Predicting Partition Coefficients of Short-Chain Chlorinated Paraffin Congeners by COSMO-RS-Trained Fragment Contribution Models

Satoshi Endo, Jort Hammer

Submitted date: 27/09/2020 • Posted date: 28/09/2020
Licence: CC BY-NC-ND 4.0

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 uses a sophisticated but time-demanding quantum chemically based method COSMO-RS and a fast-running fragment contribution approach to enable prediction of partition coefficients for a large number of short-chain chlorinated paraffin (SCCP) congeners. Fragment contribution models (FCMs) were developed using molecular fragments with a length of up to \( C_4 \) in CP molecules as explanatory variables and COSMO-RS-calculated partition coefficients as training data. The resulting FCMs 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. The FCM predictions for \( K_{ow} \) agree with experimental values for individual constitutional isomers within 1 log unit. The distribution of partition coefficients for each SCCP congener group was computed, which successfully reproduced experimental log \( K_{ow} \) ranges of industrial CP mixtures. As an application of the developed FCMs, the predicted \( K_{aw} \) and \( K_{oa} \) were plotted to evaluate the bioaccumulation potential of each SCCP congener group.

File list (3)

| File Name | Size     | View on ChemRxiv | Download file |
|-----------|----------|------------------|---------------|
| Endo_SCCPpartitioning_SI_20200926.pdf | 2.72 MiB | view on ChemRxiv | download file |
| Endo_SCCPpartitioning 20200926.pdf | 1.27 MiB | view on ChemRxiv | download file |
| SCCPpartitioning FCM Data.xlsx | 53.79 KiB | view on ChemRxiv | download file |
Supporting Information for

“Predicting partition coefficients of short-chain chlorinated paraffin congeners by COSMO-RS-trained fragment contribution models”

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

Contents
Figure S1. Fragment model fitting to the training set.
Figure S2. Residual plots for fragment model fitting to the training set.
Figure S3. Fragment contributions of (substituted) methylene groups to log $K$’s in the Level 1 model.
Figure S4. Fragment model predictions for the validation set.
Figure S5. COSMOfrag predictions for the validation set.
Figure S6. Kernel density estimates of log $K$ predicted by PLSR model.
Figure S7. Dependence of the median log $K$’s for SCCP congener groups on the numbers of C.
Figure S8. Dependence of the median log $K$’s for SCCP congener groups on the numbers of Cl.
Figure S9. Distributions of log $K_{ow}$ for CP mixtures (I).
Figure S10. Distributions of log $K_{ow}$ for CP mixtures (II).
Figure S11. Chemical space plot.
Table S1. Types of fragments used in the model development.
Table S2. Results of model training and validation.
Table S3. Quantiles of log $K$’s for each congener group (predicted by PLSR model, with double/triple Cl substitutions allowed).
Table S4. Quantiles of log $K$’s for each congener group (predicted by PLSR model, with double/triple Cl substitutions not allowed).
Table S5. Predicted and experimental log $K_{ow}$ ranges for CP mixtures. Experimental data are from Renberg et al.
Figure S1. Fragment model fitting to the training set.

Figure S2. Residual plots for fragment model fitting to the training set.
Figure S4. Fragment contributions of (substituted) methylene groups to log $K$'s in the Level 1 model. In the plot, coefficients for negative log $K_{aw}$ (thus log $K_{wa}$) are shown for a comparison purpose. Error bars indicate standard errors.

Addition description to Figure S4. 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 of -CH$_2$-, -CHCl-, and -CCl$_2$- to log $K_{ow}$ are 0.35, 0.58, and 0.93, respectively, with systematic increase with Cl (Figure S4). The contributions to log $K_{oa}$ are also systematic (0.58, 1.05, and 1.46, respectively). In contrast, the fragment contributions to log $K_{aw}$ are irregular (−0.23, −0.46, and −0.49, respectively). Thus, 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 $K_{aw}$. 
Figure S4. Fragment model predictions for the validation set.

