SUPPLEMENTARY MATERIAL FOR:

A Virtual Screening Approach to Identifying the Greenest Compound for a Task: Application to Switchable-Hydrophilicity Solvents

J. R. Vanderveen, L. Patiny, C. B. Chalifoux, M. J. Jessop, and P. G. Jessop*

Contents

1. Synthetic procedures and product characterizations

4-(diethylamino)-1-hexanol (1)

A round bottom flask was charded with 5 mL 6-bromo-1-hexanol (38 mmol), 16 mL diethylamine (150 mmol), and 20 mL acetonitrile. The mixture was refluxed overnight, cooled to room temperature, then added to 20 mL 20% aqueous NaOH. The mixture was extracted with diethyl ether (3 x 20 mL) and the organic phase was dried with magnesium sulfate and concentrated under reduced pressure to afford the crude product. Distillation yielded the pure product as a colourless oil (4.2 g, 60%); BP 77 °C (0.2 torr); ^{1}H NMR (400 MHz, $CDCl_3$) δ 1.01 (t, J = 7.1 Hz, 6H), 1.25-1.40 (m, 4H), 1.46 (p, J= 7.5 Hz, 2H), 1.56 (p, J = 6.8 Hz, 2H), 2.40 (t, J = 7.6 Hz, 2H), 2.51 (q, J = 7.1 Hz, 4H), 3.61 (t, J = 6.7 Hz, 2H); ^{13}C NMR (400 MHz, $CDCl_3$) δ 11.4, 25.7, 26.8, 27.4, 32.7, 46.8, 52.8, 62.5; ν_{max} (ATR) 3320, 2968, 2930, 2857, 2804, 1467, 1378, 1293, 1201, 1059, 769; *m/z* (EI) 173 (8), 158 (7), 86 (100), 72 (24), 58 (26), 55 (10); HRMS $C_{10}H_{23}NO$ for M^+ calculated 173.1780, found 173.1788.

N,N-di-(4-methoxybutyl)isopropylamine (2)

A round bottom flask was charged with isopropylamine (1 mL, 12 mmol) isopropylamine, 4-bromo-1-methoxybutane (5 g, 30 mmol), and di-isopropylethylamine (5.3 mL, 30 mmol). The mixture was stirred for 6 h at room temperature, then for 18 h in an oil bath heated to 40 °C, and finally for another 24 h heated to 100 °C. The mixture was cooled to room temperature and 20 mL of 20% aqueous NaOH was added. The product was extracted with diethyl ether (3 x 20 mL). The resulting organic phase was dried with magnesium sulfate and concentrated under reduced pressure to afford the crude product. Distillation yielded the pure product as a colourless liquid (1.9 g, 40%); BP 73 °C (0.2 torr); ^{1}H NMR (500 MHz, $CDCl_3$) δ 0.88 (d, J = 6.4 Hz, 6H), 1.37 (p, J = 6.6 Hz, 4H), 1.49 (p, J = 6.6 Hz, 4H), 2.29 (t, J = 6.6 Hz, 4H), 2.83 (septet, J = 6.4 Hz, 1H), 3.25 (s, 6H), 3.30 (t, J = 6.6 Hz, 4H); ^{13}C NMR (500 MHz, $CDCl_3$) δ 18.0, 25.5, 27.5, 49.6, 49.8, 58.4, 72.8; ν_{max} (ATR) 2961, 2930, 2864, 2823, 2806, 1461, 1385, 1360, 1318, 1222, 1192, 1164, 1117, 1087, 949, 898; *m/z* (EI) 231 (3), 216 (18), 158 (100), 98 (9), 87 (60), 84 (23), 55 (23); HRMS $C_{13}H_{29}NO_2$ for M^+ calculated 231.2198, found 231.2191.

4-(dipropylamino)-1-butanol (3)

To 4-chloro-1-butanol (5 mL, 50 mmol) was added dipropylamine (17 mL, 125 mmol). The mixture was stirred at 100 °C for 96 h, then cooled to room temperature. Aqueous acid (40 mL, 1 M HCl) was added to the mixture and washed with 40 mL ether. The pH of the aqueous phase was adjusted to 12 with solid sodium hydroxide and the aqueous phase was extracted with ether (3 x 40 mL). The organic phase was dried with magnesium sulfate and concentrated under reduced pressure to afford the crude product. Distillation yielded the pure product as a colourless liquid (2.54 g, 30%); BP 52 °C (0.1 torr); ^{1}H NMR (500 MHz, $CDCl_3$) δ 0.78 (t, J = 7.5 Hz, 6H), 1.39 (sextet, J = 7.6 Hz, 4H), 1.53 (m, 4H), 2.30 (m, 6H), 3.44 (t, J = 4.9 Hz, 2H), 6.39 (broad, 1H); ^{13}C (500 MHz, $CDCl_3$) δ 11.8, 19.2, 26.0, 32.5, 54.7, 55.7, 62.4; ν_{max} (ATR) 3379, 2956, 2933, 2872, 2801, 1460, 1378, 1297, 1190, 1066, 1023; *m/z* (EI) 173 (14), 144 (87), 114 (100), 100 (8), 86 (30), 72 (60), 55 (21); HRMS $C_{10}H_{23}NO$ for M^+ calculated 173.1780, found 173.1784.

4-(di-(4-methoxybutyl)amino)-1-butanol (4)

A mixture of 4-amino-1-butanol (1 mL, 10 mmol), 4-bromo-1-methoxybutane (4.2 mL, 30 mmol), and di-isopropylethylamine (3.8 mL, 20 mmol) was heated to 100 °C for 48 h. The mixture was cooled and added to 10 mL concentrated aqueous HCl. The mixture was washed with diethyl ether (2 x 10 mL). Solid NaOH was added to the aqueous phase until its pH was >12. The aqueous phase was extracted with diethyl ether (3 x 10 mL) and the resulting organic phase was dried with magnesium sulfate and concentrated under reduced pressure to afford the crude product. Distillation yielded the pure product as a colourless liquid (0.7 g, 25%); BP 89 °C (0.1 torr); ^{1}H NMR (400 MHz, $CDCl_3$) δ 1.54 (m, 8H), 1.65 (m, 4H), 2.46 (m, 6H), 3.33 (s, 6H), 3.38 (t, J = 6.0 Hz, 4H), 3.55 (t, J = 4.7 Hz, 2H), 6.20 (broad, 1H); ^{13}C NMR (400 MHz, $CDCl_3$) δ 22.7, 26.2, 27.7, 32.6, 53.4, 54.6, 58.5, 62.7, 72.6; ν_{max} (ATR) 3421, 2932, 2863, 2824, 2807, 1461, 1385, 1223, 1192, 1116, 1085, 850, 808, 737; *m/z* (EI) 261 (3), 246 (5), 202 (32), 188 (80), 116 (8), 87 (100), 84 (22), 55 (26); HRMS $C_{14}H_{31}NO_3$ for M^+ calculated 261.2304, found 261.2309.

2. Procedure for Using the Virtual Screening Method

Prior to implementing the software, a list of structural fragments that are compatible with switchable-hydrophilicity solvents (SHS) were identified. SHS are known to function poorly in the presence of acidic functional groups, which can interfere with the reversible protonation of the amine group by CO_2. Additionally, hydrolytically unstable functional groups such as esters cause SHS to degrade over time, limiting their ability to be reused. Amides were considered acceptable because they are less prone to hydrolysis than esters.

