Predicting Partition Coefficients of Short-Chain Chlorinated Paraffin Congeners by Combining COSMO-RS and Fragment Contribution Model Approaches

Satoshi Endo, Jort Hammer*

National Institute for Environmental Studies (NIES), Center for Health and Environmental Risk Research, Onogawa 16-2, 305-8506 Tsukuba, Ibaraki, Japan

*Corresponding author

Satoshi Endo, Phone/Fax: ++81-29-850-2695, endo.satoshi@nies.go.jp

Abstract

Chlorinated paraffins (CPs) are highly complex mixtures of polychlorinated *n*-alkanes with differing chain lengths and chlorination patterns. Knowledge on physicochemical properties of individual congeners is limited but needed to understand their environmental fate and potential risks. This work combines a sophisticated but time-demanding quantum chemically based method COSMO-RS and a fast-running fragment contribution approach to establish models to predict partition coefficients of a large number of short-chain chlorinated paraffin (SCCP) congeners. Molecular fragments of a length of up to C_4 in CP molecules were counted and used as explanatory variables to develop linear regression models for predicting COSMO-RS-calculated values. The resulting models can quickly provide COSMO-RS predictions for octanol–water (K_{ow}), air–water (K_{aw}), and octanol–air (K_{oa}) partition coefficients of SCCP congeners with an accuracy of 0.1–0.3 log units root mean squared errors (RMSE). The model predictions for K_{ow} agree with experimental values for individual constitutional isomers within 1 log unit. The ranges of partition coefficients for each SCCP congener group were computed, which successfully reproduced experimental log K_{ow} ranges of industrial CP mixtures. As an application of the developed approach, the predicted K_{aw} and K_{oa} were plotted to evaluate the bioaccumulation potential of each SCCP congener group.

Introduction

Chlorinated paraffins (CPs) are highly complex mixtures of polychlorinated *n*-alkanes with variable numbers of C and Cl atoms. Short-chain chlorinated paraffins (SCCPs, C_{10}–C_{13}) are considered persistent, bioaccumulative, and toxic and thus have been regulated under the Stockholm Convention since 2017.[1] Medium-chain CPs (MCCPs, C_{14}–C_{17}) and long-chain CPs (LCCPs, C_{18}+) are not under regulation at the present time, although concerns have been raised, particularly for MCCPs, whether regulation should be implemented for these longer CPs.[2] CP products contain a considerable number of congeners with different molecular structures. To date, analytical methods are not available that fully resolve individual congeners from mixtures.[3] Environmental assessments for bulk CP mixtures often use average properties. However, once diluted in the environment, each congener behaves individually following its own properties. As has been learned from decades of studies on other halogenated organic pollutants such as polychlorinated dibenzo-*p*-dioxins and polychlorinated biphenyls, environmental behavior and toxicity are often highly congener-specific. Indeed, broad bands of CP signals from chromatographic analysis[4] suggest that the properties of congeners differ substantially.

To address the environmental fate and toxicity of individual CP congeners, their partitioning properties need to be understood. Experimental determination of such properties is only possible for a handful of congeners because the availability of pure analytical standards is currently limited. Computational methods may be the only possibility to provide congener-specific information. Among such prediction models available, empirical fit models may not be useful, as congener-specific experimental data are not sufficiently available to calibrate such models.

This study applies the quantum chemically based COSMO-RS theory[5] to predict partition coefficients of CP congeners. COSMO-RS can predict partition coefficients from the molecular structure alone without any additional empirical parameter. This approach could address partition coefficients of CP congeners with differing structures even including stereoisomers. Previous studies show that COSMO-RS can predict partition coefficients for chemicals of diverse structures (but no CPs) to the accuracy of < 1 log unit root-mean squared errors (RMSE) as compared to experimental data, including chemicals with multifunctional structure.[6,7] Relative values across chemicals are expected to be even more accurate because systematic errors are canceled.[8]

The problem of using COSMO-RS for predicting a large number of chemicals is the computational time needed for the quantum chemical calculation and the conformer selection. For example, it takes several hours to generate COSMO files, necessary to calculate partition coefficients, just for a single (stereo)isomer of $C_{10}Cl_{10}$ using the supercomputer at the National Institute for Environmental Studies (HPE Apollo 2000, Intel Xeon Gold 6148 CPU, 40 CPU cores per each job). The computational time generally increases with the size of the molecule. Indeed, Glüge et al.[9] previously applied COSMO-RS to predict partition coefficients of CPs but provided predictions for only 4 structures per congener group, which are too few to address the variability of partition coefficients across congeners.

To enable the prediction of partition coefficients for hundreds of thousands of CP congeners, this study combines COSMO-RS with a fragment contribution model (FCM). An FCM counts the substructures (fragments) within the molecule and uses the fragment counts as descriptors for regression analysis. Such models have been widely adapted in the predictive model development of environmental properties.[10-13] FCMs are a linear model that can provide predictions with high speed and low electric energy consumption. In this work, we regress the COSMO-RS-predicted partition coefficients against CP's fragment counts to develop a model for predicting COSMO-RS predictions. Developing a model to predict the values that are output of another model might seem unmeaningful, because such a secondary model can only give less accurate predictions than the original model. However, such an approach is increasingly used in quantum chemistry applications where computational time is a hampering issue.[14] A secondary but fast-running fragment model could be useful particularly for CPs, and possibly other complex mixture components, because 1) experimental data for individual congeners are not available, 2) computation of the original model is too slow to cover the enormous number of congeners, and 3) the chemicals of concern are made up of relatively simple fragments and thus simple FCMs are expected to reproduce the predictions from a more sophisticated model well.

Methods

Method overview. COSMO-RS-based FCMs for the log of octanol–water (K_{ow}), air–water (K_{aw}), and octanol–air (K_{oa}) partition coefficients of SCCPs were developed by the following procedure. (1) The respective partition coefficients for a number of CP structures were calculated using the COSMO-RS method to generate training and validation sets. (2) FCMs with different combinations of fragments were calibrated using the training set. (3) Predictive performance of the calibrated FCMs was evaluated with the validation set. (4) Predictions by the FCMs were compared to available experimental data. (5) The FCMs were used to predict randomly generated SCCP congeners (1000 each for 52 congener groups) to demonstrate the variations of partition coefficients for SCCPs.

