Evaluation of Physicochemical Property Data in the ECHA Database

Originally published in:
Journal of physical and chemical reference data 52(4), https://doi.org/10.1063/5.0153030

Journal of Physical and Chemical Reference Data ARTICLE pubs.aip.org/aip/jpr

Evaluation of Physicochemical Property Data in the ECHA Database

Cite as: J. Phys. Chem. Ref. Data 52, 043101 (2023); doi: 10.1063/5.0153030
Submitted: 4 April 2023 • Accepted: 18 September 2023 •
Published Online: 13 October 2023

Juliane Glüge[1,a] and Martin Scheringer[1,2]

AFFILIATIONS

[1] Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland
[2] RECETOX, Masaryk University, 625 00 Brno, Czech Republic

[a] Author to whom correspondence should be addressed: juliane.gluege@usys.ethz.ch

ABSTRACT

The database of the European Chemicals Agency (ECHA) is one of the most important databases that contains physicochemical properties, also because these data are used for the regulation of chemicals in the European Economic Area. The present study investigates the availability and quality of the data in the ECHA database for the logarithmic octanol–water partition coefficient ($\log_{10} K_{OW}$), solubility in water (S_W), vapor pressure (p_V), air–water partition coefficient, boiling point (T_b), second-order rate constant for the degradation with OH radicals, and the soil adsorption coefficient. For the evaluation of the data, calculations were run with COSMO*therm* for the majority of the mono-constituent, neutral organic substances that are fully registered under the EU Regulation on the Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH). The COSMO*therm* data were evaluated against data from the PHYSPROP database, a manually curated database of experimental property data, to ensure that the COSMO*therm* data were free of systematic errors. The comparison between COSMO*therm* and the experimental data in the ECHA database showed that the data agree (within some variability) for many of the endpoints. However, there are also certain ranges with substantial discrepancies. These include $\log_{10} K_{OW} > 8$, $S_W < 10^{-3}$ mg/l, $p_V < 10^{-6}$ Pa, and $T_b > 400\,^{\circ}$C. The deviations between the non-experimental data and the COSMO*therm* values are for all endpoints on average higher than the deviations between the experimental data and the COSMO*therm* values. With this study, we provide COSMO*therm* data for more than 4400 substances that can be used in the future for the hazard and risk assessment of these chemicals.

CONTENTS

List of Tables

List of Figures

1. Introduction

Physicochemical properties such as the octanol–water partition coefficient (K_{OW}), the octanol–air partition coefficient (K_{OA}), vapor pressure (p_V), or the solubility in water (S_W) are important substance-specific properties that determine the environmental fate of substances. K_{OW} influences, for example, the ability of a substance to partition into membranes and can thus be used to determine the baseline toxicity of a substance.[1] The physicochemical properties of a substance are also important parameters in the planning and realization of experiments. Volatile substances, for example, have to be handled differently than non-volatile ones and substances with a low solubility in water need a different experimental setup than substances with a high water solubility.[2] Accurate experimental and non-experimental physicochemical property data are therefore important elements in the risk and hazard assessment of chemicals and various approaches have been taken in the past to evaluate the data quality of larger data sets.[3,4]

One of the largest and most important databases with physicochemical properties is the European Chemicals Agency (ECHA) database that contains data submitted under the EU regulation on

the registration, evaluation, authorization and restriction of chemicals (REACH). Starting in 2007, manufacturers and importers of chemicals had to register their chemicals under REACH if the chemical was manufactured in and/or imported to the European Economic Area (EEA) above 1 tonne/year.[5] The submitted data are very valuable and of great importance for the evaluation of the chemicals. However, the amount of data also presents a major challenge, as not only the substances have to be evaluated, but also the quality of the data has to be assessed. ECHA and the EU Member States evaluate the information substance-by-substance (sometimes also as groups of substances), either during a compliance check or a substance evaluation. However, this is very time-consuming and resource-intensive and therefore not all dossiers and substances have been evaluated so far.[6]

The aim of the present study was therefore to cross-evaluate the physicochemical property data available in the ECHA database. The evaluation here was done for all data points of one endpoint (often several 1000s) and not substance-by-substance. This allowed for the identification of systematic deviations and outliers. For the evaluation of the data, calculations were run with COSMO*therm* for all mono-constituent, neutral organic substances that are fully registered under REACH. To ensure that the COSMO*therm* data themselves did not contain systematic errors, they were evaluated beforehand with data from the PHYSPROP database, a manually curated physical properties database.[7] Based on the results of our study, suggestions for method improvements and future research are provided.

The physicochemical properties (often referred to as endpoints) evaluated in this study are those that are of high importance for the chemical risk and hazard assessments and include: The logarithmic octanol–water partition coefficient ($\log_{10} K_{OW}$), the solubility in water (S_W), the vapor pressure (p_V), the logarithmic air–water partition coefficient ($\log_{10} K_{AW}$), the logarithmic soil adsorption coefficient ($\log_{10} K_{OC}$), the normal boiling point (T_b), and the second-order rate constant for the degradation with OH radicals (k_{OH}). An exact definition of each endpoint is provided in Sec. S1 in the supplementary material 1.

It is important to note that the analysis was carried out with data from the ECHA database because these data are readily available for a large number of substances and the data are used for the regulation of the chemicals in the EEA. However, it may well be that other large data sets containing data on physicochemical properties would show similar percentages of incorrect entries and would lead to similar conclusions.

2. Methods

2.1. Substance data submitted under REACH

The non-confidential substance data submitted to ECHA under the REACH regulation were downloaded from the International Uniform ChemicaL Information Database (IUCLID) website[8] in April 2022 and contained the data from the registration dossiers as of September 15th 2021. The data contained 26 081 registration dossiers for 23 184 substances.[8] The isomeric simplified molecular-input line-entry system (SMILES) string, which describes the chemical structure in line notation, was not available in the IUCLID data. We therefore contacted ECHA and obtained a list of the registered substances with the available IUPAC names, European Community numbers (EC numbers), Chemical Abstract Service Registry Numbers® (CAS RN), molecular formulas, and SMILES notations in January 2021. Additional SMILES codes were obtained manually from the brief profiles of the registered substances in February and May 2022. For substances for which no SMILES code was publicly available in the ECHA database, the SMILES was obtained via the CAS RN from PubChem[9] and/or the CompTox Chemical Dashboard[10] and cross-checked with SciFinder^n.[11] Information on registration status, registration type, substance type, composition, and tonnage band were retrieved in April 2022 from the ECHA website.[12]

The substances investigated in this study are those that are neutral, organic, and mono-constituent with a full registration under REACH [including previously notified substances (NONS)]. Mono-constituent means here that one constituent is present at a concentration of at least 80% (w/w).[13] Intermediates were not included. Substances for which the CAS RN and SMILES notation did not correspond to each other were also not included.[14] Substances that were registered as neutral but had either an acidic pK_a value below 7.4 or a basic pK_a value above 7.4 were also not included (1300 substances), but are listed in the supplementary material 2. The final set of investigated substances consisted of 5318 substances (Table S1 in the supplementary material 1).

For substances with more than one component, only the main component was analyzed. For example, in *N*,*N*′-diallyl-1,3-diaminopropane dihydrochloride, only *N*,*N*′-diallyl-1,3-diaminopropane would be analyzed. Some of the substances with more than one component consisted also of stereoisomers that corresponded to a CAS RN without stereoinformation. The properties of one of the isomers were then used in the analysis. Additional information on the subcomponents of 687 mixtures that did not have a main component but were also not labeled as multi-constituent or unknown or variable composition, complex reaction products or biological materials (UVCB) is provided in the supplementary material 2 as well. However, these data were not included in the analysis since precise information on the composition of the mixtures is lacking. Charged substances were also not analyzed in the current study; information on their physicochemical properties will be published later.

2.2. Physicochemical property data from the ECHA database

Values of the various properties ($\log_{10} K_{OW}$, S_W, p_V, $\log_{10} K_{OA}$, k_{OH}, T_b, $\log_{10} K_{OC}$) as well as information on the physical state were extracted with Python 3.10.4 from the xml documents provided in the IUCLID files. These IUCLID-derived data were linked afterwards with the information provided on the ECHA website[12] (registration status, registration type, origin, etc.) and the SMILES codes. This was done via the EC number and/or the substance name. These two identifiers were the only ones provided in the IUCLID data sets. IUCLID data for substances without EC numbers and names (e.g., for substances that are named "no public or meaningful name is available") could not be assigned and were excluded.

For the analysis of the data, only those values were selected that were not marked as unreliable, were determined for the registered substance, and were in a certain pH and temperature range so that the data points were comparable to other experimental or calculated data points. The specific criteria for removing a study are shown in Table 1.

TABLE 1. Specific criteria for removing studies. Studies that fall under the description in the first column were removed for those endpoints that are listed in the second column

Studies/results that were removed before the data analysis	Affected endpoints
Studies labeled as "not reliable" or that were flagged by the registrants as "disregarded due to major methodological deficiencies"	All
Studies where the reference substance did not match the registered substance[a,b]	All
Studies where the reference substance name contained the string "degradation"	All
Results with the remark "not determinable"	All
Studies that were not conducted in the pH range 7.4 ± 2	$\log_{10} K_{OW}$, S_W, p_V
Studies that were not conducted in the pH range 6.5 ± 2.5	$\log_{10} K_{OC}$
Studies that were not conducted in the temperature range (20 ± 2.5) °C	S_W, p_V, $\log_{10} K_{AW}$
Studies that were not conducted in the temperature range (20 ± 5) °C	$\log_{10} K_{OW}$
Studies that were not conducted in the temperature range (25 ± 5) °C	$\log_{10} K_{OC}$
Studies that were not conducted in the pressure range (1 ± 0.1) atm	T_b
Normal boiling points above the decomposition temperature	T_b
Partition coefficients that were given in the non-logarithmic form but with a negative value	$\log_{10} K_{OW}$, $\log_{10} K_{AW}$, $\log_{10} K_{OC}$
Henry's law constants with the unit "dimensionless," as they corresponded sometimes to the logarithmic and sometimes to the non-logarithmic form	$\log_{10} K_{AW}$
Soil adsorption coefficient given as K_d	$\log_{10} K_{OC}$
Results with negative values	S_W

[a]Was not applied to read-across studies. Read-across as a method "entails the use of relevant information from analogous substances (the "source" information) to predict properties for the "target" substance(s) under consideration."[15]
[b]Substances with non-matching EC-numbers were removed.