The functional groups and the trimethylamine core were then added to *Script* and used to generate a list of structures. *Script* uses a search and replace method in conjunction with OpenChemLib to generate structures and corresponding SMILES (simplified molecular-input line-entry system) for each structure. Each structure was also given a unique structure identification code. The software also automatically calculates log K_{ow} values for each structure as they are generated. The resulting dataset was visualized using the NPellet visualizer. The initial set of molecules was screened for only molecules with log K_{ow} values between 0.9 and 2.6 by using a slider tool which allows users to select a specific range of values for a given property and view only molecules that fit within that range of values for that property. The list of structures which fit the log K_{ow} requirement was exported as a list of SMILES strings along with their corresponding structure identification code. The exported file was in the .txt format.

The text file including the list of SMILES was inputted into the Toxicity Estimation Software Tool (TEST) and ACD/Percepta to predict the properties of each structure. In TEST, the text file was imported using the software's "batch import from list of SMILES strings" option to generate a batch list of molecules. The properties of interest were predicted using this batch list and TEST's consensus prediction method. The results were outputted as .txt files that included the structures' SMILES, their structure identification codes, and their predicted values for the corresponding predictions. One .txt file was generated per prediction.

In ACD/Percepta, the text file including the list of SMILES was inputted using the "import" command in the software to generate a list of structures. The pK_a of each structure was predicted using Percepta's GALAS (base) method. The data was outputted from Percepta using the "export" command in the software. The data was outputted as a .txt file that included the structures' SMILES, their structure identification codes, and their predicted pK_a values.

Each .txt file was collected and inputted into *Script* using a parsing script. Log K_{ow} values were already available within *Script* from the initial structure generation and log K_{ow} prediction. The NPellet visualizer was modified to calculate acceptability values for each predicted property as well as the overall acceptability functions A_{perf} and A_{EHS}. The visualizer was set to display molecular structures, slider tools for each property and the overall acceptability functions, and a window to view the properties of a selected structure. The sliders for A_{perf} and A_{EHS} were adjusted to only view compounds with high values for each, then adjusted further to view compounds with lower and lower scores until structures that could be synthesized in one or two steps were identified. In this study, synthetically facile structures were identified when A_{perf} was set to be a minimum of 2.5/3 and A_{EHS} was set to be 5/7.

3. Description of property prediction methods

Eight properties were predicted using the United States Environmental Protection Agency's Toxicity Estimation Software Tool (TEST) version 4.1. These properties include LD_{50} (oral, rat), LC_{50} (fathead minnow, 96 h), LC_{50} (daphnia magna, 48 h), bioaccumulation factor (BAF), boiling point, flash point, melting point, and vapour pressure. The values used in this study were determined using the "consensus" method available in TEST. This method reports the average of five other methods: hierarchical clustering, FDA (a hierarchical model based on structurally similar compounds), single-model linear regression, group contribution, and nearest neighbour. Each model was built using the training sets and test sets already present in the software. The number of total compounds in the training and test sets varied depending on the prediction, with a minimum of 353 for LC_{50} (daphnia magna, 48 h) and a maximum of 9385 for melting point.

ACD/Percepta was used to calculate pK_{aH} values. The GALAS method was used in this study. The model was constructed using a training set already present in the software consisting of values for 17,593 compounds. The predicted values are reported with an error of ± 0.4.

OpenChemLib was used to calculate log K_{ow} values. The model was constructed using a training set already present in the software consisting of values for over 5000 compounds. The prediction is performed using an incremental system which adds contributions of individual atoms based on their atom types.

Table S1 lists the mean average error (MAE) reported for each prediction as well as the MAE found using only the amine-containing compounds used to create the acceptability functions. The MAE was calculated using the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

Where n is the number of amines used in the comparison and f_i and y_i are the predicted and experimental values, respectively, for compound *i*. The MAE was not used in the creation of the acceptability functions. Instead, the acceptability functions were made based the distribution of differences between experimental and predicted values, as described in the article.

Table S1 Mean average errors (MAEs) for different properties as reported for test and training sets and as found for amines in this study

Property	Reported MAE	MAE for amines
-log(LC_{50}) (fathead minnow, 96 h)	0.545	0.32
-log(LC_{50}) (daphnia magna, 48 h)	0.727	n/a[a]
-log(LD_{50}) (oral, rat)	0.431	0.30
log(BAF)	0.513	n/a[a]
boiling point	11.460	9.0
flash point	16.908	12.0
-log(vapour pressure)	0.466	0.45
melting point	30.207	25.7
pK_{aH}	0.4	0.2
log K_{ow}	n/a[b]	0.39

[a]Not enough experimental data available to determine a reliable MAE

[b]MAE for the OpenChemLib log K_{ow} prediction is not reported in the software's documentation.

4. Acceptability functions

Acceptability functions were created using comparisons of predicted and experimental values. First, the deviations of each predicted value from the experimentally determined value were calculated. For boiling point, flash point, melting point, pKa, and log K_{ow}, the difference between predicted and experimental values was used to evaluate absolute deviation. For vapour pressure, Fathead minnow LC_{50}, and LD_{50}, the quotient of the predicted and experimental values was used to evaluate relative deviation. Using flash point values as an example, the deviations of the predicted values from the experimental values ranged from –96 °C to +38 °C. From these deviations, every 10^{th} percentile from the 10^{th} to the 90^{th} was calculated. In the case of flash point, the 10^{th} percentile was found to be -14 °C and the 90^{th} percentile was found to be 17 °C. The 10^{th} percentile indicates that only 10% of predicted values will be >14 °C lower than the experimentally determined value. Likewise, only 10% of predicted values will be >17 °C higher than the experimentally determined value.

An acceptability function can be made by applying these percentile values to the target value for the property in question. For flash point, the target value is 80 °C. Since only 10% of predicted values will be more than 14 °C lower than the experimental values, a compound having a predicted flash point of 94 °C has a 90% chance of having an actual flash point that meets or exceeds the 80 °C target. Likewise, only 10% of predicted values will be more than 17 °C higher than the experimental values. This means that a compound having a predicted flash point of 63 °C only has a 10% chance of having an actual flash point that meets or exceeds the target value. Predicted flash points greater than 94 °C should be considered completely acceptable because they indicate high likelihood that our design criterion is met. Thus, in the acceptability function, 94 °C should return a value close to 1. Similarly, 63 °C should return a value close to 0 in the acceptability function. The 20^{th}, 30^{th}, 40^{th},…80^{th} percentiles should return acceptability values of 1/8, 2/8, 3/8,…7/8. By this metric, a predicted flashpoint at the 50^{th} percentile of deviation from the target value would be given an acceptability of 0.5, indicating that it has a 50% chance of being acceptable. When plotted on a graph of acceptability vs. predicted value, these percentile deviations from the target values with their assigned acceptabilities form an approximately sigmoidal trend from 0 to 1. The data were fit to a sigmoidal function:

