equally well.”[b]

4.3.2 Interpretation

“The k-Medoid approach for the experimental design was the only one that performed equally well on all datasets. We repeated the whole study with raw (non-normalized) and standardized descriptors, which resulted in a decreased performance of the resulting models. Both for the collections on logLC₅₀, logK_{OC} and boiling point and on the dataset for logBCF with the linear dependency between descriptor space and endpoint its performance was always among the best. Whereas the D-Optimal criterion worked best for the logBCF dataset but displayed a comparatively weak performance on the datasets with no explicit linear correlation, and whereas approaches like MDC or FFD showed a good performance on the three larger data collections but a poor initial performance or a discontinuous development of the error on the logBCF dataset, k-Medoid always displayed a fast and smooth decrease of the error, and a good initial performance. Compared to the models resulting from the selections derived by other approaches (i.e. MDC, D-Optimal, Kennard-Stone), models resulting from the selection derived by the k-Medoid approach could often reach the same performance with 40% less compounds used for training. This can be seen best in case of the logK_{OC} dataset and makes the k-Medoid approach the favorable one from an economic point of view. Furthermore, k-Medoid was the only approach that performed better than the random selection on all datasets.

To investigate the advantages on the level of PLS modeling, we executed the selection algorithms on the whole logK_{OC} dataset to draw a sample of ten compounds.

Each selection was then used for PLS modeling and the first two PLS components were examined regarding the contribution of the eight descriptors, most important for the logK_{OC} dataset. Fig. 59a-e) shows the loadings of these descriptors. The loadings derived from the model on the whole dataset are indicated black, all other

colors are accordingly to the Fig. 56-58. The grey arrows display the shift in the loadings and descriptors encircled in red indicate those, that did neither contribute to the first nor the second PLS component.” [b]

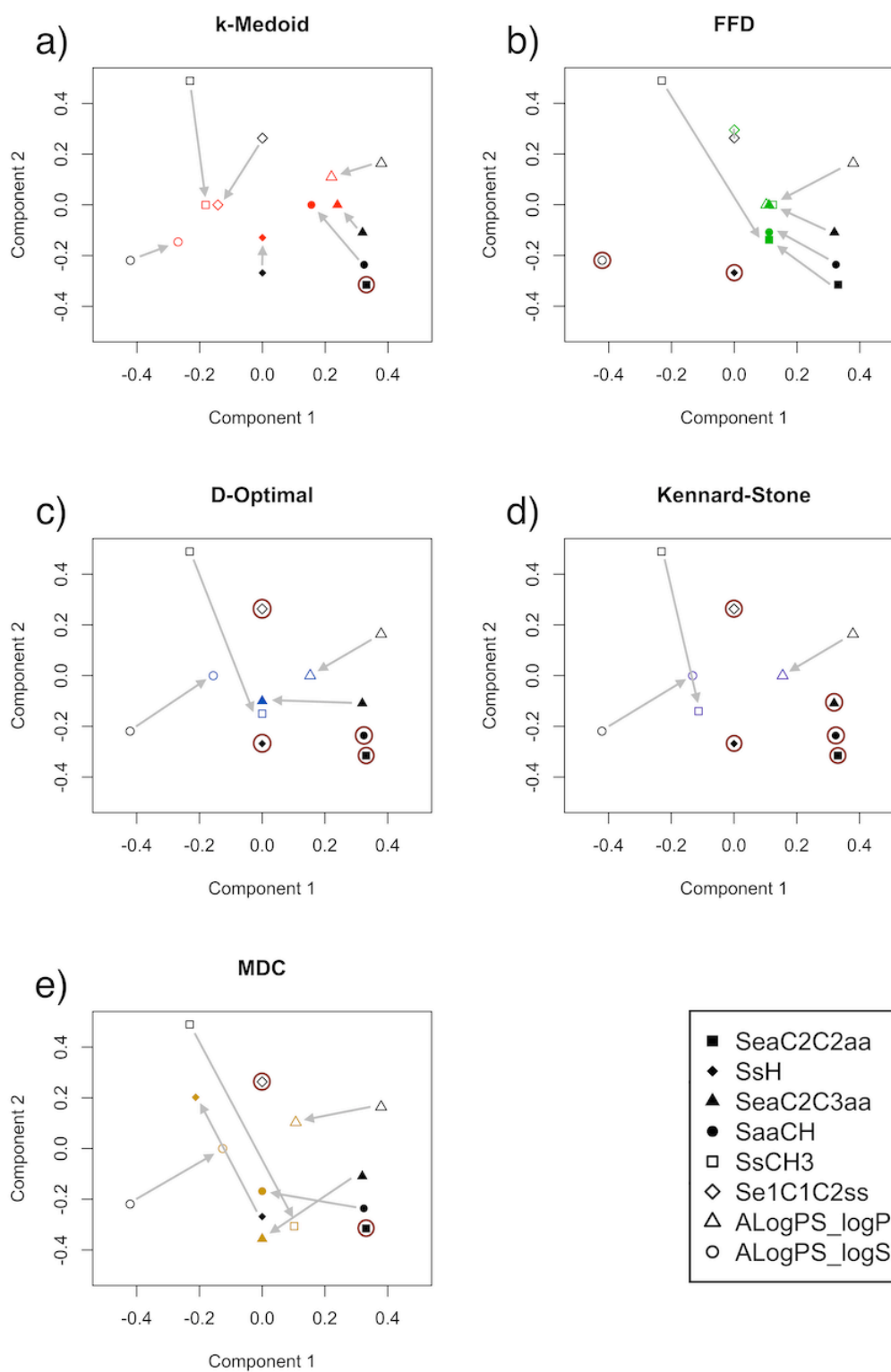


Figure 59. Shift in the PLS-loadings for the eight most important descriptors for a selection of ten compounds, using a) k-Medoid, b) full factorial design, c) D-Optimal design, d) Kennard-Stone algorithm, e) MDC selection on the logK_{OC} dataset. [b]

The closer dots connected by an arrow are and the lower the number of encircled black dots is, the closer the model resulting from a certain selection approach is to the model derived on the whole dataset. “The results show, that even for a number of only ten compounds selected, the k-Medoid approach finds the most relevant correlations from the reference model on the whole dataset. Contrary to all other approaches, the k-Medoid selection delivered a model, with only one of the relevant descriptors disregarded. Furthermore, the shift in the loadings is lower than for all other approaches.

Aside from this, the performance of the k-Medoid selection showed no statistically significant difference for a search space of three, five or seven principal components, if for the selection of the initial cluster centers a 15-fold random approach was used. Thus it is less dependent on an appropriate number of latent variables than other approaches. For the initial selection of the cluster centers with the Kennard-Stone algorithm or the D-Optimal design and for a small number of compounds selected, the problem occurs, as mentioned in the introduction, that both approaches work as ‘outlier detectors’ for a higher dimensionality of the search space. This can be best seen in the performance on the boiling point dataset, where the initial selection of the cluster centers with both approaches leads to a significantly decreased performance for seven principal components and ten compounds selected.

Table 13. The five most important descriptors and their loadings for the two main PLS-components.

	Component 1		Component 2	
	Descriptor	Loading	Descriptor	Loading
logLC₅₀	AlogPS_logS	0.524	SdO	0.360
	AlogPS_logP	-0.431	SaaCH	0.341
	SeaC2C3aa	-0.362	SssCH2	-0.307
	SaaCH	-0.280	SeaC2C2aa	0.294
	Se1C2H1a	0.232	SeaC2C3aa	0.282
logK_{oc}	AlogPS_logS	-0.421	SsCH3	0.489
	AlogPS_logP	0.379	SeaC2C2aa	-0.315
	SeaC2C2aa	0.331	SsH	-0.268
	SaaCH	0.324	Se1C1C2ss	-0.263
	SeaC2C3aa	0.319	SaaCH	-0.236
Boiling point	SeaC2C3aa	0.506	Se2C3O1s	0.419
	SaaCH	0.471	SdO	0.335
	SeaC2C2aa	0.332	Se1C2H1s	-0.267
	Se1C3Cl1a	0.273	SssO	0.254
	SsF	-0.273	Se1C4F1s	0.249

A reason for the good performance of the k-Medoid approach is that it combines the advantages of the three basic ideas and it minimizes their disadvantages:

- Like a space-filling design, it covers the whole chemical space, but one corresponding to the real distribution.
- Like approaches based on selection of the most distinct compounds, each point in the periphery of the data cloud is represented in a cluster.

- From each cluster the most representative compound is selected, as the criterion of the minimum distance is applied.

It is further worth mentioning that the k-Medoid approach is not subject to restrictions as the other approaches. Whereas for the space-filling design it is impossible to fix the number of compounds finally selected, with the k-Medoid approach the resulting number of compounds can always be precisely defined. Unlike MDC, k-Medoid has no stop criterion, and even a small number of compounds can be selected from a high dimensional search space. This 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.

Regarding the examination on the published QSAR set for logBCF and the usability of the examined approaches as tools to define the split between the training and validation set, the D-Optimal approach clearly works best for a small number of compounds selected; however, it is not an appropriate approach for this purpose. The D-Optimal criterion works by selecting those compounds that are the most distinct regarding the descriptor space, and these are the compounds that usually occupy the minimum and maximum values on the axes of the descriptor space. As the dependency between the target property and the selected descriptors is known to be linear, the compounds with extreme values regarding the descriptor space also exhibit extreme values regarding the target property. Therefore the usage of the D-Optimal criterion for the split into test and validation set will lead to an unbalanced selection of only compounds with maximum and minimum values for the target property.”[b]

4.4 Ensemble based approaches: Using the applicability domain

The three previous studies worked with a representation of the chemical space which was basically defined by descriptors. Although this is the most commonly used concept of illustrating chemical similarity between molecules, it surely is not the only one.

Our fourth study focused on selection approaches detaching from the paradigm of a descriptor based view on chemical compounds. Therefore, we used concepts obtained from the ensemble based applicability domain estimation to represent the correlations between different molecules. We developed and examined AD-Spider and AD-Fetcher two different approaches, aiming to detect and select compounds of high statistical relevance. Furthermore, we investigated the benefits of a combination of classical selection approaches with a chemical space that is defined by predicted properties.

We statistically evaluated these approaches and compared them to non-adaptive approaches, mainly focusing on the k-Medoid approach, as this turned out to be the most efficient static selection algorithm. Therefore we used four regression and two classification datasets. The considered endpoints were logBCF, logK_{OC}, -logIC₅₀, and the boiling point, as well as the AMES mutagenicity test and cytochrome inhibition. The parameterization of the validation pipeline once again was according to the default values.

4.4.1 Comparison of the performance

4.4.1.1 Regression datasets

“Fig. 60 shows the average error performance on the non-selected compounds for a) the logBCF dataset, b) the logK_{OC} dataset, c) the boiling point dataset and d) the -logIC₅₀ dataset. 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 with the dashed black line, the k-Medoid approach with the blue lines and the Kennard-Stone approach with the green lines. Referring to the underlying data, the approaches have been executed on, the coloring of the lines is dark green or blue for the static approaches on principal components derived from descriptors or bright green and blue for adaptive approaches on principal components, derived from the predicted properties. The coloring of the AD-Fetcher approach that works only on the distance to model is yellow and for the AD-Spider approach, which additionally takes the pairwise correlations between the predictions into account, the coloring is red. The same coloring was used in all further figures.

The first observation that can be deduced is that for all approaches and for all datasets, the average error is decreasing with an increasing number of selected compounds. A second observation for all datasets is that the performance of the Kennard-Stone algorithm on principal components, derived from descriptors, delivers the worst models. Although the use of principal components on predicted properties improves the performance, it is still significantly worse than most other

approaches. What is further noteworthy is the high 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 other approaches is smoother, approaching a hyperbola function.

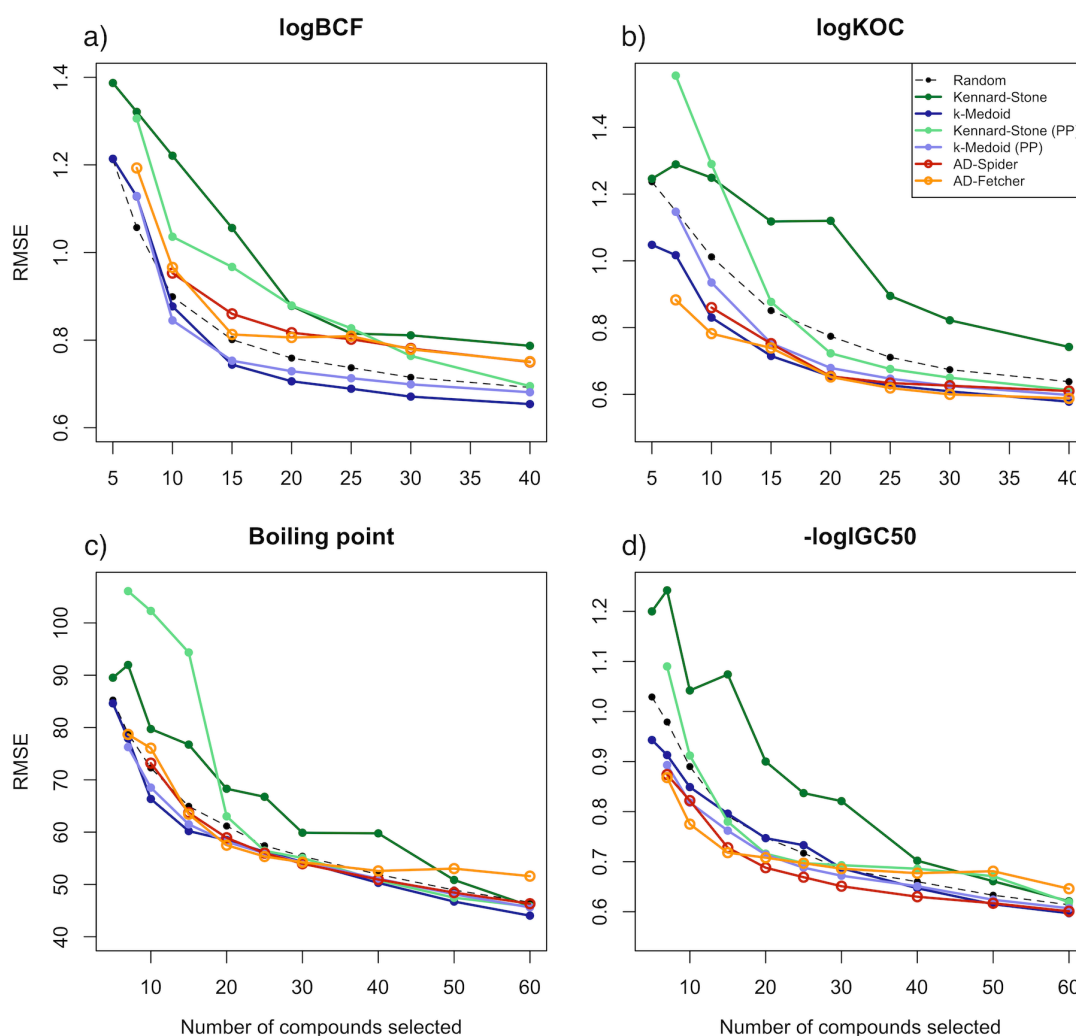


Figure 60. Comparison of the average RMSE performance on the 250 splits for the regression datasets. [e]

Referring to the logBCF dataset, the only approaches that performed significantly better than the random approach, were the two k-Medoid approaches (on principal components derived from descriptors or derived from predicted properties). All other approaches performed worse with statistical significance ($p < 0.05$). Further, the approaches using the AD estimation performed similarly and better than the Kennard-Stone approaches. On the logKOC dataset, all systematic approaches except the Kennard-Stone approaches (AD-Spider, AD-Fetcher, k-Medoid, k-Medoid on predicted properties), perform equally well and significantly better than the random approach. Referring to its low initial average error, the AD-Fetcher can be seen as the best working approach.

The observations on the boiling point dataset are similar to those on the logK_{OC} dataset with the exception that the best initial performance is derived with the clustering approaches and that the improvement to the performance of the random approach is not so significant. Further, starting from 40 selected compounds, the performance of the AD-Fetcher does not improve anymore. Finally, on the -logIC₅₀ dataset, only the AD-Spider approach performs significantly better than the random approach. The k-Medoid approaches show a similar performance as the random approach, whereas the clustering approach on predicted properties is permanently performing better than the clustering approach on descriptors. Still, this difference is not statistically significant. Comparable to the performance on the boiling point dataset, the AD-Fetcher has a good initial error performance, but reveals stagnation from 30 selected compounds.

