AquaBoxIL – computational tool for determining the environmental distribution profile for ionic liquids

Maciej Barycki¹, Anita Sosnowska¹, Tomasz Puzyn¹

¹ – University of Gdansk, Faculty of Chemistry, Department of Environmental Chemistry and Radiochemistry, Laboratory of Environmental Chemometrics

Electronic supplementary information 1

Table of contents

How to use the MM model	2
Sensitivity analysis with the FAST method	6
Recalculation of the solubility QSPR model	9
Model for biodegradation	.12
Model for the K_D partition coefficient	.19
Structural features influencing particular properties of ionic liquids	.23
References	.25

How to use the MM model

Data preparation step

Molecular models for both ions constituting the ionic liquid should be generated, and their geometries should be optimized using the appropriate method. We recommend using MOPAC 2016. The following commands should be used.

For cations:

PM7 PRECISE CHARGE=1

For anions:

PM7 PRECISE CHARGE=-1

A set of 20 molecular indices (Table ESI1) should be calculated using the DRAGON 7.0 software.

1	2	3	4	5	6	7
MW_C	MW_A	H8e_C	R7p+_C	HTi_A	ONOV_C	X5Av_C
8	9	10	11	12	13	14
L3M_A	Sv_C	J_B(p)_A	SpMin6_Bh(i)_C	ON0_C	MATS1m_C	GNar_C
15	16	17	18	19	20	
G2u_C	G3u_C	Mor19u_A	R4u+_A	C-040_A	F02[C-N]_A	

Table ESI1. Molecular descriptors required for the modeling process.

_C – descriptor for cation, _A – descriptor for anion

A vector (20×1 in size) containing all abovementioned indices in the given order should then be imported to the MATLAB program.

Installation of the tool

The tool can be downloaded from the journal's website (published as Electronic Supplementary Information 2). It is sufficient to add the path containing files extracted from the downloaded archive to MATLAB's directories library (by clicking "set path", localizing the extracted files and clicking "save"). Afterwards, the tool can be called by typing the "MMIL" command.

Example calculation



Figure ESI 1. MM model interface.

The following paragraph describes a single run of calculations using the MM tool. Figure ESI 1 presents the user interface of the tool and will be referred to in the following description.

Step-by-step instruction, how to compute a single calculation:

Provide the input data, including volumes of each compartment as well as sediment parameters and amount of ionic liquid to be deposited in the system (Panel A). Alternatively, the default values provided in the tool can be used. Then, import a vector of the IL descriptors (see data preparation step in this section). This vector should already be present in the MATLAB workspace. Chose the correct vector (Panel B – popup menu) and click LOAD (Panel B). All input data are already provided. Alternatively, the user's

own values of the IL properties can be used (for example, in cases where the experimental value of an IL property is known). For this, choose the property of interest (click the checkbox next to its name – Panel C) and type the desired value in the corresponding box right to the property's name. Click EXECUTE (Button D) to perform the calculations (which should activate as soon as the correct vector of descriptors is loaded). All predicted properties will be displayed in Panel E. The checked box next to the property (Panel E) indicates that the prediction is inside the QSPR model's applicability domain. The Insubria plot on the right (Plot F) also shows the relation of each prediction to the models' ADs. Note that the user-defined values are not taken into account in this step. Plot G shows the IL concentration in each environmental compartment. Plot H shows the amount of IL that is degraded, present as micelles or undissolved. These two plots constitute the main output of the tool. Button I will save the results, and clicking it generates three additional variables in the MATAB workspace: ILMMOUT_Concentrations – 1 × 3 vector of the IL concentration in water (1,1), sediment (1,2) and the organic phase (1,3);

ILMMOUT_Distribution – 6×1 vector of the amount of IL present in water (1,1), sediment (2,1) and organic matter (3,1) as well as the amount of IL that was degraded (4,1), present as micelles (5,1) or undissolved (6,1);

ILMMOUT_ADstatus – 5 × 2 matrix of normalized values of leverages^{*} in column 1, and normalized values of the QSPR model predictions^{**} of S (row 1), K_{OW} (row 2), CMC (row 3), K_D (row 4) and %BIO (row 5) in column 2.

