The Ionic Liquid Property Explorer: An Extensive Library of Task-Specific Solvents

Vishwesh Venkatraman *, Sigvart Evjen and Kallidanthiyil Chellappan Lethesh*

Department of Chemistry, Norwegian University of Science and Technology, 7491 Trondheim, Norway; sigvart.evjen@ntnu.no (S.E.); lethesh.k.chellappan@ntnu.no (K.C.L.)
* Correspondence: vishwesh.venkatraman@ntnu.no

Received: 25 April 2019; Accepted: 21 June 2019; Published: 21 June 2019

Abstract: Ionic liquids have a broad spectrum of applications ranging from gas separation to sensors and pharmaceuticals. Rational selection of the constituent ions is key to achieving tailor-made materials with functional properties. To facilitate the discovery of new ionic liquids for sustainable applications, we have created a virtual library of over 8 million synthetically feasible ionic liquids. Each structure has been evaluated for their-task suitability using data-driven statistical models calculated for 12 highly relevant properties: melting point, thermal decomposition, glass transition, heat capacity, viscosity, density, cytotoxicity, CO₂ solubility, surface tension, and electrical and thermal conductivity. For comparison, values of six properties computed using quantum chemistry based equilibrium thermodynamics COSMO-RS methods are also provided. We believe the data set will be useful for future efforts directed towards targeted synthesis and optimization.

Dataset: The datasets used for machine learning can be accessed at dx.doi.org/10.5281/zenodo.3251643. The SQLite database containing computed properties and a graphical user interface for querying, are available from dx.doi.org/10.5281/zenodo.3251661.

Dataset License: The data set is made available under the Creative Commons License CC BY 4.0.

Keywords: ionic liquids; machine learning; database; properties; combinatorial screening

1. Introduction

Ionic liquids (ILs) comprised of cations (mostly organic) and anions (both organic and inorganic) provide a widely applicable set of building blocks for advanced functional materials. Demonstrated applications include coatings and lubricants [1], pharmaceuticals [2], fuel cells [3], and catalysis [4]. Desirable properties such as high thermal and electrochemical stabilities, together with a negligible volatility, make them well-suited for developing novel and innovative materials. By making simple changes to the structure of the constituent ions, the chemical makeup can be altered to create the optimum solvent for a given application. The challenge, however, is to identify optimal task-specific ILs from the near-infinite combinations of the constitutive ions and functional groups [5,6]. Experimental work is limited to a small area of the ionic liquid chemical space, leaving many potentially promising compounds unexplored. Given the need to minimize experimental cost and time, the rational selection of suitable ILs from the available choices becomes paramount.

Currently, most ILs are discovered through laborious trial-and-error experiments. Given the diversity of the IL applications, the custom design of the solvents requires general knowledge of the properties,
to carry out even preliminary studies. Chemical intuition and experience play key roles in the selection. In a number of studies, approaches based on electronic structure theory have been used to understand the structure–property mechanisms [7,8]. The use of electronic structure methods [9] are, however, confined to a few systems, and despite advances in computer hardware, the associated computational costs are still prohibitive for rapid large scale screening. In recent times, machine learning (ML) based virtual screening have emerged as a powerful approach facilitating in silico searches over millions of compounds [10–12]. With the availability of data repositories such as ILThermo [13], there has been a steep rise in the use of such approaches for modelling ionic liquid properties as evidenced by recent publications [14–18].

With a view to expedite task-specific ionic liquid discovery, we have assembled a large library of ionic liquids spanning nine different cation scaffolds: ammonium, imidazolium, phosphonium, piperidinium, pyridinium, pyrrolidinium, morpholinium, azeepanium and sulphonium. These are combined with a diverse set of anions (alkylsulphonates, phenolates, phosphates, triazolides, HF₆, BF₄) yielding over 8 million compounds. For each IL, twelve properties of interest have been predicted using machine learning. The conductor-like screening model for real solvents (COSMO-RS) developed by Klamt and Eckert [19] has been shown to be a relatively robust predictive method for properties such as activity coefficients of molecular solutes in ionic liquids [20], gas separation capacity [21] and cellulose solubilities [22]. The COSMO-RS approach, however, requires computationally expensive density functional theory (DFT) calculations and has been shown to fail as much as it succeeds [23]. It nevertheless remains a popular choice for ionic liquid-based screening [24–26] and is likely to improve with better parameterization [27]. We have therefore included COSMO-RS estimates for selected properties. The database can be expanded by way of progressively adding new structures and associated properties. We expect this repository to be a playground for future work in this active research area.