The CPs considered in this work are polychlorinated *n*-alkanes (i.e., no branching, no multiple bond). In this article, we refer to individual CP structures with different chain lengths and Cl-substitution patterns as "congeners". A "congener group" collectively denotes the congeners with the same number of C and Cl atoms (i.e., isomers). Isomers of CPs include stereoisomers that have the same two-dimensional molecular structure but are not superimposable in the three-dimensional space because of the presence of chiral centers.

COSMO-RS. COSMO-RS calculates the chemical potential of solute in solution from quantum mechanics and statistical thermodynamics calculations and can thereby predict thermodynamic properties including partition coefficients.[5] For a given stereochemically specific congener, the molecular structure in the SDF format was entered into the COSMO*confX* 4.3 software (COSMO*logic*), which selected optimal conformers and generated their COSMO files using quantum chemistry program TURBOMOL 7.3 (COSMO*logic*). These COSMO files were then used in COSMO*thermX* 19.0.4

(COSMO*logic*, parameterization: BP_TZVPD_FINE_19) to calculate K_{ow}, K_{aw}, and K_{oa} at 25°C. Here, we calculated K_{ow} with wet octanol and K_{oa} with dry octanol. Note that the version of COSMO*confX* used in this work sometimes returned structures that are stereochemically inconsistent with the original structure in the SDF (i.e., incorrect *R* or *S* configuration). This problem did not occur when we used the Windows version of COSMO*confX*, switched off RDKit, and used only Balloon to generate initial candidate conformers.

The whole calculation procedure from COSMO*confX* to COSMO*thermX* is consistent, although a slight difference in the calculated partition coefficient sometimes occurs when the initial input structure entered in COSMO*confX* is in a different conformational state. We examined the extent of this "random error" using 10 starting conformations each for three arbitrarily chosen C_{10} congeners. The standard deviations for log K_{ow}, log K_{aw}, and log K_{oa} were on average 0.02, 0.14, and 0.12, respectively. These differences may represent the current precision of COSMO*therm* predictions for CPs.

Generation of training and validation sets. In this work, we used "very" short to short-chain CPs (C_5–C_{10}) as training chemicals because computational time of the COSMO*confX* optimization procedure increases with the size of molecule. By opting for short CPs, we were able to generate many congeners for model training. The validation set, in contrast, should comprise congeners that are relevant. We chose C_{10}–C_{13}, thus SCCPs in this work. Calibrating and/or testing models for MCCPs and LCCPs would also be interesting but need much more time for calculations and was thus left for future work.

The training set consisted of 815 congeners—all 315 distinct isomers of C_5 CPs and 100 randomly generated isomers for each of C_6 to C_{10} CPs. In random generation, 0 to ($2m + 2$) H atoms of C_m-*n*-alkanes were randomly substituted with Cl atoms without any restriction. Here, all H atoms were considered distinct to also generate stereoisomers. Equivalent structures (i.e., superimposable by rotation) and enantiomers (i.e., mirror images) were removed because they show the identical partition coefficient value in reality, and COSMO-RS should give the same value in theory. Diastereomers, in contrast, can have different partitioning properties and thus are considered distinct congeners. The validation set consists of 120 CP congeners (30 for each of C_{10} to C_{13} CPs) that were also randomly generated. Codes were written in the *R* language[15] to create SMILES strings for all these congeners. SMILES was then converted to SDF format using OpenBabel,[16] which was then fed to COSMO*confX* as described above.

Fragment contribution models (FCMs). Fragments in CP structures were counted using *R* with ChemmineR (3.38.0) and ChemmineOB (1.24.0) packages from Bioconducter.[17] Fragments with differing carbon-chain lengths, namely C_1, C_2, C_3, and C_4 fragments were considered (Table S1 in the Supporting Information, SI). These fragments, respectively, have 7, 19, 64, and 220 types, out of which 0, 1, 10, and 67 types describe the diastereomeric patterns. All fragments and their SMARTS queries are given in Table S2. Using these fragments, four levels of models were generated. Level 1 model used only C_1 fragments, and Level 2 model included C_1 *and* C_2 fragments. Levels 3 and 4 were

calibrated with C_1 to C_3 and C_1 to C_4, respectively. Models were calibrated with least square multiple linear regression (MLR) with fragment counts as explanatory variables and COSMO*therm* predictions as dependent variables. Since fragment counts are not fully independent and the contributions of some (or many) fragments to the partition coefficient can be insignificant, a forward and backward stepwise algorithm was used to select model variables (i.e., fragments). Akaike's Information Criterion (AIC) was considered the model evaluation metric.[18] Variable selection was first performed for the Level 1 model. Then, the selected C_1 fragments were used as the initial variable set of the variable selection procedure for the Level 2 model, and so forth. To overcome a possible over-fitting problem, partial least squares regression (PLSR) was also performed using the selected Level 4 model fragments. The randomization test method was used to decide on the number of PLS components. All these statistical analyses were performed with *R* using functions such as lm(), step(), plsr(), and selectNcomp().

Predictions of partition coefficients for congener groups. Using the FCMs calibrated with PLSR, log K_{ow}, log K_{aw}, and log K_{oa} for 1000 randomly generated isomers for each SCCP congener group (C_{10}–C_{13}, Cl_2–Cl_{14}) were predicted. Two methods were adapted to generate random isomers. In the first method, all H atoms were considered available for Cl substitution at the same likelihood. Second, all H atoms were available, but each C atom was able to carry a maximum of only one Cl atom. In other words, the first method allows double or triple Cl substitution, while the second does not. For the second case, congeners with the number of Cl > the number of C cannot be generated. Also, if Cl = C, then there is only one constitutional isomer (but with many stereoisomers). As for random generation of training and validation sets explained above, all substitution positions along the carbon-chain were considered distinct to account for stereoisomers. Duplications were allowed for random generation of 1000 isomers; this matters the most for $C_{10}Cl_2$ group, which has only 30 constitutional isomers with 46 distinct structural isomers (i.e., 16 constitutional isomers have diastereomers). Duplication occurs increasingly rarely as the number of Cl approaches that of Cl. For example, 1000 random isomers of $C_{10}Cl_{10}$ had only 10 duplications and 14 enantiomer pairs.