The "best available value" is shown in some of the graphics in the supplementary material 1 and is also given in the supplementary material 3. For T_b, $\log_{10} K_{OW}$, $\log_{10} K_{AW}$, and $\log_{10} K_{OC}$, the "best available value" is calculated as the arithmetic mean of all values from studies that were labelled as "key studies" or, if no key study existed, as the arithmetic mean of all other values. For S_W, p_V, and k_{OH}, the "best available value" is calculated as the geometric mean of all values from studies that were labelled as "key studies" or, if no key study existed, as the geometric mean of all other values. Data with qualifier <, >, ≤, or ≥ are not included in the "best available values."

2.3. Experimental physicochemical property data from the PHYSPROP database

The PHYSPROP database contains chemical structures, names, and physical properties for over 43 000 chemicals. It has been developed and is maintained by the Syracuse Research Corporation and can be accessed within the Estimation Program Interface (EPI) Suite™,[16] a property estimation tool provided by the United States Environmental Protection Agency. The entries in the PHYSPROP database are carefully selected and evaluated with the aim to create a high-quality database.[16] Experimental data for the 5318 substances considered in this study were retrieved for $\log_{10} K_{OW}$, S_W, p_V, k_{OH}, T_b, and the Henry's law constant in water (H_W). H_W given in units of atm m^3 mol^{-1} was converted into $\log_{10} K_{AW}$ using Eq. (1).

$$\log_{10} K_{AW} = \log_{10}\left(H_W \frac{c_1}{RT}\right) \quad (1)$$

where R is the universal gas constant (8.314 m^3 Pa K^{-1} mol^{-1}), T the temperature in K, and c_1 a unit conversion factor (101 325 Pa atm^{-1}).

The data from the PHYSPROP database were used to evaluate the calculations carried out with COSMO*therm* but also to define beyond which limits (for each endpoint) a deviation should be considered an outlier. A comparison of the PHYSPROP data with the data in the ECHA database was also performed and is included in the supplementary material 1. However, relatively few of the substances registered under REACH have experimental values in PHYSPROP (see Table S2 in the supplementary material 1). A validation of the data in the ECHA database just with the data from the PHYSPROP database would therefore not have covered a sufficient number of chemicals.

2.4. Calculation of pK_a values

pK_a values were calculated with MarvinSketch 22.18.[17] Substances were treated as neutral if the most acidic pK_a value was above 7.4 and/or the most basic pK_a value below 7.4. "Acidic pK_a" means pK_a values for acidic groups (e.g., COOH), "basic pK_a" means pK_a values for the conjugated acids of basic groups (e.g., NH_2).

2.5. Physicochemical property data from COSMO*therm*

COSMO*therm* and the related program COSMO*conf* are both commercial software applications that are now distributed by BIOVIA.[18] COSMO*conf* is a tool that can generate the conformers of a molecule that are most relevant for interactions with other molecules. These conformers can later be used in COSMO*therm* to estimate the physicochemical properties of a substance. Both programs are based on the Conductor-like Screening Model for Real Solvents (COSMO-RS) theory.[19,20] Compared to many other property estimation methods, the calculations with COSMO*conf* and COSMO*therm* are not based on training sets of physicochemical property data; instead they are based on chemical potential differences of molecules immersed in solvents. The data that originate from calculations with COSMO*conf* and COSMO*therm* are internally consistent and have been shown to be more accurate than the data from many other estimation programs.[21–23] A very recent publication showed that COSMO*therm* was not only superior to other estimation programs but also that the calculated $\log_{10} K_{AW}$ values for 21 per- and polyfluoroalkyl substances were within one $\log_{10}$ unit of the experimental data.[24] COSMO*therm* is also able to calculate physicochemical properties for various temperatures, which is especially important for strongly temperature-dependent endpoints such as p_V or $\log_{10} K_{OA}$.[25]

COSMO*conf* was run on the ETH "Euler Cluster" on around 200 central processing units (CPUs) in parallel over almost 3 years. Calculations in COSMO*conf* for small molecules (<12 atoms) are very fast and are done within minutes. However, the typical calculation time increases exponentially with the number of atoms in a molecule and can be in the range of weeks for molecules with 80 atoms or more. We ran COSMO*conf* therefore only for molecules that had fewer than 80 atoms. This covered around 87% of the substances. 5% of the calculations also failed and no results could be obtained. The parametrization used in COSMO*conf* was BP-TZVPD-FINE COSMO+GAS.

Calculations in COSMO*therm* are very fast and take per molecule and endpoint only a few minutes. The calculations were run for $\log_{10} K_{OW}$, $\log_{10} K_{AW}$, S_W, p_V, k_{OH}, T_b, $\log_{10} K_{OC}$, and the Henry's law constant in octanol (H_O). The parametrization used in COSMO*therm* was BP-TZVPD-FINE for all endpoints, except the $\log_{10} K_{OC}$. Data for $\log_{10} K_{OC}$ are only available with the BP-TZVP parametrization. The calculations in COSMO*therm* were performed at 20 and 25 °C for all endpoints (except $\log_{10} K_{OC}$, for which calculations were only performed at 25 °C). H_O given in units of bar was converted into $\log_{10} K_{OA}$ using Eq. (2).

$$\log_{10} K_{OA} = \log_{10}\left(\frac{RT\rho c_2}{H_O M}\right) \quad (2)$$

where R is the universal gas constant, T the temperature in K, ρ is the density of octanol (at temperature T), M the molecular weight of octanol, and c_2 a unit conversion factor (10 cm^3 bar m^{-3} Pa^{-1}). With ρ = 0.824 g/cm^3 and M = 130.23 g/mol, Eq. (2) can be shortened (at 25 °C) to

$$\log_{10} K_{OA} = \log_{10}(156.8 \text{ bars}/H_O).$$

2.6. Statistical analysis

Two kinds of statistical analysis were conducted in this study. The root-mean-square error (RMSE), R-square (r^2), and the mean absolute error (MAE) were used to measure the differences/similarity between two data sets. For the calculation of the values, the scikit-learn 1.2.2 package in Python 3.10.4 was used.[26] In addition, for each endpoint, the limits beyond which a deviation should be considered an outlier, i.e., instances where there is clearly no agreement between measurement and prediction, were determined. For a normal distribution, one could quantify this via the standard deviation. Outliers would be detected if the values are outside of three standard deviations. In this study, we adjusted this procedure slightly.

(1) For each endpoint, the errors between the experimental data from the PHYSPROP database and COSMO*therm* were calculated. The error in this case was the difference between a datapoint in the PHYSPROP database and the COSMO*therm* value. For p_V, S_W, and k_{OH}, this was done for the $\log_{10}$-transformed data.
(2) Subsequently, it was checked if the distribution of the errors followed a normal distribution. This was done by running the Shapiro-Wilk Test in Python using the SciPi package[27] and the shapiro function.
(3) For all endpoints where the error distribution did not follow a normal distribution, the error distribution was analysed visually and if it was symmetric and similar in shape to a normal distribution, the 2.3rd and 97.7th percentile of the errors were calculated. The values within these percentiles cover 95.4% of the errors which corresponds to two standard deviations in a normal distribution. We chose two standard deviations and not three (which would be the default) since we wanted to sort out not only outliers but also values with high deviation in general.
(4) To set the final threshold for the detection of outliers and other values with high deviation, the arithmetic mean of the 2.3rd and 97.7th percentile of the errors was calculated.

3. Results

3.1. Comparison between the experimental data from the PHYSPROP database and data generated with COSMO*therm*

For the substances investigated in this study, 823, 853, 852, 1050, 597, and 392 experimental data points were available in the PHYSPROP database for $\log_{10} K_{OW}$, S_W, p_V, T_b, H_W, and k_{OH}, respectively. The comparison between these data points and the ones generated with COSMO*therm* is shown in Fig. 1; the corresponding statistics are presented in Table 2. The data points agree in general quite well for $\log_{10} K_{OW}$, p_V, $\log_{10} K_{AW}$, and T_b (all have R-square values around 0.9). Higher deviations are observed for S_W and

