Influence of molecular steric factors on the sorption of organic chemicals

Dissertation

zur Erlangung des
Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät II – Chemie, Physik und Mathematik
der Martin-Luther-Universität Halle-Wittenberg

vorgelegt

von Herrn Dipl. Chem. Lukas Linden
geboren am 29. Juni 1987 in Mainz

Gutachter: 1. Prof. Dr. Kai-Uwe Goss
2. Prof. Dr.-Ing. Irina Smirnova

Verteidigt am: 8. Juni 2017

Abstract

Molecular steric effects can strongly influence a variety of processes. In the area of environmental research, they are especially relevant for partitioning processes between a homogeneous phase, e.g., water, and a heterogeneous matrix, e.g., proteins. There is no sharp distinction between a homogenous phase on the one side and a complex heterogeneous matrix on the other side (whose physiochemical properties are position-dependent and thus shows an influence of the molecular steric properties) but rather a gradual transition. Depending on the progress of this transition, it is not always necessary to consider the molecular steric effects for a successful modeling of the partition systems. Examples for partitioning systems, in which the influence of the molecular steric effects are typically neglected for the modeling, are the partitioning between natural organic matter and water and between mineral surfaces in soils and water.

The aim of this work was i) to perform partitioning experiments and quantify partitioning coefficients, which are influenced by molecular steric effects and ii) to model the investigated partition processes. The partition system α-cyclodextrin (αCD)-water was chosen, because the 3D-structure of αCD, is well-defined. CDs are used in various areas (e.g., as pharmaceutical excipients, as additives in cosmetics and food, and for the remediation of contaminated soils). Two experimental methods, both of which are mass balance based, were applied for the determination of αCD binding constants: a head space and a passive sampling method. The measured 70 neutral organic chemicals have binding constants in a range of 1.08 to 4.97 log units. The selection of the chemicals included chemicals with different functional groups and several homologous series. This selection enables a good comparability of the different binding constants. The results show that the binding to αCD is clearly influenced by steric effects, e.g., constitutional isomers have differences in their binding constants of up to 1.2 log units, which are caused by the different positions of the functional groups. Moreover the

dataset reveals that the spatial restrictions of the αCD cavity are responsible for the binding strength of differently sized, hydrophobic, aromatic chemicals.

The αCD-water partitioning system was then described with three different modeling approaches, which were evaluated in regard to the quality of their predictions, especially focusing on the respective description of the steric effects. The three modeling approaches were a) a poly-parameter linear free energy relationship, b) a comparative molecular field analysis, and c) a 3D quantitative structure activity relationship (QSAR). The COSMO (conductor like screening model) based 3D-QSAR resulted in the best predictions (R_{test}^2=0.70, $RMSE_{test}$=0.45, n=15) and it was the only modeling approach that was able to reproduce the molecular steric effects. In addition, the COSMO based 3D-QSAR gave good predictions for 88 αCD-binding data from the literature (R_{test}^2=0.64, $RMSE_{test}$=0.59). Hence, we concluded that this modeling approach can be used for the prediction of unknown αCD binding constants and it should be applicable to comparable partitioning processes.

The further applicability of the 3D-QSAR method was tested with a prominent toxicokinetic/pharmacokinetic example, the partitioning between bovine serum albumin (BSA) and water. This process is relevant for the distribution of chemicals in all vertebrates, because all vertebrates express the highly conserved protein serum albumin. This work revealed that the partitioning between BSA and water is influenced by molecular steric effects, particularly for organic anions. The COSMO based 3D-QSAR predicted experimental BSA-water partition coefficients ($K_{BSA/water}$) not only with an overall satisfying accuracy (R_{test}^2=0.52, $RMSE_{test}$=0.63, n=32) but it also captured the molecular steric effects, which are responsible for differences in the partitioning coefficients of up to 2 log units for charged isomers. The domain of applicability of this empirical model is largely determined by the used calibration (42 anions und 88 neutral chemicals). Thus, an extension and diversification of the calibration dataset, especially by including organic cations and zwitterions, would be useful to allow a broader applicability of the model. The COSMO based 3D-QSAR model can now be

used for an estimation of the distribution of anionic and neutral chemicals in vertebrates and thus enables an improved assessment of the toxicokinetics of negatively charged chemicals, as long as it is used within the domain of applicability.

Zusammenfassung

Molekulare sterische Effekte haben einen großen Einfluss auf eine Vielzahl von Prozessen. Im Bereich der Umweltforschung sind sie insbesondere relevant bei Verteilungsprozessen zwischen einer homogenen Phase, z.B. Wasser, und einer heterogenen Matrix, z.B. Proteinen. Der Übergang zwischen der homogenen Phase und der heterogenen Matrix (bei der die physiochemischen Eigenschaften ortsabhängig sind und die deshalb einen Einfluss von molekularen sterischen Effekten zeigt) ist hierbei fließend und muss, je nach Ausmaß des Effektes, nicht unbedingt für eine erfolgreiche Modellierung des Verteilungssystems berücksichtigt werden. Beispiele für Verteilungssysteme, bei denen der Einfluss der molekularen sterischen Effekte bei der Modellierung typischerweise vernachlässigt wird, sind die Verteilungen zwischen Huminstoffen und Wasser, und zwischen mineralischen Oberflächen in Böden und Wasser.

Das Ziel dieser Arbeit war es i) Verteilungsexperimente durchzuführen, die es ermöglichen Verteilungskoeffizienten, welche durch molekulare sterische Effekte beeinflusst sind, zu quantifizieren und ii) die untersuchten Verteilungsprozesse erfolgreich zu modellieren. Experimentell bestimmt wurde hierbei das Verteilungssystem α-Cyclodextrin (αCD)-Wasser, da die 3D-Struktur von αCD, die eine entscheidende Rolle für die molekularen sterischen Effekte innehat, sehr gut bekannt ist. CDe werden in verschiedensten Bereichen verwendet, z.B. als Hilfsstoffe in der Kosmetik, in Lebensmitteln, und bei Pharmaka, oder auch als Extraktionsmittel bei der Sanierung von belasteten Böden. Zwei experimentelle Messmethoden, die beide auf dem Prinzip der Massenbilanz basieren, wurden für die Bestimmung von αCD-Bindungskonstanten etabliert: eine „head space“ und eine „passive sampling“ Methode. Die insgesamt vermessenen 70 neutralen organischen Chemikalien zeigen Bindungskonstanten in einem Bereich von 1,08 bis 4,97 log-Einheiten. Es wurden Chemikalien mit unterschiedlichen funktionellen Gruppen und mehrere homologe Reihen

ausgewählt. Diese Auswahl ermöglichte eine gute Vergleichbarkeit der verschiedenen Bindungskonstanten. Die Resultate zeigten, dass der Bindungsprozess zu αCD deutlich von molekularen sterischen Effekten beeinflusst ist, z.B. haben Konstitutionsisomere einen Unterschied in der Bindungskonstante von bis zu 1,2 log-Einheiten, der durch die verschiedenen Positionen der funktionalen Gruppe in den Isomeren verursacht wird. Außerdem sind die räumlichen Begrenzungen der αCD-Kavität ausschlaggebend für die Bindungsstärke von unterschiedlich großen, hydrophoben, aromatischen Chemikalien.

Das αCD-Wasser Verteilungssystem wurde anschließend mit drei Modellierungsansätzen beschrieben, die hinsichtlich ihrer Vorhersagequalität, mit besonderem Fokus auf die molekularen sterischen Effekte, evaluiert wurden. Die drei Modellierungsansätzen waren a) eine Polyparameter lineare Freie Energie Beziehung, b) eine „comparative molecular field analysis“, und c) eine 3D quantitative Struktur Aktivität Beziehung (QSAR). Das COSMO (conductor like screening model) basierte 3D-QSAR Modell lieferte die beste Vorhersage (R_{test}^2=0,70, $RMSE_{test}$=0,45, n=15) und schloss als einzige Methode die molekularen sterischen Effekte hinreichend mit ein. Außerdem war es in der Lage weitere 88 Literaturdaten erfolgreich vorherzusagen (R_{test}^2=0,64, $RMSE_{test}$=0,59). Die COSMO 3D-QSAR Methode kann also zur Vorhersage von unbekannten αCD-Bindungskonstanten verwendet werden und sollte sich auch auf analoge Verteilungsprobleme anwenden lassen.

Die weitere Anwendbarkeit der 3D-QSAR Methode wurde mit einem prominenten Verteilungsprozess aus dem Bereich der Toxikokinetik/Pharmakokinetik getestet, der Verteilung zwischen bovinem Serumalbumin (BSA) und Wasser. Dieser Prozess ist für die Verteilung von Chemikalien im Körper von allen Wirbeltieren relevant, da alle Wirbeltiere das stark konservierte Protein Serumalbumin exprimieren. Die Verteilung zwischen BSA und Wasser ist insbesondere für organische Anionen stark durch molekulare sterische Effekte beeinflusst. Die COSMO 3D-QSAR konnte erfolgreich experimentelle Verteilungskoeffizienten zwischen BSA und Wasser ($K_{BSA/Wasser}$) vorhersagen (R_{test}^2=0,52,

$RMSE_{test}$=0,63, n=32) und umfasste auch die molekularen sterischen Effekte, die für Unterschiede von bis zu 2 log-Einheiten bei den Verteilungskoeffizienten von geladenen Isomeren verantwortlich sind. Die Applikationsdomäne dieses empirischen Models hängt weitgehend von der zugrunde liegenden Kalibrierung ab (42 Anionen und 88 neutrale Chemikalien). Daher wäre eine Erweiterung der Kalibrierung hilfreich um eine breitere Anwendbarkeit zu ermöglichen, z.B. durch die Inklusion von organischen Kationen und Zwitterionen. Das COSMO 3D-QSAR Modell kann folglich, im Bereich der Applikationsdomäne, für eine Abschätzung der Verteilung von anionischen und neutralen Chemikalien in Wirbeltieren genutzt werden und damit zu einer verbesserten Einschätzung der Toxikokinetik von negativ geladenen organischen Chemikalien beitragen.

Preface

The present work was performed between March 2013 to June 2016 at the Helmholtz Centre for Environmental Research, Leipzig at the Department of Analytical Environmental Chemistry. The thesis was written in a cumulative form and is based on the following articles:

Linden, Lukas, Kai-Uwe Goss, and Satoshi Endo: “Exploring 3D structural influences of aliphatic and aromatic chemicals on α-cyclodextrin binding.” Journal of colloid and interface science 468 (2016): 42-50.
(SI available at: http://www.sciencedirect.com/science/article/pii/S0021979716300339)

Linden, Lukas, Kai-Uwe Goss, and Satoshi Endo. "3D-QSAR predictions for α-Cyclodextrin binding constants using quantum mechanically based descriptors." Chemosphere 169 (2017): 693-699.
(SI available at: http://www.sciencedirect.com/science/article/pii/S0045653516316587)

Linden, Lukas, Kai-Uwe Goss, and Satoshi Endo. "3D-QSAR predictions for bovine serum albumin–water partition coefficients of organic anions using quantum mechanically based descriptors." Environmental Science: Processes & Impacts (2017).
(SI available at:
http://pubs.rsc.org/en/content/articlelanding/2017/em/c6em00555a#!divAbstract)

Note that text passages and figures in the summary are partly taken from the original publication without further indication. The original publications were included at the end.

Contents

1 Summary: Influence of molecular steric factors on the sorption and partitioning of organic chemicals

1.1 Introduction

Information about the partitioning and binding behavior of organic chemicals is necessary for a broad range of fields. In environmental sciences, the distribution of chemicals and their environmental fate is largely determined by the partitioning of the chemicals between several phases like water, air, and soil; in toxicology, the distribution of chemicals and hence their effect concentration (freely dissolved concentration) is determined through their partitioning in lipids and membranes, where in simple models octanol is often used as a surrogate phase; and in medical science, the binding to macromolecules plays a crucial role for plasma protein binding and drug formulation. These examples can be divided in to two cases, the first case is partitioning between two phases (e.g., air-water, octanol-water, lipid-water) and the second case is the partitioning between a macromolecule and water. For the first case the 3D-structure of the solute and the solvent is of minor concern for the interaction energy, because steric hindrance is negligible and the solute and the solvent can interact in all possible ways. This results in prediction models, so called poly-parameter linear free energy relationship (pp-LFER), that use descriptors that characterize the interaction properties of the whole molecule without considering the molecular geometry[1-3] and they can theoretically be used to predict all cases of partitioning between two phases.

For the second case, the 3D-structure of the solute and the solvent and steric effects influence the partitioning or binding but the effects are not always clearly distinguishable from the partitioning between phases if the binding to the macromolecule was not investigated systematically. A deeper understanding of the influence of molecular steric factors on the sorption and partitioning of organic chemicals can be achieved best by starting with a good

test system like cyclodextrins (CDs) and use it for a systematic investigation. Cyclodextrins are ideal candidates for such a test system because their 3D-structure is well investigated and they are known to form inclusion complexes (host-guest complexes) with many chemicals. CDs are conic ring oligosaccharides and in the common cyclodextrin family (i.e., α-, β-, γ-) αCD may be the most suitable starting material for studying 3D-effects on binding, as it has the smallest cavity and thus the highest restriction for host-guest complexation. αCD is built of six 1-4-linked glucopyranose units that form a conic ring with a diameter of 5 Å. In water, all hydroxyl groups are positioned on the outside of the αCD ring, resulting in a hydrophobic cavity inside[4], which enables αCD to form host-guest complexes. Formation of CD complexes[5] can improve the solubility of chemicals[6], construct supramolecular polymers[7], or mask taste and odor compounds[8], underpinning the vast application areas of cyclodextrins.

Apart from the qualitative understanding of the effects that influence the binding to αCD, a successful development of a model is necessary, first, for the prediction of unknown binding constants but more importantly as a general possible solution to other binding problems that are influenced by molecular steric effects. For the evaluation of the model, pp-LFER can be used as a good reference model that helps to distinguish factors that do not appear in the partitioning between phases. A modeling approach that seems promising for the description of molecular steric effects is 3D quantitative structure activity relationship (3D-QSAR) which establishes a correlation between a macroscopic property (e.g., receptor affinity, binding constant) and 3D-structural features of the solute molecules. A widely used 3D-QSAR tool is comparative molecular field analysis (CoMFA)[9]. CoMFA uses 3D-discretized molecular field properties, called molecular interaction fields (MIFs), as descriptors for a statistical method (e.g., partial least square, PLS). Recently, Klamt et al. proposed the COSMOsar3D method[10], which uses 3D-gridded COSMO surface polarization charge densities as a new set of MIFs. This extension of CoMFA emerges from the quantum mechanically-based COSMO-RS

(conductor-like screening model for real solvent) method[2,11], which predicts the properties of a chemical by using the surface polarization charge densities (called sigma surface) of the molecule calculated quantum mechanically in a virtual conductor. For each molecule, the calculated sigma surface can be condensed into a sigma profile, a histogram of all the 'partial' charges (or charge-patches) of the molecule. The sigma surface and the sigma profile of a chemical appear to accurately describe the abilities of the molecule to undergo intermolecular interactions including electrostatic, hydrogen-bond, and van der Waals interactions[12] and are typically used to predict the partitioning between two phases. To extend this concept to 3D-QSARs, COSMOsar3D computes the sigma profiles at grid points within the 3D space to give the local sigma profiles (LSPs)[13]. The LSP is thus a 4-dimensional histogram that contains information about the sigma surface of a specific part of the molecule. Considering the theoretical basis and the proven accuracy of COSMO-RS for partitioning between liquids, it is anticipated that the LSPs are ideal MIFs for 3D-QSAR modeling of the binding free energy that is strongly influenced by the molecular geometry of solutes.

Another more advanced and for the field of ecotoxicology more relevant example of the second case (the partitioning between a macromolecule and water) is the binding to serum albumin. Serum albumin is of major importance for the toxicokinetic behavior of organic chemicals because it is the most abundant blood protein of mammals and often a predominating sorption phase in blood.[14] Additionally, fetal bovine serum is the most commonly used serum supplement for cell culture assays, where bovine serum albumin (BSA) has a strong impact on the freely dissolved concentration of the test chemical in the assays.[15] Apart from neutral organic chemicals that bind to BSA[16], Henneberger et al. published BSA/water partition coefficients ($K_{BSA/water}$) for a broad set of ionogenic chemicals measured in a consistent condition[17]. These reported ionic partition data show specific molecular steric effects, which cannot easily be described by common methods for the

prediction of partition coefficients such as pp-LFERs[18]. This highly relevant example would be an ideal candidate for another application of the tested 3D-QSAR modeling approach.

1.2 Objective of this study

The aim of this study was to identify, understand, and describe molecular steric effects that influence the binding to αCD and then develop an appropriate model that is capable of covering these effects and ideally is applicable to similar problems. For this goal, a consistent experimental data set for αCD binding of neutral organic chemicals was measured. Two approaches for the determination of αCD binding constants and binding mode (i.e., 1:1 or 2:1 binding) were tested: a) headspace and b) passive sampling. The binding constants of several isomers and homologous series were measured, which should enable a useful comparison of the chemicals and the respective binding constants and thus a good identification of molecular steric effects. Several modeling approaches like correlations with log K_{OW}, pp-LFER, and different 3D-QSARs were tested for the most thorough description of the binding to αCD. A particular emphasis of the model evaluation was the inclusion of the detected molecular steric effects.

Finally, the most successful modeling approach was applied to the partitioning of neutral and anionic organic chemicals to BSA. Again, the inclusion of molecular steric effects, which were responsible for log $K_{BSA/water}$ differences between structural isomers of up to two log units, were the main focus of the model evaluation.

1.3 Experimental identification of molecular steric effects that influence the binding to αCD

The 1:1 binding constants (K_{a1} [M^{-1}]) of organic chemicals were determined with the help of the thermodynamic cycle. K_{a1} can be expressed as,

$$K_{a1} = \frac{[S\alpha CD\,]}{[S][\alpha CD]} \quad (1)$$

where *S* is the substrate (guest) and *SαCD* is the 1:1 complex. Two methods were applied to the αCD test system. In both methods, the unbound, freely dissolved concentration of the chemical was determined via the measurement of a third phase, either air (headspace approach, see Fig. 1) or a polyacrylate (PA) or poly(dimethylsiloxane) (PDMS) fiber (passive sampling approach). All binding experiments were performed at 30 °C, which was the lowest possible temperature that the sample tray of the GC autosampler was able to control.

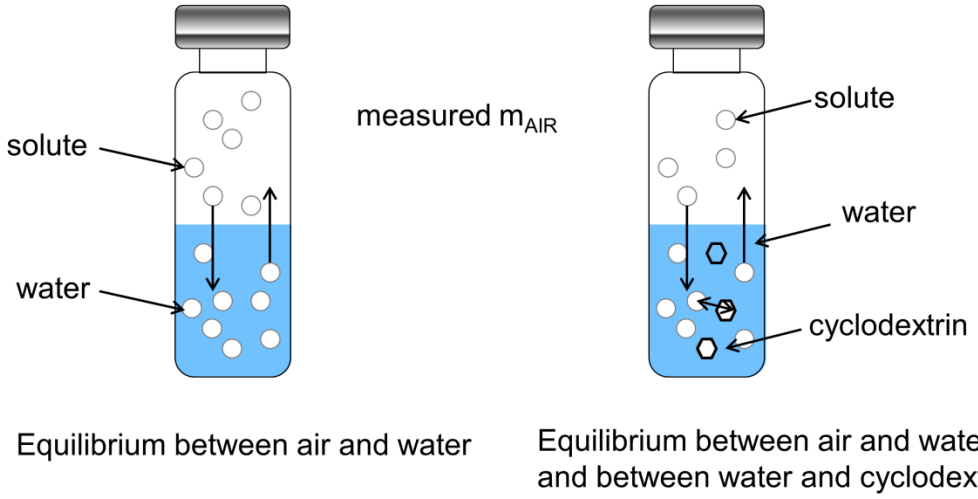


Figure 1 Experimental setting for the headspace approach. The reference phase is air, in case of the passive sampling approach the reference phase was a fiber (PDMS or PA) and no air phase was present.

1.3.1 Headspace approach

Air was the common third phase (reference phase) for this approach[19]. Two groups of weighed 20 mL vials were prepared with four vials per group. One group was filled with 5 mL water and the other was filled with 5 mL αCD solution (2 - 15 g/L). The vials were spiked

with 10 or 25 µL of methanolic stock solution of the selected chemicals and were immediately closed with a PTFE- or aluminum-lined silicone septum to prevent loss of the chemicals. From the experience of preliminary experiments, the equilibrium time was set to a minimum of four hours: first three hours on a horizontal shaker at 30 °C with 300 rpm and then at least one hour on the GC-sample tray at 30 °C with low shaking speed. Then the headspace was probed with a 100 µL sampling loop or a 250 µL syringe and injected into the GC and measured with GC-FID/ECD or GC-MS.

1.3.2 Passive sampling approach

The passive sampling approach was used for chemicals which are not volatile enough for the headspace approach. PA or PDMS fiber is the common reference phase for this approach[20,21]. The experimental setting was similar to that with the headspace approach except for the following changes. The volume of the solutions and the vials was 10 mL, each vial received 5 or 10 cm of PA- or PDMS-coated fiber and the equilibrium time was 72 hours at 30 °C. Previous studies[22,23] confirmed that this equilibrium time is sufficient for a wide range of chemicals. After equilibrium was reached, the fibers were removed from the vials and carefully wiped with a clean tissue. Then the fibers were extracted overnight on a roller mixer using 200 µL of cyclohexane (for PDMS) or ethyl acetate (for PA). The concentrations of the extracts were quantified with a GC-MS system using an external calibration.

1.3.3 Detected molecular steric effects

The K_{a1} values of 70 chemicals were determined in batch experiments. The chemical set comprises: 19 alcohols, 19 ketones, 9 polycyclic aromatic hydrocarbons (PAHs), 6 chlorobenzenes, 5 alkylbenzenes, 4 ethers, 4 nitroalkanes, and 4 phosphates/phosphonates. These chemicals have various functional groups but relatively simple molecular structures, which facilitates interpretation of the results. Moreover, the data set includes multiple series of chemicals with increasing number of structural units (i.e., -CH_2-, Cl-, aromatic ring),

enabling the assessment of incremental effects on the binding behavior. The measured log K_{a1} values span over a wide range, from 1.08 (pentachlorobenzene) to 4.97 (1-dodecanol). For 10 chemicals, K_{a1} was too small to be measured with the applied method.

The logarithmic octanol-water partition coefficient (log K_{OW}) is often related to log $K_{CD/water}$[24-27] and was even proposed as a descriptor for predictions[25]. In Fig. 2 the log K_{a1} values measured in this study are compared to log K_{OW}. The correlation between log K_{OW} and log K_{a1} is weak (R^2 of 0.19). Correlation is particularly weak in the high K_{OW} range (i.e., log K_{OW} > 3). For example, 1-dodecanol and pentachlorobenzene have similar log K_{OW} values (5.13 and 5.17, respectively) but differ more than 3 log units in their K_{a1} values (4.96 and 1.08, respectively). Conversely, nitroethane and phenanthrene have > 4 log units difference in log K_{OW} values (0.18 and 4.46, respectively) but the respective log K_{a1} values are both < 1.3. This shows that log K_{OW} is neither useful for the understanding of the specific binding processes to αCD, nor for estimating log $K_{CD/water}$ if different chemical classes are considered.

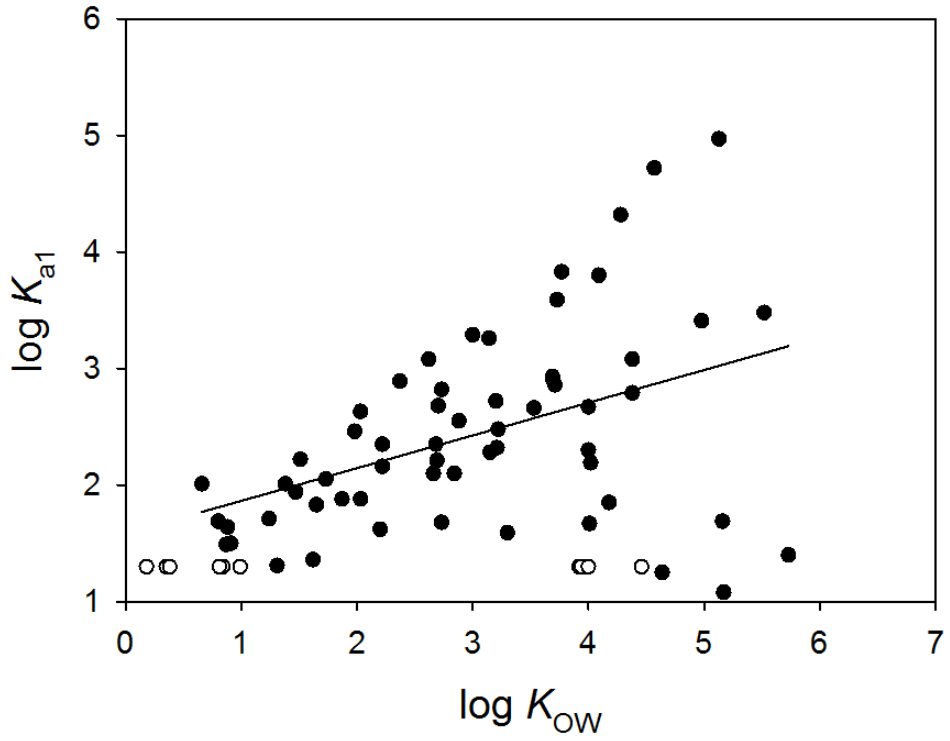


Figure 2 Experimental 1:1 αCD binding coefficients versus octanol/water partition coefficients.

An interesting finding is that the position of the functional group has a high influence on the log K_{a1} values. In general, log K_{a1} increases linearly with the number of carbon atoms within each homologous group: (1) linear aliphatic compounds with the polar functional group at the end of the molecule, i.e., R-OH, R-C(=O)CH_3, and R-NO_2, where R is a linear alkyl chain of differing lengths, (2) aliphatic compounds with the polar functional group in the middle of the molecule, i.e., R-C(OH)-R', R-C(=O)-R', and R-O-R, where R' = R or R-CH_2- (i.e., one unit longer), and (3) trialkyl phosphates (i.e., PO_4-RRR). But chemicals with the functional group at the end of the molecule have generally higher K_{a1} than chemicals with the same functional group in the middle, when compared at the same number of carbon atoms and there is a substantial difference in the slopes between end-substituted and middle-substituted classes (0.40 and 0.26 log units/C on average, respectively). Such a differential increase per C does not occur with solvent-water partition coefficients such as K_{OW} and thus has to be caused by steric effects. We hypothesize that this occurs mainly because the polar functional group of the bound guest molecule stays outside the hydrophobic cavity and interacts with the surrounding water or with one of the hydroxyl groups of the αCD rims. Thus, the polar functional group restricts the location and the orientation of the guest relative to αCD and can thereby hinder the optimal interactions of the alkyl chain(s) with the αCD cavity. It is plausible that the polar functional group stays outside the cavity, because the polar functional group of the free, unbound chemical can undergo strong hydrogen bonding interactions with water molecules, whereas hydrogen bonds cannot be formed inside the hydrophobic cavity of αCD. Thus, the polar functional group could enter the cavity only if that leads to a free energy gain that is larger than the free energy loss due to the breakup of hydrogen bonds with water. Assuming that the polar functional group has to be outside the cavity, end-substituted chemicals may still fully insert their alkyl chain into the cavity, whereas middle-substituted chemicals may not insert both chains well in the cavity.

Furthermore, the interaction with αCD is highest if the alkyl chain of a chemical is linear and non-branched as can be seen by the comparison of several isomers (Fig. 3, 1-octanol, 2-ethyl-1-hexanol, 4-octanol, 3-ethyl-3-hexanol, and 4-ethyl-3-hexanol). While K_{a1} of a chemical with an ethyl-branched alkyl chain is lower than that of its non-branched isomer, the energetic contribution of the additional ethyl group is always positive. Hence, log K_{a1} is higher for 2-ethyl-1-hexanol (2.81) than for 1-hexanol (2.62), and log K_{a1} of 3-ethyl-3-hexanol and 4-ethyl-3-hexanol is higher than that of 3-hexanol. It is thus apparent that the branched ethyl group can also interact with CD and has a significant contribution to K_{a1}.

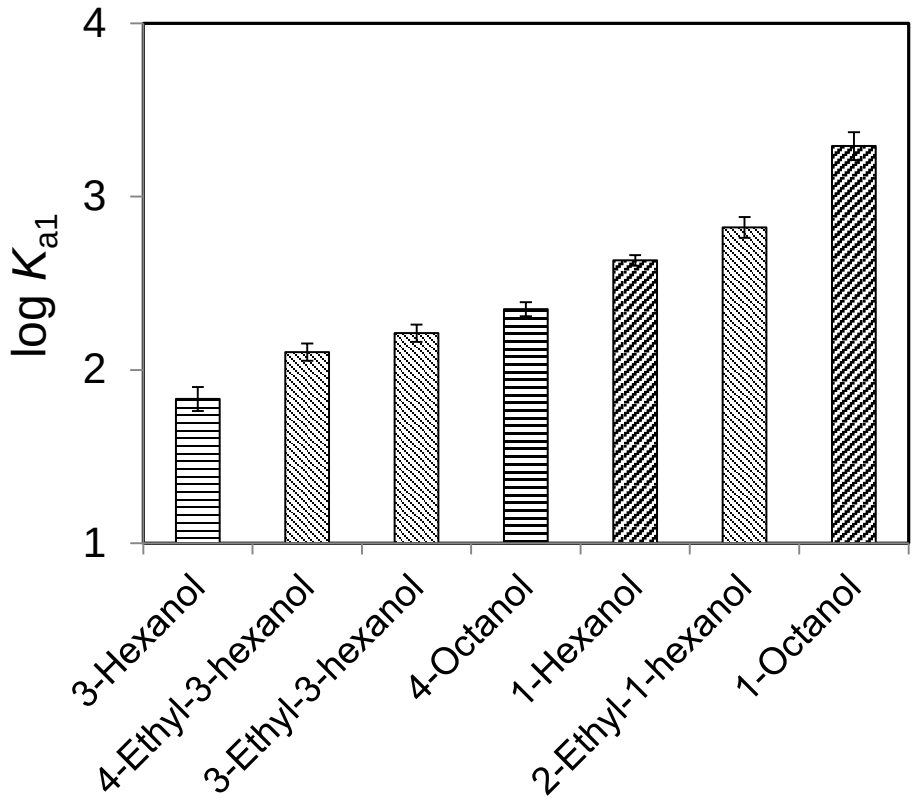


Figure 3 Comparison of K_{a1} for 2 C_6-alcohols and 5 C_8-alcohols.