We are aware that existing studies have shown that Cl substitution patterns are not random in commercial CP mixtures. A recent study suggested that the first, second, and third carbons from an end of the chain and central carbons all have differing likelihood of chlorination.[19] Also, it has been known that chlorination occurs less likely to the neighbors of the carbon that is already chlorinated due to a steric effect,[20,21] which is also inferred by GC retention measurements for CP mixtures.[4,22] Nevertheless, in highly chlorinated CP mixtures, dichloro-substituted carbons and trichloromethyl groups have also been identified.[19,23] Since general rules for positions of Cl for CPs of different lengths and chlorination degree are still under investigation, we opted for the fully random and "one Cl per C" rules to generate congener sets for this work.

Results and discussion

FCM training and validation. For all of log K_{ow}, log K_{aw}, and log K_{oa}, increasing the number of

183 fragments for MLR from Level 1 to Level 4 models improved the fitting quality, as indicated by R^2, root
184 mean squared errors (RMSE), and AIC (Figures 1, S1, S2, Table S3). Hence, Level 4 model resulted in
185 the best fit. It is interesting that C_4 fragments do have statistically significant contributions to the
186 partition coefficients, suggesting that the molecular interaction properties of CPs cannot fully be
187 reduced to the shorter fragments, at least in the COSMO-RS calculation. In the variable selection
188 procedure, about half (49-61%) of the total fragments were removed for Level 2 to 4 models. This is
189 not surprising, because many fragments share common substructures and thus are interrelated. PLSR
190 with the Level 4 fragments resulted in a similar fitting quality as compared to the least square MLR,
191 although the PLSR has more restrictions (i.e., a lower degree of freedom) when deriving fitting
192 coefficients. The good fit indicated by low RMSE (0.05, 0.12, and 0.09 for log K_{ow}, log K_{aw}, and log K_{oa},
193 respectively) with the Level 4 model or its PLSR version show that FCMs can accurately fit
194 COSMO*therm* calculated values for CPs. These RMSE values are similar to the precision of
195 COSMO*therm* for CPs mentioned in the Methods section and may thus be considered the best
196 achievable fit. All resulting fitting coefficients are presented in Table S4. We note that some fragments
197 that describe diastereometric structures were also significant.
198 External validation leads to the same conclusions as above. Thus, the Level 4 model showed
199 the best statistics, and the statistics were better in order of log K_{ow}, log K_{oa}, and log K_{aw}. (Figures 1 and
200 S3, Table S3). PLSR and the Level 4 model predicted the validation set equally well. While PLSR
201 typically is considered more robust when the number of variables is large, it was just similar to the
202 least square MLR-based Level 4 model. RMSE was 0.12, 0.29, and 0.21 for log K_{ow}, log K_{aw}, and log K_{oa},
203 respectively, being 2.2–2.4 times higher than RMSE for fitting. Use of the calibrated FCMs thus causes
204 additional prediction errors of 0.1 to 0.3 RMSE in log *K*'s of SCCPs, as compared to the direct use of
205 COSMO*therm*.
206

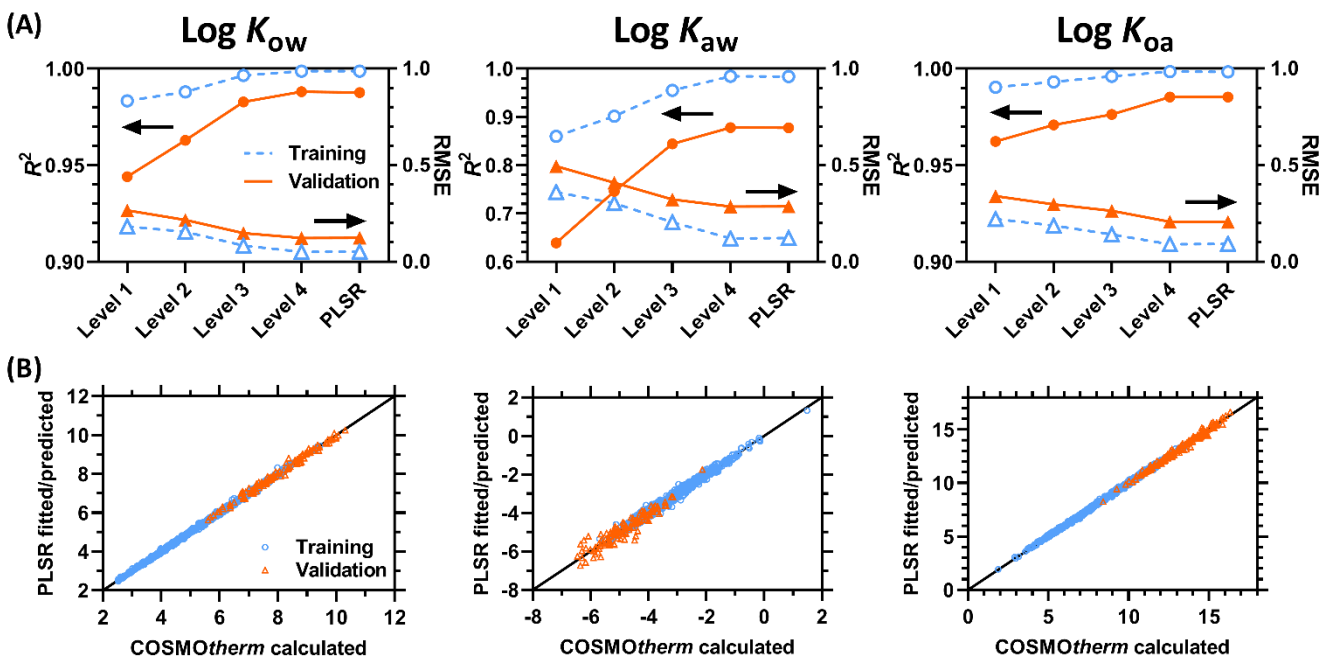


Figure 1. Statistics of model fitting and validation (A) and comparison of FCM (PLSR)-calculated values to training and validation data (B). Arrows indicate the axes that the data points refer to. Larger figures (Figures S1–S3) and a table with statistics for all models (Table S3) are presented in the SI.