Figure S5. COSMOfrag predictions for the validation set. Plots for the default and updated COSMOfrag databases (DB) are shown.
Figure S6. Kernel density estimates of log $K$ predicted by PLSR model. For each molecular formula, 1000 structures were considered. “All” indicates the results of random generation of isomers with double and triple Cl substitution allowed, while “1 Cl/C” with only one Cl substitution per C at maximum. Vertical lines indicate the 2.5, 50 (median), and 97.5%iles. Data points are the minimum and maximum of COSMOtherm predictions from Glüge et al.$^3$
Figure S7. Dependence of the median log $K$'s for SCCP congener groups on the numbers of C.
Figure S8. Dependence of the median log $K$'s for SCCP congener groups on the numbers of Cl.
Figure S9. Distributions of log $K_{ow}$ for CP mixtures (I). Blue bars indicate experimental data from Renberg (TLC), where available. Curves show the predictions from the FCMs (PLSR, double/triple Cl allowed). Congener groups are displayed with different line types (e.g., dashed, dotted). SCCP mix 1, 2, 3, and 4 are, respectively, the following mixtures considered by Yuan et al.: SCCP 51.5 %Cl + C\textsubscript{13} 50.23 %Cl, SCCP 55.5 %Cl + C\textsubscript{13} 55.03 %Cl, SCCP 63.0 %Cl + C\textsubscript{11} 65.25 %Cl, SCCP 63.0 %Cl + C\textsubscript{12} 69.98 %Cl.
Figure S10. Distributions of log $K_{ow}$ for CP mixtures (II). The same plot as Figure S9, except that the FCM-predicted distributions are based on the “one Cl per C” rule (i.e., double/triple Cl NOT allowed). For congeners with the number of Cl > that of C, however, double/triple Cl is allowed.
Figure S11. Chemical space plot. The chemical space for a high Arctic contamination and bioaccumulation potential AC-BAP (>10%, 70 days)\textsuperscript{1} was enclosed with lines, as in ref 2.
### Table S1. Types of fragments used in the model development.

| Fragment type | Example (terminal) | Example (middle) | Total number |
|---------------|--------------------|------------------|--------------|
| C₁            | -CH₃, -CH₂Cl       | -CH₂-, -CHCl-    | 7            |
|               |                    | -CH₂-CH₂-, -CHCl-CHCl- | 19(1)      |
| C₂            | -CH₂-CH₃, -CCl₂-CH₂Cl | -CH₂-CH₂-CH₂-, -CHCl-CHCl- | 64(10)    |
| C₃            | -CH₂-CH₂-CH₂-CH₃, -CHCl-CCl₂-CH₂Cl | -CH₂-CH₂-CH₂-CH₂-, -CHCl-CHCl-CCl₂- | 220(67)  |
| C₄            | -CH₂-CH₂-CH₂-CH₃, -CHCl-CCl₂-CHCl-CCl₃ | -CH₂-CH₂-CH₂-CH₂-, -CHCl-CH₂-CCl₂- | 220(67)  |

**Model**

| Level       | Fragments used |
|-------------|----------------|
| Level 1     | C₁             |
| Level 2     | C₁, C₂         |
| Level 3     | C₁, C₂, C₃     |
| Level 4     | C₁, C₂, C₃, C₄ |
| PLSR        | (selected fragments for the Level 4 model) |

The number in parentheses indicates the number of fragments that describe diastereomers.
| Level | $R^2$  | RMSE  | AIC  | Number of variables | $R^2$  | RMSE  |
|-------|--------|-------|------|---------------------|--------|-------|
| **log $K_{ow}$** |        |       |      |                     |        |       |
| Level 1 | 0.9833 | 0.183 | -442 | 6                   | 0.9441 | 0.266 |
| Level 2 | 0.9879 | 0.155 | -694 | 13                  | 0.9629 | 0.216 |
| Level 3 | 0.9965 | 0.083 | -1641| 46                  | 0.9828 | 0.148 |
| Level 4 | 0.9987 | 0.051 | -2278| 123                 | 0.9881 | 0.123 |
| PLSR   | 0.9986 | 0.053 | *    | 123(22)             | 0.9876 | 0.125 |
| COSMOfrag, default DB |        |       |      |                     | 0.8179 | 1.308 |
| COSMOfrag, updated DB |        |       |      |                     | 0.9427 | 0.265 |

| **log $K_{aw}$** |        |       |      |                     |        |       |
| Level 1 | 0.8597 | 0.361 | 667  | 5                   | 0.6389 | 0.494 |
| Level 2 | 0.9015 | 0.303 | 395  | 13                  | 0.7456 | 0.410 |
| Level 3 | 0.9548 | 0.205 | -192 | 37                  | 0.8440 | 0.322 |
| Level 4 | 0.9843 | 0.121 | -876 | 125                 | 0.8780 | 0.286 |
| PLSR   | 0.9837 | 0.123 | *    | 125(17)             | 0.8775 | 0.287 |
| COSMOfrag, default DB |        |       |      |                     | 0.3308 | 2.194 |
| COSMOfrag, updated DB |        |       |      |                     | 0.5095 | 0.709 |

| **log $K_{oa}$** |        |       |      |                     |        |       |
| Level 1 | 0.9904 | 0.221 | -134 | 6                   | 0.9623 | 0.340 |
| Level 2 | 0.9931 | 0.187 | -386 | 13                  | 0.9709 | 0.297 |
| Level 3 | 0.9960 | 0.142 | -788 | 39                  | 0.9763 | 0.263 |
| Level 4 | 0.9984 | 0.090 | -1361| 122                 | 0.9853 | 0.207 |
| PLSR   | 0.9983 | 0.093 | *    | 122(16)             | 0.9853 | 0.206 |
| COSMOfrag, default DB |        |       |      |                     | 0.7628 | 3.238 |
| COSMOfrag, updated DB |        |       |      |                     | 0.8204 | 0.721 |