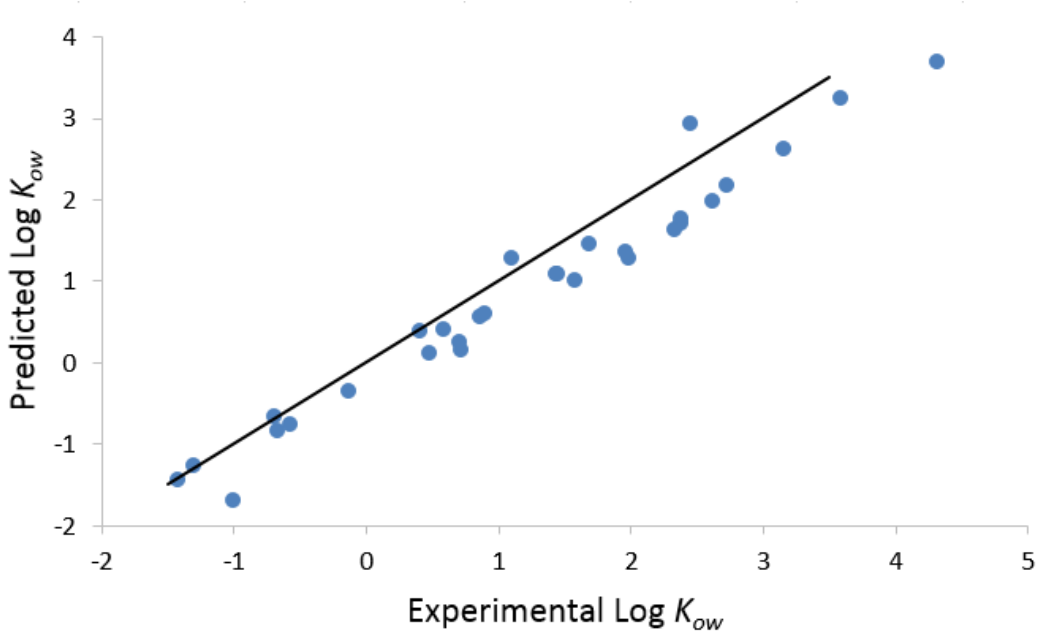
$$A_x = \frac{1}{1 + \left(\frac{C}{P_x + E}\right)^D}$$

Where A_x is the acceptability value for property x and P_x is the predicted value for property x. The values were fit by altering the constants C, D and E to minimize the sum of the root mean square deviation of the 10^{th}-90^{th} percentile acceptabilities from the acceptabilities calculated from the equation.

Log K_{ow}:

$$A_{logKow} = \frac{1}{1 + \left(\frac{25.85}{logKow + 24.29}\right)^{196.96}}$$

a)



b)

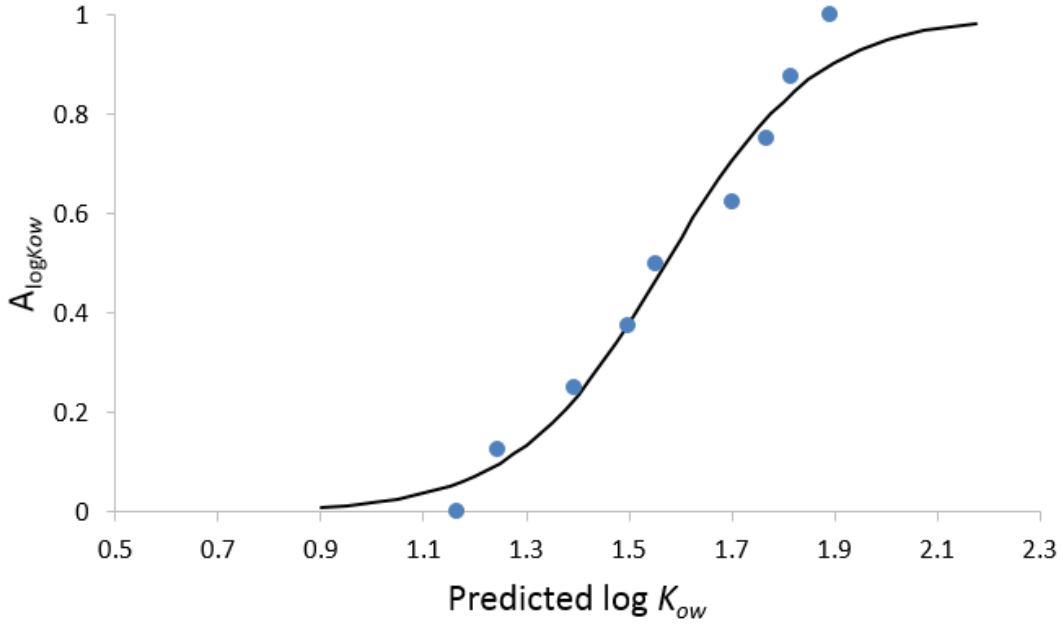
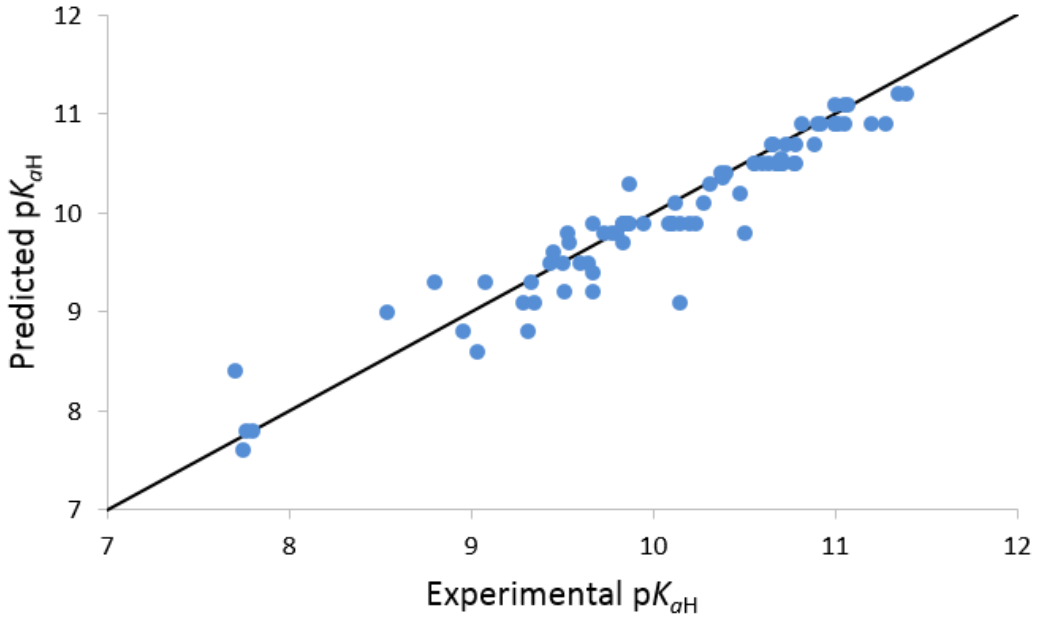


Fig. S1 a) Comparison of predicted and experimental log K_{ow} values (black line represents a perfect prediction) and b) the sigmoidal acceptability function (black line) fit to the acceptability values assigned to every 10th percentile from the 10th to the 90th (blue dots).

pK_{aH}:

$$A_{pKa} = \frac{1}{1 + \left(\frac{9.06}{pKa - 0.0339}\right)^{107.28}}$$

a)



b)