Fragment contributions to log *K*'s. The fact that the level 4 model performs the best suggests that the actual contribution of each C type (e.g., -CH_2-, -CHCl-) to log *K*'s depends on its neighboring structure. Nevertheless, lower level models may also be useful to illustrate the average contributions of the C types to log *K*'s. For instance, the Level 1 model (with only C_1 fragments) shows that the fragment contributions to log K_{ow} and log K_{oa} are fairly systematic (Figure S4): The -CH_2- increment increases log K_{ow} and log K_{oa}, while substituting H of -CH_2- or CH_3- with Cl also increases log K_{ow} and log K_{oa}. In contrast, the fragment contributions to log K_{aw} are irregular. Substituting one H in -CH_2- with Cl to form -CHCl- decreases log K_{aw}, but further substitution to -CCl_2- would not change log K_{aw}. Similarly, Cl-substitution of CH_3- to CH_2Cl- decreases log K_{aw}, but further substitution to $CHCl_2$- has no influence, and that to CCl_3- even increases log K_{aw}.

Comparison to experimental data. There are some experimentally determined log K_{ow} and log K_{aw} for specific constitutional isomers in the literature. The predictions by the FCM (PLSR-calibrated) agree with the literature data for log K_{ow} within 1 log unit difference (Figure 2). The FCM tends to overpredict log K_{ow} of CP congeners with five or more chlorinated C atoms. The predictions by the original COSMO*therm* deviate from the experimental data to a similar extent. Thus, the observed overpredictions for some log K_{ow} data should be related to the inaccuracy in the original COSMO*therm* calculations or due to experimental errors. The cited experimental data were derived from HPLC retention measurements using 10 hydrophobic aromatic compounds as calibration

compounds. While this is a standard approach, the resulting K_{ow} may not be as accurate for aliphatic chemicals with appreciable polarity like CPs as for hydrophobic aromatic compounds.

The two congener-specific experimental log K_{aw} are underpredicted by the FCM by 1–1.5 log units. The original COSMO*therm* predictions agree better with the experimental data in this case. As Figure 2 shows, the predictions for K_{aw} by FCM and COSMO*therm* differ by ca 1 log unit, which is close to the maximal error (0.94 log units) found in the model validation presented above. The reason for the model disagreement specifically for 1,2,9,10-$C_{10}Cl_4$ and 1,2,10,11-$C_{10}Cl_4$ is unknown, but we speculate that an extended sequence of non-Cl-substituted -CH_2- units rarely occurs in our random isomer generation for the training set and thus may have been under-represented in the model training. Indeed, our training set contained only two congeners with a -CH_2-CH_2-CH_2-CH_2- fragment and none with a longer CH_2 chain. That said, the experimental K_{aw} value for 1,2,9,10-$C_{10}Cl_4$ could also be somewhat too high, as log K_{aw} of −2 in combination with log K_{ow} of 5 would result in log K_{oa} of 7 (via $K_{oa} = K_{ow}/K_{aw}$) and this is even smaller than an experimentally measured log hexadecane–air partition coefficient (L) of 8.4 for 1,2,9,10-$C_{10}Cl_4$.[22] $K_{oa} \gtrsim L$ is generally expected because CPs interact with octanol via additional polar interactions that do not occur with apolar hexadecane.

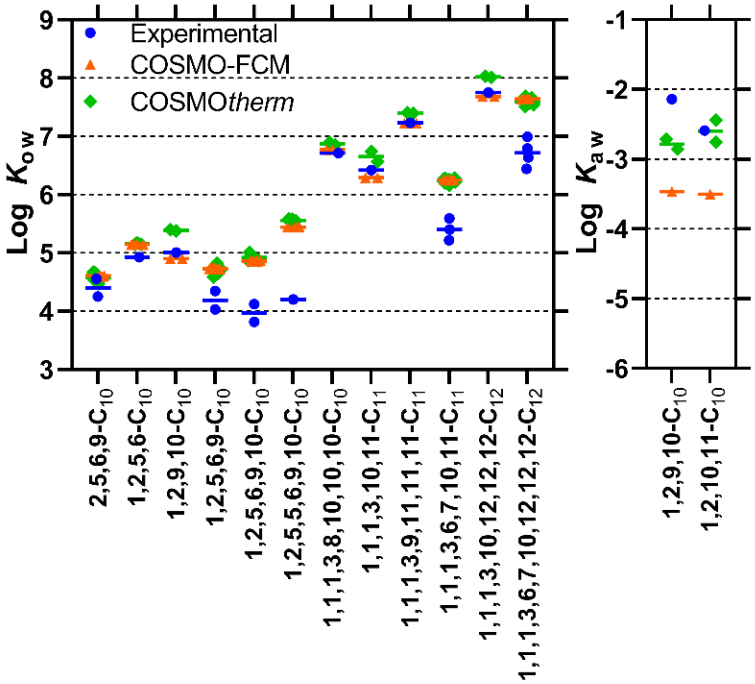


Figure 2. Comparison of predicted with experimental K values. Predictions were derived from COSMO-RS based FCMs (with PLSR calibration) as well as directly from COSMO*therm* software. K_{ow} data are from Hilger et al.[24] K_{aw} data are from Drouillard et al.[25] for 23°C. There are multiple data points both for predictions and experimental data because of the presence of diastereomers. Bars indicate the mean.

Distributions of K_{ow}, K_{aw}, and K_{oa} for SCCP congener groups. Using the FCM based on PLSR calibration, log *K*'s for 1000 isomers per congener group were predicted. These predictions were used to estimate the distributions of K_{ow}, K_{aw}, and K_{oa} for SCCP congener groups (Figures 3, S5; here double/triple Cl is not allowed). The 2.5, 25, 50 (median), 75, and 97.5%iles of log *K*'s for each SCCP congener group are presented in Tables S5.

Log K_{ow} and log K_{oa} for each congener group are within a relatively narrow range (1 to 2 log units), whereas log K_{aw} for each congener group spreads over 1.5 to 3 log units. The median log K_{ow} values of different SCCP congener groups range over 4 log units (5.0 to 8.9) and log K_{aw} also over 4 log units (−5.7 to −1.6), whereas the median log K_{oa} spans over 8 log units (6.7 to 14.8).