$R^2$ is the coefficient of determination (not adjusted). RMSE, root mean squared error; AIC, Akaike's Information Criterion; Values in parentheses indicate the number of PLSR components selected for the final model. *Not applicable. For COSMOfrag (default DB), $n = 68$ (see the main text).
| Quantiles | log Kon | log Kaw | log Koa |
|-----------|---------|---------|---------|
| 0.025     | 0.25    | 0.5     | 0.75    | 0.975   | 0.025   | 0.25    | 0.5     |
| 1.00      | 1.30    | 1.91    | 2.02    | 2.18    | 2.47    | 3.00    | 3.50    |
| 5.00      | 1.35    | 1.92    | 2.03    | 2.19    | 2.48    | 3.01    | 3.51    |
| 15.00     | 1.38    | 1.93    | 2.04    | 2.20    | 2.49    | 3.02    | 3.52    |
| Quantiles | log Kaw | log Kaw | log Koa |
|-----------|---------|---------|---------|
| 0.025     | 4.71    | 4.85    | 5.13    |
| 0.025     | 4.50    | 4.84    | 5.03    |
| 0.025     | 4.49    | 4.82    | 5.07    |
| 0.025     | 4.61    | 4.76    | 5.14    |
| 0.025     | 5.14    | 5.32    | 5.56    |
| 0.025     | 5.49    | 5.65    | 5.88    |
| 0.025     | 5.81    | 6.00    | 6.24    |
| 0.025     | 6.20    | 6.38    | 6.65    |
| 0.025     | 5.17    | 5.32    | 5.57    |
| 0.025     | 4.96    | 5.13    | 5.49    |
| 0.025     | 4.86    | 5.10    | 5.27    |
| 0.025     | 4.97    | 5.16    | 5.32    |
| 0.025     | 5.16    | 5.35    | 5.49    |
| 0.025     | 5.42    | 5.62    | 5.74    |
| 0.025     | 5.75    | 5.91    | 6.03    |
| 0.025     | 6.08    | 6.26    | 6.39    |
| 0.025     | 6.44    | 6.61    | 6.78    |
| 0.025     | 6.82    | 7.00    | 7.11    |
| 0.025     | 5.62    | 5.77    | 5.87    |
| 0.025     | 5.41    | 5.59    | 5.75    |
| 0.025     | 5.32    | 5.53    | 5.68    |
| 0.025     | 5.34    | 5.56    | 5.74    |
| 0.025     | 5.49    | 5.70    | 5.88    |
| 0.025     | 5.73    | 5.94    | 6.08    |
| 0.025     | 5.99    | 6.20    | 6.34    |
| 0.025     | 6.35    | 6.53    | 6.65    |
| 0.025     | 6.70    | 6.88    | 7.02    |
| 0.025     | 7.03    | 7.25    | 7.38    |
| 0.025     | 7.42    | 7.63    | 7.77    |
| 0.025     | 6.08    | 6.23    | 6.33    |
| 0.025     | 5.84    | 6.04    | 6.20    |
| 0.025     | 5.73    | 5.96    | 6.15    |
| 0.025     | 5.72    | 5.98    | 6.15    |
| 0.025     | 5.82    | 6.06    | 6.25    |
| 0.025     | 6.05    | 6.27    | 6.45    |
| 0.025     | 6.30    | 6.51    | 6.66    |
| 0.025     | 6.62    | 6.82    | 7.06    |
| 0.025     | 6.94    | 7.14    | 7.27    |
| 0.025     | 7.30    | 7.49    | 7.63    |
| 0.025     | 7.63    | 7.85    | 7.99    |
| 0.025     | 8.01    | 8.23    | 8.39    |
Table S5. Predicted and experimental log $K_{ow}$ ranges for CP mixtures. Experimental data are from Renberg et al.\(^5\)

| Experimental data | FCM prediction (PLSR, double/triple Cl allowed) |
|-------------------|-----------------------------------------------|
|                   | min   | max   | Percentiles | 0.025 | 0.25  | 0.5   | 0.75  | 0.975 |
| Cereclor 50LV     | 4.39  | 6.93  | Cereclor 50LV | 4.91  | 5.46  | 5.78  | 6.16  | 6.99  |
| Cereclor 70L      | 5.68  | 8.01  | Cereclor 70L | 6.16  | 7.06  | 7.54  | 7.97  | 8.75  |
| Hüls 60C          | 4.48  | 7.38  | Hüls 60C   | 5.20  | 5.82  | 6.21  | 6.61  | 7.49  |
| Hüls 70L          | 5.89  | 8.69  | Hüls 70   | 6.14  | 6.95  | 7.50  | 8.09  | 8.94  |
| Witaclor 63       | 5.47  | 7.30  | Witaclor 63 | 5.36  | 6.04  | 6.45  | 6.90  | 7.81  |
| Witaclor 71P      | 5.37  | 8.69  | Witaclor 171P | 6.21  | 7.10  | 7.63  | 8.18  | 9.04  |
References

1. Czub, G.; Wania, F.; McLachlan, M. S., Combining Long-Range Transport and Bioaccumulation Considerations to Identify Potential Arctic Contaminants. *Environ. Sci. Technol.* **2008**, *42*, (10), 3704-3709.

