SciForum
Mol2Net

Chemoinformatics Profiling of Ionic Liquids Cytotoxicity—From Machine Learning to Network-Like Similarity Graphs †

Maykel Cruz-Monteagudo [1,2,*], Eduardo Tejera [2], Cesar Paz-y-Miño [2], Yunierkis Perez-Castillo [3,4], Aminael Sánchez-Rodríguez [5], Fernanda Borges [1] and M. Natália D. S. Cordeiro [6]

[1] CIQUP/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal; E-Mail: fborges@fc.up.pt (F.B.)

[2] Instituto de Investigaciones Biomédicas (IIB), Universidad de Las Américas, 170513 Quito, Ecuador; E-Mails: eduardo.tejera@udla.edu.ec (E.T.); cesar.pazymino@udla.edu.ec (C.P.-Y.-M.)

[3] Sección Físico Química y Matemáticas, Departamento de Química, Universidad Técnica Particular de Loja, San Cayetano Alto S/N, EC1101608 Loja, Ecuador; E-Mail: yunierkis@gmail.com

[4] Molecular Simulation and Drug Design Group, Centro de Bioactivos Químicos (CBQ), Central University of Las Villas, Santa Clara, 54830, Cuba

[5] Departamento de Ciencias Naturales, Universidad Técnica Particular de Loja, Calle París S/N, EC1101608 Loja, Ecuador; E-Mail: asanchez2@utpl.edu.ec

[6] REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal; E-Mail: ncordeir@fc.up.pt

* Author to whom correspondence should be addressed; E-Mail: gmailkelcm@yahoo.es; Tel.: +351-220-402-000; Fax: +351-220-402-009.

Received: 20 September 2015 / Accepted: 20 September 2015 / Published: 2 December 2015

Abstract: Ionic liquids (ILs) possess a unique physicochemical profile providing a wide range of applications. However, their "greenness", specifically their claimed relative non toxicity has been frequently questioned, hindering their REACH registration processes and so, their final application. In this work we introduce a reliable, predictive, simple and chemically interpretable classification and regression tree (CART) classifier enabling the prioritization of ILs with a favourable cytotoxicity profile. By inspecting the structure of the CART several moieties that can be regarded as "cytotoxicophores" were identified and used to establish a set of SAR trends specifically aimed to prioritise low cytotoxicity ILs. We also demonstrated the suitability of the joint use of the CART classifier and a group fusion similarity search as a virtual screening strategy for the automatic prioritisation of safe ILs disperse in a data set of ILs of moderate to very high cytotoxicity. Additionally, we decided to complement the quantitative results already obtained by applying the network-like similarity graphs (NSG) approach to the mining of relevant structure-

cytotoxicity relationships (SCR) trends. Finally, the SCR information concurrently gathered by both, quantitative (CART classifier) and qualitative (NSG) approaches was used to design a focused combinatorial library enriched with potentially safe ILs.

Mol2Net YouTube channel*: http://bit.do/mol2net-tube*

† The full content of this communication can be found in Cruz-Monteagudo M, Ancede-Gallardo E, Jorge M, Dias Soeiro Cordeiro MN. Toxicol Sci. 2013; 136(2):548-65 and Cruz-Monteagudo M, Cordeiro MN. Toxicol Sci. 2014; 138(1):191-204.

1. Introduction

Ionic liquids (ILs) constitute one of the hottest areas in chemistry since they have become increasingly popular as reaction and extraction media [1]. Their almost limitless structural possibilities, as opposed to limited structural variations within molecular solvents, make ILs ‘‘designer solvents’’ [2]. They have also been widely promoted as “green solvents” [3] but such a “greenness” has been frequently questioned [1].

Despite the scarcity of reports of the prediction of the cytotoxicity of ILs by using a classification approach [4], we consider that a computational prediction system based on the use of classification methods is well justified and can offer a practical tool for the identification of new and safe ILs. So, in this work we intend to introduce a computational system allowing a fully automatic and chemically interpretable IPC-81 cytotoxicity profiling of ILs. In addition to test the predictive capabilities of the system, its potential as the core of a virtual screening (VS) strategy directed to prioritise safe ILs will be demonstrated. Additionally, we complemented these results by applying the NSG approach to the mining of SAR trends relevant for the cytotoxicity of ILs, namely, structure-cytotoxicity relationships (SCR) trends which can be used as useful tips guiding the molecular design of new and safe ILs. Finally, the SCR information concurrently gathered by both, quantitative (CART classifier) and qualitative (NSG) approaches was used to design a focused combinatorial library enriched with potentially safe ILs.