It is interesting that the medians of log K_{ow}, K_{aw}, and K_{oa} show different dependence on the numbers of C and Cl. All three log *K*'s are linearly dependent on the number of C, although the slopes differ depending on the partitioning phases and partially on the number of Cl (Figure S6). In contrast, dependence on the number of Cl is nonlinear (Figure 3; more clearly in Figure S7). Log K_{ow} is fairly constant from Cl_2 to ~Cl_5, above which it increases with ca 0.35 log units/C. Log K_{aw} has the opposite trend; it decreases from Cl_2 to ~Cl_{10} by 2.5–3.5 log units and thereafter stays nearly the same. Log K_{oa} monotonically increases but in a concave downward shape. The increase is ca 0.8 log unit/Cl from Cl_2 to Cl_3 whereas only 0.4 log units/Cl from C_{13} to C_{14}.

We also derived the distributions of partition coefficients for CP congeners without double and triple Cl substitutions (Figure S8, Table S6). The distribution peaks are sometimes slightly sharper, but overall, the results are just similar to the distributions of CP congeners including double/triple Cl substitutions. The median for each congener group is different by only 0.13 log unit on average and by 0.39 at most. The similarity between the cases with and without double/triple Cl is expected for low-chlorinated congeners (Cl_2–Cl_4), because random generation forms limited numbers of double and triple substitutions, even if allowed. The similarity for higher chlorinated congeners is interesting and suggests that this difference is not important for partition coefficients when the congener groups are considered as a whole.

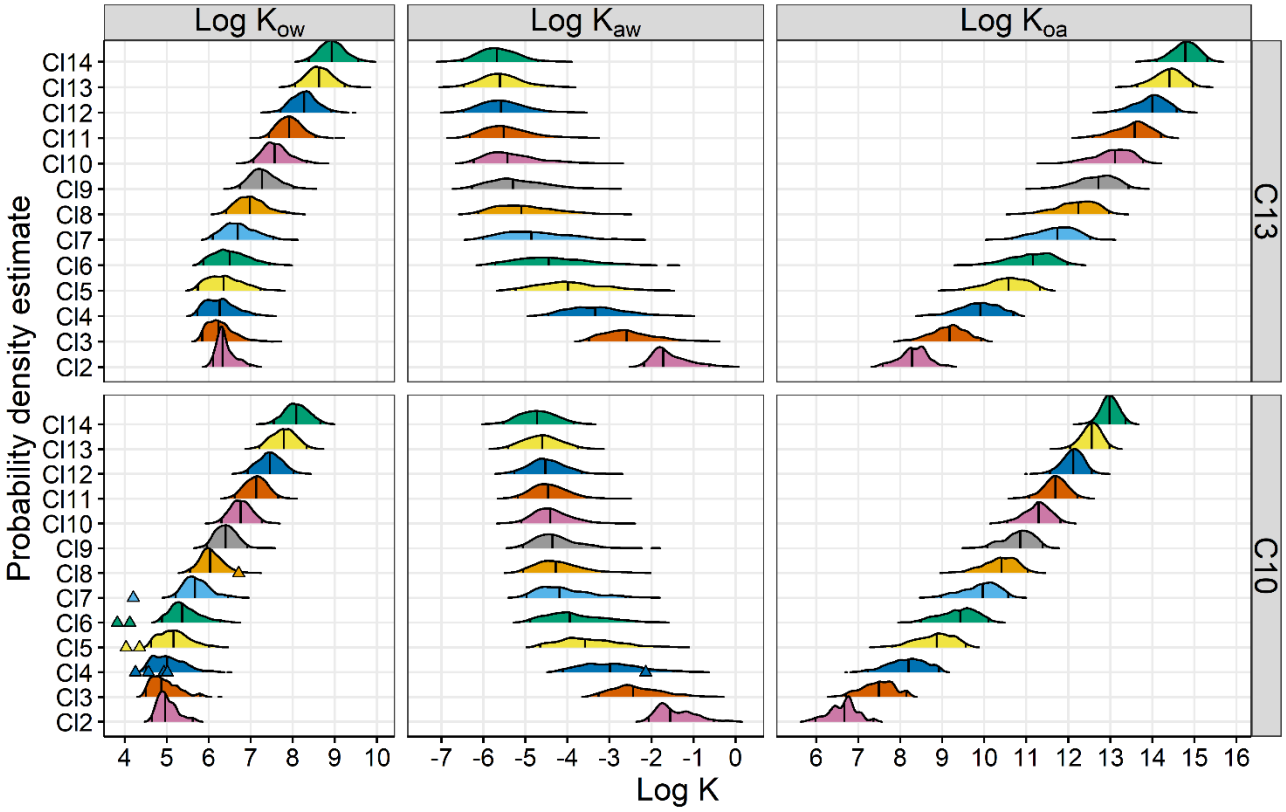


Figure 3. Kernel density estimates resulting from 1000 structures (with double and triple Cl substitution allowed) for each molecular formula (predicted by PLSR model). Vertical lines indicate the 2.5, 50 (median), and 97.5%iles. Data points are experimental data from Hilger et al.[24] and Drouillard et al.[25] for specific isomers. Plots with more congener groups and without double and triple Cl substitutions are shown in Figures S5 and S8.

Predicting log K_{ow} of SCCP mixtures. The log K_{ow} distributions predicted above for all relevant SCCP congener groups were used in combination with the compositions of SCCP mixtures experimentally derived from Yuan et al.[26] to predict log K_{ow} ranges of bulk CP mixtures (Figure 4; more plots in Figures S9, S10). Here, the predicted log K_{ow} distributions for congener groups were weighted by their relative abundance (i.e., mole fractions) in the mixture and were then summed. The results agreed with the experimental data from Renberg et al.,[27] who used retentions on thin layer chromatography to estimate the ranges of log K_{ow} for CP mixtures. The lower bound of the experimental data agrees with the predicted 2.5%ile and the upper bound with the predicted 97.5%ile within 0.84 log units (Table S7). These results serve as additional validation of COSMO*therm* and the FCMs for predicting log K_{ow} of CPs.