The evaluation of the performance referring to the correlation revealed no insight beyond. The observations were equivalent to those on RMSE therefore it is 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 dataset for all endpoints and methods.

To be enabled to do a comparison of the stability and reliability of the approaches with one another, we calculated the standard deviation of the RMSE for all approaches on all datasets. The results can be seen in Fig. 61a-d).

We explicitly choose not to show the standard deviation in the same plot as the average RMSE, as this implicates the possibility to evaluate the significance of an improved performance by overlapping intervals. In fact, due to the preceding random exclusion of 16% of compounds from each design set, this is not the case. A sampling on design sets showed that the performance derived from different splits differs with more than two standard deviations. Still, the standard deviation of the models derived with the selection approaches is a valid measurement to estimate the uncertainty within one selection approach and compare it to that of other approaches.

The observations on the standard deviation are similar for all datasets. Compared to the other approaches, the initial standard deviation of the random approach is higher, but furthermore, it is decreasing the fastest. For all other systematic approaches the standard deviation decreases (and thereby the reliability of the resulting models increases) with a growing number of selected compounds. The standard deviation of the Kennard-Stone algorithm on descriptors reveals quite an inconsistent development. For the logBCF it is permanently growing, for the logK_{OC} dataset, it has a peak at 20 selected compounds, for the boiling point dataset, the peak is reached at 40 selected compounds. Only for the -logIC₅₀ dataset the development is similar to that of other systematic approaches. Furthermore the model error development on the boiling point dataset is remarkable when considering the reliability of the AD-Spider approach. For the whole range from 15 to 40 compounds, we compared the models derived on each of the 250 validation splits with the model derived in the previous step. We found

that on the whole range, a minimum of 200 models (80%) improved with any additional selected compound.”[e]

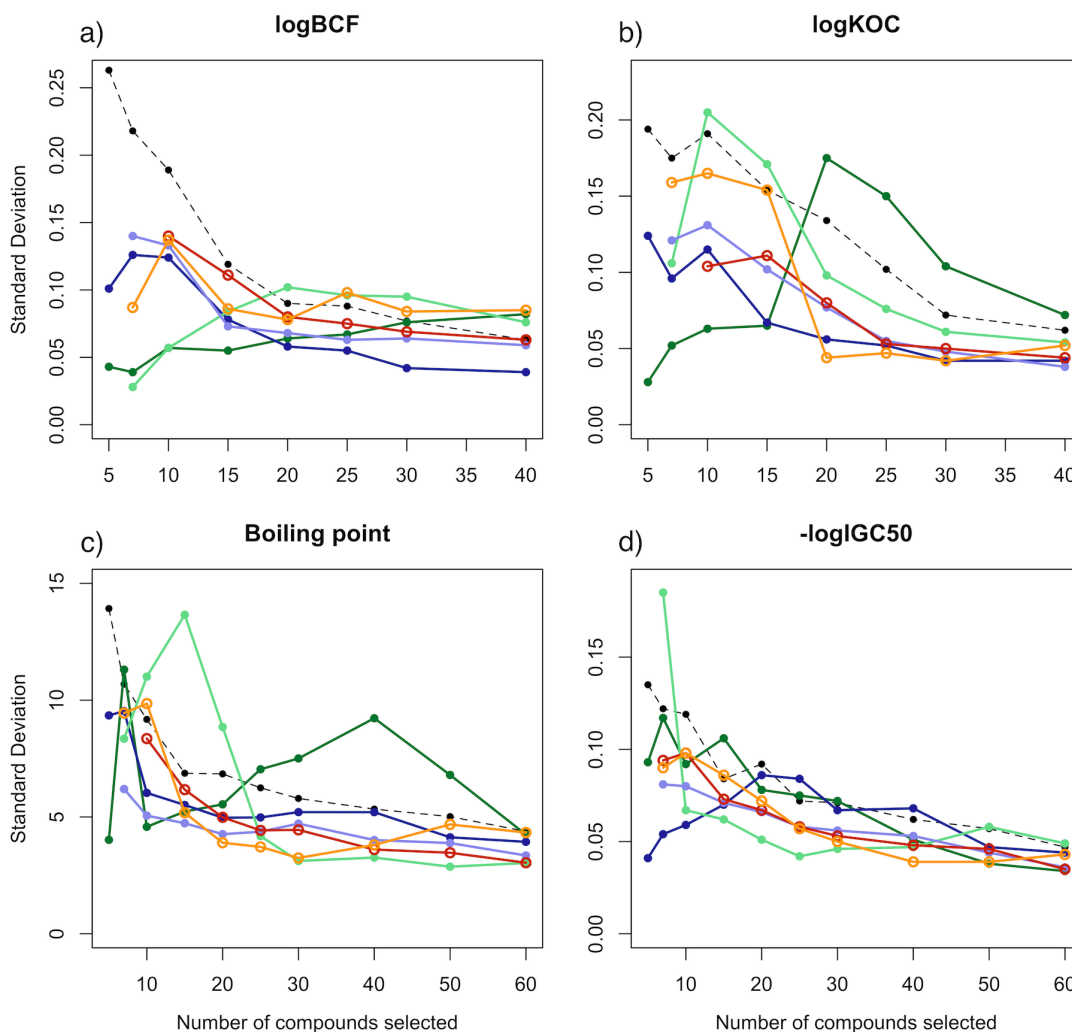


Figure 61. Comparison of the standard deviation derived from the RMSE performance. [e]

4.4.1.2 Classification datasets

“Referring to the size of the CYP-inhibition dataset, the computational costs of the AD-Fetcher approach (an overall number of 130,000 PLS models is required) and taking into consideration its poor performance, we disclaimed a full statistical validation of the approach on the classification dataset. The performance of the other approaches is shown in Fig. 62a) and the according standard deviation in Fig. 62b). The y-axis shows the development of the balanced accuracy.

Similar to the results derived on the regression datasets, the performance of the Kennard-Stone approaches was significantly worse than that of other approaches. Further, the k-Medoid approaches are within the best methods for compound selection. The performance of AD-Spider is significantly worse than that of the clustering approach and also worse than the results derived from a random

selection. A comparison regarding the F-measure as criterion of prediction quality resulted in the same observations.”[e]

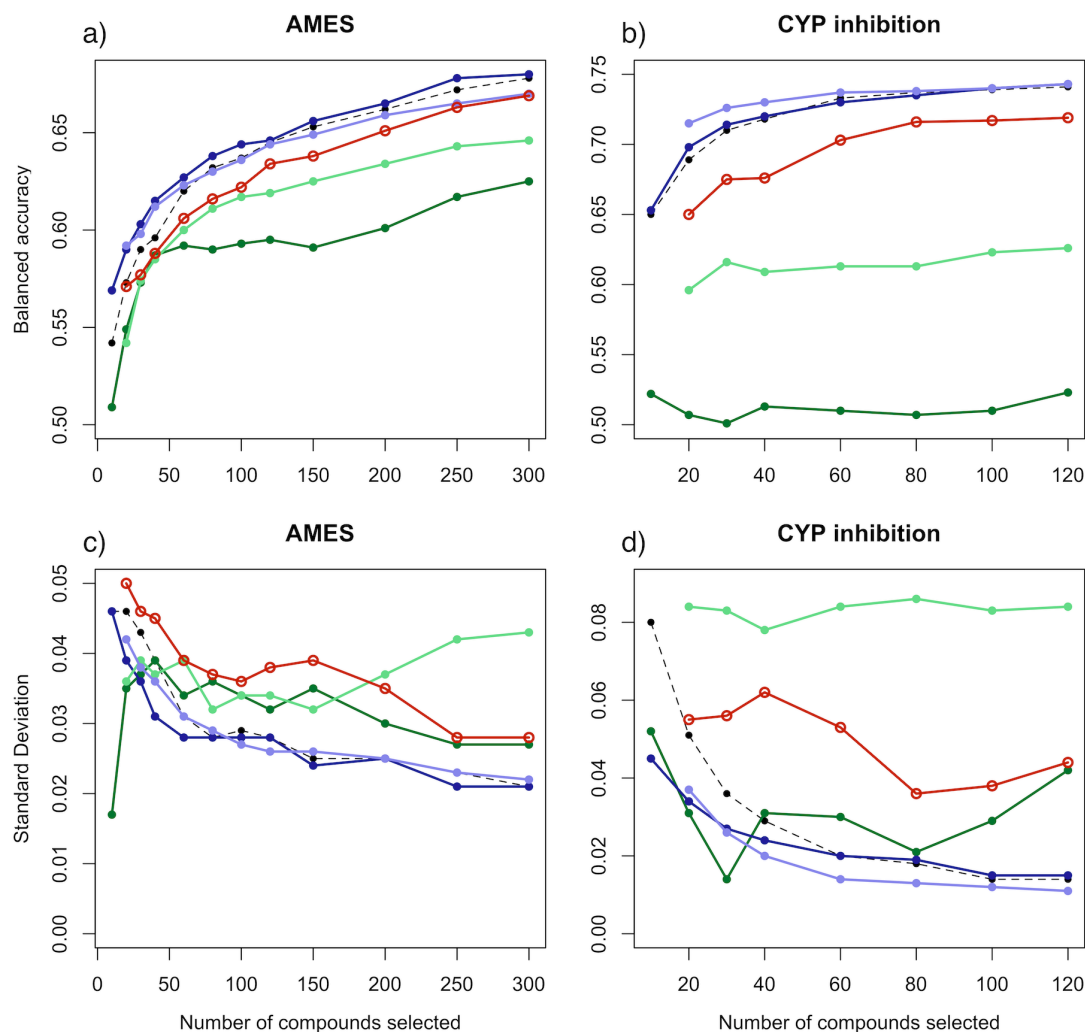


Figure 62. RMSE and according standard deviation for the classification dataset. [e]

4.4.2 Interpretation

4.4.2.1 AD-Spider

“The AD-Spider approach, which takes the variance and the correlation of predictions into account, performs significantly better than the performance of a random selection on the datasets for $\log K_{OC}$, boiling point and $-\log IGC_{50}$. Furthermore, its performance was equally well as that of the k-Medoid approach on the $\log K_{OC}$ dataset and for the boiling point. In case of the $-\log IGC_{50}$ dataset, it performed even better with statistical significance.

Contrary, the average AD-Spider performance on the $\log BCF$ dataset was significantly worse than for a random selection. The reason therefore can be found in a depiction of the principal components derived from the E-State indices for the

dataset. Fig. 63a) shows that most compounds are within a small subspace and the remaining ones widely scattered and sparsely filling the rest of the chemical space.

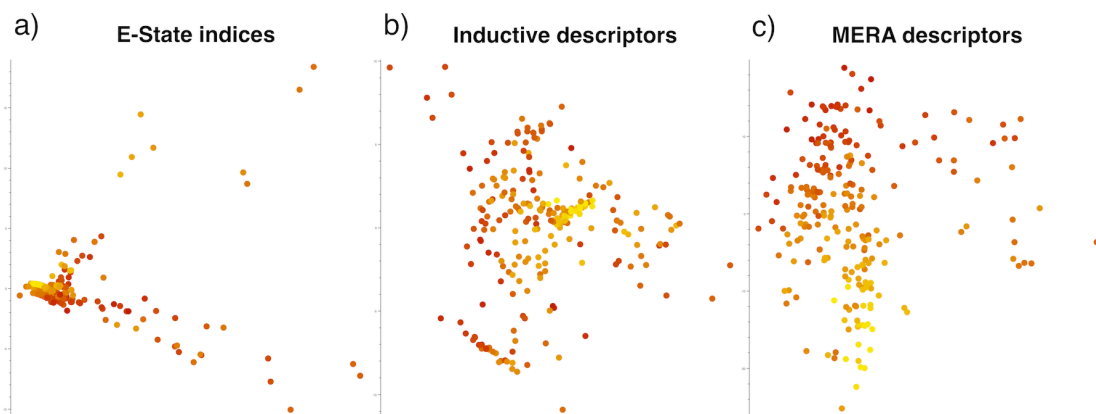


Figure 63. The principal components of the logBCF dataset, derived from different descriptor sets. [e]

We therefore tried a comparison of the examined approaches on the same dataset, but with different (not fragment based) descriptors. It was decided to use Inductive descriptors¹⁶⁶ and MERA descriptors^{161,162} for the representation of the compounds. A depiction of the main principal components can be seen in Fig. 63b) and 63c).

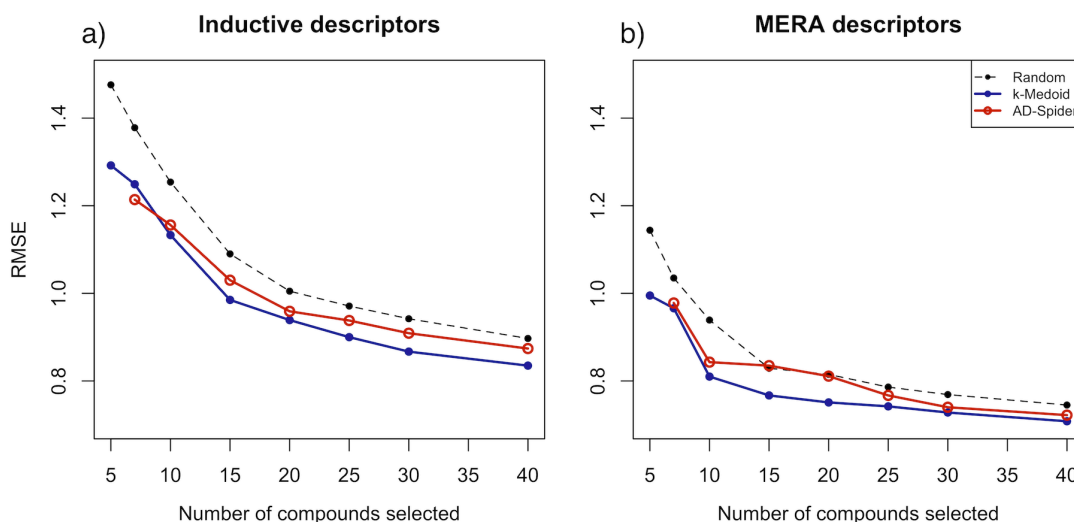


Figure 64. RMSE performance for a random selection, k-Medoid and AD-Spider using inductive and MERA descriptors to represent the logBCF dataset. [e]

We repeated the study with these descriptors, comparing the random selection, the k-Medoid approach and AD-Spider. The results are shown in Fig. 64. Both for the Inductive descriptors (Fig. 64a), as for the MERA descriptors (Fig. 64b) the quality of the models derived from the approach increases, compared to the two other selection approaches. It performs significantly better than a random approach on both datasets and approaches the performance of the selection derived from the k-Medoid clustering.

We also examined the performance for other descriptors which delivered a similarly scattered depiction as the E-State indices did and derived a performance worse than that derived by a random selection. Obviously the AD-Spider approach is not appropriate for scattered compound distributions.

Furthermore, taking into consideration that in comparison to the k-Medoid approach, the AD-Spider performs significantly worse for a dataset of 238 compounds, equally well for a dataset of 648 compounds and significantly better for a set of 1093 compounds, implicates that there is a correlation between the size of the dataset and the performance of AD-Spider on it. Such a dependency seems logical as a chemical space defined by a lower number of compounds is less densely populated. Therefore also the probability to find pairwise correlations in predictions between the compounds is decreased or just arbitrary. As the approach works on these correlations, small datasets affect its performance.