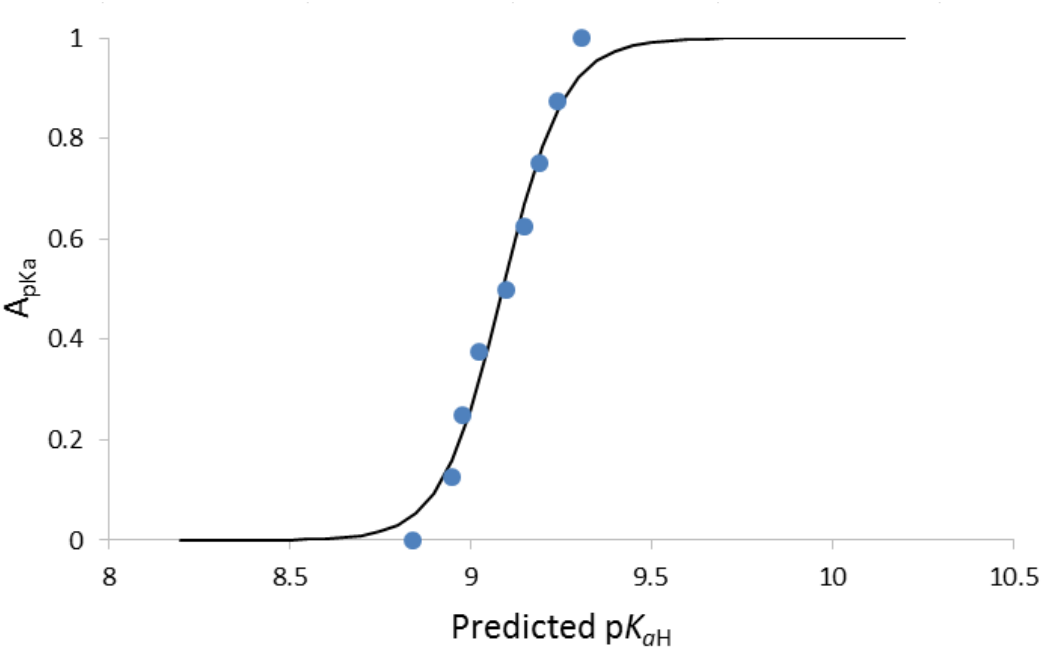


Fig. S2 a) Comparison of predicted and experimental pK_{aH} values (black line represents a perfect prediction) and b) the sigmoidal acceptability function (black line) fit to the acceptability values assigned to every 10th percentile from the 10th to the 90th (blue dots).

Melting point (°C):

$$A_{MP} = \frac{1}{1 + \left(\frac{59.53}{75 - MP}\right)^5}$$

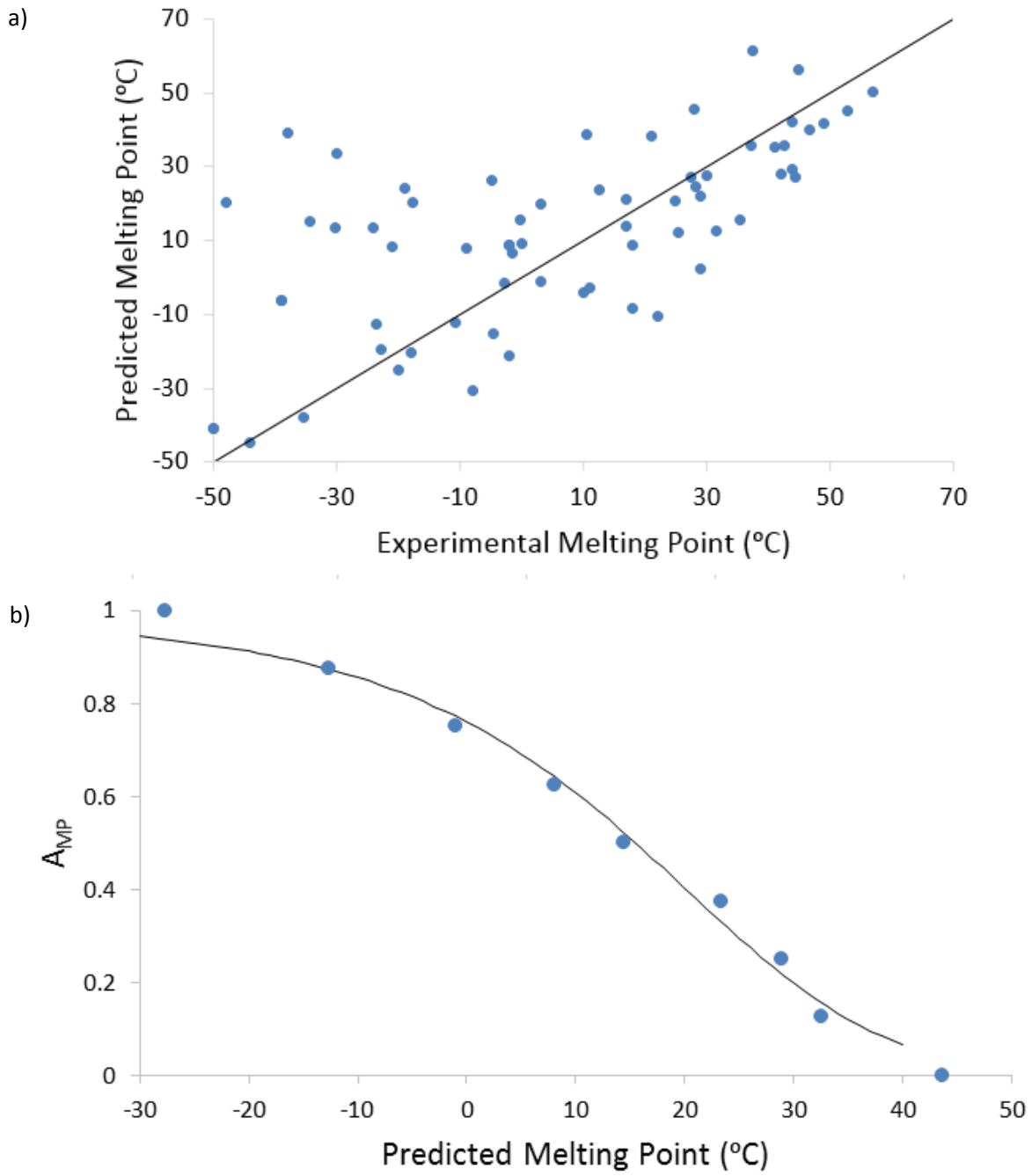
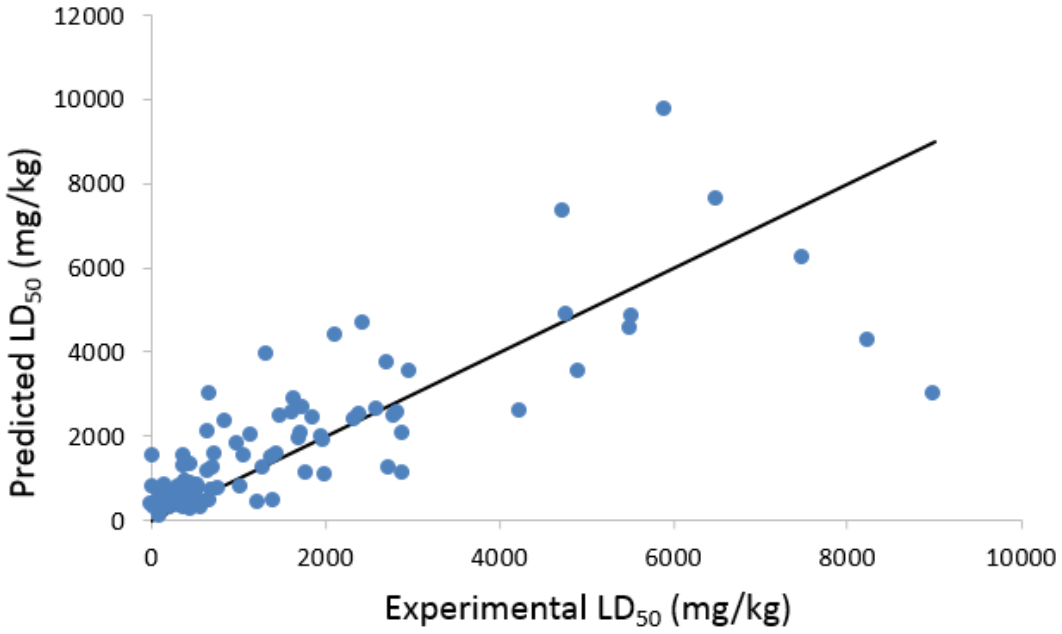


Fig. S3 a) Comparison of predicted and experimental melting points (black line represents a perfect prediction) and b) the sigmoidal acceptability function (black line) fit to the acceptability values assigned to every 10th percentile from the 10th to the 90th (blue dots).