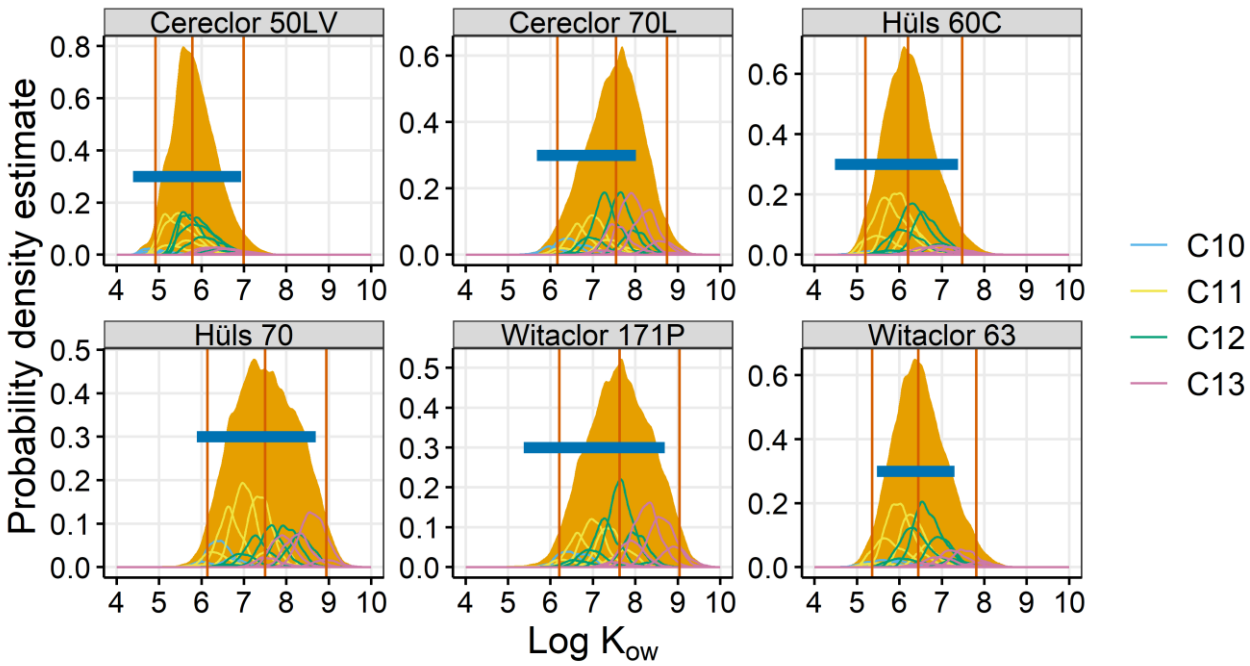


Figure 4. Comparison of predicted distributions of log K_{ow} for CP mixtures (filled curves) and experimental data from Renberg et al.[27] (horizontal bars). The 2.5, 50, and 97.5%iles of the predications for mixtures (vertical lines) and the predictions for each congener groups (unfilled curves) are also shown. The predictions were derived from the FCMs (PLSR, double/triple Cl allowed).

Implications. This study presents, for the first time, a time-efficient method to predict partition coefficients for a large number of CP congeners on the basis of quantum chemical calculations. We provided the ranges of partition coefficients for CP congener groups and bulk CP mixtures, grounded on the predictions for individual congeners. These new pieces of information should improve our understanding on the environmental fate of CPs. As an example, SCCP congeners were plotted in the chemical space that indicates the Arctic bioaccumulation potential using predicted log K_{aw} and log K_{oa}, following the approach by Czub et al.[28] and Brown and Wania[29] (Figure 5; plots for each SCCP congener group is shown in Figure S10). Figure 5 shows that relatively low chlorinated (Cl_2–Cl_6) SCCPs fall into the chemical space where high Arctic bioaccumulation is expected, assuming perfect persistence. In contrast, SCCPs with relatively high molecular weight (C + Cl ≥ 20) do not fall in this zone. Previously, Gawor and Wania[30] presented various chemical space plots for CPs using log K_{aw} and log K_{oa} predicted by ACD/ADME Suite prediction tools and came up with conclusions that are in part similar to those in this work. It would be interesting to repeat their analysis with the predicted partition coefficients from this work, which is however beyond the scope of this article.

The presented approach of course has room for further improvement. First, the current FCMs have been calibrated with relatively short CPs (≤ C_{10}) and thus would have to be extrapolated for M/LCCPs. Extension to M/LCCPs requires lengthy COSMO*confX* calculations for long molecules, which

322 will be conducted as a next step. Second, while the current study demonstrated good model
323 predictions for log K_{ow} of a dozen of individual constitutional isomers and six bulk SCCP mixtures,
324 validation for more of specific (stereo)isomers and other partition phases would be desirable. This
325 statement applies both to the original COSMO-RS approach and the FCMs presented here. Availability
326 of isomer-specific CP standards is being improved and more data are expected in the future (e.g., ref
327 22). Third, the current work used randomly generated congeners from all Cl substitution patterns or
328 excluding double and triple Cl substitutions to represent the congener composition of each CP
329 congener group, but this is a first approximation. As more and more knowledge regarding Cl
330 substitution patterns in the bulk CP mixtures is becoming available,[19,23,31] congener compositions
331 used for the prediction of partition coefficients could be elaborated further.
332 This study demonstrated that the approach that combines COSMO-RS and FCM methods can
333 provide accurate predictions for SCCP partitioning coefficients. As the most time-consuming
334 COSMO*confX* step that generates COSMO files has been completed for a number of CP congeners, it
335 is possible to run COSMO*therm* and derive new FCMs quickly for other partition coefficients or other
336 properties of CPs that are related to the chemical potential in solvent. Our approach may be useful
337 for other highly complex mixtures as well, as partitioning properties of complex mixtures are generally
338 difficult to determine both experimentally and computationally.
339

