SAND2016-3631J

Screening for High Conductivity/Low Viscosity Ionic Liquids Using Product Descriptors

Shawn Martin, Harry D. Pratt III, and Travis M. Anderson*

Sandia National Laboratories, Albuquerque, New Mexico, 87185, USA

Abstract. Ionic liquids (ILs) consist of cation-anion pairs. Despite this fact, current efforts to predict IL properties using quantitative structure property relationships (QSPRs) treat the cations and anions separately, ignoring potential cross-correlations. Here we consider a method for treating ILs as pairs using product descriptors for QSPRs, a concept borrowed from the prediction of protein-protein interactions in bioinformatics. We demonstrate the method by predicting electrical conductivity, viscosity, and melting point on a dataset taken from the ILThermo database on June 18th, 2014. The dataset consists of 3,926 measurements taken from 165 ILs made up of 72 cations and 34 anions. We benchmark our QSPRs on the known values in the dataset then extend our predictions to screen all 2,448 possible cation-anion pairs in the dataset.

Introduction

Ionic liquids (ILs) are highly modifiable molten salts (usually considered to have melting points below 100 °C) used for a range of basic and applied studies[1-4]. ILs can exhibit high thermal stability, negligible vapor pressure, wide electrochemical window, and the ability to dissolve a range of organic and inorganic compounds. They have been studied in the context of separations, catalysis, and electrochemistry[5-10]. Although many of their properties can be systematically varied by compositional and structural changes, there is considerable ongoing effort to develop ionic liquids that simultaneously exhibit low viscosity and high conductivity.

In this paper, we consider the use of quantitative structure property relationships (QSPRs) to screen for high conductivity/low viscosity ILs. QSPRs have been shown to predict various properties of ILs[11,12], including viscosity[13-17], conductivity[18-22], and melting point[23-26]. These studies have used different datasets, generally available from the public domain; different methods for generating descriptors, including group contribution[18,22], ISIDA fragments[13,25], Gaussian 03[14,20,24], MOPAC[19,23], Chem3D[19], CODESSA[14,20,23,24], and DRAGON[16,25]; and different regression algorithms including Genetic Algorithms[16,18,19], Neural Networks[13,25,26], Multiple Linear Regression[14,15,20,23,24], and Support Vector Machines (SVMs)[21,22,25]. One commonality between these studies is the use of concatenation to produce feature vectors describing the ILs. Descriptors are computed for either the entire IL, or for the cation and anion separately before combination by concatenation into a vector. In the first case, it can be difficult

[*] Corresponding Author: Email smartin@sandia.gov; Phone 1(505)845-9644; Fax 1(505)845-8506.

to generalize prediction to ILs not in the original dataset, and in the second case, potential cross-correlations between cation and anion descriptors are ignored.

In this paper we consider QSPRs which incorporate descriptor cross-correlation between cations and anions in ILs. This approach treats ILs as pairs using product descriptors for QSPRs, a concept borrowed from the prediction of protein-protein, enzyme-metabolite, and drug-target interactions in bioinformatics[27,28]. We apply this method to screening for high conductivity/low viscosity ILs. We show that this approach yields qualitative improvements over a comparable non-product based method when extrapolating beyond the training set.

Materials and Methods

Data. A dataset was downloaded from the ILThermo database (http://ilthermo.boulder.nist.gov) on June 18th, 2014. The full dataset consisted of 8,508 measurements: 2,396 of conductivities; 5,672 of viscosities; and 440 of melting points. ILs with unusual composition (e.g. including Hg or Ge), with non-standard naming not easily parsed by computer, or with measurements not at standard pressure were removed, yielding 7,014 measurements: 1,980 conductivities; 4,624 viscosities; and 410 melting points. Measurements duplicated by different studies were averaged and retained only if the minimum and maximum measurements were within the average error (typically reported as a standard deviation) of the average measurement, i.e.

$$\bar{m} + \overline{m_{err}} < m_{max} \tag{1}$$

and

$$\bar{m} - \overline{m_{err}} > m_{min}, \tag{2}$$

where $\bar{m}$ is the average duplicated measurement, $\overline{m_{err}}$ is the average measurement error for the duplicated measurements reported by the different studies, m_{min} is the minimum measurement for the duplicated measurements, and m_{max} is the maximum duplicated measurement. The resulting dataset consisted of 6,569 measurements: 1,863 conductivity; 4,296 viscosity; and 410 melting point. These measurements were taken from 491 ILs made from 277 cations and 103 anions. Figure 1(a) shows the distribution of the different measurement types in the full dataset.