LD_{50} (oral, rat, mg/kg):

$$A_{LD50} = \frac{1}{1 + \left(\frac{1255.33}{LD_{50_{rat}} - 121.86}\right)^{2.89}}$$

a)



b)

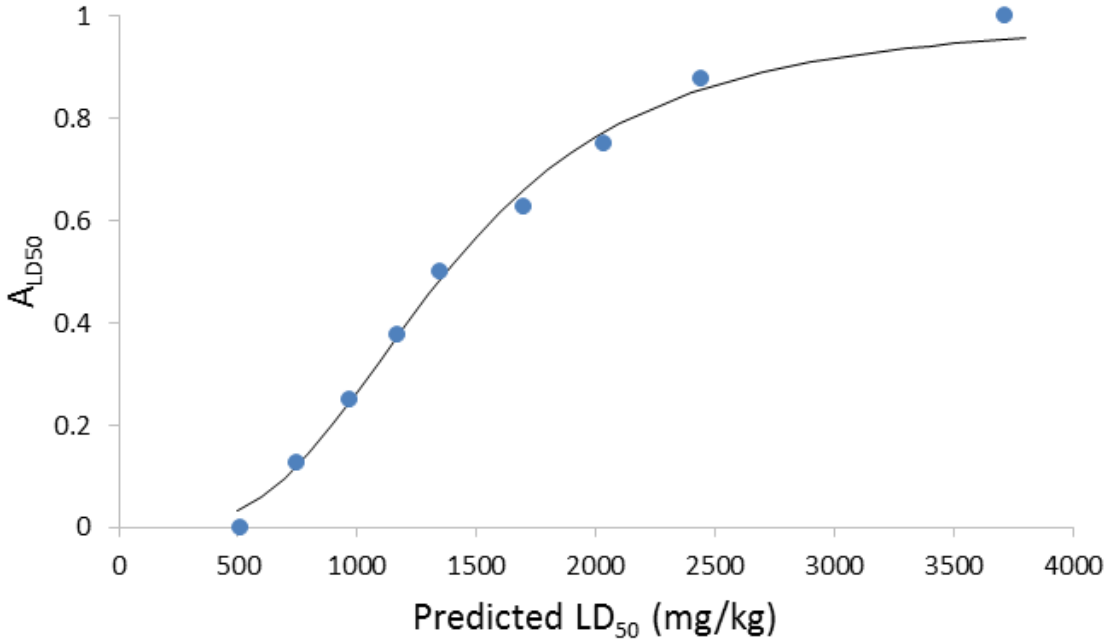
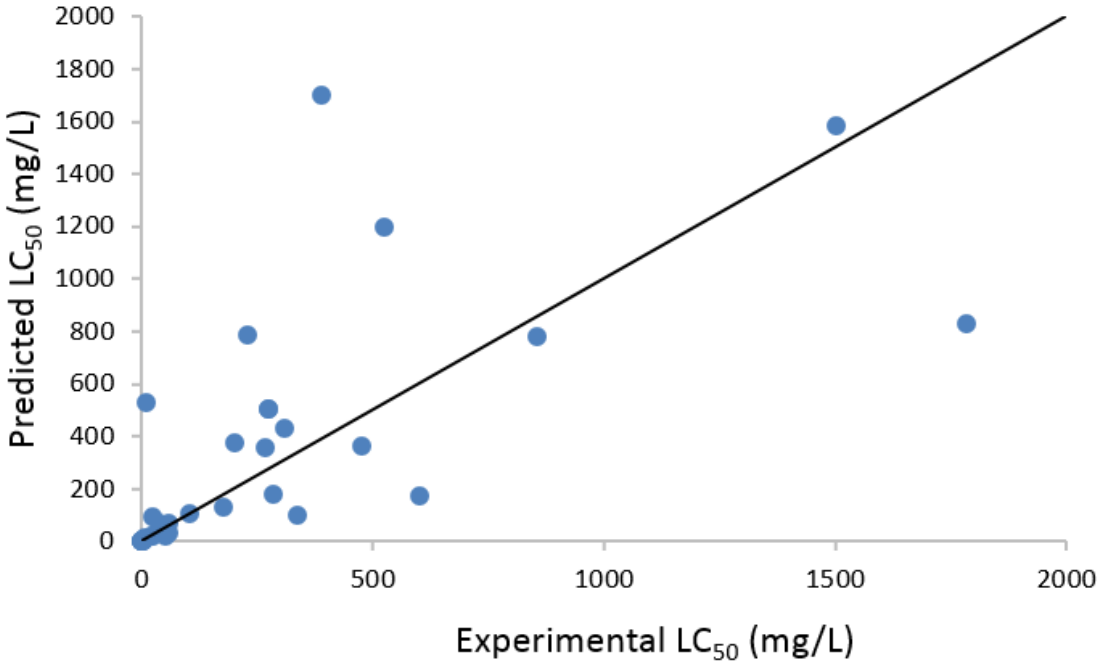


Fig. S4 a) Comparison of predicted and experimental LD_{50} (oral, rat) values (black line represents a perfect prediction) and b) the sigmoidal acceptability function (black line) fit to the acceptability values assigned to every 10^{th} percentile from the 10^{th} to the 90^{th} (blue dots).

LC_{50} (fathead minnow, 96 h, mg/L):

$$A_{FMLC50} = \frac{1}{1 + \left(\frac{73.92}{FMLC_{50} + 3.24}\right)^{2.6}}$$

a)



b)