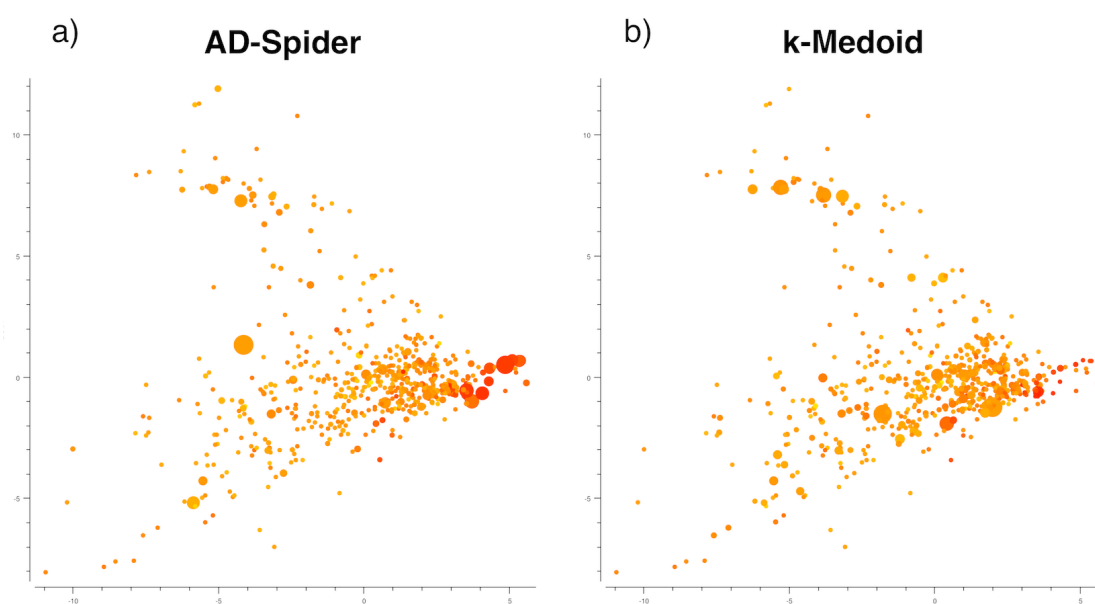


Figure 65. Selection of highly representative compounds. [e]

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

To gain a deeper insight into the mechanistic within the approach, we investigated those compounds within the logK_{OC} dataset, which have a high contribution to the quality of the resulting model. Therefore we built 7000 models on the dataset, each with 20 randomly selected compounds. We used these models to predict the remaining compounds, which have not been used for model building and calculated the RMSE. For each of the 648 compounds in the dataset, we calculated the average RMSE of all models it contributed to and used it as a measurement of representativeness. Finally we used the selected compounds from

the 250 validation trials, using AD-Spider to draw 20 compounds and counted in how many cases each molecule had been chosen.

The result of this analysis is shown in Fig. 65a). The axes depict the principal components and each data point represents one compound. Highly representative compounds, which contributed to good models, are colored red, those contributing to poor models are colored yellow. The size of the data points indicates how often a compound was selected with the AD-Spider approach. Remarkably, almost all frequently selected compounds have a high or very high representative quality. Fig. 65b) shows the same correlation for the k-Medoid selection. Although also this approach is favoring the selection of compounds with a good representativeness, it is neither so successful in the highly representative compounds and the AD-Spider, nor is its selection so specific to certain compounds. This allows the conclusion that the good performance of the k-Medoid approach is in large part resulting from its good statistical coverage of the chemical space, but that the good performance of the AD-Spider approach is resulting from its ability to recognize highly representative compounds. We repeated this comparison also for the boiling point and the $-\log\text{IGC}_{50}$ dataset and observed the same correlations.”[e]

4.4.2.2 *AD-Fetcher*

“The performance on the $\log\text{BCF}$ dataset is comparable to the one of AD-Spider and over the whole range significantly worse than the random approach. The dataset on $\log\text{K}_{\text{OC}}$ is the only one where AD-Fetcher could perform similar to AD-Spider and the k-Medoid approach and where it performs significantly better than the random approach. For the boiling point dataset, it performed initially well, but starting with 40 selected compounds, the performance of the approach is, contrary to all other tried approaches, not improving significantly anymore. The same observation can be made for the $-\log\text{IGC}_{50}$ dataset. The approach works for less than 30 selected compounds, but starting from this point, the performance is significantly worse than for AD-Spider or the k-Medoid approach.

This observation, which is conflicting with the observation on the example in the materials and methods section, where the approach was very stable, can be explained by the changed parameterization. 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, we do not have a mechanism to avoid drawing redundant information within a cycle.”[e]

4.4.2.3 *Predicted properties*

“Regarding the regression datasets, the use of a three dimensional PCA space derived from predicted properties instead of search space defined by descriptors or their orthogonal transformation has to be interpreted in two ways. Firstly, in case of the k-Medoid clustering the performance did not significantly change. Neither for the error performance, nor the standards deviation, nor the correlation coefficient a clear tendency towards descriptor space or predicted property space

was observable. Only when regarding the reliability in terms of improvement, there is a slight (but not significant) bias towards favoring the predicted properties. The switch in the search space representation neither improved nor declined the performance of the selection approach. The robustness of the k-Medoid approach against the dimensionality of the search space has already been shown in our previous study. Furthermore, the results of this study indicate that it also has no influence on the approaches performance, if the search space takes information about the target property into account.

Secondly in case of the Kennard-Stone approach, the switch in the search space significantly improved the performance on all regression sets, regarding error performance and correlation. In case of the boiling point dataset and the logK_{oc} dataset, the initial performance with less than 20, respectively 15 selected compounds the use of stepwise approach on predicted properties could not improve the performance, but starting from this point it significantly improved. For the two other datasets the performance was improved when using predicted properties instead of principal components starting with only seven selected compounds.

Regarding the classification dataset for cytochrome inhibition, the use of a predicted property space could not just improve the performance of the Kennard-Stone algorithm, but also the balanced accuracy of the k-Medoid approach could be significantly improved for 20 to 60 selected compounds.”[e]

4.4.2.4 Comparison with models on the whole dataset

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

Table 14. Reference models on the whole dataset.

Dataset	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±2.2	-	50	60	60
-logIGC ₅₀	0.62±0.04	-	40	40	30
CYP	-	75.4±1.0	150	120	N/A

We investigated for the k-Medoid approach, the k-Medoid-Approach on predicted properties and the AD-Spider approach, the number of required compounds to reach a model of the same accuracy. The results can be seen in Tab.14. The first column indicates the dataset, the second and third column contain information on

the average performance and according uncertainty and the following columns display the number of compounds required to build a model within one standard deviation of the reference model. The best approaches, referring to the number of required compounds, are indicated with a green background.

The AD-Spider approach delivers the best performance for the $\log K_{OC}$ and the $-\log IC_{50}$ dataset and it delivers models with similar performance for only 20 out of 648 (3.1%) and for 30 out of 1093 compounds (2.7%). The k-Medoid approach on predicted properties is the best performing approach on the cytochrome dataset with 120 out of 7481 compounds (1.6%).”[e]

4.5 Exhaustive comparison

The four previous studies exemplified that the use of meaningful selection approaches and an adaptive data representation can significantly improve the performance of an experimental design. Still, the picture is not complete. Although the results show the advantages for the combination of certain selection approaches with a certain data representation, an exhaustive comparison is missing. Finally, in this study, I present the results of all selection approaches used in this study carried into execution on all regression datasets and for all data representation techniques that were presented in this study.

The D-Optimal criterion, the Kennard-Stone algorithm, space filling design, the developed similarity selection and the k-Medoid approach are conducted on principal components, latent variables, selected descriptors and predicted properties. Additionally, to complete the overview, the random selection, MDC, AD-Spider and AD-Fetcher are taken into consideration as well. The performance is compared respective to two different characteristics: firstly, the used selection paradigm; and secondly, the underlying data representation.

As the previous results indicate that none of the developed approaches is optimal for classification problems, the binary classification datasets are not taken into consideration in this study. The parameterization for the validation procedure (number of dataset splits, split size, modeling procedure, etc.) uses the default values given in paragraph 3.5.

4.5.1 Method based view

Fig. 66 shows the performance of all approaches on the five regression datasets. The coloring is in reference to the selection approaches. The approaches using the Kennard-Stone algorithm are shown in bright blue, whereas those based on the D-Optimal criterion are shown in a dark blue. The similarity selection is shown in dark green the space filling design is depicted in a bright green and the cluster-based approach is represented in ocher. The underlying data representation is shown by differently dashed lines, whereas the random selection is represented by a solid black line and AD-Spider, AD-Fetcher, as well as the MDC selection are colored in shades of grey.

The first observation is that the best working approaches, regardless of the structure of the dataset and the underlying data representation appear to be those using the k-Medoid approach, whereas the approaches based on the space filling design reveal a comparably poor performance.

The performance reached on the logBCF dataset is in line with this finding. Furthermore, the performance of the Kennard-Stone approaches is not optimal as well. The similarity selection performs clearly better, and the performance of D-Optimal criterion is somewhere in between. The MDC selection works exceptionally well whereas AD-Fetcher and AD-Spider cannot reach the performance of the random selection.

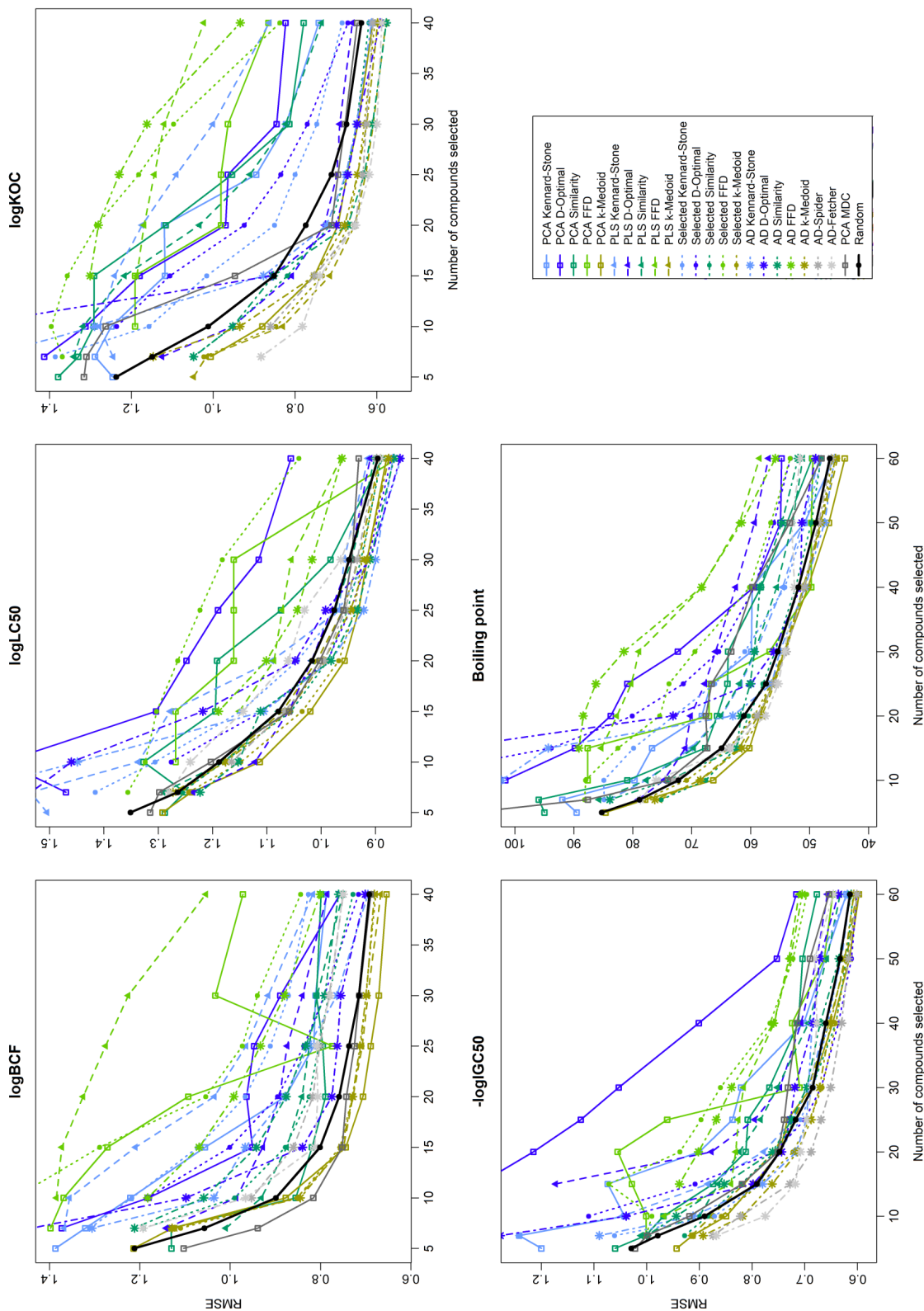


Figure 66. Performance of different approaches on the used regression datasets. The coloring is in reference to the underlying selection approach. Especially the k-Medoid approach performs well for all datasets and on any data representation.

This is reasonable, as the dataset is comparably small. This results in a limited number of observable co-variances, which are essential for the two aforementioned approaches. Furthermore, the logBCF dataset has to be identified

as exceptional, as it was customized for modeling purposes, which is displayed by scattered compound distribution in PCA space.

Respective to the logLC₅₀ dataset, the space filling approach is the worst, whereas the approaches based on the similarity selection work well. Especially the performance of the approaches based on the D-Optimal criterion is worth to be paid attention, as there is no clear tendency in the observable performance. Although a widespread applicability is shown, the PLS-Optimal as it was developed for the first study performs well. This indicated that the initial idea to combine latent variables with a dissimilarity selection was reasonable. Furthermore, worth mentioning, numerous approaches perform better than a random selection. This was not the case for the logBCF dataset, where almost only the cluster-based approaches performed better than a random selection.

The observations on the logK_{OC} dataset reveal similar observations. Again, numerous approaches perform better than the random selection. Especially for the k-Medoid approaches, this is true, as well as for implementations of the D-Optimal criterion and the similarity selection, whereas the Kennard-Stone approach cannot improve the performance of the random approach for any data representation.

The space filling design delivers the worst performance and AD-Spider, as well as AD-Fetcher perform well, both for the logK_{OC} dataset and the -logIGC₅₀ dataset. Both dissimilarity-based approaches (Kennard-Stone, D-Optimal) perform similar. The number of approaches performing better than a random selection is not as high as for the logK_{OC} and the logLC₅₀ dataset. Still it is clearly higher than for the logBCF dataset.

Respective to the observations for the space filling approaches, as well as for the cluster-based approaches, the boiling point dataset reassures the prior findings. Furthermore, the number of approaches improving the performance of the random selection is limited.

4.5.2 Property based view

Fig. 67 shows the performance of all approaches on the five regression datasets. The coloring is in reference to the data representation paradigm. The PCA-based approaches are colored in yellow, whereas PLS-based approaches are shown in orange. The approaches based on selected descriptors are shown in pink and the approaches using predicted properties are represented in purple. The coloring of the random selection, the AD-Spider and the AD-Fetcher approach, as well as of the MDC selection is the same as in Fig. 66, underlying selection approaches are indicated by differently dashed lines.

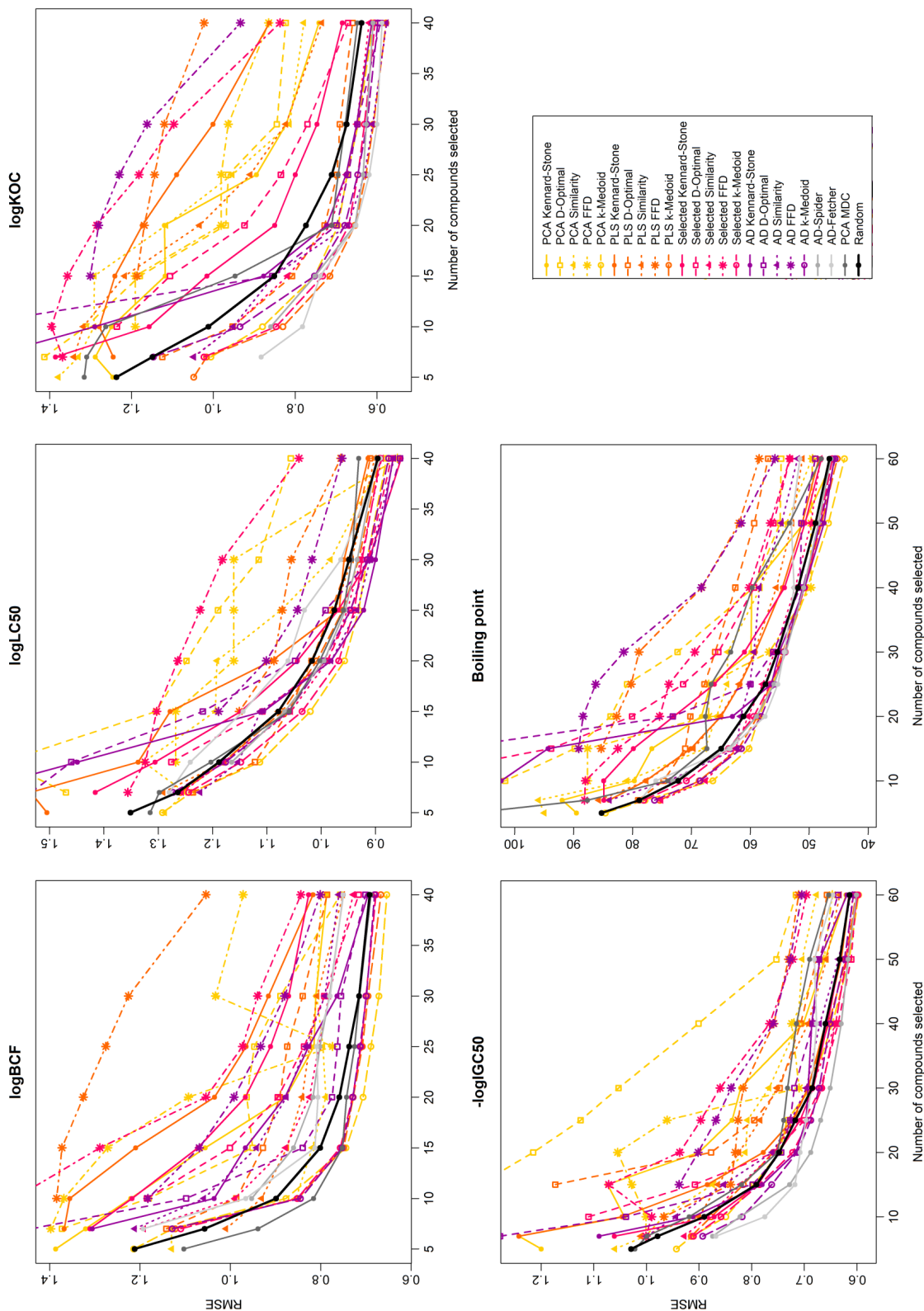


Figure 67. Performance of different approaches on the used regression datasets. The coloring is in reference to the underlying data representation. The interpretation is complex, but in general the data representation with predicted properties delivers good and stable results.

