Journal of Hazardous Materials 352 (2018) 17–26

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids

Lingdi Cao[a,1], Peng Zhu[b,c,1], Yongsheng Zhao[b,c,*], Jihong Zhao[d,e,**]

[a] Forschungszentrum Jülich GmbH, Helmholtz Institute Erlangen-Nürnberg for Renewable Energy (IEK-11), Egerlandstr. 3, 91058, Erlangen, Germany
[b] Key Laboratory for Thin Film and Microfabrication of Ministry of Education, Department of Micro/Nano-electronics, Shanghai Jiao Tong University, Shanghai, 200240, China
[c] Department of Chemical Engineering, University of California, Santa Barbara, CA, 93106-5080, USA
[d] Collaborative Innovation Center of Environmental Pollution Control and Ecological Restoration, Zhengzhou, Henan, 450001, China
[e] Xuchang University, Xuchang, Henan, 461001, China

ARTICLE INFO

Keywords:
Toxicity
Ionic liquids
Quantitative structure-activity relationship
Extreme learning machine
Quantum chemistry descriptors

ABSTRACT

Large-scale application of ionic liquids (ILs) hinges on the advancement of designable and eco-friendly nature. Research of the potential toxicity of ILs towards different organisms and trophic levels is insufficient. Quantitative structure-activity relationships (QSAR) model is applied to evaluate the toxicity of ILs towards the leukemia rat cell line (ICP-81). The structures of 57 cations and 21 anions were optimized by quantum chemistry. The electrostatic potential surface area (S_{EP}) and charge distribution area ($S_{\sigma\text{-profile}}$) descriptors are calculated and used to predict the toxicity of ILs. The performance and predictive aptitude of extreme learning machine (ELM) model are analyzed and compared with those of multiple linear regression (MLR) and support vector machine (SVM) models. The highest R^2 and the lowest AARD% and RMSE of the training set, test set and total set for the ELM are observed, which validates the superior performance of the ELM than that of obtained by the MLR and SVM. The applicability domain of the model is assessed by the Williams plot.

1. Introduction

Ionic liquids (ILs) have been considered as the environmentally friendly materials based on the principles of green engineering [1]. Intensive research interests on ILs were given to develop their potential applications in many areas, such as gas capture and separation [2–6], organic synthesis [7–9], electrochemical reaction [10–12], catalysis modification [13–16], and so on. However, the properties of ILs, such as thermal stability and non-volatility [17], might pose environmental threats due to their nature of slow degradation. In the context of mass commercialization, the toxic evaluation of perceived environmentally friendly ILs has aroused broad attention except for their physical properties. It is now realized that the previously acknowledged notion of low toxicity for ILs has been proved to be inaccurate and to some extent ILs have hazard potentials for the human being and the environment [18]. It is reported that there is a direct relationship between the toxicity of ILs and their hydrophobicity [19]. For example, the release of ILs from different industrial processes into environments will cause water contamination [20]. Docherty et al. stated that studies of toxicity testing of ILs will definitely offer insights for engineers on tailoring IL-synthesis to particular industrial processes, instead of releasing harmful compounds to the environment [21]. Therefore, it is imperative to evaluate the toxic effects of ILs and determine their further consequences to the environmental fate.

Since the work of Jastorff et al. in 2000 [22], many attempts were extensively devoted to understanding and identifying the toxicity of the ILs [20,21,23–25]. Romero and co-workers determined the acute toxicity and EC_{50} values of each compound in the aqueous solution using the Microtox® standard procedure. Their results showed that the short chain length of side chain R_2 on imidazolium cation is positive to the low toxic effect, while the anion has a minor effect on the toxicity of ILs [26]. To better understand the toxic influence of anion, Pereira et al. presented their toxicological assessment of a group of environmentally friendly ILs with the benign cholinium cation and linear alkanoate anions, using filamentous fungi as model eukaryotic organisms [20]. They found that 1) toxicity of ILs increase with the elongation of the