* - normalized to 0=0, h*=1;

^{**} - normalized by the minimal prediction by the QSPR model for the training set = 0 and the maximal prediction of the QSPR model for the training set = 1

Figure ESI 2 shows the exemplary MM calculation for n-octylimidazolium hexafluorophosphate using default environmental properties, where the K_{OW} value was arbitrarily set to 7.5 by the user [for the purpose of demonstrating the utility].



Figure ESI 2. Example of the IL distribution in the environmental simulation using the presented tool.

Sensitivity analysis with the FAST method

In the FAST method, the model's response is a function $Y = f(X_1, ..., X_N)$.¹ The space $\Omega = [0,1]^N$, describing the variance of the input parameters, is searched through search curve K. The distribution of the measurement points is planned according to the following equation:²

$$X_i(s) = \frac{1}{2} + \frac{1}{\pi} \arcsin(\sin\omega_i s)$$

where

 X_i – matrix of input values for the corresponding measurement points ω_i – frequencies s – vector f measurement points

The number of measurement points was set according to the following equation:²

$$N_s = 2 * N_{harm} * max\omega + 1$$

where N_s – number of measurement points N_{harm} – interference coefficient max ω – maximal value of the selected frequencies

The variance of model's output can be presented as:¹

$$V = D^{2}(Y) = \lim_{T \to +\infty} \frac{1}{2T} \int_{-T}^{T} f^{2}(s) ds - \left[\lim_{T \to +\infty} \frac{1}{2T} \int_{-T}^{T} f(s) ds \right]^{2}$$

Assuming $\omega_i \in N+$, function Y = f(s) is periodical in the range of 2π , the variance can be approximated as:

$$V = \frac{1}{2\pi} \int_{-\pi}^{\pi} f^{2}(s) ds - \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) ds \right]^{2}$$

Using Fourier's transformation on function Y = f(s) we get:¹

$$f(s) = \sum_{j=-\infty}^{+\infty} (A_j cos(js) + B_j sin(js))$$

Fourier's coefficients are:²

$$A_{j} = \begin{cases} \frac{1}{N_{s}} \left\{ f(sN_{o}) + \sum_{q=1}^{N_{q}} \left[f(sN_{o+q}) + f(sN_{o-q}) \right] * \cos\left(j\frac{\pi}{N_{s}}q\right) \right\} \text{ for je} \\ 0 \text{ for jo} \end{cases}$$

and

$$B_{j} = \left\{ \begin{array}{l} 0 \text{ for } je \\ \frac{1}{N_{s}} \left\{ f(sN_{o}) + \sum_{q=1}^{N_{q}} \left[f(sN_{o+q}) + f(sN_{o-q}) \right] * sin\left(j\frac{\pi}{N_{s}}q\right) \right\} \text{ for } jo \end{cases} \right\}$$

where

Nq - (Ns-1)/2 No - (Ns+1)/2 je - even values of j jo - odd values of j

According to Saltelli's reasoning, assuming $\Lambda_j = A_j^2 + B_j^2$, and based on Parsewal's theorem, we can write (1):

$$\sum_{j \in \mathbb{Z}} \Lambda_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} f^2(s) ds$$

Comparing equations (1) and (2), we obtain an estimator of variance Y:¹

$$V = 2\sum_{j=1}^{+\infty} \Lambda_j$$

and an estimator of the conditional variance in the expected Y under condition $\rm Xi:^1$

$$V_i = 2\sum_{p=1}^{+\infty} \Lambda_{p\omega i}$$

The influence of each input variable on the output can be described by the sensitivity index according to the equation:

$$S_i = \frac{V_i}{V}$$

Recalculation of the solubility QSPR model

In the QSPR model, the solubility in water was adopted from Freire et al.³ Because the two molecular descriptors used in the model are highly correlated (Sv and Sp; r = 0.99), we reduced the number of descriptors in the equation and recalculated it. The model was validated internally and externally. The new model equation is as follows:

$$\ln(S) = 1.97 (\pm 0.77) - 0.34 (\pm 0.02)$$
 Sv_C - 1.32 (± 0.16) J_B(p)_A

where

Sv_C – sum of the atomic van der Waals volumes (scaled to a carbon atom) – descriptor for cation J_B(p) – Balaban-like index from the Burden matrix weighted by polarizability – descriptor for anion

The model parameters are presented in Table ESI2.