Overall, the interpretation, respective to the underlying data representation is more complex, compared to the interpretation respective to selection approaches. This is particularly observable for the logBCF dataset. No clear order or bias

towards a certain data representation paradigm is identifiable. This might be once again due to the aforementioned, exceptional data distribution.

Taking a look at that logLC₅₀ dataset, the observations are more distinct. The PCA-based approaches show the worst performance, the AD-based approaches perform better than the random selection for a higher number of selected compounds, whereas the best initial performance is reached with latent variables and selected descriptors. It has to be taken into consideration, that the principal components derived on this dataset were not correlated with the target property at all. The poor performance of the PCA-based approaches is a consequence thereof.

A further indication that the correlation of the principal components to the target property is crucial for the performance of the PCA-based approaches is given by the logKOC dataset. In spite of a similarly balanced distribution of the compounds in the chemical space (compared to the logLC₅₀ dataset), the performance of the PCA-based approaches is clearly better. This is due to a natural order within the target property along the principal components.

The performance on the boiling point dataset confirms this observation. As there is some order in the chemical space, respective to the concerned endpoint, no clear bias towards a certain linear data representation is observable. Latent variables, selected descriptors and principal components enable a performance of similar quality.

On the contrary, for the -logIGC₅₀ dataset, which is lacking of such a correlation, the performance of the PCA-based approaches is not optimal. Furthermore worth mentioning, on this dataset, the approaches based on predicted properties work similarly well, compared to the approaches using selected descriptors. Generally, for all of the datasets, except the one on logBCF, the AD-based approaches appear to be performing best.

4.5.3 General remarks

Overall, the most explicit finding is that the cluster-based approaches, using the k-Medoid partition perform best. The partition-based approach is extremely robust and its performance is almost independent of the data representation. This was observable on all five datasets. Furthermore, the k-Medoid approaches enabled a smooth development of the average error curve without exception.

Contrary, the use of the space filling design appears to be deprecated, as it did not work well on any of the datasets. Even a change in the dataset representation did not improve the poor performance. Usually it would be expected that the space filling design reveals advantages for datasets with equally distributed compounds, but the evidence gathered within this data contradicts this assumption.

Respective to the paradigm of data representation, except for the space filling selection approach, the predicted properties constantly enable a smooth and reliable decrease in the average error curve. Furthermore, latent variables and

selected descriptors perform similarly well clearly better than principal components.

The PCA-based representation is the most inconsistent one, with numerous deviations. Additionally, we found indications that a natural order of the target property in PCA space is required to make the selection approaches work properly on principal components. As soon as a dataset lacks of this correlation, the PCA-based approaches reveal problems.

Furthermore, our results indicate that the data distribution is a crucial criterion for a reliable experimental design. It is remarkable that especially for those datasets with a plain distribution and without outliers ($\log LC_{50}$, $\log K_{OC}$) a lot of combinations could perform clearly better than a random approach, whereas the number of well working approaches drastically increased for datasets with outliers.

4.6 Practical application

The purpose of this practical application was to design the experiments to select representative compounds to build a reliable prediction model for the EC (electrochemistry) starting voltage. The EC starting voltage is a physico-chemical endpoint that was not taken into consideration in the previous studies. This is caused by the circumstance that only little information on this endpoint is available, which makes a statistical evaluation impossible.

The EC starting voltage describes the voltage that is required to initiate a molecule's decomposition. In order to measure this endpoint, a mass spectrometer is combined with a commercial EC-cell setup. The working electrode potential is applied within the range from 0 mV to 2,500 mV and mass spectra are recorded after each change of the cell potential.¹⁹⁴ The impressed voltage, when the mass spectrometer shows a decrease in the amount of the starting material and metabolites appear is then defined as the EC starting voltage.

The following study concentrates on the oxidative starting voltage detected with electron-capture mass spectrometry. The motivation to investigate on this endpoint is its potential relevance to predict the ecological fate of a chemical compound, amongst others. It is intended to function as a simulation of oxidative processes in the environment. Knowledge about the EC starting voltage might be beneficial to prioritize hardly degradable persistent organic pollutants and to improve the predictive quality of degradability models.

The starting point for our study was a list of available chemicals and twelve EC starting values for twelve previously measured compounds.

4.6.1 Evaluation of previous results

The initial step within the prioritization of compounds to increase the applicability domain and accuracy of a model to predict the EC starting voltage of organic compounds was the evaluation of the previously presented results of Kamel Mansouri, as well as the model performance that could be derived with the known values for twelve previously measured compounds.

Our analysis of relevant Dragon descriptors¹³² revealed a strong bias to numerous autocorrelation descriptors (Geary, Moran, Broto-Moreau) for lags of various length. As these autocorrelation descriptors were not specific towards a certain property (mass, polarizability, van der Waals volume) or a certain lag length the descriptors are most likely to represent all the same principal property, which can be referred to as 'molecular size'.

A second, frequently detected group of descriptors was atom count descriptors. Also these descriptors are closely related to the molecular size of a compound. This finding was expected, as the results presented from Jülich indicate a correlation between the number of aromatic rings and the EC starting voltage.

The suggested models were calculated with descriptors, which were preselected by their correlation on the whole dataset. To give a realistic insight into the reliability of the descriptor selection, we believe, it has to be emphasized, that the measurements contained one compound (PCB31) that revealed an EC starting value which was three times higher than the second highest value in the data set. Taking a closer look at the best suggested descriptor (GATS2m) the R^2 on the whole dataset is 0.84, but applying a leave-one-out cross-validation (CV_{Loo}), the correlation coefficient decreases to 0.61, a value which is comparable to that derived with a CV_{Loo} on the dataset without PCB31 and therefore more realistic.

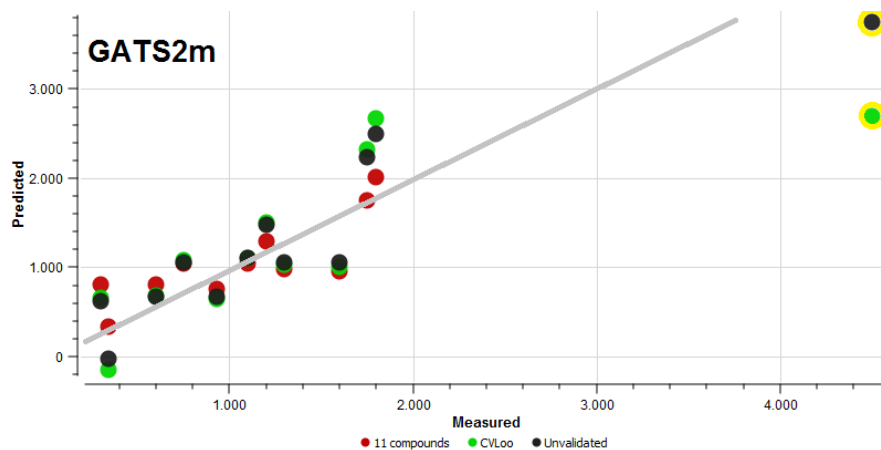


Figure 68. The correlation of GATS2m with EC starting voltage. Black dots represent the prediction values for regression model trained on the whole dataset, green ones represent prediction values resulting from a CV_{Loo} . It is obvious that the predictions for PCB 31 (highlighted with yellow circles) get worse, for a CV_{Loo} .

Fig. 68 shows the measured EC starting voltage on the x-axis and the predicted values on the y-axis. Predictions derived without validation are depicted black, those from a CV_{Loo} on all twelve compounds are depicted green. The predictions for the compound PCB 31 are highlighted with a yellow circle and the correlation analysis for all compounds, except PCB 31, are depicted red.

Table 15. R^2 values derived from a CV_{Loo} on eleven measurements.

Descriptors	FSMLR	PLS
CDK	0.3	0.1
Dragon6	0.4	0
AlogPS + E-States	0	0
ISIDA	0	0.2
GSFrag	0.2	0.1
MERA + MerSy	0	0.3
Chemaxon	0.4	0.4
Inductive	0.1	0.2
Adriana	0	0.3
Spectrophores	0.2	0.4
ShapeSignatures	n/a	n/a
QNPR	0	0.3
Mass + nC	0	0.1

Furthermore, we believe that the use of four or five descriptors to describe a relation of twelve instances is not appropriate. Also literature recommends the usage of not more than one or two variables for such an amount of data.^{45,186,195} Additionally, we found that the underlying dataset contained a duplication of the compound tetrazene.

We used OCHEM⁸⁶ to build properly validated linear OSAR models on the eleven compounds (PCB 31 was excluded). Both FSMLR¹⁹⁶ and PLS regression⁴⁵ were applied to various descriptor sets, but no approach exceeded an R^2 of 0.4. The derived results are shown in Tab. 15.

4.6.2 Descriptor representation

The use of Dragon descriptors to find correlations in such a small amount of prior knowledge (twelve measurements) is not unproblematic, as the Dragon package provides almost 5,000 descriptors. This increases the probability of 'chance' correlations. We therefore decided to use raw (non-normalized) E-State indices,^{84,85} to represent the chemical compounds. E-State indices contain implicit information about a compound's

- substructures / fragments
- size / weight
- electrotopological states

All three properties might be highly relevant for the target property EC starting value and the E-State indices might therefore be optimal to span the chemical space for relevant compound.

4.6.3 Data set cleaning

Referring to the list of available compounds, we took only those into consideration, for which a CAS registry number and therefore unambiguous structural information was available. In the first step we performed an automatic exclusion of compounds containing metals and uploaded the remaining 450 compounds to OCHEM.

Table 16. The five principle components that characterized the dataset and the most relevant descriptors.

Property	Desc1	Desc2	Desc3	Desc4	Desc5
Size	SsCH3	SssO	SsH	Se102P4sd	SdsssP
Branchedness, nitrogen	SeaC3N2aa	SaaN	SssNH	Se1C3H1s	Se1C3C4ss
Aromaticity, alcohol	SaaCH	SeaC2C3aa	SeaC2C2aa	Se1C2H1a	Se1H101s
Oxygen, halogens	SsCl	SdO	Se1C4Cl1s	Se2C301s	SdssC
Phosphor	SssCH2	Se1C1H1s	SdsssP	Se102P4sd	SaasC

In the second step, we manually examined the compounds once again to exclude inorganic compounds, polymers, compounds containing extraordinary atoms (e.g. Ge, Si, etc.) and those compounds containing metal, which were not detected in the automated approach. The resulting dataset contained 417 compounds.

We applied principle component analysis³⁵ to the E-State descriptors to detect the five most descriptive principle properties of the dataset and to enable the exclusion of structural outliers. This step was performed three times and resulted in the exclusion of 93 further compounds and a dataset containing 324 relevant compounds, each of them organic and containing no other atoms than C, H, O, N, S, P, and halogens. Fig. 69 shows the distribution of these compounds in the chemical space.

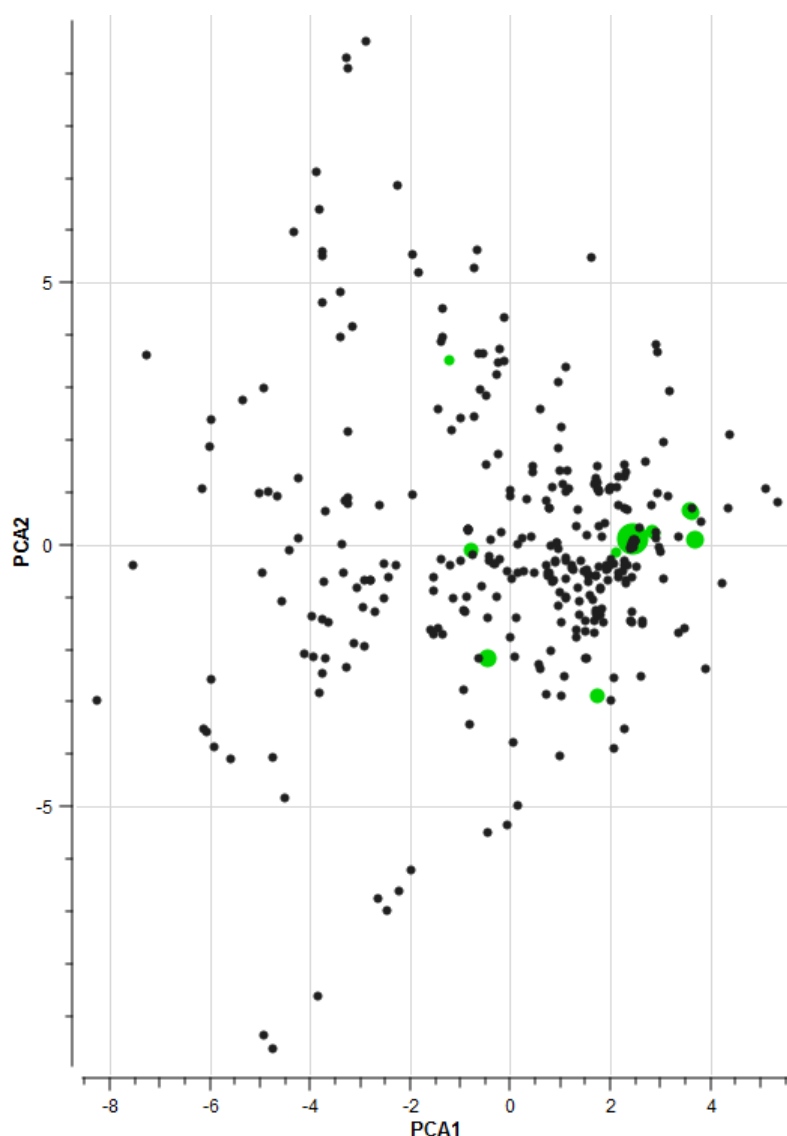


Figure 69. The compound distribution referring to the first principal components. The twelve measured compounds are indicated green and the dot size correlates to the EC value.

An analysis of the derived principle components revealed that the most significant property within the dataset was the compound size, but also aromaticity,

branchedness, and the content of certain compounds contributed significantly to characterize the dataset. Tab. 16 shows one principle component per row and the five most contributing descriptors.

4.6.4 Experimental design

Our previous studies indicate that for datasets of comparable size (>300 compounds), a number of 30 systematically selected compounds delivers a model which is not anymore different with statistical significance from a model derived on the whole dataset. As we already have a seed of twelve measured compounds, but as we do not have any information about the underlying concept these compounds were selected with, and as these compounds might therefore not be free of redundant information, we decided to increase the number of required measurements to 35. Therefore 23 further compounds need to be selected.

Furthermore, we decided for a two-step selection strategy. In the first step we applied an approach that takes the available information about the target property into account and in the second step we apply a PCA-based approach to ensure a good statistical coverage of the whole chemical space of interest.

To guarantee a maximum level of reliability and stability in the selection process, we generated 250 random samples, each containing 90% of the available 324 compounds. These 250 samples were used to investigate the frequency, each compound was selected and thereby to estimate the representativeness of this compound. This information was used to select the most reliable combination of compounds within the 250 samples.

4.6.4.1 *Property oriented selection step*

In the first step the AD-Spider approach was used to select a fixed number of six compounds from each sample. The AD-Spider approach uses an ensemble of PLS predictions derived from an n-fold bagging approach (n=64) on the 12 compounds with measurement values. This ensemble of predictions enables to calculate the standard deviation within the predictions for each compound, which is referred to as the distance to model (DM). The DM is then combined with the observed correlations within the predictions. The AD-Spider approach was shown to select compounds of high significance for resulting models. Still, due to the comparably small number of measurements and the poor knowledge of the target property, we decided to limit the number of compounds selected with this approach to only 50% of already available measurements.

Referring to the 250 generated samples, we counted the number of occurrences in the resulting selection for each compound and used this number as a score of representativeness for the compound. Out of all available selected combinations derived from the 250 samples, we chose the one with the highest sum of representativeness scores.

The selected compounds, presented in Tab. 17, consisted of trichloroethanoic acid, chlorosulfuron, alpha-naphthylflavone, coumafene, 3-aminopyrene and ethynyl estradiol.

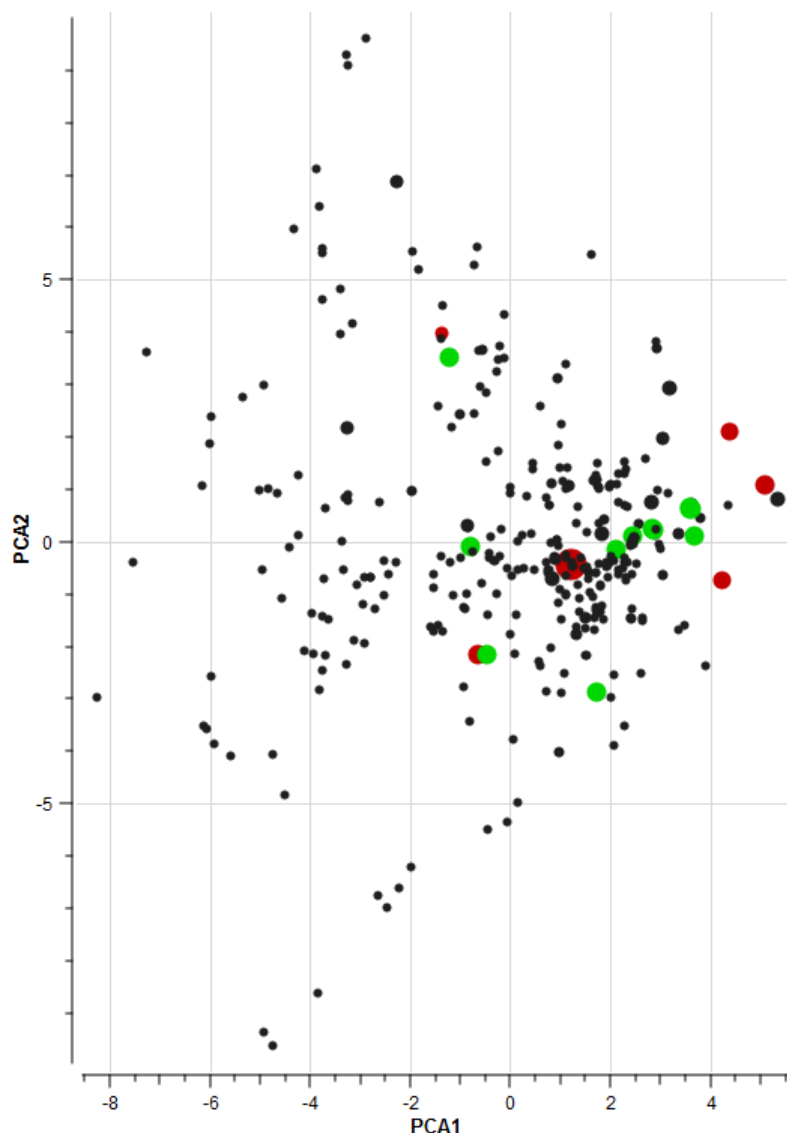
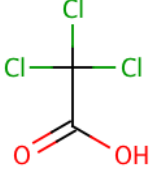
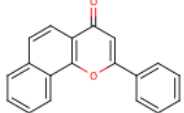
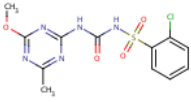
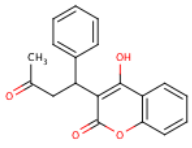
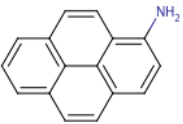
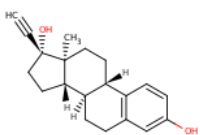


Figure 70. Compounds selected with the AD-Spider approach indicated red. The compounds previously measured are indicated green and the dot size represents the selection frequency in 250 trials.

Fig. 70 shows the results of the selection in a space spanned by the first two principal components. The measured compounds are highlighted green, the compounds selected by the AD-Spider approach are highlighted red and the dot size represents the selection frequency. The most frequently selected compound (trichloroethanoic acid) is located in the center of a densely crowded cluster with a reasonable distance to all previously measured compounds. The selection of this compound can be interpreted as a contribution to stabilizing the model, whereas all other selected compounds are located beyond the periphery of the subspace with experimentally measured compounds. Their contribution is therefore most likely to increase the applicability domain of the resulting model.

Table 17. Compounds selected with the AD-Spider approach and the respective representativeness score.

CAS-RN	Structure	Count	CAS-RN	Structure	Count
76-03-9		204	604-59-1		92
64902-72-3		92	81-81-2		88
1606-67-3		80	57-63-6		47

4.6.4.2 Descriptor oriented selection step

In the second step we used the k-Medoid clustering on five principal components derived from the E-State descriptors. The twelve compounds with available measured values were used as an initial seed of cluster centers, as well as the six compounds that were selected with the AD-Spider approach. Depending on these data points 35 clusters were assigned to the dataset, so that an additional 17 compounds (the newly detected cluster centers) were selected.

The k-Medoid clustering was shown to work highly reliable, regardless of the dimensionality of the underlying search space. It benefits from the good statistical coverage of the chemical space and works with high robustness on datasets of different distribution. The selection of the most representative clustering derived from the 250 was done exactly as for the AD-Spider approach.

Fig. 71 shows the compound selected with the k-Medoid approach on the first principal components. The coloring is similar to Fig. 69 and Fig. 70 and the newly selected compounds are highlighted blue. The dot size indicates the selection frequency within the 250 samples.

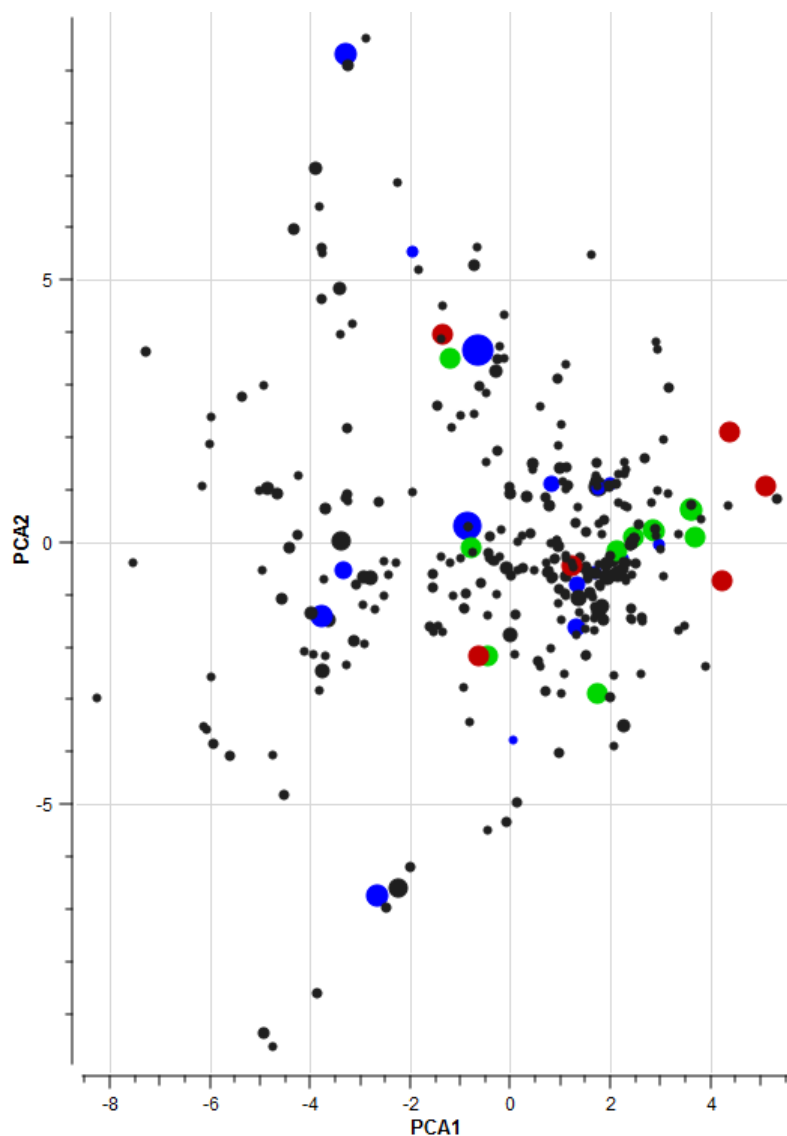
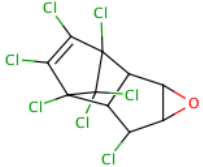
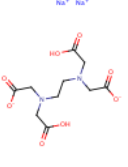
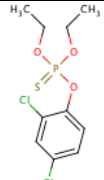
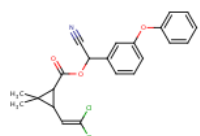
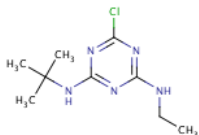
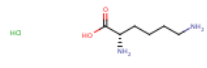
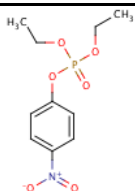
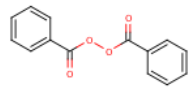
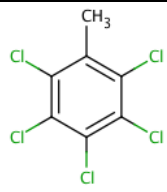
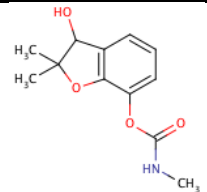
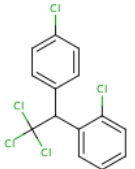
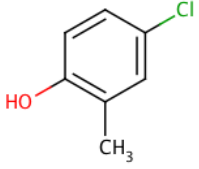
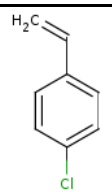
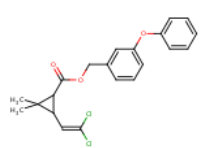
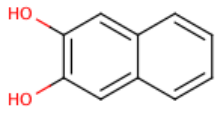
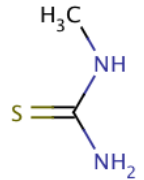
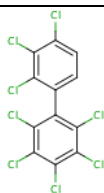


Figure 71. Compounds selected with the k-Medoid approach indicated blue. The coloring of the other compounds and the meaning of the dot size is similar to Fig. 70.

The full list of compounds that were selected with the k-Medoid approach can be seen in Tab. 18. It apparently consists of a variety of structurally diverse compounds. Referring to the PCA plot most of the selected compounds are located in those areas of the chemical space, which is not yet covered by the previously selected compounds, but which is still sufficiently crowded.

Table 18. Compounds selected with the k-Medoid approach.

CAS-RN	Structure	Count	CAS-RN	Structure	Count
1024-57-3		199	6381-92-6		177

97-17-6		125	52315-07-8		124
5915-41-3		115	657-27-2		83
311-45-5		81	94-36-0		72
877-11-2		64	16655-82-6		48
789-02-6		45	1570-64-5		38
1073-67-2		34	51877-74-8		29
92-44-4		22	598-52-7		3
55722-26-4		1			

4.6.5 Summary

Knowledge about the EC starting voltage of the suggested compounds should enable to build a model of sufficient prediction quality which is not limited to

aromatic compounds, but applicable to the full range of small organic compounds as they are contained in the dataset.

We employed a selection procedure to cover the whole range of the five most significant characteristics of the underlying dataset. Experimental measuring of the compounds we suggest should enable the detection of statistically reliable correlations between structural features and the target property. Although it suggests itself to use a combination of PLS regression and E-State descriptors for the subsequent modeling, we recommend the application of a more exhaustive comparison of different machine learning techniques and descriptor sets.

5 Conclusion and outlook

5.1 Summary of presented work

In this thesis I presented the results of five different studies, four of them published, one of them still in preparation, at the point of time this thesis was finished, and a practical application, that is currently in the stage of experimental testing.

I firstly used a model on aquatic toxicity¹⁹⁷ to carve out and expose the requirements and the course of action of current QSAR research with respect to the REACH legislation.

The main focus in this study was on the improvement of existing experimental design approaches and the development of new ones that we presented in four published studies. Three of these studies were dedicated to examine the benefits of stepwise, adaptive approaches, with two of them using classic selection approaches to a property correlated data representation based on descriptors^{198,199} and one of them introducing a newly developed concept of predicted properties and a compound representation by the applicability domain estimation.²⁰⁰ The fourth study investigated the usability of a static, cluster based approach in terms of a representative subset selection.²⁰¹

The results of these four studies were subject to a concluding comparison, respective to the choice of an optimal selection approach and the choice of an optimal data representation.

5.2 Significant findings

“The results of our study show that stepwise approaches, which take the correlation to the target property into consideration, significantly improved the quality of experimental design in terms of QSAR modeling.”[c]

The PLS-optimal design and DescRep operate in the property-correlated space; “therefore, the selection of compounds is not only based on their structural properties but is also tuned for a specific endpoint.”[a] Furthermore, “the variance in prediction can not only be used to estimate the applicability domain of a model, but it can also be used 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.”[e]

“With respect to the structural outlier, it was dramatic to see how the majority of selection procedures were strongly affected with the inclusion of only one compound, which was not representative of the analyzed set. This resulted in higher variability of models developed with such sets. Compared to the static approaches, the selection within stepwise approaches is not so focused on certain

compounds, but on a harmonious context within the selection. Thus small variations in the dataset get buffered in an efficient way.”[c]

Within the static approaches, used in this study, the k-Medoid selection was the only systematic concept, which performed significantly better than a random selection. Moreover, it performed similarly well, compared to the stepwise approaches. The observed criteria of quality hereby were:

- The performance of the k-Medoid approach was one of the the best for all examined datasets.
- Its performance was less dependent on the number of latent variables than most other approaches.
- It also performed well with a small number of compounds selected.
- The error decreased constantly with an increasing number of compounds selected.

For several of the examined datasets ($\log LC_{50}$, $\log K_{OC}$, $-\log IGC_{50}$), the stepwise approaches I developed were shown to result in models of the same average performance with only 30%-50% of required compounds, compared to models derived with the most commonly used static approaches. In case of the k-Medoid approach, the number of required measurements for a comparably performing model on these endpoints could be decreased to even 25%-35%.

5.3 Interpretation of the observations

“We recommend, whenever this is feasible, to design experiments in a stepwise manner. Especially in the case of high cost experiments, e.g. measuring aquatic bio-concentration factor,²⁰² that allow only a limited number of tests, the stepwise approaches can significantly decrease the financial effort to produce models of the same predictive quality. These models can be used to predict the molecules without measurements thus decreasing costs and time.

The PLS-Optimal approach is an appropriate choice for compounds and endpoints, where a linear correlation between the target property and the descriptor space is expected. For other kinds of dependencies, DescRep shows a fast convergence in error, a reliable performance with a low standard deviation, and a high robustness against structural outliers.”[c]

“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, combining this variance with correlated development in predictions, which putatively indicates a common mode of action, produced very good results. Especially for sufficiently big regression datasets (more than 500 compounds) with a non-scattered distribution of compounds, we could show the efficiency in this approach.”[e]

“The analyzed step-wise approaches explore different ideas for selection of compounds based on similarity and dissimilarity measures. Still, they produced comparable results. Thus, we can conclude that the major contribution to their

performance was not the selection method, but the accounting for the resulting property, i.e. informational basis on which the selection was performed. Similar observations were done for QSAR modeling, where the underlying data, but not the chosen machine learning method or descriptors determined the accuracy of models.^{203,204,131}[c]

For initial situations that cast a stepwise experimental design into doubt, for example resulting from financially cheap, but time consuming experimental procedures, we suggest the k-Medoid approach. Deduced from the results of this thesis, it “seems to be an excellent choice in terms of efficiency. It enables the calculation of models with the same predictive quality with a low number of tested compounds. Thereby the reliability, represented by the variance, is not negatively affected.”[b]

5.4 Outlook

The sequential approaches are surely not limited to the use of selected descriptors, predicted properties or PLS latent variables. In principle any representation based on prior knowledge about the target property might be appropriate. Moreover, the use of prior knowledge suggests the application of a Bayesian experimental design, with the representation techniques we examined in this thesis. Measurements, collected by a literature search could act as the basis of such variants.