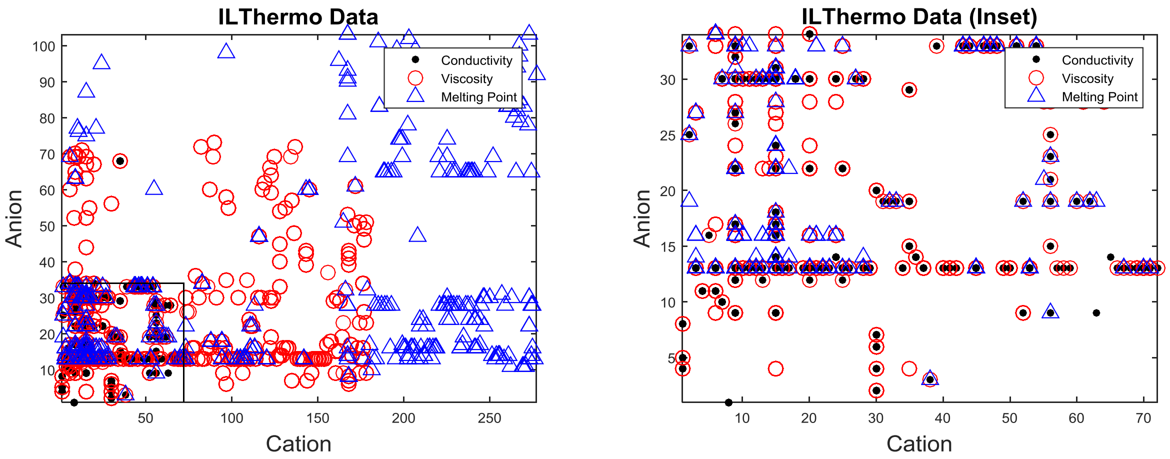


Figure 1. ILThermo Dataset. On the left (a), we show the distribution of measurement types in the full dataset downloaded from ILThermo on June 18th, 2014. The cations and anions are ordered so that the maximum measurement type intersection is shown by the inset on the lower left. On the right (b), the inset is shown. This subset of the ILThermo database is used in the QSPR analysis. It represents the maximal intersection of ILs with conductivity, viscosity, and melting point measurements.

As seen in Figure 1(a), there is a limited overlap between the three properties taken from the ILThermo database, where conductivity is the most limiting property. To maximize this overlap, we obtained our final dataset by restricting to ILs with conductivity measurements. Our final dataset consisted of 3,926 measurements: 1,853 conductivities; 2,584 viscosities; and 130 melting points. Conductivities were in the range 3×10^{-5} to 144.6 S/m; viscosities were in the range 1.059×10^{-5} to 364 Pa$*$s, and melting points were in the range 188 to 399.4 K. These measurements were taken from 165 ILs made from 72 cations and 34 anions. The distribution of the measurements is shown in Figure 1(b). This is the dataset used in the QSPR analysis. The measurements can be found Supplement 1, and the cation/anion structures can be found in Supplement 2.

Descriptor Matrices. We used PubChem (https://pubchem.ncbi.nlm.nih.gov/) and chemicalize.org (http://www.chemicalize.org) to obtain SMILES strings representing the structures obtained from the ILThermo database. We then used AMPAC 10 (http://www.semichem.com) and CODESSA III (http://www.semichem.com) to produce chemical descriptors for each cation and anion in our dataset. For the cations, we assumed a charge of +1, singlet bonds, and performed a Hessian minimum energy calculation using AMPAC. For the anions, we assumed a charge of -1. The results from AMPAC were input to CODESSA, which was used to compute various chemical descriptors (ranging from compositional and topological to quantum chemical). After removing constant valued descriptors, and descriptors which were identical, we were left with 320 cation descriptors and 222 anion descriptors. Each of these descriptors was mean-centered and scaled to have unit variance.

From the CODESSA descriptors, we formed two matrices describing the IL dataset. The first matrix was obtained by concatenating the cation and anion descriptors for each IL cation-anion pair. This operation resulted in a $3{,}926 \times 542$ concatenated matrix. In the case of temperature dependent properties (such as conductivity and viscosity), we added an extra column giving temperature. This matrix is the standard used for QSPRs in conductivity, viscosity and melting point predictions. It encodes information about IL chemical structure as well as temperature for use by the QSPRs.

The second matrix was obtained by using a tensor product of the cation descriptors with the anion descriptors for each IL pair. Briefly, if the vector $\boldsymbol{c}$ contains the cation descriptor values and the vector $\boldsymbol{a}$ contains anion descriptor values then

$$P = \boldsymbol{c}\boldsymbol{a}^T \tag{3}$$

gives a product matrix containing the product of every pair of cation-anion descriptor values in $\boldsymbol{c}$ and $\boldsymbol{a}$. By re-forming P as a vector for each IL cation-anion pair, we obtain a $3{,}926 \times 71{,}040$ product matrix. Again for temperature dependent quantities we added an extra column providing temperature.