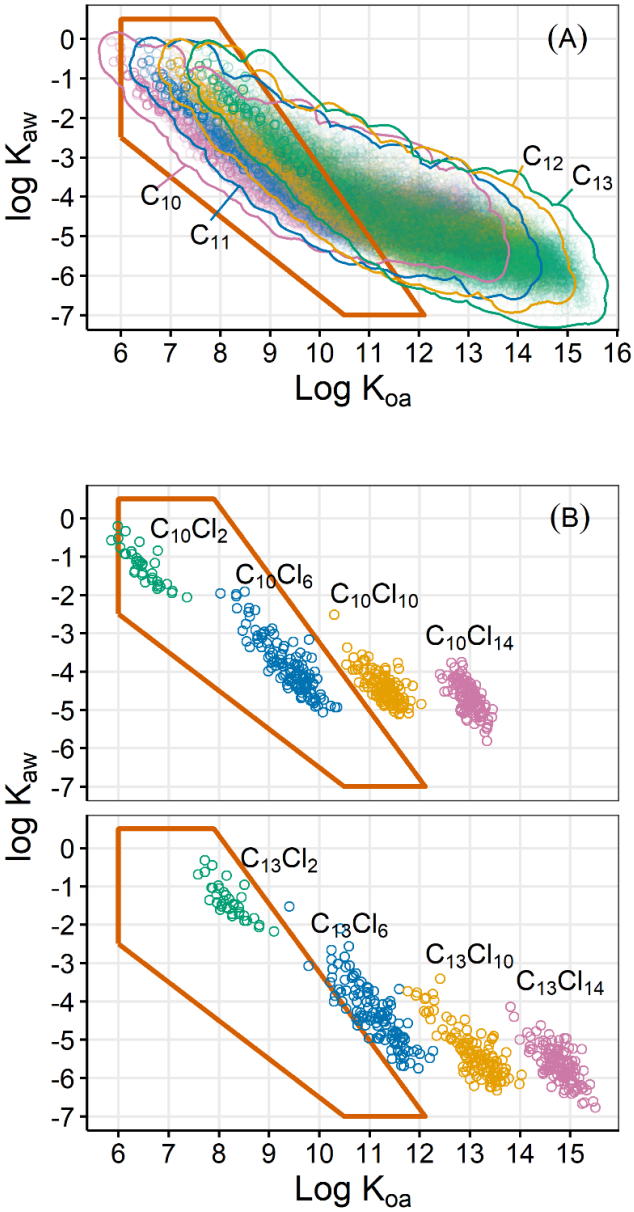


Figure 5. Chemical space plots for (A) all and (B) selected SCCP congener groups. The chemical space for a high Arctic contamination and bioaccumulation potential AC-BAP (>10%, 70 days)[28] was enclosed with lines, as in ref 29.

Associated Content

Supporting Information

Additional figures and tables for the results of model fitting and validation, model predictions for log *K*'s, distributions of log K_{ow}, and chemical space plots. This material is available free of charge via the Internet at http://pubs.acs.org.

Author Information

Corresponding author
Satoshi Endo
Phone/Fax: ++81-29-850-2695
endo.satoshi@nies.go.jp
ORCID: 0000-0001-8702-1602

Notes

The authors declare no competing financial interest.

Author contributions

Study design: SE. COSMO-RS calculations: SE, JH. Statistical analysis: SE. Data evaluation: SE. Drafting of manuscript: SE. Revising of manuscript: SE, JH.

Acknowledgements

We thank Frank Wania, Trevor Brown, and Kai-Uwe Goss for their valuable comments on the manuscript. COSMO*confX* and TURBOMOL calculations were performed with the NIES supercomputer system. This work was supported by the Environment Research and Technology Development Fund (SII-3-1) of the Environmental Restoration and Conservation Agency, Japan.

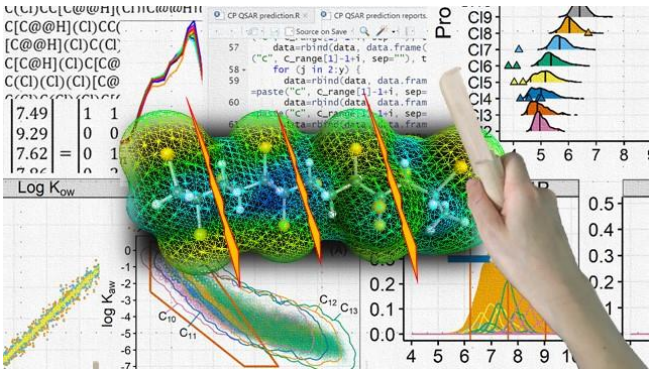
References

1. UNEP, Decision SC-8/11: Listing of short-chain chlorinated paraffins. 2017, UNEP/POPS/COP.8/SC-8/11.
2. Glüge, J.; Schinkel, L.; Hungerbuhler, K.; Cariou, R.; Bogdal, C., Environmental risks of medium-chain chlorinated paraffins (MCCPs): A Review. *Environ. Sci. Technol.* **2018,** *52*, (12), 6743-6760.
3. van Mourik, L. M.; Lava, R.; O'Brien, J.; Leonards, P. E. G.; de Boer, J.; Ricci, M., The underlying challenges that arise when analysing short-chain chlorinated paraffins in environmental matrices. *J. Chromatogr. A* **2020,** *1610*, 460550.
4. Korytar, P.; Parera, J.; Leonards, P. E.; Santos, F. J.; de Boer, J.; Brinkman, U. A., Characterization of polychlorinated *n*-alkanes using comprehensive two-dimensional gas chromatography--electron-capture negative ionisation time-of-flight mass spectrometry. *J. Chromatogr. A* **2005,** *1086*, (1-2), 71-82.
5. Klamt, A., 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.
6. Stenzel, A.; Goss, K.-U.; Endo, S., Prediction of partition coefficients for complex environmental contaminants: Validation of COSMO*therm*, ABSOLV, and SPARC. *Environ. Toxicol. Chem.* **2014,** *33*, (7), 1537-1543.