Measure	Calibration	Cross- validation (loo)	External validation	Criteria
F	106.08			dependent ^a
\mathbb{R}^2	0.942			> 0.7
\mathbf{Q}^2		0.916	0.944 ^b	> 0.6
CCC	0.970	0.957	0.973	>0.85
R ² _{SCR}		0.131		dependent ^a
RMSE	0.252	0.304	0.270	dependent ^a
MAE	0.204	0.251	0.170	dependent ^a
average r _m ²			0.923	> 0.5
Δr_m^2			0.014	< 0.2
$(r^2 - r_0^2) / r^2$			0.028	< 0.1
$(r^2 - r'_0^2) / r^2$			0.033	< 0.1
k			1.001	$0.85 \le k \le 1.15$
k'			0.997	$0.85 \leq k' \leq 1.15$
$ \mathbf{r}_0^2 - \mathbf{r'}_0^2 $			0.005	< 0.3

Table ESI2. Calibration and validation measures calculated for the QSPR model predicting the solubility of the ionic liquid.

[---] – measure does not refer to the procedure; a – F criterion is dependent of the number of compounds in the set, should be higher than tabulated value of F at chosen significance, R^2_{SCR} – is dependent of value of R^2 of calibration, should be significantly lower than R^2 ; RMSE and MAE are dependent of the range of modelled values, should be as low as possible; b – value of Q^2 parameter in the validation procedure refers to $Q^2_{\rm F2}$

Figure ESI 3 presents the goodness-of-fit of the training and the test set. A William's plot (Figure ESI 4) and Insubria plot (Figure ESI 5) are presented for the compounds from the training and test sets in the space of model's applicability domain.



Figure ESI 3. Plot presenting the observed vs. predicted solubilities, which confirms the quality of the model calibration.



Figure ESI 4. William's plot presenting the applicability domain of the QSPR solubility model. $\pm 3\sigma$ refers to ± 3 standard deviations. h* is the critical leverage value.



Figure ESI 5. Insubria plot presenting the applicability domain of the QSPR solubility model. Ypmax and Ypmin are maximal and minimal values, respectively, of the model prediction for the training set. h* is the critical leverage value.

Model for biodegradation

Data describing the biodegradation of the ionic liquids were collected from the literature.^{4–9} The percent of biodegradation was determined experimentally with the use of the OECD 310 protocol.¹⁰ The histogram of the collected data was visibly bimodal (Figure ESI 6).



Figure ESI 6. Histogram of the biodegradability data.

This suggested the existence of two separate subsets. According to the OECD 310 test, substances with a percent of biodegradation under 60% are considered not readily biodegradable. We split the data into two sets and developed one QSPR model for each group (readily and not readily biodegradable ILs). Additionally, we created a classification model (classification tree). This model can distinguish between readily and not readily biodegradable ILs and therefore help decide which QSPR model should be used for a particular IL.

Classification model



Figure ESI 7. Classification tree.

Figure ESI 7 presents the classification model. The classification is based entirely on the cation descriptors. The model validation parameters are presented in Table ESI3.

Confusion matrices for the training set (T) and test set (V)					
Т	Not Readily B.	Readily B.	V	Not Readily B.	Readily B.
Not Readily B.	36	0	Not Readily B.	17	0
Not Readily B.	0	16	Not Readily B.	1	7
ACC	10	0%	ACC		96%
TPR	10	0%	TPR		94%
TNR	10	0%	TNR		100%

Table ESI 3. Model validation parameters.