The product matrix tracks every potential first-order interaction between the cation and anion descriptors in the IL dataset. It encodes more detailed IL chemical structure information than the standard matrix, as well as temperature for use by the QSPRs. It has been shown previously to improve prediction performance in the case of pairwise data, including protein-protein, enzyme-metabolite, and drug-target interactions[27,28].

SVM Kernels. One problem with using the product descriptor is the combinatorial growth of the pair-wise descriptors. In our case, for example, the 320 cation and 222 anion descriptors yield $320 \times 222 = 71{,}040$ product descriptors. Fortunately, this problem can be circumvented by implementing the method in the context of SVMs. A SVM takes as input a kernel function which specifies the similarity of two objects under consideration. For ILs, we are comparing two cation-anion pairs. Following the original work on protein-protein interactions[27], we assume that our first IL is denoted $(\boldsymbol{c}_1, \boldsymbol{a}_1)$ and our second IL is denoted $(\boldsymbol{c}_2, \boldsymbol{a}_2)$. Then we define the kernel product measuring the similarity of the two ILs as

$$k_p\big((\boldsymbol{c}_1, \boldsymbol{a}_1), (\boldsymbol{c}_2, \boldsymbol{a}_2)\big) = k(\boldsymbol{c}_1, \boldsymbol{c}_1)k(\boldsymbol{a}_1, \boldsymbol{a}_1), \tag{4}$$

where $k(\boldsymbol{c}_1, \boldsymbol{c}_1)$ and $k(\boldsymbol{a}_1, \boldsymbol{a}_1)$ are given by the standard dot products $\boldsymbol{c}_1\boldsymbol{c}_1^T$ and $\boldsymbol{a}_1\boldsymbol{a}_1^T$. This definition follows algebraically[27] from the use of the tensor product in Eq. (3). It should be noted that the concatenated descriptor can also be expressed using kernels by addition

$$k_c\big((\boldsymbol{c}_1, \boldsymbol{a}_1), (\boldsymbol{c}_2, \boldsymbol{a}_2)\big) = k(\boldsymbol{c}_1, \boldsymbol{c}_1) + k(\boldsymbol{a}_1, \boldsymbol{a}_1), \tag{5}$$

We can further use a Gaussian version of the kernels in (4) and (5). The Gaussian kernel version of (4) is given by

$$k_G(L_1, L_2) = \exp(-\gamma\left(k_p(L_1, L_1) - 2k_p(L_1, L_2) + k_p(L_2, L_2)\right)), \tag{6}$$

where $L_1 = (\boldsymbol{c}_1, \boldsymbol{a}_1)$ is the first IL and $L_2 = (\boldsymbol{c}_2, \boldsymbol{a}_2)$ is the second IL.

Finally, using SVMs also allows us to compare our results with the leading method for computing QSPRs for conductivity in ILs[21,22], which uses Least Squares SVMs, or LS-SVMs[29]. The LS-SVM implementation used here, and by Gharaghezi *et al.*[21,22], is available as LS-SVM Lab (http://www.esat.kuleuven.be/sista/lssvmlab/). This package includes automatic parameter tuning for regularization, Gaussian kernels, and scaling of the original descriptors. Details on the actual LS-SVM algorithm can be found elsewhere[29,30].

Performance Metrics. There are a host of metrics available for assessing the accuracy of QSPR predictions[31-33], and a number of these were applied by Gharagheizi *et al.* to the LS-SVM models for predicting conductivity in ILs[32,33]. Of these metrics, we use two very common statistics (R^2 and Q^2) along with a lesser known metric (confidence) designed to judge extrapolation quality of predictions farther from the training set.

The goodness-of-fit measure R^2 is computed according to the formula

$$R^2 = 1 - \frac{\sum_i (e_i - p_i)^2}{\sum_i (e_i - \bar{e})^2}, \qquad (7)$$

where e_i are measured (experimental) values, p_i are predicted values and $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$, given n measured values.

We also compute Q^2, which is a measure of the generalization ability of a QSPR model. Q^2 is typically computed using a leave-one-out strategy, although we use a 10-fold data split due to the relatively large size of our dataset. To compute ten-fold Q^2, we first divide the dataset into ten equal subsets. For a given subset, denoted the test set, we train a model on the remaining nine subsets of the dataset. