Electronic supplementary information-1 (ESI-1) for "Refinement and extension of COSMO-RS-trained fragment contribution models for predicting partition properties of $C_{10–20}$ chlorinated paraffin congeners"

Satoshi Endo

Health and Environmental Risk Division, National Institute for Environmental Studies (NIES), Onogawa 16-2, 305-8506 Tsukuba, Ibaraki, Japan
Phone: ++81-29-850-2695, endo.satoshi@nies.go.jp

SI-A. Supplementary information for the variability of COSMO*confX* optimization

COSMO*therm* prediction for CPs sometimes depends on the original input structure entered in the COSMO*confX* program. Figure S1 shows predictions of K_{ow}, K_{aw} and K_{oa} of two congeners as examples. For each congener, all calculation procedure was repeated ten times, each time with a different input structure (i.e., starting conformer). Log K_{ow} was stable over the ten trials (SD < 0.04), whereas log K_{aw} and log K_{oa} varied to some extent (SD < 0.29). As another test, the three log *K*'s of 10 pairs of enantiomers were predicted and compared (Figure S2). Enantiomers (i.e., mirror-imaged molecules) should give the identical property values in the theory of COSMO-RS. However, the actual calculations of log K_{aw} and log K_{oa} for enantiomer pairs sometimes show an appreciable difference (< 0.44 log units). It appears that COSMO*confX* sometimes cannot find the optimal conformation of a CP molecule in the gas phase, which results in sporadically large solvent/air (or low air/solvent) partition coefficients. By simply increasing the number of candidate conformers to consider (ca 4 times) within the COSMO*confX* algorithm, it was possible to reduce this artifact and obtain a more repeatable prediction (SD of 10 trials < 0.10 , enantiomer difference < 0.27, Figures S1, S2). Typically, such large dependence on the starting structure is absent in COSMO*confX*. Difficulty to find the optimal conformer(s) may be characteristic for CPs, as CP congeners have many rotatable bonds and thus an enormous number of possible conformers.

SI-B. Preliminary runs for Monte Carlo model

A Monte Carlo simulation with too few molecules may not result in a repeatable set of CP congener structures and thus a representative distribution of property values for a give mixture. However, the computational time increases with increasing number of molecules. To optimize the starting number of *n*-alkane molecules, test simulations were performed with 300, 1000, 3000, 10,000, and 20,000 molecules. Simulations with 300–10,000 molecules were conducted 4 times each, and with 20,000 molecules only once. Mixtures simulated were C_{13} alkanes with 30 or 70 wt% Cl and C_{18} alkanes with 30 or 70 wt% Cl. The set of congener structures obtained from each simulation was transferred to the FCM models, which calculated log K_{ow}, log K_{aw}, and log K_{oa} for all molecules. The resulting partition coefficients were sorted according to the congener groups (i.e., molecular formulae), and 2.5, 25, 50, 75, and 97.5 percentiles of the property values were calculated for each group. The results are presented in Figures S3A–D. The results show that percentiles of property values are variable (i.e., low repeatability) when the number of molecules for a given congener group is small. Variability is relatively high for C_{18} with 70 wt% Cl, likely because of the large number of possible congener structures. Variation of percentile values becomes smaller with increasing number of molecules, and the percentiles obtained with 100 molecules or more are stable and not different from the values obtained with a larger number of molecules. From this observation, the number of molecules for Monte Carlo simulations was set to 10,000 so that representative property distributions can be obtained for congener groups that make up more than 1 mol% of the mixture.

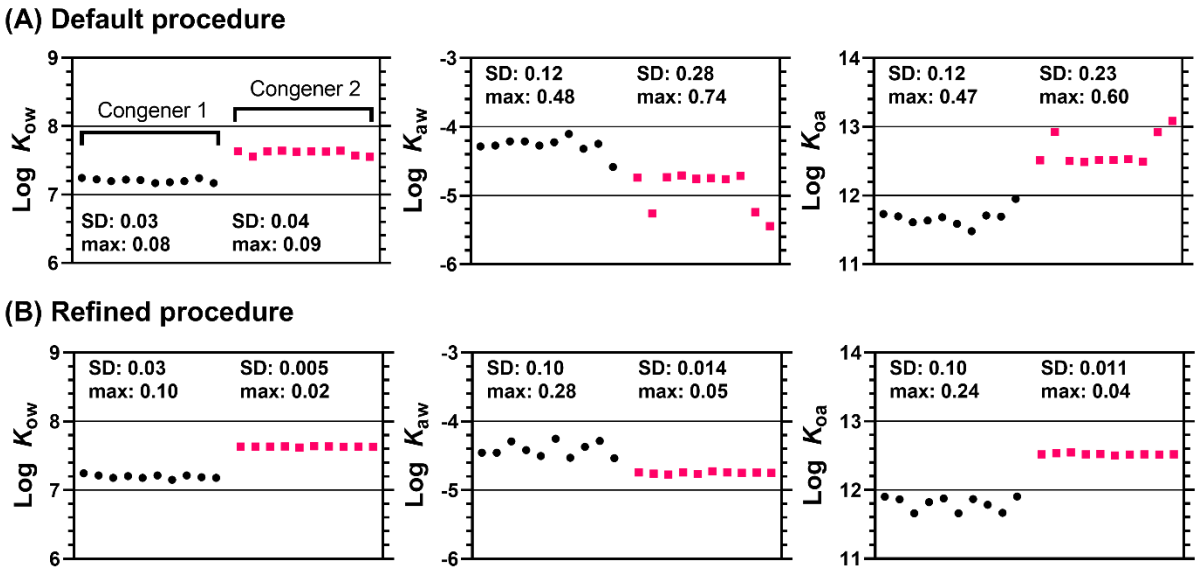

Figure S1. Influence of the initial conformer entered in COSMO*confX* on the eventual prediction results by COSMO*therm*. (A) The results of the default conformer optimization procedure implemented in COSMO*confX* (the RDkit conformer generation step was removed. See the main text). (B) The results of the refined conformer optimization procedure used in this work. For each congener and procedure, prediction was performed 10 times (thus 10 data points in each case), each time with a different starting conformer. Congener 1, (4*S*,5*S*,6*S*,7*S*,9*S*)-1,4,5,6,7,8,8,9,10,10,10-undecachlorodecane; congener 2, (2*R*,4*R*,5*S*,6*S*,7*R*,8*S*,9*S*)-1,1,1,2,3,3,4,5,6,7,8,9-tridecachlorodecane; SD, standard deviation; max, maximal difference.