The aromatic chemicals studied in this work are nine PAHs, six chlorobenzenes, and five alkylbenzenes and show a different behavior than the aliphatic chemicals discussed above. The alkylbenzenes contain one linear alkyl chain of increasing length, but log K_{a1} is not a simple linear function of the number of C atoms, in contrast to the polar aliphatic compounds shown above. The benzene ring does not form a strong H-bond with water and thus can

favorably enter the hydrophobic cavity of αCD. As the benzene ring occupies a fraction of the cavity, alkylbenzenes possessing an alkyl chain with three or more carbon atoms appear to experience a steric effect that lowers the log K_{a1} increase per carbon atom. Chlorobenzenes represent an even more pronounced example of the influence of steric restriction. The log K_{a1} values are above two for chemicals possessing one to three chlorine atoms after which log K_{a1} starts to decrease with an increasing number of chlorine atoms. Mono and 1,3-dichlorobenzenes appear to fit into the cavity, whereas 1,2,4-trichlorobenzene already experiences a negative steric effect. The log K_{a1} for 1,2,4,5-tetrachlorobenzene is even lower than that of monochlorobenzene, suggesting that the three additional chlorine-substitutions hinder the interactions of the benzene ring and the original chlorine atom with αCD. Due to the summarized results, we decided to model the binding to αCD with a modeling approach that includes the 3D information of the chemicals, namely 3D-QSARs.

1.4 3D-QSAR modeling of the binding to αCD

We evaluated the predictive performance of two different modeling approaches focusing on the detected molecular steric effects on the binding to αCD: I) pp-LFER[28,29] and II) 3D-QSAR. The performance of the 3D-QSAR modeling approach was thoroughly investigated in a way that we performed a standard CoMFA[9] and then extended it with two methods, COSMOsim3D[13] and COSMOsar3D[10].

1.4.1 Methods

1.4.1.1 Selection procedures for training and test sets

For generation and evaluation of each model (i.e., pp-LFER and 3D-QSARs), the used data set was split into training and test sets. The training set was used for model calibration and selection, while the performance of the resulting model was validated with regard to the prediction of the test set. Prediction of data that were not part of the training set is essential as a control and should be considered the more important quality feature for 3D-QSARs[30].

For the general model evaluation, the training and test sets were generated with the log K_{a1} hierarchic bin system[31] (procedure 1)**.** This classifies 25% chemicals of the data set to the test set. The rest of the chemicals formed the training set. The procedure was repeated five times, resulting in five random training sets and the corresponding test sets.

In order to evaluate varying steric effects within homologous series of chemicals and isomers, the following modified procedure was used to generate constructed test sets (procedure 2)**.** As in the first procedure, the chemicals were sorted by log K_{a1} and four chemicals in a row were grouped into one bin. Then, the numbers 1 to 4 were given randomly to the four chemicals of a bin. In the first run of chemical selection, the chemicals with the number 1 embodied the test set, while the rest of the chemicals were used as the training set. In the second run, the chemicals with the number 2 were the test set, and so forth. In comparison to procedure 1, the

randomness of the selection is reduced, whereas each chemical is part of a test set once and the other three times it belonged to the training set.

1.4.1.2 pp-LFER

The pp-LFER is among the most accurate and robust models to describe solute partitioning between liquids or liquid and gas phases, where molecular interactions are not sterically restricted. In a practical sense, a 3D-QSAR model may be considered meaningful only if it gives better predictions than the pp-LFER model, which is simple and quick as long as the solute descriptors are known. The pp-LFER used here appears,

$$\log K_{a1} = c + sS + aA + bB + vV + lL \quad (2)$$

where S is the polarizability/dipolarity parameter, A the solute H-bond acidity, B the solute H-bond basicity, V the McGowan characteristic volume ($cm^3\ mol^{-1}/100$) and L the logarithm of the hexadecane-air partitioning coefficient. In this work, the pp-LFER solute descriptors (capital letters in eq. 1) were obtained from the UFZ-LSER database[32] and the system parameters (lower case letters in eq. 1) were fitted with multiple linear regression analysis using the experimental data for log K_{a1} of training chemicals.

1.4.1.3 3D-QSAR

The 3D-QSAR modeling followed the workflow shown in Fig. 4. Modeling generally takes the following steps: 3D-structure generation, alignment, MIFs generation, model calibration with PLS, and model evaluation using the test set. There are multiple options for each step, as explained below, and different combinations were tested in this work for comprehensive evaluation of the methods.

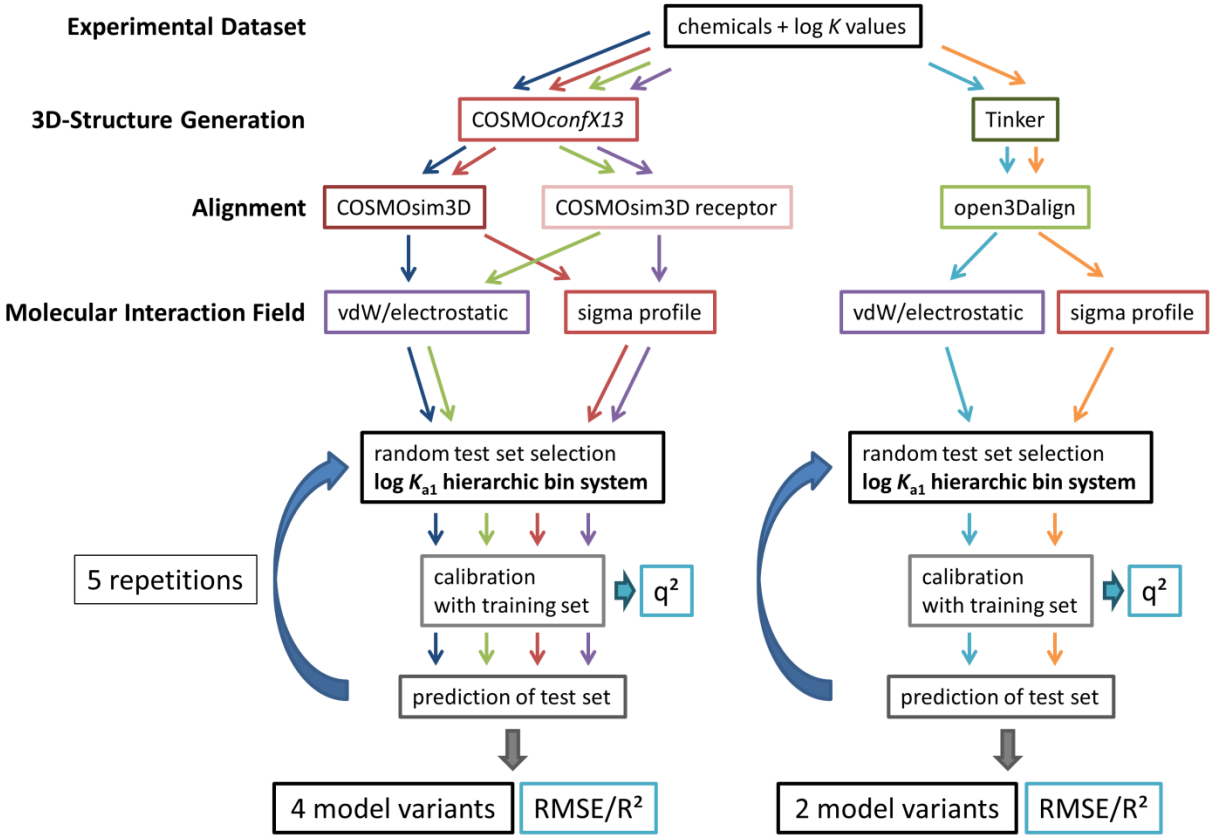


Figure 4 3D-QSAR modeling workflow. Each colored line indicates one specific model variant.

1.4.1.3.1 3D structure generation

The 3D structures of all chemicals were generated with Tinker or COSMOconfX13. Tinker[33] is a molecular modeling package implemented in Open3Dalign v. 2.3 (O3A)[34]. COSMOconfX13 is a tool box that uses Turbomole[35] for the quantum mechanics calculations of COSMO files.

1.4.1.3.2 Alignments

The 3D structures of chemicals need to be aligned in the 3D space before performing statistical analysis. Ideally, the resulting position and orientation of a chemical in the 3D space corresponds to the optimal interaction possibility between the chemical and αCD. In a target-based approach, the structure or a substructure of αCD is used as the template to which all molecules are aligned. In a ligand-based approach, the template is generated with the help of chemicals that bind strongly to αCD (i.e., with high log K_{a1} values). For all approaches, up

to ten conformers of each chemical were considered and the conformer with the highest alignment score and, if there are multiple conformers with the highest score, then that with the lowest energy was chosen for the model. In this study, the following three alignment procedures were applied.

1. The O3A alignment maximizes the overlap of atoms of the template chemicals and of the remaining chemicals. This is a ligand-based method and a standard alignment for CoMFA approaches and was performed here by using O3A v. 2.3[34]. The seven chemicals with the largest log K_{a1} values of this study, namely 1-dodecanol, 1-undecanol, 1-decanol, 1-nonanol, 2-undecanone, 2-decanone, and hexylbenzene were used as template chemicals.
2. The COSMOsim3D alignment[13] maximizes the overlap between the sigma surfaces of the chemical and the template. Hereby, the template is an averaged sigma profile of the template chemicals. The template chemicals used were the same as in the previous alignment method.
3. The COSMOsim3D receptor alignment is a target-based approach that maximizes the overlap between the inverted sigma surface of αCD (which is the sigma charge value of each surface patch multiplied with -1) and the sigma surface of the chemicals of the data set. The sigma surface of αCD needs to be inverted because the alignment algorithm maximizes the overlap of like sigma charges in a ligand-based approach. The inversion therefore places the chemicals in a position where greatest interaction energies between both αCD and the respective chemical occur, as the interaction energy is greatest when the difference between the sigma charges of two interacting surface segments is maximal. This alignment already considers the steric restrictions of the αCD cavity because the chemicals cannot be placed at the same position as the αCD. Two sources for an input structures, the αCD and an exemplary ligand, were used in our approach to test the dependence of the COSMOsim3D receptor alignment

on the input structure: a) X-ray measurement[36] and b) molecular dynamics simulation (MDsim).

1.4.1.3.3 MIFs

Two sets of MIFs were used as independent variables for the partial least squares (PLS) regression analysis.

1. The van der Waals (vdW) and the electrostatic (ele) fields are the two standard CoMFA variables. Molecular mechanics calculations using the Merck force field (MMFF94) were performed with Open3DQSAR v. 2.3[37] to derive the vdW and ele fields. A sp^3 carbon atom was used as the probe. A grid spacing of 1 Å was used with a 5 Å gap, i.e., the minimal distance to the box, around the chemicals.
2. LSPs were derived from the cosmo files by COSMOsar3D[10]. For the 3D-QSAR model used here the LSPs were split into several consecutive profiles, each covering a range of 0.006 e/Å². Thus, MIFs 1, 2, …, and 7 cover sigma values from -0.024 to -0.018 e/ Å², -0.018 to -0.012 e/ Å², …, and, 0.012 to 0.018 e/ Å², respectively. In the end, the LSPs, thus the amount of the surface area within a certain sigma charge interval and a space interval, serves as the value for the independent variable. A grid spacing of 2 Å was used in a box that leaves at least a 5 Å gap around the chemicals.

1.4.1.3.4 Statistical tool

The independent variables, i.e., the MIFs, of the training set chemicals were correlated with the log K_{a1} values using PLS regression analysis. Prior to PLS regression analysis, the number of independent variables was reduced in a way that potential meaning less variables were excluded. PLS analysis was performed to derive one to five PLS components. Leave-two-out cross validation was performed with each model and then the model with the minimum of the root mean square error (RMSE) value was selected for further evaluation against the test set.

1.4.2 Internal validation of the modeling approaches methods

The general performance of the modeling approaches were evaluated using the αCD data set described above (called Linden data set in the following), which data is of high quality and consistency, and the test set selection procedure 1. Table 1 shows the statistical results for evaluation of the modeling approaches. RMSE and R^2 calculated with the test sets are considered more important evaluation criteria than q^2.

Table 1. Comparison of the statistical results of the different modeling approaches for the prediction of log K_{a1} of the Linden data set.

Modeling approach	Method	Alignment	Field	$q^2 \pm SD$	RMSE ± SD	$R^2 \pm SD$
M1	pp-LFER				0.52 ± 0.05	0.68 ± 0.07
M2	3D-QSAR	O3A	LSP	0.63 ± 0.03	0.54 ± 0.08	0.56 ± 0.17
M3	3D-QSAR	O3A	vdW ele	0.58 ± 0.08	0.53 ± 0.11	0.53 ± 0.11
M4	**3D-QSAR**	**COSMOsim3D**	**LSP**	**0.83 ± 0.02**	**0.45 ± 0.06**	**0.70 ± 0.08**
M5	3D-QSAR	COSMOsim3D	vdW ele	0.70 ± 0.01	0.56 ± 0.06	0.53 ± 0.12
M6a	3D-QSAR	COSMOsim3D receptor X-ray	LSP	0.66 ± 0.06	0.51 ± 0.06	0.61 ± 0.09
M6b	3D-QSAR	COSMOsim3D receptor MDsim	LSP	0.71 ± 0.04	0.49 ± 0.04	0.64 ± 0.07
M7	3D-QSAR	COSMOsim3D receptor X-ray	vdW ele	0.51 ± 0.08	0.55 ± 0.08	0.56 ± 0.13

O3A means open3DALIGN, q^2 is the coefficient of determination for the leave-two-out cross validation using the training set, RMSE is the root mean square error of the test set in log units, and R^2 is the coefficient of determination of the test set. LSP, vdW, and ele indicate the usage of local sigma profiles, van der Waals interaction field, and electrostatic interaction field as molecular interaction field, respectively, SD is standard deviation, and MDsim is molecular dynamics simulation.

1.4.2.1 pp-LFER

First, the pp-LFER equation (eq. 2) was fitted to all experimental αCD binding constants of the αCD Linden data set (i.e., no test and training set selection) to have an idea to what extent the 2D model can describe the whole data set. The fit of the pp-LFER equation usually results

in a standard deviation of 0.1 to 0.2 log units for homogeneous solvent-water partition systems, which are not influenced by steric effects, and a larger standard deviation for partitioning or binding to heterogeneous materials such as serum albumin and natural organic matter[23,38]. The RMSE for predicted binding to αCD is 0.48, being comparable to fits for other heterogeneous materials[38].

The pp-LFER fits for training sets extracted from the Linden data set resulted in system parameters similar to those for the complete data set. The predictions for the corresponding test sets (Table 1, M1) were surprisingly accurate (RMSE = 0.52 ± 0.05 and R^2 = 0.68 ± 0.07). This result was unexpected because the experimental results do suggest strong steric effects, whereas the pp-LFER model does not capture such effects[39]. A closer examination of the results revealed that systematic prediction errors do exist for binding constants, e.g., log K_{a1} values for end-substituted chemicals were systematically underestimated and those for middle-substituted chemicals were overestimated. In addition, chemicals that are not expected to fit into the αCD cavity due to the steric hindrance were over-predicted by the pp-LFER, e.g., the log K_{a1} value of 1-chloronaphthalene is predicted as 2.13, while the experiment suggests that it is < 1.3[39].

1.4.2.2 3D-QSARs

Seven 3D-QSAR model variants were constructed using different combinations of structure generation, alignment, and MIF methods and evaluated with the Linden data set, as explained in the method section (Fig. 4, Table 1). The results show the following trends: (i) RMSE and R^2 of the 3D-QSAR model variants for test set predictions were 0.45–0.56 and 0.53–0.70, respectively. While the best 3D-QSAR model (M4) performed slightly better than the pp-LFER, the statistics were similar on average. (ii) The models that used the LSPs[10] as independent variables tended to result in better predictions than those using the vdW and ele MIFs for a given alignment (i.e., O3A, COSMOsim3D, or COSMOsim3d receptor). These outcomes suggest that LSPs are more suitable descriptors to describe the binding to αCD than

the tested CoMFA variables. This interpretation is in line with the claim that LSPs are theoretically more relevant for linear regression models, like PLS, to describe the interaction energy[10].

Of the 3D-QSARs tested, the model that uses the COSMOsim3D alignment with the LSP variables (M4, Table 1) was the best model variant (i.e., with the lowest RMSE). No improvement was observed for the use of the 3D-structure of αCD as the template for the alignment (compare M6a and M6b to M4). The fact that no improvement was observed by the use of the target-dependent alignment suggests that the selected seven template chemicals were sufficient for aligning the 60 chemicals in the Linden data set. This result, however, may not be general; alignments with a binding site structure are expected to be advantageous particularly if the data availability is limited.

1.4.2.2.1 Predictions of specific molecular steric effects

To evaluate the performance of the 3D-QSAR modeling approaches for predicting particular types of chemicals, four training and test sets were generated from the Linden data set according to test set selection procedure 2 (see the method section) and all prediction procedures were redone. Model approaches M3, M4, M5, and M6b were evaluated here because they performed best in the random evaluation above and allow comparison of the classical CoMFA approach and the new COSMO-based approach. The resulting statistics (i.e., q^2, RMSE, R^2) were similar to those obtained above with test set selection procedure 1 (Table 1), except for M3, for which the test set selection procedure 2 resulted in worse predictions. Fig. 5 compares the experimental data and the predictions by the best model variant (M4, with COSMOsim3D + LSPs) for individual chemicals.

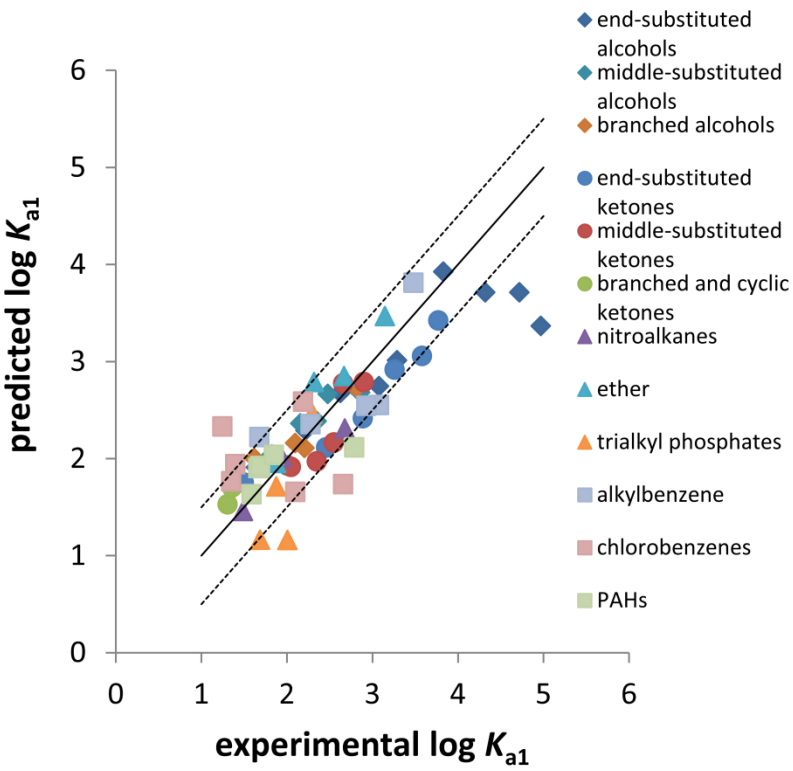


Figure 5 Prediction of log K_{a1} of 60 chemicals with COSMOsim3D alignment and local sigma profiles as variables

Many trends of the data that are related to steric effects were quantitatively described in the best 3D-QSAR model variant we found (M4). For example: experimental data show relatively large differences in log K_{a1} between isomeric chemicals with the functional group at the terminal and the middle positions such as 1-heptanol and 4-heptanol. These chemicals are predicted successfully by M4, e.g., 1-heptanol (log K_{a1} exper. 3.08, pred. 2.75) and 4-heptanol (log K_{a1} exper. 2.16, pred. 2.36). Also, elongation of the alkyl chain in only one direction resulted in a higher increase of log K_{a1} than elongation in two or more directions (Fig. 6), correctly reproducing the findings of the experimental data. The 3D-QSAR model variants M3, M5, and M6b were not able to describe the differences between these alcohols so well as M4 (Fig. 6). The comparison between M4 and M5 shows that the use of LSPs instead of vdW and ele not only minimizes the overall prediction errors but helps distinguish structural isomers of alcohols. The standard CoMFA model (M3) underestimates most of these alcohols and is not able to capture the molecular steric effects. M6b uses LSPs as variables, but it

appears that the target-based alignment cannot as accurately reproduce the trend of alcohol data as the ligand-based alignment in this case.

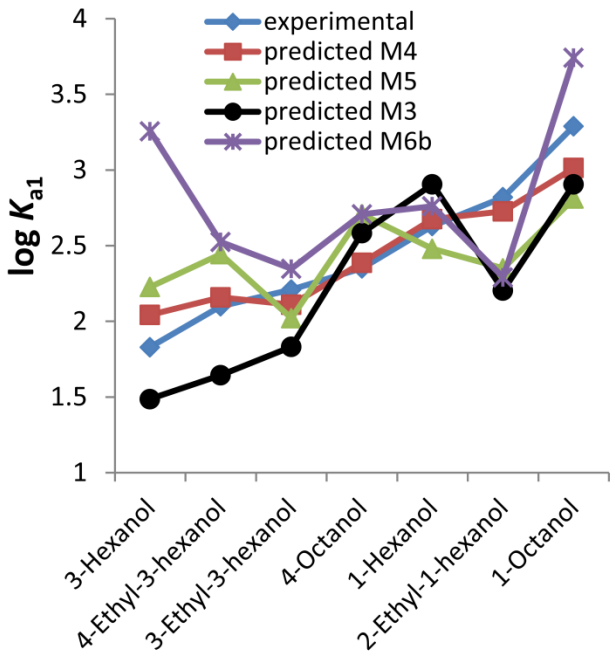


Figure 6 Experimental and predicted log K_{a1} for αCD binding of two C6-alcohols and five C8-alcohols.

1.4.3 External validation of the modeling approaches

For an external evaluation of each modeling approach, models were generated using all self-measured data (Linden data set) as the training set and evaluated with a literature data set (Suzuki data set) as an external test set. The Suzuki data set[40] includes 87 neutral aliphatic and aromatic chemicals (range of log K_{a1}: -0.09–3.81, mean: 1.95, SD: 0.81). The prediction of the Suzuki data by the pp-LFER calibrated with the Linden data (Table 2, M1) was substantially worse (RMSE = 1.08, R^2 = 0.16), as compared to the test set predictions of the Linden data set (Table 1, M1). This RMSE is even greater than the SD of the Suzuki data. It is notable that the pp-LFER, which does not include steric terms, does show promising statistics when evaluated with the Linden set alone (Table 1, M1), whereas the model calibrated with the Linden set does not extrapolate well to the external Suzuki set.

The 3D-QSAR models handled the external prediction better than the pp-LFER model, but RMSE values for the predictions of the Suzuki data set (Table 2, M2-M7) were 0.13-0.19 log units higher than the test set predictions for the Linden data set. The model variant that uses the COSMOsim3D alignment and LSPs (Table 2, M4) achieved an RMSE of 0.59 and an R^2 of 0.61, while all other models had RMSE > 0.68 and R^2 < 0.5. For a given alignment, LSPs resulted in better or equivalent statistics as compared to vdW and ele. These results are in line with the findings we obtained from the model evaluation with the Linden data set only.

Table 2 Comparison of the statistical results for the prediction of the Suzuki data set. All Linden data were used as the training set.

Modeling approach	Method	Alignment	Field	q^2	RMSE	R^2
M1	pp-LFER				1.09	0.19
M2	3D-QSAR	O3A	LSP	0.8	0.69	0.44
M3	3D-QSAR	O3A	vdW ele	0.69	0.72	0.39
M4	**3D-QSAR**	**COSMOsim3D**	**LSP**	**0.83**	**0.59**	**0.61**
M5	3D-QSAR	COSMOsim3D	vdW ele	0.71	0.72	0.32
M6b	3D-QSAR	COSMOsim3D receptor MDsim	LSP	0.58	0.68	0.48
M7	3D-QSAR	COSMOsim3D receptor MDsim	vdW ele	0.73	0.68	0.49

O3A is open3Dalign, MDsim is molecular dynamics simulation, q^2 is the coefficient of determination for the leave-two-out cross validation using the training set, RMSE is the root mean square error of the test set in log units, and R^2 is the coefficient of determination of the test set. LSP, vdW, and ele indicate the use of local sigma profiles, van der Waals interaction field, and electrostatic interaction field, respectively, as molecular interaction fields.

1.5 3D-QSAR modeling of the binding to BSA

The modeling approach that performed best with the prediction of the αCD binding (COSMOsim3D + COSMOsar3D) was applied to another partitioning example that is influenced by steric effect: the partitioning between BSA and water[16,17]. Especially, anionic chemicals show distinct steric effects that are responsible for up to two log units differences in log $K_{BSA/water}$ ([L_{water}/kg_{BSA}]) between structural isomers. The partition coefficient is defined as

$$K_{BSA/water} = \frac{c_{BSA}}{c_{free}} \quad (3)$$

where c_{BSA} is the concentration of the chemical bound to BSA [mol/kg_{BSA}] and c_{free} is the freely dissolved concentration of the chemical in water [mol/L_{water}]. Depending on the field, partitioning or binding to BSA is also reported as a binding constant K_a [M^{-1}], again defined for the 1:1 binding as

$$K_{a1} = \frac{[S - BSA\,]}{[S][BSA]} \quad (4)$$

where *S* is the substrate and *S-BSA* is the 1:1 complex. Thus the binding constant can be derived from the partition coefficient using the following equation:

$$K_{a1} = K_{BSA/water} MW_{BSA} \quad (5)$$

where MW_{BSA} is the molecular weight of albumin (~67 kg/mol).

It is important to notice that prior to building a model, we had to generate a common binding hypothesis, i.e., a common 3D alignment, between the solutes and BSA. The exact position and orientation of the solute is of minor influence in case of solvent partitioning because steric effects do not hinder the possible interactions between small solvent molecules and a solute. In contrast, the sorption to proteins, like BSA, is influenced by the spatial structure of the sorption sites and any possibly resulting steric hindrance. This means that a modeling approach needs to represent the spatial structure and the chemical environment of the sorption sites. Because we wanted to construct a model that is as generally applicable as possible, we

chose an approach that assumes that the different reported sorption sites of BSA are alike and their spatial structure and interaction possibilities can be expressed through one characteristic binding site[41]. To identify the optimal alignment in the characteristic binding site, we used those five chemicals from the experimental data sets with the strongest binding to BSA (so called template chemicals) and a rigid structure, assuming that they would represent a nearly optimal positioning at the binding site. The software COSMOsim3D[13] generated an averaged sigma surface (including the 3D information) from the sigma surfaces of the template chemicals, benzo[*g,h,i*]perylene, chrysene, pyrene, naphthalene-2-sulfonate, and 2-naphthaleneacetate, which represents the characteristic binding site and which was used for the optimal alignment of the chemicals of the data set. These five chemicals are a reasonable choice for the template because a high partition coefficient corresponds to a good interaction with BSA and rigid structure helps to delineate the binding site better than flexible structure. Obviously, choice of template chemicals is always limited through the data availability of binding chemicals, which may partially limit the domain of applicability of the resulting model. The 3D similarity between the averaged sigma surface of the five template chemicals and the sigma surface of each chemical was maximized through the translation and rotation of the 3D-COSMOfiles of each chemical in the 3D space; this corresponds to an optimization of the best possible interaction with BSA. This optimization procedure was carried out using a grid with a 0.5 Å spacing. Analog to the modeling procedure of the binding to αCD, the conformer with the highest alignment score was selected for further modeling and if there were multiple conformers with the same alignment score, then the conformer with the lowest internal energy was used.