Based on the classification, one of the following models can be used to quantitatively predict the biodegradability of the IL:

QSPR model for <u>not readily biodegradable</u> ionic liquids

 $\sqrt{{}^{\% BIO}}$ = 23.90 (±4.17) – 8.84 (±1.71) **GNar_C** – 27.10 (±8.79) **G2u_C** + 6.33 (±1.01) **G3u_C** + 1.59 (±0.22) **Mor19u_A**

where

GNar_C – Narumi geometric topological index – descriptor for cation

G2u_C – 2nd compartment symmetry directional WHIM index / unweighted – descriptor for cation

G3u_C – 3rd compartment symmetry directional WHIM index / unweighted – descriptor for cation Mor19u_A – 3D-MoRSE descriptor describing signal 19 / unweighted – descriptor for anion

The model parameters are presented in Table ESI4.

Measure	Calibration	Cross- validation (loo)	External validation	Criteria
F	23.20			dependent ^a
\mathbb{R}^2	0.726			> 0.7
\mathbf{Q}^2		0.663	0.848	> 0.6
CCC	0.841	0.807	0.906	>0.85
R ² _{SCR}		0.099		dependent ^a
RMSE	0.954	1.058	0.659	dependent ^a
MAE	0.788	0.884	0.538	dependent ^a
average r _m ²			0.663	> 0.5
Δr_m^2			0.152	< 0.2
$(r^2 - r_0^2) / r^2$			0.123	< 0.1
$(r^2 - r'_0^2) / r^2$			0.027	< 0.1
k			1.014	$0.85 \le k \le 1.15$
k'			0.958	$0.85 \leq k' \leq 1.15$
$ r_0^2 - r'_0^2 $			0.084	< 0.3

Table ESI4. Calibration and validation measures calculated for the QSPR model predicting the percent of biodegradation of persistent ionic liquids.

[---] – measure does not refer to the procedure; a – F criterion is dependent of the number of compounds in the set, should be higher than tabulated value of F at chosen significance, R^2_{SCR} – is dependent of value of R^2 of calibration, should be significantly lower than R^2 ; RMSE and MAE are dependent of the range of modelled values, should be as low as possible; b – value of Q^2 parameter in the validation procedure refers to Q^2_{F2}

Figure ESI 8 presents the goodness-of-fit of the training and test sets. A William's plot (Figure ESI 9) and Insubria plot (Figure ESI 10) are presented for the compounds from the training and test sets in the space of model's applicability domain.



Figure ESI 8. Plot presenting the observed vs. predicted values of the biodegradability of persistent ionic liquids, which confirms the quality of the model's calibration.



Figure ESI 9. William's plot presenting the applicability domain of the QSPR model of the biodegradability of persistent ionic liquids . $\pm 3\sigma$ refers to ± 3 standard deviations. h* is the critical leverage value.



Figure ESI 10. Insubria plot presenting the applicability domain of the QSPR model of the biodegradability of persistent ionic liquids . Ypmax and Ypmin are the maximal and minimal values, respectively, of the model prediction for the training set. h* is the critical leverage value.

QSPR model for readily biodegradable ionic liquids

%BIO = 63.43 (±1.33) + 124.77 (±25.34) **R4u+_A** + 12.24 (±1.94) **C-040_A** – 2.56 (±0.69) **F02[C-N]_A**

where

R4u+_A – R maximal autocorrelation of lag 4 / unweighted – descriptor for anion C-040_A – atom-centered fragment descriptor indicating the presence of R-C(=X)-X / R-C#X / X=C=X structures – descriptor for anion F02[C-N]_A – frequency of C-N at topological distance 2 – descriptor for anion

The model parameters are presented in Table ESI5.

Table ESI5. Calibration and validation measures calculated for the QSPR model predicting the biodegradation of non-persistent ionic liquids.

Measure	Calibration	Cross- validation (loo)	External validation	Criteria
F	34.64			dependent ^a
R ²	0.881			> 0.7

\mathbf{Q}^2		0.836	0.699	> 0.6
CCC	0.937	0.913	0.883	>0.85
R ² _{SCR}		0.175		dependent ^a
RMSE	2.490	2.923	2.626	dependent ^a
MAE	2.061	2.549	2.116	dependent ^a
average r _m ²			0.714	> 0.5
Δr_m^2			0.144	< 0.2
$(r^2 - r_0^2) / r^2$			0.098	< 0.1
$(r^2 - r'_0^2) / r^2$			0.023	< 0.1
k			0.980	$0.85 \le k \le 1.15$
k'			1.020	$0.85 \leq k' \leq 1.15$
$ \mathbf{r}_0^2 - \mathbf{r}'_0^2 $			0.145	< 0.3