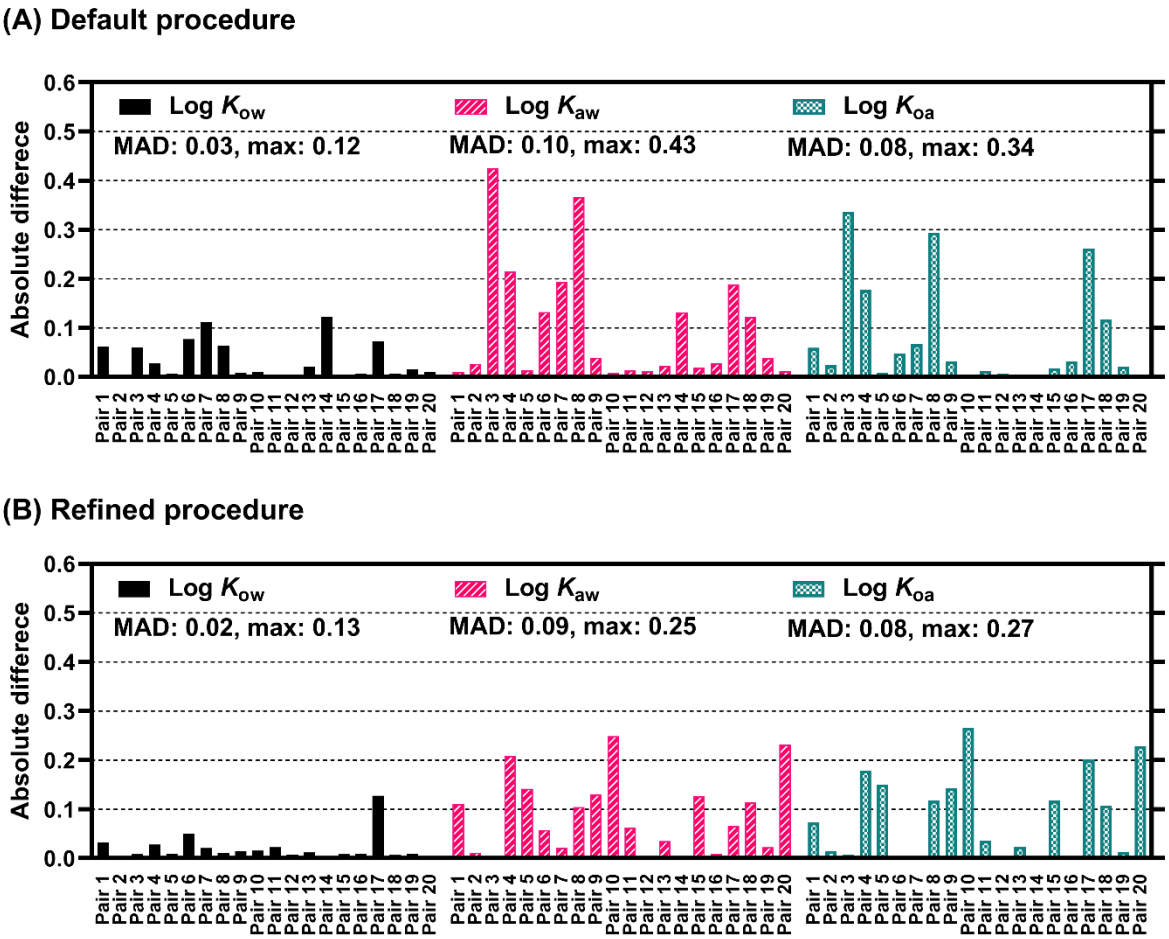

Figure S2. Difference of COSMO*therm*-predicted log *K* values between enantiomer pairs. (A) The results of the default conformer optimization procedure implemented in COSMO*confX* (the RDkit conformer generation step was removed. See the main text). (B) The results of the refined conformer optimization procedure used in this work. MAD, mean absolute difference of 20 enantiomer pairs; max, maximal difference. Congeners considered are C_{10} with varying number of Cl.

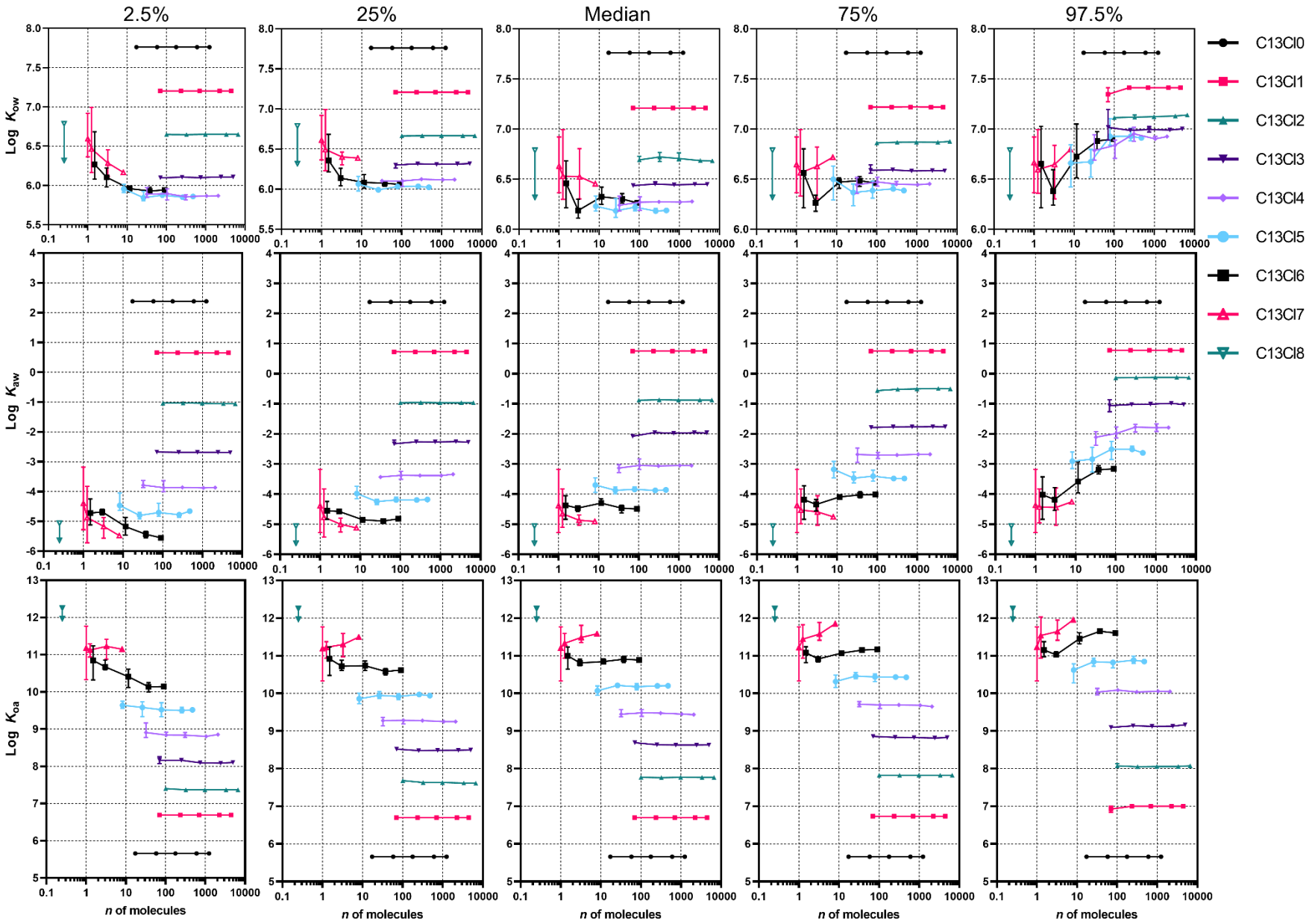

Figure S3A. Preliminary tests for Monte Carlo simulations (C_{13}, 30 wt% Cl). Error bars indicate the range of four simulations.