1.5.1 Results

Five 3D-QSAR models were calibrated from different subsets of the available experimental data to describe the partitioning to BSA and to predict the respective test sets. Again, we

decided to use several combinations of training and test sets to account for the dependency of the statistical results of 3D-QSAR modeling on the combination of training and test sets. Fig. 7 gives examples of the test set predictions that resulted from different combinations of training and test sets. The prediction of the five random test sets resulted in an RMSE of 0.63 ± 0.10 and an R^2 of 0.52 ± 0.15 (the values represent the mean ± standard deviation). The neutral chemicals (n=21) of the test set were predicted with an RMSE of 0.59 ± 0.04 while anionic chemicals (n=11) were predicted with an RMSE of 0.68 ± 0.23. In general, the neutral chemicals are better predicted compared to the anionic chemicals, which might be caused by the disproportion of the training sets (62 neutral chemicals and 32 anionic chemical). However, the neutral chemicals in the calibration set appear to improve the description of the partitioning of anionic chemicals to BSA, as modeling using solely the anionic chemicals was less successful than that with the combined data set. Reasons for this outcome could be the small number of anionic chemicals that is not enough to calibrate the model, and the higher diversity of the neutral data set that helps also to predict log $K_{BSA/water}$ of less diverse, and even anionic, chemicals as long as the 3D-structures of the anionic chemicals are similar to those of the neutral chemicals. The binding mechanism behind the 3D-QSAR model can be examined with the help of the contributions of the different LSPs/MIFs to the overall model. The positive influence of anionic partial charges on the partitioning to BSA, which is expressed in the experimental data, is captured in the model. Other important interactions identified by the model are van der Waals interactions and the hydrophobic effect.

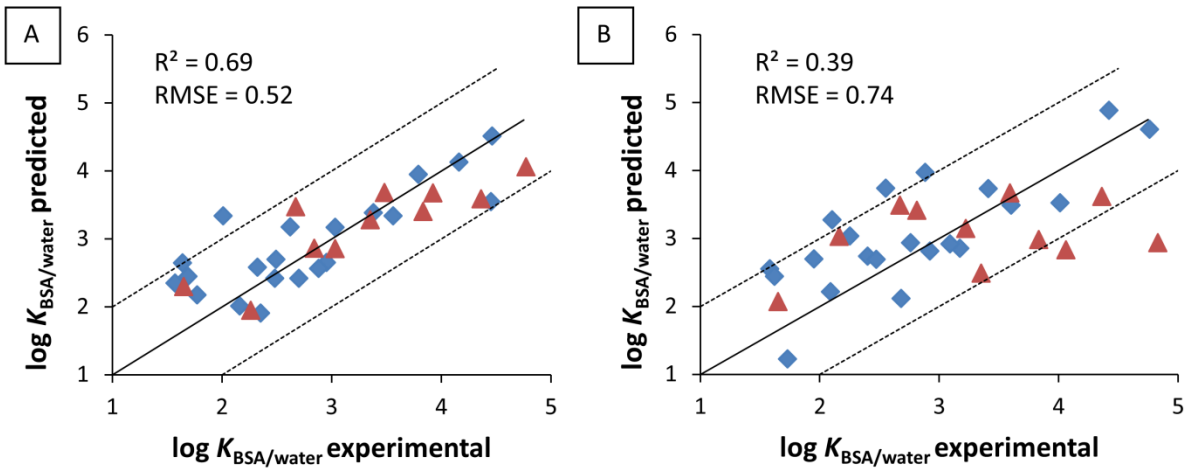


Figure 7 (A) Best and (B) worst prediction of log $K_{BSA/water}$ of 21 neutral and 11 anionic chemicals of five random test sets. The blue diamonds indicate the neutral chemicals and the red triangles indicate the anionic chemicals. The solid line indicates the 1:1 line and the dashed lines indicate a deviation of 1 log unit from the 1:1 line.

1.5.2 Prediction of molecular steric effects

The important steric effects in the anionic data were investigated separately using the comparison of the prediction of different isomers. In experimental data, several isomer pairs show similar steric effects: an ortho-substitution of benzoate decreases log $K_{BSA/water}$ substantially compared to a para- or meta-substitution (2-chlorobenzoate vs. 4-chlorobenzoate, 2,6-dichlorobenzoate vs. 3,4-dichlorobenzoate, 2-methylbenzoate vs 4-methylbenzoate) and a substitution at the alpha-position of naphthalene decreases log $K_{BSA/water}$ while a substitution at the beta-position increases log $K_{BSA/water}$, particularly if the substitution group is negatively charged (1-naphthoic acid anion vs. 2-naphthoic acid anion, 1-naphthalenacetic acid anion vs. 2-naphthalenacetic acid anion). The steric hindrance of the ortho-position results in a twist of the carboxylate group[17], which was speculated as a possible reason for the observed specificity. The relative sorption behavior of these isomer pairs with steric effects was predicted correctly by the models (Fig. 8). Even quantitative predictions (errors < 0.8) were achieved for three of the five isomer pairs. The other two had relatively large prediction errors: log $K_{BSA/water}$ of 3,4-dichlorobenzoate is underestimated (1.26 log units) and log $K_{BSA/water}$ of 4-methylbenzoate is overestimated (0.85 log units).

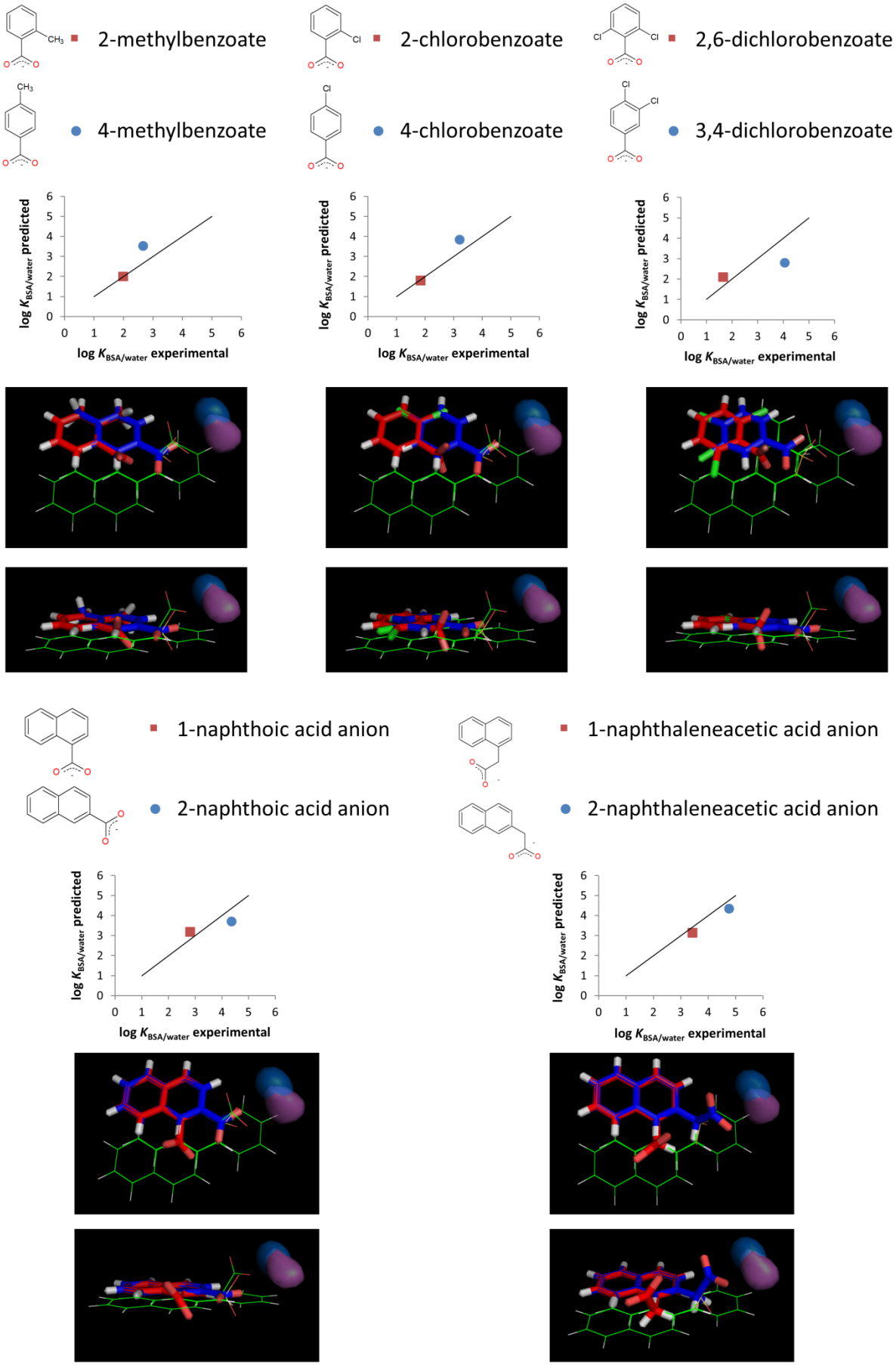

2-methylbenzoate
4-methylbenzoate
2-chlorobenzoate
4-chlorobenzoate
2,6-dichlorobenzoate
3,4-dichlorobenzoate
log $K_{BSA/water}$ predicted
log $K_{BSA/water}$ experimental
1-naphthoic acid anion
2-naphthoic acid anion
1-naphthaleneacetic acid anion
2-naphthaleneacetic acid anion

Figure 8 Experimental and the average predicted log $K_{BSA/water}$ values of the modified test sets for several isomer pairs. The black line in the graphs indicates the 1:1 line, the red squares indicate the ortho- or alpha-substituted isomer, and the blue squares indicate the para- or beta-substituted isomer. The green lines in the pictures show the alignment chemicals/templates while the blue sticks show the ortho- or alpha-substituted isomer and the red sticks show the para- or beta-substituted isomer. The teal (LSP 7) and the violet (LSP 8) area indicate the space where the models identified a positive interaction of an anionic partial charge with BSA. The alignment figures were generated using Pymol[42].

The alignment of the chemicals was an important factor for the distinction of the isomer pairs. The green lines in the pictures of Fig. 8 show the five alignment chemicals while the sticks show the respective isomers. The alignments of the five template chemicals resulted in superimposed atoms and bonds and hereby in stacked aromatic π-systems. In addition, the anionic groups of naphthalene-2-sulfonate and 2-naphthaleneacetate are located at the same position, which could represent a possible interaction with a positively charged or electron-withdrawing group of BSA.[43] Indeed, all isomers of Fig. 8 with the higher log $K_{BSA/water}$ value have their charged group located close to this position (this interaction space is indicated in Fig. 8 by the teal and violet areas as it is expressed in the model). The isomers of Fig. 8 with the lower log $K_{BSA/water}$ value (marked with red squares) have their anionic group at different positions, which seems to be inevitable for maximizing the overlapping of the rest of the structure to the template but seems to lead to omission of the interaction between the charged group of the chemical and BSA in the model. This difference in the positions of the anionic groups, which is caused by the twist of the carboxylate group, can explain the different log $K_{BSA/water}$ values of the isomers.

Another pair of chemicals that is of interest is 2,4,6-trimethylbenzene sulfonate and 2,4,6-trimethylbenzoate, which have a 2.3 log units difference between their experimental log $K_{BSA/water}$ values. This difference is also predicted correctly but it might not be solely caused by the steric hindrance of the carboxylate group. In comparison to the superimposition of the other aromatic chemicals, 2,4,6-trimethylbenzene sulfonate has a shifted position in the

alignment. This could be a hint for a different binding mode of 2,4,6-trimethylbenzene sulfonate (log $K_{BSA/water}$ exper.: 4.23 pred.: 3.52) caused by: A closer inspection of the sigma surface of 2,4,6-trimethylbenzene sulfonate shows: a) its aromatic ring exhibits a lower electron density than that of 2,4,6-trimethylbenzoate (log $K_{BSA/water}$ exper.: 1.99 pred.: 2.00) and b) the C-SO^{3-} bond (1.8 Å) is longer than the C-CO^{2-} bond (1.5 Å).[42] The latter structural feature might allow 2,4,6-trimethylbenzene sulfonate to undergo an interaction with the charged group even in the presence of the steric hindrance of the neighboring methyl groups. Furthermore, the sulfonate group has more interaction possibilities than the carboxylate group because the sulfonate group has an additional oxygen atom and the C-SO_3^- bond is better rotatable than the C-CO_2^- bond. Thus, the positions and interactions of the sp^2 orbitals of the oxygens are more flexible in case of the 2,4,6-trimethylbenzene sulfonate. These flexibilities of 2,4,6-trimethylbenzene sulfonate in the positioning and the interaction possibilities may result in a higher experimental and predicted log $K_{BSA/water}$ value compared to 2,4,6-trimethylbenzoate.

These results show that the 3D-QSAR model with LSPs as descriptors is capable of describing and predicting log $K_{BSA/water}$ for anionic and neutral chemicals. The steric effects, especially for the anionic chemicals, are successfully captured by the model. Thus, the model may be used for the prediction of unknown $K_{BSA/water}$ for neutral and anionic chemicals, which is helpful for a qualified environmental and toxicological assessment of these chemicals.

1.5.1 Domain of applicability

The domain of applicability was assessed with the help of the Tanimoto indices. Tanimoto indices[44] calculate the similarity of a test chemical against the training set. For the LSPs of two different chemicals (X and Y), the Tanimoto index is calculated as:

$$T_j(x,y) = \frac{\sum X_{ij}\, Y_{ij}}{\sum X_{ij}^2 + \sum Y_{ij}^2 - \sum X_{ij} Y_{ij}} \quad (6)$$

with X_{ij} and Y_{ij}, the j-th field values at the i-th grid point. The arithmetic mean of the Tanimoto indices of the LSP 1 to 10 (i.e., the j-th field value in eq. 1) of a test chemical was calculated against each of the chemicals in the training set. Then, the mean of the five highest values was calculated (Tanimoto index mean). Data were grouped for every Tanimoto index mean value of 0.1 (called Tanimoto groups) and compared in regard to the prediction errors of the different Tanimoto groups. The statistical difference between the variances of two Tanimoto groups was determined with a Brown-Forsythe analysis[45] and the statistical difference between the medians of two Tanimoto groups was determined with a Mann-Whitney U analysis[46]. These statistical tests were selected because the data are, most likely, not normally distributed.

The median of the prediction errors for the five random test sets apparently decreases with increasing Tanimoto index mean (Fig. 9). This may suggest that the reliability of the prediction rises with increasing Tanimoto index mean. For statistical evaluation, we chose the second highest range of Tanimoto index mean (0.60-0.70) as the reference group and tested the differences in prediction errors of all the other groups from it. We did not consider the group 0.70-0.80 because it comprises only four chemicals. Compared to the reference group, the median of the prediction errors is only significantly larger for the Tanimoto group of 0.30-0.40. No group has a significantly different variance than the reference group. Note, however, that the prediction error depends strongly on the combination of test and training sets.

Three anions (1-bromo-2-naphthoic acid anion, bromoxynil, pentachlorophenolate) that were not part of the model calibration set were used as additional validation chemicals. The prediction is accurate for 1-bromo-2-naphthoic acid anion (prediction error 0.08 log units) despite a relatively small Tanimoto index mean of 0.34. In contrast, bromoxynil anion and pentachlorophenolate were predicted with 2.47 and 2.33 log units off, respectively. Both chemicals have a Tanimoto index mean value of 0.16, which indicates a higher chance for a large prediction error. The large prediction errors for these two phenolates can be expected

because the training set does not contain any phenolate, and their low Tanimoto index means reasonably explain the outlying behavior of these chemicals. In the alignment, bromoxynil anion and pentachlorophenolate are displaced compared to the other aromatic chemicals, which might be caused by the different nature of the anionic groups of the template chemicals and of these two phenolates. For a future successful prediction of log $K_{BSA/water}$ for phenolates more experimental data for phenolates and thus a better calibration through phenolates in the training set of the 3D-QSAR model appear to be needed. Moreover, template chemicals may also need to include at least one phenolate.

Other chemicals that are expected to be out of the domain of applicability of the presented model are zwitterions and cations because they have no representation in the training set. Multiply charged anions may also be difficult to predict because the effect of the second charged group is probably not covered by the model. Other examples of chemicals that should be out of the domain of applicability are big bulky chemicals (e.g., monensin Tanimoto index mean 0.07, perfluorononanoic carboxylate Tanimoto index mean 0.09) including oligosaccharides (e.g., maltotriose Tanimoto index mean 0.12), long tertiary and quaternary organic chemicals (e.g., 4-butyl-4-pentylnonanal Tanimoto index mean 0.14), because they are not part of the current calibration set and might bind to BSA through another mechanism. The same holds true for fatty acids, which bind to a specific binding site of BSA[47] (e.g., undecane carboxylate Tanimoto index mean 0.19).

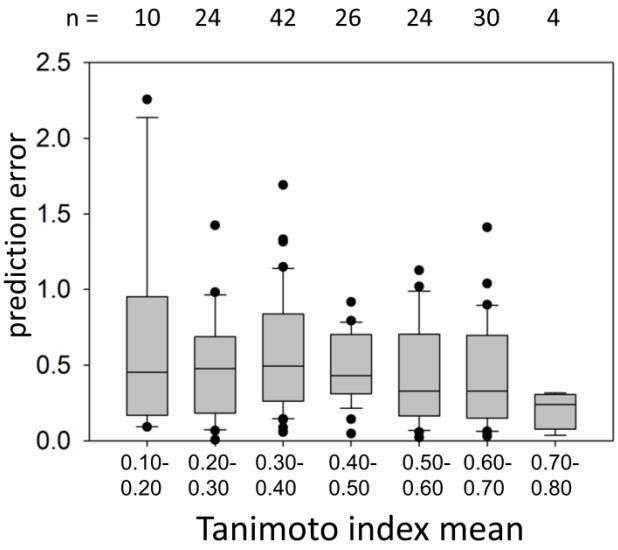


Figure 9 Prediction errors of the 3D-QSAR model plotted against the Tanimoto index range of the five most similar chemicals of the training set. The boxes outline the 25th to 75th percentiles, the lines through the centers represent the median, the whiskers indicate the 90th and 10th percentiles, and the dots indicate outlying points. The results for all five random test sets are plotted.

1.6 Conclusions

In this work, we determined αCD binding constants for systematically selected neutral organic chemicals to gain more insight into the influence of 3D steric effects on the binding to αCD. Based on the acquired data set, we established a new method for the determination of CD binding constants. The obtained results show clear steric restrictions which influence the binding process to αCD. Particularly, hydrophobic aromatic chemicals indicated clear size limitations. Another strong effect on the binding constant is caused by the position of the functional group, which restricts the length of the alkyl chain that interacts with the αCD cavity. This insight might be helpful for practical applications of CD, e.g., high affinity of αCD for linear aliphatic compounds relative to branched, inflexible compounds could be used for selective binding and separation of these chemicals.

Modeling the binding to αCD was the next step after the results of an often (over)used approach, a correlation with log K_{OW}, were less than convincing. Thus, a thorough evaluation of three different modeling approaches was done. As assumed, the description of the binding

to αCD needs to include the 3D-structure of the solutes because the 3D-QSAR model worked much better than the simple correlation with log K_{OW} and better than the 2D-QSAR model (pp-LFER) considered here. Because the COSMO 3D-QSAR performed also better than tested standard CoMFA, it can be concluded that the LSPs are more suitable variables for 3D-QSAR modeling of the binding process to αCD and probably for other binding processes as well, e.g., binding to other types of cyclodextrin with a different application range. The positive results, especially the coverage of the steric effects, suggested an applicability of the successful modeling approach to similar problems, i.e., binding processes that are also highly influenced by steric effects and not appropriately describable with other modeling approaches (log K_{OW}, pp-LFER).

An example is the partitioning between BSA and water for organic chemicals, which is strongly influenced by steric effects particularly for anionic organic chemicals. The applied 3D-QSAR successfully captured the steric effects that are responsible for up to two log units differences in log $K_{BSA/water}$ between structural isomers. The assumptions behind the generated characteristic binding site (i.e., several localized binding sites with similar chemical environments and the interaction possibilities of the sites can be expressed as an averaged characteristic binding site) appear to be adequate for the 3D-QSAR modeling approach. The discrimination between different binding sites was not necessary for successful modeling for the data set used in this work. Thus, the obtained model may be used for the prediction of unknown $K_{BSA/water}$ for neutral and anionic chemicals, which is helpful for a qualified environmental and toxicological assessment of these chemicals. Possible examples are the assessment of the freely dissolved concentration of chemicals in typical cell assays and the estimation of the bioaccumulation potential of organic anions, provided that other sorption phases such as phospholipid membranes are considered as well. The presented work is the first application of COSMO-based 3D-QSARs for binding/partitioning problems and could be the basis for applications to similar or even more advanced problems.

1.7 References

(1) Abraham, M. H.; Ibrahim, A.; Zissimos, A. M. Determination of Sets of Solute Descriptors from Chromatographic Measurements. *J. Chromatogr. A* **2004**, *1037*, 29-47.
(2) Klamt, A. Conductor-Like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224-2235.
(3) Karickhoff, S. W.; McDaniel, V. K.; Melton, C.; Vellino, A. N.; Nute, D. E.; Carreira, L. A. Predicting Chemical Reactivity by Computer. *Environ. Toxicol. Chem.* **1991**, *10*, 1405-1416.
(4) Cox, G. S.; Turro, N. J.; Yang, N. C. C.; Chen, M. J. Intramolecular Exciplex Emission from Aqueous B-Cyclodextrin Solutions. *J. Am. Chem. Soc.* **1984**, *106*, 422-424.
(5) Connors, K. A. The Stability of Cyclodextrin Complexes in Solution. *Chem. Rev.* **1997**, *97*, 1325-1357.
(6) Hedges, A. R. Industrial Applications of Cyclodextrins. *Chem. Rev.* **1998**, *98*, 2035-2044.
(7) Bai, Y.; Fan, X.-d.; Yao, H.; Yang, Z.; Liu, T.-t.; Zhang, H.-t.; Zhang, W.-b.; Tian, W. Probing into the Supramolecular Driving Force of an Amphiphilic B-Cyclodextrin Dimer in Various Solvents: Host–Guest Recognition or Hydrophilic–Hydrophobic Interaction? *J. Phys. Chem. B* **2015**, *119*, 11893-11899.
(8) Del Valle, E. M. Cyclodextrins and Their Uses: A Review. *Process Biochem.* **2004**, *39*, 1033-1046.
(9) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (Comfa). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *Journal of the American Chemical Society* **1988**, *110*, 5959-5967.
(10) Klamt, A.; Thormann, M.; Wichmann, K.; Tosco, P. Cosmosar3d: Molecular Field Analysis Based on Local Cosmo Σ-Profiles. *Journal of Chemical Information and Modeling* **2012**, *52*, 2157-2164.
(11) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of Cosmo-Rs. *J. Phys. Chem. A* **1998**, *102*, 5074-5085.
(12) Klamt, A. The Cosmo and Cosmo-Rs Solvation Models. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 699-709.
(13) Thormann, M.; Klamt, A.; Wichmann, K. Cosmosim3d: 3d-Similarity and Alignment Based on Cosmo Polarization Charge Densities. *Journal of Chemical Information and Modeling* **2012**, *52*, 2149-2156.
(14) Kragh-Hansen, U. Molecular Aspects of Ligand Binding to Serum Albumin. *Pharmacological Reviews* **1981**, *33*, 17-53.
(15) Gülden, M.; Mörchel, S.; Tahan, S.; Seibert, H. Impact of Protein Binding on the Availability and Cytotoxic Potency of Organochlorine Pesticides and Chlorophenols in Vitro. *Toxicology* **2002**, *175*, 201-213.
(16) Endo, S.; Goss, K.-U. Serum Albumin Binding of Structurally Diverse Neutral Organic Compounds: Data and Models. *Chemical Research in Toxicology* **2011**, *24*, 2293-2301.
(17) Henneberger, L.; Goss, K.-U.; Endo, S. Equilibrium Sorption of Structurally Diverse Organic Ions to Bovine Serum Albumin. *Environmental Science & Technology* **2016**, *50*, 5119-5126.
(18) Endo, S.; Goss, K. U. Applications of Polyparameter Linear Free Energy Relationships in Environmental Chemistry. *Environmental Science & Technology* **2014**, *48*, 12477-12491.
(19) Lantz, A. W.; Wetterer, S. M.; Armstrong, D. W. Use of the Three-Phase Model and Headspace Analysis for the Facile Determination of All Partition/Association Constants for

Highly Volatile Solute-Cyclodextrin-Water Systems. *Analytical and bioanalytical chemistry* **2005**, *383*, 160-166.
(20) Doong, R.; Chang, S.; Sun, Y. Solid-Phase Microextraction for Detemnining the Distribution of Sixteen Us Environmental Protection Agency Polycyclic Aromatic Hydrocarbons in Water Samples. *J. Chromatogr. A* **2000**, *879*, 177-188.
(21) Kloskowski, A.; Pilarczyk, M.; Namieśnik, J. Membrane Solid-Phase Microextraction—a New Concept of Sorbent Preparation. *Analytical Chemistry* **2009**, *81*, 7363-7367.
(22) Endo, S.; Droge, S. T. J.; Goss, K.-U. Polyparameter Linear Free Energy Models for Polyacrylate Fiber−Water Partition Coefficients to Evaluate the Efficiency of Solid-Phase Microextraction. *Analytical Chemistry* **2011**, *83*, 1394-1400.
(23) Endo, S.; Goss, K. U. Serum Albumin Binding of Structurally Diverse Neutral Organic Compounds: Data and Models. *Chem. Res. Toxicol.* **2011**, *24*, 2293-2301.
(24) Sanemasa, I.; Takuma, T.; Deguchi, T. Association of Some Polynuclear Aromatic-Hydrocarbons with Cyclodextrins in Aqueous-Medium. *Bulletin of the Chemical Society of Japan* **1989**, *62*, 3098-3102.
(25) Wang, X. B., Mark. Solubilization of Some Low-Polarity Organic Compounds by Hydroxy Propy I-A-Cyclodextrin. *Environmental Science & Technology* **1993**, *27*, 2821-2825.
(26) Tanada, S.; Nakamura, T.; Kawasaki, N.; Torii, Y.; Kitayama, S. Removal of Aromatic Hydrocarbon Compounds by Hydroxypropyl-Cyclodextrin. *Journal of Colloid and Interface Science* **1999**, *217*, 417-419.
(27) Kim, S.-J.; Kwon, J.-H. Determination of Partition Coefficients for Selected Pahs between Water and Dissolved Organic Matter. *Clean-Soil Air Water* **2010**, *38*, 797-802.
(28) Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. Hydrogen Bonding. Part 34. The Factors That Influence the Solubility of Gases and Vapours in Water at 298 K, and a New Method for Its Determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777-1791.
(29) Goss, K.-U. Predicting the Equilibrium Partitioning of Organic Compounds Using Just One Linear Solvation Energy Relationship (Lser). *Fluid Phase Equilib.* **2005**, *233*, 19-22.
(30) Gramatica, P. Principles of Qsar Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26*, 694-701.
(31) Kauffman, G. W.; Jurs, P. C. Qsar and K-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553-1560.
(32) Endo, S.; Watanabe, N.; Ulrich, N.; Bronner, G.; Goss, K.-U. Ufz-Lser Database V 2.1; Helmholtz Centre for Environmental: Leipzig, Germany, 2015.
(33) Marinescu, L.; Bols, M. Cyclodextrin Derivatives That Display Enzyme Catalysis. *Trends Glycosci. Glycotechnol.* **2009**, *21*, 309-323.
(34) Tosco, P.; Balle, T.; Shiri, F. Open3dalign: An Open-Source Software Aimed at Unsupervised Ligand Alignment. *J. Comput. Aided Mol. Des.* **2011**, *25*, 777-783.
(35) Sijm, D.; Kraaij, R.; Belfroid, A. Bioavailability in Soil or Sediment: Exposure of Different Organisms and Approaches to Study It. *Environ. Pollut.* **2000**, *108*, 113.
(36) Stanier, C. A.; O'Connell, M. J.; Clegg, W.; Anderson, H. L. Synthesis of Fluorescent Stilbene and Tolan Rotaxanes by Suzuki Coupling. *Chem. Commun.* **2001**, 493-494.
(37) Tosco, P.; Balle, T. Open3dqsar: A New Open-Source Software Aimed at High-Throughput Chemometric Analysis of Molecular Interaction Fields. *J Mol Model* **2011**, *17*, 201-208.
(38) Bronner, G.; Goss, K.-U. Predicting Sorption of Pesticides and Other Multifunctional Organic Chemicals to Soil Organic Carbon. *Environ. Sci. Technol.* **2011**, *45*, 1313-1319.

(39) Linden, L.; Goss, K.-U.; Endo, S. Exploring 3d Structural Influences of Aliphatic and Aromatic Chemicals on A-Cyclodextrin Binding. *Journal of Colloid and Interface Science* **2016**, *468*, 42-50.
(40) Suzuki, T. A Nonlinear Group Contribution Method for Predicting the Free Energies of Inclusion Complexation of Organic Molecules with A- and B-Cyclodextrins. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1266-1273.
(41) Abou-Zied, O. K.; Al-Lawatia, N.; Elstner, M.; Steinbrecher, T. B. Binding of Hydroxyquinoline Probes to Human Serum Albumin: Combining Molecular Modeling and Förster's Resonance Energy Transfer Spectroscopy to Understand Flexible Ligand Binding. *The Journal of Physical Chemistry B* **2013**, *117*, 1062-1074.
(42) Schrodinger, LLC. The Pymol Molecular Graphics System, Version 1.3r1, 2010.
(43) Peters Jr, T. All About Albumin. In *All About Albumin*; Academic Press: San Diego, 1995; pp xv-xvii.
(44) Monev, V. Introduction to Similarity Searching in Chemistry. *MATCH Commun. Math. Comput. Chem* **2004**, *51*, 7-38.
(45) Brown, M. B.; Forsythe, A. B. Robust Tests for the Equality of Variances. *Journal of the American Statistical Association* **1974**, *69*, 364-367.
(46) Mann, H. B.; Whitney, D. R. On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other. **1947**, 50-60.
(47) van der Vusse, G. J. Albumin as Fatty Acid Transporter. *Drug Metabolism and Pharmacokinetics* **2009**, *24*, 300-307.