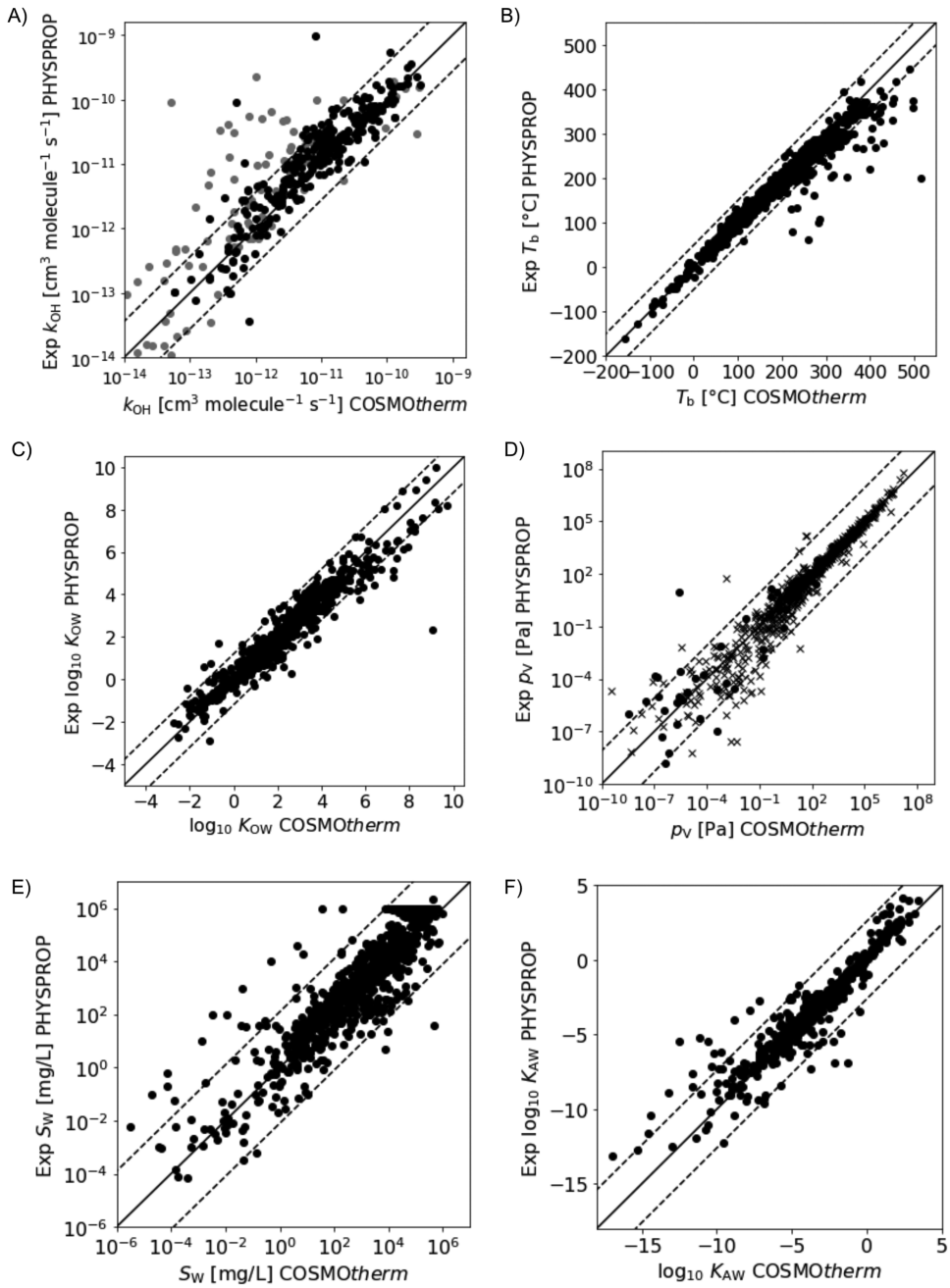

FIG. 1. Experimental data from the PHYSPROP database for the substances investigated in this study compared to data generated with COSMO*therm*. (A) Second-order rate constant for the degradation with OH radicals. Data points in grey are for substances that contain nitrogen, sulfur, phosphor, or fluorine. (B) Normal boiling point. (C) Logarithmic octanol–water partition coefficient. (D) Vapor pressure. x values for 25 °C, • values for 20 °C. (E) Solubility in water. (F) Logarithmic air–water partition coefficient. The solid line in all plots is the 1:1 line. The dashed lines indicate the thresholds for the outlier detection/values with high deviation, as given in Table 2.

TABLE 2. Statistics for the comparison of the experimental data points from the PHYSPROP database and COSMO*therm* for the substances investigated in this study

	$\log_{10} K_{OW}$	$\log_{10} p_V$ ($\log_{10}$ Pa)	$\log_{10} K_{AW}$	$\log_{10} S_W$ ($\log_{10}$ mg/l)	T_b (°C)	$\log_{10} k_{OH}$ [$\log_{10}$ cm^3/(molecule s)]
RMSE	0.58	0.89	1.1	1.04	32.9	0.36
r^2	0.92	0.91	0.88	0.76	0.90	0.79
MAE	0.38	0.48	0.63	0.70	16.0	−0.14
Does the error distribution follow a normal distribution?	No (close)	No	No	No (close)	No	Possibly
Two times standard deviation of the normal distribution	1.2	⋯	⋯	2.0	⋯	0.66
2.3th percentile (of errors)	−0.99	−1.5	−3.03	−2.41	−19.7	−0.72
50th percentile (of errors)	0.04	0.05	−0.19	−0.27	3.06	−0.1
97.7th percentile (of errors)	1.36	2.22	2.20	1.69	81.2	0.4
Threshold for the detection of outliers and other values with high deviation	1.2	1.9	2.6	2.1	50	0.56

TABLE 3. Validated ranges for the COSMO*therm* data based on the comparison with the experimental data from the PHYSPROP database

Endpoint	Validated range
Logarithmic octanol–water partition coefficient ($\log_{10} K_{OW}$)	−3 to 10
Solubility in water (S_W)	10^{-4} to 10^{6} mg/l
Logarithmic air–water partition coefficient ($\log_{10} K_{AW}$)	−10 to 4
Vapor pressure (p_V)	10^{-8} to 10^{8} Pa
Normal boiling point (T_b)	−200 to 400 °C
Second-order rate constant for the degradation with OH radicals (k_{OH})	10^{-13} to 5×10^{-9} cm^3/(molecule s) for substances that include the elements C, H, and O

k_{OH} with R-square values of 0.76 and 0.79, respectively. A detailed analysis of the data is provided in the supplementary material 1 Sec. S3.

From this comparison, we derived for each endpoint a validated range for the COSMO*therm* data (Table 3). The validated range gives for each endpoint the lower and upper bound for which PHYSPROP data were available; not counting those values from the PHYSPROP database that were outside of the thresholds defined in Table 2. The data outside the validated range are not *per se* inaccurate, but the uncertainty in these data is larger. A discussion of these data per endpoint is provided in Sec. 4.2.

3.2. Availability of data in the ECHA database

The available data in the ECHA database differ between full registrations and NONS (Fig. 2). NONS registrations are for substances that were already registered under the previous regulation (Directive 67/548/EEC). These substances were automatically transferred into REACH and are regarded as already registered. Companies were then able to claim the registrations as their own.

For substances with full registrations, the largest numbers of data points are available for $\log_{10} K_{OW}$, T_b, S_W, and p_V. These endpoints are required for all substances that are manufactured or imported in quantities of 1 tonne/year or more.[28] Most of the submitted data for these four endpoints were experimental data (Fig. 2 left). Data for $\log_{10} K_{AW}$ and k_{OH} are not required for any tonnage band and were therefore only provided in some registration dossiers. The submitted data for these two endpoints are a mixture of experimental data, (Quantitative) Structure-Activity Relationships [(Q)SAR] results, and other calculations.

Most of the submitted data for the NONS only include the main result but no further information on the study type [experimental study, calculation, results from (Q)SAR, etc.] or other details (Fig. 2 right). This is because dossiers for NONS complying with the standard information requirements need to be submitted to ECHA only when the dossiers are updated to increase the tonnage band.[29,30] According to Chemsafe-Consulting (2021), this happened until 2021 in only 8% of the cases. Otherwise, the dossier does not need to include information that was not required under the previous legislation.[30]

It is important to note that, for many of the endpoints shown in Fig. 2, more than half of the studies submitted under REACH (and shown in Fig. 2) are either labeled as not reliable, were not conducted for the registered substance, or were outside the selected pH or temperature range. The final number of investi-

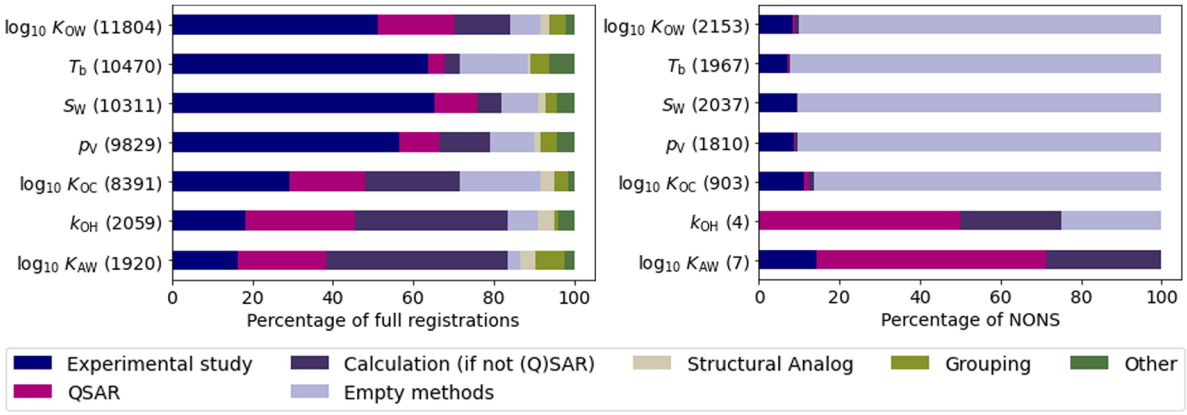

FIG. 2. Data availability for organic mono-constituent substances. Left: Full registrations. Right: NONS (notifications of new substances). The study types are shown with different colors.

gated substances and studies (excluding those just mentioned) is provided in Table S2 in the supplementary material 1. The studies that are listed in Table S2 are also the ones that were analyzed in Sec. 3.3.

3.3. Evaluation of the data in the ECHA database

The calculated COSMO*therm* values that were used to evaluate the data from the ECHA database are provided in the supplementary material 2. Moreover, the supplementary material 3 includes for all endpoints those substances for which the "best available value" deviates more than the endpoint-specific threshold from the COSMO*therm* values. The evaluation of the $\log_{10} K_{OC}$ values is provided in Sec. S5.7 in the supplementary material 1; all other endpoints are described in the following subsections.

3.3.1. Octanol-water partition coefficient

Relevant experimental data for $\log_{10} K_{OW}$ were submitted for 2425 substances and included a total of 3390 studies (Table S2, supplementary material 1). 2994 of these studies could be analyzed by comparison with the COSMO*therm* data. Relevant non-experimental data for $\log_{10} K_{OW}$ were submitted for 2702 substances and included a total of 4140 studies. Out of these, 3448 could be analyzed by comparison with the COSMO*therm* data. Figure 3 shows the data points for the experimental as well as non-experimental $\log_{10} K_{OW}$ values.