2. Brown, T. N.; Wania, F., Screening Chemicals for the Potential to be Persistent Organic Pollutants: A Case Study of Arctic Contaminants. *Environ. Sci. Technol.* **2008**, *42*, (14), 5202-5209.

3. Glüge, J.; Bogdal, C.; Scheringer, M.; Buser, A. M.; Hungerbühler, K., Calculation of Physicochemical Properties for Short- and Medium-Chain Chlorinated Paraffins. *J. Phys. Chem. Ref. Data* **2013**, *42*, (2), 023103.

4. Yuan, B.; Bogdal, C.; Berger, U.; MacLeod, M.; Gebbink, W. A.; Alsberg, T.; de Wit, C. A., Quantifying Short-Chain Chlorinated Paraffin Congener Groups. *Environ. Sci. Technol.* **2017**, *51*, (18), 10633-10641.

5. Renberg, L.; Sundstrom, G.; Sundhnygard, K., Partition-Coefficients of Organic-Chemicals Derived from Reversed Phase Thin-Layer Chromatography - Evaluation of Methods and Application on Phosphate-Esters, Polychlorinated Paraffins and Some PCB-Substitutes. *Chemosphere* **1980**, *9*, (11), 683-691.
Predicting Partition Coefficients of Short-Chain Chlorinated Paraffin Congeners by COSMO-RS-Trained Fragment Contribution Models

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 uses a sophisticated but time-demanding quantum chemically based method COSMO-RS and a fast-running fragment contribution approach to enable prediction of partition coefficients for a large number of short-chain chlorinated paraffin (SCCP) congeners. Fragment contribution models (FCMs) were developed using molecular fragments with a length of up to C₄ in CP molecules as explanatory variables and COSMO-RS-calculated partition coefficients as training data. The resulting FCMs can quickly provide COSMO-RS predictions for octanol–water (Kₗw), air–water (Kₐw), and octanol–air (Kₒa) partition coefficients of SCCP congeners with an accuracy of 0.1–0.3 log units root mean squared errors. The FCM predictions for Kₗw agree with experimental values for individual constitutional isomers within 1 log unit. The distribution of partition coefficients for each SCCP congener group was computed, which successfully reproduced experimental log Kₗw ranges of industrial CP mixtures. As an application of the developed FCMs, the predicted Kₐw and Kₒa 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. Industrial CP mixtures usually have a chlorination degree from 30 to 70% by weight. 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. 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. 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. 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 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 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. Relative values across chemicals are expected to be even more accurate because systematic errors are canceled.

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. previously applied COSMO-RS to predict partition coefficients of CPs but performed predictions for only 4 structures per congener group, which may be 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 develops a fragment contribution model (FCM) for CPs. 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.\textsuperscript{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.\textsuperscript{14} CPs are made up of relatively simple fragments and thus simple FCMs are expected to reproduce the predictions from a more sophisticated model well. The trained FCMs are then compared to COSMO\textsuperscript{frag}, an existing method that also quickly produces approximate results of COSMO-RS prediction.\textsuperscript{15,16} Finally, the newly developed FCMs are used to provide distributions of partition coefficients for CP congener groups and CP technical mixtures.

Methods

\textbf{Method overview.} COSMO-RS-trained FCMs for the log of octanol–water ($K_{ow}$), air–water ($K_{oa}$), 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 distributions 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.

\textbf{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.\textsuperscript{5} For a given stereochemically specific congener, the molecular structure in the SDF format was entered into the COSMO\textsuperscript{confX} 4.3 software (COSMologic), which selected optimal conformers and generated their COSMO files using quantum chemistry program TURBOMOL 7.3 (COSMologic). These COSMO files were then used in COSMO\textsuperscript{thermX} 19.0.4...
(COSMOlogic, 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 COSMOconfX 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 COSMOconfX, switched off RDKit, and used only Balloon to generate initial candidate conformers.

In the course of work, we noticed that the calculated partition coefficient sometimes depends slightly on the conformation of the initial input structure entered in COSMOconfX. 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 COSMOtherm predictions for CPs.

**Generation of training and validation sets.** 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. We used “very” short to short-chain CPs as training chemicals because computational time of COSMOconfX increases with the size of molecule. The validation set, in contrast, should comprise congeners that are relevant. We used 120 SCCP congeners (30 for each of C$_{10}$ to C$_{12}$ CPs) that were also randomly generated. 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.