1.8 Abbreviations

3D-QSAR	3D quantitative structure activity relationship
αCD	alpha-cyclodextrin
BSA	bovine serum albumin
CDs	cyclodextrins
CoMFA	comparative molecular field analysis
COSMO	conductor like screening model
COSMO-RS	conductor-like screening model for real solvent
ele	electrostatic
log KOW	logarithmic octanol-water partition coefficient
LSPs	local sigma profiles
MIFs	molecular interaction fields
O3A	Open3Dalign
PA	polyacrylate
PAHs	polycyclic aromatic hydrocarbons
PDMS	poly(dimethylsiloxane)
PLS	partial least square
pp-LFER	poly-parameter linear free energy relationship
QSAR	quantitative structure activity relationship
RMSE	root mean square error
vdW	van der Waals

2 Original publications

2.1 Exploring 3D structural influences of aliphatic and aromatic chemicals on α-cyclodextrin binding

Journal of Colloid and Interface Science 468 (2016) 42–50

Contents lists available at ScienceDirect

Journal of Colloid and Interface Science

journal homepage: www.elsevier.com/locate/jcis

Exploring 3D structural influences of aliphatic and aromatic chemicals on α-cyclodextrin binding

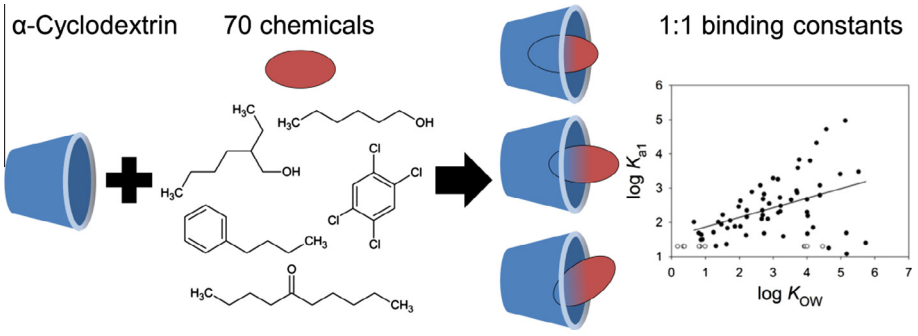
Lukas Linden [a], Kai-Uwe Goss [a,b], Satoshi Endo [a,c,*]

[a] Helmholtz Centre for Environmental Research UFZ, Permoserstr. 15, D-04318 Leipzig, Germany
[b] University of Halle-Wittenberg, Institute of Chemistry, Kurt Mothes Str. 2, D-06120 Halle, Germany
[c] Osaka City University, Urban Research Plaza & Graduate School of Engineering, Sugimoto 3-3-138, Sumiyoshi-ku, 558-8585 Osaka, Japan

HIGHLIGHTS

- Determination of 70 primary α-cyclodextrin binding constants (K_{a1}).
- Interpretation of different steric effects depending on the 3D structure of the solutes.
- The position of the functional group identified as a critical factor for log K_{a1}.
- The correlation between log K_{a1} and log K_{ow} is weak.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 10 December 2015
Revised 14 January 2016
Accepted 15 January 2016
Available online 16 January 2016

Keywords:
α-Cyclodextrin (CD)
Steric effect
Binding constant
Inclusion complex
Cyclodextrin water partitioning
Solute 3D structure

ABSTRACT

Binding of solutes to macromolecules is often influenced by steric effects caused by the 3D structures of both binding partners. In this study, the 1:1 α-cyclodextrin (αCD) binding constants (K_{a1}) for 70 organic chemicals were determined to explore the solute-structural effects on the αCD binding. K_{a1} was measured using a three-part partitioning system with either a headspace or a passive sampler serving as the reference phase. The K_{a1} values ranged from 1.08 to 4.97 log units. The results show that longer linear aliphatic chemicals form more stable complexes than shorter ones, and that the position of the functional group has a strong influence on K_{a1}, even stronger than the type of the functional group. Comparison of linear and variously branched aliphatic chemicals indicates that having a sterically unhindered alkyl chain is favorable for binding. These results suggest that only one alkyl chain can enter the binding cavity. Relatively small aromatic chemicals such as 1,3-dichlorobenzene bind to αCD well, while larger ones like tetrachlorobenzene and 3-ring aromatic chemicals show only a weak interaction with αCD, which can be explained by cavity exclusion. The findings of this study help interpret cyclodextrin binding data and facilitate the understanding of binding processes to macromolecules.

1. Introduction

The binding of small molecules to macromolecules is important in numerous processes such as enzymatic reactions, receptor binding, plasma protein binding, and drug formulation with excipients.

* Corresponding author at: Osaka City University, Urban Research Plaza & Graduate School of Engineering, Sugimoto 3-3-138, Sumiyoshi-ku, 558-8585 Osaka, Japan.
E-mail address: satoshi.endo@urban.eng.osaka-cu.ac.jp (S. Endo).

http://dx.doi.org/10.1016/j.jcis.2016.01.032
0021-9797/

While partitioning of small molecules between various homogenous phases such as solvents are well understood and quantified according to the contributions of specific molecular interactions [1], this is yet not the case for the binding to macromolecules. In contrast to homogeneous partitioning systems, the influence of the three-dimensional (3D) structure plays a decisive role in the sorption process to macromolecules. Generally, a good fit between the small molecule and the macromolecule is important for the efficiency of the binding process [2].

An example of macromolecules that are used to bind smaller chemicals are cyclodextrins (CDs). CDs are conic ring oligosaccharides and are also present naturally. CDs are usually made of 6, 7, and 8 glucopyranose units, which are named α-, β-, and γ-CD, respectively. The conic ring structure of CD generates a cavity. Its surface is mostly formed by the hydrophobic parts of the molecule [3]. The molecular structure of CDs can also be modified to increase their particular applicability, e.g., six modified CDs are widely used as excipients for clinical purposes [4]. Here, one advantage of CD as excipient is its low toxicity [4]; orally applied, CDs have shown low absorption to the blood circulation and therefore exerted no toxic effect [5].

In solution, CDs commonly form inclusion complexes (host–guest complexes) with many chemicals. Typically studied guests are drugs whose molecular mass ranges from 100 to 400 Da [6]. The specificity of CD binding appears to be relatively low, and the association constants (K [M^{-1}]) vary widely across different guest molecules: for example, protonated aniline has a log K of 0.36 for the association with α-CD [7] while decyltrimethylammonium bromide has a log K of 3.57 [8], and nucleotides can have log $K \geqslant 6$ for the association with aminocyclodextrins [7]. While CDs are sometimes considered like a normal, homogeneous phase [9], the 3D structure of the small molecules appears to play a critical role for the formation of the host–guest complex with CD [10] as for other macromolecular binding.

Energetics associated with the formation of the host–guest complex with CD are discussed in the literature based on two concepts: (a) Direct intermolecular interactions between host and guest via van der Waals forces and hydrogen bonding which are influenced by the fit between the guest and the CD cavity, and (b) additional positive energy gains through the formation of the host–guest complex. The latter includes mechanisms such as: the release of bound water from the cavity to bulk water, and the relief of conformational stress of the cyclodextrin [11]. The relative importance of the different factors on the partition process should depend on the guest molecule.

In addition to the beneficial use for clinical and other industrial purposes, CD is often considered a model macromolecule to study host–guest complexation. An advantage of using CD for studying molecular steric effects on binding behavior is its well-investigated 3D structure. The binding is flexible to some degree but restricted in the conic main structure [12]. The angles between the glucopyranose units vary depending on the solvation medium, the host–guest complex, and the aggregate state. Binding coefficients of CD give direct indications of the strength of binding, but experimental data found in the literature (e.g., [10,13]) are derived from many sources that use different methods. Thus, the data are not always comparable and the composition of the data set might not be designed to answer specific questions regarding the influence of the 3D structure.

In this study we experimentally determined a large, consistent dataset of binding constants for αCD with 70 aliphatic and aromatic chemicals such as alcohols, ethers and chlorobenzenes. The aim of this study was to identify the 3D-structural features of guest molecules that influence the binding affinity to αCD. Particularly, we sought for explanations for substantially different binding constants that we found for apparently similar chemicals.

2. Materials and methods

2.1. Materials

The chemicals were purchased from various providers- and their purity was at least 94% and mostly >98%, as listed in the supporting information (SI). There were some chiral chemicals, the chirality of which was not specified. All test chemicals used for binding experiments were first dissolved in methanol to make stock solutions. Three to five chemicals of one compound class were mixed into one stock solution. Only those chemicals that were distinctly separated through the gas-chromatographic (GC) system (see below) were mixed together. The concentration of each chemical in methanol stock solution did not exceed 10% of the water solubility so that the final concentration after dilution in water was well below the solubility limit. For all experiments pure water produced by a MilliQ Gradient A10 system (Millipore) was used. Polyacrylate (PA, coating thickness 36 μm, volume of the coating 16.5 μL/m)- and poly(dimethylsiloxane) (PDMS, coating thickness 30 μm, volume of the coating 13.2 μL/m)-coated glass fibers produced by Polymicro Technologies Inc. (Phoenix, AZ) were purchased from Optronics GmbH (Kehl, Germany). αCD was obtained from Wacker Chemie AG with a purity of at least 98.0% and a maximum residual complexant (1-decanol) of 20 ppm. αCD (0.5–2 g) was weighed into a 100 mL volumetric flask and dissolved with MilliQ water to prepare CD stock solution, which was diluted further before the binding experiment.

2.2. Instruments

The following equipment was used for the quantitative analysis: Hewlett Packard GC System HP 6890 series gas-chromatographs with a flame ionization detector (FID) or an electron capture detector (ECD), both systems connected to an HP 7694 Headspace Sampler; an Agilent 7890A GC System equipped with a 5975C inert MSD Triple Axis Detector and a Gerstel Multi Purpose Sampler (MPS 2XL). The chemicals were analyzed on either of the following two columns from Agilent Technologies: HP-1 (30 m × 0.32 mm i.d., 4 μm film thickness), or HP-5MS (30 m × 0.25 mm i.d., 0.25 μm film thickness).

2.3. Binding experiments

Binding constants for 70 chemicals were measured in batch systems. The used methods have been described in detail previously [14] and are briefly explained below. In both methods, the unbound, freely dissolved concentration of the chemical was determined via the measurement of a third phase, either air (headspace approach) or a PA or PDMS fiber (passive sampling approach). All binding experiments were performed at 30 °C, which was the lowest possible temperature that the sample tray of the GC autosampler was able to control.

2.3.1. Headspace approach

Air was the common third phase (reference phase) for this approach [15]. Two groups of weighed 20 mL vials were prepared with four vials per group. One group was filled with 5 mL water and the other was filled with 5 mL αCD solution (2–15 g/L). The vials were spiked with 10 or 25 μL of methanolic stock solution of the selected chemicals and were immediately closed with a PTFE- or aluminum-lined silicone septum to prevent loss of the chemicals. From the experience of preliminary experiments, the equilibrium time was set to a minimum of four hours: first three hours on a horizontal shaker at 30 °C with 300 rpm and then at least one hour on the GC-sample tray at 30 °C with low shaking

speed. Then the headspace was probed with a 100 μL sampling loop or a 250 μL syringe and injected into the GC and measured with GC–FID/ECD or GC–MS.

2.3.2. Passive sampling approach

The passive sampling approach was used for chemicals which are not volatile enough for the headspace approach. PA or PDMS fiber is the common reference phase for this approach [16,17]. The experimental setting was similar to that with the headspace approach except for the following changes. The volume of the solutions and the vials was 10 mL, each vial received 5 or 10 cm of PA- or PDMS-coated fiber and the equilibrium time was 72 h at 30 °C. Previous studies [18,19] confirmed that this equilibrium time is sufficient for a wide range of chemicals. After equilibrium was reached, the PA/PDMS fibers were removed from the vials and carefully wiped with a clean tissue. Then the fibers were extracted overnight on a roll mixer using 200 μL of cyclohexane (for PDMS) or ethyl acetate (for PA). The concentrations of the extract were quantified with the GC–MS system using external calibration.

2.4. Data analysis

The results of GC analyses were evaluated using the same approach as Geisler et al. [14]. Thus, the partition coefficient between the αCD solution and water ($K_{CD\ solution/water}$ [$L_{water}/L_{CD\ solution}$]) was determined from the relative GC peak areas of vials with and without αCD. In cases where partitioning of the chemical into air or the fiber contributed substantially to the mass balance, this was considered in the calculation of $K_{CD\ solution/water}$ based on known air–water or fiber–water partition coefficient of the chemical [14]. The resulting $K_{CD\ solution/water}$ was used to derive the partition coefficient between αCD and water ($K_{CD/water}$ [L_{water}/kg_{CD}]) according to:

$$K_{CD\ solution/water} = K_{CD/water} c_{CD} + f_{water} \quad (1)$$

c_{CD} is the concentration of αCD in the αCD solution [$kg_{CD}/L_{CD\ solution}$] and f_{water} is the volume fraction of water in the αCD solution [$L_{water}/L_{CD\ solution}$]. As c_{CD} was only up to 0.015 kg/L in this work, f_{water} was considered unity.

2.5. Stoichiometry of the complexes

The stoichiometry of the CD host–guest complexes can differ depending on the types and concentrations of CD and guest [20–22]. The common stoichiometry is 1:1 and 2:1 but other kinds of complexes can be found in the literature [8,23–27]. Previous studies show that smaller chemicals tend to form only 1:1 complexes, while larger ones can also form 2:1 complexes [28]. For studying interactions of CD with various guest molecules, it is advantageous to compile consistent data for 1:1 complexes.

Complexation constants for 1:1 and 2:1 binding can be expressed as,

$$K_{a1} = \frac{[S\alpha CD]}{[S][\alpha CD]} \quad (2)$$

$$K_{a2} = \frac{[S(\alpha CD)_2]}{[S\alpha CD][\alpha CD]} \quad (3)$$

$$K_{a1}K_{a2} = \frac{[S(\alpha CD)_2]}{[S][\alpha CD]^2} \quad (4)$$

where K_{a1} and K_{a2} [M^{-1}] are the formation constants for the 1:1 and 2:1 complexes, respectively. S is the substrate (guest), and SαCD and $S(\alpha CD)_2$ are the 1:1 and 2:1 complexes, respectively. Assuming that 1:1 and 2:1 binding is dominant, the partition coefficient introduced in Eq. (1) has the following relationship with the binding constants:

$$K_{CD\ solution/water} = f_{water}\left(K_{a1}[\alpha CD] + K_{a1}K_{a2}[\alpha CD]^2 + 1\right) \quad (5)$$

If K_{a2} is very small, then the second term in the parentheses and therefore 2:1 binding become negligible and $K_{CD\ solution/water}$ depends linearly on the concentration of αCD. This implies that a linear relationship between $K_{CD\ solution/water}$ and [αCD] indicates a dominance of 1:1 binding. The relative importance of 2:1 binding becomes higher with higher [αCD], as indicated by the squared concentration in the second term of Eq. (5). We experimentally tested the influence of the αCD concentration (0.5, 1, 5, 10, 15, 20 g/L) on $K_{CD\ solution/water}$ for chemicals with a long linear alkyl chain (i.e., 6-undecanone, 2-undecanone, 5-decanone, 2-decanone, dihexylether, dipentylether, dibutylether), because they are more likely to form 2:1 complexes with αCD than their shorter analogues [29]. Also, we tested the influence of [αCD] on the binding of three aromatic chemicals of different sizes (chlorobenzene, 1,3-dichlorobenzene, pentachlorobenzene) to ensure 1:1 binding. Eq. (5) was fitted to these concentration dependent data for $K_{CD\ solution/water}$ to obtain K_{a1} and K_{a2}. For the other chemicals, single concentration data for $K_{CD\ solution/water}$ were measured and used to derive K_{a1}, assuming no significant 2:1 binding (i.e., $K_{a2} = 0$ in Eq. (5)).

3. Results and discussion

3.1. Measurements of K_{a1}

The concentration dependent measurements (Table 1; also see figures in SI) show that $K_{CD\ solution/water}$ increases linearly with [CD] for chlorobenzene, 1,3-dichlorobenzene, and pentachlorobenzene, suggesting dominant 1:1 binding. For 6-undecanone, 2-undecanone, dihexylether, dipentylether, dibutylether, 5-decanone, and 2-decanone, the relationship was nonlinear to varying degree. Nevertheless, these chemicals also form mostly 1:1 complexes if the αCD concentration is smaller than 5 g/L (i.e., 5 mM), as suggested by Eq. (5) in combination with K_{a1} and K_{a2} obtained by fitting. At [CD] ⩽ 2 mM, the second term in Eq. (5) has only a small contribution, and K_{a1} could also be derived from a single concentration measurement with an error below 0.15 log units. For these reasons, we conclude that it is valid to assume a dominance of 1:1 binding for the studied aromatic chemicals in the [CD] range we used and for the tested aliphatic chemicals in the low [CD] range, which allows the use of Eq. (5) with $K_{a2} = 0$ to derive K_{a1}. This assumption is consistent with the literature, where it is reported that short chain surfactants (C8) do not form 2:1 complexes to a considerable degree [30] and that long chain surfactants (C10 and C12) form mostly 1:1 complexes with some contributions from 2:1 complexes [8,31,32].

Table 1
K_{a1} and K_{a2} derived from the concentration dependent measurement.

Chemical	K_{a1} (M^{-1})	Standard error	K_{a2} (M^{-1})	Standard error
6-Undecanone	8.13×10^2	0.24×10^2	7.6×10^1	0.5×10^1
2-Undecanone	6.31×10^3	0.46×10^3	9.7×10^1	1.8×10^1
Dihexylether	2.55×10^3	0.28×10^3	1.28×10^2	0.27×10^2
Dipentylether	4.65×10^2	0.61×10^2	1.10×10^2	0.25×10^2
Dibutylether	2.09×10^2	0.19×10^2	4.4×10^1	1.4×10^1
5-Decanone	5.24×10^2	0.62×10^2	4.5×10^1	1.6×10^1
2-Decanone	3.86×10^3	0.03×10^3	6.7×10^1	0.2×10^1
Chlorobenzene	1.25×10^2	0.24×10^2	0.9×10^1	1.5×10^1
1,3-Dichlorobenzene	4.69×10^2	0.14×10^2	0.6×10^1	0.2×10^1
Pentachlorobenzene	1.2×10^1	0.4×10^1	0	

The K_{a1} values of 70 chemicals were determined in batch experiments. The chemical set comprises: 19 alcohols, 19 ketones, 9 polycyclic aromatic hydrocarbons (PAHs), 6 chlorobenzenes, 5 alkylbenzenes, 4 ethers, 4 nitroalkanes, and 4 phosphates/phosphonates. These chemicals have various functional groups but relatively simple molecular structures, which should facilitate interpretation of the results. Moreover, the data set includes multiple series of chemicals with increasing number of structural units (i.e., —CH_2—, Cl—, aromatic ring), enabling the assessment of incremental effects on the binding behavior. We avoided measuring complex chemicals like drugs, because they have many different chemical features which could cause indistinguishable effects on the binding to αCD. The measured log K_{a1} values span over a wide range, from 1.08 (pentachlorobenzene) to 4.97 (1-dodecanol). For 10 chemicals, K_{a1} was too small to measure with the applied method (denoted with log K_{a1} < 1.3 in Table 2).

3.2. Correlation with the octanol–water partition coefficient

The log octanol–water partition coefficient (K_{ow}) is often related to log $K_{CD/water}$ [33–36] and was even proposed as a descriptor for predictions [34]. The log K_{a1} values measured in this study are compared to log K_{ow} in Fig. 1. The experimental log K_{ow} values were taken from the EPIsuite 4.1 database [37] if available, and otherwise, KOWWIN was used to estimate log K_{ow}.

As shown in Fig. 1, the correlation between log K_{ow} and log K_{a1} is weak, with R^2 of 0.19. Correlation is particularly weak in the high K_{ow} range (i.e., log K_{ow} > 3). For example, 1-dodecanol and pentachlorobenzene have similar log K_{ow} values (5.13 and 5.17, respectively) but differ more than 3 log units in their K_{a1} values (4.96 and 1.08, respectively). Conversely, nitroethane and phenanthrene have >4 log units different log K_{ow} values (0.18 and 4.46, respectively) but log K_{a1} values are both <1.3. As has been demonstrated by others, the correlation between log K_{ow} and log $K_{CD/water}$ can be higher if only one chemical class is considered [35,38]. However, log K_{ow} is not useful for the understanding of the specific binding process to αCD, nor for estimating log $K_{CD/water}$ if different classes are considered. Obviously, log K_{ow} does not account for the 3D structure effects. Therefore, in the following sections, we analyze the data further and try to identify the specific steric factors influencing the binding process to αCD.

3.3. Aliphatic compounds

The experimentally derived binding coefficients of chemicals with one or more linear alkyl chains (14 alcohols, 14 ketones, 4 nitroalkanes, 4 ethers, and 3 phosphates) are plotted in log units against the number of carbon atoms (Fig. 2). More specifically, we include here (1) linear aliphatic compounds with the polar functional group at the end of the molecule, i.e., R—OH, R—C(=O)CH_3, and R—NO_2, where R is a linear alkyl chain of differing lengths, (2) aliphatic compounds with the polar functional group in the middle of the molecule, i.e., R—C(OH)—R′, R—C(=O)—R′, and R—O—R, where R′ = R or R—CH_2— (i.e., one unit longer), and (3) trialkyl phosphates (i.e., PO_4-RRR). The log K_{a1} values within each homologous group increase linearly with the number of carbon atoms. It is remarkable that log K_{a1} of *n*-alkan-1-ols (R-OH) increase linearly from 1-butanol to 1-decanol. The torus of αCD has a height of 8 Å [11], whereas the distance between the atomic nuclei of C1 and C10 of decanol is around 11.5 Å for the stretched conformer [39]. Theoretically, around 3 Å of the stretched 1-decanol molecule should stick out of the αCD cavity and experience water as the surrounding phase, from which there is no energy gain to be expected for the sorption process. Thus, the linear increase of log K_{a1} for *n*-alkan-1-ols suggests that the alkyl chain or αCD–or both–adapt their conformation, which enables optimal interactions between the alkyl chain and αCD for molecules that are too long to fit into the cavity in their stretched conformation. Comparable data have been published for surfactants (alkyltrimethylammonium bromide/chloride) in regard to the chain length; K_{a1} values larger than 10^4 M^{-1} are reported for 12 carbon atoms [8]. The K_{a1} values of 1-undecanol and 1-dodecanol measured in this study are even higher than that of 1-decanol but the stoichiometry might not be solely 1:1, as the concentration dependent measurements in this study do not cover chemicals with a linear alky chain length >9. We were not able to perform these experiments with 1-undecanol and 1-dodecanol because the GC peak shapes were too distorted. A similar trend for alcohols has been reported also in the literature (see SI) but with 1-nonanol as the largest chemical.

Chemicals with the functional group at the end of the molecule have generally higher K_{a1} than chemicals with the same functional group in the middle, when compared at the same number of C (Fig. 2). Moreover, the slopes of the three end-substituted chemical classes are similar, and so are the slopes of the three middle-substituted classes. However, there is a substantial difference in the slopes between end-substituted and middle-substituted classes (0.40 and 0.26 log units/C on average, respectively). That means that elongation of the molecule in one direction increases K_{a1} more than elongation in two directions, per carbon atom. Trialkyl phosphates (elongation in three directions) exhibit an even smaller slope (0.1 log units/C). Such a differential increase per C does not occur with solvent–water partition coefficients such as K_{ow} and thus has to be caused by steric effects.

The observations above are consistent with the concept available in the literature [40,41] that the polar functional group of the bound guest molecule stays outside the hydrophobic cavity and interacts with the surrounding water or with one of the hydroxyl groups of the CD rims. Thus, the polar functional group restricts the location and the orientation of the guest relative to CD and can thereby hinder the optimal interactions of the alkyl chain(s) with the CD cavity. It is plausible that the polar functional group stays outside the cavity, because the polar functional group of the free, unbound chemical can undergo strong hydrogen bonding interactions with water molecules, whereas hydrogen bonds cannot be formed inside the hydrophobic cavity of CD. Thus, the polar functional group could enter the cavity only if that leads to a free energy gain that is larger than the free energy loss due to the breakup of hydrogen bonds with water. Assuming that the polar functional group has to be outside the cavity, end-substituted chemicals may still fully insert their alkyl chain into the cavity, whereas middle-substituted chemicals may not insert both chains well in the cavity.

It is also worth noting that the regression lines for end- and middle-substituted chemicals intersect at 2 and 3 carbon atoms for alcohols and ketones, respectively (Fig. 2). These intersections correspond to ethanol and acetone. This is reasonable, as the structural difference between end- and middle substitutions diminishes with decreasing number of carbon atoms.

To obtain more insights into the binding mechanisms of aliphatic compounds to αCD, we estimated the number of carbon atoms within the cavity based on the following, very simple assumptions: 1. The polar functional group cannot enter the cavity. 2. Only one alkyl chain per molecule can enter the cavity, and the number of encapsulated carbon atoms is not restricted by the height of the CD torus. 3. If there are two chains or more with differing lengths, then the longest one enters the cavity (see Fig. 3; only the numbered carbon atoms are assumed to be in the cavity). In Fig. 4, log K_{a1} is plotted against the number of encapsulated carbon atoms estimated this way.

There are several findings in Fig. 4. First, plots for end-functionalized alcohols, ketones and nitroalkanes overlap each

Table 2
Experimental αCD/water binding coefficients of the measured chemicals.

Class	Chemical	log K_{a1}	SD or SE
End-substituted alcohols	1-Butanol	1.64	0.09
	1-Pentanol	2.22	0.04
	1-Hexanol	2.63	0.03
	1-Heptanol	3.08	0.04
	1-Octanol	3.29	0.08
	1-Nonanol	3.83	0.21
	1-Decanol	4.32	0.06
	1-Undecanol	4.72	0.06
	1-Dodecanol	4.97	0.07
Middle-substituted alcohols	3-Hexanol	1.83	0.07
	4-Heptanol	2.16	0.05
	4-Octanol	2.35	0.04
	5-Nonanol	2.48	0.04
	5-Decanol	2.86	0.12
Branched alcohols	2-Methyl-2-propanol	<1.3	
	3-Ethyl-3-pentanol	1.62	0.11
	2-Ethyl-1-hexanol	2.82	0.06
	3-Ethyl-3-hexanol	2.21	0.05
	4-Ethyl-3-hexanol	2.10	0.05
End-substituted ketones	2-Pentanone	1.50	0.02
	2-Hexanone	2.01	0.01
	2-Heptanone	2.46	0.01
	2-Octanone	2.89	0.01
	2-Nonanone	3.26	0.06
	2-Decanone[a]	3.59	0.01
	2-Undecanone[a]	3.80	0.03
Middle-substituted ketones	3-Pentanone	<1.3	
	3-Hexanone	1.71	0.08
	4-Heptanone	2.05	0.06
	4-Octanone	2.35	0.05
	5-Nonanoe	2.55	0.06
	5-Decanone[a]	2.72	0.05
	6-Undecanone[a]	2.91	0.01
Branched and cyclic ketones	3-Methyl-2-butanone	<1.3	
	4-Methyl-2-pentanone	1.31	0.09
	Cyclopentanone	<1.3	
	Cyclohexanone	<1.3	
	Cycloheptanone	1.36	0.38
Nitroalkanes	Nitroethane	<1.3	
	1-Nitropropane	1.49	0.05
	1-Nitrobutane	1.94	0.03
	1-Nitrohexane	2.68	0.04
Ethers	Dipropyl ether	1.88	0.31
	Dibutyl ether[a]	2.32	0.04
	Dipentyl ether[a]	2.67	0.05
	Dihexyl ether[a]	3.41	0.04
Trialkyl phosphates	Triethyl phosphate	1.69	0.15
	Tri-*n*-propyl phosphate	1.88	0.09
	Tri-*n*-butyl phosphate	2.30	0.07
	Diethyl ethylphosphonate	2.01	0.17
Alkylbenzenes	Toluene	1.68	0.05
	Ethylbenzene	2.28	0.02
	n-Propylbenzene	2.93	0.01
	n-Butylbenzene	3.08	0.05
	n-Hexylbenzene	3.48	0.13
Chlorobenzenes	Chlorobenzene[a]	2.10	0.08
	1,3-Dichlorobenzene[a]	2.67	0.01
	1,2,4-Trichlorobenzene	2.19	0.03
	1,2,4,5-Tetrachlorobenzene	1.25	0.07
	Pentachlorobenzene[a]	1.08	0.12
	Hexachlorobenzene	1.40	0.26
PAHs	Naphthalene	1.59	0.07
	Acenaphthylene	<1.3	
	Acenaphthene	<1.3	
	Biphenyl	1.67	0.16
	1-Chloronaphthalene	<1.3	
	Fluorene	1.85	0.25
	Phenanthrene	<1.3	
	Dibenzothiophene	2.79	0.10
	Fluoranthene	1.69	0.21

[a] Data are derived from the concentration dependent measurement. The error presented is SE for these data. For the other data, the error shown is SD.