2. Results and Discussion

Cytotoxicity CART Classifier. The main goal of this work is to derive a reliable tool for the automatic prioritisation of safe (low cytotoxicity) ILs. The decision tree corresponding to the simplest best performing CART classifier found is shown in Figure 1.

In general terms, the classifier exhibits a good classification performance. The levels of accuracy (ILs correctly classified), sensitivity (Class_1 ILs correctly classified) and specificity (Class_0 ILs correctly classified) achieved by the CART were around 86%, evidencing the discrimination power and statistical significance of the pattern found. See details in Table 1.

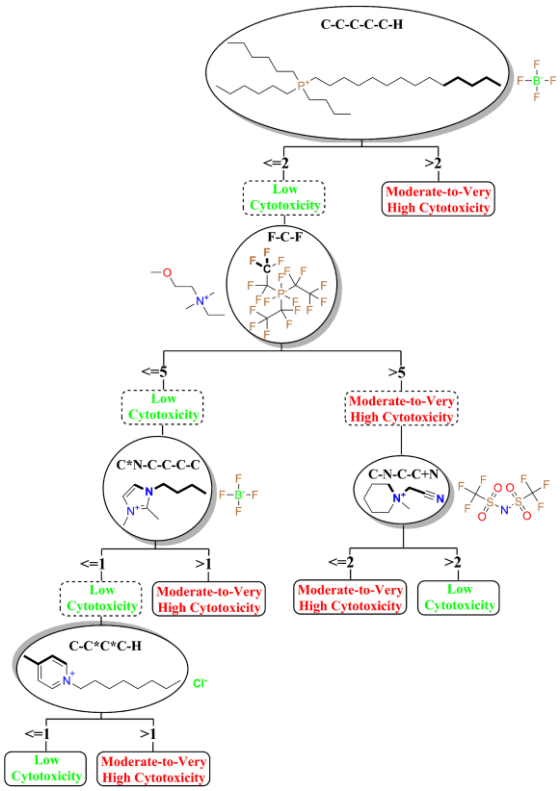

Figure 1. Chemically interpretable decision tree corresponding to the best preforming CART classifier found.

Table 1. Classification matrix and classification performance metrics of the CART classifier for the training, test and external evaluation sets.

		TRAINING SET		TEST SET		EXTERNAL EVALUATION SET	
		Observed					
		0	*1*	*0*	*1*	*0*	*1*
Predicted	*0*	**125**	8	**20**	2	**25**	3
	1	21	**51**	4	**8**	5	**9**
Acc. (%)		85.85		82.35		80.95	
Se. (%)		86.44		80.00		75.00	
Sp. (%)		85.62		83.33		83.33	
FN (%)		13.56		20.00		25.00	
FP (%)		14.38		16.67		16.67	
MCC (%)		68.34		60.39		55.90	
$F_{\text{Class 1}}$ *(%)*		77.86		72.73		69.23	
$F_{\text{Class 0}}$ *(%)*		89.6		86.96		86.21	

Acc.: Accuracy; *Se.*: Sensitivity or true positives (*TP*) rate; *Sp.*: Specificity or true negatives (*TN*) rate; *FN*: False negatives (*FN*) rate; *FP*: False positives (*FP*) rate; *MCC*: Matthews correlation coefficient; $F_{Class\ 1}$: F-measure for Class

Cytotoxicophores Identification. The influence of the SMFs must be interpreted as a function of the level occupied by the respective SMFs in the decision tree (the influence of the SMF decreases from the base to the leaf of the tree). So, considering the structure of the decision tree depicted in Figure 1 and the structural information of the SMFs it is possible to identify several moieties on ILs inducing a moderate-to-very high cytotoxicity that can be regarded as "cytotoxicophores". According to this analysis, in order of influence, the cytotoxicophores identified are:

- Cationic linear alkyl side chain of length > 5.
- Anions with highly fluorinated alkyl side chains (a fluorocarbonated side chain of length ≥ 2 or two or more trifluoromethyl groups).
- Cationic aromatic N-heterocycles with linear alkyl side chain of length ≥ 4.
- Six membered aromatic rings with a methyl substituent, which can be either the cation head group or its substituent.