3.3.1.1. Experimental data points. The $\log_{10} K_{OW}$ values from the ECHA database with no qualifier or qualifier "ca." agree with a R-square value of 0.70 (Table S5, supplementary material 1) with the calculated values from COSMO*therm* (Fig. 3, left). 4.3% of the

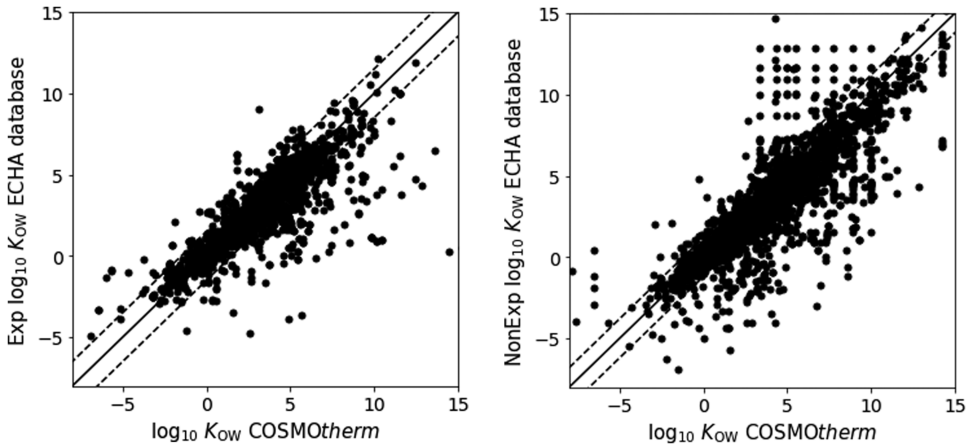

FIG. 3. Comparison of the experimental (left) and non-experimental (right) $\log_{10} K_{OW}$ values from the ECHA database with no qualifier or qualifier "ca." with the $\log_{10} K_{OW}$ from COSMO*therm* for neutral mono-constituent organic substances (full registrations and NONS). The solid line is the 1:1 line; the dashed lines indicate a deviation of 1.2 units. COSMO*therm* values refer to 20 °C, values from the ECHA database to (20 ± 5) °C.

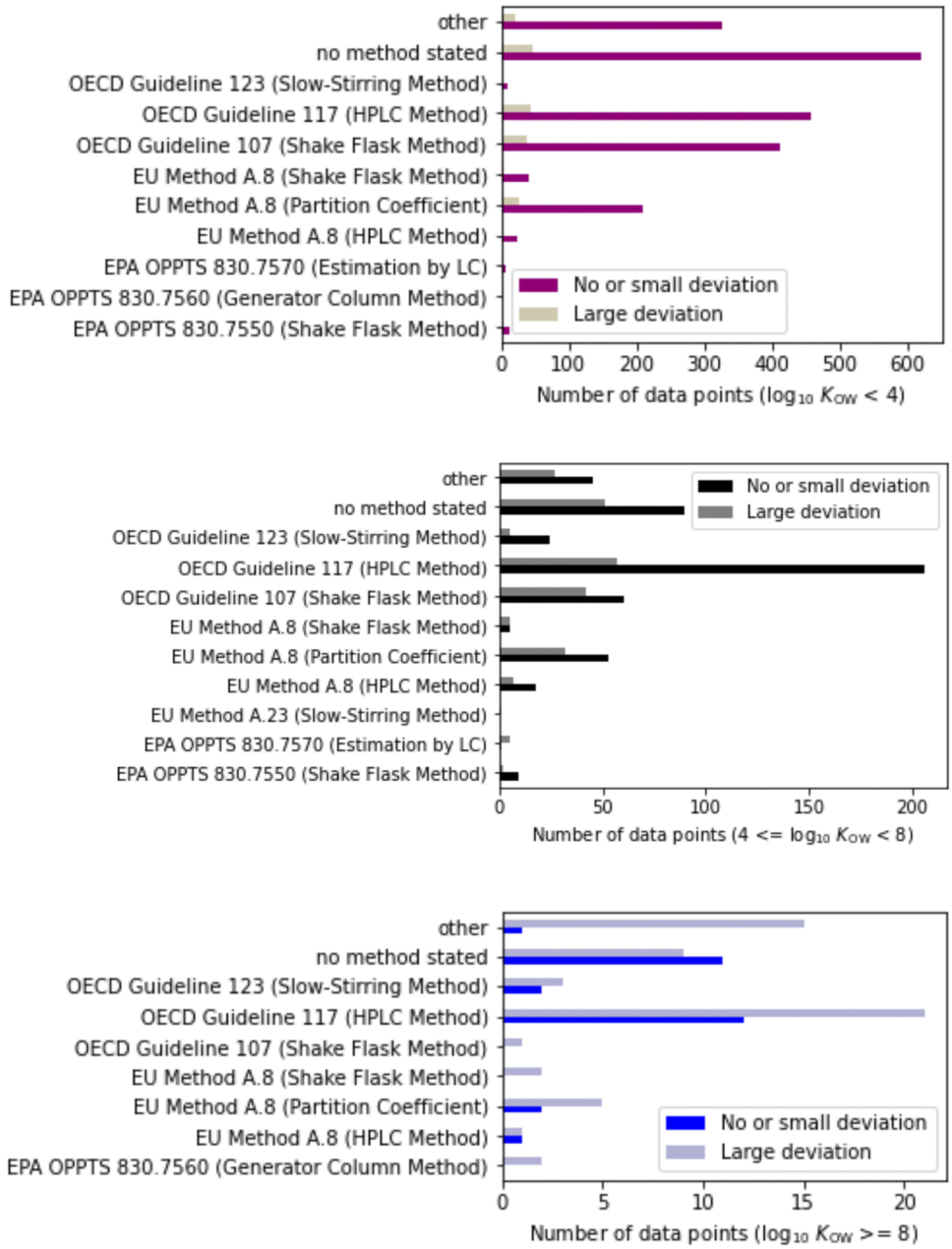

FIG. 4. Methods used for the determination of $\log_{10} K_{OW}$. "No or small deviation" and "Large deviation" refer to the deviation from the COSMO*therm* values. Large deviations means that the deviation is more than 1.2 units.

ECHA data are more than 1.2 units above the COSMO*therm* values, 11% are more than 1.2 units below the COSMO*therm* data points. Higher deviations are observed for the more hydrophobic substances. A few of the detected outliers and values with higher deviation are outside the range validated here with the data from the PHYSPROP database ($\log_{10} K_{OW} > 10$), but most of them are within the range. More information on the outliers and values with high deviation can be gained from a closer look at the K_{OW} measurement methods.

The guidance on information requirements and chemical safety assessment - Chapter R.7a: Endpoint specific guidance[31] recommends to use certain methods for specific ranges of $\log_{10} K_{OW}$. Specifically, the shake flask method (EU A.8, OECD TG 107) for $-2 < \log_{10} K_{OW} < 4$; the high-performance liquid chromatography

(HPLC) method (EU A.8, OECD TG 117) for $0 < \log_{10} K_{OW} < 6$ (10 in exceptional cases); and the slow-stirring method (OECD TG 123) for $\log_{10} K_{OW}$ values up to 8.3. The plots in Fig. 4 are therefore also divided into $\log_{10} K_{OW}$ ranges (<4, 4–8, and >8).

All methods perform (judged by comparison to the COSMO*therm* values) quite well for substances with $\log_{10} K_{OW} < 4$ (Fig. 4, top). There are only a couple of data points that deviate by more than 1.2 units from the COSMO*therm* values and no method seems to be better or worse than the other. For substances with $\log_{10} K_{OW}$ between 4 and 8, most measurements were performed with the HPLC method, which seems to be the best-suited method together with the slow stirring method (OECD TG 123). However, some data points were also obtained with the shake flask method (which is not recommended for this range) or without a method stated, and these partition coefficients deviate in more than one-third of the cases by more than 1.2 units from the COSMO*therm* values. It is therefore highly recommended to use the HPLC or slow stirring method for future measurements in the $\log_{10} K_{OW}$ range 4–8. No method performed very well for substances with $\log_{10} K_{OW} > 8$. Most of the results were again obtained with the HPLC methods, but 21 of the 33 partition coefficients obtained by HPLC deviate by more than 1.2 units from the COSMO*therm* values.

The $\log_{10} K_{OW}$ values from the ECHA database with qualifier > or ≥ and qualifier < and ≤ are shown in Fig. S5 in the supplementary material 1. The experimental $\log_{10} K_{OW}$ values for substances with a calculated $\log_{10} K_{OW} < 5$ and qualifier > or ≥ are very similar to the COSMO*therm* values. However, more than half of the experimental $\log_{10} K_{OW}$ values with a (calculated) $\log_{10} K_{OW} > 5$ and a qualifier > or ≥ are lower by more than 1.2 units than the COSMO*therm* values. For the experimental data points with qualifier < or ≤, the deviations are less severe.

3.3.1.2. Non-experimental data points. The partition coefficients that are not from experimental studies but from (Q)SARS, calculation, read across, and unspecified methods are shown in Fig. 3 (right). The R-square value is 0.64 and thus lower than the value for the experimental partition coefficients. One reason for this is that the registration dossiers of five alkenes contained 75 or 150 non-experimental data points for $\log_{10} K_{OW}$ that ranged for all five alkenes from 1.5 to 16.3 (vertical lines in Fig. 3 right). These $\log_{10} K_{OW}$ values were obtained from (Q)SARs and read-across of structurally similar compounds that had different chain lengths. However, it is even stated in the summary of e.g., dodec-1-ene that "Partition coefficient increases with increasing carbon number across the category, with a very similar trend for both the alpha olefins and olefins." The partition coefficients of substances with different chain lengths should therefore not have been used for the registered compounds.

If the partition coefficients for the five alkenes are disregarded and the remaining $\log_{10} K_{OW}$ values are compared to the COSMO*therm* values, 20% of the non-experimental data points in the ECHA database are still at least 1.2 units below the COSMO*therm* values and 6.3% are at least 1.2 units above the COSMO*therm* values. These values are still higher than the comparable values for the experimental data points and show that (Q)SARs and calculation methods need to be selected more carefully than currently done.