2. Data Description

2.1. Workflow

Figure 1 provides a schematic overview of the database. Using SmiLib [28] as the combinatorial enumeration engine, 219,216 cations spanning nine different cation families were created. These were combined with 38 anions to yield a total of 8,333,096 ionic liquids. For each IL, values for twelve properties of interest are reported: melting point ($T_m$), glass transition temperature ($T_g$), thermal decomposition temperature ($T_d$), viscosity ($\eta$), density ($\rho$), heat capacity ($C_p$), CO₂ capacity ($x_{CO_2}$), electrical ($\kappa$) and thermal conductivity ($\lambda$), cytotoxicity towards the leukemia rat cell line IPC-81 (log₁₀EC₅₀), surface tension ($\sigma$) and refractive index ($n_D$). For each property, an estimate of the uncertainty associated (±1 standard deviation) with the prediction is included as a way to assess when the model is likely to be more accurate. Large standard deviations are typically associated with those that cannot be trusted [17,29].

2.2. Ionic Liquid Library

The library spans a wide range of functionalities, from simple alkyl-functionalized cations to more exotic structures. Moieties were selected based on ease of synthesis i.e., those that can be readily prepared, such as the availability of halide precursors for SN2-reaction preparation. Amine, ether and alcohol groups are able to coordinate to metals. Alcohols and amines can engage in H-bonding, allowing for selective interactions with hydrophilic compounds. Amines with different levels of basicity were incorporated into the database, facilitating more application-specific tuning of IL $pK_a$, e.g., selective probes for acid gases [30].
Figure 1. Fragments and other functional moieties used in the construction of the ionic liquid library.

The customized cation libraries were built using combinatorial enumeration (using the software SmiLib [28]) of building blocks attached to different scaffolds (see Table 1). The counterions were selected among common anion groups for ILs, as well as organic anions, which possess several interesting properties. For instance, acetates have a high cellulose solubility and phenolates can efficiently be used for extraction of acids [31,32]. Different carboxylates allow for the tuning of the IL properties. Acetylacetonate-based ILs could be used for selective metal extraction [33,34]. Sulfonate anions allow for the extraction of hydrocarbons [35], and have potential applications in batteries [36].

Table 1. The second column shows the number of cations obtained using combinatorial library enumeration. The final column gives the number of ionic liquids obtained.

<table>
<thead>
<tr>
<th>Cation</th>
<th>#Molecules</th>
<th># Ionic Liquids</th>
</tr>
</thead>
<tbody>
<tr>
<td>ammonium</td>
<td>179466</td>
<td>6819708</td>
</tr>
<tr>
<td>azepanium</td>
<td>5460</td>
<td>207480</td>
</tr>
<tr>
<td>imidazolium</td>
<td>5460</td>
<td>207480</td>
</tr>
<tr>
<td>morpholinium</td>
<td>5460</td>
<td>207480</td>
</tr>
<tr>
<td>phosphonium</td>
<td>7914</td>
<td>300732</td>
</tr>
<tr>
<td>piperidinium</td>
<td>5460</td>
<td>207480</td>
</tr>
<tr>
<td>pyridinium</td>
<td>1040</td>
<td>39520</td>
</tr>
<tr>
<td>pyrrolidinium</td>
<td>5460</td>
<td>207480</td>
</tr>
<tr>
<td>sulphonium</td>
<td>3576</td>
<td>135888</td>
</tr>
</tbody>
</table>

2.3. Graphical Summary of the Data Set

The variation in the values for the ML-predicted properties are shown in Figures 2–4. Values for \( C_p, \eta, \rho, \kappa, \lambda, x_{CO_2} \) are estimated at standard room temperature and pressure. For many properties, the estimated values mirror the trends observed in literature. For instance, the \( T_m \) for phosphonium- and ammonium-based ILs are higher than those based on the imidazolium scaffold [37–39]. These differences
are related to the highly symmetric nature of the tetraalkylammonium and tetraalkylphosphonium cations. The compounds in the library are thermally stable, with phosphonium-based ILs in particular showing higher values of $T_d$ compared to the other cationic cores (see Figure 2). This is attributed to the difference in the thermal degradation mechanisms [40]. In contrast, sulphonium based ILs show poorer stabilities owing to the unstable nature of the sulphur atom [41].