Only one moiety was found to have a positive influence on the cytotoxicity profile of ILs, reducing their cytotoxicity from moderate-to-very high to low:

- Short alkyl side chains functionalized with polar nitrile groups on (essentially although not restricted to) aliphatic cation head groups containing nitrogen atoms.

It is important to highlight that the five SMFs identified can also be directly used as cytotoxicophores suitable for automatic procedures of ILs prioritisation such as expert systems, in addition to the cytotoxicity CART classifier proposed in this work.

Joint Use of CART Classifiers and Group Fusion Similarity Searches for the Automatic Prioritization of Safe ILs. The use of the cytotoxicity CART as a virtual screening tool could provide a practical solution to the automatic prioritisation of safe (poorly cytotoxic) ILs.

First, a group fusion similarity search (GFSS) approach [5] was applied. The set of reference structures consist of 20 structurally diverse ILs of lowest cytotoxicity, specially focused on the anion species. The degree of structural proximity by the corresponding values of 1 – the normalized Euclidean distance (1–ED). Finally, the set of 1–ED values between each reference IL and each database IL is combined into a fused similarity score (ε) by averaging the 20 corresponding 1–ED values. In this way, ε "captures" the structural patterns determining ILs of low cytotoxicity and thus can be used independently as a ranking criterion in a GFSS task. However, ε was derived to modify PP_{Class_1} and attain the variability required for library ranking. So, the result of using ε as a weighing factor of PP_{Class_1} is a new scoring metric that quantifies the likelihood of an IL to exhibit a favourable cytotoxicity profile based on probabilistic (PP_{Class_1}) and structural similarity (ε) criteria. This new scoring metric will be denoted from now on as Π and it is defined as the geometric mean of PP_{Class_1} and ε ($\Pi = \sqrt{PP_{Class\_1} \times \varepsilon}$).

So, decided to simulate an experiment to evaluate the ability of the approach to retrieve just those 12 ILs of low cytotoxicity (Class 1) of the external evaluation set dispersed in the full set of 200 ILs of moderate-to-very high cytotoxicity (Class 0). For comparison purposes we decided to estimate also the enrichment ability of the independent use of the GFSS approach by using as ranking criterion the fused similarity score ε.

The respective values of *AUAC* and *ROC* metrics obtained from the application of the CART-GFSS approach suggest that it is able to rank a safe IL earlier than an IL of moderate-to-very high cytotoxicity with a probability > 0.85. Instead, the values of these metrics obtained for the GFSS approach show a still good overall enrichment performance (*ROC* = 0. 78), but inferior to the CART-GFSS approach by about 8%.

Table 2. Classic and early recognition enrichment metrics computed to evaluate the enrichment performance of the CART-GFSS and GFSS approaches, respectively.[a]

Metric	*CART-GFSS*	*GFSS*
Classic Enrichment Metrics		
AUAC	0.8557(±0.0014)	0.7775(±0.0013)
ROC	0.8771(±0.0015)	0.7942(±0.0014)
$EF_{1\%}$	11.7778(±0.3200)	11.7778(±0.3200)
$EF_{5\%}$	6.4242(±0.0766)	4.8182(±0.0574)
$EF_{10\%}$	5.6212(±0.0449)	3.2121(±0.0257)
$EF_{20\%}$	3.6977(±0.0197)	3.2868(±0.0175)
Early Recognition Metrics		
$RIE_{1\%}$	6.9142(±0.1692)	6.9116(±0.1691)
$RIE_{5\%}$	6.5692(±0.0717)	5.8978(±0.0643)
$RIE_{10\%}$	5.3265(±0.0397)	4.3204(±0.0322)
$RIE_{20\%}$	3.9049(±0.0190)	3.1222(±0.0152)
$BEDROC_{1\%}$	0.3914(±0.0234)	0.3913(±0.0234)
$BEDROC_{5\%}$	0.4435(±0.0053)	0.3982(±0.0047)
$BEDROC_{10\%}$	0.5042(±0.0028)	0.4089(±0.0023)
$BEDROC_{20\%}$	0.6065(±0.0014)	0.4849(±0.0012)

a: The relative error associated to each enrichment metric is reported. *AUAC*: area under the accumulation curve; *ROC*: area under the ROC curve; $EF_{1\%/5\%/10\%/20\%}$: enrichment factor at χ = 1%/5%/10%/20%, respectively; $RIE_{1\%/5\%/10\%/20\%}$: robust initial enhancement at χ = 1%/5%/10%/20%, respectively; $BEDROC_{1\%/5\%/10\%/20\%}$: Boltzmann-enhanced discrimination of ROC at χ = 1%/5%/10%/20%, respectively.