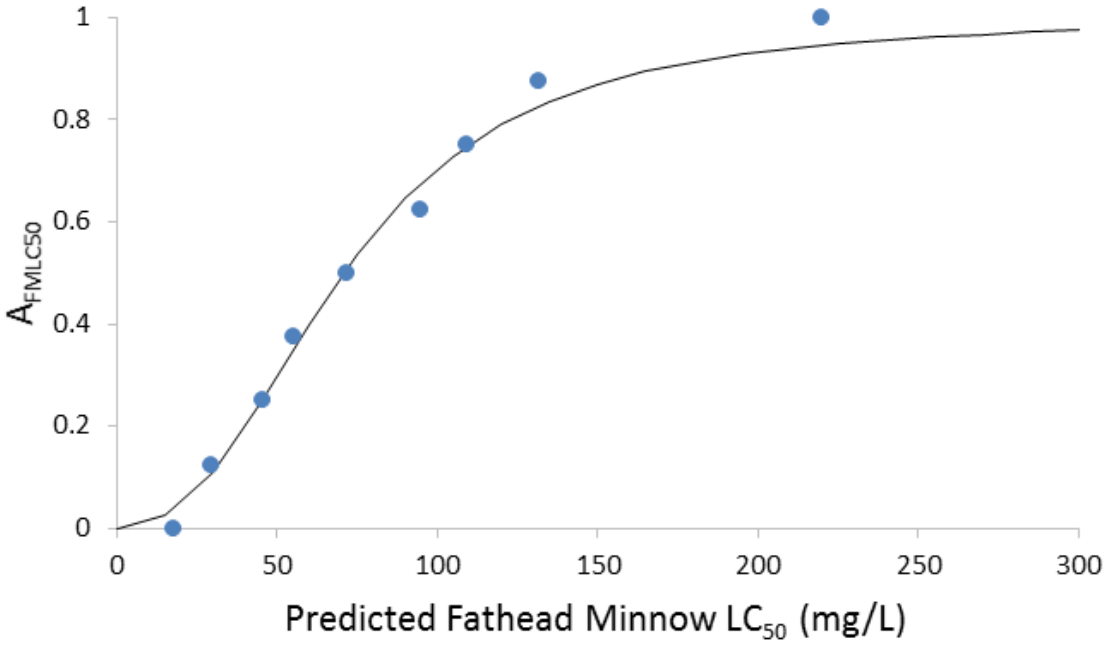
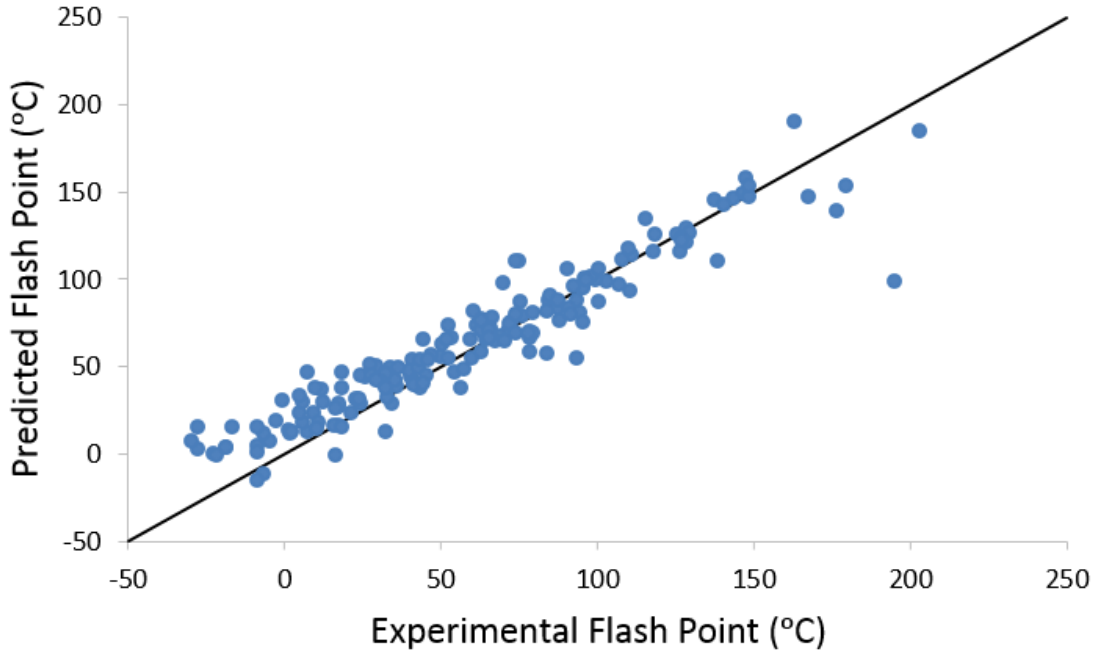


Fig. S5 a) Comparison of predicted and experimental LC_{50} (fathead minnow, 96 h) values (black line represents a perfect prediction) and b) the sigmoidal acceptability function (black line) fit to the acceptability values assigned to every 10^{th} percentile from the 10^{th} to the 90^{th} (blue dots).

Flash point (°C):

$$A_{FP} = \frac{1}{1 + \left(\frac{74.01}{FP - 2.87}\right)^{16.73}}$$

a)



b)

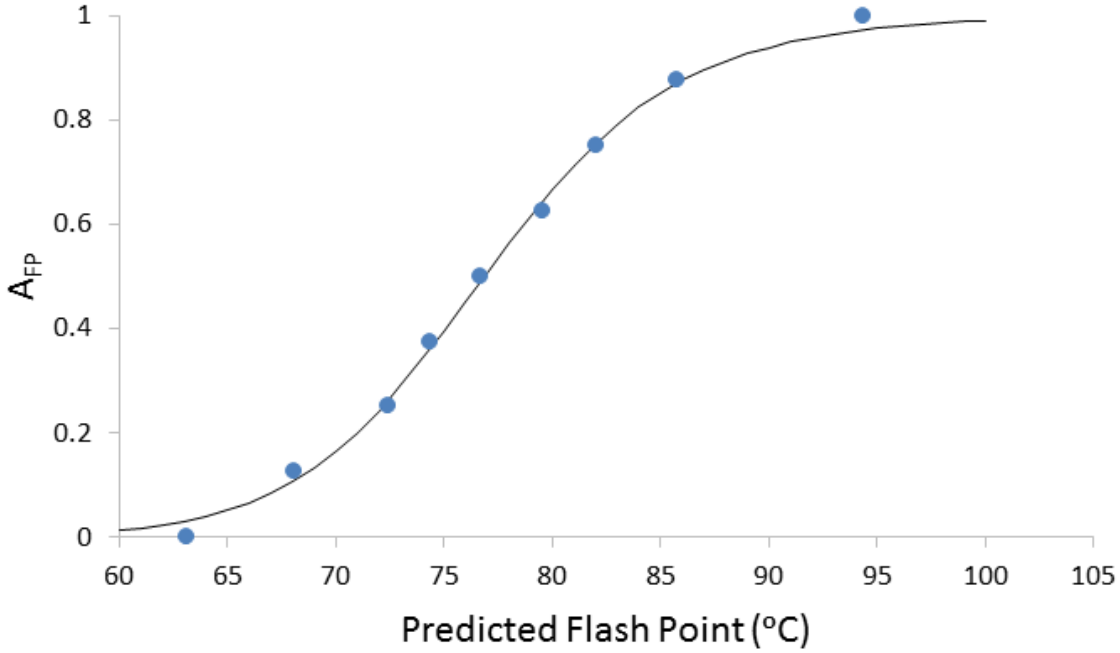


Fig. S6 a) Comparison of predicted and experimental flash points (black line represents a perfect prediction) and b) the sigmoidal acceptability function (black line) fit to the acceptability values assigned to every 10th percentile from the 10th to the 90th (blue dots).

Boiling point (°C):