* Corresponding author at: Key Laboratory for Thin Film and Microfabrication of Ministry of Education, Department of Micro/Nano-electronics, Shanghai Jiao Tong University, Shanghai, 200240, China.
** Corresponding author at: Collaborative Innovation Center of Environmental Pollution Control and Ecological Restoration, Zhengzhou, Henan, 450001, China.
E-mail addresses: yzhao01@ucsb.edu (Y. Zhao), zjh@zzuli.edu.cn (J. Zhao).
[1] These authors contributed equally.

https://doi.org/10.1016/j.jhazmat.2018.03.025
Received 1 November 2017; Received in revised form 2 March 2018; Accepted 14 March 2018
Available online 15 March 2018
0304-3894/

linear chain in the anion, which is the similar conclusion drawn by Lima and Coutinho [27]; 2) branching resulted in lower toxicity because it usually depresses lipophilicity [20]. At multiple levels, Wang et al. investigated the interactions of fullerene C60 with imidazolium-based ILs with different alkyl side chain lengths and anionic types [28,29]. Their results showed that π-cation interaction contributed to the mechanism of the C60-IL interaction more than π-anion interaction. Wang et al. also predicted the joint toxicity effects in the green alga Scenedesmus obliquus exposed to binary mixtures of intrinsic graphene (iG)/graphene oxide (GO) and five ILs. They found that the isolated ILs in the binary mixtures were the main contributors to toxicity. They believed that the mechanism of the joint toxicity may be associated with the adsorption capability of the graphenes for the ILs, the dispersion stability of the graphenes in aquatic media, and modulation of the binary mixtures-induced oxidative stress [30]. Quantitative structure-property relationships (QSPR) or quantitative structure-activity relationships (QSAR) studies have already been used in correlation and prediction of properties [24,31–35]. Zhang's group established a comprehensive database on the toxicity of ILs up to 4000 pieces and then used QSAR model to qualitatively analyze the relationship between the structure and toxicity of ILs [25]. They proposed that the nonlinear model developed by supported vector machine (SVM) algorithm is more reliable in the prediction of toxicity of ILs, which will be meaningful in designing novel environmentally benign agents. Singh et al. studied the chemical attributes of a wide variety of ILs towards their inhibitory potential of acetyl cholinesterase enzyme (AChE) using SVM and cascade correlation network (CCN) [36]. Their results showed that proposed QSAR models bear more statistical confidence, especially with respect to external validation which has not been focused on in other studies. This work successfully proposed QSAR in predicting different toxicity classes and precise toxicity end-point of ILs.

It is noted that the above-mentioned algorithms are still insufficient, and therefore, more advanced models should be developed. Additionally, it is proved that molecular descriptors play vital roles in building models [37]. The distribution area of the σ-profile ($S_{\sigma\text{-profile}}$) has been considered as an *a* priori quantum-chemical descriptor that quantitatively represents the molecule's polar surface screen charge on the polarity scale and it can be achieved from the histogram function σ-profile given by COSMO computation [38]. It has been effectively used as a parameter in QSAR models to predict the toxicity [39] and viscosity [38] of ILs, and the advantages have been proved. The electrostatic potential surface for molecules, which means the molecular surface areas in the interval of different electrostatic potential, has the ability to show the rich information at electron level and therefore it is expected to be used as descriptors for predicting properties of materials. Therefore, the above-mentioned two kinds of descriptors have been utilized as parameters to build models. Herein, we hope to build models using electrostatic potential surface area (S_{EP}) and the $S_{\sigma\text{-profile}}$ calculated based on quantum chemistry as the input parameters, combined with a novel algorithm, namely, extreme learning machine (ELM). First, the ELM was employed to predict the toxicity of ILs towards leukemia rat cell line (IPC-81) using the calculated descriptors. Then the results were compared with those of different QSAR models, including SVM and MLR. Our work establishes a platform offering insights on the limited information of a huge number of ILs and assessment of the environmental impact.