3.3.2. Solubility in water

Relevant experimental data for the solubility in water were submitted for 2170 substances and included a total of 2851 studies (Table S2, supplementary material 1). 2358 of these studies could be analyzed by comparison with the COSMO*therm* data. Relevant non-experimental data for the solubility in water were submitted for 1324 substances and included a total of 1656 studies. Out of these, 1321 studies could be analyzed by comparison with the COSMO*therm* data. Figure 5 shows the data points for the experimental as well as non-experimental solubilities in water.

3.3.2.1. Experimental data points. The solubilities in water from the ECHA database with no qualifier or qualifier "ca." agree with a R-square value of 0.72 (Table S5, supplementary material

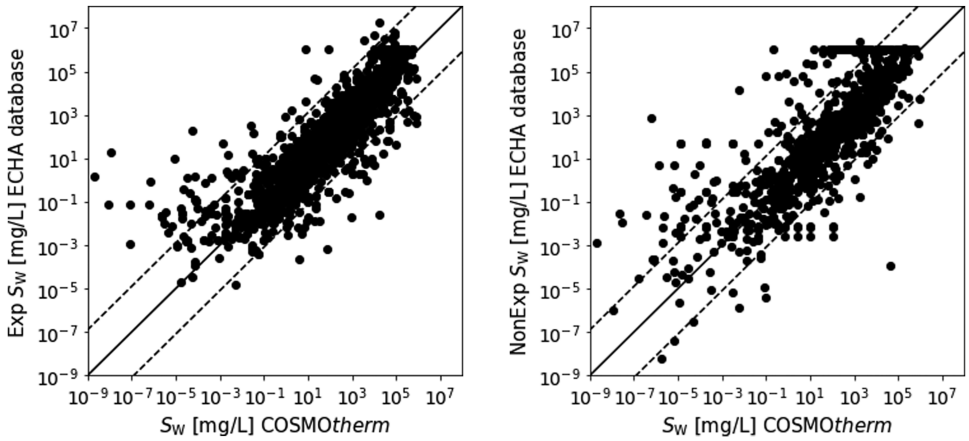

FIG. 5. Comparison of the experimental (left) and non-experimental (right) values for the solubility in water from the ECHA database with no qualifier or qualifier "ca." with the solubility from COSMO*therm* for neutral organic substances with one component (full registrations and NONS). The solid line is the 1:1 line; the dashed lines indicate a deviation of 2.1 orders of magnitude. COSMO*therm* values refer to 20 °C, values from the ECHA database to (20 ± 2.5) °C.

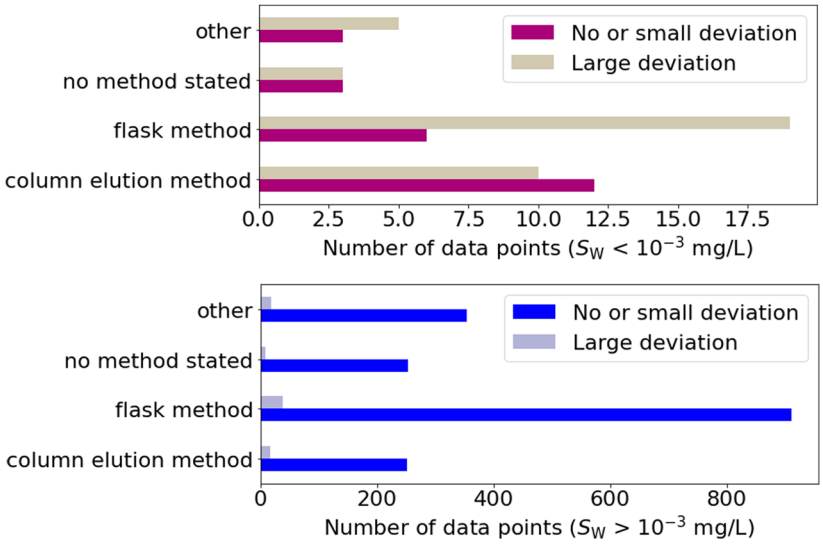

FIG. 6. Methods used for the determination of the solubility in water. "No or small deviation" and "Large deviation" refer to the deviation to the COSMO*therm* values. Large deviations means that the deviation is more than 2.1 orders of magnitude.

1) with the calculated values from COSMO*therm*. 3.9% of the ECHA data are more than 2.1 orders of magnitude above the COSMO*therm* values; 3.4% are more than 2.1 orders of magnitude below the COSMO*therm* values. The highest deviations are observed for substances with a solubility in water below 10^{-3} mg/l. Some of these solubilities are outside the range validated here with the PHYSPROP data ($S_W < 10^{-4}$ mg/l). A discussion of the accuracy of these data is provided in Sec. 4.2.

Figure 6 shows the method types (flask method, column elution method, other) underlying the experimental data in the ECHA database. The general test guidelines (OECD TG 105, EU method A.6, EPA OPPTS 830.7840) are also provided in IUCLID, however all three guidelines include both the flask and column elution method and are therefore not really useful for method selection. According to the guidance on information requirements and chemical safety assessment - Chapter R.7a,[31] the column elution method and the flask method with slow stirring are appropriate for low-solubility test substances ($S_W < 10$ mg/l). The flask method with fast stirring is appropriate for higher-solubility test substances ($S_W > 10$ mg/l).[31] No specific recommendations are given for substances with a very low solubility ($S_W < 10^{-3}$ mg/l).

In the data from the ECHA database, no distinction is possible between the flask method with fast and slow stirring (Fig. 6). This makes it very difficult to judge the performance of the individual methods. However, we can say that both the flask method and the column elution method worked well (with the expected variation) for substances with solubilities above 10^{-3} mg/l. For substances for which COSMO*therm* calculated a solubility below 10^{-3} mg/l, neither of the methods worked well. But again, it was not possible to distinguish between fast and slow stirring so it cannot be said *per se* that also the flask method with slow stirring would not work.

The solubilities in water from the ECHA database with qualifier > or ≥ and qualifier < and ≤ are shown in Fig. S7 in the supplementary material 1. Almost all experimental solubilities in water with qualifier > or ≥ deviate less than 2.1 orders of magnitude from the COSMO*therm* values. The water solubilities in the ECHA database with qualifier < or ≤ show much larger deviations. Especially the values for substances with a calculated solubility of less than 10^{-4} mg/l are in almost all cases more than 2.1 orders of magnitude above the COSMO*therm* values. This is not problematic *per se* because the qualifier says that the values are the upper limit, but this also means that they should not be seen as real values, but as upper bounds.

3.3.2.2. Non-experimental data points. The non-experimental values for the solubility in water from the ECHA database are shown in Fig. 5 (right). The R-square value is 0.57. There are 58 data points that have a water solubility of 10^6 mg/l, i.e., 1 kg per kg of water in the ECHA database, many of them far above the solubility calculated by COSMO*therm*. It was not possible to determine what caused these many occurrences of the same value. 10^6 mg/l seems to be the maximum solubility that WATERNT in EPI Suite predicts.[32] However, this does not explain why many of the data points are much higher than the COSMO*therm* values. The meaning of these values is entirely unclear, and we recommend that they should not be used.

Without these values, 5.1% of the non-experimental water solubilities in the ECHA database are still at least 2.1 orders of magnitude below the COSMO*therm* values and 9.0% are at least 2.1 orders of magnitude above the COSMO*therm* values. The deviations are significantly higher for substances with a very low solubility in water ($S_W < 10^{-3}$ mg/l).

3.3.3. Vapor pressure

Relevant experimental data for the vapor pressure were submitted for 1924 substances and included a total of 2515 studies (Table S2, supplementary material 1). 2254 of these studies could

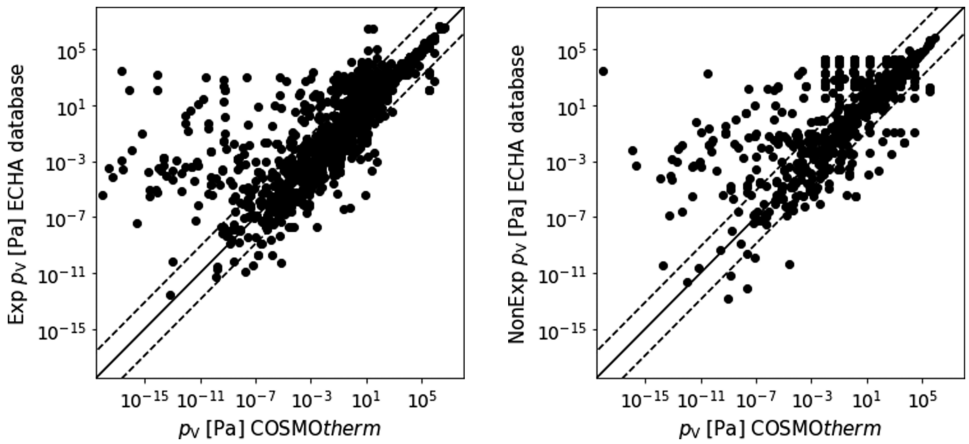

FIG. 7. Comparison of the experimental (left) and non-experimental (right) vapor pressure from the ECHA database with no qualifier or qualifier "ca." with the data from COSMO*therm* for neutral organic substances with one component (full registrations and NONS). The solid line is the 1:1 line, the dashed lines indicate a deviation of 1.9 orders of magnitude. COSMO*therm* values refer to 20 °C, values from the ECHA database to (20 ± 5) °C.

be analyzed by comparison with the COSMO*therm* data. Relevant non-experimental data for the vapor pressure were submitted for 775 substances and included a total of 1112 studies. Out of these, 962 could be analyzed by comparison with the COSMO*therm* data. Figure 7 shows the data points from the ECHA database for the experimental as well as non-experimental vapor pressure compared to the data from COSMO*therm*.

3.3.3.1. Experimental data points. The experimental vapor pressures from the ECHA database with no qualifier or qualifier "ca." show large deviations for some substances (for some, more than 10 orders of magnitude) from the calculated values from COSMO*therm* (Fig. 7). The overall R-square value is 0.63. 10.5% of the ECHA data are more than 1.9 orders of magnitude above the COSMO*therm* values; 4.3% are more than 1.9 orders of magnitude below the COSMO*therm* data points. Higher deviations are observed for the less volatile substances. Many of these data points are outside the range validated here with the data from the PHYSPROP database for COSMO*therm* ($<10^{-8}$ Pa). However, high deviations are also seen within the validated range.