The analysis of *RIE* at the respective top 1%, 5%, 10% and 20% fractions also points to an attractive early recognition ability of both approaches, consistently favouring the CART-GFSS approach. This pattern is also observed when the metric analysed is *BEDROC*. See details in Table 2.

Network-like Similarity Graph SAR Mining. The analysis was directed to detect in the ILs NSG highly discontinuous regions (clusters of ILs)

encoding minimal structural variations leading to significant cytotoxicity changes, with a special interest on those containing cytotoxicity cliffs. Figure 2 shows the NSG obtained at 95% similarity threshold.

This network is characterised by the coexistence of regions with continuous and discontinuous SARs. Regions of continuous SAR are characterized by clusters of small green nodes whereas discontinuous regions involve clusters composed of large green and red nodes (highlighted with a circle around). Clusters of ILs combining large red and green nodes connected by an edge are cytotoxicity cliff markers that can be easily identified. These types of cluster were visually inspected in order to identify the key structure-cytotoxicity relationship (SCR) trends dominating this ILs network.

The most significant cytotoxicity cliff pair in this network (see Figure 2) it is constituted by *N-Methyl-N,N-dioctyl-1-octanaminium bis(trifluoromethylsulfonyl)imide* and *N-Ethyl-N,N-dimethyl-1-butanaminium bis(trifluoromethylsulfonyl)imide* (nodes 1 and 2 in the network) which is a clear example of the influence of the alkyl side chain length over the cytotoxicity of ILs [6-10].

The cluster including the ILs represented by nodes 3, 4 and 5 clearly suggest the effect of highly fluorinated anions over cytotoxicity. A quite explicit correlation between the degree of fluorination of the anion and cytotoxicity it is observed, as reported in previous studies [11].

The last cluster analysed (nodes 6, 7, 8 and 9) suggest a weak influence of the cation head group over cytotoxicity. However, another previous finding can be confirmed in this cluster: the relatively higher cytotoxicity of aromatic cation head groups [12, 13].

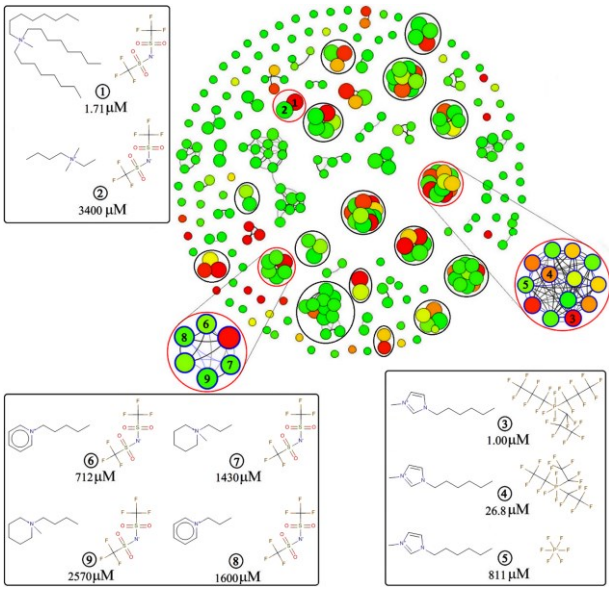

Figure 2. NSG constructed with the software SARANEA for a set of 281 ionic liquids using a Tanimoto similarity threshold of 0.95. The molecular structure and the respective EC50 values for IPC-81 leukaemia rat cell lines of the nine ionic liquids conforming the three clusters analysed are highlighted in three respective square boxes. These three clusters are highlighted in the NSG by red circles while the rest of clusters including ILs inducing a high discontinuity (connected large red and green nodes) are highlighted with black circles and further subjected to SAR pathway analysis.

Design and Assembling of a Focused Combinatorial Library Enriched with Potentially Safe ILs. Finally, the SCR trends identified were used to assemble a focused combinatorial library enriched with potentially safe ILs. The final result is a focused combinatorial library of 697748 ILs. We estimate the quality of the library assembled based on the use of a combined scoring metric (Π) that quantifies the likelihood of a IL to exhibit a favourable cytotoxicity profile. The values of Π near to 1 will be obtained for ILs with a high probability of exhibiting a favourable cytotoxicity profile. The analysis of the combinatorial library revealed that 75.57% of the ILs in the library exhibited values of $\Pi \geq 0.8$, while just 17.72%

exhibited values of $\Pi < 0.5$. The mean value of Π obtained for the library was of 0.67. Considering these values one can expect that an IL randomly selected from the library assembled will have a probability of exhibiting a favourable cytotoxicity profile around 67%.