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. Codes were written in the R language$^{17}$ to create SMILES strings for all these congeners. SMILES was then converted to SDF format using OpenBabel,$^{18}$ which was then fed to COSMOconfX 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 Bioconductor.$^{19}$ 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. Structures of all fragments are given as SMARTS strings in the supplemental Excel data file. Using these fragments, four levels of models were generated. Level 1 model used only C$_1$ fragments, and Level 2 model C$_1$ and C$_2$ fragments. Levels 3 and 4 were calibrated with C$_1$ to C$_3$ and C$_1$ to C$_4$ fragments, respectively. Models were calibrated with least square multiple linear regression (MLR) with fragment counts as explanatory variables and COSMOtherm predictions as dependent variables. 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.\textsuperscript{20} Variable selection was first performed for the Level 1 model. Then, the selected C\textsubscript{1} fragments were used as the initial variable set of the variable selection procedure for the Level 2 model, and so forth. To avoid 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 \textit{R} using functions such as \texttt{lm()}, \texttt{step()}, \texttt{plsr()}, and \texttt{selectNcomp()}.

\textbf{Predictions of partition coefficients for congener groups.} Using the FCMs calibrated with PLSR, log \textit{K}_{\text{ow}}, log \textit{K}_{\text{sw}}, and log \textit{K}_{\text{oa}} for 1000 randomly generated isomers for each SCCP congener group (C\textsubscript{10}–C\textsubscript{13}, Cl\textsubscript{2}–Cl\textsubscript{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. 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\textsubscript{10}Cl\textsubscript{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 C. For example, 1000 random isomers of C\textsubscript{10}Cl\textsubscript{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.\textsuperscript{21} 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,\textsuperscript{22,23} which is also inferred by GC retention measurements for CP mixtures.\textsuperscript{4,24} Nevertheless, in highly chlorinated CP mixtures, dichloro-substituted carbons and trichloromethyl groups have also been identified.\textsuperscript{21,25} 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.

\textbf{COSMOfrag.} COSMOfrag, as implemented in BIOVIA COSMO\textit{quick} software (Version 1.8; Dassault Systemes), was compared to COSMOTHERM and FCM calculations. COSMOfrag builds a meta COSMO file of the new molecule by collecting and assembling fragments of molecules for which COSMO files are already available. This process is very quick (<1 s per molecule), as it does not require any additional quantum chemical calculation. COSMO\textit{quick} then performs COSMOTHERM calculations to predict various properties including partition coefficients. The default database from which to take fragments is composed of >190,000 molecules. Note that the current version of COSMO\textit{quick} does not consider conformers and diastereomers and supports only the TZVP level, which is less accurate than TZVPD.
Results and discussion

FCM training and validation. For all of log \(K_{ow}\), log \(K_{aw}\), and log \(K_{oa}\), increasing the number of fragments for MLR from Level 1 to Level 4 models improved the quality of fitting on the 815 training data, as indicated by \(R^2\), root mean squared errors (RMSE), and AIC (Figures 1, S1, S2, Table S2). Hence, Level 4 model resulted in the best fit on the training data set. It is interesting that C₄ fragments do have statistically significant contributions to the partition coefficients, suggesting that the molecular interaction properties of CPs cannot fully be reduced to the shorter fragments, and that the actual contribution of each C type (e.g., -CH₂-, -CHCl-) to log \(K\)'s depends on its neighboring structure (see Figure S3 and additional discussion on fragment contributions in the SI). In the variable selection procedure, about half (49-61%) of the total fragments were removed for Level 2 to 4 models. This is not surprising, because many fragments share common substructures and thus are interrelated. PLS with the Level 4 fragments resulted in a similar fitting quality as compared to the least square MLR, although the PLSR has more restrictions (i.e., a lower degree of freedom) when deriving fitting coefficients. The good fit on the training data as indicated by low RMSE (0.05, 0.12, and 0.09 for log \(K_{ow}\), log \(K_{aw}\), and log \(K_{oa}\), respectively) with the Level 4 model or its PLSR version show that FCMs can accurately fit COSMOtherm calculated values for CPs. These RMSE values are similar to the precision of COSMOtherm for CPs mentioned in the Methods section and may thus be considered the best achievable fit. All resulting fitting coefficients for all levels are presented in the supplementary Excel file. We note that some fragments that describe diastereometric structures were also significant.

External validation with 120 SCCP congeners leads to the same conclusions as those with the training data presented above. Thus, the Level 4 model showed the best statistics, and the statistics were better in order of log \(K_{ow}\), log \(K_{oa}\), and log \(K_{aw}\). (Figures 1 and S4, Table S2). PLSR and the Level 4 model predicted the validation set equally well. While PLSR typically is considered more robust when the number of variables is large, it was just similar to the least square MLR-based Level 4 model for validation. RMSE for validation was 0.12, 0.29, and 0.21 for log \(K_{ow}\), log \(K_{aw}\), and log \(K_{oa}\), respectively, being 2.2–2.4 times higher than RMSE for training. The results of validation thus suggest that use of the calibrated FCMs causes additional prediction errors of 0.1 to 0.3 RMSE in log \(K\)'s of SCCPs, as compared to the direct use of COSMOtherm.
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, S2, S4) and a table with statistics for all models (Table S2) are presented in the SI.