Figure 8 shows the method types used for the experimental data in the ECHA database. Commission Regulation No. 761/2009 Annex I (A.4 Vapor pressure) recommends – with some restrictions – for substances with a vapor pressure above 1 Pa the static, dynamic, isoteniscope, or the gas saturation method.[33] For substances with a vapor pressure between 0.001 and 1 Pa, one of the three effusion methods, the gas saturation methods, or the spinning rotor method are recommended. For substances with a vapor pressure below 0.001 Pa, the Knudson cell effusion method, the isothermal thermogravimetry effusion method, the gas saturation method, and, in part, the spinning rotor method are recommended.[33]

In the ECHA database, substances with a vapor pressure (calculated with COSMO*therm*) above 1 Pa were mainly analyzed with the static or dynamic method and the deviations are quite small compared to the values from COSMO*therm* (Fig. 8). Although the dynamic and the static method are not recommended for substances with a vapor pressure below 1 Pa, they were also used in 29% of the cases for substances with a calculated vapor pressure between 0.001 and 1 Pa. Especially the results from the static method show large deviations from the COSMO*therm* values and we recommend not using these values. For substances with a calculated vapor pressure below 0.001 Pa, no method performed well. Most data in the ECHA database for these substances were derived with the vapor pressure balance method, although this method is not recommended for substances with a vapor pressure below 0.001 Pa. However, none of the methods that were recommended in Commission Regulation No. 761/2009 for a vapor pressure below 0.001 Pa performed better (judged by comparison to the COSMO*therm* results). The worst results were obtained with the static method. Only 1 out of 39 values shows a deviation from the COSMO*therm* values of less than 1.9 orders of magnitude.

3.3.3.2. Non-experimental data points. The non-experimental values for the vapor pressure from the ECHA database are shown in Fig. 7, right. The R-square value is 0.47. 6.3% of the vapor pressures in the ECHA database are at least 1.9 orders of magnitude below the COSMO*therm* values and 22.3% are at least 1.9 orders of magnitude above the COSMO*therm* values. The deviations are significantly higher for substances with a low vapor pressure ($<10^{-6}$ Pa).

3.3.4. Air–water partition coefficient

Relevant experimental data for the Henry's law constant in water were submitted for 54 substances and included a total of 62 studies (Table S2, supplementary material 1). All of these studies could be analyzed by comparison with the COSMO*therm* data. Relevant non-experimental data for the Henry's law constant in water were submitted for 227 substances and included a total of 322 studies. Out of these, 314 could be analyzed by comparison with the COSMO*therm* data. The experimental and non-experimental data agreed with a R-square value of 0.74 and 0.70, respectively (Fig. S11).

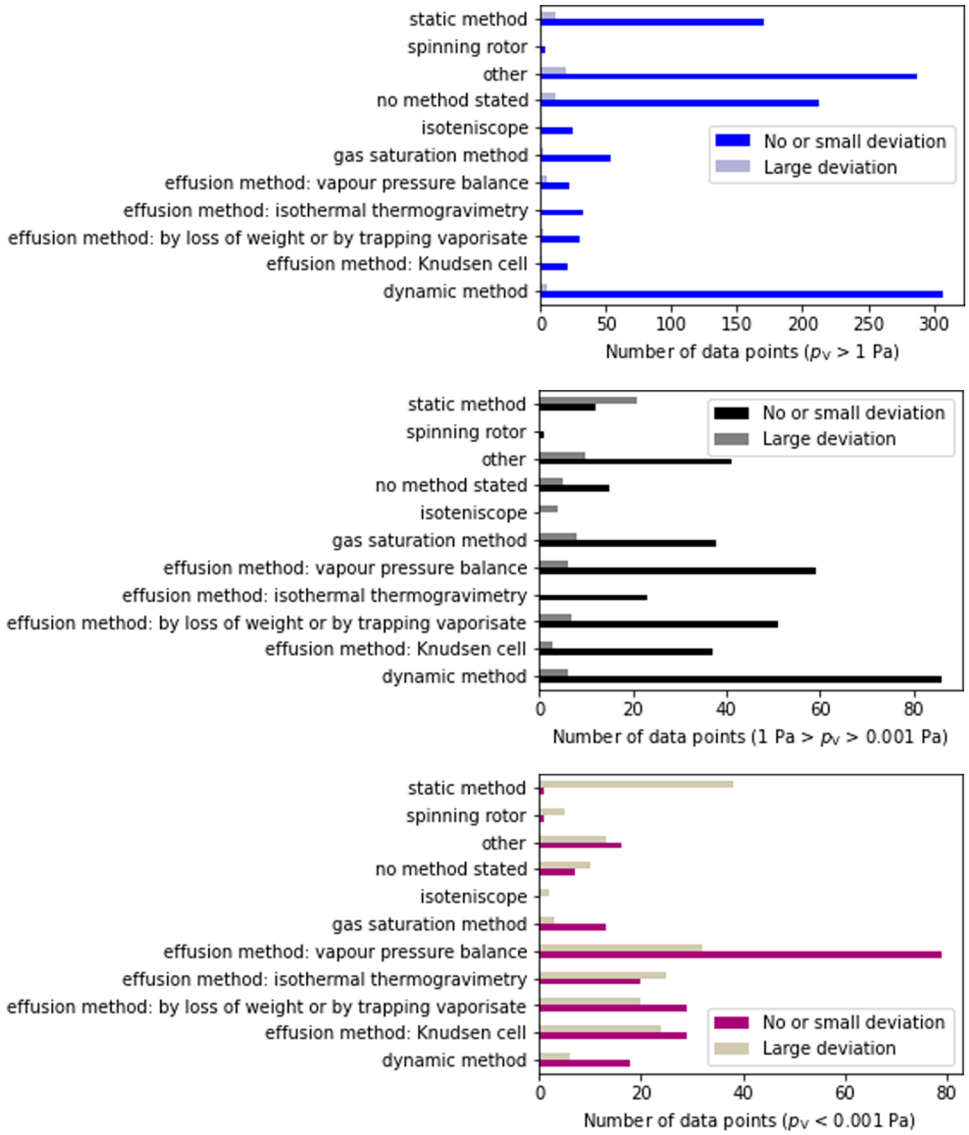

FIG. 8. Methods used for the determination of the vapor pressure. "No or small deviation" and "Large deviation" refer to the deviation to the COSMO*therm* values. Large deviations means that the deviation is more than 1.9 orders of magnitude.

3.3.5. Normal boiling point

Relevant experimental data for the normal boiling point were submitted for 2263 substances and included a total of 4295 studies (Table S2, supplementary material 1). 4035 of these studies could be analyzed by comparison with the COSMO*therm* data. Relevant non-experimental data for the normal boiling point were submitted for 1432 substances and included a total of 2431 studies. Out of these, 2152 could be analyzed by comparison with the COSMO*therm* data. Figure 9 shows the experimental as well as non-experimental normal boiling points.

3.3.5.1. Experimental data points. The experimental normal boiling points from the ECHA database with no qualifier or qualifier "ca." show large deviations from the COSMO*therm* values for some substances (Fig. 9 left). The overall R-square value is 0.67. 0.4% of the ECHA data are more than 50 °C above the COSMO*therm* values; 12% (476 of 3851 data points) are more than 50 °C below the COSMO*therm* data points. The highest deviations are visible for substances with (calculated) normal boiling points above 500 °C, which is also outside the range validated here with the data from the PHYSPROP database. A discussion of the accuracy of these points is provided in Sec. 4.2. Some of the calculated normal boiling

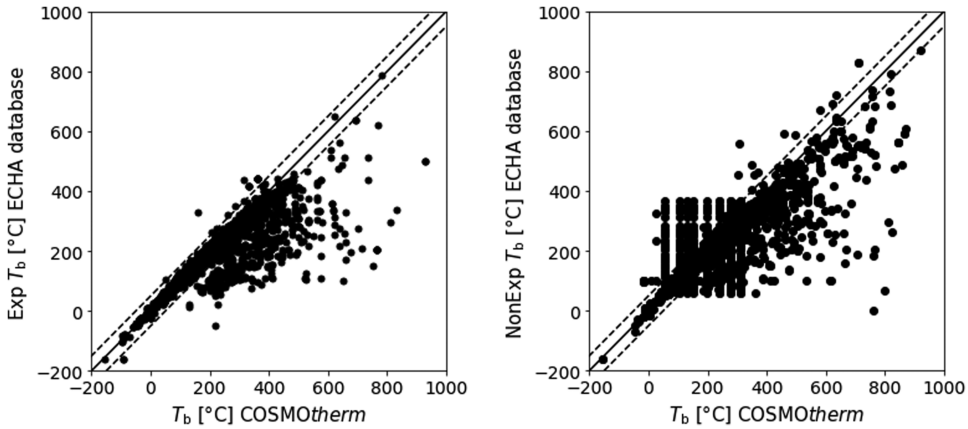

FIG. 9. Comparison of the experimental (left) and non-experimental (right) normal boiling points from the ECHA database with no qualifier or qualifier "ca." with the data from COSMO*therm* for neutral organic substances with one component (full registrations and NONS). The solid line is the 1:1 line; the dashed lines indicate a deviation of 50 °C.

points might be above the decomposition temperature of the substance. However, we were not able to check this in all cases because decomposition temperatures were not available for all substances. No differences in accuracy were observed regarding the applied methods (Fig. S14). This is also in line with the guidance on information requirements and chemical safety assessment - Chapter R.7a, where it is stated that any determination method may be used within the scope and applicability specifications.[31]

3.3.5.2. Non-experimental data points. The non-experimental values for the normal boiling point from the ECHA database are shown in Fig. 9 right. The R-square value is 0.56. Similar to the octanol–water partition coefficient, the registration dossiers of several alkenes (this time eight alkenes) contained 46 or 92 non-experimental data points for the normal boiling point that ranged for all eight alkenes from 60 to 367 °C (vertical lines in Fig. 9 right). The normal boiling points were obtained from QSARs and read-across of structurally similar compounds that had different chain lengths and were actually not suitable for read-across.