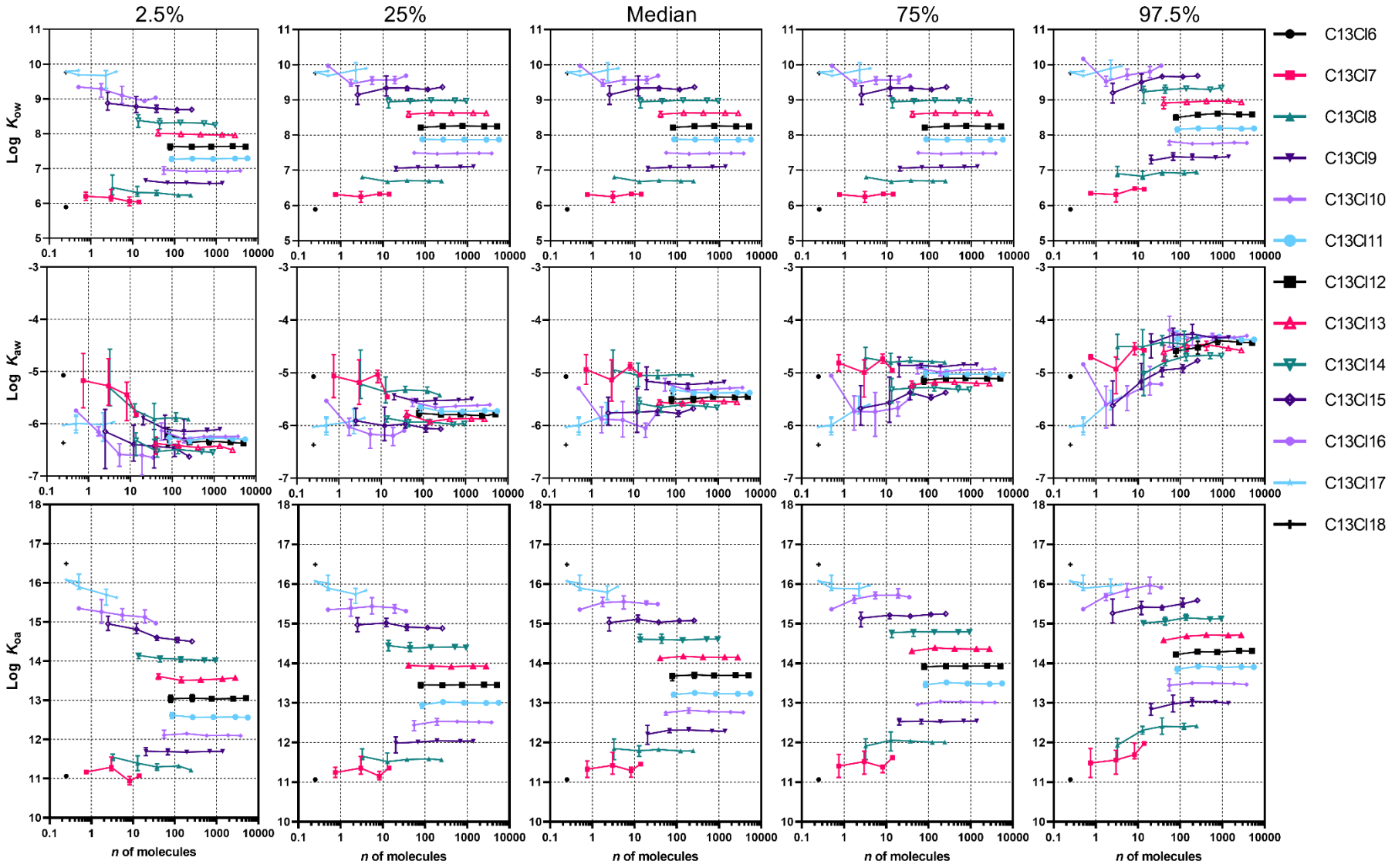

Figure S3B. Preliminary tests for Monte Carlo simulations (C_{13}, 70 wt% Cl). Error bars indicate the range of four simulations.

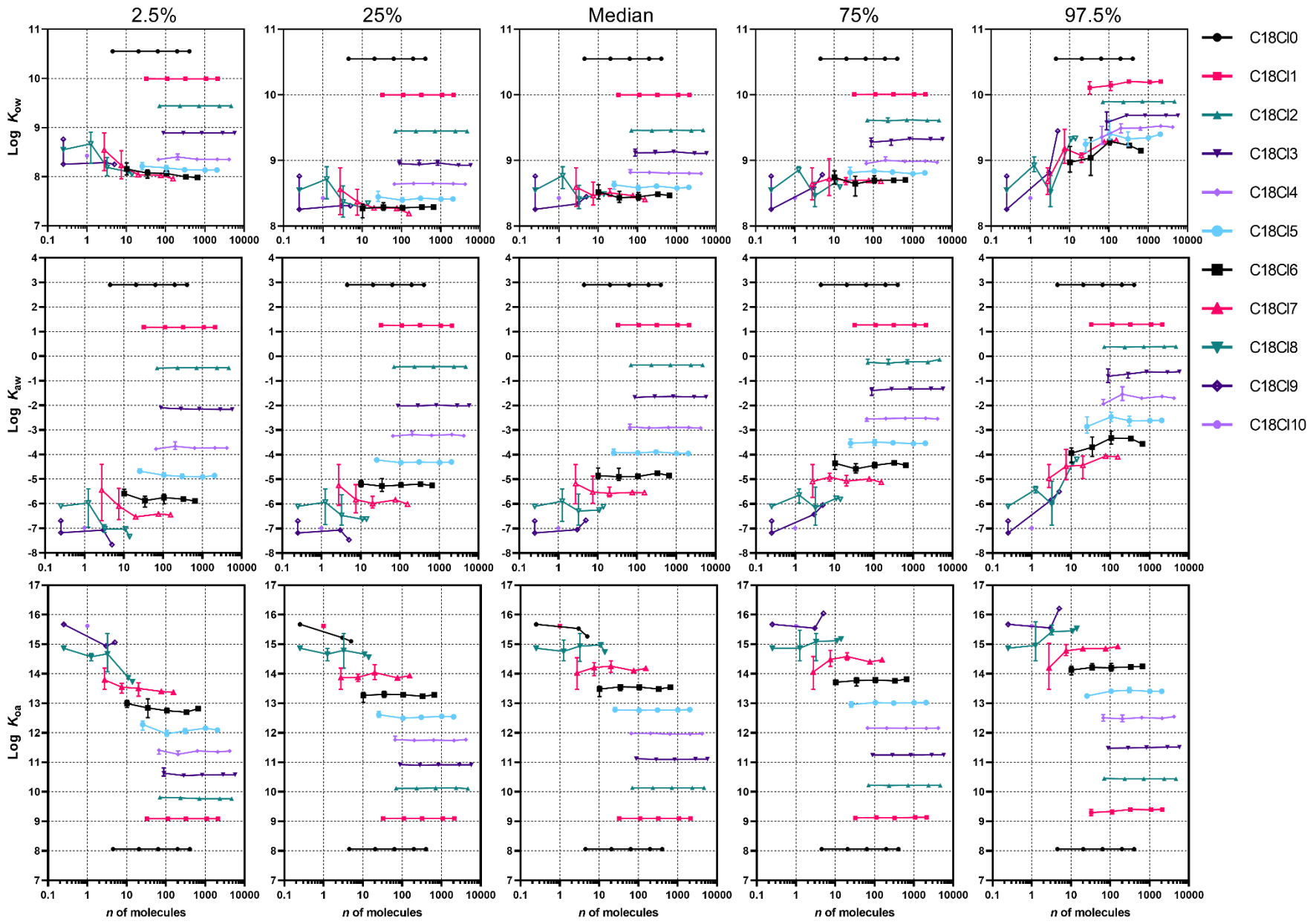

Figure S3C. Preliminary tests for Monte Carlo simulations (C_{18}, 30 wt% Cl). Error bars indicate the range of four simulations.

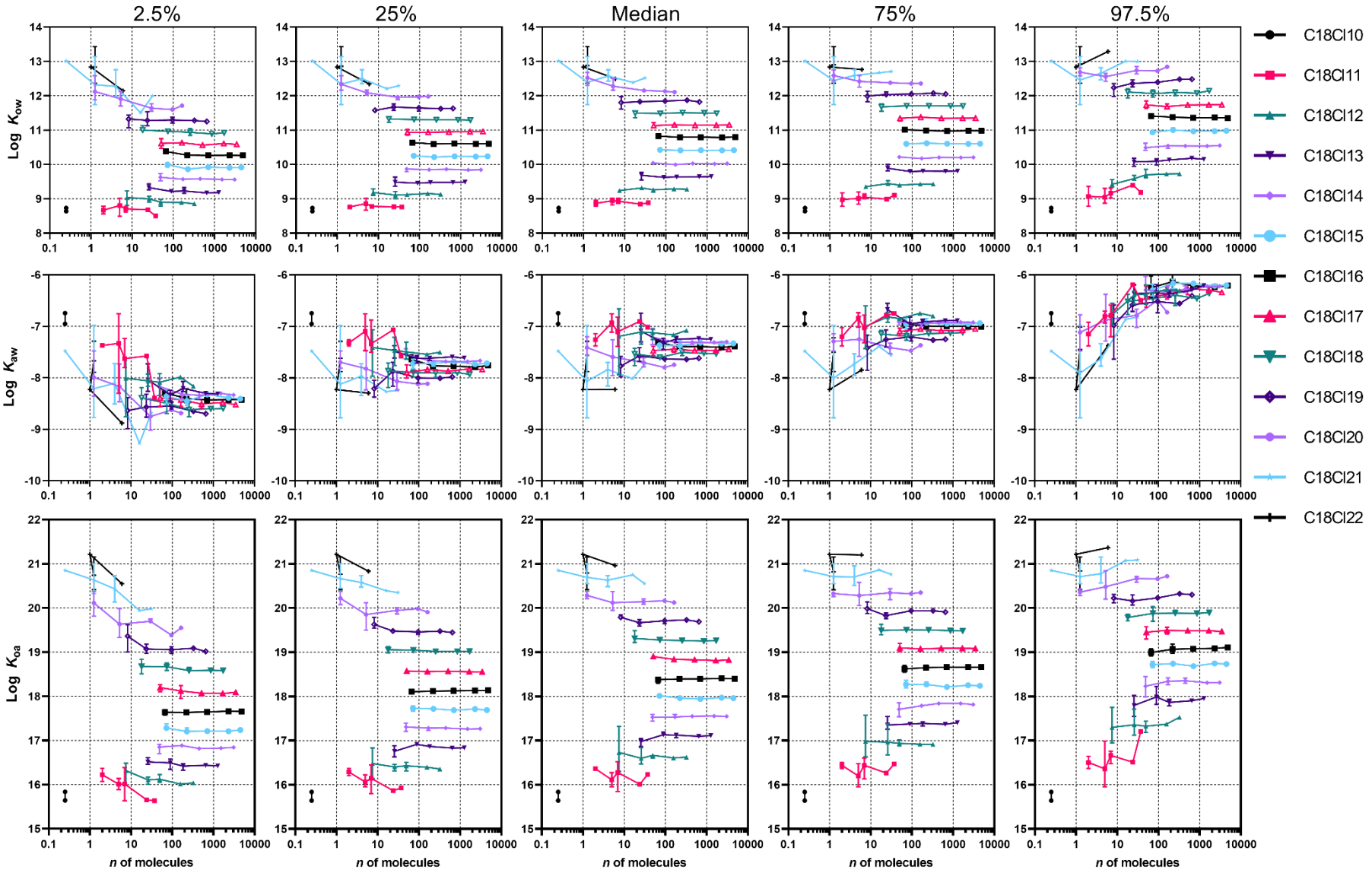

Figure S3D. Preliminary tests for Monte Carlo simulations (C_{18}, 70 wt% Cl). Error bars indicate the range of four simulations.