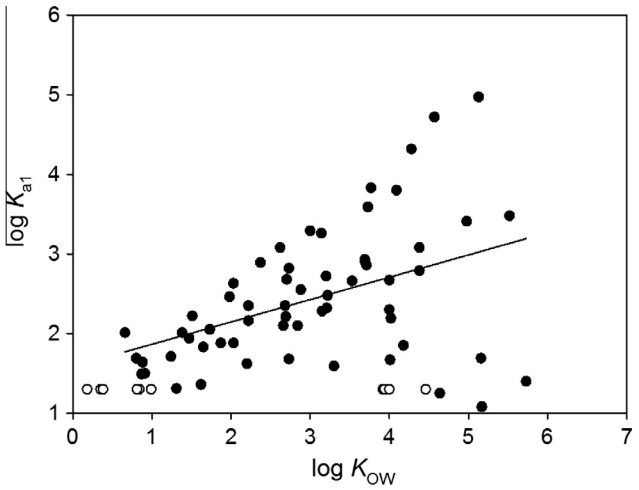


Fig. 1. Experimental 1:1 αCD binding coefficients versus octanol/water partition coefficients. The log K_{a1} values were determined at 30 °C and the log K_{ow} values were from the EPIsuite 4.1 database or predicted with KOWWIN. The white circles indicate chemicals with log K_{a1} < 1.3. The solid line indicates the linear regression of the black circles.

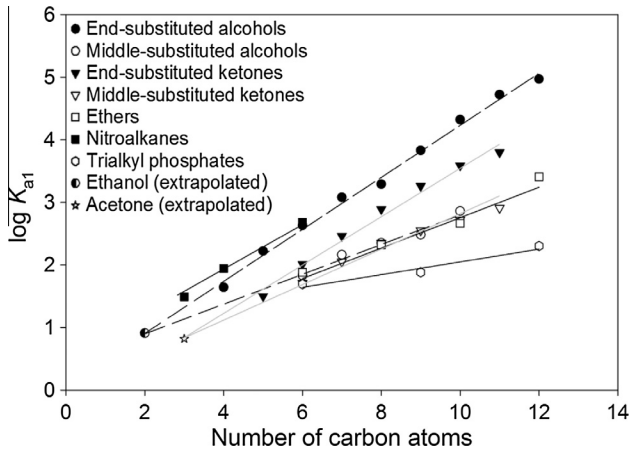


Fig. 2. Experimental K_{a1} for αCD versus number of carbon atoms. The lines indicate the linear regressions. The dashed lines are the regressions for end- and middle-substituted alcohols which are extrapolated to the intersection. The grey lines are the regressions for end- and middle-substituted ketones which are extrapolated to the intersection. The data points for ethanol and acetone were derived by extrapolation and have not been measured.

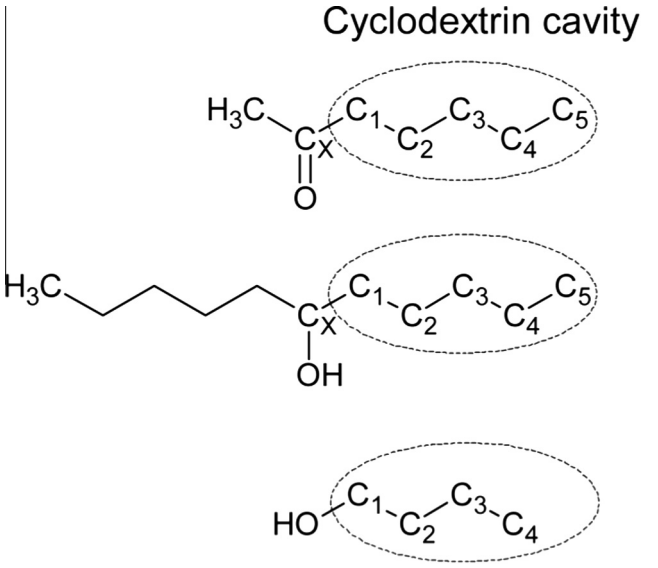


Fig. 3. Scheme for estimating the number of encapsulated carbon atoms by the αCD cavity. The numbered carbon atoms are assumed to be in the CD cavity, which is represented as the dotted circle.

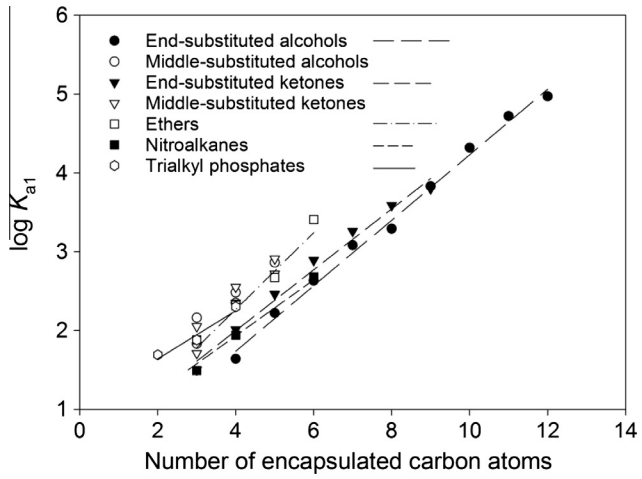


Fig. 4. Experimental K_{a1} versus estimated number of encapsulated carbon atoms. The lines indicate the linear regressions. The linear regressions for middle-substituted alcohols and ketones are omitted for clarity.

other; that is, end-functionalized linear chemicals from different classes have similar log K_{a1} when the estimated number of carbon atoms in the cavity is identical. This shows that the functional group influences K_{a1} similarly for alcohols, ketones and nitroalkanes. If the functional group interacts with water in both bound and unbound states, its net energy contribution to the binding is close to zero, and thus the type of functional group has only a limited influence on K_{a1}, agreeing with the data in Fig. 4. Note that the functional groups considered here are all polar and can form hydrogen bonds with water. Second, slopes and positions of middle-functionalized chemicals and trialkyl phosphates are now close to those of end-functionalized chemicals. Third, log K_{a1} of the middle-substituted chemicals are consistently higher than those of the end-substituted chemicals when compared at the same number of estimated C. This shift is roughly between 0.2–0.5 log units and smaller than the increment per C for the end-functionalized chemicals. This supports the assumption that only one alkyl chain enters the cavity. The shift itself may be explained by some additional energy gain from the alkyl chain that is assumed outside the cavity or the carbon atom that is designated as C_X in Fig. 4. Perhaps, these carbon atoms can still interact with the rim or the outside of the CD molecule. But energy gain from the non-encapsulated carbon atoms appears to be much less than that of the encapsulated ones.

It has to be repeated that the middle-substituted alcohols and ketones plotted in Fig. 4 are not all symmetric but include those chemicals with one chain that is one CH_2 unit longer than the other chain. The middle-substituted alcohols and ketones do show linear increase depending on the encapsulated number of carbon atoms in the cavity but the data scatter more than end-substituted chemicals. This also suggests some interactions between the non-encapsulated alkyl group and αCD.

Overall, the concept that polar functional group stays outside the cavity and only one alkyl chain enters the cavity can explain

the data trends of mono-functional, linear aliphatic chemicals very well, but additionally, αCD seems to interact with the parts of the guest molecule that are not considered encapsulated in the cavity.

3.4. Aliphatic chemicals with branched alkyl chain

We measured five alcohols with 8 carbon atoms: 1-octanol, 2-ethyl-1-hexanol, 4-octanol, 3-ethyl-3-hexanol, and 4-ethyl-3-hexanol. The log K_{a1} values decrease in this order (Fig. 5) and also the length of the longest linear, non-branched alkyl chain within the molecule. These data also imply a favorable interaction between αCD and a linear alkyl chain, although not excluding possible interactions of the rest of the molecule with αCD.

While K_{a1} of a chemical with an ethyl-branched alkyl chain is lower than that of its non-branched isomer, the energetic contribution of the additional ethyl group is always positive. Hence, log K_{a1} is higher for 2-ethyl-1-hexanol (2.81) than for 1-hexanol (2.62), and log K_{a1} of 3-ethyl-3-hexanol and 4-ethyl-3-hexanol is higher than that of 3-hexanol. It is thus apparent that the branched ethyl group can also interact with CD and has a significant contribution to K_{a1}.

3.5. Aromatic chemicals

The aromatic chemicals studied in this work are nine PAHs, six chlorobenzenes, and five alkylbenzenes. The alkylbenzenes contain one linear alkyl chain of increasing length, but log K_{a1} is not a simple linear function of the number of C atoms (Fig. 6), in contrast to the polar aliphatic compounds shown above. The increment in log K_{a1} is 0.62 per additional methylene unit from toluene to propylbenzene, which is higher than the mean increment of 0.40 for the end-substituted linear aliphatic compounds. From propylbenzene to hexylbenzene the increment in log K_{a1} per additional methylene unit is 0.19. The benzene ring does not form a strong H-bond with water and thus can favorably enter the hydrophobic cavity of αCD. As the benzene ring occupies a fraction of the cavity, alkylbenzenes possessing an alkyl chain with three or more carbon atoms appear to experience a steric effect that lowers the log K_{a1} increase per C.

Chlorobenzenes represent an even more pronounced example of the influence of steric restriction. The log K_{a1} values are above 2 for chemicals possessing one to three chlorine atoms; 1,3-dichlorobenzene has the highest log K_{a1} (2.66), after which log K_{a1} starts to decrease with an increasing number of chlorine atoms. The 1,2,4-trichlorobenzene has a log K_{a1} (2.19) that is around 0.8 log units higher than the log K_{a1} of 1,2,4,5-tetrachlorobenzene, pentachlorobenzene, and hexachlorobenzene. Mono and 1,3-dichlorobenzenes appear to fit into the cavity, whereas 1,2,4-trichlorobenzene already experiences a negative steric effect. The log K_{a1} for 1,2,4,5-tetrachlorobenzene is even lower than that of monochlorobenzene, suggesting that the three additional chlorine-substitutions hinder the interactions of the benzene ring and the original chlorine atom with αCD.

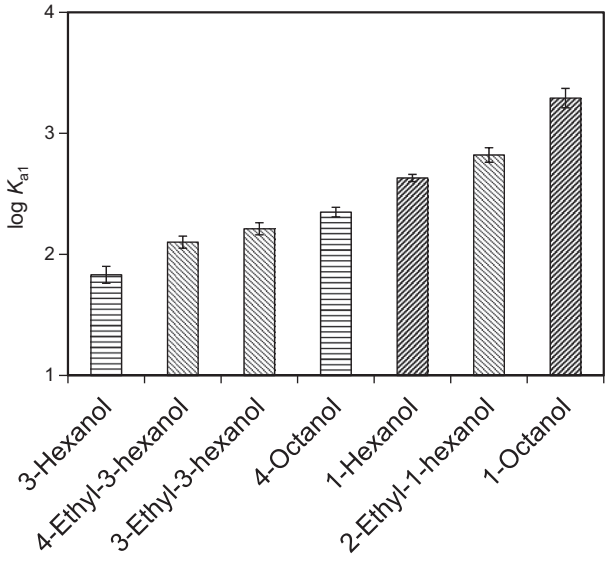


Fig. 5. Experimental K_{a1} for αCD binding of 2 C_6-alcohols and 5 C_8-alcohols.

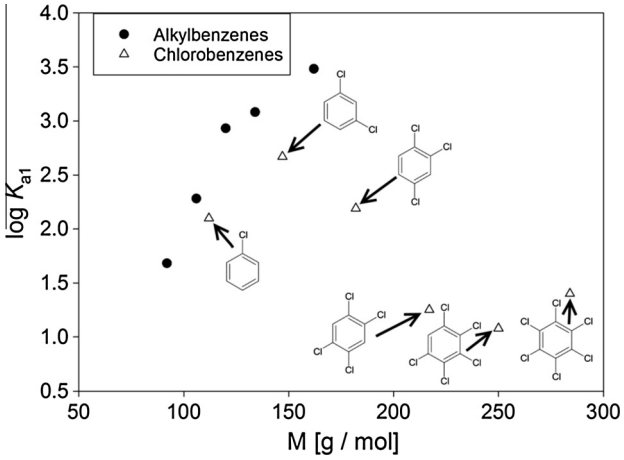


Fig. 6. Experimental K_{a1} versus molecular weight for alkylbenzenes (circles) and chlorobenzenes (triangles). The molecular structures show the corresponding chlorobenzenes.

PAHs also exhibit clear steric restriction. Toluene (1.68), biphenyl (1.67) and naphthalene (1.59) have similar log K_{a1} values, even though the latter two molecules are much larger, indicating that more than one aromatic six-ring cannot fully fit into the αCD cavity (Table 2). Moreover, naphthalene has a higher log K_{a1} value than 1-chloronaphthalene, acenaphthene, acenaphthylene, and phenanthrene. This suggests that there is no room left in the cavity for an element other than hydrogen around the naphthalene structure.

Dibenzothiophene, an aromatic three-ring system with a sulfur atom, exhibits the highest log K_{a1} (2.79) of all PAH-related compounds tested. It is interesting that its log K_{a1} value is 0.94 log units higher than that of fluorene (1.85), the structural analogue of dibenzothiophene with S being replaced by C. It is unknown why there is such a large difference in K_{a1} between dibenzothiophene and fluorene. The 3D structure and the electron distribution appear similar, as indicated by the sigma profiles and the COSMO files generated by quantum chemical software Turbomole [44]. Our experimental observation is consistent with the literature data for βCD binding constants of fluorene (log $K_{\beta CD}$ = 3.08 [42]) and dibenzothiophene (log $K_{\beta CD}$ = 3.48 [43]). The difference is smaller than that of αCD but the trend is similar. We also tried to measure the binding constants of pyrene, chrysene, benzo[*a*]pyrene and benzo[*b*]fluoranthene, but the binding to αCD was too weak to measure (pyrene) or concentrations in the PDMS fiber were below our detection limit.

4. Conclusions

In this study, we observed clear steric restrictions which influence the binding process to αCD. Particularly, hydrophobic aromatic chemicals indicated clear size limitations. The width of the αCD cavity may be represented by 1,3-dichlorobenzene or

naphthalene, and any bulkier chemicals do not fit well into the cavity. Considering this clear size effect on aromatic compounds, it is surprising that linear aliphatic chemicals with an alkyl chain longer than nine carbon atoms still exhibit high binding coefficients. The question remains: how is the long alkyl chain able to fit into the cavity or how else is the high interaction energy explainable? There should be an adaption of the alkyl chain to the cavity as well as some other parts of the αCD. This would require some bending of the alkyl chain. Crystallographic data or molecular dynamics simulations would be needed to obtain a better insight into this phenomenon.

Another finding is that the binding of polar aliphatic chemicals strongly depends on the position of the functional group which restricts the length of the alkyl chain interacting with the αCD cavity. In general, a long unhindered alkyl chain appears to act as an anchor on a bigger molecule and should be able to bind to αCD. The general trends identified in this data set of organic chemicals should provide useful information for practical applications of αCD. For example, high affinity of αCD for linear aliphatic compounds relative to branched, inflexible compounds could be used for selective binding and separation of these chemicals.

The experimental data of this work are highly consistent and diverse and thus are useful in a quantitative modeling approach for log K_{a1}. The simple correlation with log K_{ow} was found to be too low to be useful for prediction purposes. Our results strongly indicate that any modeling approach should consider the 3D structure of the αCD and the guest molecule. We are currently working on the use of 3D quantitative structure activity relationships for modeling of the binding data presented above, which will be reported in an upcoming article.

Acknowledgements

The authors thank Wacker Chemie AG for donation of α-cyclodextrin, Andrea Pfennigsdorff for lab assistance and the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE) for financial support. SE acknowledges the financial support from the MEXT/JST Tenure Track Promotion Program.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jcis.2016.01.032.

References

[1] S. Endo, K.U. Goss, Applications of polyparameter linear free energy relationships in environmental chemistry, Environ. Sci. Technol. 48 (21) (2014) 12477–12491.
[2] A. Herrmann, Dynamic combinatorial/covalent chemistry: a tool to read, generate and modulate the bioactivity of compounds and compound mixtures, Chem. Soc. Rev. 43 (6) (2014) 1899–1933.
[3] G.S. Cox et al., Intramolecular exciplex emission from aqueous β-cyclodextrin solutions, J. Am. Chem. Soc. 106 (2) (1984) 422–424.
[4] T. Loftsson, M.E. Brewster, Cyclodextrins as functional excipients: methods to enhance complexation efficiency, J. Pharm. Sci. 101 (9) (2012) 3019–3032.
[5] T. Loftsson, M.E. Brewster, Pharmaceutical applications of cyclodextrins: effects on drug permeation through biological membranes, J. Pharm. Pharmacol. 63 (9) (2011) 1119–1135.
[6] A. Vyas, S. Saraf, S. Saraf, Cyclodextrin based novel drug delivery systems, J. Incl. Phenom. Macrocycl. Chem. 62 (2008) 23–42.
[7] H.J. Schneider, Binding mechanisms in supramolecular complexes, Angew. Chem. Int. Ed. Engl. 48 (22) (2009) 3924–3977.
[8] A.J. Valente, O. Soderman, The formation of host-guest complexes between surfactants and cyclodextrins, Adv. Colloid Interface Sci. 205 (2014) 156–176.
[9] M.H. Abraham, Characterization of some GLC chiral stationary phases: LFER analysis, Anal. Chem. 69 (1997) 613–617.
[10] T. Suzuki, M. Ishida, W.M. Fabian, Classical QSAR and comparativemolecular field analyses of the host-guest interaction of organic molecules with cyclodextrins, J. Comput. Aided Mol. Des. 14 (2000) 669–678.
[11] H.M.C. Marques, A review on cyclodextrin encapsulation of essential oils and volatiles, Flavour Fragr. J. 25 (5) (2010) 313–326.
[12] R. Arun, Cyclodextrins as drug carrier molecule: a review, Sci. Pharm. 76 (4) (2008) 567–598.
[13] J.B. Ghasemi et al., Docking and 3D-QSAR study of stability constants of benzene derivatives as environmental pollutants with α-cyclodextrin, J. Incl. Phenom. Macrocycl. Chem. 73 (1–4) (2012) 405–413.
[14] A. Geisler, S. Endo, K.U. Goss, Partitioning of polar and non-polar neutral organic chemicals into human and cow milk, Environ. Int. 37 (7) (2011) 1253–1258.
[15] A.W. Lantz, S.M. Wetterer, D.W. Armstrong, Use of the three-phase model and headspace analysis for the facile determination of all partition/association constants for highly volatile solute-cyclodextrin-water systems, Anal. Bioanal. Chem. 383 (2) (2005) 160–166.
[16] R. Doong, S. Chang, Y. Sun, Solid-phase microextraction for detemnining the distribution of sixteen US Environmental Protection Agency polycyclic aromatic hydrocarbons in water samples, J. Chromatogr. A 879 (2) (2000) 177–188.
[17] A. Kloskowski, M. Pilarczyk, J. Namieśnik, Membrane solid-phase microextraction—a new concept of sorbent preparation, Anal. Chem. 81 (17) (2009) 7363–7367.
[18] S. Endo, S.T.J. Droge, K.-U. Goss, Polyparameter linear free energy models for polyacrylate fiber–water partition coefficients to evaluate the efficiency of solid-phase microextraction, Anal. Chem. 83 (4) (2011) 1394–1400.
[19] S. Endo, K.U. Goss, Serum albumin binding of structurally diverse neutral organic compounds: data and models, Chem. Res. Toxicol. 24 (12) (2011) 2293–2301.
[20] L.X. Song et al., Inclusion complexation, encapsulation interaction and inclusion number in cyclodextrin chemistry, Coord. Chem. Rev. 253 (9–10) (2009) 1276–1284.
[21] M. Másson et al., investigation of drug-cyclodextrin complexes by a phase-distribution method: some theoretical and practical considerations, Chem. Pharm. Bull. 53 (8) (2005) 958–964.
[22] W. Daniel, F.N. Armstrong, A. Larry, Spino, D. Golden Teresa, Efficient detection and evaluation of cyclodextrin multiple complex formation, J. Am. Chem. Soc. 108 (7) (1986) 1418–1421.
[23] K.A. Connors, The stability of cyclodextrin complexes in solution, Chem. Rev. 97 (5) (1997) 1325–1357.
[24] H. Dodziuk, Cyclodextrins and their Complexes: Chemistry, Analytical Methods, Applications, John Wiley & Sons, 2006.
[25] X.M. Qiu et al., A study on complexation between α-cyclodextrin and bis-quaternary ammonium surfactants, Acta Phys. Chim. Sin. 21 (12) (2005) 1415–1418.
[26] T. Tominaga, D. Hachisu, M. Kamado, Interactions between the Tetradecyltrimethylammonium ion and α-, β-, and γ-cyclodextrin in water as studied by a surfactant-selective electrode, Langmuir 10 (12) (1994) 4676–4680.
[27] R. De Lisi, S. Milioto, N. Muratore, Thermodynamic evidence of cyclodextrin-micelle interactions, J. Phys. Chem. B 106 (35) (2002) 8944–8953.
[28] T. Akita, K. Yoshikiyo, T. Yamamoto, Formation of 1:1 and 2:1 host–guest inclusion complexes of α-cyclodextrin with cycloalkanols: A 1H and 13C NMR spectroscopic study, J. Mol. Struct. 1074 (2014) 43–50.
[29] H. Ohtsuki et al., 13C NMR spectroscopy on the complexation of α-cyclodextrin with 1-alkanols and 1-alkanoate Ions, J. Incl. Phenom. Macrocycl. Chem. 50 (1–2) (2004) 25–30.
[30] N. Funasaki, S. Ishikawa, S. Neya, Proton NMR study of α-cyclodextrin inclusion of short-chain surfactants, J. Phys. Chem. B 107 (37) (2003) 10094–10099.
[31] E. Saint Aman, D. Serve, A conductimetric study of the association between cyclodextrins and surfactants—application to the electrochemical study of a mixed aqueous system: Substrate, cyclodextrin, surfactant, J. Colloid Interface Sci. 138 (2) (1990) 365–375.
[32] N. Funasaki, S. Ishikawa, S. Neya, 1:1 and 1:2 complexes between long-chain surfactant and α-cyclodextrin studied by NMR, J. Phys. Chem. B 108 (28) (2004) 9593–9598.
[33] I. Sanemasa, T. Takuma, T. Deguchi, Association of some polynuclear aromatic-hydrocarbons with cyclodextrins in aqueous-medium, Bull. Chem. Soc. Jpn. 62 (10) (1989) 3098–3102.
[34] X.B. Wang, Mark, Solubilization of some low-polarity organic compounds by hydroxy propy I-α-cyclodextrin, Environ. Sci. Technol. 27 (1993) 2821–2825.
[35] S. Tanada et al., Removal of aromatic hydrocarbon compounds by hydroxypropyl-cyclodextrin, J. Colloid Interface Sci. 217 (2) (1999) 417–419.
[36] S.-J. Kim, J.-H. Kwon, Determination of partition coefficients for selected PAHs between water and dissolved organic matter, Clean-Soil Air Water 38 (9) (2010) 797–802.
[37] US EPA, Estimation Programs Interface Suite™ for Microsoft® Windows, W. United States Environmental Protection Agency, DC, USA, Editor; 2012.
[38] W.J. Blanford et al., Solubility enhancement and QSPR correlations for polycyclic aromatic hydrocarbons complexation with α, β, and γ cyclodextrins, J. Incl. Phenom. Macrocycl. Chem. 78 (1–4) (2014) 415–427.
[39] Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.3r1; 2010.
[40] P.D. Ross, M.V. Rekharsky, Thermodynamics of hydrogen bond and hydrophobic interactions in cyclodextrin complexes, Biophys. J. 71 (4) (1996) 2144–2154.

[41] D. Hallen et al., Microcalorimetric titration of α-cyclodextrin with some straight-chain alkan-1-ols at 288.15, 298.15 and 308.15 K, J. Chem. Soc., Faraday Trans. 88 (19) (1992) 2859–2863.
[42] S. Hamai, 1:1:1 Inclusion compounds of β-cyclodextrin with fluorene and alcohols or nitriles in aqueous solution, Bull. Chem. Soc. Jpn. 62 (9) (1989) 2763–2767.
[43] R. Carpignano et al., QSAR study of inclusion complexes of heterocyclic compounds with β-cyclodextrin, Anal. Chim. Acta 348 (1–3) (1997) 489–493.
[44] TURBOMOLE V6.5 2013, in a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH: since 2007.

2.2 3D-QSAR Predictions for α-Cyclodextrin Binding Constants Using Quantum Mechanically Based Descriptors

Chemosphere 169 (2017) 693–699

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

3D-QSAR predictions for α-cyclodextrin binding constants using quantum mechanically based descriptors

Lukas Linden [a], Kai-Uwe Goss [a, b], Satoshi Endo [a, c, *]

[a] Helmholtz Centre for Environmental Research UFZ, Permoserstr. 15, D-04318 Leipzig, Germany
[b] University of Halle-Wittenberg, Institute of Chemistry, Kurt Mothes Str. 2, D-06120 Halle, Germany
[c] Osaka City University, Urban Research Plaza & Graduate School of Engineering, Sugimoto 3-3-138, Sumiyoshi-ku, 558-8585 Osaka, Japan

HIGHLIGHTS

- Successful prediction of α-cyclodextrin binding constants with a new 3D-QSAR model.
- The 3D-QSAR model uses local sigma profiles that emerge from the COSMO-RS theory.
- Comparison of the new 3D-QSAR model with two standard models.
- Accurate predictions of steric effects that influence the binding by the new model.
- Validation of the modeling approaches with a literature data set.

ARTICLE INFO

Article history:
Received 6 October 2016
Received in revised form
18 November 2016
Accepted 21 November 2016

Handling Editor: I. Cousins

Keywords:
α-Cyclodextrin (CD)
Binding constant
Inclusion complex
Prediction

ABSTRACT

Binding of organic chemicals to α-cyclodextrin (αCD) is a typical example for host-guest complexation that is influenced by the 3D-structure of both the binding site (host) and the solute (guest). Prediction of the binding constant is challenging and requires a successful representation of the binding site-solute interactions in the 3D-space. In this study, we tested if a 3D quantitative structure activity relationship (3D-QSAR) model with quantum mechanically based local sigma profiles (LSPs) derived from the COSMOsar3D method is capable of predicting αCD binding constants from the most recent literature and how the model performs in comparison to a standard comparative molecular field analysis and to a reference 2D-QSAR. The results showed that the new 3D-QSAR model was more predictive than both reference models (RMSE 0.45 vs 0.53/0.52, R^2 0.70 vs 0.53/0.68). Furthermore, only the new model captured the differences in the binding constants between structural isomers of aliphatic alcohols and allowed an extrapolation of the prediction to another literature data set. The high performance of the 3D-QSAR model with LSPs tested in this study and its theoretical robustness suggest that this modeling approach should be applicable to other binding processes including protein binding.

1. Introduction

Binding of organic chemicals to macromolecules is of high relevance in environmental science and related fields. For example, binding to macromolecular sorbents such as cyclodextrins (CDs) can be utilized for remediation of contaminated materials. Moreover, binding to proteins including binding proteins, enzymes, transporters, and receptors has strong impacts on toxicity of chemicals. Prediction of binding coefficients poses a major challenge, as the three-dimensional (3D) structure of both the solute and the binding site strongly influences the binding free energy, thus the binding constant (Herrmann, 2014). This is in contrast to the partition coefficients between liquids, for which the free energy is sufficiently well predicted by descriptors that characterize the interaction properties of the whole molecule without considering the molecular geometry (Karickhoff et al., 1991; Klamt, 1995;

Abbreviations: αCD, α-cyclodextrin; QSAR, quantitative structure activity relationship; LSPs, local sigma profiles; COSMO, conductor-like screening model; CoMFA, comparative molecular field analysis; MIFs, molecular interaction fields; PLS, partial least square; COSMO-RS, conductor-like screening model for real solvents; LSER, linear solvation energy relationship; pp-LFER, poly-parameter linear free energy relationship; O3A, Open3DALIGN; MDsim, molecular dynamics simulation; ele, electrostatic.

* Corresponding author. Osaka City University, Urban Research Plaza & Graduate School of Engineering, Sugimoto 3-3-138, Sumiyoshi-ku, 558-8585 Osaka, Japan.
E-mail address: satoshi.endo@urban.eng.osaka-cu.ac.jp (S. Endo).

http://dx.doi.org/10.1016/j.chemosphere.2016.11.115
0045-6535/

Abraham et al., 2004; Endo and Goss, 2014).

3D quantitative structure activity relationships (3D-QSARs) attempt to establish a correlation between a macroscopic property (e.g., binding constant, receptor affinity) and 3D-structural features of the solute molecules. A widely used 3D-QSAR tool is comparative molecular field analysis (CoMFA) (Cramer et al., 1988). CoMFA uses 3D-discretized molecular field properties, called molecular interaction fields (MIFs), as descriptors for a statistical method (e.g., partial least square, PLS). Recently, Klamt et al. proposed the COSMOsar3D method (Klamt et al., 2012), which uses 3D-gridded COSMO surface polarization charge densities as a new set of MIFs. This extension of CoMFA emerges from the quantum mechanically-based COSMO-RS (conductor-like screening model for real solvents) method (Klamt, 1995; Klamt et al., 1998), which predicts the properties of a chemical by using the surface polarization charge densities (called sigma surface) of the molecule calculated quantum mechanically in a virtual conductor. For each molecule, the calculated sigma surface can be condensed into a sigma profile, a histogram of all the 'partial' charges (or charge-patches) of the molecule. The sigma surface and the sigma profile of a chemical appear to accurately describe the abilities of the molecule to undergo intermolecular interactions including electrostatic, hydrogen-bond, and van der Waals interactions (Klamt, 2011). To extend this concept to 3D-QSARs, COSMOsar3D computes the sigma profiles at grid points within the 3D space to give the local sigma profiles (LSPs) (Thormann et al., 2012). The LSP is thus a histogram that contains information about the sigma surface of a specific part of the molecule. Considering the theoretical basis and the proven accuracy of COSMO-RS for partitioning between liquids, it is anticipated that the LSPs are ideal MIFs for 3D-QSAR modeling of the binding free energy that is strongly influenced by the molecular geometry of solutes. Nevertheless, the COSMOsar3D method has only been tested against standard sets of enzymatic inhibition activities by the developers and there has been no attempt to apply this method to equilibrium binding constants.

In this study, COSMOsar3D is used to model data sets of α-cyclodextrin (αCD) binding constants. αCD is built of six 1-4-linked glucopyranose units that form a conic ring with a diameter of 5 Å. In water, all hydroxyl groups are positioned on the outside of the αCD ring, resulting in a hydrophobic cavity inside (Cox et al., 1984), which enables αCD to form host-guest complexes. Formation of CD complexes (Connors, 1997) can improve the solubility of chemicals (Hedges, 1998), clean waste gas streams (Blach et al., 2008), remediate contaminated soils (Villaverde et al., 2005; Flaherty et al., 2013), and mask taste and odor compounds (Del Valle, 2004). Further, CDs can be used to enhance the bioavailability of organic pollutants (Liu et al., 2013), remove them from aqueous media (Sawicki and Mercier, 2006), and extract dyes from sand (De Lisi et al., 2007). CDs are also considered a useful test material for investigating macromolecular binding because of their relatively simple and well-studied structure as well as evidences of substantial molecular steric effects on the binding constants (Tabushi, 1982; Ishiwata and Kamiya, 1999; Schneider, 2009). In the common cyclodextrin family (i.e., α-, β-, γ-), αCD may be the most suitable starting material for studying 3D-effects on binding, as it has the smallest cavity and thus the highest restriction for host-guest complexation.

The purpose of this study is to evaluate the LSP-based 3D-QSAR (i.e., COSMOsar3D) for predicting αCD binding constants in comparison to a standard CoMFA model that uses steric and electrostatic fields as MIFs. In addition, these 3D-QSARs are compared to a well-established 2D-QSAR, namely the linear solvation energy relationship (LSER), which is a polyparameter linear free energy relationship (pp-LFER) model using Abraham's descriptors (Abraham et al., 1994; Goss, 2005). Since the LSER does not explicitly include descriptors that describe molecular geometry, this comparison serves to evaluate whether taking into account the molecular 3D geometry improves the accuracy of predictions for αCD binding constants.

2. Methods

2.1. Data sets

Two data sets of 1:1 αCD binding constants (K_{a1}) [M^{-1}] were considered in this study. The first has been measured in our laboratory under a consistent experimental condition, as reported previously (Linden et al., 2016). This data set, referred to as the "Linden data set", was used for the calibration and the first evaluation of the modeling approaches, because we consider these data of high quality and consistency. The second data set was from Suzuki (2001), who assembled literature data for αCD binding constants. The Suzuki data set was used for an additional external validation of the modeling approaches.

The Linden data set (Linden et al., 2016) consists of 60 neutral aliphatic and aromatic chemicals (range of log K_{a1}: 1.25–4.97, mean: 2.42, standard deviation (SD): 0.83). It contains several groups of isomers, e.g., 1-hexanol (i.e., end-substituted alcohol) and 3-hexanol (i.e., middle-substituted alcohol) as well as homologous series of chemicals (e.g., alcohols, ketones, ether, chlorobenzenes). The Suzuki data set (Suzuki, 2001) includes 87 neutral aliphatic and aromatic chemicals (range of log K_{a1}: −0.09–3.81, mean: 1.95, SD: 0.81). Ionic or partly ionic chemicals were not considered here to avoid uncertainty associated with the actual charge state of the bound molecule (i.e., ionic or neutral) and different descriptions of ionic molecules between MIFs. The chemicals and the respective log K_{a1} values are listed in Table SI 1 and Table SI 2. Five alcohols, namely 1-butanol, 1-pentanol, 1-heptanol, 1-hexanol, and 1-octanol exist in both data sets. Their reported log K_{a1} values are 0.28–0.51 log units higher in the Suzuki data set than in the Linden data set. The difference in log K_{a1} might be, in part, caused by the different experimental temperatures (Suzuki data 25 °C, Linden data 30 °C). Linden data were measured at 30 °C which was the lowest adjustable temperature in the experimental setting. This minor difference in temperature should be borne in mind when the results are evaluated (see below).

2.2. Selection procedures for training and test sets

For generation and evaluation of each model (i.e., 2D- and 3D-QSARs), the Linden data set was split into training and test sets. The training set was used for model calibration and selection, while the performance of the resulting model was validated with regard to the prediction of the test set. Prediction of data that were not part of the training set is essential as a control and should be considered the more important quality feature for 3D-QSARs (Gramatica, 2007).

For the general model evaluation, the training and test sets were generated with the log K_{a1} hierarchic bin system (Kauffman and Jurs, 2001) (procedure 1, see Fig. SI 3 for a scheme). In this system, the data set was sorted according to the log K_{a1} values of the chemicals and then, from highest to lowest, four consecutive chemicals were placed in one bin. One chemical from each bin was selected randomly and placed in the test set. This classifies 25% chemicals of the data set to the test set. The rest of the chemicals formed the training set. The procedure was repeated five times, resulting in five random training sets and the corresponding test sets.

In order to evaluate varying steric effects within homologous series of chemicals and isomers, the following modified procedure

was used to generate constructed test sets (procedure 2). As in the first procedure, the chemicals were sorted by log K_{a1} and four chemicals in a row were grouped into one bin. Then, the numbers 1 to 4 were given randomly to the four chemicals of a bin. In the first run of chemical selection, the chemicals with the number 1 embodied the test set, while the rest of the chemicals were used as the training set. In the second run, the chemicals with the number 2 were the test set, and so forth. This procedure resulted in four test and training set combinations. In comparison to procedure 1, the randomness of the selection is reduced, whereas each chemical is part of a test set once and the other three times it belonged to the training set.

2.3. 3D-QSARs

The 3D-QSAR modeling followed the workflow shown in Fig. SI 1. Modeling generally takes the following steps: 3D-structure generation, alignment, MIFs generation, model calibration with PLS, and model evaluation using the test set. There are multiple options for each step, as explained below, and different combinations were tested in this work for comprehensive evaluation of the methods.

2.3.1. 3D structure generation

The 3D structures of all chemicals were generated with Tinker or COSMO*conf*X13. Tinker (Marinescu and Bols, 2009) is a molecular modeling package implemented in Open3DALIGN v. 2.3 (O3A) (Tosco et al., 2011) and generates the structure-data files of the conformers for the O3A alignment. The quenched molecular dynamics conformational search of Tinker was performed with an implicit solvent calculation and a dielectric constant of 24, which is the dielectric constant of βCD (Yu et al., 2002), while for the rest of the parameters the default setting was chosen.

COSMO*conf*X13 is a tool box that uses Turbomole (Sijm et al., 2000) for the quantum mechanics calculations of COSMO files. The default COSMO*conf* procedure was modified so that it creates more conformers than usual (see SI). That is to say, the total number of possible conformers was increased, the energetic distance between conformers was reduced, and the clustering steps were loosened. These modifications were intended to account for the flexibility of the chemicals, which is more important for the αCD binding than for bulk phase partitioning.

2.3.2. Alignments

The 3D structures of chemicals need to be aligned in the 3D space before performing statistical analysis. Ideally, the resulting position and orientation of a chemical in the 3D space corresponds to the optimal interaction possibility between the chemical and αCD. In a target-based approach, the structure or a substructure of αCD is used as the template to which all molecules are aligned. In a ligand-based approach, the template is generated with the help of chemicals that bind strongly to αCD (i.e., with high log K_{a1} values). For all approaches, up to ten conformers of each chemical were considered and the conformer with the highest alignment score and, if there are multiple conformers with the highest score, then that with the lowest energy was chosen for the model. In this study, the following three alignment procedures were applied.

1. The O3A alignment maximizes the overlap of atoms of the template chemicals and of the remaining chemicals. This is a ligand-based method and a standard alignment for CoMFA approaches and was performed here by using O3A v. 2.3 (Tosco et al., 2011). The seven chemicals with the largest log K_{a1} values of the Linden data set, namely 1-dodecanol, 1-undecanol, 1-decanol, 1-nonanol, 2-undecanone, 2-decanone, and hexylbenzene were used as template chemicals. These chemicals were pre-aligned against each other and then each conformer of the remaining chemicals was aligned against the pre-aligned conformers of each template chemical. In the end, the position of the chemical/conformer with the highest score against any of the template chemicals was chosen.
2. The COSMOsim3D alignment (Thormann et al., 2012) maximizes the overlap between the sigma surfaces of the chemical and the template. Hereby, the template is an averaged sigma profile of the template chemicals. The template chemicals used were the same as in the previous alignment method.
3. The COSMOsim3D receptor alignment is a target-based approach that maximizes the overlap between the inverted sigma surface of αCD (which is the sigma charge value of each surface patch multiplied with −1) and the sigma surface of the chemicals of the data set. The sigma surface of αCD needs to be inverted because the alignment algorithm maximizes the overlap of like sigma charges in a ligand-based approach. The inversion therefore places the chemicals in a position where greatest interaction energies between both αCD and the respective chemical occur, as the interaction energy is greatest when the difference between the sigma charges of two interacting surface segments is maximal. This alignment already considers the steric restrictions of the αCD cavity because the chemicals cannot be placed at the same position as the αCD. The input structure for the COSMOsim3D receptor alignment is the 3D structure of αCD and the position of an exemplary ligand, the latter defines the starting position in the alignment procedure for all chemicals that need to be aligned. Two input structures were used in our approach to test the dependence of the COSMOsim3D receptor alignment on the input structure:
   (3a) The 3D structure of αCD and the position of a ligand (poly-*p*-phenylene rotaxane) were obtained from an X-ray measurement (Stanier et al., 2001) (three different views of the complex are shown in Fig. SI 4). The cosmo file of the αCD structure was derived with a single point calculation using COSMO*confX13*.
   (3b) The 3D structure of αCD and the position of a ligand (1-dodecanol) were estimated by a molecular dynamics simulation (MDsim), which was kindly provided by Sven Jakobtorweihen at Hamburg University of Technology. The complex with the smallest distance between the center of mass of αCD and that of 1-dodecanol was chosen as the template for the alignment (Fig. SI 5). The cosmo file for the resulting αCD structure was derived with a single point calculation using COSMO*confX13*.

2.3.3. MIFs

Two sets of MIFs were used as independent variables for the PLS regression analysis.

1. The van der Waals (vdW) and the electrostatic (ele) fields are the two standard CoMFA variables. Molecular mechanics calculations using the Merck force field (MMFF94) were performed with Open3DQSAR v. 2.3 (Tosco and Balle, 2011) to derive the vdW and ele fields. A sp^3 carbon atom was used as the probe. A grid spacing of 1 Å was used with a 5 Å gap, i.e., the minimal distance to the box, around the chemicals.
2. LSPs were derived from the cosmo files by COSMOsar3D (Klamt et al., 2012). For the 3D-QSAR model used here the LSPs were split into several consecutive profiles, each covering a range of 0.006 e/Å^2. Thus, MIFs 1, 2, …, and 7 cover sigma values from −0.024 to −0.018 e/Å^2, -0.018 to −0.012 e/Å^2, …, and, 0.012 to 0.018 e/Å^2, respectively (Fig. SI 2). In the end, the integral of

each LSP serves as the value for the independent variable. A grid spacing of 2 Å was used in a box that leaves at least a 5 Å gap around the chemicals.

2.3.4. Statistical tool

The independent variables, i.e., the MIFs, of the training set chemicals were correlated with the log K_{a1} values using PLS regression analysis. Prior to PLS regression analysis, the number of independent variables was reduced as following. An energy cutoff was set at ± 30 kcal/mol (Kim, 1995), and variables that have a SD below a level of 0.1 among all training chemicals were excluded. The different MIFs were scaled before the PLS procedure using block unscaled weighting (Kastenholz et al., 2000). Moreover, fractional factorial design selection (Baroni et al., 1992, 1993) was used to reduce the number of variables.

PLS analysis was performed to derive one to five PLS components. Thus, each run resulted in five different models that used one to five PLS components. Leave-two-out cross validation was performed with each model and then the model with the minimum of the root mean square error (RMSE) value was selected for further evaluation against the test set.

2.3.5. pp-LFER

The pp-LFER is among the most accurate and robust models to describe solute partitioning between liquids or liquid and gas phases, where molecular interactions are not sterically restricted. In a practical sense, a 3D-QSAR model may be considered meaningful only if it gives better predictions than the pp-LFER model, which is simple and quick as long as the solute descriptors are known. The pp-LFER used here appears,

$$\log K_{a1} = c + sS + aA + bB + vV + lL \quad (1)$$

where S is the polarizability/dipolarity parameter, A the solute H-bond acidity, B the solute H-bond basicity, V the McGowan characteristic volume (cm^3 $mol^{-1}/100$) and L the logarithm of the hexadecane-air partitioning coefficient. In this work, the pp-LFER solute descriptors (capital letters in Eq. (2)) were obtained from the UFZ-LSER database (Endo et al., 2015) and the system parameters (lower case letters in Eq. (2)) were fitted with multiple linear regression analysis using the experimental data for log K_{a1} of training chemicals.

3. Results & discussion

Table 1 shows the statistical results for evaluation of the modeling approaches using the Linden data set. RMSE and R^2 calculated with the test sets are considered more important evaluation criteria than q^2. Each value in the table represents the mean (± standard deviation) of five runs with five different training and test sets generated by test set selection procedure 1. In the following, the results of the pp-LFER approach are discussed first and then the results of the 3D-QSAR approach.

3.1. pp-LFER

First, the pp-LFER equation (Eq. (2)) was fitted to all experimental αCD binding constants of the Linden data set (i.e., no test and training set selection) to have an idea to what extent the 2D model can describe the whole data set (Fig. SI 4). This fit resulted in the equation

$$\log K_{a1} = -0.32(\pm 0.44) + 2.04(\pm 0.63)S + 3.15(\pm 0.63)A - 3.01(\pm 0.50)B + 6.01(\pm 0.88)V - 1.10(\pm 0.21)L \quad (2)$$

The fit of the pp-LFER equation usually results in a standard deviation of 0.1–0.2 log units for homogeneous solvent-water partition systems, which are not influenced by steric effects, and a larger standard deviation for partitioning or binding to heterogeneous materials such as serum albumin and natural organic matter (Bronner and Goss, 2011; Endo and Goss, 2011). The RMSE for the binding to αCD (Fig. SI 4) is 0.48, being comparable to fits for other heterogeneous materials (Bronner and Goss, 2011).

The pp-LFER fits for training sets extracted from the Linden data set resulted in system parameters similar to those for the complete Linden data set (Table SI 3). The predictions for the corresponding test sets (Table 1, M1) were surprisingly accurate (RMSE = 0.52 ± 0.05 and R^2 = 0.68 ± 0.07). This result was unexpected because the experimental results do suggest strong steric effects, whereas the pp-LFER model does not capture such effects (Linden et al., 2016). A closer examination of the results revealed that systematic prediction errors do exist for binding constants, e.g., log K_{a1} values for end-substituted chemicals were systematically underestimated and those for middle-substituted chemicals were overestimated, which is an indication that the pp-LFER model is not able to cover the underlying steric effects. In addition, chemicals that are not expected to fit into the αCD cavity due to the steric hindrance were over-predicted by the pp-LFER, e.g., the log K_{a1} value of 1-chloronaphthalene is predicted as 2.13, while the experiment showed that it is < 1.3 (Linden et al., 2016).

3.2. 3D-QSARs

Seven 3D-QSAR model variants were constructed using different combinations of structure generation, alignment, and MIF methods and evaluated with the Linden data set, as explained in the method

Table 1
Comparison of the statistical results of the different modeling approaches for the prediction of log K_{a1} of the Linden data set using test set selection procedure 1. The modeling approach written in bold performed best of all investigated modeling variants.

Modeling approach	Method	Alignment	Field	q^2 ± SD	RMSE ±SD	R^2 ± SD
M1	pp-LFER				0.52 ± 0.05	0.68 ± 0.07
M2	3D-QSAR	O3A	LSP	0.63 ± 0.03	0.54 ± 0.08	0.56 ± 0.17
M3	3D-QSAR	O3A	vdW ele	0.58 ± 0.08	0.53 ± 0.11	0.53 ± 0.11
M4	**3D-QSAR**	**COSMOsim3D**	**LSP**	**0.83 ± 0.02**	**0.45 ± 0.06**	**0.70 ± 0.08**
M5	3D-QSAR	COSMOsim3D	vdW ele	0.70 ± 0.01	0.56 ± 0.06	0.53 ± 0.12
M6a	3D-QSAR	COSMOsim3D receptor X-ray	LSP	0.66 ± 0.06	0.51 ± 0.06	0.61 ± 0.09
M6b	3D-QSAR	COSMOsim3D receptor MDsim	LSP	0.71 ± 0.04	0.49 ± 0.04	0.64 ± 0.07
M7	3D-QSAR	COSMOsim3D receptor X-ray	vdW ele	0.51 ± 0.08	0.55 ± 0.08	0.56 ± 0.13

O3A means open3DALIGN, q^2 is the coefficient of determination for the leave-two-out cross validation using the training set, RMSE is the root mean square error of the test set in log units, and R^2 is the coefficient of determination of the test set. LSP, vdW, and ele indicate the usage of local sigma profiles, van der Waals interaction field, and electrostatic interaction field as molecular interaction field, respectively, SD is standard deviation, and MDsim is molecular dynamics simulation.

section (Fig. SI 1, Table 1). The results show the following trends: (i) RMSE and R^2 of the 3D-QSAR model variants for test set predictions were 0.45–0.56 and 0.53–0.70, respectively. While the best 3D-QSAR model (M4) performed slightly better than the pp-LFER, the statistics were similar on average. (ii) The models that used the LSPs (Klamt et al., 2012) as independent variables tended to result in better predictions than those using the vdW and ele MIFs for a given alignment (i.e., O3A, COSMOsim3D, or COSMOsim3d receptor). These outcomes suggest that LSPs are more suitable descriptors to describe the binding to αCD than the tested CoMFA variables. This interpretation is in line with the claim that LSPs are theoretically more relevant for linear regression models, like PLS, to describe the interaction energy (Klamt et al., 2012).

Of the 3D-QSARs tested, the model that uses the COSMOsim3D alignment with the LSP variables (M4, Table 1) was the best model variant (i.e., with the lowest RMSE). No improvement was observed for the use of the 3D-structure of αCD as the template for the alignment (compare M6a and M6b to M4). Moreover, no difference was observed between the use of the two αCD structures (M6a (X-Ray) vs. M6b (MDsim)) for the target-dependent alignment. The fact that no improvement was observed by the use of the target-dependent alignment suggests that the selected 7 template chemicals were sufficient for aligning the 60 chemicals in the Linden set. This result, however, may not be general; alignments with a binding site structure are expected to be advantageous particularly if the data availability is limited. Note that, in principle, MDsim could directly calculate binding coefficients (Gebhardt and Hansen, 2016; Sancho et al., 2016) but such calculations would be time consuming for a larger number of chemicals, although these calculations are more and more automated and routinely performed.

The possibility of a chance correlation for the best modeling approach (M4) was evaluated by scrambling of the dependent log K_{a1} values in two sorted bins (this means each chemical got a permuted log K_{a1} value) (Tropsha et al., 2003; Rücker et al., 2007), which resulted in non-predictive models ($R^2_{training}$ = 0.40, q^2_{LTO} = −0.0030, the mean of 10 times evaluation).

To infer binding mechanisms, the contributions of the MIFs (vdW and ele, or LSPs) to the PLS components are examined. The percentage contributions of the seven LSPs to the M4 PLS model are shown in Fig. SI 6. MIF 4 (−0.012 to 0 e/Å^2, Fig. SI 2) had the highest contribution to the PLS components. This is an indication for the importance of vdW interactions and the hydrophobic effect for the binding to αCD (Marques, 2010). The contribution of MIF 4 decreases slightly with increasing PLS component number, whereas the contributions of the other MIFs rather increased with increasing PLS component number. The PLS component 1 in this example already explained 70% of the variance in the log K_{a1} data, while the other four PLS components added up to an explained variance of 27%, i.e., the PLS components 2–5 serve for fine tuning of the model. The field contributions of model variants that used vdW and ele variables support the mechanistic interpretation obtained from the LSPs; the contribution of the vdW field is around 90% for the models.

3.2.1. Predictions of specific molecular steric effects

To evaluate the performance of the 3D-QSAR modeling approaches for predicting particular types of chemicals, four training and test sets were generated from the Linden data set according to test set selection procedure 2 (see the method section) and all prediction procedures were redone. Model approaches M3, M4, M5, and M6b were evaluated here because they performed best in the random evaluation above and allow comparison of the classical CoMFA approach and the new COSMO-based approach. The resulting statistics (i.e., q2, RMSE, R2) were similar to those obtained above with test set selection procedure 1 (Table 1), except for M3, for which the test set selection procedure 2 resulted in worse predictions (see Table SI 5). Fig. 1 compares the experimental data and the predictions by the best model variant (M4, with COSMOsim3D + LSPs) for individual chemicals.

Many trends of the data that are related to steric effects were quantitatively described in the best 3D-QSAR model variant we found (M4). For example: experimental data show relatively large differences in log K_{a1} between isomeric chemicals with the functional group at the terminal and the middle positions such as 1-heptanol and 4-heptanol. These chemicals are predicted successfully by M4, e.g., 1-heptanol (log K_{a1} exper. 3.08, pred. 2.75) and 4-heptanol (log K_{a1} exper. 2.16, pred. 2.36). Also, as is the case in the experimental data, elongation of the alkyl chain in only one direction resulted in a higher increase of log K_{a1} than elongation in two or more directions (Fig. 2). The 3D-QSAR model variants M3, M5, and M6b were not able to describe the differences between these alcohols as well as M4 (Fig. 2). The comparison between M4 and M5 shows that the use of LSPs instead of vdW and ele not only minimizes the overall prediction errors but helps distinguish structural isomers of alcohols. The standard CoMFA model (M3) underestimates most of these alcohols and is not able to capture the steric effects. M6b uses LSPs as variables, but it appears that the target-based alignment cannot as accurately reproduce the trend of alcohol data as the ligand-based alignment in this case.

Experimental data for chlorobenzenes showed a distinct substitution effect on the αCD binding constant. K_{a1} increases with chlorine substitution up to two chlorine atoms, whereas a further substitution decreases K_{a1}, which can be explained by the size limitation of the cavity. This effect is not well described by any 3D-QSAR model tested here. For example, 1,2,4,5-tetrachlorobenzene and 1,3-dichlorobenzene showed a prediction error larger than 0.6 log units with the best model variant, M4. The use of the αCD target structure (COSMOsim3D receptor alignment, M6b), the CoMFA variables vdW and ele (M5), and the standard CoMFA model (M3) did not improve the prediction of chlorobenzenes. A reason for the inaccurate predictions for chlorobenzenes could be the

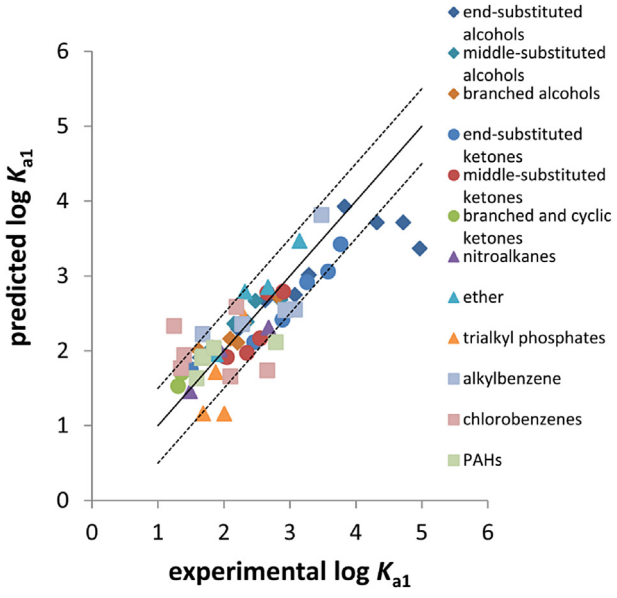


Fig. 1. Prediction of log K_{a1} of 60 Linden's chemicals with COSMOsim3D alignment and local sigma profiles as variables (M4). Test sets were selected with test set selection procedure 2. The solid line indicates the 1:1 line and the dashed lines indicate a deviation of 0.5 log units from the 1:1 line.

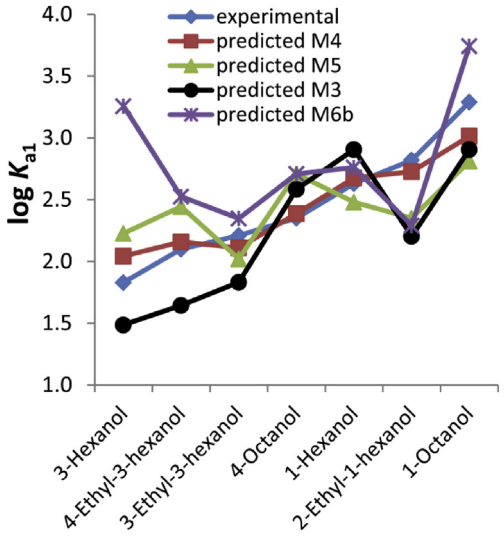


Fig. 2. Experimental and predicted log K_{a1} for αCD binding of two C_6-alcohols and five C_8-alcohols.

small number of calibration data that showed strong effects of steric restrictions. As shown in the previous work (Linden et al., 2016), K_{a1} for chemicals that undergo strong steric restrictions tend to have K_{a1} values that are too low to measure and thus such chemicals cannot be included in the data set for model calibration.

The end-substituted chemical 1-dodecanol was the biggest outlier in all predictions. A reason could be that 1-dodecanol has the longest alkyl chain and the largest K_{a1} in the data set. Therefore, the positive interaction between the long alkyl chain and αCD may not be covered by the models. Additionally, the 3D-QSAR models in this work only consider one selected conformer of each chemical, which neglects the influence of different binding modes for predictions of flexible molecules like 1-dodecanol. Furthermore, a recent MDsim study showed that 1-dodecanol interacts substantially with the water surrounding αCD and that the explicit consideration of the water molecules is necessary for a successful prediction of long chain alcohols (Gebhardt and Hansen, 2016). Note that, while the data we considered are for 1:1 binding constants, 2:1 binding can become more important for chemicals with long alkyl chain(s).

3.3. Predictions of the Suzuki data set

For a further evaluation of each modeling approach, models were generated using all Linden data as the training set and evaluated with the Suzuki data as an external test set. The prediction of the Suzuki data by the pp-LFER calibrated with the Linden data (Table SI 6, M1) was substantially worse (RMSE = 1.09, $R^2 = 0.13$), as compared to the test set predictions of the Linden data set (Table 1, M1). This RMSE is even greater than the SD of the Suzuki data. It is notable that the pp-LFER, which does not include steric terms, does show promising statistics when evaluated with the Linden set alone (Table 1, M1), whereas the model calibrated with the Linden set does not extrapolate well to the external Suzuki set. We have tried the reversed evaluation (i.e., using the Suzuki set as the training set and the Linden set as the test set, Table SI 6) and obtained similar statistics but substantially different regression coefficients.

The 3D-QSAR models handled the external prediction better than the pp-LFER model, but RMSE values for the predictions of the Suzuki data set (Table SI 6, M2-M7) were 0.13–0.19 log units higher than the test set predictions for the Linden data set. The model variant that uses the COSMOsim3D alignment and LSPs (Table SI 6, M4) achieved an RMSE of 0.59 and an R^2 of 0.61, while all other models had RMSE > 0.68 and $R^2 < 0.5$. For a given alignment, LSPs resulted in better or equivalent statistics as compared to vdW and ele. These results are in line with the findings we obtained from the model evaluation with the Linden data set only. Note that systematic under-predictions for the Suzuki data were not found; thus, the temperature difference is not a significant reason for the increased RMSE. We obtained similar statistics for the reversed evaluation (i.e., using the Suzuki set as training set and the Linden set as test set, Table SI 6). We also found that, if both Linden and Suzuki sets are combined and split to training and test sets, statistics for the test set prediction improves (RMSE, R^2), which suggests that there are significant differences in the chemical domains that are covered by the two data sets (Table SI 6). As an example, the Suzuki data set includes phenols and phenyl acetates, which are chemical classes not included in the Linden data set. On the other hand, only the Linden data set includes ethers and ketones. Moreover, the Suzuki data set is predominated by aromatic chemicals while the proportion of aromatic and aliphatic chemicals is comparable in the Linden data set.

We further tested if the steric restriction through the cavity can correctly be described by the model variant M4. The binding coefficients were predicted for the ten chemicals for which we were able to determine only the upper limit of log K_{a1} (<1.3) in the previous work (Linden et al., 2016). These chemicals are most likely too large to fit into the αCD cavity. Eight of the ten chemicals had predicted log K_{a1} values of 1.3 ± 0.4, which is in a semi-quantitative agreement with our experiments. Log K_{a1} values for 1-chloronaphthalene (predicted log K_{a1} 2.67) and acenaphthene (predicted log K_{a1} 2.42) were overestimated by > 1 log unit. In contrast, the prediction of a similar chemical, acenaphthylene resulted in a predicted log K_{a1} of 1.7. The COSMOsim3D alignment placed acenaphthene and acenaphthylene in different positions, which likely explains the deviation in the predictions.