3. Materials and Methods

Data Collection. The UFT/Merck IL DB reports the half cytotoxic concentration (EC_{50}) values (expressed in micromolar units) towards the rat leukemia cell line IPC-81 for 309 ILs and related salts.

Structure Codification. The structural codification was conducted by using the approach proposed by Prof. Varnek´s group and depicted in [14].

Design of the Experiment. The dataset of 281 ILs was subdivided by applying an EC_{50} threshold of 5000 μM into 81 safe or low toxicity ILs (Class_1) and 200 ILs with moderate-to-very high toxicity (Class_0). Once the classes were assigned, we proceeded to split the dataset into three subsets: training, test and external evaluation sets, as part of the model validation scheme [15].

Feature selection, Modelling and Validation. The full vector of ISIDA SMFs was reduced by means of the mRMR software [16] to a minimally redundant vector of size 50 composed of 9/41 anion/cation SMFs. Once this subset was found, the definitive subset of features, and consequently the final classification model, was directly determined by using the Classification and Regression Trees (CART) approach implemented on the *Data Mining* module of STATISTICA 8.0. Both the learning and predictive ability of the classification tree model were assessed by checking overall and class-specific performance measures on training, test and external evaluation sets, respectively [17].

Enrichment Analysis. The main goal in a virtual screening effort is to select a subset from a large pool of compounds (typically a compound database or a virtual library) and try to maximise the number of known actives in this subset. That is, to select the most "enriched" subset as possible. Several enrichment metrics have been proposed in the literature to measure the enrichment ability of a VS protocol [18]. In this work, we use some of the most extended metrics.

Network-like Similarity Graphs Analysis. For this task we resort to SARANEA [9], a freely available program that implements a graphical user interface to NSGs and NSG-based data mining techniques. In SARANEA, as a criterion for edges between nodes in NSGs, connected ILs needed to exceed a predefined Tanimoto similarity threshold value. To search for highly discontinuous regions in the network containing "cytotoxicity cliffs" pairs encoding critical structure variations for cytotoxicity we used a Tanimoto similarity threshold of 0.95.

Combinatorial Library Generation. The assembling of the focused combinatorial library was based on three sets of 15 cationic head groups, 20 cationic side chains and 31 anions previously identified as favouring the cytotoxicity behaviour of ILs. A combinatorial library of 22508 unique cations was generated with the aid of the SmiLib software [19] by using as inputs the corresponding SMILES notation of the two sets of head groups and side chains. The SmiLib software generated an SDF file comprising 22508 unique cations. The SDF file comprising the 31 anions was generated by using the ChemAxon´s JChem for Excel software [20]. Both, cation's and anion's SDF files were submitted to the ISIDA Fragmentor software [21] to compute the corresponding 371/2136 SMFs used to establish the structural reference space for the similarity assessment of the initial set of 281 ILs. Finally, the corresponding SVM output files provided by the ISIDA Fragmentor were converted to a fixed format/length vector file and concatenated into a unique vector file of size 2507 (including the

corresponding vector files of 371/2136 anion/cation SMFs) for each one of the 697748 ILs of the combinatorial library. The similarity assessment and the corresponding basic statistical analysis of the combinatorial library were conducted by using a MatLab implementation developed in our group.

Conclusions

In this work we have derived a reliable, predictive, simple and chemically interpretable CART classifier enabling the prioritisation of ILs with a favourable cytotoxicity profile. The analysis of the structure of the corresponding decision tree allowed us to identify several moieties that can be regarded as "cytotoxicophores. We also demonstrated the suitability of the joint use of the CART classifier and a group fusion similarity search (the CART-GFSS approach) as a virtual screening strategy for the automatic prioritisation of safe On the other hand, the NSG approach and NSG-based data mining techniques implemented on SARANEA have proved to be an efficient tool to mine relevant SCR information guiding the design of potentially safe ILs. The adaptation of the NSG approach proposed here to the particular and special case of disconnected molecular structures such as ILs also contributes to the integration of approaches like the traditional T-SAR analysis and the computational mining and visualisation of relevant SCRs of this interesting family of chemicals. Finally, the SCR information gathered from both quantitative (CART classifier) and qualitative (NSG) approaches guided the design of a focused combinatorial library of about 700000 ILs with a likelihood to exhibit a favourable cytotoxicity profile of about 80%. Such a virtual library represents a valuable decision making element for the development of ILs for various technical applications that fulfil the principles of green chemistry.