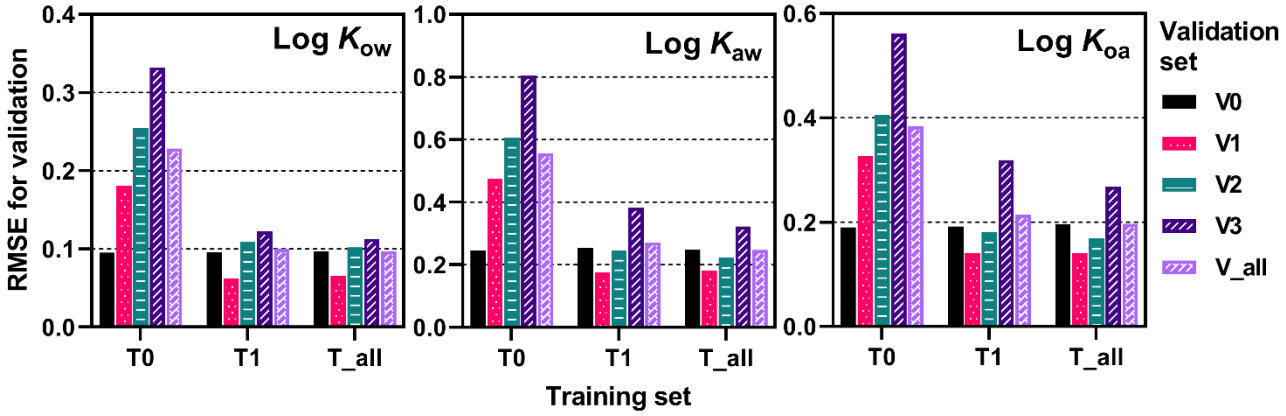

Figure S4. Root mean squared errors (RMSE) for validation of FCM models.

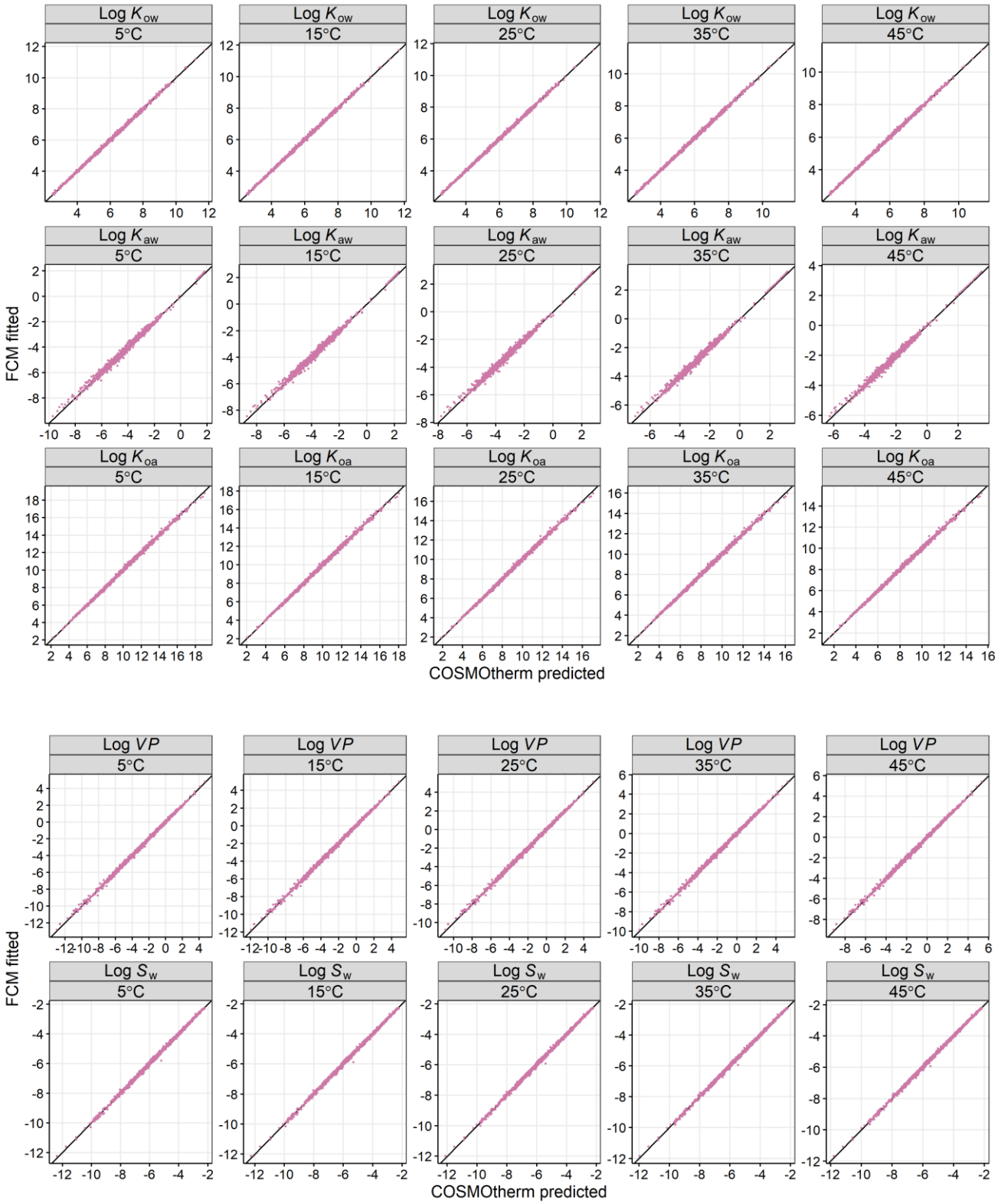

Figure S5A. The results of FCM fitting to the whole training data set (T_all).

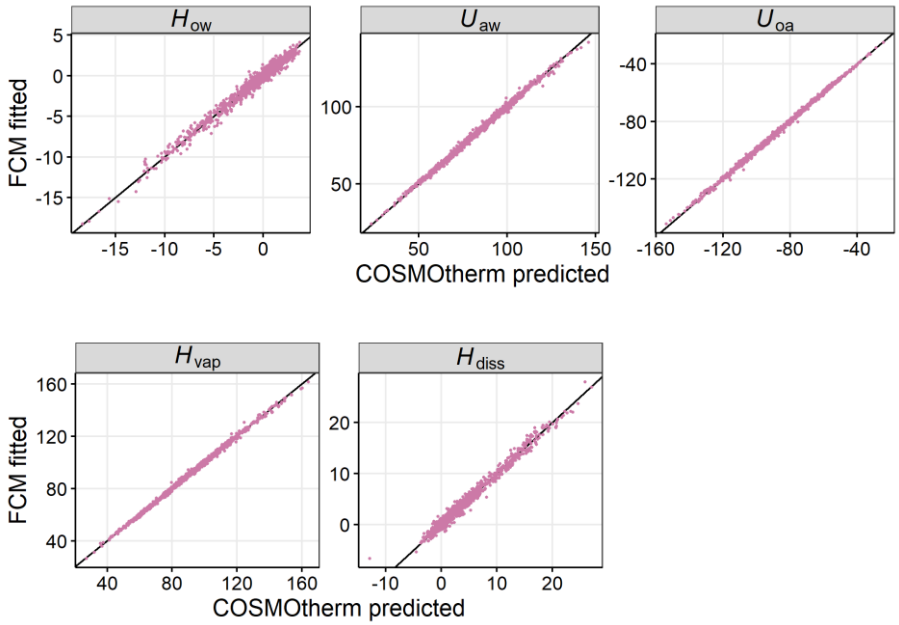

Figure S5B. The results of FCM fitting to the whole training data set (T_all). Values are in kJ/mol.