4. Conclusions

A 3D-QSAR model with COSMOsim3D (Thormann et al., 2012) for alignment and LSPs for independent variables in PLS regression analysis was capable of predicting αCD binding constants for organic chemicals with an RMSE of 0.45 log units. This model can be used for the prediction of unknown αCD binding constants for neutral organic chemicals and covers the most important steric effects that influence the binding to αCD (Linden et al., 2016). As assumed, the description of the binding to αCD needs to include the 3D-structure of the solutes because the 3D-QSAR model worked much better than the simple correlation with log K_{OW} (Linden et al., 2016) and better than the 2D-QSAR model (pp-LFER) considered here. Hence, it can be concluded that the LSPs are more suitable variables for 3D-QSAR modeling of the binding process to αCD and probably for other binding processes as well, e.g., binding to other types of cyclodextrin with a different application range. Use of 7 out of 60 chemicals as templates for the alignment appeared to be sufficient, also with regard to the prediction for 84 external data (Suzuki, 2001). Consequently, the combination of COSMOsim3D and COSMOsar3D may be applicable to similar binding systems with an unknown or flexible target-structure, as far as data for some strongly binding chemicals are available. In an upcoming study, we will apply the 3D-QSAR modeling approaches tested in this study to model the binding to serum albumin, which also showed specific 3D effects.

Acknowledgements

The authors thank the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE) for financial support and Sven Jakobtorweihen at Hamburg University of Technology for providing the molecular dynamics simulations. SE acknowledges the financial support from the MEXT/JST Tenure Track Promotion Program. The authors thank Nadin Ulrich for helpful comments on an early version of the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.chemosphere.2016.11.115.

References

Abraham, M.H., Andonian-Haftvan, J., Whiting, G.S., Leo, A., Taft, R.S., 1994. Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. J. Chem. Soc. Perkin Trans. 2, 1777–1791.

Abraham, M.H., Ibrahim, A., Zissimos, A.M., 2004. Determination of sets of solute descriptors from chromatographic measurements. J. Chromatogr. A 1037, 29–47.

Baroni, M., Clementi, S., Cruciani, G., Costantino, G., Riganelli, D., Oberrauch, E., 1992. Predictive ability of regression models. Part II: selection of the best predictive PLS model. J. Chemom. 6, 347–356.

Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R., Clementi, S., 1993. Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. Quant. Struct. Act. Relat. 12, 9–20.

Blach, P., Fourmentin, S., Landy, D., Cazier, F., Surpateanu, G., 2008. Cyclodextrins: a new efficient absorbent to treat waste gas streams. Chemosphere 70, 374–380.

Bronner, G., Goss, K.-U., 2011. Predicting sorption of pesticides and other multifunctional organic chemicals to soil organic carbon. Environ. Sci. Technol. 45, 1313–1319.

Connors, K.A., 1997. The stability of cyclodextrin complexes in solution. Chem. Rev. 97, 1325–1357.

Cox, G.S., Turro, N.J., Yang, N.C.C., Chen, M.J., 1984. Intramolecular exciplex emission from aqueous β-cyclodextrin solutions. J. Am. Chem. Soc. 106, 422–424.

Cramer, R.D., Patterson, D.E., Bunce, J.D., 1988. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. 110, 5959–5967.

De Lisi, R., Lazzara, G., Milioto, S., Muratore, N., 2007. Adsorption of a dye on clay and sand. Use of cyclodextrins as solubility-enhancement agents. Chemosphere 69, 1703–1712.

Del Valle, E.M., 2004. Cyclodextrins and their uses: a review. Process Biochem. 39, 1033–1046.

Endo, S., Goss, K.U., 2011. Serum albumin binding of structurally diverse neutral organic compounds: data and models. Chem. Res. Toxicol. 24, 2293–2301.

Endo, S., Goss, K.U., 2014. Applications of polyparameter linear free energy relationships in environmental chemistry. Environ. Sci. Technol. 48, 12477–12491.

Endo, S., Watanabe, N., Ulrich, N., Bronner, G., Goss, K.-U., 2015. UFZ-LSER Database V. 2.1. Helmholtz Centre for Environmental Research, Leipzig, Germany. www.ufz.de/lserd/.

Flaherty, R.J., Nshime, B., DeLaMarre, M., DeJong, S., Scott, P., Lantz, A.W., 2013. Cyclodextrins as complexation and extraction agents for pesticides from contaminated soil. Chemosphere 91, 912–920.

Gebhardt, J., Hansen, N., 2016. Calculation of binding affinities for linear alcohols to α-cyclodextrin by twin-system enveloping distribution sampling simulations. Fluid Phase Equilib. 422, 1–17.

Goss, K.-U., 2005. Predicting the equilibrium partitioning of organic compounds using just one linear solvation energy relationship (LSER). Fluid Phase Equilib. 233, 19–22.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701.

Hedges, A.R., 1998. Industrial applications of cyclodextrins. Chem. Rev. 98, 2035–2044.

Herrmann, A., 2014. Dynamic combinatorial/covalent chemistry: a tool to read, generate and modulate the bioactivity of compounds and compound mixtures. Chem. Soc. Rev. 43, 1899–1933.

Ishiwata, S., Kamiya, M., 1999. Cyclodextrin inclusion: catalytic effects on the degradation of organophosphorus pesticides in neutral aqueous solution. Chemosphere 39, 1595–1600.

Karickhoff, S.W., McDaniel, V.K., Melton, C., Vellino, A.N., Nute, D.E., Carreira, L.A., 1991. Predicting chemical reactivity by computer. Environ. Toxicol. Chem. 10, 1405–1416.

Kastenholz, M.A., Pastor, M., Cruciani, G., Haaksma, E.E.J., Fox, T., 2000. Grid/cpca: A new computational tool to design selective ligands. J. Med. Chem. 43, 3033–3044.

Kauffman, G.W., Jurs, P.C., 2001. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. J. Chem. Inf. Comput. Sci. 41, 1553–1560.

Kim, K.H., 1995. Comparative molecular field analysis (CoMFA). In: Dean, P.M. (Ed.), Molecular Similarity in Drug Design. Springer, Netherlands, Dordrecht, pp. 291–331.

Klamt, A., 1995. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. J. Phys. Chem. 99, 2224–2235.

Klamt, A., 2011. The COSMO and COSMO-RS solvation models. Wiley Interdiscip. Rev. Comput. Mol. Sci. 1, 699–709.

Klamt, A., Jonas, V., Bürger, T., Lohrenz, J.C.W., 1998. Refinement and parametrization of COSMO-RS. J. Phys. Chem. A 102, 5074–5085.

Klamt, A., Thormann, M., Wichmann, K., Tosco, P., 2012. COSMOsar3D: molecular field analysis based on local COSMO σ-profiles. J. Chem. Inf. Model. 52, 2157–2164.

Linden, L., Goss, K.-U., Endo, S., 2016. Exploring 3D structural influences of aliphatic and aromatic chemicals on α-cyclodextrin binding. J. Colloid Interface Sci. 468, 42–50.

Liu, H., Cai, X., Chen, J., 2013. Mathematical model for cyclodextrin alteration of bioavailability of organic pollutants. Environ. Sci. Technol. 47, 5835–5842.

Marinescu, L., Bols, M., 2009. Cyclodextrin derivatives that display enzyme catalysis. Trends Glycosci. Glycotechnol. 21, 309–323.

Marques, H.M.C., 2010. A review on cyclodextrin encapsulation of essential oils and volatiles. Flavour Fragr. J. 25, 313–326.

Rücker, C., Rücker, G., Meringer, M., 2007. Y-randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. 47, 2345–2357.

Sancho, M.I., Andujar, S.A., Porasso, R.D., Enriz, R.D., 2016. Theoretical and experimental study of inclusion complexes of β-cyclodextrins with chalcone and 2',4'-dihydroxychalcone. J. Phys. Chem. B 120, 3000–3011.

Sawicki, R., Mercier, L., 2006. Evaluation of mesoporous cyclodextrin-silica nanocomposites for the removal of pesticides from aqueous media. Environ. Sci. Technol. 40, 1978–1983.

Schneider, H.J., 2009. Binding mechanisms in supramolecular complexes. Angew. Chem. Int. Ed. 48, 3924–3977.

Sijm, D., Kraaij, R., Belfroid, A., 2000. Bioavailability in soil or sediment: exposure of different organisms and approaches to study it. Environ. Pollut. 108, 113.

Stanier, C.A., O'Connell, M.J., Clegg, W., Anderson, H.L., 2001. Synthesis of fluorescent stilbene and tolan rotaxanes by Suzuki coupling. Chem. Commun. 493–494.

Suzuki, T., 2001. A nonlinear group contribution method for predicting the free energies of inclusion complexation of organic molecules with α- and β-cyclodextrins. J. Chem. Inf. Comput. Sci. 41, 1266–1273.

Tabushi, I., 1982. Cyclodextrin catalysis as a model for enzyme action. Acc. Chem. Res. 15, 66–72.

Thormann, M., Klamt, A., Wichmann, K., 2012. COSMOsim3D: 3D-similarity and alignment based on COSMO polarization charge densities. J. Chem. Inf. Model. 52, 2149–2156.

Tosco, P., Balle, T., 2011. Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. J. Mol. Model. 17, 201–208.

Tosco, P., Balle, T., Shiri, F., 2011. Open3DALIGN: an open-source software aimed at unsupervised ligand alignment. J. Comput. Aided Mol. Des. 25, 777–783.

Tropsha, A., Gramatica, P., Gombar, V.K., 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22, 69–77.

Villaverde, J., Pérez-Martínez, J.I., Maqueda, C., Ginés, J.M., Morillo, E., 2005. Inclusion complexes of α- and γ-cyclodextrins and the herbicide norflurazon: I. Preparation and characterisation. II. Enhanced solubilisation and removal from soils. Chemosphere 60, 656–664.

Yu, J.S., Wei, F.D., Gao, W., Zhao, C.C., 2002. Thermodynamic study on the effects of beta-cyclodextrin inclusion with berberine. Spectrochim. Acta Part A 58, 249–256.

2.3 3D-QSAR predictions for bovine serum albumin-water partition coefficients of organic anions using quantum mechanically based descriptors

Environmental Science Processes & Impacts

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c6em00555a

3D-QSAR predictions for bovine serum albumin–water partition coefficients of organic anions using quantum mechanically based descriptors†

Lukas Linden,[a] Kai-Uwe Goss[ab] and Satoshi Endo*[ac]

Received 10th October 2016
Accepted 1st December 2016

DOI: 10.1039/c6em00555a

rsc.li/process-impacts

Ionic organic chemicals are a class of chemicals that is released in the environment in a large amount from anthropogenic sources. Among various chemical and biological processes, binding to serum albumin is particularly relevant for the toxicokinetic behavior of ionic chemicals. Several experimental studies showed that steric effects have a crucial influence on the sorption to bovine serum albumin (BSA). In this study, we investigated whether a 3D quantitative structure–activity relationship (3D-QSAR) model can accurately account for these steric effects by predicting the BSA–water partition coefficients ($K_{BSA/water}$) of neutral and anionic organic chemicals. The 3D-QSAR tested here uses quantum mechanically derived local sigma profiles as descriptors. In general, the 3D-QSAR model was able to predict the partition coefficients of neutral and anionic chemicals with an acceptable quality ($RMSE_{test\ set}$ 0.63 ± 0.10, $R^2_{test\ set}$ 0.52 ± 0.15, both for log $K_{BSA/water}$). Particularly notable is that steric effects that cause a large difference in the log $K_{BSA/water}$ values between isomers were successfully reproduced by the model. The prediction of unknown $K_{BSA/water}$ values with the proposed model should contribute to improved environmental and toxicological assessments of chemicals.

Environmental impact

Ionic and ionogenic chemicals are used in a substantial amount in our daily life and thus released in the environment. Accurately assessing their partitioning and distribution behaviour in organisms is necessary for a qualified assessment of their toxicological and bioaccumulation potential. Serum albumin is an important target for the partitioning of anionic organic chemicals in blood. The results of this work demonstrate that the constructed 3D-QSAR model can be used to predict unknown bovine serum albumin (BSA)–water partitioning coefficients for both neutral and anionic chemicals. The used modelling approach should be applicable to other partitioning processes that are also highly influenced by steric effects.

Introduction

Ionic organic chemicals are common types of chemicals in industry and our daily life. They are, among others, used as pesticides; *e.g.*, 2,4-dichlorophenoxyacetic acid (2,4-D) and methylchlorophenoxypropionic acid (mecoprop) are among the most widely used herbicides[1,2] and both are anionic at typical environmental and physiological pH. Many pharmaceuticals are also ionic; *e.g.*, ibuprofen is an anionic chemical under neutral pH and is one of the most commonly taken nonsteroidal anti-inflammatory drugs.[3] The wide spread of ionic chemicals is also reflected in the general statistics, *e.g.*, under REACH (the Registration, Evaluation, Authorization and Restriction of Chemicals) around 50% of the preregistered chemicals are estimated to be ionogenic.[4] Nevertheless, the ecotoxicological and environmental assessment of organic chemicals (including modeling of their fate) has its focus on neutral species and usually treats the ionic species in a simplistic manner, *i.e.*, with the assumption that ionic species only occur in aqueous phases and do not partition into other phases. However, a number of experimental studies demonstrate that even a rather strong sorption of organic cations to natural organic matter and mineral surfaces in soils[5] and of both cations and anions to phospholipids and proteins in biological tissues may occur.[6–10]

A biological phase particularly relevant for the toxicokinetic behavior of ionic chemicals is serum albumin, the most abundant blood protein of mammals and often a predominating sorption phase in blood.[11] Through its relatively low specificity and strong binding for many chemicals, serum albumin influences the transport and the distribution of many organic ions

[a]Helmholtz Centre for Environmental Research UFZ, Permoserstr. 15, D-04318 Leipzig, Germany

[b]University of Halle-Wittenberg, Institute of Chemistry, Kurt Mothes Str. 2, D-06120 Halle, Germany

[c]Osaka City University, Urban Research Plaza & Graduate School of Engineering, Sugimoto 3-3-138, Sumiyoshi-ku, 558-8585 Osaka, Japan. E-mail: satoshi.endo@urban.eng.osaka-cu.ac.jp; Fax: +81-6-6605-2763; Tel: +81-6-6605-2763

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6em00555a

in organisms. Particularly, anionic chemicals including perfluorinated alkyl acids[12,13] and nonsteroidal anti-inflammatory drugs[14] are known to bind strongly to serum albumin. It is also noted that fetal bovine serum is the most commonly used serum supplement for cell culture assays, where bovine serum albumin (BSA) has a strong impact on the freely dissolved concentration of the test chemical in the assays.[15] Recently, Henneberger *et al.* published BSA/water partition coefficients ($K_{BSA/water}$ [L_{water} kg_{BSA}^{-1}]) for a broad set of ionic chemicals measured under consistent conditions.[7] $K_{BSA/water}$ data help to assess the chemical distribution in organisms and bioassay systems.[16] The reported ionic partition data to serum albumin show specific steric effects, which cannot easily be described by common methods for the prediction of partition coefficients such as polyparameter linear free energy relationships (pp-LFERs).[17] Prediction of $K_{BSA/water}$ may become even more challenging when one aims for a model that can be used both for neutral[18] and ionic organic chemicals. In this study, we aim to construct a model that (i) is capable of predicting log $K_{BSA/water}$ of neutral and ionic chemicals, (ii) can cover the specific 3D effects that influence the binding to BSA, and (iii) can be used to estimate log $K_{BSA/water}$ for the (eco)toxicological and environmental assessment of organic chemicals.

A modeling tool that is conceptually capable of predicting steric effects on sorption is the 3D quantitative structure–activity relationship (3D-QSAR), which correlates 3D-structural features of the chemicals with the property of interest. This approach has been developed since the late 80s[19] and is a well-established ligand-based approach to generate a predictive model.[20] Recently, Klamt *et al.* combined an existing 3D-QSAR method with quantum chemically based molecular descriptors, the local sigma profiles (LSPs).[21] The LSPs emerge from a solid theoretical basis, the COSMO-RS (conductor-like screening model for real solvents) method.[22,23] The COSMO-RS method uses the COSMO surface polarization charge densities (also called the sigma surface) to calculate, among others, partition coefficients and was successfully applied to numerous partition systems.[24,25] The sigma surface describes the abilities of a molecule to undergo intermolecular interactions including electrostatic, hydrogen bond, and van der Waals interactions with its neighboring molecules.[26] The COSMOsim3D method discretizes the sigma surface into LSPs. The LSPs are 4-dimensional histograms describing the amount of surface area within a certain sigma interval in a specific part of the molecule.[21] Klamt *et al.* suggested that LSPs are theoretically more suitable for a linear regression model than the standard comparative molecular field analysis (CoMFA) descriptors,[21] the latter use a van der Waals and an electrostatic potential derived from a molecular mechanics calculation.[27] LSPs were already applied by us to predict the binding to α-cyclodextrin,[28] which is also influenced by 3D effects[29] and is a typical test system that shows specific binding.[30,31] LSPs resulted in a better prediction than the standard CoMFA descriptors for α-cyclodextrin binding data.[28] In this study, we test whether steric effects that influence the partitioning to BSA[7] can also be modeled by the LSPs.

Methods

Dataset

Two datasets of $K_{BSA/water}$ were combined in our study: the dataset from Endo *et al.*[18] with 83 neutral chemicals (log $K_{BSA/water}$ 1.48–4.76) and the dataset from Henneberger *et al.*[7] with 43 anionic chemicals (log $K_{BSA/water}$ 1.65–5.03). The dataset from Henneberger *et al.* includes many benzoic acid anions and naphthoic acid anions with different substitutions and is thus suitable for investigating 3D structural effects on BSA binding. The four cationic chemicals from the Henneberger dataset were not used in this work, because their number is too small for meaningful evaluation.

3D-QSAR

The sorption to binding proteins such as BSA is influenced by the spatial structure of the sorption sites and any possible steric hindrance. This means that a modeling approach needs to represent the spatial structure and the chemical environment of the sorption sites. It should be noted that BSA has multiple binding sites and that the most favorable binding site may depend on the solutes. Thus, to apply 3D-QSARs for BSA binding constants, we have to set the working hypothesis that the different reported sorption sites of BSA are alike and that their spatial structure and interaction possibilities can be expressed through one characteristic binding site.[32]

In general, 3D-QSAR modeling takes the following steps: (1) 3D-structure generation for the sorbing chemicals, (2) alignment, (3) generation of independent variables, (4) training and test set selection in the experimental dataset, (5) model generation by the training set with partial least square (PLS) regression analysis, and (6) model evaluation using the test set. Here, we used the method combination that performed the best in terms of the overall statistics and the qualitative descriptions in the previous publication for α-cyclodextrin binding.[28]

Local sigma profiles

The LSPs are a spatial representation of the surface polarization charge densities and thereby of the interaction possibilities of a chemical (for a graphical explanation see Fig. SI1†). The LSPs are derived from the 3D-COSMO files of the chemicals. The LSPs can be used for the alignment of the chemicals and for the PLS regression model as independent variables. Each LSP was split into sections of 0.006 e $Å^{-2}$ to capture the spatial distribution of surface segments with similar charge densities. LSP 1 starts with the most negative sigma value (in this work, −0.024 to −0.018 e $Å^{-2}$) (note that a negative sigma charge value corresponds to a positive partial charge and *vice versa*) and the LSP with the highest index (in this work, 10) represents the most positive sigma charge values of the molecular surface (0.030–0.036 e $Å^{-2}$).

Alignment

Prior to building a model, we had to generate a common binding hypothesis, *i.e.*, a common 3D alignment, between the

solutes and BSA. For this purpose, we chose those five chemicals from the experimental datasets with the strongest binding to BSA and the most rigid structure, *i.e.*, chemicals with at least one conjugated two-ring aromatic structure, which reduces the degrees of freedom for the alignment. These five chemicals are referred to as template chemicals. 3D structures of one to ten conformers of all chemicals were generated with COSMOconfX15 in combination with Turbomole (v. 7.0)[33] that performs the quantum mechanics calculations generating 3D-COSMO files. For more details, see ref. 28. The software COSMOsim3D[34] generated an averaged sigma surface (including the 3D information) from the sigma surfaces of the five template chemicals, namely benzo[*g*,*h*,*i*]perylene, chrysene, pyrene, naphthalene-2-sulfonate, and 2-naphthaleneacetate. This averaged sigma surface is assumed to describe the 3D interaction requirements of the BSA binding site and was used for the alignment of the chemicals of the dataset. These five chemicals are a reasonable choice for the template because a high partition coefficient corresponds to a good interaction with BSA and a rigid structure helps to delineate the binding site better than a flexible structure. Obviously, the choice of template chemicals is always limited through the data availability of binding chemicals, which may partially limit the domain of applicability of the resulting model. The 3D similarity between the averaged sigma surface of the five template chemicals and the sigma surface of each chemical was maximized through the translation and rotation of each chemical in the 3D space; conceptually, this corresponds to a search for the chemical's relative position that is optimal for interactions with BSA. This optimization procedure was carried out using a grid with a 0.5 Å spacing. All conformers generated for each chemical were aligned. The conformer with the highest alignment score was selected for further modeling and if there were multiple conformers with the same alignment score, then the conformer with the lowest internal energy was used.

Independent variables

The independent variables for the model are the LSPs, *i.e.*, the amount of the surface area within a certain sigma charge interval and a space interval. The LSPs were derived at each grid point of a box with a grid spacing of 2 Å and a size that includes a 5 Å space around the chemicals. In the end, there were ten LSP intervals and 2730 grid points, which gave 27 300 independent variables but on average 2910 active independent variables (variables whose values are unequal to zero). The number of independent variables was then further reduced by an exclusion of variables that have a SD below a level of 0.1 among all training chemicals and by a fractional factorial design selection.[35,36]

Selection procedures for training and test sets

The quality and the predictive power of the 3D-QSAR models were assessed with the test sets whose chemicals were not part of the respective training sets and thus did not influence the construction of the respective model. The statistical results of 3D-QSAR modeling depend highly on the combination of training and test sets. We decided to use several combinations of training and test sets (see the next paragraph) to capture this dependency and to obtain statistical results that represent the entire dataset. The two phenolates in Henneberger's set, namely pentachlorophenolate and bromoxynil anion were used as additional validation chemicals, because we wanted to test how the model performs with the extrapolation to external data that are not represented in the training set in terms of the ionic functional group. In addition, 1-bromo-2-naphthoic acid was also used for additional model validation in order to evaluate the model performance for an external test chemical that has the same ionic functional group as some of the test chemicals. It is worth noting that the critical settings of the alignment, *i.e.*, the grid dimension and the choice of template chemicals, were defined before the selection of test and training sets.

The test sets included eleven anionic and 21 neutral chemicals. These correspond to 25.6% and 25.3%, respectively, of the data available. The dataset was sorted according to the charge state, *i.e.*, neutral or anionic, and to the log $K_{BSA/water}$ values of the chemicals.[37] Then, four consecutive chemicals with the same charge state were placed in one bin – the last bin of the ions contained five chemicals and the last bin of the neutral chemicals contained three chemicals (due to the fact that the total numbers of anionic and neutral chemicals were not multiples of four). A random chemical from each bin was selected and placed in a test set and the rest of the chemicals were put in the training set; this was repeated until five test and training sets were generated. In addition to these random sets, modified test sets were generated by placing some structurally interesting chemicals (*e.g.*, both chemicals from a pair of isomers) always in the test set while the rest of the test sets were selected randomly. This was also performed five times. Modified test sets were prepared to study specific structural effects on log $K_{BSA/water}$.

Statistical tool

PLS regression analysis correlates the independent variables, *i.e.*, the LSPs, with the dependent variables, *i.e.*, the log $K_{BSA/water}$ values, of the training set. For the PLS regression analysis the program Open3DQSAR[38] was used, see ref. 28. Models with one to five PLS components (PCs) were generated and leave-two-out cross-validation was used to select the best model in terms of predictive power and least chance of overfitting. This selected model is then used to predict the test set.

Domain of applicability

Tanimoto indices[39] were applied to calculate the similarity of a test chemical to the training set. For the LSPs of two different chemicals (X and Y), the Tanimoto index is calculated as:

$$T_j(x,y) = \frac{\sum X_{ij}\ Y_{ij}}{\sum {X_{ij}}^2 + \sum {Y_{ij}}^2 - \sum X_{ij} Y_{ij}} \quad (1)$$

with X_{ij} and Y_{ij}, the j-th field values at the i-th grid point. The arithmetic mean of the Tanimoto indices of the LSP 1 to 10 (*i.e.*, the j-th field value in eqn (1)) of a test chemical was calculated against each of the chemicals in the training set. Then, the

mean of the five highest values was calculated (Tanimoto index mean). Data were grouped for every Tanimoto index mean value of 0.1 (called Tanimoto groups). We then compared the prediction errors of the different Tanimoto groups. The statistical difference between the variances of two Tanimoto groups was determined with a Brown–Forsythe analysis[40] and the statistical difference between the medians of two Tanimoto groups was determined with a Mann–Whitney U analysis.[41] These statistical tests were selected because the data are, most likely, not normally distributed.

Results and discussion

General performance of the models for log $K_{BSA/water}$

Five 3D-QSAR models were generated from different subsets of the available experimental data to describe the partitioning to BSA and to predict the respective test sets (using on average 230 independent variables). Fig. 1 gives examples of the test set predictions that resulted from different combinations of training and test sets. Panel A of Fig. 1 shows the best of the five predictions, while panel B shows the worst. All chemicals lay closer to the 1 : 1 line in panel A than in panel B. The biggest outlier of all predictions was flufenamic acid anion with a prediction error of 1.9 log units (shown in panel B). The prediction of the five random test sets resulted in an RMSE of 0.63 ± 0.10 and an R^2 of 0.52 ± 0.15 (the values represent the mean ± standard deviation). The neutral chemicals ($n = 21$) of the test set were predicted with an RMSE of 0.59 ± 0.04 while anionic chemicals ($n = 11$) were predicted with an RMSE of 0.68 ± 0.23. In general, the neutral chemicals are better predicted compared to the anionic chemicals, which might be caused by the disproportion of the training sets (62 neutral chemicals and 32 anionic chemicals). However, the neutral chemicals in the calibration set appear to improve the description of the partitioning of anionic chemicals to BSA, as modelling using solely the anionic chemicals was less successful (data not shown) than that with the combined dataset. Reasons for this outcome could be the small number of anionic chemicals that is not sufficient to calibrate the model, and the higher diversity of the neutral dataset that helps also to predict log $K_{BSA/water}$ of less diverse, though anionic, chemicals.

A chance correlation of the models can be excluded based on the results of ten scrambling runs[42,43] using two log $K_{BSA/water}$ sorted bins, *i.e.*, the log $K_{BSA/water}$ values of the chemicals of each bin were permuted in the respective bin prior to each run. The resulting statistics of leave one out cross-validation indicate non-predictive models (mean $R^2 = 0.44$, mean $q_{LOO}^2 = 0.002$).

The binding mechanism behind the 3D-QSAR model can be examined with the help of the contributions of the different LSPs to the overall model. Fig. 2 shows the percentage contributions of the LSPs 2 to 9 to the PCs that were generated with the training set of the best prediction (Fig. 1A). The LSPs 1 and 10 contributed to the PCs only to a negligible degree and thus are not shown. The LSP 8 (representing a part of the anionic interactions) contributes 20% to the PC 1, which explains 48.7% of the variance in log $K_{BSA/water}$. Thus, the positive influence of anionic partial charges on the partitioning to BSA, which is apparent in the experimental data, is captured in the model. Other important interactions identified by the model are van der Waals interactions and the hydrophobic effect (LSPs 4, 5).

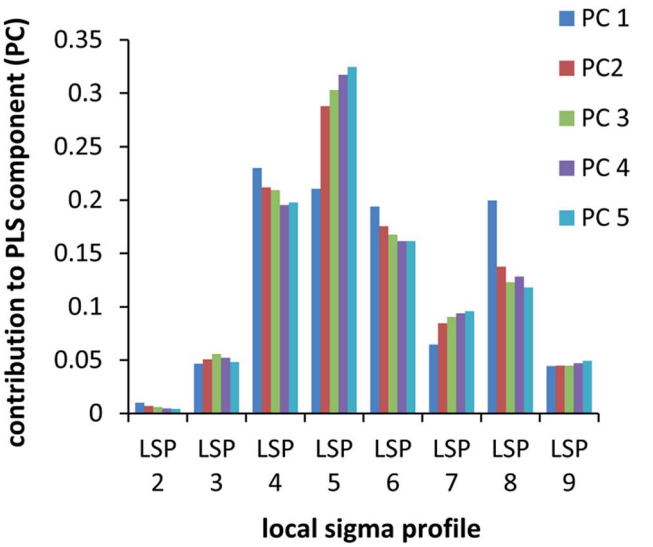


Fig. 2 Contribution of the local sigma profiles for the different PLS components of the 3D-QSAR model.

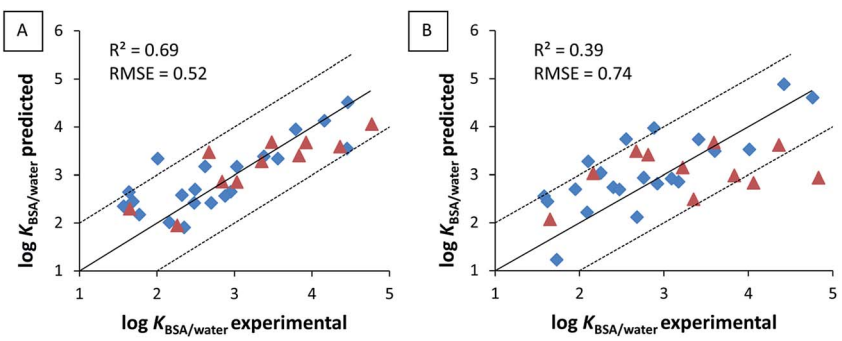


Fig. 1 (A) Best and (B) worst prediction of log $K_{BSA/water}$ for 21 neutral and 11 anionic chemicals in five random testsets. The blue diamonds indicate the neutral chemicals and the red triangles indicate the anionic chemicals. The solid line indicates the 1 : 1 line and the dashed lines indicate a deviation of 1 log unit from the 1 : 1 line.