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 COSMOtherm 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 COSMOtherm 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 COSMOtherm predictions agree better with the experimental data in this case. As Figure 2 shows, the predictions for $K_{aw}$ by FCM and COSMOtherm 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-C14Cl4. $K_{oa} \geq L$ 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-trained FCMs (with PLSR calibration), directly from COSMOtherm software, and from COSMOfrag with the default and updated databases. $K_{ow}$ data are from Hilger et al.26 $K_{aw}$ data are from Drouillard et al.27 for 23°C. Multiple data points both for predictions and experimental data indicate the presence of data for diastereomers. Bars indicate the mean. COSMOfrag with the default database resulted in a program error for 1,1,1,3,6,7,10,12,12,12-C12Cl10; thus, no data is shown.

**COSMOfrag.** COSMOfrag was used to predict log $K_{ow}$, $K_{aw}$, and $K_{oa}$ for the 120 validation CP congeners. COSMOfrag with the default database resulted in charge inconsistency and could not provide predictions for 52 out of 120 validation congeners, which may have resulted from the absence of similar compounds in the database.16 For the remaining 68 congeners, $R^2$ was 0.33–0.82 and RMSE was 1.3–3.2 log units, as compared to the direct COSMOtherm predictions (Table S2, Figure S5). Apparently, COSMOfrag with the default database with >190,000 compounds cannot reproduce the calculations of COSMOtherm accurately. We then added COSMO files for the most stable conformers (i.e., c0.cosmo files) of the 815 calibration CPs to this database. This update of the database substantially improved the predictions. $R^2$ and RMSE became 0.51–0.94 and 0.26–0.72, respectively, without any program error for all 120 validation congeners. COSMOfrag predictions were also compared to the available experimental data (Figure 2). For C10 congeners, the COSMOfrag-predicted $K$ values agree well with the experimental values and also with the predictions from direct
COSMOtherm and the FCMs. For C11 and C12, however, COSMOfrag systematically overestimates the experimental log $K_{ow}$ values by ca 2 log units when the default database is used, and by ca 1 log unit when the updated database is used.

The results presented above show that the newly developed FCMs can better reproduce direct COSMOtherm predictions than COSMOfrag does, even with additional 815 calibration compounds to the database. Additionally, COSMOfast needed ca 30 min to predict partition coefficients of 1000 CP congeners, while the FCM (once calibrated) only needed a few seconds to minutes using a PC (Windows, Intel® Core™ i7-8550U). Thus, for the purpose of predicting partition coefficients of a large number of CP congeners, the FCMs do have an advantage.

**Distributions of $K_{ow}$, $K_{aw}$, and $K_{oa}$ for SCCP congener groups.** Using the FCMs based on PLSR calibration, $K_{ow}$, $K_{aw}$, and $K_{oa}$ for 1000 isomers per SCCP congener group were predicted (Figures 3, S6). The 2.5, 25, 50 (median), 75, and 97.5%iles of log $K$’s for each SCCP congener group are presented in Tables S3 and S4.

We derived the distributions of partition coefficients for CP congeners with and without double and triple Cl substitutions. The distributions in the two cases turned out to be similar. The median for each congener group is different by 0.13 log unit on average and by 0.39 at most between the cases with and without double/triple Cl. The similarity is particularly high for low-chlorinated congeners (Cl2–Cl4), which may be expected because random generation forms only a limited number of double and triple substitutions even if allowed. As the number of Cl increases, the distribution peaks of log $K_{ow}$ and log $K_{oa}$ without double/triple Cl become slightly sharper and shifted as compared to the corresponding peaks with double/triple Cl. Still, the presence of double/triple Cl influences log $K$’s only by < 1 log unit and thus is not highly important for calculation of partition coefficients when the congener groups are considered as a whole. An important difference between the two cases is that congeners with the number of Cl > the number of C cannot be generated under the one Cl per C rule. Also, if Cl = C, then there is only one constitutional isomer that fulfills the one Cl per C rule (but with many stereoisomers), which makes the distribution peak comparatively sharp.

Log $K_{ow}$ and log $K_{oa}$ values 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).

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 S7). In contrast, dependence on the number of Cl is nonlinear (Figure 3; more clearly in Figure S8). Log $K_{ow}$ is fairly constant from Cl2 to Cl5, above which it increases with ca 0.35 log units/C. Log $K_{aw}$ has the opposite trend; it decreases from Cl2 to Cl10 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$_{14}$.