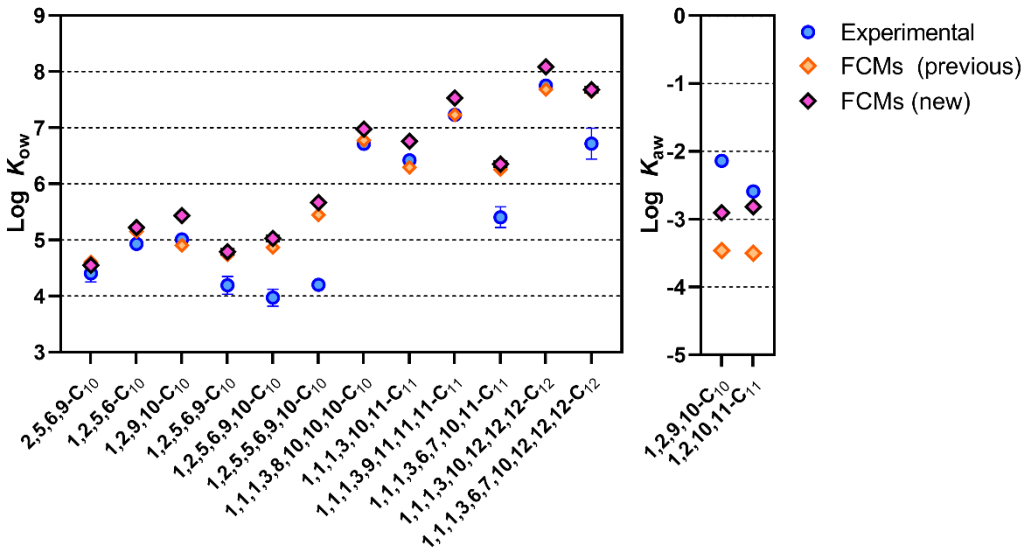

Figure S6. Comparison of experimental data and predicted values from FCMs. The previous FCMs trained in the past study[1] and the new FCMs from this study were used for prediction. Experimental K_{ow} data are from Hilger et al.[2] Experimental K_{aw} data are from Drouillard et al.[3] for 23°C.

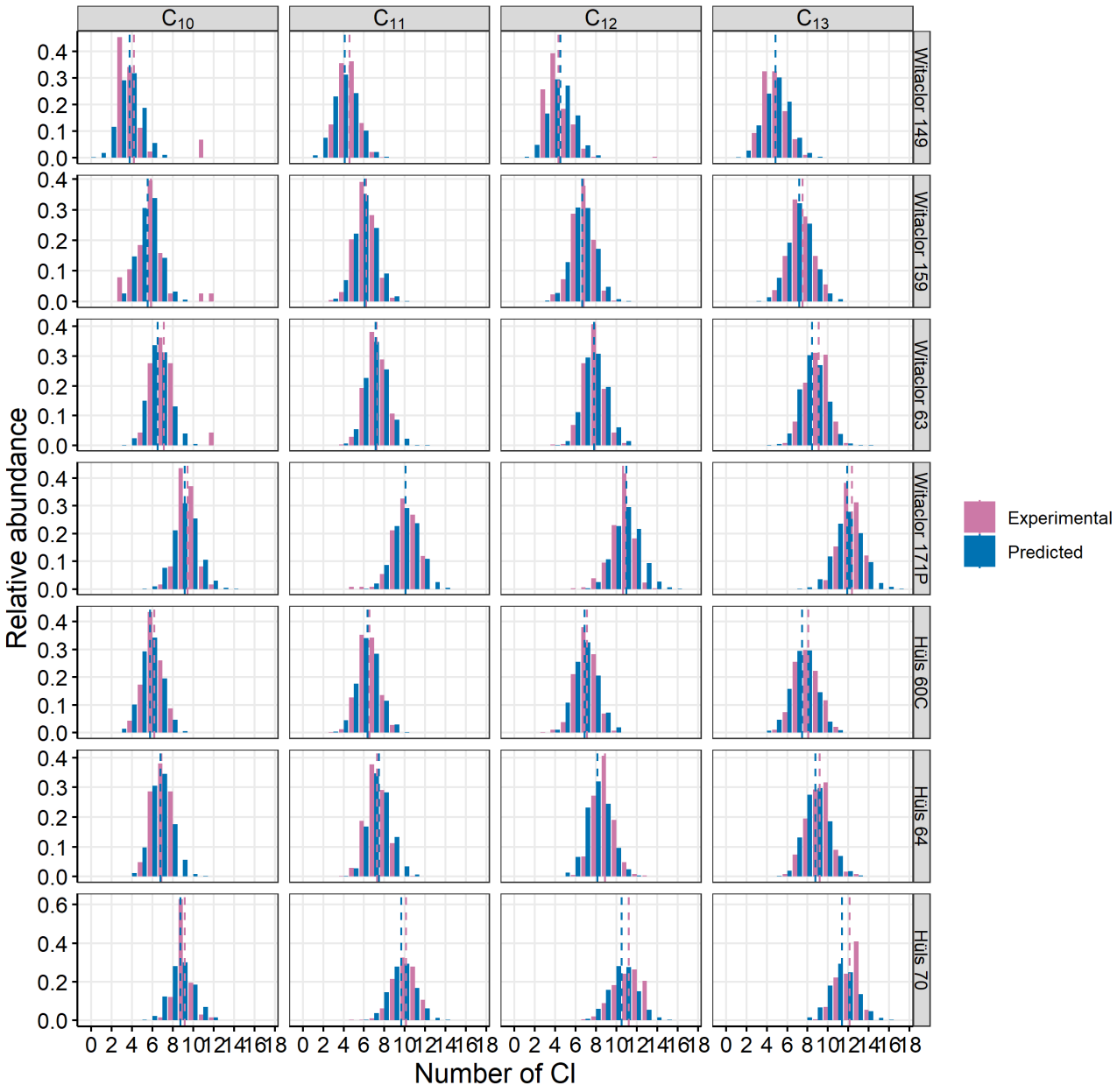

Figure S7A. Experimental and predicted congener profiles in CP mixtures.