Prediction of molecular steric effects

For further evaluation of the modeling approach, model performance was investigated for isomer pairs using the modified test sets. In experimental data, several isomer pairs show similar steric effects: an *ortho*-substitution of benzoate decreases $\log K_{BSA/water}$ substantially compared to a *para*- or *meta*-substitution (2-chlorobenzoate *vs.* 4-chlorobenzoate, 2,6-dichlorobenzoate *vs.* 3,4-dichlorobenzoate, 2-methylbenzoate *vs.* 4-methylbenzoate) and a substitution at the *alpha*-position of naphthalene decreases $\log K_{BSA/water}$ while a substitution at the *beta*-position increases $\log K_{BSA/water}$, particularly if the substitution group is negatively charged (1-naphthoic acid anion *vs.* 2-naphthoic acid anion, 1-naphthaleneacetic acid anion *vs.* 2-naphthaleneacetic acid anion). The steric hindrance by the *ortho*-substitution results in a twist of the carboxylate group,[7] which was speculated as a possible reason for the observed specificity. The relative sorption behavior of these isomer pairs with steric effects was predicted correctly by the models (Fig. 3). Even quantitative predictions (errors < 0.8) were achieved for three of the five isomer pairs. The other two had relatively large prediction errors: $\log K_{BSA/water}$ of 3,4-dichlorobenzoate is underestimated (by 1.26 ± 0.22 log units) and $\log K_{BSA/water}$ of 4-methylbenzoate is overestimated (by 0.85 ± 0.04 log units). Another pair of chemicals that is of interest is 2,4,6-trimethylbenzene sulfonate and 2,4,6-trimethylbenzoate, the former has a 2.3 log units higher $\log K_{BSA/water}$ value than the latter. This difference is also predicted correctly but it might not be solely caused by the steric hindrance of the carboxylate group, which is explained in the following.

The alignment of the chemicals had an important role in the distinction of the isomer pairs (Fig. 3). The green lines in the pictures of Fig. 3 show the five chemicals used as alignment templates (see Methods, Alignment) while the sticks show the respective isomers. In addition, the anionic groups of naphthalene-2-sulfonate and 2-naphthaleneacetate are located at the same position, which could represent a possible interaction with a positively charged or electron-withdrawing group of BSA.[45] Indeed, all isomers of Fig. 3 with the higher $\log K_{BSA/water}$ value have their charged group located close to this position (this interaction space is indicated in Fig. 3 by the teal and violet areas as it is expressed in the model). The isomers of Fig. 3 with the lower $\log K_{BSA/water}$ value (marked with red squares) have their anionic group at different positions, which seems to be inevitable for maximizing the overlapping of the rest of the structure to the template but seems to lead to omission of the interaction between the charged group of the chemical and BSA in the model. This difference in the positions of the anionic groups, which is caused by the twist of the carboxylate group, can explain the different $\log K_{BSA/water}$ values of the isomers (see Fig. SI9–11† for conformations of the isomers).

In comparison to the superimposition of the other aromatic chemicals, 2,4,6-trimethylbenzene sulfonate has a shifted position in the alignment (Fig. SI3†). This could be a hint for a different binding mode of 2,4,6-trimethylbenzene sulfonate ($\log K_{BSA/water}$ exper.: 4.23 pred.: 3.52). A closer inspection of the sigma surface of 2,4,6-trimethylbenzene sulfonate shows that: (a) its aromatic ring exhibits a lower electron density than that of 2,4,6-trimethylbenzoate ($\log K_{BSA/water}$ exper.: 1.99 pred.: 2.00) (Fig. SI12 and 13†) and (b) the C–SO^{3-} bond (1.8 Å) is longer than the C–CO^{2-} bond (1.5 Å).[44] The latter structural feature might allow 2,4,6-trimethylbenzene sulfonate to undergo an interaction with the charged group even in the presence of the steric hindrance of the neighboring methyl groups.[44] Furthermore, the sulfonate group has higher interaction possibilities than the carboxylate group because the sulfonate group has an additional oxygen atom and the C–SO_3^- bond is better rotatable than the C–CO_2^- bond. Thus, the positions and interactions of the sp^2 orbitals of the oxygens are more flexible in the case of 2,4,6-trimethylbenzene sulfonate. These flexibilities of 2,4,6-trimethylbenzene sulfonate in the positioning and the interaction possibilities may result in a higher experimental and predicted $\log K_{BSA/water}$ value as compared to 2,4,6-trimethylbenzoate. These inferences are based on the alignment results, which led to successful modeling, but additional insight from further experimental data or direct modeling tools, like molecular dynamics simulation,[46] would be desirable.

Domain of applicability

The domain of applicability was assessed with the help of the Tanimoto indices. The median of the prediction errors for the five random test sets apparently decreases with increasing Tanimoto index mean (Fig. 4). This may suggest that the reliability of the prediction rises with increasing Tanimoto index mean. For statistical evaluation, we chose the second highest range of Tanimoto index mean (0.6–0.7) as the reference group and tested the differences in prediction errors of all the other groups from it (Table SI1†). We did not consider the group 0.7–0.8 because it comprises only four chemicals. Compared to the reference group, the median of the prediction errors is only significantly larger for the Tanimoto group of 0.3–0.4. No group has a significantly different variance than the reference group. Note, however, that the prediction error depends strongly on the combination of test and training sets. We also compared prediction errors and Tanimoto index means using test and training sets generated by a slightly less random procedure. This procedure (see the ESI† for details) uses each chemical once as a test set chemical. Although the resulting plot appears comparable to that presented in Fig. 4, the medians of the prediction errors are significantly larger for all Tanimoto groups <0.50 than for the reference group (Table SI2†), showing that the Tanimoto index means could indicate the domain of applicability. We do not know why Tanimoto index works in one case but not the other. Possible reasons include: the data size is not sufficiently large to show a statistical significance, and the Tanimoto index mean calculated in this study (*i.e.*, the mean of the top five Tanimoto indices) is not suitable.

The three anions that were not part of the model calibration set, nor included in Fig. 4, were used as additional validation chemicals. The prediction is accurate for 1-bromo-2-naphthoic acid anion (prediction error 0.08 log units) despite a relatively small Tanimoto index mean of 0.34. In contrast, bromoxynil anion and pentachlorophenolate were predicted with 2.47 and

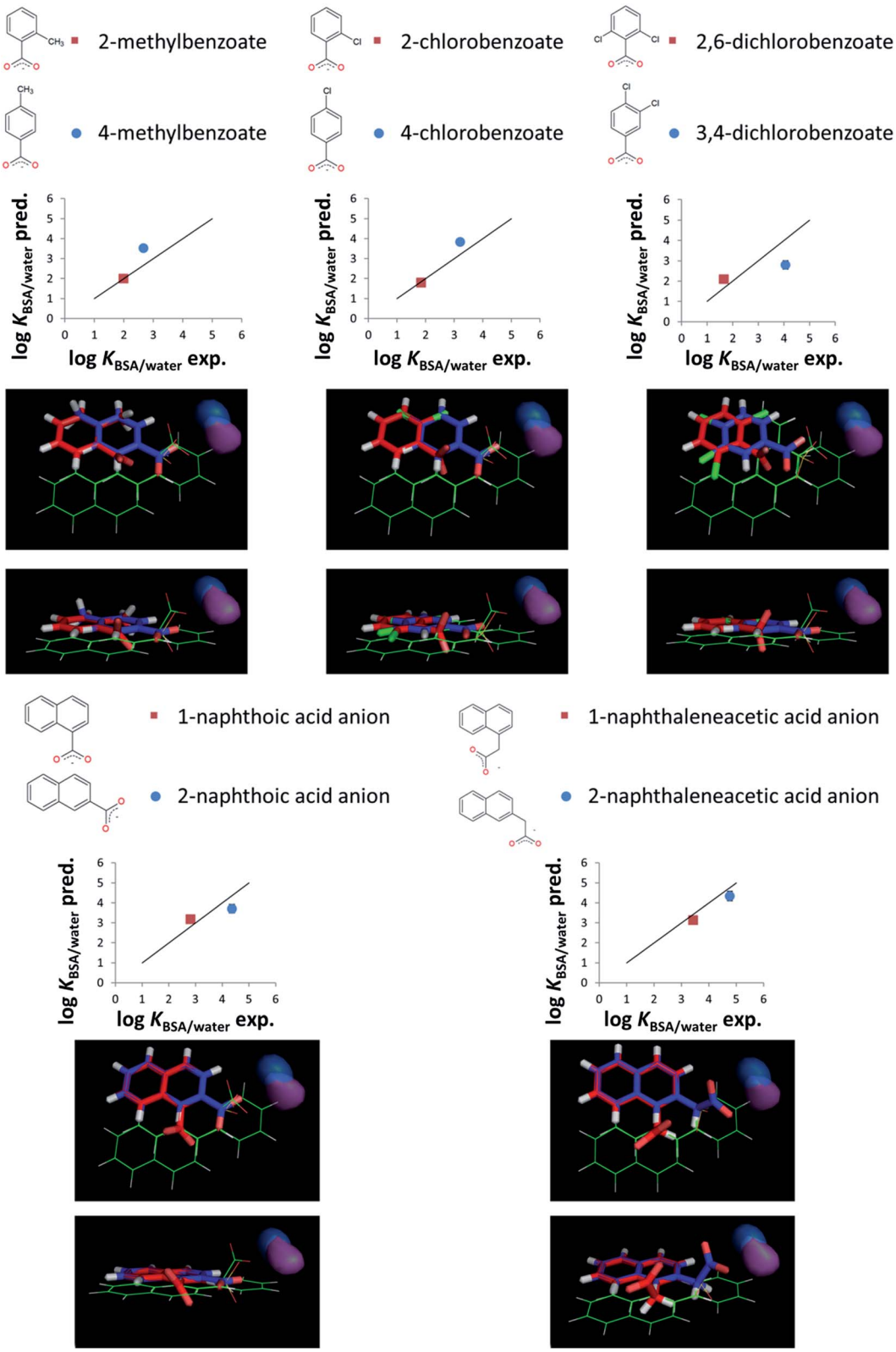


Fig. 3 Experimental and the average predicted log $K_{BSA/water}$ values of the modified test sets for several isomer pairs. The black line in the graphs indicates the 1 : 1 line, the red squares indicate the *ortho*- or *alpha*-substituted isomer, and the blue squares indicate the *para*- or *beta*-substituted isomer. The error bars indicate the respective standard deviation of the averaged predicted log $K_{BSA/water}$ values (mostly not visible). The green lines in the pictures show the alignment of the template chemicals while the blue sticks show the *ortho*- or *alpha*-substituted isomer and the red sticks show the *para*- or *beta*-substituted isomer. The teal (LSP 7) and the violet (LSP 8) areas indicate the space where the models identified a positive interaction of an anionic partial charge with BSA. The alignment figures were generated using Pymol.[44]

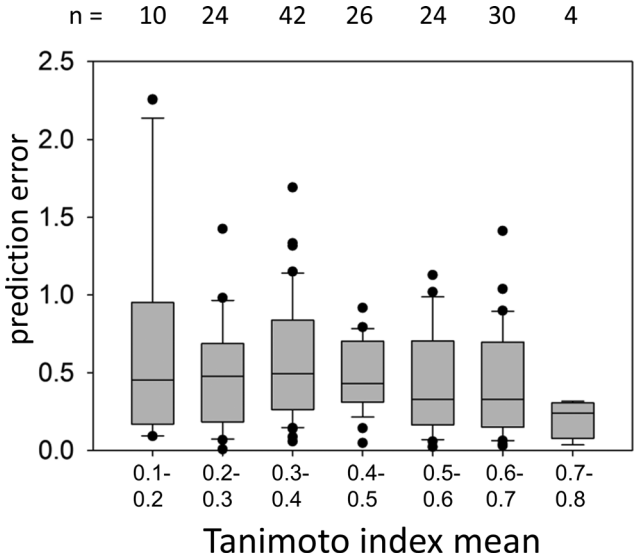


Fig. 4 Prediction errors of the 3D-QSAR model plotted against the mean of Tanimoto indices of the five most similar chemicals of the training set. The boxes outline the 25th to 75th percentiles, the lines through the centers represent the median, the whiskers indicate the 90th and 10th percentiles, and the dots indicate outlying points. The results for all five random test sets are plotted.

2.33 log units off, respectively. Both chemicals have a Tanimoto index mean value of 0.16, which indicates a higher chance for a large prediction error. The large prediction errors for these two phenolates can be expected because the training set does not contain any phenolate, and their low Tanimoto index means reasonably explain the outlying behavior of these chemicals. In the alignment, bromoxynil anion and pentachlorophenolate are displaced compared to the other aromatic chemicals, which might be caused by the different nature of the anionic groups of the template chemicals and of these two phenolates. For a future successful prediction of log $K_{BSA/water}$ for phenolates more experimental data for phenolates and thus a better calibration through phenolates in the training set of the 3D-QSAR model appear to be needed. Moreover, template chemicals may also need to include at least one phenolate.

Other chemicals that are expected to be out of the domain of applicability of the presented model are zwitterions and cations because they have no representation in the training set. Multiply charged anions may also be difficult to predict because the effect of the second charged group is probably not covered by the model. Other examples of chemicals that should be out of the domain of applicability are big bulky chemicals (*e.g.*, monensin, Tanimoto index mean 0.07, perfluorononanoic carboxylate, Tanimoto index mean 0.09) including oligosaccharides (*e.g.*, maltotriose, Tanimoto index mean 0.12), long tertiary and quaternary organic chemicals (*e.g.*, 4-butyl-4-pentylnonanal, Tanimoto index mean 0.14), because they are not part of the current calibration set and might bind to BSA through another mechanism. The same holds true for fatty acids, which bind to a specific binding site of BSA[47] (*e.g.*, undecane carboxylate, Tanimoto index mean 0.19).

Conclusions

The 3D-QSAR model with LSPs as descriptors is capable of describing and predicting log $K_{BSA/water}$ for anionic and neutral chemicals. The assumptions behind the generated characteristic binding site (*i.e.*, several localized binding sites with similar chemical environments and the interaction possibilities of the sites can be expressed as an averaged characteristic binding site) appear to be adequate for the 3D-QSAR modelling approach. The discrimination between different binding sites was not necessary for successful modeling for the dataset used in this work. The steric effects that are responsible for up to two log units differences in log $K_{BSA/water}$ between structural isomers are successfully captured by the model. Thus, the model may be used for the prediction of unknown $K_{BSA/water}$ for neutral and anionic chemicals, which is helpful for a qualified environmental and toxicological assessment of these chemicals. As an example, in an upcoming study the 3D-QSAR model developed in this work will be used to assess the freely dissolved concentration of chemicals in a typical cell assay.[16] Furthermore, the model could contribute to an estimation of the bioaccumulation potential of organic anions, provided that other sorption phases such as phospholipid membranes are considered as well. Whereas serum albumin appears not to be the most important plasma binding protein for many cationic chemicals,[48] an extension of the model with more cationic chemicals is still desirable because there are cations that bind strongly to serum albumin.[11] An inclusion of zwitterions is another interesting example of possible extensions of the model applicability domain. The availability of accurately and consistently measured data will be the key to such future work.

Acknowledgements

The authors thank the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE) for financial support. SE acknowledges the financial support from the MEXT/JST Tenure Track Promotion Program. The authors thank Nadin Ulrich for helpful comments on an early version of the manuscript and the anonymous referees for their constructive comments on the manuscript.

References

1 A. D. D. Grube, T. Kiely and L. Wu, *Pesticides Industry Sales and Usage: 2006 and 2007 Market Estimates*, United States Environmental Protection Agency, Washington, DC, 2011.
2 US Environmental Protection Agency, Reregistration Eligibility Decision for Mecoprop-p, https://archive.epa.gov/pesticides/reregistration/web/pdf/mcpp_red.pdf, 2007.
3 IMS Health Rezeptfreie Schmerzmittel, http://www.imshealth.de/files/web/Germany/Publikationen/Infografiken/2014_9_IMS_Infografik_%20Schmerzmittel.pdf, accessed 03.08 2016.
4 A. Franco, A. Ferranti, C. Davidsen and S. Trapp, An unexpected challenge: ionizable compounds in the REACH chemical space, *Int. J. Life Cycle Assess.*, 2010, **15**(4), 321–325.

5 S. T. J. Droge and K.-U. Goss, Development and Evaluation of a New Sorption Model for Organic Cations in Soil: Contributions from Organic Matter and Clay Minerals, *Environ. Sci. Technol.*, 2013, **47**(24), 14233–14241.
6 K. Bittermann, S. Spycher and K. U. Goss, Comparison of different models predicting the phospholipid-membrane water partition coefficients of charged compounds, *Chemosphere*, 2016, **144**, 382–391.
7 L. Henneberger, K.-U. Goss and S. Endo, Equilibrium Sorption of Structurally Diverse Organic Ions to Bovine Serum Albumin, *Environ. Sci. Technol.*, 2016, **50**(10), 5119–5126.
8 C. A. Ng and K. Hungerbühler, Bioconcentration of Perfluorinated Alkyl Acids: How Important is Specific Binding?, *Environ. Sci. Technol.*, 2013, **47**(13), 7214–7223.
9 L. Henneberger, K.-U. Goss and S. Endo, Partitioning of Organic Ions to Muscle Protein: Experimental Data, Modeling, and Implications for in Vivo Distribution of Organic Ions, *Environ. Sci. Technol.*, 2016, **50**(13), 7029–7036.
10 J. M. Kremer, J. Wilting and L. H. Janssen, Drug binding to human alpha-1-acid glycoprotein in health and disease, *Pharmacol. Rev.*, 1988, **40**(1), 1–47.
11 U. Kragh-Hansen, Molecular aspects of ligand binding to serum albumin, *Pharmacol. Rev.*, 1981, **33**(1), 17–53.
12 H. N. Bischel, L. A. MacManus-Spencer and R. G. Luthy, Noncovalent Interactions of Long-Chain Perfluoroalkyl Acids with Serum Albumin, *Environ. Sci. Technol.*, 2010, **44**(13), 5263–5269.
13 H. N. Bischel, L. A. MacManus-Spencer, C. Zhang and R. G. Luthy, Strong associations of short-chain perfluoroalkyl acids with serum albumin and investigation of binding mechanisms, *Environ. Toxicol. Chem.*, 2011, **30**(11), 2423–2430.
14 F. Lapicque, N. Muller, E. Payan, N. Dubois and P. Netter, Protein Binding and Stereoselectivity of Nonsteroidal Anti-Inflammatory Drugs, *Clin. Pharmacokinet.*, 1993, **25**(2), 115–125.
15 M. Gülden, S. Mörchel, S. Tahan and H. Seibert, Impact of protein binding on the availability and cytotoxic potency of organochlorine pesticides and chlorophenols in vitro, *Toxicology*, 2002, **175**(1–3), 201–213.
16 F. Fischer, L. Henneberger, M. König, K. Bittermann, L. Linden, K. U. Goss and B. I. Escher, Modelling freely dissolved and internal cellular effect concentrations in the Tox21 in vitro bioassays, 2016, in preparation.
17 S. Endo and K. U. Goss, Applications of Polyparameter Linear Free Energy Relationships in Environmental Chemistry, *Environ. Sci. Technol.*, 2014, **48**(21), 12477–12491.
18 S. Endo and K.-U. Goss, Serum Albumin Binding of Structurally Diverse Neutral Organic Compounds: Data and Models, *Chem. Res. Toxicol.*, 2011, **24**(12), 2293–2301.
19 R. D. Cramer, D. E. Patterson and J. D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.*, 1988, **110**(18), 5959–5967.
20 G. Lambrinidis, T. Vallianatou and A. Tsantili-Kakoulidou, In vitro, in silico and integrated strategies for the estimation of plasma protein binding. A review, *Adv. Drug Delivery Rev.*, 2015, **86**, 27–45.
21 A. Klamt, M. Thormann, K. Wichmann and P. Tosco, COSMOsar3D: Molecular Field Analysis Based on Local COSMO σ-Profiles, *J. Chem. Inf. Model.*, 2012, **52**(8), 2157–2164.
22 A. Klamt, Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena, *J. Phys. Chem.*, 1995, **99**(7), 2224–2235.
23 A. Klamt, V. Jonas, T. Bürger and J. C. W. Lohrenz, Refinement and Parametrization of COSMO-RS, *J. Phys. Chem. A*, 1998, **102**(26), 5074–5085.
24 A. Klamt, F. Eckert and W. Arlt, COSMO-RS: An Alternative to Simulation for Calculating Thermodynamic Properties of Liquid Mixtures, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**(1), 101–122.
25 M. Diedenhofen and A. Klamt, COSMO-RS as a tool for property prediction of IL mixtures—a review, *Fluid Phase Equilib.*, 2010, **294**(1–2), 31–38.
26 A. Klamt, The COSMO and COSMO-RS solvation models, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**(5), 699–709.
27 C. C. Melo, R. C. Braga and C. H. Andrade, 3D-QSAR Approaches in Drug Design: Perspectives to Generate Reliable CoMFA Models, *Curr. Comput.-Aided Drug Des.*, 2014, **10**(2), 148–159.
28 L. Linden, K.-U. Goss and S. Endo, 3D-QSAR predictions for α-Cyclodextrin binding constants using quantum mechanically based descriptors, *Chemosphere*, 2017, **169**, 693–699.
29 L. Linden, K.-U. Goss and S. Endo, Exploring 3D structural influences of aliphatic and aromatic chemicals on α-cyclodextrin binding, *J. Colloid Interface Sci.*, 2016, **468**, 42–50.
30 I. Tabushi, Cyclodextrin catalysis as a model for enzyme action, *Acc. Chem. Res.*, 1982, **15**(3), 66–72.
31 H. J. Schneider, Binding mechanisms in supramolecular complexes, *Angew. Chem.*, 2009, **48**(22), 3924–3977.
32 O. K. Abou-Zied, N. Al-Lawatia, M. Elstner and T. B. Steinbrecher, Binding of Hydroxyquinoline Probes to Human Serum Albumin: Combining Molecular Modeling and Förster's Resonance Energy Transfer Spectroscopy to Understand Flexible Ligand Binding, *J. Phys. Chem. B*, 2013, **117**(4), 1062–1074.
33 D. Sijm, R. Kraaij and A. Belfroid, Bioavailability in soil or sediment: exposure of different organisms and approaches to study it, *Environ. Pollut.*, 2000, **108**(1), 113.
34 M. Thormann, A. Klamt and K. Wichmann, COSMOsim3D: 3D-Similarity and Alignment Based on COSMO Polarization Charge Densities, *J. Chem. Inf. Model.*, 2012, **52**(8), 2149–2156.
35 M. Baroni, S. Clementi, G. Cruciani, G. Costantino, D. Riganelli and E. Oberrauch, Predictive ability of regression models. Part II: selection of the best predictive PLS model, *J. Chemom.*, 1992, **6**(6), 347–356.
36 M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi, Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems, *Quant. Struct.-Act. Relat.*, 1993, **12**(1), 9–20.

37 G. W. Kauffman and P. C. Jurs, QSAR and k-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors, *J. Chem. Inf. Comput. Sci.*, 2001, **41**(6), 1553–1560.
38 P. Tosco and T. Balle, Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields, *J. Mol. Model.*, 2011, **17**(1), 201–208.
39 V. Monev, Introduction to similarity searching in chemistry, *MATCH Commun. Math. Comput. Chem.*, 2004, **51**, 7–38.
40 M. B. Brown and A. B. Forsythe, Robust Tests for the Equality of Variances, *J. Am. Stat. Assoc.*, 1974, **69**(346), 364–367.
41 H. B. Mann and D. R. Whitney, On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other, *Ann. Math. Stat.*, 1947, **18**(1), 50–60.
42 A. Tropsha, P. Gramatica and V. K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sci.*, 2003, **22**(1), 69–77.
43 C. Rücker, G. Rücker and M. Meringer, y-Randomization and its Variants in QSPR/QSAR, *J. Chem. Inf. Model.*, 2007, **47**(6), 2345–2357.
44 Schrodinger, LLC, *The PyMOL Molecular Graphics System, Version 1.3r1*, 2010.
45 T. Peters Jr, All About Albumin, in *All about Albumin*, Academic Press, San Diego, 1995, pp. xv–xvii.
46 B. Sudhamalla, M. Gokara, N. Ahalawat, D. G. Amooru and R. Subramanyam, Molecular Dynamics Simulation and Binding Studies of β-Sitosterol with Human Serum Albumin and Its Biological Relevance, *J. Phys. Chem. B*, 2010, **114**(27), 9054–9062.
47 G. J. van der Vusse, Albumin as Fatty Acid Transporter, *Drug Metab. Pharmacokinet.*, 2009, **24**(4), 300–307.
48 G. L. Trainor, The importance of plasma protein binding in drug discovery, *Expert Opin. Drug Discovery*, 2007, **2**(1), 51–64.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Weiterhin erkläre ich, dass ich noch keine vergeblichen Promotionsversuche unternommen habe und die Dissertation weder in der gegenwärtigen noch in einer anderen Fassung bereits einer anderen Fakultät vorgelegen hat.

Leipzig, 10. Januar 2017

Lukas Linden

Angaben zur Person und zum Bildungsgang

Name	Lukas Linden
Geburtsdatum	29.06.1987
Geburtsort	Mainz

03/2013 – 09/2016 **Helmholtz Zentrum für Umweltforschung**
Department: Analytische Umweltchemie
Promotion: “Influence of molecular steric factors on the sorption of organic chemicals”

04/2007 – 09/2012 **Johannes-Gutenberg-University Mainz**
Fach: bio-medizinische Chemie
Hauptfächer: Organische Chemie, Analytik, Biochemie, Pharmakologie und Toxikologie

12/2011 – 09/2012 Diplom Arbeit "Analysen zur Regulation der KSRP-Expression in humanen DLD-1-Zellen - Effekte auf die pro-inflammatorische Genexpression" at the institute of pharmacology

08/2010 – 01/2011 **University of Aberdeen**
Erasmus Praktikum Programm

- Forschungspraktikum am “Department of analytical chemistry”
- Themea: "Mercury uptake in rice plants - Determination in different parts"

08/1997 - 03/2006 **Willigis-Gymnasium Mainz**

Publikationsliste

Veröffentlichungen

- Linden, L.; Goss, K.-U.; Endo, S. Exploring 3D Structural Influences of Aliphatic and Aromatic Chemicals on α-Cyclodextrin Binding. J. Colloid Interface Sci. 2016, 468, 42-50
- Linden, L.; Goss, K.-U.; Endo, S. 3D-QSAR Predictions for α-Cyclodextrin Binding Constants Using Quantum Mechanically Based Descriptors. Chemosphere 2017, 169, 693-699
- Linden, L.; Goss, K.-U.; Endo, S. 3D-QSAR predictions for bovine serum albumin-water partition coefficients of organic anions using quantum mechanically based descriptors. Environ. Sci.: Processes Impacts 2017, 10.1039/C6EM00555A

Konferenzbeiträge

20. - 24.09.2015 ICCE, Leipzig

- L. Linden, S. Endo, K.-U. Goss: 3D-QSAR: a better tool for predicting partition coefficients influenced by molecular steric effects?, Vortrag

16. - 18.03.2015 COSMO Symposium, Bonn

- K.-U. Goss, L. Linden: Physicochemical Property Prediction for Environmental Research, Vortrag

11. - 15.05.2014 SETAC Europe, Basel

- L. Linden, S. Endo, K.-U. Goss: Influence of 3D molecular structure on partitioning and sorption behaviour of organic chemicals – αcyclodextrin binding as a test case, Poster

Danksagung

Mein Dank gilt vor allem meinen Betreuern Prof. Dr. Kai-Uwe Goss und Dr. Satoshi Endo, für die stets hervorragende Betreuung, für ihr Engagement und die Unterstützung bei der Erstellung dieser Arbeit. Insbesondere die gute und weitreichende Betreuung durch Satoshi Endo, trotz der beginnenden Elternschaft und der räumlichen Distanz, war beeindruckend. Außerdem möchte ich mich bei meinen Kollegen Angelika Stenzel, Luise Henneberger, Anett Krause, Kai Bittermann, Wolfgang Larisch, Sophia Krause, Nadin Ulrich und Trevor Brown für die gute Zusammenarbeit und die angenehme Atmosphäre bedanken. Danke besonders an Nadin Ulrich und Kai Bittermann für die konstruktiven Diskussionen und Korrekturen, die immer einen erweiterten Blickwinkel auf die eigene Arbeit ermöglichten. Andrea Pfennigsdorff muss für das gute Labormanagement, die Hilfe im Labor und die moralische, herzliche Unterstützung gedankt werden.

Gedankt sei auch den anderen „Leidensgenossen“ des UFZs für die fachübergreifenden Diskussionen, die interessanten Perspektiven, die gegenteilige Anteilnahme und natürlich die fröhlichen Runden beim Volleyball und anderen Gelegenheiten. Nicht zuletzt danke ich meiner Freundin, die mich immer unterstützt und sehr unkompliziert ermutigt hat, und meiner Familie.