![Diagram showing distributions of thermophysical properties](image)

**Figure 2.** Violin plots showing the distribution of the ML predictions for thermophysical properties. Predicted values for $C_p$ and $\lambda$ are calculated at standard room temperature and pressure.

Low melting ILs are expected to have low viscosities. Analysis of the data suggests that less than 2% of the ILs have $\eta < 300$ cP and $T_m < 30^\circ$C. It has been reported that the viscosities of imidazolium and sulphonium ILs are on the lower side because of the asymmetric nature of the imidazolium cation and bulky nature of the sulphur atom [41,42]. The predicted viscosities show similar trends (shown in Figure 3). High refractive index ILs ($n_D > 1.60$) are much needed in optical microscopy studies of minerals [43]. While the predictions of $n_D$, in particular for imidazolium and pyridinium ILs, are in accordance with experimentally observed data for similar compounds [44,45], none of the predictions exceed 1.60. Imidazolium-based ILs generally have higher electrical conductivities compared to other ILs based on other cationic cores, owing to their lower viscosity [46]. In the predicted data, over 100 low-viscosity imidazolium ILs show promise as electrolyte materials in battery applications.
In all cases, anions play a crucial role in determining the behaviour of the ionic liquids. Fluorinated anions (PF₆, NTf₂, hfac), for instance, show higher CO₂ solubilities [47] (see Figure 4). Although considered environmentally friendly, many ILs are soluble in water which can be hazardous to aquatic organisms, if released into the aqueous system [48,49]. It has been shown that toxicity increases with the length of the alkyl chain (varying between 2 and 12 carbon atoms) attached to the cationic cores. The effect of anions has not been investigated broadly to allow a conclusive analysis. The graph for cytotoxicity shown in Figure 4 suggests that, of the 8.33 million combinations examined, only 0.2% of the ILs showed log₁₀ EC₅₀ > 3.4 μM) [50].

Figure 3. Violin plots showing the distribution of the ML-predictions for volumetric properties. The predicted values are calculated at standard room temperature and pressure.
Figure 4. Distribution of the predicted CO$_2$ mole fractions and log$_{10}$EC$_{50}$ with respect to the anion groups.

3. Methods

3.1. Machine Learning

Each structure was subjected to geometry optimization at the semi-empirical PM6 level using MOPAC [51]. Although fragment/group contribution descriptors have been popular, models trained on such variables often fail when presented with new fragments for which they were not trained. We have therefore chosen molecular descriptors that focus on charge distributions and geometrical indices which have been shown to yield good predictive performances for IL properties such as melting points [17], thermal decomposition temperatures [15], refractive indices [18] and CO$_2$ solubilities [52]. The variables were calculated independently for each cation and anion using the software KrakenX [53,54]. The top ranked variables (ranked according to the contribution of the variable to the response) in the models included the charged partial surface area descriptors (summarize the charge distribution in the ion), chemical reactivity parameters such as the HOMO/LUMO energies that are closely related to electrophilic/nucleophilic attack and the charge distribution in the ion, and softness (inverse of the HOMO-LUMO gap) which are indicative of the cation-anion electrostatic (nucleophilic-electrophilic) interactions. Experimental data for the properties was taken from various literature sources [13,17,18,52,55]. Machine learning models were evaluated for 12 different properties: melting points ($T_m$), glass transition temperatures ($T_g$), thermal decomposition temperatures ($T_d$), viscosities ($\eta$), densities ($\rho$), heat capacities ($C_p$), CO$_2$ capacity ($x_{CO2}$), electrical ($\kappa$) and thermal conductivities ($\lambda$), cytotoxicities towards the leukemia rat cell line IPC-81 (log$_{10}$EC$_{50}$), surface tension ($\sigma$) and refractive indices ($n_D$). The models were trained on 67% of the available data and a 5-fold cross validation technique was used to obtain performance statistics. Three different ML models were employed: generalized boosted regression models (GBM), random forests
(RF) and Cubist methods. For each property, the best performing model across both calibration and test data was determined based on standard evaluation metrics and used for further predictions.

For the obtained models, performance metrics including the squared coefficient of correlation ($R^2$), root mean square error (RMSE) and the mean absolute error (MAE) are reported in Table 2. The supplementary material lists the performances of all the ML models used for prediction. For most IL properties, the experimental and predicted values are in agreement with reported studies [14,15,17,18,52]. The larger deviations for $T_m$, $T_g$ and $\eta$ can be attributed to experimental variations, presence of impurities and water [56]. Model applicability domain methods often rely on the chemical similarity of a test set structure to members of the training set [57]. Here, we have chosen to associate each prediction with a bootstrapped uncertainty estimate [17,18]. In bootstrapping [58], for example, the training set is randomly sampled with replacement, and a model is built for each bootstrapped sample. For computational expediency, a total of 100 models were built, and each model was applied to a given test set compound to obtain a distribution of predictions. The uncertainty associated with the prediction was then calculated as the standard deviation of the distribution [29]. Working on the assumption that ILs with small uncertainties are likely to have small prediction errors, one can exclude compounds with moderate to high prediction uncertainties.

Table 2. Summary of the ML model performances for 12 different properties. In each case, the number of data points available for calibration and testing are provided. The metrics are reported the best performing model. EC50 values correspond to Rat cell line toxicities. $N_C$ and $N_A$ are the number of cations and anions, respectively.

<table>
<thead>
<tr>
<th>Property</th>
<th>$N_C$</th>
<th>$N_A$</th>
<th>ML</th>
<th>$N_{cal}$</th>
<th>RMSE (MAE)</th>
<th>$R^2_{cv}$</th>
<th>$N_{val}$</th>
<th>(MAE)</th>
<th>$R^2_{cv}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_m$ (°C)</td>
<td>1369</td>
<td>141</td>
<td>RF</td>
<td>1486</td>
<td>44 (15)</td>
<td>0.67</td>
<td>726</td>
<td>45 (33)</td>
<td>0.66</td>
</tr>
<tr>
<td>$T_g$ (°C)</td>
<td>327</td>
<td>109</td>
<td>Cubist</td>
<td>442</td>
<td>21 (2)</td>
<td>0.62</td>
<td>202</td>
<td>19 (12)</td>
<td>0.62</td>
</tr>
<tr>
<td>$T_d$ (°C)</td>
<td>538</td>
<td>192</td>
<td>RF</td>
<td>833</td>
<td>39 (12)</td>
<td>0.77</td>
<td>455</td>
<td>35 (25)</td>
<td>0.8</td>
</tr>
<tr>
<td>$\log_{10}(\eta)$ (mPa·s)</td>
<td>847</td>
<td>227</td>
<td>GBM</td>
<td>2465</td>
<td>0.17 (0.06)</td>
<td>0.94</td>
<td>3994</td>
<td>0.35 (0.23)</td>
<td>0.76</td>
</tr>
<tr>
<td>$\rho$ (kg/m³)</td>
<td>333</td>
<td>120</td>
<td>RF</td>
<td>9225</td>
<td>12.23 (2.65)</td>
<td>0.99</td>
<td>7731</td>
<td>49.10 (28.34)</td>
<td>0.93</td>
</tr>
<tr>
<td>$\ln(C_p)$ (J/K/mol)</td>
<td>115</td>
<td>48</td>
<td>GBM</td>
<td>6320</td>
<td>0.042 (0.012)</td>
<td>0.99</td>
<td>2763</td>
<td>0.28 (0.19)</td>
<td>0.91</td>
</tr>
<tr>
<td>$\gamma$ (N/m)</td>
<td>412</td>
<td>80</td>
<td>GBM</td>
<td>1464</td>
<td>0.006 (0.002)</td>
<td>0.97</td>
<td>1456</td>
<td>0.017 (0.011)</td>
<td>0.83</td>
</tr>
<tr>
<td>$x_{CO_2}$</td>
<td>78</td>
<td>74</td>
<td>GBM</td>
<td>6084</td>
<td>0.03 (0.01)</td>
<td>0.98</td>
<td>4839</td>
<td>0.09 (0.06)</td>
<td>0.86</td>
</tr>
<tr>
<td>$\log_{10}(EC_{50})$ µM</td>
<td>114</td>
<td>25</td>
<td>Cubist</td>
<td>157</td>
<td>0.52 (0.05)</td>
<td>0.79</td>
<td>70</td>
<td>0.40 (0.30)</td>
<td>0.86</td>
</tr>
<tr>
<td>$\kappa$ (S/m)</td>
<td>158</td>
<td>80</td>
<td>GBM</td>
<td>1433</td>
<td>0.05 (0.01)</td>
<td>0.98</td>
<td>1251</td>
<td>0.15 (0.10)</td>
<td>0.79</td>
</tr>
<tr>
<td>$\lambda$ (W/m/K)</td>
<td>28</td>
<td>28</td>
<td>GBM</td>
<td>326</td>
<td>0.005 (0.002)</td>
<td>0.95</td>
<td>147</td>
<td>0.009 (0.006)</td>
<td>0.89</td>
</tr>
</tbody>
</table>