We focused on the investigation of the benefits of linear techniques, but we did not take non-linear kernel methods into consideration. With respect to the PLS-Optimal approach, e.g. kernel PLS, could be an interesting work to further extend the method we have developed in this article.

Furthermore, although not presented in this study, we conducted research on the combination of data mining techniques, such as the apriori algorithm with a compound representation using IUPAC names and hierarchical clustering techniques. The results indicated that such approaches are not applicable to comparably small, structurally diverse datasets, as they were used in this study. Still, in terms of large-scale in-vitro scanning, such approaches might be valuable. The application of structural alerts might be a promising field of study as well. As the alerts contain information about the target property these alerts could, for example, be used as preselected binary descriptors.

5.5 Final remarks

Finally, it has to be considered that the most important factor in the application of statistical experimental design is the human expertise. None of the presented approaches can be applied as a universal solution to any experimental design problem. The presented concepts are intended to serve as a sufficient support to thinking scientists.

The development of statistical models is carried out with a computer, but requires the supervision and control. Datasets have to be cleaned and maintained, the

selection of representative descriptors must be purposive, the decision for a machine-learning algorithm needs to be target-oriented and resulting models require an appropriate interpretation.

The same is true in terms of experimental design. Data cleaning, data representation and appropriate selection criteria have to be subject to a human decision. A permanent interaction is essential.

References

- (1) E.P. Council. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999/45/EC and Repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as Well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union* **2006**, 3–280.
- (2) Rovidá, C.; Hartung, T. Re-evaluation of Animal Numbers and Costs for in Vivo Tests to Accomplish REACH Legislation Requirements for Chemicals - a Report by the Transatlantic Think Tank for Toxicology (t(4)). *ALTEX* **2009**, *26*, 187–208.
- (3) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*.
- (4) Öberg, T. A QSAR for the Hydroxyl Radical Reaction Rate Constant: Validation, Domain of Application, and Prediction. *Atmos. Environ.* **2005**, *39*, 2189 – 2200.
- (5) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862–865.
- (6) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59–67.
- (7) Stenberg, M.; Linusson, A.; Tysklind, M.; Andersson, P. L. A Multivariate Chemical Map of Industrial Chemicals – Assessment of Various Protocols for Identification of Chemicals of Potential Concern. *Chemosphere* **2009**, *76*, 878 – 884.
- (8) Lahl, U.; Gundert-Remy, U. The Use of (Q)SAR Methods in the Context of REACH. *Toxicol. Mech. Method.* **2008**, *18*, 149–158.
- (9) Öberg, T.; Iqbal, M. S. The Chemical and Environmental Property Space of REACH Chemicals. *Chemosphere* **2012**, *87*, 975 – 981.
- (10) Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, Å.; Pettersen, J.; Bergman, R. Experimental Design and Optimization. *Chemometr. Intell. Lab.* **1998**, *42*, 3 – 40.
- (11) Eichler, U.; Ertl, P.; Gobbi, A.; Rohde, B. Definition of an Optimal Subset of Organic Substituents. Interactive Visual Comparison of Various Selection Algorithms. *Internet J. Chem.* **1999**, *2*.
- (12) Daszykowski, M.; Walczak, B.; Massart, D. L. Representative Subset Selection. *Anal. Chim. Acta* **2002**, *468*, 91 – 103.
- (13) Hopfinger, A. J. A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines Based Upon Molecular Shape Analysis. *Journal of the American Chemical Society* **1980**, *102*, 7196–7206.
- (14) Cowan, C. E.; Federle, T. W.; Larson, R. J.; Feijtel, T. C. J. Impact of Biodegradation Test Methods on the Development and Applicability of Biodegradation Qsars. *SAR and QSAR in Environmental Research* **1996**, *5*, 37–49.

- (15) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 276–285.
- (16) Thomsen, M.; Rasmussen, A. G.; Carlsen, L. SAR/QSAR Approaches to Solubility, Partitioning and Sorption of Phthalates. *Chemosphere* **1999**, *38*, 2613 – 2624.
- (17) Duprat, A. F.; Huynh, T.; Dreyfus, G. Toward a Principled Methodology for Neural Network Design and Performance Evaluation in QSAR. Application to the Prediction of LogP. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 586–594.
- (18) Paschke, A.; Popp, P.; Schüürmann, G. Water Solubility and Octanol/water-partitioning of Hydrophobic Chlorinated Organic Substances Determined by Using SPME/GC. *Fresenius' Journal of Analytical Chemistry* **1998**, *360*, 52–57.
- (19) Dearden, J. C. The QSAR Prediction of Melting Point, a Property of Environmental Relevance. *Science of The Total Environment* **1991**, *109*–*110*, 59 – 68.
- (20) Hall, L. H.; Story, C. T. Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks†. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 1004–1014.
- (21) Öberg, T.; Liu, T. Global and Local PLS Regression Models to Predict Vapor Pressure. *QSAR & Combinatorial Science* **2008**, *27*, 273–279.
- (22) Meyer, H. Zur Theorie der Alkoholnarkose. *Archiv für experimentelle Pathologie und Pharmakologie* **1899**, *42*, 109–118.
- (23) Mills, E. J. XXIII. On Melting-point and Boiling-point as Related to Chemical Composition. *Philosophical Magazine Series 5* **1884**, *17*, 173–187.
- (24) Hansch, C.; Fujita, T. P- σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society* **1964**, *86*, 1616–1626.
- (25) Hansch, C.; Leo, A. *Exploring QSAR*; 1st ed.; American Chemical Society: Washington, DC, USA, 1995.
- (26) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I. The Present Status of QSAR in Toxicology. *Journal of Molecular Structure: THEOCHEM* **2003**, *622*, 23 – 38.
- (27) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; 1st ed.; John Wiley & Sons: New York, USA, 2000.
- (28) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; 1st ed.; Wiley-VCH: Weinheim, Germany, 2001.
- (29) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; 2nd ed.; Wiley-VCH: Weinheim, Germany, 2009.
- (30) Estrada, E.; González, H. What Are the Limits of Applicability for Graph Theoretic Descriptors in QSPR/QSAR? Modeling Dipole Moments of Aromatic Compounds with TOPS-MODE Descriptors. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 75–84.
- (31) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solovev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided Drug Des.* **2008**, *4*, 191–198.

- (32) Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *Journal of Computer-Aided Molecular Design* **1990**, *4*, 1–103.
- (33) Veith, G. D.; Mekenyan, O. G.; Ankley, G. T.; Call, D. J. A QSAR Analysis of Substituent Effects on the Photoinduced Acute Toxicity of PAHs. *Chemosphere* **1995**, *30*, 2129 – 2142.
- (34) Thormann, M.; Vidal, D.; Almstetter, M.; Pons, M. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Name. *Open Appl. Inf. J.* **2007**, *1*, 28–32.
- (35) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37 – 52.
- (36) Domingos, P.; Pazzani, M. On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Machine Learning* **1997**, *29*, 103–130.
- (37) Williams, C. K. I.; Barber, D. Bayesian Classification with Gaussian Processes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **1998**, *20*, 1342–1351.
- (38) Quinlan, J. R. Induction of Decision Trees. *Machine Learning* **1986**, *1*, 81–106.
- (39) Quinlan, J. R. Simplifying Decision Trees. *International Journal of Man-Machine Studies* **1987**, *27*, 221 – 234.
- (40) Apté, C.; Damerau, F.; Weiss, S. M. Automated Learning of Decision Rules for Text Categorization. *ACM Trans. Inf. Syst.* **1994**, *12*, 233–251.
- (41) Dencœux, T. A k-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*; Yager, R.; Liu, L., Eds.; Studies in Fuzziness and Soft Computing; Springer Berlin Heidelberg, 2008; Vol. 219, pp. 737–760.
- (42) Minsky, M. Steps Toward Artificial Intelligence. *Proceedings of the IRE* **1961**, *49*, 8–30.
- (43) Krauth, W.; Mezard, M. Learning Algorithms with Optimal Stability in Neural Networks. *Journal of Physics A: Mathematical and General* **1987**, *20*, L745.
- (44) Eriksson, L.; Johansson, E. Multivariate Design and Modeling in QSAR. *Chemometr. Intell. Lab.* **1996**, *34*, 1 – 19.
- (45) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a Basic Tool of Chemometrics. *Chemometr. Intell. Lab.* **2001**, *58*, 109 – 130.
- (46) Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When Is “Nearest Neighbor” Meaningful? In *Database Theory — ICDT’99*; Beer, C.; Buneman, P., Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg, 1999; Vol. 1540, pp. 217–235.
- (47) Aha, D.; Kibler, D.; Albert, M. Instance-based Learning Algorithms. *Machine Learning* **1991**, *6*, 37–66.
- (48) Dudani, S. A. The Distance-Weighted k-Nearest-Neighbor Rule. *Systems, Man and Cybernetics, IEEE Transactions on* **1976**, *SMC-6*, 325–327.
- (49) Weinberger, K. Q.; Blitzer, J.; Saul, L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *In NIPS*; MIT Press, 2006.
- (50) Keller, J. M.; Gray, M. R.; Givens, J. A. A Fuzzy K-nearest Neighbor Algorithm. *Systems, Man and Cybernetics, IEEE Transactions on* **1985**, *SMC-15*, 580–585.
- (51) Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees*; 1st ed.; Chapman and Hall, 1984.
- (52) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

- (53) Cortes, C.; Vapnik, V. Support-vector Networks. *Machine Learning* **1995**, *20*, 273–297.
- (54) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory; COLT '92*; ACM: New York, NY, USA, 1992; pp. 144–152.
- (55) Crammer, K.; Singer, Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *J. Mach. Learn. Res.* **2002**, *2*, 265–292.
- (56) Schölkopf, B. *Statistical Learning and Kernel Methods*; 2000.
- (57) Rosenblatt, F. *The Perceptron, a Perceiving and Recognizing Automaton*; Cornell Aeronautical Laboratory, 1957.
- (58) Rosenblatt, F. Perceptron Simulation Experiments. *Proceedings of the IRE* **1960**, *48*, 301–309.
- (59) Anlauf, J. K.; Biehl, M. The AdaTron: An Adaptive Perceptron Algorithm. *EPL (Europhysics Letters)* **1989**, *10*, 687.
- (60) Pineda, F. J. Generalization of Back-propagation to Recurrent Neural Networks. *Phys. Rev. Lett.* **1987**, *59*, 2229–2232.
- (61) Cun, L.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. Handwritten Digit Recognition with a Back-Propagation Network. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann, 1990; pp. 396–404.
- (62) Schüürmann, G.; Muller, E. Back-propagation Neural Networks-recognition Vs. Prediction Capability. *Environmental Toxicology and Chemistry* **1994**, *13*, 743–747.
- (63) Gruau, F.; I, L. C. B.; Doctorat, O. A. D. D.; Demongeot, M. J.; Cosnard, E. M. M.; Mazoyer, M. J.; Peretto, M. P.; Whitley, M. D. *Neural Network Synthesis Using Cellular Encoding And The Genetic Algorithm.*; 1994.
- (64) Haikonen, P. O. Associative Neural Networks. In *Robot Brains*; John Wiley & Sons, Ltd, 2007; pp. 17–43.
- (65) Tetko, I. V. Associative Neural Network. *Methods in Molecular Biology* **2009**, *458*, 180–197.
- (66) Ziliak, S. T.; McCloskey, D. N. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*; University of Michigan Press, 2008.
- (67) Payton, M. E.; Greenstone, M. H.; Schenker, N. Overlapping Confidence Intervals or Standard Error Intervals: What Do They Mean in Terms of Statistical Significance? *J. Insect Sci.* **2003**, *3*.
- (68) Mason, J.; Pickett, S. Partition-based Selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85–114.
- (69) Lajiness, M. S. Dissimilarity-based Compound Selection Techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65–84.
- (70) Chaudhuri, B. B. How to Choose a Representative Subset from a Set of Data in Multi-dimensional Space. *Pattern Recognition Letters* **1994**, *15*, 893 – 899.
- (71) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J.; Osman, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285–289.
- (72) Roy, A.; Ghosal, S.; Rosenberger, W. F. Convergence Properties of Sequential Bayesian D-optimal Designs. *J. Stat. Plan. Infer.* **2009**, *139*, 425 – 440.
- (73) Chaloner, K.; Verdinelli, I. Bayesian Experimental Design: A Review. *Stat. Sci.* **1995**, *10*, pp. 273–304.

- (74) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148.
- (75) Baroni, M.; Clementi, S.; Cruciani, G.; Kettaneh-Wold, N.; Wold, S. D-Optimal Designs in QSAR. *Quant. Struct.-Act. Relat.* **1993**, *12*, 225–231.
- (76) Wold, S.; Josefson, M.; Gottfries, J.; Linusson, A. The Utility of Multivariate Design in PLS Modeling. *J. Chemometr.* **2004**, *18*, 156–165.
- (77) Van Den Berg, J.; Curtis, A.; Trampert, J. Optimal Nonlinear Bayesian Experimental Design: An Application to Amplitude Versus Offset Experiments. *Geophys. J. Int.* **2003**, *155*, 411–421.
- (78) Pronzato, L. One-step Ahead Adaptive D-optimal Design on a Finite Design Space Is Asymptotically Optimal. *Metrika* **2010**, *71*, 219–238.
- (79) Olsson, I.-M.; Gottfries, J.; Wold, S. D-optimal Onion Designs in Statistical Molecular Design. *Chemometr. Intell. Lab.* **2004**, *73*, 37 – 46.
- (80) Norrington, F. E.; Hyde, R. M.; Williams, S. G.; Wootton, R. Physicochemical-activity Relations in Practice. 1. Rational and Self-consistent Data Bank. *Journal of Medicinal Chemistry* **1975**, *18*, 604–607.
- (81) Wootton, R.; Cranfield, R.; Sheppey, G. C.; Goodford, P. J. Physicochemical-activity Relations in Practice. 2. Rational Selection of Benzenoid Substituents. *J. Med. Chem.* **1975**, *18*, 607–613.
- (82) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136–1145.
- (83) Tetko, I. V.; Poda, G. I.; Ostermann, C.; Mannhold, R. Large-Scale Evaluation of Log P Predictors: Local Corrections May Compensate Insufficient Accuracy and Need of Experimentally Testing Every Other Compound. *Chemistry & Biodiversity* **2009**, *6*, 1837–1844.
- (84) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharmaceut. Res.* **1990**, *7*, 801–807.
- (85) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
- (86) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; et al. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J. Comput. Aid. Mol. Des.* **2011**, *25*, 533–554.
- (87) US EPA. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.10, 2011.
- (88) Hilal, S. H.; Karickhoff, S. W.; Carreira, L. A. Prediction of the Vapor Pressure Boiling Point, Heat of Vaporization and Diffusion Coefficient of Organic Compounds. *QSAR & Combinatorial Science* **2003**, *22*, 565–574.
- (89) Dearden, J. C. Quantitative Structure-property Relationships for Prediction of Boiling Point, Vapor Pressure, and Melting Point. *Environmental Toxicology and Chemistry* **2003**, *22*, 1696–1709.
- (90) Bhatarai, B.; Teetz, W.; Liu, T.; Öberg, T.; Jeliaskova, N.; Kochev, N.; Pukalov, O.; Tetko, I. V.; Kovarich, S.; Papa, E.; et al. CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Molecular Informatics* **2011**, *30*, 189–204.