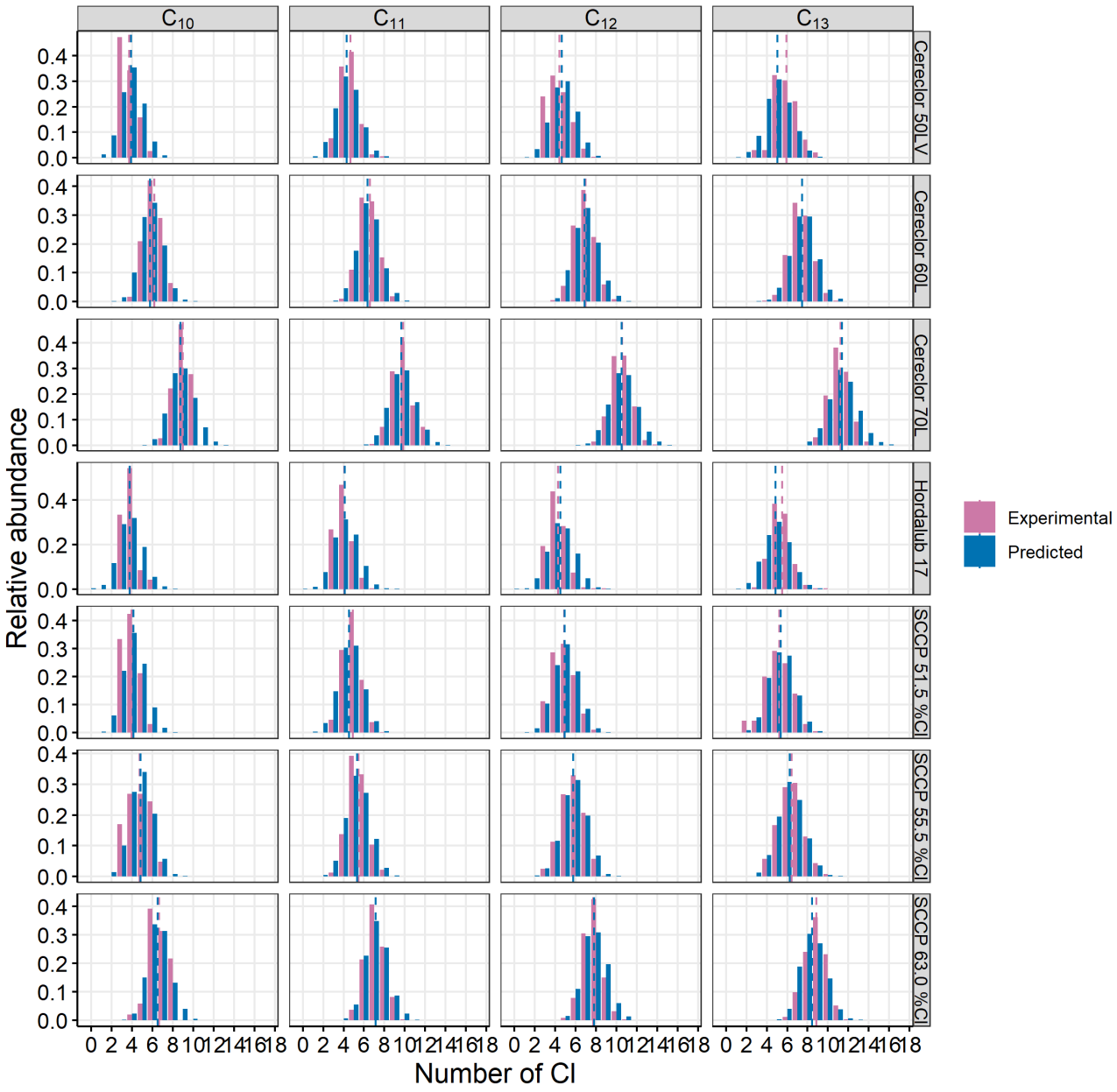

Figure S7B. Experimental and predicted congener profiles in CP mixtures.

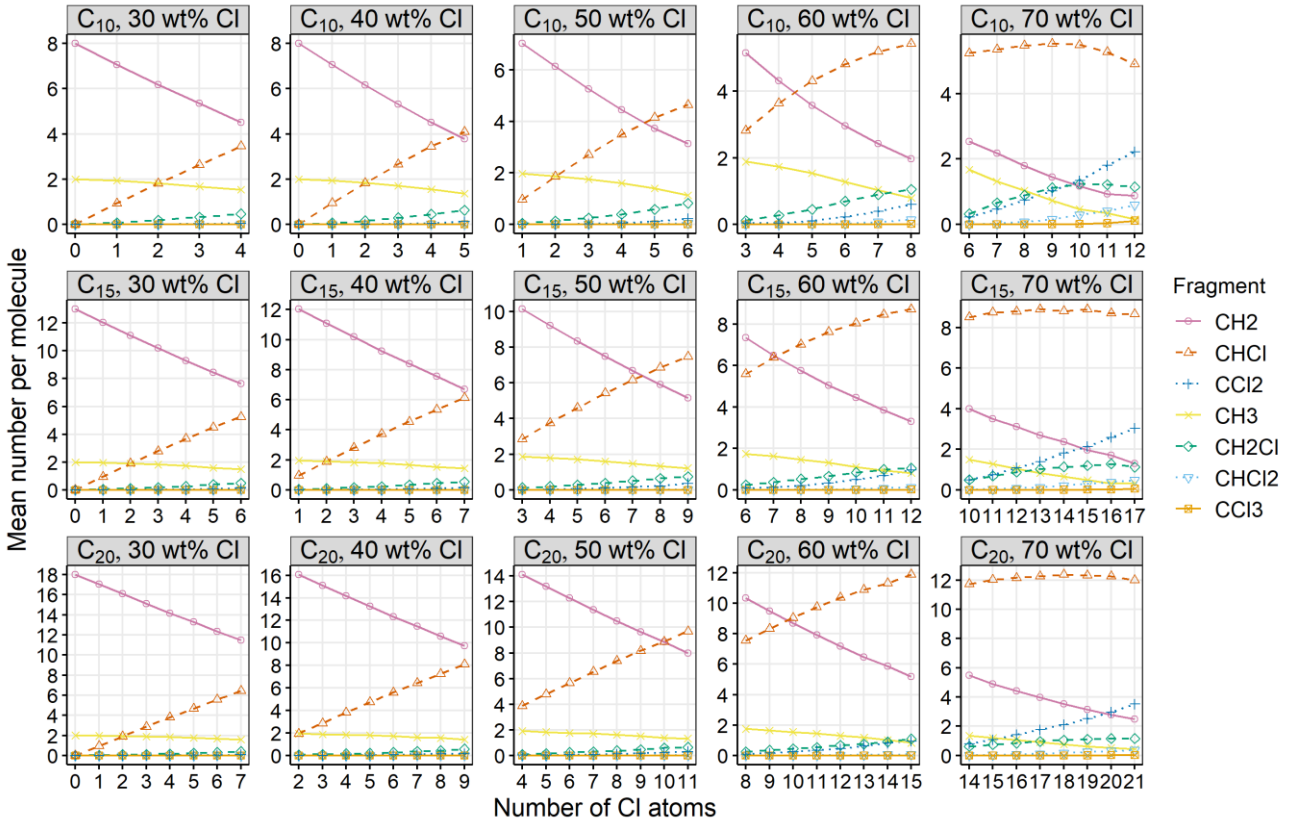

Figure S8. Mean numbers of C_1-fragments per molecule. Fragments with the mean number per molecule < 0.5 were omitted to avoid overcrowded plots.

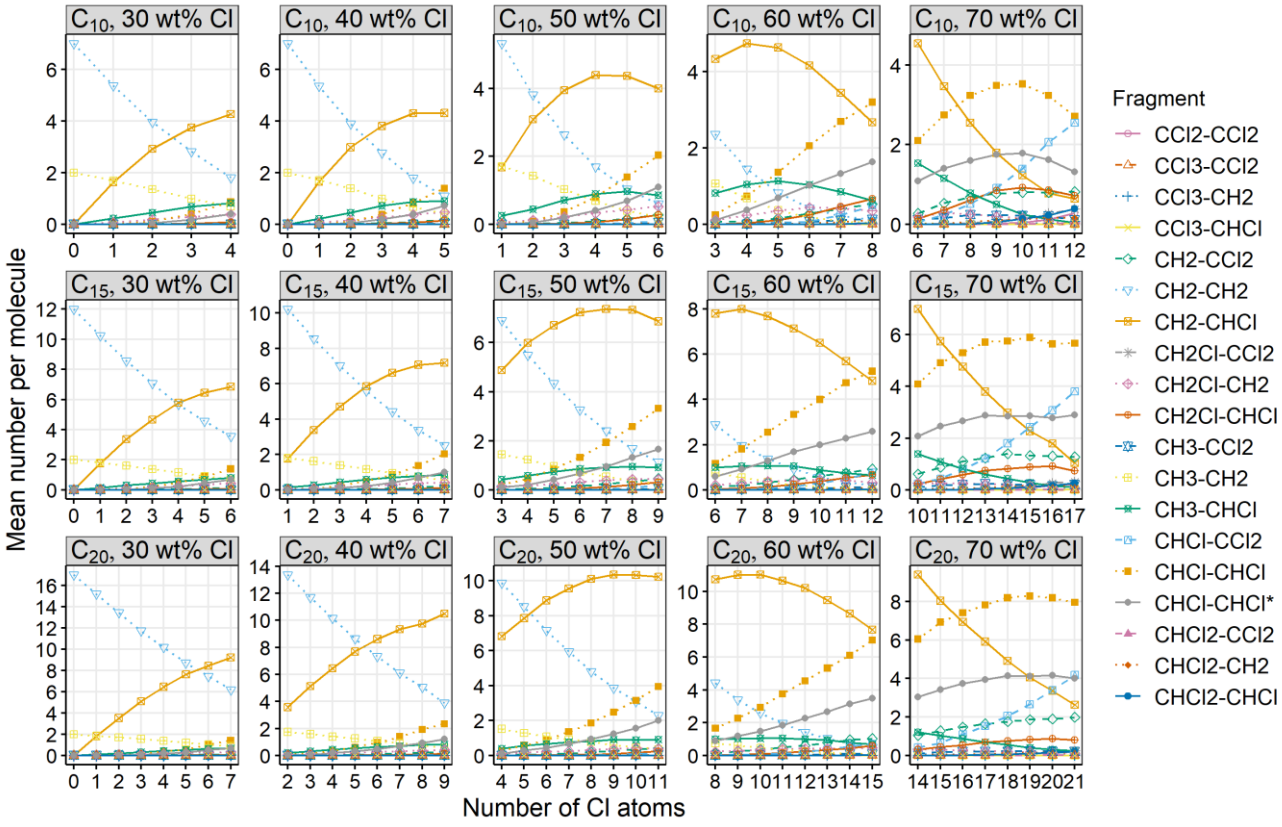

Figure S9. Mean numbers of C_2-fragments per molecule. Fragments with the mean number per molecule < 0.5 were omitted to avoid overcrowded plots. "CHCl-CHCl*" is a diastereomerically specific fragment with the two C atoms that are in the same rotational configuration.