3.2. COSMO-RS Evaluation

COSMO-RS, a quantum chemistry based method, was used to evaluate selected properties [59]. For each cation and anion, geometry optimizations using the density functional theory (DFT) functional B88-PW86 with a triple zeta valence polarized basis set [60] (TZVP) and the resolution of identity standard approximation were performed. Values of $\eta$, $\rho$, $C_p$, $x_{CO_2}$, $\kappa$ and $T_m$ were calculated using the COSMOtherm software with the parameterization set BP_TZVP_C30_01601) [61].

4. Database Exploration and Use

An interactive graphical user interface written in the Java programming language is provided as a way to query the library (see Figure 5). All data including machine learning and COSMO-RS predictions have been compiled into an SQLite database [62]. Structures can be searched (exact or substructure) as SMILES strings and IUPAC names of moieties. Alternatively, structure files in MOL format can be
uploaded. For conversion of IUPAC names to SMILES representations, the functionality in the OPSIN [63] library has been used. Filtering of compounds according to select property cut-offs allows for a targeted search. Query results may be additionally saved as a sortable HTML table for future inspection.

Figure 5. Interface to the Ionic Liquid Property Explorer is through a graphical user interface that allows for searching the database. Where available, COSMO-RS values are displayed as tool tips. The results can also be exported to a sortable HTML table.

Supplementary Materials: The following are available online at http://www.mdpi.com/2306-5729/4/2/88/s1.

Author Contributions: V.V. performed the data curation, developed the models and carried out the computations. S.E. and K.C.L. proposed the molecular library and tested the software. V.V wrote the paper with contributions from S.E. and K.C.L.

Funding: The Norwegian Research Council (NFR) is acknowledged for financial support through the CLIMIT program (Grant No. 233776).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IL Ionic Liquids  
ML Machine Learning  
DFT Density Functional Theory  
RF Random Forest  
GBM Generalized Boosted Models  
COSMO-RS Conductor like Screening Model for Real Solvents  
hfac Hexafluoroacetylacetonate  
DCA Dicyanamide  
NTf2 Bis(trifluoromethanesulfonyl)imide  
PF6 Hexafluorophosphate  
HOMO Highest Occupied Molecular Orbital  
LUMO Lowest Unoccupied Molecular Orbital
References


22. Kahlen, J.; Masuch, K.; Leonhard, K. Modelling cellulose solubilities in ionic liquids using COSMO-RS. Green Chem. 2010, 12, 2172. [CrossRef]


50. Torrecilla, J.S.; Palomar, J.; Lemus, J.; Rodríguez, F. A quantum-chemical-based guide to analyze/quantify the cytotoxicity of ionic liquids. *Green Chem.* 2010, 12, 123–134. [CrossRef]


52. Venkatraman, V.; Alsberg, B.K. Predicting CO2 capture of ionic liquids using machine learning. *J. CO2 Util.* 2017, 21, 162–168. [CrossRef]