Without these substances, 26% of the normal boiling points in the ECHA database are at least 50 °C below the COSMO*therm* values and 8.0% are at least 50 °C above the COSMO*therm* values. Again, the deviations are significantly higher for substances with a normal boiling point above 500 °C.

3.3.6. Photodegradation in air with OH radicals

For the photodegradation in air with OH radicals, the results are often stored in fields like "Remarks on Results" that are not exported to IUCLID. It was therefore only possible to extract experimental rate constants from 34 studies although 195 were labeled as experimental studies, which indicates that there might be more information on the ECHA website. We therefore manually extracted additional results and added these to the IUCLID data. The finally analyzed data set contains experimental data for 110 substances and includes in total 141 studies (Table S2). 136 of these studies could be analyzed by comparison with the COSMO*therm* data. At least 20 studies were also labeled as experimental although they originated from (Q)SARs or calculations.

Non-experimental data for the photodegradation in air with OH radicals were available in IUCLID for 167 substances and included a total of 183 studies. Out of these, 164 could be analyzed by comparison with the COSMO*therm* data. No additional data points were extracted manually for the non-experimental data. The available data points agree for some substances quite well with the calculated values from COSMO*therm* [Figs. S11(c) and S11(d)]. However, some studies report very unrealistic values like −12 cm^3 $molecule^{-1}$ s^{-1} (CAS RN 507-20-0) or 1.24 cm^3 $molecule^{-1}$ s^{-1} (CAS RN 420-46-2), which also causes the very low R-square value of 0.05.

4. Discussion

4.1. Data basis for the validation

The approach that was chosen here – validation of the COSMO*therm* data with data from the manually curated PHYSPROP database and a subsequent validation of the data in the ECHA database with the COSMO*therm* data – has the advantage that far more data can be validated in the ECHA database than if only the PHYSPROP data had been used (Table S2). COSMO*therm* is therefore an important element of our approach because the calculations in COSMO*therm* can be used to evaluate data where no other reliable measurements or (Q)SARs are available. This is for example visible for the 665 substances whose "best available" $\log_{10} K_{OW}$ values deviate more than expected (based on the variation normally observed for $\log_{10} K_{OW}$ values from the COSMO*therm* values). For these 665 substances, only 60 have experimental data in the PHYSPROP database and could have been validated with these values (supplementary material 3).

Additionally, it is important to note that some of the values from the PHYSPROP database may also be contained in the ECHA database. Including these values directly in the validation of the data

from the ECHA database leads to a bias, suggesting an overall accuracy of the data in the ECHA database that might not be real. This means that even if we see high R-square values for the comparison, this does not imply that all other values in the ECHA database would also be very accurate. However, for completeness, the comparison between the experimental ECHA data and those in the PHYSPROP database is shown in Sec. S4 of the supplementary material 1.

4.2. Accuracy of the COSMO*therm* values outside the validated range

The data in the PHYSPROP database did not cover for all endpoints the entire range of values that were calculated by COSMO*therm*. The following sections therefore discuss the accuracy of the values outside of this validated range. The air–water partition coefficients and the second-order rate constants for the photodegradation in air with OH radicals are not discussed, as the data points in the PHYSPROP database cover the same range as the COSMO*therm* values for these endpoints.

4.2.1. Octanol–water partition coefficient

The $\log_{10} K_{OW}$ values in the PHYSPROP database cover the range from −3 to 10. For a few substances, COSMO*therm* calculated $\log_{10} K_{OW}$ values that are two $\log_{10}$ units below or above this range. We assume that also the values outside this validated range are reasonable, because the values from COSMO*therm* agree over the entire $\log_{10} K_{OW}$ range very well with the data from the PHYSPROP database.

4.2.2. Solubility in water

COSMO*therm* calculated solubilities in water that are up to two $\log_{10}$ units below the validated range (10^{-4} to 10^{6} mg/l), for a few substances even five orders of magnitude below. Especially the very low calculated solubilities in water (10^{-6} to 10^{-9} mg/l) are all below the corresponding experimental values, sometimes up to 5 or 6 orders of magnitude. Without a confirmation with experimental values, it is very difficult to judge how reliable the low solubilities in water calculated by COSMO*therm* are. On the other hand, solubilities in water in the ng/l range are also not easy to measure and there is *per se* no reason why the COSMO*therm* values should be incorrect as they are all based on the same calculation principle. We would therefore still recommend using the very low solubilities from COSMO*therm*, but stating that they are associated with a higher uncertainty.

4.2.3. Vapor pressure

The vapor pressures from COSMO*therm* in the range 10^{-18} to 10^{-8} Pa, which are outside the validated range, are all significantly below the experimental values from the ECHA database. However, COSMO*therm* also calculated values below those from the ECHA database for 22% of the substances that are in the validated range of 10^{-3} to 10^{-8} Pa and for which the other data from COSMO*therm* agree well with the experimental data from the PHYSPROP database. The vapor pressures below the validated range ($<10^{-8}$ Pa) from COSMO*therm* are therefore not necessarily incorrect, but show that there is an urgent need for accurate measurements in this low-volatility range. If those measurements are not feasible, then it would at least be good to indicate that the measured values are upper bounds and not exact values.

4.2.4. Normal boiling point

The calculated normal boiling points from COSMO*therm* are for 31% of the substances considered here outside the range that could be validated with the data from the PHYSPROP database. It is therefore very difficult to judge how reliable the calculated normal boiling points above 350 °C from COSMO*therm* are that deviate from the measured ones. The values from COSMO*therm* outside the validated range should therefore be regarded as quite uncertain.

4.3. Accuracy of the experimental data in the ECHA database

Our analysis shows that there are some endpoints, and ranges within the endpoints, where the COSMO*therm* values deviate considerably from the experimental values in the ECHA database. A summary of the findings is given in Table 4. These include, e.g., $\log_{10} K_{OW}$ values above 8, solubilities in water below 10^{-3} mg/l, vapor pressures below 10^{-6} Pa and normal boiling points above 500 °C. These are also the ranges for which various problems have been reported with measurements.[31,34–36] Specific examples included e.g., decabromodiphenyl ethane for which a measured $\log_{10} K_{OW}$ of 3.55 has been reported,[37] but which is actually much higher (11.1–13.6) (Refs. 38 and 39). Also, for trisiloxane, 1,1,1,3,5,5,5-heptamethyl-3-tetradecyl-, a measured solubility in water of 19.1 mg/l has been reported,[40] but the solubility in water is expected to be much lower based on the structure of the compound. WSKOW v.1.4.1 in EPI Suite calculates for trisiloxane, 1,1,1,3,5,5,5-heptamethyl-3-tetradecyl-, a solubility in water of 6.5×10^{-9} mg/l. In the case of $\log_{10} K_{OW}$, S_W, and p_V, we presume that the large differences were mainly caused by inaccuracies and uncertainties in the measurements. We recommend therefore that newly developed measurement techniques should be applied for these ranges to generate more reliable data. Examples for those techniques for $\log_{10} K_{OW}$ are the slow-stirring dual-flask/solid-phase microextraction (SPME) method[36] or correlating the octanol–water partition coefficient to measured *n*-butanol–water partition coefficients.[35] In terms of solubility, the slow-stir water solubility method[41] has been developed for substances with solubilities below 0.1 mg/l and although the method is recommended in the ECHA Guidance[31] for low-solubility test substances, it is unclear how often it has been really used for the substances registered under REACH; this is mainly because there is no differentiation between fast and slow-stirring methods in the IUCLID data.

Methods for vapor pressures as low as 10^{-10} Pa are available,[33] however none of the recommended methods performed very well in the in the range $<10^{-8}$ Pa compared to the COSMO*therm* data. The static method (which is not recommended for this low-vapor-pressure range) performed particularly poorly. We recommend therefore that at least the results of the static method for low-volatile substances should be repeated with other, more suitable methods.

4.4. Accuracy of the non-experimental data in the ECHA database

The average deviations between the non-experimental data and the COSMO*therm* values are for all endpoints greater than the average deviations between the experimental data and the COSMO*therm*

TABLE 4. Summary of the general reliability of the ECHA data (compared to the COSMO*therm* values) and recommendations for methods for specific endpoints (and ranges). MAE: mean absolute error

Endpoint	Range of values	General reliability of ECHA data (compared to COSMO*therm* values)	Recommendations for methods (based on the results in this study)
$\log_{10} K_{OW}$	<4	Good (MAE = 0.47)	None in particular, all methods performed well
$\log_{10} K_{OW}$	4–8	Mixed (MAE = 1.11)	OECD TG 117 (HPLC method) or OECD TG 123 (slow stirring method)
$\log_{10} K_{OW}$	>8	Not good (MAE = 3.33)	None of the standard methods, none of them performed well
S_W (mg/l)	$<10^{-3}$	Not good [MAE (of $\log_{10} S_W$) = 3.06]	Difficult to judge as it was not possible to distinguish between the flask method with fast and slow stirring
S_W (mg/l)	$\geq 10^{-3}$	Ok [MAE (of $\log_{10} S_W$) = 0.68]	None in particular, flask and column elution method both work well
p_V (Pa)	$<10^{-6}$	Not good [MAE (of $\log_{10} p_V$) = 5.69]	None of the standard methods, none of them performed well
p_V (Pa)	10^{-6}–10^{-3}	Not good [MAE (of $\log_{10} p_V$) = 1.52]	Best results from effusion method: vapor pressure balance, dynamic method, gas saturation method, and effusion method: by loss of weight or by trapping vaporisate; do not use static method
p_V (Pa)	10^{-3}–1	Mixed [MAE (of $\log_{10} p_V$) = 1.04]	Do not use the static method, all other performed very similar
p_V (Pa)	>1	Good [MAE (of $\log_{10} p_V$) = 0.43]	None in particular, all performed well
T_b (°C)	<200	Good (MAE = 8.76)	None in particular, all performed well
T_b (°C)	200–500	Mixed (MAE = 35.24)	No method was especially good or bad
T_b (°C)	≥500	Not good (MAE = 381.6)	None of the standard methods, none of them performed well (or maybe COSMO*therm* cannot calculate the values well)

values. This shows that more guidance is needed that details which estimation methods or (Q)SARs should or can be used for which endpoint. It would probably also be useful to check/calculate if a substance is within the applicability domain of a model and rate those data that are outside the applicability domain as less certain.[42] Read-across from substances with deviating carbon-chain lengths should be avoided, because the values for most of the endpoints increase or decrease with changing carbon-chain length.