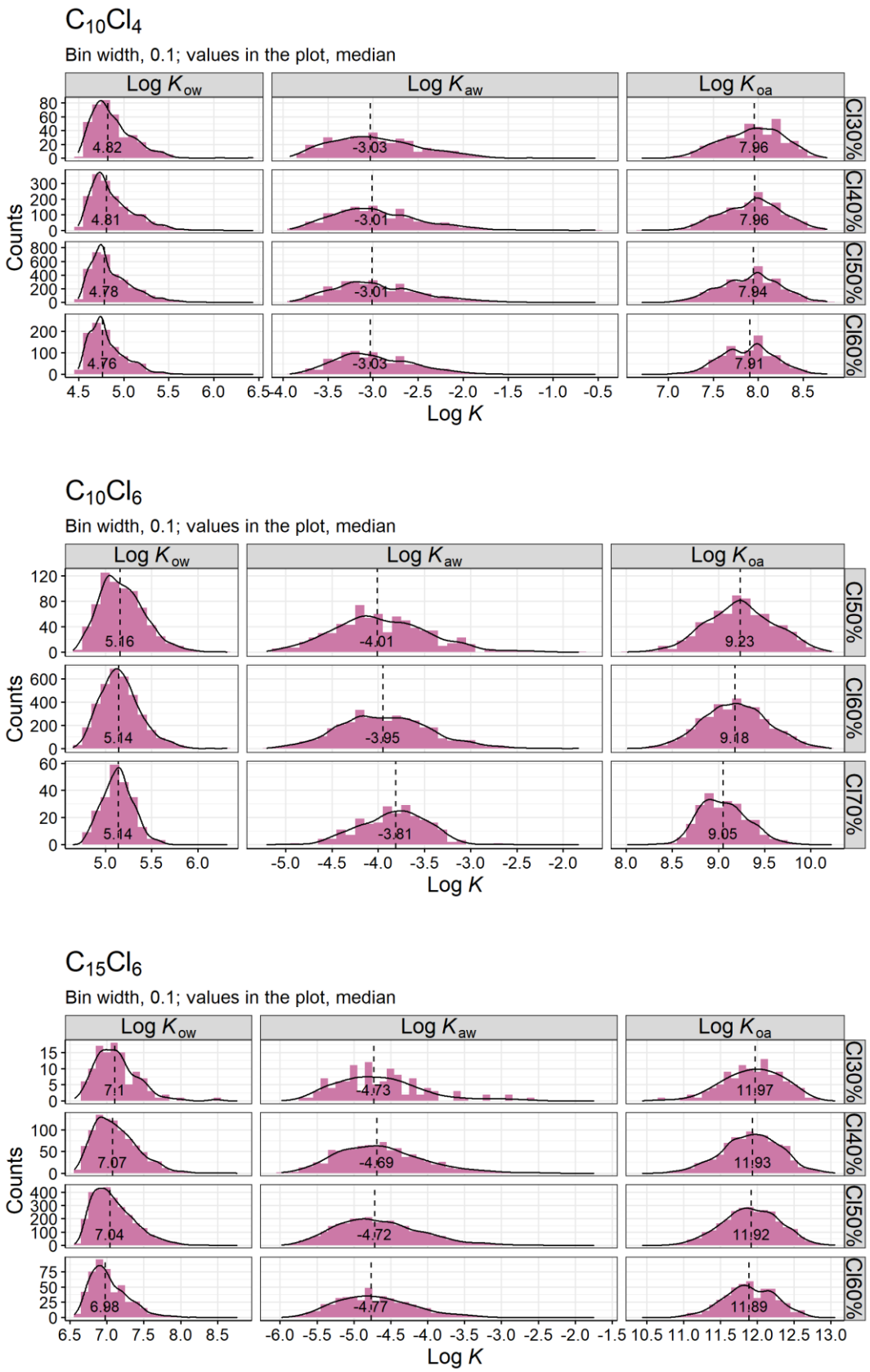

Figure S10A. Property distributions of a congener group in various technical mixtures predicted by FCMs and the Monte Carlo model.

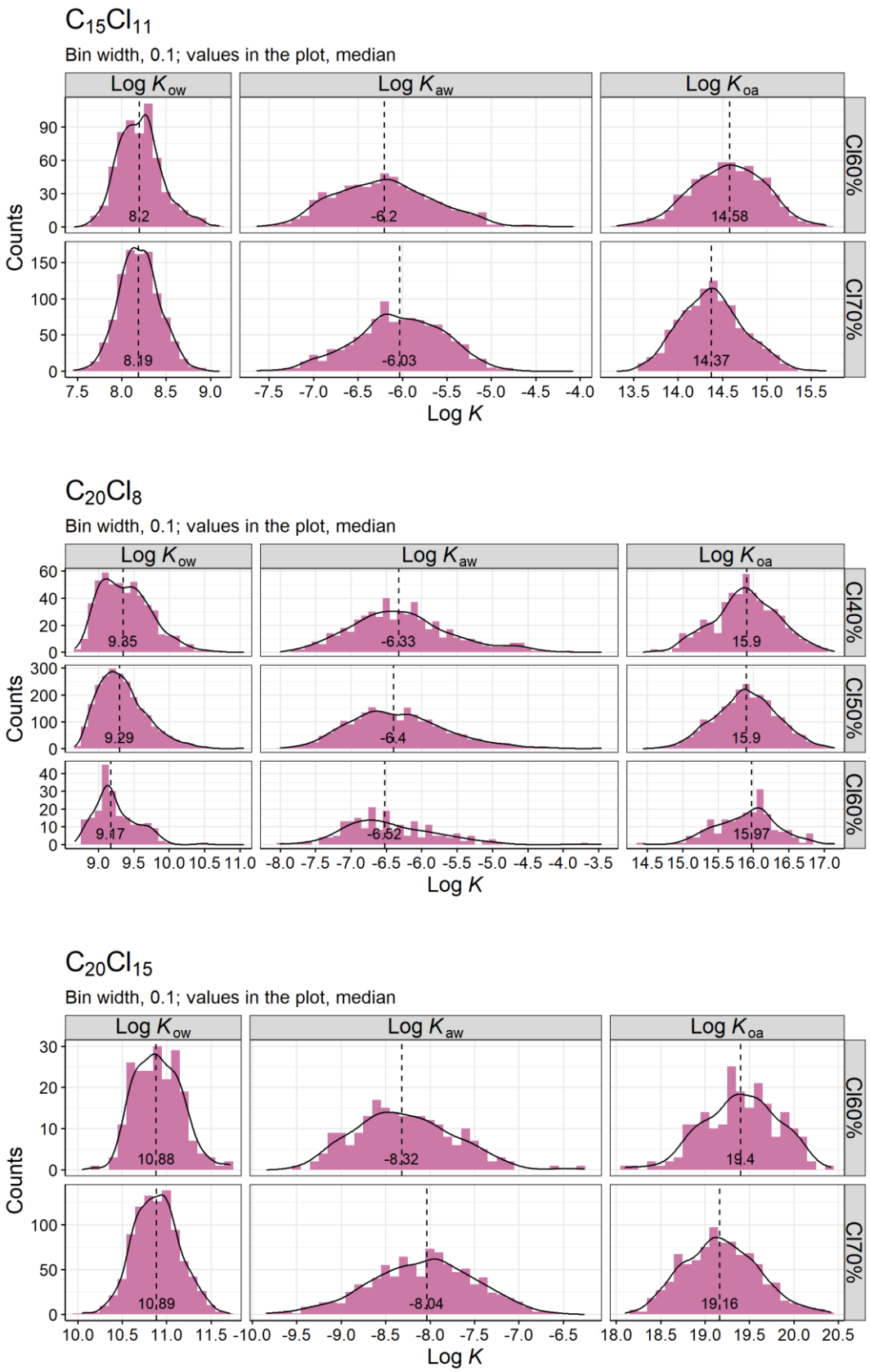

Figure S10B. Property distributions of a congener group in various technical mixtures predicted by FCMs and the Monte Carlo model.