[---] – measure does not refer to the procedure; a – F criterion is dependent of the number of compounds in the set, should be higher than tabulated value of F at chosen significance, R^2_{SCR} – is dependent of value of R^2 of calibration, should be significantly lower than R^2 ; RMSE and MAE are dependent of the range of modelled values, should be as low as possible; b – value of Q^2 parameter in the validation procedure refers to Q^2_{F2}

Figure ESI 11 present the goodness-of-fit of the training and test sets. A William's plot (Figure ESI 12) and Insubria plot (Figure ESI 13) are presented for the compounds from the training and test sets in the space of model's applicability domain.



Figure ESI 11. Plot presenting the observed vs. predicted values of the biodegradability of non-persistent ionic liquids, which confirms the quality of the model calibration.



Figure ESI 12. William's plot presenting the applicability domain of the QSPR model of the biodegradability of non-persistent ionic liquids. $\pm 3\sigma$ refers to ± 3 standard deviations. h* is the critical leverage value.



Figure ESI 13. Insubria plot presenting the applicability domain of the QSPR model of the biodegradability of non-persistent ionic liquids. Ypmax and Ypmin are the maximal and minimal values, respectively, of the model prediction for the training set. h* is the critical leverage value.

Model for the K_D partition coefficient

The QSPR model for the K_D coefficient could not be developed using the standardized protocol due to the strong influence of external parameters (physicochemical characteristics of the sediment) on the value of the modeled parameter. These characteristics had to be included in the model equation; therefore, the QSPR model created for K_D describes the dependence of the partition coefficient on both the chemical structure of the ionic liquid and the characteristics of the sediment. This model is not a QSPR model in its strict definition.

Data were collected from Mrozik et al.¹¹ The in32_G index was created to describe different aspects of the tested set of adsorption matrices and can be written as:

$$in32_G = \left(\frac{\begin{pmatrix}g_1\\100\end{pmatrix} + \begin{pmatrix}g_2\\300\end{pmatrix}}{2}\right)^2$$

where g_1 – clay content (%) g_2 – cation exchange capacity (meq g⁻¹)

The developed QSPR equation is as follows:

where

in32_G – index describing the properties of the adsorption matrix

SpMin6_Bh(i)_C – smallest eigenvalue n = 6 of the Burden matrix weighted by the ionization potential – descriptor for cation

The model parameters are presented in Table ESI6.

Table ESI6. Calibration and validation measures calculated for the QSPR model for predicting the K_D of an ionic liquid.

Measure	Calibration	Cross- validation (loo)	External validation	Criteria
F	54.63			dependent ^a
\mathbb{R}^2	0.657			> 0.7
\mathbf{Q}^2		0.623	0.744	> 0.6
CCC	0.793	0.773	0.870	>0.85
R ² _{SCR}		0.035		dependent ^a
RMSE	0.220	0.231	0.164	dependent ^a
MAE	0.186	0.196	0.139	dependent ^a
average r _{m²}			0.695	> 0.5
Δr_m^2			0.091	< 0.2
$(r^2 - r_0^2) / r^2$			0.030	< 0.1
$(r^2 - r'_0^2) / r^2$			0.002	< 0.1
k			1.031	$0.85 \leq k \leq 1.15$
k'			0.887	$0.85 \leq k' \leq 1.15$
$ \mathbf{r}_0^2 - \mathbf{r'}_0^2 $			0.091	< 0.3

[---] – measure does not refer to the procedure; a – F criterion is dependent of the number of compounds in the set, should be higher than tabulated value of F at chosen significance, R^2_{SCR} – is dependent of value of R^2 of calibration, should be significantly lower than R^2 ; RMSE and MAE are dependent of the range of modelled values, should be as low as possible; b – value of Q^2 parameter in the validation procedure refers to Q^2_{F2}

Figure ESI 14 presents the goodness-of-fit of the training and test sets. A William's plot (Figure ESI 15) and Insubria plot (Figure ESI 16) are presented for the compounds from the training and test sets in the space of the model's applicability domain.



