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

Lukas Linden, Kai-Uwe Goss, Satoshi Endo

Lukas Linden, Kai-Uwe Goss, Satoshi Endo. (2017). 3D-QSAR predictions for α-cyclodextrin binding constants using quantum mechanically based descriptors. *Chemosphere*. 169, 693-699.

"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

*corresponding author: phone/fax: ++81-6-6605-2763; e-mail: satoshi.endo@urban.eng.osaka-cu.ac.jp

Abbreviations:
α-cyclodextrin αCD; quantitative structure activity relationship QSAR; local sigma profiles LSPs; COSMO conductor-like screening model; comparative molecular field analysis CoMFA; molecular interaction fields MIFs; partial least square PLS; COSMO-RS conductor-like screening model for real solvent; linear solvation energy relationship LSER; poly-parameter linear free energy relationship pp-LFER; Open3Dalign O3A; molecular dynamics simulation MDsim; electrostatic ele

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.

Keywords:

α-Cyclodextrin (CD); Binding constant; Inclusion complex; Prediction

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; 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 solvent) 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 such 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 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 (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 COSMOconfX13. 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/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 (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; Baroni et al., 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.

2.4 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 hexandecane-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.

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.

Modeling	Method	Alignment	Field	q^2 ± SD	RMSE ± SD	R^2 ± SD

approach						
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.

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

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

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 (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 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 ($R_{training}^2$ = 0.40, q_{LTO}^2 = -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/Å², 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.

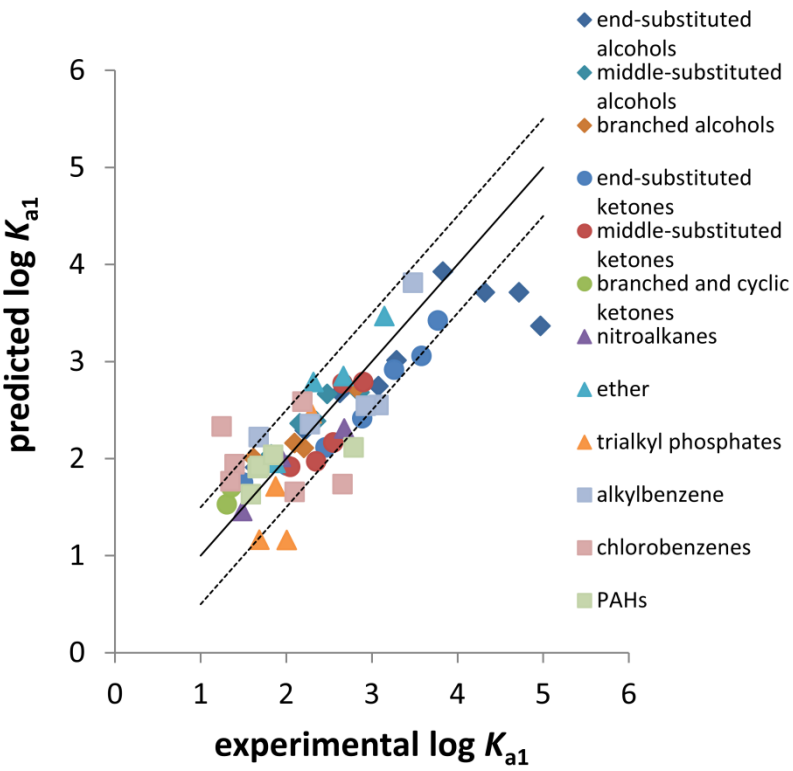

Figure 1. Prediction of log K_{a1} of 60 Linden's chemicals with COSMOsim3D alignment and local sigma profiles as variables (M4). Test sets were selected with test set selection procedure 2. The solid line indicates the 1:1 line and the dashed lines indicate a deviation of 0.5 log units from the 1:1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

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 so 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.

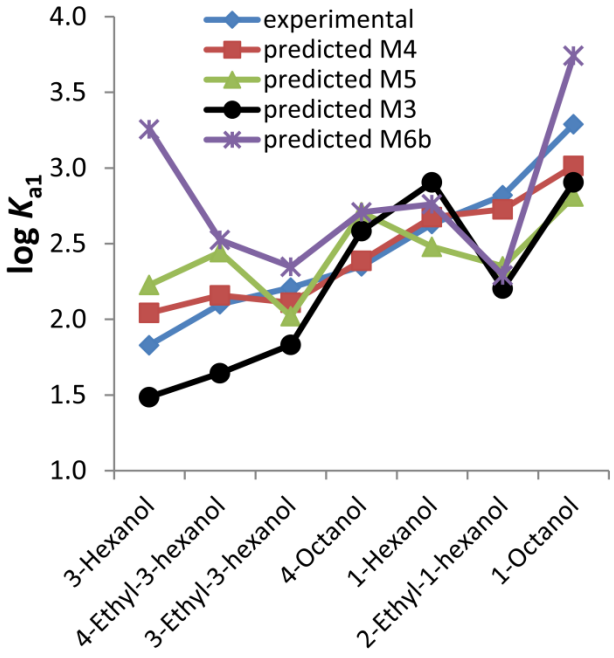

Figure 2. Experimental and predicted log K_{a1} for αCD binding of two C6-alcohols and five C8-alcohols.

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 small number of 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

406 tried the reversed evaluation (i.e., using the Suzuki set as the training set and the Linden set
407 as the test set, Table SI 6) and obtained similar statistics but substantially different
408 regression coefficients.

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

427 We further tested if the steric restriction through the cavity can correctly be described by
428 the model variant M4. The binding coefficients were predicted for the ten chemicals for
429 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 material

Supplementary data associated with this article can be found, in the online version, at …

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. Journal of the Chemical Society, Perkin Transactions 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. Environmental Science & Technology 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, Leipzig, Germany.

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 & Combinatorial Science 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 Interdisciplinary Reviews: Computational Molecular Science 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. Journal of Chemical Information and Modeling 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. Environmental Science & Technology 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 Fragrance J. 25, 313-326.
Rücker, C., Rücker, G., Meringer, M., 2007. Y-randomization and its variants in qspr/qsar. Journal of Chemical Information and Modeling 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.
Sawicki, R., Mercier, L., 2006. Evaluation of mesoporous cyclodextrin-silica nanocomposites for the removal of pesticides from aqueous media. Environmental Science & Technology 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. Journal of Chemical Information and Modeling 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 & Combinatorial Science 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.

All tables:

Table 2. 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.

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.