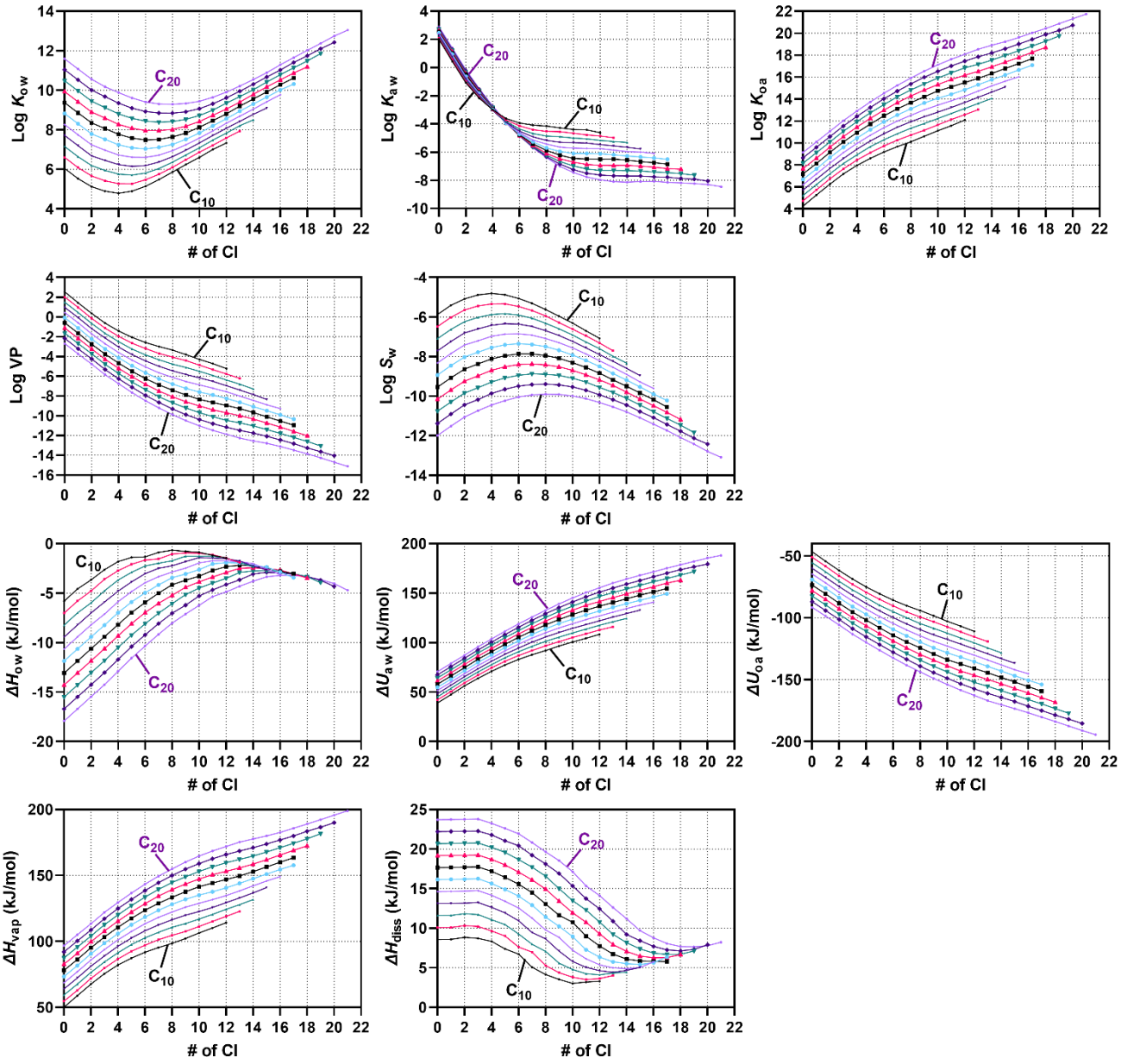

Figure S11. Medians of property values at 25 °C for each congener group predicted by COSMO-RS-trained FCMs and Monte Carlo model.

Table S1. Root mean squared errors (RMSE) of FCM training and validation

Training set	RMSE of training	RMSE of validaton Validation set				
		V0	V1	V2	V3	V_all
Log K_{ow} @ 25°C						
T0	0.044	0.095	0.181	0.255	0.332	0.228
T1	0.048	0.096	0.062	0.109	0.122	0.100
T_all	0.052	0.097	0.066	0.102	0.113	0.097
Log K_{aw} @ 25°C						
T0	0.111	0.246	0.475	0.607	0.804	0.556
T1	0.122	0.254	0.176	0.246	0.382	0.271
T_all	0.135	0.248	0.182	0.223	0.322	0.247
Log K_{oa} @ 25°C						
T0	0.087	0.190	0.327	0.405	0.561	0.384
T1	0.095	0.192	0.141	0.180	0.319	0.214
T_all	0.107	0.196	0.141	0.169	0.268	0.197

Table S2. Results of training and validation of FCMs.

	Training						Validation	
	R^2	SD	RMSE	n	Number of fragments	AIC	R^2	RMSE
Log K_{ow} 5°C	0.9989	0.0578	0.0541	1070	130	-2940	0.9949	0.1017
Log K_{ow} 15°C	0.9989	0.0566	0.0530	1070	129	-2986	0.9953	0.0981
Log K_{ow} 25°C	0.9989	0.0556	0.0521	1070	130	-3023	0.9954	0.0961
Log K_{ow} 35°C	0.9989	0.0546	0.0511	1070	134	-3057	0.9954	0.0950
Log K_{ow} 45°C	0.9990	0.0537	0.0503	1070	130	-3097	0.9957	0.0917
Log K_{aw} 5°C	0.9913	0.1569	0.1474	1070	124	-809	0.9776	0.2745
Log K_{aw} 15°C	0.9904	0.1511	0.1427	1070	115	-896	0.9771	0.2582
Log K_{aw} 25°C	0.9900	0.1434	0.1352	1070	118	-1005	0.9762	0.2457
Log K_{aw} 35°C	0.9894	0.1378	0.1298	1070	120	-1089	0.9749	0.2371
Log K_{aw} 45°C	0.9888	0.1323	0.1246	1070	119	-1178	0.9739	0.2276
Log K_{oa} 5°C	0.9983	0.1257	0.1182	1070	123	-1283	0.9937	0.2205
Log K_{oa} 15°C	0.9983	0.1195	0.1127	1070	117	-1397	0.9938	0.2074
Log K_{oa} 25°C	0.9983	0.1140	0.1075	1070	118	-1497	0.9937	0.1981
Log K_{oa} 35°C	0.9983	0.1087	0.1024	1070	120	-1597	0.9937	0.1880
Log K_{oa} 45°C	0.9982	0.1043	0.0985	1070	114	-1691	0.9936	0.1827
Log VP 5°C	0.9980	0.1409	0.1336	1070	106	-1054	0.9931	0.2355
Log VP 15°C	0.9980	0.1342	0.1272	1070	108	-1156	0.9930	0.2251
Log VP 25°C	0.9980	0.1278	0.1213	1070	105	-1263	0.9930	0.2139
Log VP 35°C	0.9980	0.1221	0.1161	1070	101	-1365	0.9929	0.2058
Log VP 45°C	0.9979	0.1171	0.1111	1070	105	-1451	0.9928	0.1965
Log S_w 5°C	0.9982	0.0801	0.0753	1070	123	-2248	0.9938	0.1258
Log S_w 15°C	0.9984	0.0764	0.0718	1070	124	-2348	0.9941	0.1221
Log S_w 25°C	0.9985	0.0729	0.0684	1070	128	-2444	0.9943	0.1187
Log S_w 35°C	0.9986	0.0705	0.0662	1070	125	-2520	0.9945	0.1150
Log S_w 45°C	0.9986	0.0677	0.0635	1070	130	-2600	0.9949	0.1093
ΔH_{ow}	0.9794	0.528	0.497	1070	121	1788	0.9507	0.878
ΔU_{aw}	0.9963	1.324	1.247	1070	119	3751	0.9895	2.140
ΔU_{oa}	0.9981	1.022	0.964	1070	118	3197	0.9936	1.701
ΔH_{vap}	0.9978	1.136	1.079	1070	104	3411	0.9929	1.854
ΔH_{diss}	0.9831	0.766	0.723	1070	116	2578	0.9641	1.150

References

1. Endo, S.; Hammer, J., Predicting Partition Coefficients of Short-Chain Chlorinated Paraffin Congeners by COSMO-RS-Trained Fragment Contribution Models. *Environ. Sci. Technol.* **2020,** *54*, (23), 15162-15169.
2. 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.
3. Drouillard, K. G.; Tomy, G. T.; Muir, D. C. G.; Friesen, K. J., Volatility of chlorinated n-alkanes (C_{10}-C_{12}): Vapor pressures and Henry's law constants. *Environ. Toxicol. Chem.* **1998,** *17*, (7), 1252-1260.