- (91) Krzyzaniak, J. F.; Myrdal, P. B.; Simamora, P.; Yalkowsky, S. H. Boiling Point and Melting Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds. *Industrial & Engineering Chemistry Research* **1995**, *34*, 2530–2535.
- (92) Öberg, T. Boiling Points of Halogenated Aliphatic Compounds: A Quantitative Structure–Property Relationship for Prediction and Validation. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 187–192.
- (93) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 1054–1060.
- (94) Meylan, W.; Howard, P. H.; Boethling, R. S. Molecular Topology/fragment Contribution Method for Predicting Soil Sorption Coefficients. *Environ. Sci. Technol.* **1992**, *26*, 1560–1567.
- (95) Gawlik, B. M.; Sotiriou, N.; Feicht, E. A.; Schulte-Hostede, S.; Kettrup, A. Alternatives for the Determination of the Soil Adsorption Coefficient, KOC, of Non-ionic organic Compounds - a Review. *Chemosphere* **1997**, *34*, 2525 – 2551.
- (96) Fiedler, H.; Schramm, K.-W. QSAR Generated Octanol-water Partition Coefficients of Selected Mixed Halogenated Dibenzodioxins and Dibenzofurans. *Chemosphere* **1990**, *20*, 1597 – 1602.
- (97) Reddy, K. N.; Locke, M. A. Relationships Between Molecular Properties and Log P and Soil Sorption (Koc) of Substituted Phenylureas: QSAR Models. *Chemosphere* **1994**, *28*, 1929 – 1941.
- (98) Sabljic, A.; Güsten, H.; Verhaar, H.; Hermens, J. QSAR Modelling of Soil Sorption. Improvements and Systematics of Log KOC Vs. Log KOW Correlations. *Chemosphere* **1995**, *31*, 4489 – 4514.
- (99) Hansen, B. G.; Paya-Perez, A. B.; Rahman, M.; Larsen, B. R. QSARs for KOW and KOC of PCB Congeners: A Critical Examination of Data, Assumptions and Statistical Approaches. *Chemosphere* **1999**, *39*, 2209 – 2228.
- (100) Gramatica, P.; Papa, E. An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors. *QSAR & Combinatorial Science* **2005**, *24*, 953–960.
- (101) (Q)SAR Model Reporting Format Inventory <http://qsardb.jrc.it/qmrf/> (accessed Sep 11, 2012).
- (102) Lu, X.; Tao, S.; Hu, H.; Dawson, R. W. Estimation of Bioconcentration Factors of Nonionic Organic Compounds in Fish by Molecular Connectivity Indices and Polarity Correction Factors. *Chemosphere* **2000**, *41*, 1675 – 1688.
- (103) Gramatica, P.; Papa, E. QSAR Modeling of Bioconcentration Factor by Theoretical Molecular Descriptors. *QSAR & Combinatorial Science* **2003**, *22*, 374–385.
- (104) Veith, G. D.; DeFoe, D. L.; Bergstedt, B. V. Measuring and Estimating the Bioconcentration Factor of Chemicals in Fish. *Journal of the Fisheries Research Board of Canada* **1979**, *36*, 1040–1048.
- (105) Mccarty, L. S. The Relationship Between Aquatic Toxicity QSARs and Bioconcentration for Some Organic Chemicals. *Environmental Toxicology and Chemistry* **1986**, *5*, 1071–1080.
- (106) Geyer, H.; Rimkus, G.; Scheunert, I.; Kaune, A.; Schramm, K.-W.; Kettrup, A.; Zeeman, M.; Muir, D. G.; Hansen, L.; Mackay, D. Bioaccumulation and Occurrence of Endocrine-Disrupting Chemicals (EDCs), Persistent Organic Pollutants (POPs), and Other Organic Compounds in Fish and Other Organisms Including Humans. In

- Bioaccumulation – New Aspects and Developments*; Beek, B., Ed.; The Handbook of Environmental Chemistry; Springer Berlin Heidelberg, 2000; Vol. 2J, pp. 1–166.
- (107) Khadikar, P. V.; Singh, S.; Mandloi, D.; Joshi, S.; Bajaj, A. V. QSAR Study on Bioconcentration Factor (BCF) of Polyhalogenated Biphenyls Using the PI Index. *Bioorganic & Medicinal Chemistry* **2003**, *11*, 5045 – 5050.
- (108) Papa, E.; Dearden, J. C.; Gramatica, P. Linear QSAR Regression Models for the Prediction of Bioconcentration Factors by Physicochemical Properties and Structural Theoretical Molecular Descriptors. *Chemosphere* **2007**, *67*, 351 – 358.
- (109) Schreiber, R.; Altenburger, R.; Paschke, A.; Schüürmann, G.; Küster, E. A Novel in Vitro System for the Determination of Bioconcentration Factors and the Internal Dose in Zebrafish (*Danio Rerio*) Eggs. *Chemosphere* **2009**, *77*, 928 – 933.
- (110) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales Promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948–967.
- (111) Cronin, M. T. D.; Dearden, J. C. QSAR in Toxicology. 1. Prediction of Aquatic Toxicity. *Quantitative Structure-Activity Relationships* **1995**, *14*, 1–7.
- (112) Neely, W. B. An Analysis of Aquatic Toxicity Data: Water Solubility and Acute LC50 Fish Data. *Chemosphere* **1984**, *13*, 813 – 819.
- (113) McCarty, L. S. Model Validation in Aquatic Toxicity Testing: Implications for Regulatory Practice. *Regulatory Toxicology and Pharmacology* **2012**, *63*, 353 – 362.
- (114) Raevsky, O. A.; Grigor'ev, V. Y.; Dearden, J. C.; Weber, E. E. Classification and Quantification of the Toxicity of Chemicals to Guppy, Fathead Minnow, and Rainbow Trout. Part 2. Polar Narcosis Mode of Action. *QSAR & Combinatorial Science* **2009**, *28*, 163–174.
- (115) Russom, C. L.; Breton, R. L.; Walker, J. D.; Bradbury, S. P. An Overview of the Use of Quantitative Structure-activity Relationships for Ranking and Prioritizing Large Chemical Inventories for Environmental Risk Assessments. *Environ. Toxicol. Chem.* **2003**, *22*, 1810–1821.
- (116) Schüürmann, G. QSAR Analysis of the Acute Fish Toxicity of Organic Phosphorothionates Using Theoretically Derived Molecular Descriptors. *Environmental Toxicology and Chemistry* **1990**, *9*, 417–428.
- (117) Papa, E.; Villa, F.; Gramatica, P. Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in *Pimephales Promelas* (Fathead Minnow). *J. Chem. Inf. Model.* **2005**, *45*, 1256–1266.
- (118) Pavan, M.; Netzeva, T. I.; Worth, A. P. Validation of a QSAR Model for Acute Toxicity. *SAR and QSAR in Environmental Research* **2006**, *17*, 147–171.
- (119) In, Y.-Y.; Lee, S.-K.; Kim, P.-J.; No, K.-T. Prediction of Acute Toxicity to Fathead Minnow by Local Model Based QSAR and Global QSAR Approaches. *Bull. Korean Chem* **2012**, *33*, 613–619.
- (120) Martin, T. M.; Young, D. M. Prediction of the Acute Toxicity (96-h LC50) of Organic Compounds to the Fathead Minnow (*Pimephales Promelas*) Using a Group Contribution Method. *Chemical Research in Toxicology* **2001**, *14*, 1378–1385.
- (121) Ankley, G. T.; Villeneuve, D. L. The Fathead Minnow in Aquatic Toxicology: Past, Present and Future. *Aquatic Toxicology* **2006**, *78*, 91 – 102.
- (122) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity Against *Tetrahymena Pyriformis*: Focusing on

- Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.
- (123) Schultz, T. W. Tetratox: Tetrahymena Pyriformis Population Growth Impairment Endpoint – a Surrogate for Fish Lethality. *Toxicol. Mech. Method.* **1997**, *7*, 289–309.
- (124) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. Evaluation of QSARs for Ecotoxicity: A Method for Assigning Quality and Confidence. *SAR and QSAR in Environmental Research* **2004**, *15*, 385–397.
- (125) Aptula, A. O.; Roberts, D. W.; Cronin, M. T. D.; Schultz, T. W. Chemistry–Toxicity Relationships for the Effects of Di- and Trihydroxybenzenes to Tetrahymena Pyriformis. *Chemical Research in Toxicology* **2005**, *18*, 844–854.
- (126) Schultz, T. W.; Hewitt, M.; Netzeva, T. I.; Cronin, M. T. D. Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. *QSAR & Combinatorial Science* **2007**, *26*, 238–254.
- (127) Environmental Toxicity Prediction Challenge <http://cadaster.eu/node/65> (accessed Apr 11, 2013).
- (128) Ames, B. N.; Lee, F. D.; Durston, W. E. An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens. *Proceedings of the National Academy of Sciences* **1973**, *70*, 782–786.
- (129) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K.-R. Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077–2081.
- (130) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; et al. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094–2111.
- (131) Novotarskyi, S.; Sushko, I.; Körner, R.; Pandey, A. K.; Tetko, I. V. Classification of CYP450 1A2 Inhibitors Using PubChem Data. *Journal of Cheminformatics* **2010**, *2*, 1–1.
- (132) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon Software: An Easy Approach to Molecular Descriptor Calculations. *Match Communications In Mathematical And In Computer Chemistry* **2006**, *56*, 237–248.
- (133) Puzyn, T.; Gajewicz, A.; Rybacka, A.; Haranczyk, M. Global Versus Local QSPR Models for Persistent Organic Pollutants: Balancing Between Predictivity and Economy. *Structural Chemistry* **2011**, *22*, 873–884.
- (134) Fernandez, P.; Grimalt, J. O. On the Global Distribution of Persistent Organic Pollutants. *CHIMIA International Journal for Chemistry* **2003**, *57*, 514–521.
- (135) Damgaard, I. N.; Skakkebaek, N. E.; Toppari, J.; Virtanen, H. E.; Shen, H.; Schramm, K.-W.; Petersen, J. H.; Jensen, T. K.; Main, K. M. Persistent Pesticides in Human Breast Milk and Cryptorchidism. *Environmental Health Perspectives* **2006**, *114*, 1133–1138.
- (136) Shen, H.; Main, K. M.; Andersson, A.-M.; Damgaard, I. N.; Virtanen, H. E.; Skakkebaek, N. E.; Toppari, J.; Schramm, K.-W. Concentrations of Persistent Organochlorine Compounds in Human Milk and Placenta Are Higher in Denmark Than in Finland. *Human Reproduction* **2008**, *23*, 201–210.
- (137) Sáez, M.; Voogt, P.; Parsons, J. Persistence of Perfluoroalkylated Substances in Closed Bottle Tests with Municipal Sewage Sludge. *Environmental Science and Pollution Research* **2008**, *15*, 472–477.

- (138) Frömel, T.; Knepper, T. Biodegradation of Fluorinated Alkyl Substances. In *Reviews of Environmental Contamination and Toxicology Volume 208*; De Voogt, P., Ed.; Reviews of Environmental Contamination and Toxicology; Springer New York, 2010; Vol. 208, pp. 161–177.
- (139) Ahrens, L.; Xie, Z.; Ebinghaus, R. Distribution of Perfluoroalkyl Compounds in Seawater from Northern Europe, Atlantic Ocean, and Southern Ocean. *Chemosphere* **2010**, *78*, 1011 – 1016.
- (140) N, K. S.; E, T. C.; Konstanze, G.; Ibrahim, C. *Developmental Exposure to Low-dose PBDE-99: Effects on Male Fertility and Neurobehavior in Rat Offspring*; Environmental health perspectives; US Department of Health and Human Services: Research Triangle Park, NC, ETATS-UNIS, 2005; Vol. 113.
- (141) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.
- (142) Worth, A. P.; Van Leeuwen, C. J.; Hartung, T. The Prospects for Using (Q)SARs in a Changing Political Environment–high Expectations and a Key Role for the European Commission’s Joint Research Centre. *SAR and QSAR in Environmental Research* **2004**, *15*, 331–343.
- (143) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J.; Kahn, S.; Klopman, G.; Marchant, C. A.; et al. Current Status of Methods for Defining the Applicability Domain of (quantitative) Structure–activity Relationships. *Altern. Lab. Anim.* **2005**, *33*, pp. 155–173.
- (144) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADME–Tox Predictions? *Drug Discovery Today* **2006**, *11*, 700 – 707.
- (145) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *Journal of Chemical Information and Modeling* **2005**, *45*, 839–849.
- (146) Nikolova-Jeliazkova, N.; Jaworska, J. *An Approach to Determining Applicability Domains for QSAR Group Contribution Models: An Analysis of SRC KOWWIN*; ATLA. Alternatives to laboratory animals; Fund for the Replacement of Animals in Medical Experiments: Nottingham, ROYAUME-UNI, 2005; Vol. 33.
- (147) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability Domain for in Silico Models to Achieve Accuracy of Experimental Measurements. *J. Chemometr.* **2010**, *24*, 202–208.
- (148) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1912–1928.
- (149) OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models
<http://www.oecd.org/dataoecd/33/37/37849783.pdf>.
- (150) Scior, T.; Medina-Franco, J. L.; Do, Q.-T.; Martinez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies:A Critical Review. *Current Medicinal Chemistry* **2009**, *16*, 4297–4313.
- (151) Akaike, H. A New Look at the Statistical Model Identification. *Automatic Control, IEEE Transactions on* **1974**, *19*, 716–723.

- (152) Akaike, H. Likelihood and the Bayes Procedure. *Trabajos de Estadística Y de Investigación Operativa* **1980**, *31*, 143–166.
- (153) Schwarz, G. Estimating the Dimension of a Model. *Annals of Statistics* **1978**, *6*, 461–464.
- (154) Hannan, E. J.; Quinn, B. G. The Determination of the Order of an Autoregression. *J. R. Statist. Soc* **1979**, *41*, 190–195.
- (155) Schüürmann, G.; Ebert, R.-U.; Kühne, R. Quantitative Read-Across for Predicting the Acute Fish Toxicity of Organic Compounds. *Environmental Science & Technology* **2011**, *45*, 4616–4622.
- (156) He, L.; Jurs, P. C. Assessing the Reliability of a QSAR Model's Predictions. *Journal of Molecular Graphics and Modelling* **2005**, *23*, 503 – 523.
- (157) Brust, K. Toxicity of Aliphatic Amines on the Embryos of Zebrafish Danio Rerio - Experimental Studies and QSAR. *Saechsische Landesbibliothek - Staats- und Universitaetsbibliothek Dresden* **2002**.
- (158) PPDB: Pesticide Properties DataBase
<http://sitem.herts.ac.uk/aeru/footprint/en/index.htm> (accessed Sep 11, 2012).
- (159) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: a New General Purpose Quantum Mechanical Molecular Model. *Journal of the American Chemical Society* **1985**, *107*, 3902–3909.
- (160) Ghafourian, T.; Dearden, J. C. The Use of Atomic Charges and Orbital Energies as Hydrogen-bonding-donor Parameters for QSAR Studies: Comparison of MNDO, AM1 and PM3 Methods. *Journal of Pharmacy and Pharmacology* **2000**, *52*, 603–610.
- (161) Potemkin, V.; Grishina, M. A New Paradigm for Pattern Recognition of Drugs. *Journal of Computer-Aided Molecular Design* **2008**, *22*, 489–505.
- (162) Potemkin, V. A.; Pogrebnoy, A. A.; Grishina, M. A. Technique for Energy Decomposition in the Study of “Receptor-Ligand” Complexes. *Journal of Chemical Information and Modeling* **2009**, *49*, 1389–1406.
- (163) Chemaxon (2010) Chemaxon — toolkits and desktop applications for chemoinformatics: calculator Plugins.
<http://www.chemaxon.com/library/scientific-presentations/calculator-plugins/> (accessed May 24, 2011).
- (164) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design* **2006**, *12*, 2111–2120.
- (165) Stewart, J. P. Optimization of Parameters for Semiempirical Methods IV: Extension of MNDO, AM1, and PM3 to More Main Group Elements. *Journal of Molecular Modeling* **2004**, *10*, 155–164.
- (166) Cherkasov, A.; Ban, F.; Santos-Filho, O.; Thorsteinson, N.; Fallahi, M.; Hammond, G. L. An Updated Steroid Benchmark Set and Its Application in the Discovery of Novel Nanomolar Ligands of Sex Hormone-Binding Globulin. *Journal of Medicinal Chemistry* **2008**, *51*, 2047–2056.
- (167) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. The Electronegativity Equalization Method I: Parametrization and Validation for Atomic Charge Calculations. *The Journal of Physical Chemistry A* **2002**, *106*, 7887–7894.