$$A_{BP} = \frac{1}{1 + \left(\frac{138.8}{BP - 38.22}\right)^{35.62}}$$

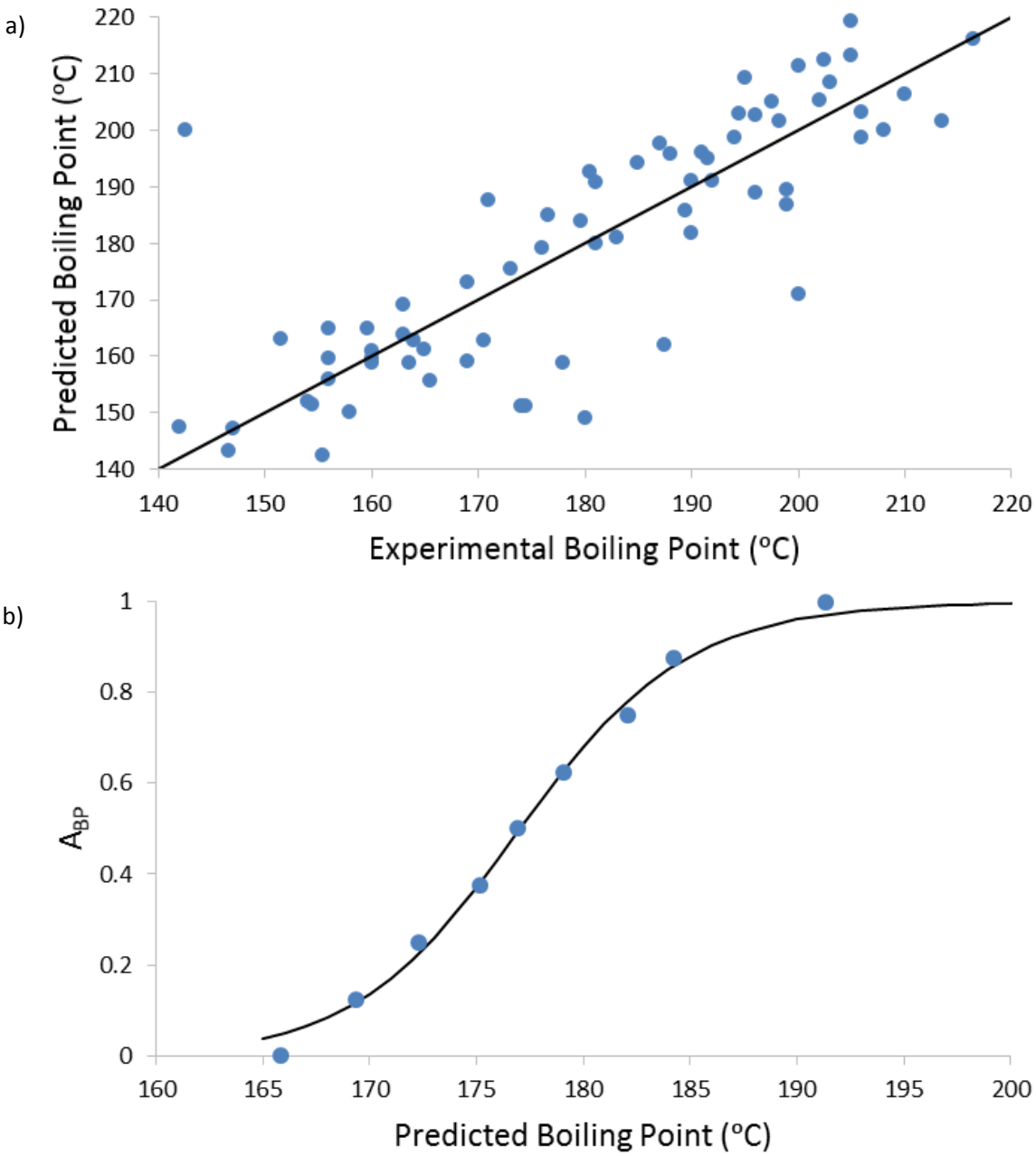


Fig. S7 a) Comparison of predicted and experimental boiling points (black line represents a perfect prediction) and b) the sigmoidal acceptability function (black line) fit to the acceptability values assigned to every 10th percentile from the 10th to the 90th (blue dots).

Bioaccumulation factor:

Not enough experimentally determined values were found to characterize the accuracy of the predictions, therefore any compound with a bioaccumulation factor less than 500 was assigned an acceptability (A_{BAF}) of 1, while any compound with a bioaccumulation factor 500 or greater was assigned an acceptability of 0.

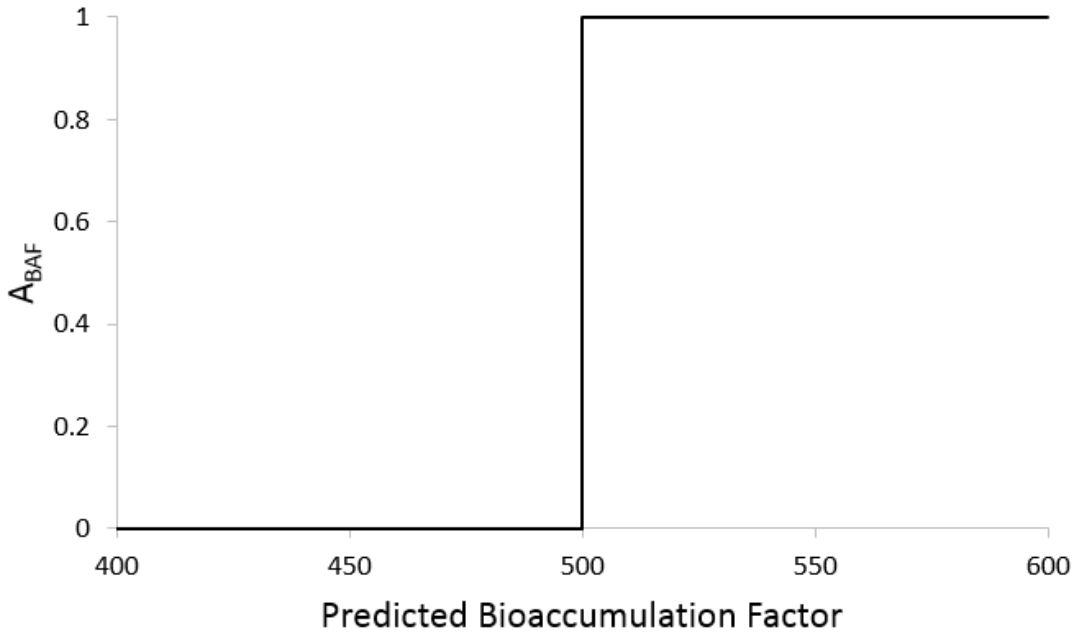


Fig. S8 The acceptability function (black line) applied to bioaccumulation factor predictions.

LC_{50} (daphnia magna, 48h, mg/L):

Not enough experimentally determined values were found to characterize the accuracy of the predictions, therefore compounds were assigned acceptabilities based on a linear acceptability function, where LC_{50} values <10 mg/L were assigned an acceptability value of 0, LC_{50} values >200 mg/L were assigned an acceptability value of 1, and intermediate LC_{50} values were assigned an acceptability according to the following function:

$$A_{DMLC50} = 0.0053 \times DM_{LC50} - 0.06$$

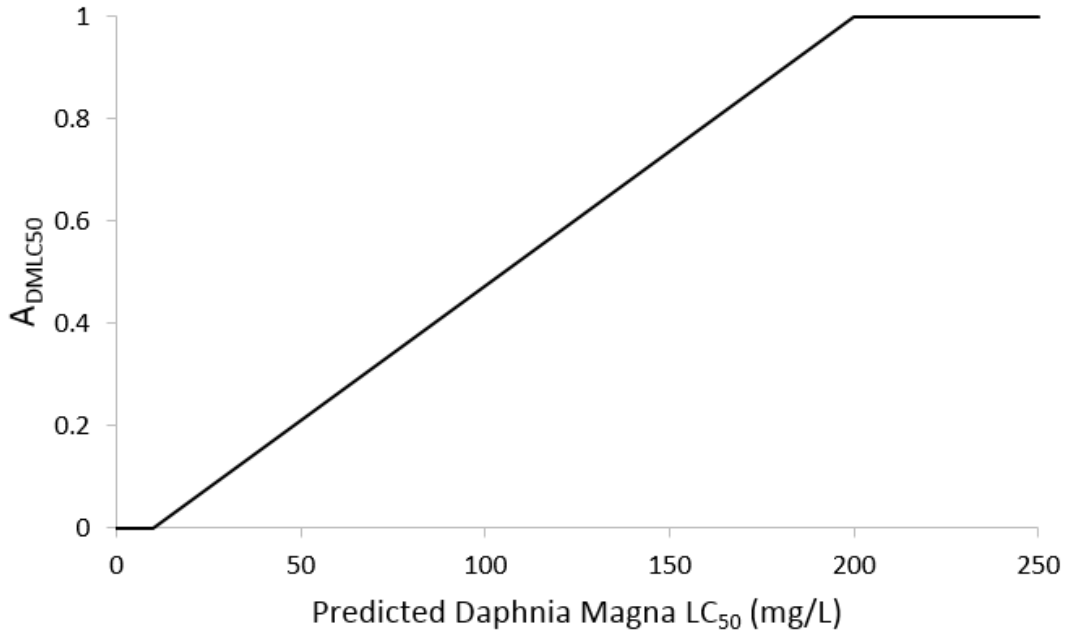


Fig. S9 The acceptability function (black line) applied to LC_{50} (daphnia magna, 48 h) predictions.