5. Conclusions

The ECHA database is one of the largest and most important databases with physicochemical properties. However, the quality of the data is very variable, and we also identified certain ranges that show values (compared to the COSMO*therm* values, which we assessed as reliable) that are systematically too low or too high. This can be problematic as, for example, a high $\log_{10} K_{OW}$ value is an indicator for a bioaccumulative substance. An underestimation of $\log_{10} K_{OW}$, therefore, likely also leads to an incorrect hazard assessment. With the publication of the COSMO*therm* data for more than 4400 substances, this work contributes therefore also to more accurate and trustworthy non-experimental physicochemical property data than currently available that can be used in the future for the hazard and risk assessment of these chemicals.

6. Supplementary Material

The supplementary material 1 is a pdf and contains additional data and graphics related to the methods and results in this article. Supplementary material 2 is an Excel workbook and contains the physicochemical property data calculated with COSMO*therm*. Supplementary material 3 is also an Excel workbook and contains for each endpoint those datapoints from the ECHA database that were identified as outliers or values with high deviation.

7. Acknowledgments

J.G. acknowledges funding from the Swiss Federal Office for the Environment. We thank Oleksandr Yushchenko, Rachel London, Katharina Sodnikar, Maya Amacha, Narain Ashta, and Sarah Partanen (all ETH Zürich or previously ETH Zürich) for their help with the COSMO*conf* calculations and Elvira Rudin for interesting discussions.

8. Author Declarations

8.1. Conflict of interest

The authors have no conflicts to disclose.

8.2. Author contributions

Data curation, formal analysis, methodology, visualization, and writing of the original draft was done by J.G. Conceptualization, funding acquisition, methodology, supervision, and reviewing and editing of the draft was done by M.S.

9. Data Availability

The data that support the findings of this study are available within the article and its supplementary material.

10. References

[1] V. Maeder, B. I. Escher, M. Scheringer, and K. Hungerbühler, "Toxic ratio as an indicator of the intrinsic toxicity in the assessment of persistent, bioaccumulative, and toxic chemicals," Environ. Sci. Technol. **38**, 3659–3666 (2004).

[2] OECD, "Test no. 305: Bioaccumulation in fish: Aqueous and dietary exposure" (OECD, Paris, 2012)

[3] C. Wittekindt and K.-U. Goss, "Screening the partition behavior of a large number of chemicals with a quantum-chemical software," Chemosphere **76**, 460–464 (2009).

[4] N. Ulrich, K.-U. Goss, and A. Ebert, "Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation," Commun. Chem. **4**, 90 (2021).

[5] EC, REACH implementation, https://ec.europa.eu/environment/chemicals/reach/implementation_en.htm, 2022.

[6] ECHA, Progress in evaluation, https://echa.europa.eu/overall-progress-in-evaluation, 2022.

[7] SRC, PHYSPROP database, https://www.epa.gov/tsca-screening-tools/download-epi-suitetm-estimation-program-interface-v411, 2017.

[8] ECHA, IUCLID6, https://iuclid6.echa.europa.eu/reach-study-results, 2021.

[9] NIH, PubChem, https://pubchem.ncbi.nlm.nih.gov/, 2022.

[10] USEPA, CompTox chemicals dashboard, https://comptox.epa.gov/dashboard/, 2022.

[11] CAS, SciFinder-n, https://scifinder-n.cas.org/, 2022.

[12] ECHA, REACH–registered substances dactsheets, https://echa.europa.eu/information-on-chemicals/registered-substances, 2022.

[13] ECHA, Guidance for identification and naming of substances under REACH and CLP, https://echa.europa.eu/view-article/-/journal_content/title/guidance-for-identification-and-naming-of-substances-under-reach-and-clp, 2017.

[14] J. Glüge, K. McNeill, and M. Scheringer, "Getting the SMILES right: Identifying inconsistent chemical identities in the ECHA database, PubChem and the CompTox Chemicals Dashboard," Environ. Sci. Adv. **2**, 612–621 (2023).

[15] ECHA, "Read-across assessment framework (RAAF): Considerations on multi-constituent substances and UVCBs," https://echa.europa.eu/documents/10162/13630/raaf_uvcb_report_en.pdf/3f79684d-07a5-e439-16c3-d2c8da96a316 (2017).

[16] USEPA, Estimation programs interface SuiteTM for Microsoft® Windows, v 4.11 (updated), 2017.

[17] Chemaxon, pKa plugin, https://docs.chemaxon.com/display/docs/pka-plugin.md, 2022.

[18] Biovia, COSMOtherm, https://www.3ds.com/products-services/biovia/products/molecular-modeling-simulation/solvation-chemistry/biovia-cosmotherm/, 2020.

[19] A. Klamt, "Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena," J. Phys. Chem. **99**, 2224–2235 (1995).

[20] F. Eckert and A. Klamt, "Fast solvent screening via quantum chemistry: COSMO-RS approach," AIChE J. **48**, 369–385 (2002).

[21] C. Loschen, J. Reinisch, and A. Klamt, "COSMO-RS based predictions for the SAMPL6 logP challenge," J. Comput.-Aided Mol. Des. **34**, 385–392 (2020).

[22] A. Stenzel, K.-U. Goss, and S. Endo, "Prediction of partition coefficients for complex environmental contaminants: Validation of COSMOtherm, ABSOLV, and SPARC," Environ. Toxicol. Chem. **33**, 1537–1543 (2014).

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

[24] S. Endo, J. Hammer, and S. Matsuzawa, "Experimental determination of air/water partition coefficients for 21 per- and polyfluoroalkyl substances reveals variable performance of property prediction models," Environ. Sci. Technol. **57**, 8406–8413 (2023).

[25] J. M. Parnis, D. Mackay, and T. Harner, "Temperature dependence of Henry's law constants and K_{OA} for simple and heteroatom-substituted PAHs by COSMO-RS," Atmos. Environ. **110**, 27–35 (2015).

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[27] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G. L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," Nat. Methods **17**, 261–272 (2020).

[28] ECHA, Information-requirements, https://echa.europa.eu/regulations/reach/registration/information-requirements, 2021.

[29] Chemsafe-Consulting, Unclaimed notification of new substances, https://www.chemsafe-consulting.com/2021/07/06/reach-unclaimed-nons-notification-of-new-substances/, 2021.

[30] ECHA, How to update your previously notified substance (NONS), https://echa.europa.eu/support/registration/how-to-update-your-previously-notified-substance, 2023.

[31] ECHA, "Guidance on information requirements and chemical safety assessment—Chapter R.7a: Endpoint specific guidance," https://echa.europa.eu/documents/10162/13632/information_requirements_r7a_en.pdf/e4a2a18f-a2bd-4a04-ac6d-0ea425b2567f (2017).

[32] ECHA, 2,2-dimethylpropane-1,3-diamine, https://echa.europa.eu/da/registration-dossier/-/registered-dossier/22090/4/9/?documentUUID=5f2290c9-4e82-487e-bae8-296102e39589, 2023.

[33] EC, Commission regulation (EC) no. 761/2009–method A.4 vapour pressure, Official Journal of the European Union, L 220/1, 2009.

[34] J. Pontolillo and R. P. Eganhouse, *The Search for Reliable Aqueous Solubility (S_w) and Octanol-Water Partition Coefficient (K_{OW}) Data for Hydrophobic Organic Compounds: DDT and DDE as a Case Study*, Water Resources Investigations Report 01-4201, U.S. Geological Survey, https://pubs.usgs.gov/wri/wri014201/ (2001).

[35] K. B. Hanson, D. J. Hoff, T. J. Lahren, D. R. Mount, A. J. Squillace, and L. P. Burkhard, "Estimating *n*-octanol-water partition coefficients for neutral highly hydrophobic chemicals using measured *n*-butanol-water partition coefficients," Chemosphere **218**, 616–623 (2019).

[36]M. T. O. Jonker, "Determining octanol–water partition coefficients for extremely hydrophobic chemicals by combining 'slow stirring' and solid-phase microextraction," Environ. Toxicol. Chem. **35**, 1371–1377 (2016).
[37]ECHA, Partitioning coefficient of 1,1′-(ethane-1,2-diyl)bis[pentabromobenzene], https://echa.europa.eu/registration-dossier/-/registered-dossier/15001/4/8/?documentUUID=3715905c-51db-415f-827a-cd332e172e11, 2023.
[38]Canada, screening assessment certain organic flame retardants substance grouping benzene, 1,1′-(1,2-ethanediyl)bis [2,3,4,5,6-pentabromodecabromodiphenyl ethane (DBDPE)], https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/screening-assessment-certain-organic-flame-retardants-substance-grouping-benzene-ethanediyl-bis-pentabromo-decabromodiphenyl-ethane-dbdpe.html, 2019.
[39]AMAP, Decabromodiphenylethane (DBDPE), https://chemicals.amap.no/chemicals/decabromodiphenylethane-dbdpe/, 2023.
[40]ECHA, Water solubility of trisiloxane, 1,1,1,3,5,5,5-heptamethyl-3-tetradecyl-, https://echa.europa.eu/registration-dossier/-/registered-dossier/11782/4/9/?documentUUID=8966f879-6292-409a-87b7-4845af16cb2a, 2023.
[41]D. J. Letinski, A. D. Redman, H. Birch, and P. Mayer, "Inter-laboratory comparison of water solubility methods applied to difficult-to-test substances," BMC Chem. **15**, 52 (2021).
[42]C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.-C. De Hemptinne, P. Ungerer, B. Rousseau, and C. Adamo, "A general guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes," Chem. Rev. **115**, 13093–13164 (2015).