- (168) Bultinck, P.; Langenaeker, W.; Carbó-Dorca, R.; Tollenaere, J. P. Fast Calculation of Quantum Chemical Molecular Descriptors from the Electronegativity Equalization Method. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 422–428.
- (169) Schüürmann, G.; Ebert, R.-U.; Chen, J.; Wang, B.; Kühne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean Vs Training Set Activity Mean. *Journal of Chemical Information and Modeling* **2008**, *48*, 2140–2145.
- (170) Holmes, G.; Donkin, A.; Witten, I. H. WEKA: a Machine Learning Workbench. In; 1994; pp. 357–361.
- (171) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Model Selection Based on Structural Similarity–Method Description and Application to Water Solubility Prediction. *Journal of Chemical Information and Modeling* **2006**, *46*, 636–641.
- (172) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Chemical Domain of QSAR Models from Atom-Centered Fragments. *Journal of Chemical Information and Modeling* **2009**, *49*, 2660–2669.
- (173) Netzeva, T. I.; Pavan, M.; Worth, A. P. Review of (Quantitative) Structure–Activity Relationships for Acute Aquatic Toxicity. *QSAR & Combinatorial Science* **2008**, *27*, 77–90.
- (174) Pirovano, A.; Huijbregts, M. A. J.; Ragas, A. M. J.; Hendriks, A. J. Compound Lipophilicity as a Descriptor to Predict Binding Affinity (1/Km) in Mammals. *Environmental Science & Technology* **2012**, *46*, 5168–5174.
- (175) Veith, G. D.; Greenwood, B.; Hunter, R. S.; Niemi, G. J.; Regal, R. R. On the Intrinsic Dimensionality of Chemical Structure Space. *Chemosphere* **1988**, *17*, 1617–1630.
- (176) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *Journal of Chemical Information and Modeling* **2012**, *52*, 2310–2316.
- (177) Nilsson, H.; Kärrman, A.; Westberg, H.; Rotander, A.; van Bavel, B.; Lindström, G. A Time Trend Study of Significantly Elevated Perfluorocarboxylate Levels in Humans after Using Fluorinated Ski Wax. *Environmental Science & Technology* **2010**, *44*, 2150–2155.
- (178) Aguiar, P. F. de; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Thau-Luu, R. D-optimal Designs. *Chemometr. Intell. Lab.* **1995**, *30*, 199–210.
- (179) Fedorov, V. *Theory of Optimal Experiments*; Academic Press, 1972.
- (180) Wikel, J. H.; Dow, E. R. The Use of Neural Networks for Variable Selection in QSAR. *Bioorganic & Medicinal Chemistry Letters* **1993**, *3*, 645–651.
- (181) Burden, F. R.; Winkler, D. A. An Optimal Self-Pruning Neural Network and Nonlinear Descriptor Selection in QSAR. *QSAR & Combinatorial Science* **2009**, *28*, 1092–1097.
- (182) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. *Journal of Chemical Information and Computer Sciences* **1995**, *35*, 77–84.
- (183) Yasri, A.; Hartsough, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1218–1227.

- (184) Godden, J. W.; Bajorath, J. An Information-Theoretic Approach to Descriptor Selection for Database Profiling and QSAR Modeling. *QSAR & Combinatorial Science* **2003**, *22*, 487–497.
- (185) Lundstedt, T.; Thelin, B. A Multivariate Strategy for Optimizing a Two-step Process. *Chemometr. Intell. Lab.* **1995**, *29*, 255 – 261.
- (186) Geladi, P.; Kowalski, B. R. Partial Least-squares Regression: a Tutorial. *Analytica Chimica Acta* **1986**, *185*, 1 – 17.
- (187) Rännar, S.; Andersson, P. L. A Novel Approach Using Hierarchical Clustering To Select Industrial Chemicals for Environmental Impact Assessment. *J. Chem. Inf. Model.* **2010**, *50*, 30–36.
- (188) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. Cluster-based Design in Environmental QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 383–390.
- (189) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Computer 1996*, 226–231.
- (190) Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; Sander, J. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Rec.* **1999**, *28*, 49–60.
- (191) Mahajan, M.; Nimbhorkar, P.; Varadarajan, K. The Planar k-Means Problem Is NP-Hard. In *WALCOM: Algorithms and Computation*; Das, S.; Uehara, R., Eds.; Lecture Notes in Computer Science; Springer Berlin / Heidelberg, 2009; Vol. 5431, pp. 274–285.
- (192) Valle, S.; Li, W.; Qin, S. J. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods†. *Ind. Eng. Chem. Res.* **1999**, *38*, 4389–4401.
- (193) Qin, S. J.; Dunia, R. Determining the Number of Principal Components for Best Reconstruction. *J. Process Contr.* **2000**, *10*, 245 – 250.
- (194) Hoffmann, T.; Hofmann, D.; Klumpp, E.; Küppers, S. Electrochemistry-mass Spectrometry for Mechanistic Studies and Simulation of Oxidation Processes in the Environment. *Analytical and Bioanalytical Chemistry* **2011**, *399*, 1859–1868.
- (195) III, W. J. D.; Scott, D. R.; Glen, W. G. Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Computer Methodology* **1989**, *2*, 349 – 376.
- (196) Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S. Fragmental Descriptors with Labeled Atoms and Their Application in QSAR/QSPR Studies. *Doklady Chemistry* **2007**, *417*, 282–284.
- (197) Brandmaier, S.; Teetz, W.; Sahigara, F.; Ebert, R.-U.; Abdelaziz, A.; Salmina, E.; Kühne, R.; Ehret, J.; Tetko, I. V.; Schramm, K.-W.; et al. Estimating Aquatic Toxicity in Fish: A Three Descriptor Solution. *In preparation* **2013**.
- (198) Brandmaier, S.; Sahlin, U.; Tetko, I. V.; Öberg, T. PLS-Optimal: A Stepwise D-Optimal Design Based on Latent Variables. *J. Chem. Inf. Model.* **2012**, *52*, 975–983.
- (199) Brandmaier, S.; Tetko, I. V. Robustness in Experimental Design: A Study on the Reliability of Selection Approaches. *Computational and Structural Biotechnology Journal* **2013**, *7*.
- (200) Peijnenburg, W.; Tetko, I. V. Exemplification of the Implementation of Alternatives to Experimental Testing in Chemical Risk Assessment – Case Studies from Within the CADASTER Project. *ATLA Alternatives to Laboratory Animals* **2013**, *41*, 13–17.

- (201) Brandmaier, S.; Tetko, I. V.; Öberg, T. An Evaluation of Experimental Design in QSAR Modelling Utilizing the K-medoid Clustering. *Journal of Chemometrics* **2012**, *26*, 509–517.
- (202) Fleischer, M. Testing Costs and Testing Capacity According to the REACH Requirements : Results of a Survey of Independent and Corporate GLP Laboratories in the EU and Switzerland. *Journal of Business Chemistry* **2007**, *4*, 96–114.
- (203) Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Model.* **2004**, *44*, 1257–1266.
- (204) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure–Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.

List of cited publications and studies

- a. **Brandmaier, S.**; Sahlin, U.; Tetko, I. V.; Öberg, T. PLS-Optimal: A Stepwise D-Optimal Design Based on Latent Variables. *J. Chem. Inf. Model.* 2012, 52, 975–983. (submitted version)
- b. **Brandmaier, S.**; Tetko, I. V.; Öberg, T. An evaluation of experimental design in QSAR modelling utilizing the k-medoid clustering. *J. Chemom.* 2012, 26, 509–517. (submitted version)
- c. **Brandmaier, S.**; Tetko, I. V. Robustness in experimental design: A study on the reliability of selection approaches. *Comput. Struct. Biotechnol. J.* 2013, 40 (1), pp 33–47. (submitted version)
- d. **Brandmaier, S.**; Teetz, W.; Sahigara, F.; Ebert, R.-U.; Abdelaziz, A.; Salmina, E.; Kühne, R.; Ehret, J.; Tetko, I. V.; Schramm, K.-W.; Schüürmann, G. Estimating aquatic toxicity in fish: A three descriptor solution. In preparation. 2013.
- e. **Brandmaier, S.**; Novotarskyi, S.; Sushko, Y.; Tetko, I. V. From Descriptors to Predicted Properties: Experimental Design by Using Applicability Domain Estimation. *Atla Altern. Lab. Anim.* 2013, 41, 33–47. (submitted version)

Hiermit erkläre ich an Eides statt, dass ich alleiniger, federführender Hauptautor der fünf oben genannten Publikationen und Studien bin, die in dieser Arbeit wörtlich zitiert wurden. Die betreffenden Passagen wurden ausschließlich von mir verfasst.

Ort, Datum

Unterschrift

Software used

Figures 1, 3 were partially generated with **Jmol: an open-source Java viewer for chemical structures in 3D** (<http://www.jmol.org/>)

Figures 2, 8, 12, 18 were partially generated with **yED**, Copyright © 2013 yWorks (http://www.yworks.com/de/products_yed_about.html)

Figures 4, 5, 14, 15, 35, 41, 43, 46-48, 50-62, 64, 66, 67 were partially generated with **R**, Copyright © 2004-2013 The R Foundation for Statistical Computing (<http://www.R-project.org>)

Figures 7, 9, 11 were partially generated with **Preview**, Copyright © 2002–2012 Apple Inc. (<http://www.apple.com/>)

Figures 6, 39, 45 were partially generated with **Powerpoint**, Copyright © 2013 Microsoft Corporation (<http://office.microsoft.com/>)

Figures 10, 13, 16, 17, 19-33, 37, 38, 40, 42, 44, 49, 63, 65, 68- 71 were partially generated with **Orange, Data Mining Fruitful & Fun** (<http://orange.biolab.si/>)

Molecule depictions in figures 34, 36, 38 and tables 17, 18 were taken from **OCHEM** (<https://ochem.eu/>) and originally generated with **Instant JChem**, Copyright © 1998-2013 ChemAxon Ltd. (<http://www.chemaxon.com/>)

List of publications

- Arnold, R.; **Brandmaier, S.**; Kleine, F.; Tischler, P.; Heinz, E.; Behrens, S.; Niinikoski, A.; Mewes, H.-W.; Horn, M.; Rattei, T. Sequence-Based Prediction of Type III Secreted Proteins. *Plos Pathog* 2009, 5, e1000376.
- Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; **Brandmaier, S.**; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* 2011, 25, 533–554.
- **Brandmaier, S.**; Sahlin, U.; Tetko, I. V.; Öberg, T. PLS-Optimal: A Stepwise D-Optimal Design Based on Latent Variables. *J. Chem. Inf. Model.* 2012, 52, 975–983.
- **Brandmaier, S.**; Tetko, I. V.; Öberg, T. An evaluation of experimental design in QSAR modelling utilizing the k-medoid clustering. *J. Chemom.* 2012, 26, 509–517.
- Cassani, S.; Kovarich, S.; Papa, E.; Roy, P. P.; Rahmberg, M.; Nilsson, S.; Sahlin, U.; Jeliaskova, N.; Kochev, N.; Pukalov, O.; Tetko, I. V.; **Brandmaier, S.**; Kos Durjava, M.; Kolar, B.; Peijnenburg, W.; Gramatica, P. Evaluation of CADASTER QSAR Models for the Aquatic Toxicity of (Benzo)triazoles and Prioritisation by Consensus Prediction. *Atla Altern. Lab. Anim.* 2013, 41, 49–64.
- Tetko, I. V.; Sopasakis, P.; Kunwar, P.; **Brandmaier, S.**; Novotarskyi, S.; Charochkina, L.; Prokopenko, V.; Peijnenburg, W. J. G. M. Prioritisation of Polybrominated Diphenyl Ethers (PBDEs) by Using the QSPR-THESAURUS Web Tool. *Atla Altern. Lab. Anim.* 2013, 41, 127–135.
- **Brandmaier, S.**; Novotarskyi, S.; Sushko, Y.; Tetko, I. V. From Descriptors to Predicted Properties: Experimental Design by Using Applicability Domain Estimation. *Atla Altern. Lab. Anim.* 2013, 41, 33–47.
- **Brandmaier, S.**; Tetko, I. V. Robustness in experimental design: A study on the reliability of selection approaches. *Comput. Struct. Biotechnol. J.* 2013, 40 (1), pp 33–47.
- **Brandmaier, S.**; Teetz, W.; Sahigara, F.; Ebert, R.-U.; Abdelaziz, A.; Salmina, E.; Kühne, R.; Ehret, J.; Tetko, I. V.; Schramm, K.-W.; Schüürmann, G. Estimating aquatic toxicity in fish: A three descriptor solution. In preparation. 2013.
- Pirovano, A.; **Brandmaier, S.**; Huijbregts, M. A. J.; Ragas, A. M. J.; Veltman, K.; Tetko, I. V.; Hendriks, A. J. The utilization of structural descriptors to assess the metabolic affinity (1/Km) and maximum velocity (Vmax) of xenobiotics in mammals. In preparation. 2013.
- **Brandmaier, S.**; Peijnenburg, W.; Kos Durjava, M.; Kolar, B.; Gramatica, P.; Papa, E.; Bhatarai, B.; Kovarich, S.; Cassani, S.; D'Onofrio, E.; Rahmberg, M.; Öberg, T.; Jeliaskova, N.; Golsteijn, L.; Comber, M.; Ruggiu, F.; Novotarskyi, S.; Sushko, I.; Kunwar, P.; Abdelaziz, A.; Tetko, I. V. The QSPR-Thesaurus: The online platform of the CADASTER project. Accepted by *Atla Altern. Lab. Anim.* 2013
- Tetko, I. V.; Schramm, K.-W.; Knepper, T.; Peijnenburg, W. J. G. M.; Hendriks, A. J.; Navas, J. M.; Nicholls, I. A.; Öberg, T.; Todeschini, R.; Schlosser, E.; **Brandmaier, S.** Preface to the Special ATLA Edition. Accepted by *Atla Altern. Lab. Anim.* 2013

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die bei der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt (promotionsführende Einrichtung) der TUM zur Promotionsprüfung vorgelegte Arbeit mit dem Titel:

Experimental design methods to increase the accuracy of in silico models

im Institut für Bioinformatik und System Biologie des Helmholtz Zentrums München unter der Anleitung durch:

Prof. Dr. Hans-Werner Mewes

und unter der Betreuung durch:

Dr. Igor V. Tetko

ohne sonstige Hilfe erstellt und bei der Abfassung nur die gemäß § 6 Abs. 6 und 7 Satz 2 angegebenen Hilfsmittel benutzt habe.

- Ich habe keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen und Betreuer für die Anfertigung von Dissertationen sucht, oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt.
- Ich habe die Dissertation in dieser oder ähnlicher Form in keinem anderen Prüfungsverfahren als Prüfungsleistung vorgelegt
- Ich habe den angestrebten Doktorgrad noch nicht erworben und bin nicht in einem früheren Promotionsverfahren für den angestrebten Doktorgrad endgültig gescheitert

Die öffentlich zugängliche Promotionsordnung der TUM ist mir bekannt, insbesondere habe ich die Bedeutung von § 28 (Nichtigkeit der Promotion) und § 29 (Entzug des Doktorgrades) zur Kenntnis genommen. Ich bin mir der Konsequenzen einer falschen Eidesstattlichen Erklärung bewusst. Mit der Aufnahme meiner personenbezogenen Daten in die Alumni-Datei bei der TUM bin ich einverstanden

Ort, Datum

Unterschrift

Curriculum Vitae

Persönliche Daten

Stefan Josef Brandmaier
Am Klösterlmoos 3
85716 Unterschleißheim
Tel.: 089 / 37 50 84 46
Geboren am 05.01.1977 in Augsburg, ledig

Akademische Ausbildung

Januar 2013	- gegenwärtig	Wissenschaftler am Helmholtz Zentrum München
März 2009	- Dezember 2012	Doktorand am Helmholtz Zentrum München
September 2002	- September 2008	Studium der Bioinformatik an der TU / LMU, München Abschluss: Diplom (Abschlussnote 1,5)
September 1998	- März 2002	Studium der Humanmedizin an der LMU, München
September 1987	- Juni 1996	Carl-Orff-Gymnasium, Unterschleißheim Abschluss: Allgemeinen Hochschulreife (Abschlussnote 1,9)
September 1983	- September 1987	Raiffeisen-Grundschule, Unterschleißheim

Zivildienst

März 1997	- April 1998	Zivildienstes im GSF-Forschungszentrum, Abteilung für Technischen Strahlenschutz
-----------	--------------	--

Nebentätigkeiten

Februar 2008	- April 2008	Anstellung als studentische Hilfskraft am WZW, Lehrstuhl für Genomorientierte Bioinformatik (Expressionsdatenanalyse)
April 2005	- Dezember 2005	Anstellung als studentische Hilfskraft am IFI der LMU, LFE Bioinformatik (Textmining, Machine-Learning-Anwendungen)
Mai 2000	- Mai 2003	Anstellung als Sitz und Sonderwache, im MKI
März 2000	- April 2000	Pflegepraktikum in der Medizinischen Klinik, Gastroskopie

Besondere Kenntnisse

Sprachen	Deutsch, Englisch
EDV-Kenntnisse	Sehr gute Programmier-Kenntnisse in Java, Perl, SQL