Figure ESI 14. Plot presenting the observed vs. predicted values of K_D, which confirms the quality of the model calibration.



Figure ESI 15. William's plot presenting the applicability domain of the QSPR model for

 $K_{\text{D}}.$ ±3 σ refers to ±3 standard deviations. h* is the critical leverage value.



Figure ESI 16. Insubria plot presenting the applicability domain of the QSPR model for K_D. Ypmax and Ypmin are the maximal and minimal values, respectively, of the model prediction for the training set. h* is the critical leverage value.

Structural features influencing particular properties of ionic liquids

Table ESI7. Features of cations and anions constituting ILs with high/low values of certain physicochemical properties. Table based on the interpretation of the QSPR models used in this work as sources of input data.

	Low value of property	High value of property	
	CATION: larger particles, larger	CATION: smaller particles composed	
Solubility	number of atoms	of fewer atoms	
	ANION: larger, linear shape	ANION: smaller, more spherical	
	CATION: smaller, shorter, branched,	CATION: long substituents, linear,	
	multiple chemical bonds near the	containing heteroatoms near the	
K _{ow}	terminus of the particle	terminus	
	ANION: smaller, with lower	ANION: larger, with higher	
	molecular mass	molecular mass	
	CATION: short alkyl chains, more	CATION: long chains straight best	
	spherical particles, if present alkyl		
СМС	chains are bent	hydrophobic alkyl substituents	
	ANION: small anion	ANION: large anion	

	CATION: homogeneous structure of	CATION: heterogeneous, different	
	the cation, following atoms are the	atoms following each other in the	
	same, not branched, having an axis	structure, highly branched, highly	
	of symmetry on the second	symmetric	
Biodegradation	directional axes (WHIM indices)		
		ANION: containing oxygen next to	
	ANION: fluorinated particle, similar	the alkyl chain, containing the	
	to	following groups: C=X-X, X=C=X, R-	
	bis(trifluoromethylsulfonyl)amide	C#X	
	CATION: less hydrophobic, smaller	CATION: hydrophobic cations, long	
V.	particles, fewer atoms	with long alkyl chains	
К _D			
	ANION:	ANION:	

References

- 1 A. Charzyńska, A. Nałęcz, M. Rybiński and A. Gambin, *J. Biotechnol. Comput. Biol. Bionanotechnol.*, 2012, **93**, 291–308.
- A. Saltelli, T. Stefano and K. Chan, *Technometrics*, 1999, **41**, 39–56.
- M. G. Freire, C. M. S. S. Neves, S. P. M. Ventura, M. J. Pratas, I. M. Marrucho, J.
 Oliveira, J. a P. Coutinho and A. M. Fernandes, *Fluid Phase Equilib.*, 2010, 294, 234–240.
- 4 L. Ford, J. R. Harjani, F. Atefi, M. T. Garcia, R. D. Singer and P. J. Scammells, *Green Chem.*, 2010, **12**, 1783.
- 5 R. G. Gore, L. Myles, M. Spulak, I. Beadham, T. M. Garcia, S. J. Connon and N. Gathergood, *Green Chem.*, 2013, **15**, 2747.
- C. Pretti, M. Renzi, S. E. Focardi, A. Giovani, G. Monni, B. Melai, S. Rajamani and C.
 Chiappe, *Ecotoxicol. Environ. Saf.*, 2011, **74**, 748–53.
- 7 F. Atefi, M. T. Garcia, R. D. Singer and P. J. Scammells, *Green Chem.*, 2009, **11**, 1595.
- D. Coleman, M. Špulák, M. T. Garcia and N. Gathergood, *Green Chem.*, 2012, 14, 1350.
- 9 X. D. Hou, Q. P. Liu, T. J. Smith, N. Li and M. H. Zong, *PLoS One*, 2013, **8**.
- 10 OECD, Test No. 310: Ready Biodegradability CO2 in sealed vessels (Headspace Test), 2006.
- W. Mrozik, A. Kotłowska, W. Kamysz and P. Stepnowski, *Chemosphere*, 2012, 88, 1202–1207.