Figures 3 and S6 also show COSMOtherm predictions from Glüge et al.,$^9$ who reported the maximum and minimum of four predictions they made per congener group. Despite the limited number of congeners they considered, their reported values are in line with the trends based on 1000 congeners in this work. That said, there are notable differences between the two studies. First, some irregular patterns exist in the predictions of the cited work$^9$ (see, e.g., log $K_{ow}$ of C$_{13}$Cl$_9$; log $K_{aw}$ of C$_{13}$Cl$_5$), which are absent in the predicted distributions of this work. Second, Glüge et al.$^9$ generated congeners with one Cl per C at maximum, and as such, they considered only congeners with the number of Cl $<$ the number of C. Third and most importantly, the cited work$^9$ only provides a range, not a distribution. This is crucial, because the distributions appear to be often highly skewed, and the mean of the maximum and minimum does not capture the most frequently occurring values. In this regard, the median of the distribution presented in this work may be considered a more representative value for a congener group (see Tables S3, S4).

Figure 3. Kernel density estimates of log $K$ predicted by PLSR model. For each molecular formula, 1000 structures were considered. “All” indicates the results of random generation of isomers with double and triple Cl substitution allowed, while “1 Cl/C” with only one Cl per C at maximum. Vertical lines
indicate the 2.5, 50 (median), and 97.5%iles. Data points are the maximum and minimum of COSMOTHERM predictions from Glüge et al.\textsuperscript{9} Plots with more congener groups are shown in Figure S6.

**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.\textsuperscript{28} to predict log $K_{ow}$ distributions 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.,\textsuperscript{29} 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 S5). These results serve as additional validation of COSMOTHERM and the FCMs for predicting log $K_{ow}$ of CPs.

![Figure 4](image-url). Comparison of predicted distributions of log $K_{ow}$ for CP mixtures (filled curves) and experimental data from Renberg et al.\textsuperscript{29} (horizontal bars at arbitrary vertical positions). The 2.5, 50, and 97.5%iles of the predictions 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. The newly developed FCMs can provide realistic estimates for log $K$'s of individual congeners and thereby log $K$ distributions for congener groups. 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. and Brown and Wania (Figure 5; see Figure S11 for individual SCCP congener groups). 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 $\geq$ 20) do not fall in this zone. Previously, Gawor and Wania presented various chemical space plots for CPs using log $K_{aw}$ and log $K_{oa}$ predicted by ACD/ADME Suite prediction tools. The plots appear in part similar but not identical to those in this work. For example, ACD/ADME appears to predict log $K_{oa} < 6$ for many low-chlorinated SCCP congeners, but such data points are absent in Figure 5. 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.

In future work, the presented approach may be further improved at least in three aspects: (A) Extension to M/LCCPs, (B) further experimental validation, and (C) use of more realistic CP compositions. (A) is currently undertaken, and (B) will be possible in the future, as availability of isomer-specific CP standards is being improved and more data are expected (e.g., ref 24). (C) also has a good hope, as more and more knowledge regarding Cl substitution patterns in the bulk CP mixtures is becoming available, and congener compositions used for the prediction of partition coefficients could be elaborated further.

This study demonstrates that FCMs trained by COSMO-RS can accurately reproduce direct COSMO-RS predictions as well as available experimental data for CPs. As the most time-consuming COSMOconfX step that generates COSMO files has been completed for a number of CP congeners, it is possible to run COSMOtherm and derive new FCMs quickly for other partition coefficients or other properties of CPs that are related to the chemical potential in solvent. Our approach may be useful for other highly complex mixtures as well, as partitioning properties of complex mixtures are generally difficult to determine both experimentally and computationally.
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) was enclosed with lines, as in ref 31.

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; an Excel data file presenting all fragment structures and FCM coefficients. 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. COSMOconfX 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 (JPMEERF18S20300) 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.; Hungerbühler, 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 COSMOTHERM, ABSOLV, and SPARC. Environ.
Toxicol. Chem. 2014, 33, (7), 1537-1543.

7. Loschen, C.; Reinisch, J.; Klamt, A., COSMO-RS based predictions for the SAMPL6 logP challenge. J. Comput. Aided Mol. Des. 2020, 34, (4), 385-392.

8. Goss, K.-U.; Arp, H. P. H.; Bronner, G.; Niederer, C., Partition behavior of hexachlorocyclohexane isomers. J. Chem. Eng. Data 2008, 53, (3), 750-754.

9. Glüge, J.; Bogdal, C.; Scheringer, M.; Buser, A. M.; Hungerbühler, K., Calculation of physicochemical properties for short- and medium-chain chlorinated paraffins. J. Phys. Chem. Ref. Data 2013, 42, (2), 023103.

10. Meylan, W. M.; Howard, P. H., Atom/fragment contribution method for estimating octanol-water partition coefficients. J. Pharm. Sci. 1995, 84, (1), 83-92.

11. Brown, T. N.; Arnot, J. A.; Wania, F., Iterative fragment selection: a group contribution approach to predicting fish biotransformation half-lives. Environ. Sci. Technol. 2012, 46, (15), 8253-8260.