Acknowledgments

This work is supported by projects financed by FEDER through COMPETE – Programa Operacional Factores de Competitividade and by FCT – Fundação para a Ciência e a Tecnologia [PTDC/EQU-FTT/104195/2008, Pest-C/EQB/LA0006/2011, Pest-C/EQB/LA0020/2011]. M.C-M acknowledges also FCT for the grant [SFRH/BPD/90673/2012] co-financed by the European Social Fund.

Author Contributions

All the authors contributed equally.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Ranke, J.; Stolte, S.; Stormann, R.; Arning, J.; Jastorff, B., Design of sustainable chemical products--the example of ionic liquids. *Chem. Rev.* **2007**, 107, (6), 2183-206.
2. Sheldon, R. A., Green solvents for sustainable organic synthesis: state of the art. *Green Chem.* **2005,** 7, (5), 267-278.
3. Rogers, R. D.; Seddon, K. R., *Ionic Liquids As Green Solvents: Progress and Prospects*. American Chemical Society: 2003; Vol. 856, p 620.
4. Alvarez-Guerra, M.; Irabien, A., Design of ionic liquids: an ecotoxicity (Vibrio fischeri) discrimination approach. *Green Chem.* **2011,** 13, (6), 1507.
5. Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006,** 11, (23-24), 1046-53.
6. Stumpfe, D.; Bajorath, J., Methods for SAR visualization. *RSC Advances* **2012,** 2, 369-378.

7. Wawer, M.; Bajorath, J., Extracting SAR Information from a Large Collection of Anti-Malarial Screening Hits by NSG-SPT Analysis. *ACS Med. Chem. Lett.* **2011,** 2, 201-206.
8. Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J., Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discov. Today* **2010,** 15, (15/16), 630-639.
9. Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J., SARANEA: a freely available program to mine structure–activity and structure–selectivity relationship information in compound data sets. *J. Chem. Inf. Model.* **2009,** 50, 68-78.
10. Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J., Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.* **2008,** 51, 6075-6084.
11. Stolte, S.; Arning, J.; Bottin-Weber, U.; Matzke, M.; Stock, F.; Thiele, K.; Uerdingen, M.; Welz-Biermann, U.; Jastorff, B.; Ranke, J., Anion effects on the cytotoxicity of ionic liquids. *Green Chem.* **2006,** 8, (7), 621.
12. Ranke, J.; Muller, A.; Bottin-Weber, U.; Stock, F.; Stolte, S.; Arning, J.; Stormann, R.; Jastorff, B., Lipophilicity parameters for ionic liquid cations and their correlation to in vitro cytotoxicity. *Ecotoxicol. Environ. Saf.* **2007,** 67, (3), 430-8.
13. Stolte, S.; Arning, J.; Bottin-Weber, U.; Müller, A.; Pitner, W.-R.; Welz-Biermann, U.; Jastorff, B.; Ranke, J., Effects of different head groups and functionalised side chains on the cytotoxicity of ionic liquids. *Green Chem.* **2007,** 9, (7), 760.
14. Billard, I.; Marcou, G.; Ouadi, A.; Varnek, A., In silico design of new ionic liquids based on quantitative structure-property relationship models of ionic liquid viscosity. *J. Phys. Chem. B* **2011,** 115, (1), 93-8.
15. Tropsha, A., Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010,** 29, 476-488.
16. Peng, H.; Long, F.; Ding, C., Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Machine Intel.* **2005,** 27, (8), 1226-1238.
17. Witten, I. H.; Frank, E., Chapter 5: Credibility: Evaluating what's been learned. In *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Gray, J., Ed. Morgan Kaufman: San Francisco, CA, 2005; pp 143-186.
18. Truchon, J. F.; Bayly, C. I., Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007,** 47, (2), 488-508.
19. Schüller, A.; Schneider, G.; Byvatov, E., SMILIB: Rapid Assembly of Combinatorial Libraries in SMILES Notation. *QSAR & Comb Science* **2003,** 22, 719-721.
20. ChemAxon *JChem for Excel*, version 5.10.2.725; 2012.
21. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I.; G., M., ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008,** 4, 191-198.