2. Methodology and database

2.1. Dataset and structural descriptors

In this work, toxicity data set of 119 ILs for IPC-81 (EC_{50} values) was chosen from the widely acknowledged ILs database [40,41]. The structures of 119 ILs (57 cations and 21 anions) were geometrically optimized by the *Gaussian* 09 software B.01 at the theoretical level of B3LYP/6-31 + + G(d, p) and specially LanL2DZ was used for I^- ion [42]. The optimized results were listed in Table S1-2. EC_{50} values (μM) for IPC-81 are converted into the form of the logarithm of half maximal effective concentration, written as $\log_{10}(EC_{50})$. The whole data set is divided into two parts: a training set of 80% ILs to build the model and a set of the remaining of 20% ILs to evaluate the model's prediction ability. Namely, the compounds numbered from 1 to 99 belong to the training set and the remaining compounds numbered from 100 to 119 are test set. Each category of the total data is separated by random selection. Usually, the related compound is represented by theoretic molecular descriptors. The choice of descriptors is of vital significance to the performance of predictive models. In this work, two kinds of descriptors: S_{EP} and $S_{\sigma\text{-profile}}$ are applied to develop the models. Herein, S_{EP} and $S_{\sigma\text{-profile}}$ are calculated by different programs based on the optimal structures of cations and anions. The electrostatic potential V(*r*) is produced at the point *r* around a molecule, via its nuclear and electrons, i.e. it is calculated via the static distribution of a molecule [43]. The molecular electrostatic potential is expressed rigorously by Eq. (1)

$$V(r) = \sum_A \frac{Z_A}{|r_A - r|} - \int \frac{\rho(r')dr'}{|r' - r|} \quad (1)$$

where Z_A denotes the charge of the nuclear A, located at r_A, $|r_A - r|$ stands for the distance between nucleus A and r, $\rho(r)$ represents the electronic density function for the molecule, $\rho(r')dr'$ represents the electronic charge increment in each volume element and $|r' - r|$ is its distance from r. As can be seen from the Eq. (1), the electrostatic potential is composed of two parts, atomic charge and electron density contribution. First, the S_{EP} files of the corresponding cations and anions were calculated using the Mutiwfn software with the electrostatic potential range of 0–150 kcal/mol for cations, −150 to 0 kcal/mol for anions [44]. Second, the COSMO files of cations and anions were calculated using *Gaussian* 03 software [45] based on the structures optimized by *Gaussian* 09. Third, the $S_{\sigma\text{-profile}}$ files of the corresponding cations and anions were calculated by a MATLAB code. Each σ-profile ranges from −0.03 to 0.03 e/Å^2 and the step size is 0.001 e/Å^2. For convenience, we used the intermediate value to stand for the S_{EP} of each step, for instance, $S_{EP0.25}$ represents the electrostatic potential surface areas in the range of 0–0.5 kcal/mol. Specifically, we take cation 1-(cyanomethyl)-1-methylpiperidinium and anion 2-(2-methoxyethoxy)ethylsulfate as examples to show the representative S_{EP} and $S_{\sigma\text{-profile}}$ in this work, as seen in Figs. 1 and 2, respectively. It can be seen from Fig. 1 that the darker color in blue for cation 1-(cyanomethyl)-1-methylpiperidinium or red for anion 2-(2-methoxyethoxy)ethylsulfate means the much stronger polarity. The similar situation can be found for distribution area of the σ-profile as revealed in Fig. 2.

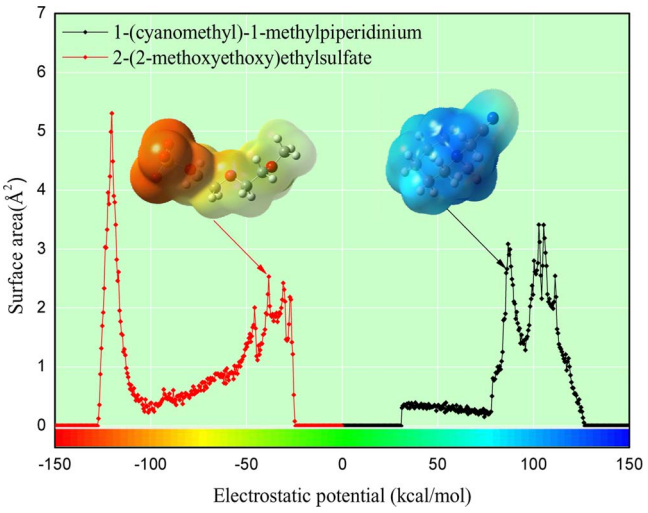

Fig. 1. S_{EP} of a representative cation and anion of IL.