12. OECD, Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. 2007, ENV/JM/MONO(2007)2.

13. Dearden, J. C.; Cronin, M. T.; Kaiser, K. L., How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ. Res. 2009, 20, (3-4), 241-266.

14. Ramakrishnan, R.; von Lilienfeld, O. A., Machine learning, quantum chemistry, and chemical space. In Reviews in Computational Chemistry, Vol 30, Parrill, A. L.; Lipkowitz, K. B., Eds. Wiley-Blackwell: Malden, 2017; Vol. 30, pp 225-256.

15. Loschen, C.; Klamt, A., COSMOQuick: A novel interface for fast σ-profile composition and its application to COSMO-RS solvent screening using multiple reference solvents. Ind. Eng. Chem. Res. 2012, 51, (43), 14303-14308.

16. Hornig, M.; Klamt, A., COSMOfrag: A Novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry. J. Chem. Inf. Modeling 2005, 45, (5), 1169-1177.

17. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 2019.

18. O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. J. Cheminformatics 2011, 3, (1), 33.

19. Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, T.; Girke, T., ChemmineR: a compound mining framework for R. Bioinformatics 2008, 24, (15), 1733-1734.

20. Burnham, K. P.; Anderson, D. R., Multimodel Inference. Social. Methods Res. 2016, 33, (2), 261-304.

21. Yuan, B.; Lysak, D. H.; Soong, R.; Haddad, A.; Hisatsune, A.; Moser, A.; Golotvin, S.; Argyropoulos, D.; Simpson, A. J.; Muir, D. C. G., Chlorines are not evenly substituted in chlorinated paraffins: A predicted NMR pattern matching framework for isomeric
discrimination in complex contaminant mixtures. *Environ. Sci. Technol. Letters* **2020**.
doi.org/10.1021/acsestlett.0c00244.

22. Fredricks, P. S.; Tedder, J. M., Free-radical substitution in aliphatic compounds. Part II.
Halogeneration of the $n$-butyl halides. *J. Chem. Soc. (Resumed)* **1960**, 144-150.

23. Colebourne, N.; Stern, E. S., The chlorination of some $n$-alkanes and alkyl chlorides. *J. Chem. Soc. (Resumed)* **1965**, 3599-3605.

24. Hammer, J.; Matsukami, H.; Endo, S., Congener-specific partition properties of chlorinated paraffins evaluated with COSMOTHERM and GC-retention indices. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12752312.v1

25. Sprengel, J.; Wiedmaier-Czerny, N.; Vetter, W., Characterization of single chain length chlorinated paraffin mixtures with nuclear magnetic resonance spectroscopy (NMR). *Chemosphere* **2019**, **228**, 762-768.

26. Hilger, B.; Fromme, H.; Volkel, W.; Coelhan, M., Effects of chain length, chlorination degree, and structure on the octanol-water partition coefficients of polychlorinated $n$-alkanes. *Environ Sci Technol* **2011**, **45**, (7), 2842-9.

27. Drouillard, K. G.; Tomy, G. T.; Muir, D. C. G.; Friesen, K. J., Volatility of chlorinated $n$-alkanes (C10-C12): Vapor pressures and Henry’s law constants. *Environ. Toxicol. Chem.* **1998**, **17**, (7), 1252-1260.

28. Yuan, B.; Bogdal, C.; Berger, U.; MacLeod, M.; Gebbink, W. A.; Alsberg, T.; de Wit, C. A., Quantifying short-chain chlorinated paraffin congener groups. *Environ. Sci. Technol.* **2017**, **51**, (18), 10633-10641.

29. Renberg, L.; Sundstrom, G.; Sundhnygard, K., Partition coefficients of organic chemicals derived from reversed phase thin-layer chromatography: Evaluation of methods and application on phosphate esters, polychlorinated paraffins and some PCB-substitutes. *Chemosphere* **1980**, **9**, (11), 683-691.

30. Czub, G.; Wania, F.; McLachlan, M. S., Combining long-range transport and bioaccumulation considerations to identify potential Arctic contaminants. *Environ. Sci. Technol.* **2008**, **42**, (10), 3704-3709.

31. Brown, T. N.; Wania, F., Screening chemicals for the potential to be persistent organic pollutants: A case study of Arctic contaminants. *Environ. Sci. Technol.* **2008**, **42**, (14), 5202-5209.

32. Gawor, A.; Wania, F., Using quantitative structural property relationships, chemical fate models, and the chemical partitioning space to investigate the potential for long range transport and bioaccumulation of complex halogenated chemical mixtures. *Environ. Sci. Process Impacts* **2013**, **15**, (9), 1671-1684.

33. Jensen, S. R.; Brown, W. A.; Heath, E.; Cooper, D. G., Characterization of polychlorinated alkane mixtures-A Monte Carlo modeling approach. *Biodegradation* **2007**, **18**, (6), 703-717.
Other files

SCCPpartitioning FCM Data.xlsx (53.79 KiB)