Vapour pressure (torr):

The vapour pressure acceptabilities were assigned based on a linear scale based on the likely strength of the odour of the compound. Vapour pressures less than 0.03 torr, or -1.5 on a logarithmic scale, were assumed to correspond to an odourless compound and were assigned an acceptability of 1. Vapour pressures more than 0.5 torr, or -0.3 on a logarithmic scale, were assigned an acceptability of 0. Compounds with intermediate vapour pressures were assigned an acceptability according to the following function (based on the log of their vapour pressure):

$$A_{VP} = -1 - 0.6667 \times logVP$$

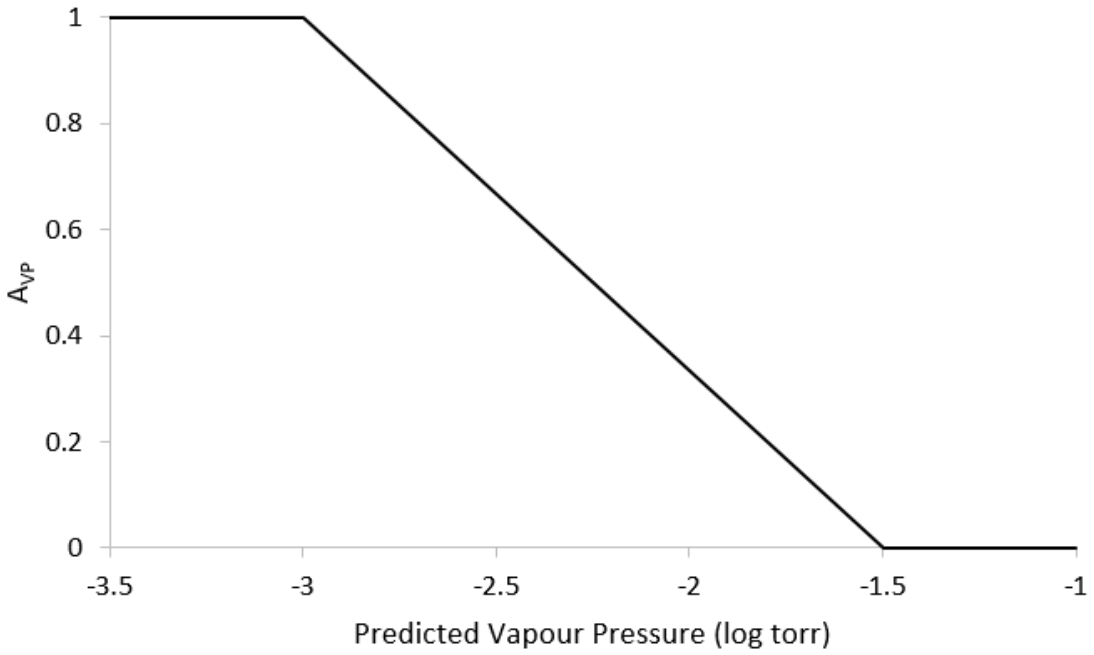


Fig. S10 The acceptability function (black line) applied to vapour pressure predictions.

Performance acceptability:

$$A_{performance} = A_{logKow} + A_{pKa} + A_{MP}$$

EHS acceptability:

$$A_{EHS} = A_{LD50} + A_{FMLC50} + A_{DMLC50} + A_{BAF} + A_{FP} + A_{BP} + A_{VP}$$

5. Comparison of experimental and predicted amine solubilities in water:

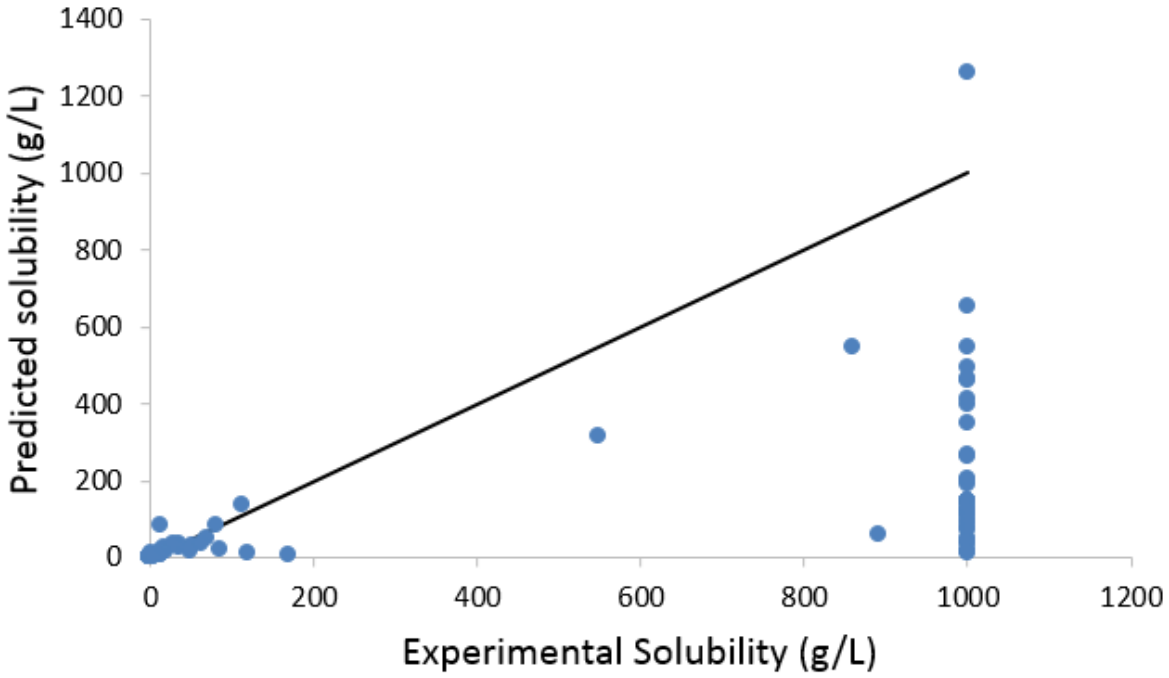


Fig. S11 Comparison of experimental amine solubilities in water with values predicted by TEST. The black line represents a perfect prediction.

Solubility in water is a useful property to use in switchable hydrophilicity solvent (SHS) screening because a compound cannot be an SHS if it is completely miscible with water, often reported as a solubility of 1000 g/L. The solubility predictions performed by TEST proved to be a poor predictor of miscibility with water as shown in Fig. S8. Compounds reported to be completely miscible with water have predicted values as low as 8 g/L. Other solubility predictors, including EPIWEB and ACD/Labs predictors, had similarly poor performance. For this reason, solubility predictions were not used in this study.