390 7. Loschen, C.; Reinisch, J.; Klamt, A., COSMO-RS based predictions for the SAMPL6 logP
391 challenge. *J. Comput. Aided Mol. Des.* **2020,** *34*, (4), 385-392.
392 8. Goss, K.-U.; Arp, H. P. H.; Bronner, G.; Niederer, C., Partition behavior of
393 hexachlorocyclohexane isomers. *J. Chem. Eng. Data* **2008,** *53*, (3), 750-754.
394 9. Glüge, J.; Bogdal, C.; Scheringer, M.; Buser, A. M.; Hungerbühler, K., Calculation of
395 physicochemical properties for short- and medium-chain chlorinated paraffins. *J. Phys.*
396 *Chem. Ref. Data* **2013,** *42*, (2), 023103.
397 10. Meylan, W. M.; Howard, P. H., Atom/fragment contribution method for estimating octanol-
398 water partition coefficients. *J. Pharm. Sci.* **1995,** *84*, (1), 83-92.
399 11. Brown, T. N.; Arnot, J. A.; Wania, F., Iterative fragment selection: a group contribution
400 approach to predicting fish biotransformation half-lives. *Environ. Sci. Technol.* **2012,** *46*, (15),
401 8253-8260.
402 12. OECD, Guidance document on the validation of (quantitative) structure-activity relationships
403 [(Q)SAR] models. 2007, ENV/JM/MONO(2007)2.
404 13. Dearden, J. C.; Cronin, M. T.; Kaiser, K. L., How not to develop a quantitative structure-activity
405 or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009,** *20*, (3-4),
406 241-266.
407 14. Ramakrishnan, R.; von Lilienfeld, O. A., Machine learning, quantum chemistry, and chemical
408 space. In *Reviews in Computational Chemistry, Vol 30*, Parrill, A. L.; Lipkowitz, K. B., Eds.
409 Wiley-Blackwell: Malden, 2017; Vol. 30, pp 225-256.
410 15. *R* Core Team, *R: A language and environment for statistical computing. R* Foundation for
411 Statistical Computing, Vienna, Austria. https://www.R-project.org/. 2019.
412 16. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R.,
413 Open Babel: An open chemical toolbox. *J. Cheminformatics* **2011,** *3*, (1), 33.
414 17. Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, T.; Girke, T., ChemmineR: a compound mining
415 framework for *R*. *Bioinformatics* **2008,** *24*, (15), 1733-1734.
416 18. Burnham, K. P.; Anderson, D. R., Multimodel Inference. *Sociol. Methods Res.* **2016,** *33*, (2),
417 261-304.
418 19. Yuan, B.; Lysak, D. H.; Soong, R.; Haddad, A.; Hisatsune, A.; Moser, A.; Golotvin, S.;
419 Argyropoulos, D.; Simpson, A. J.; Muir, D. C. G., Chlorines are not evenly substituted in
420 chlorinated paraffins: A predicted nmr pattern matching framework for isomeric
421 discrimination in complex contaminant mixtures. *Environ. Sci. Technol. Letters* **2020**.
422 doi.org/10.1021/acs.estlett.0c00244.
423 20. Fredricks, P. S.; Tedder, J. M., 28. Free-radical substitution in aliphatic compounds. Part II.
424 Halogenation of the *n*-butyl halides. *J. Chem. Soc. (Resumed)* **1960**, 144-150.
425 21. Colebourne, N.; Stern, E. S., 660. The chlorination of some *n*-alkanes and alkyl chlorides. *J.*
426 *Chem. Soc. (Resumed)* **1965**, 3599-3605.
427 22. Hammer, J.; Matsukami, H.; Endo, S., Congener-specific partition properties of chlorinated

428 paraffins evaluated with COSMO*therm* and GC-retention indices. (Manuscript in
429 preparation).
430 23. Sprengel, J.; Wiedmaier-Czerny, N.; Vetter, W., Characterization of single chain length
431 chlorinated paraffin mixtures with nuclear magnetic resonance spectroscopy (NMR).
432 *Chemosphere* **2019,** *228*, 762-768.
433 24. Hilger, B.; Fromme, H.; Volkel, W.; Coelhan, M., Effects of chain length, chlorination degree,
434 and structure on the octanol-water partition coefficients of polychlorinated *n*-alkanes.
435 *Environ Sci Technol* **2011,** *45*, (7), 2842-9.
436 25. Drouillard, K. G.; Tomy, G. T.; Muir, D. C. G.; Friesen, K. J., Volatility of chlorinated n-alkanes
437 (C-10-C-12): Vapor pressures and Henry's law constants. *Environ. Toxicol. Chem.* **1998,** *17*,
438 (7), 1252-1260.
439 26. Yuan, B.; Bogdal, C.; Berger, U.; MacLeod, M.; Gebbink, W. A.; Alsberg, T.; de Wit, C. A.,
440 Quantifying short-chain chlorinated paraffin congener groups. *Environ. Sci. Technol.* **2017,**
441 *51*, (18), 10633-10641.
442 27. Renberg, L.; Sundstrom, G.; Sundhnygard, K., Partition coefficients of organic chemicals
443 derived from reversed phase thin-layer chromatography: Evaluation of methods and
444 application on phosphate esters, polychlorinated paraffins and some PCB-substitutes.
445 *Chemosphere* **1980,** *9*, (11), 683-691.
446 28. Czub, G.; Wania, F.; McLachlan, M. S., Combining long-range transport and bioaccumulation
447 considerations to identify potential Arctic contaminants. *Environ. Sci. Technol.* **2008,** *42*, (10),
448 3704-3709.
449 29. Brown, T. N.; Wania, F., Screening chemicals for the potential to be persistent organic
450 pollutants: A case study of Arctic contaminants. *Environ. Sci. Technol.* **2008,** *42*, (14), 5202-
451 5209.
452 30. Gawor, A.; Wania, F., Using quantitative structural property relationships, chemical fate
453 models, and the chemical partitioning space to investigate the potential for long range
454 transport and bioaccumulation of complex halogenated chemical mixtures. *Environ. Sci.*
455 *Process Impacts* **2013,** *15*, (9), 1671-1684.
456 31. Jensen, S. R.; Brown, W. A.; Heath, E.; Cooper, D. G., Characterization of polychlorinated
457 alkane mixtures-A Monte Carlo modeling approach. *Biodegradation* **2007,** *18*, (6), 703-717.
458
459

TOC


CP QSAR prediction.R
CP QSAR prediction reports.
Source on Save
data=rbind(data, data.frame(
("C", C_range[1]-1+i, sep=""), t
for (j in 2:y) {
data=rbind(data, data.fram
=paste("C", C_range[1]-1+i, sep=
data=rbind(data, data.fram
Log K_OW
log K_aw
C10
C11
C12
C13
Cl9
Cl8
Cl7
Cl6
Cl5
Cl4
Cl3
7.49
9.29
7.62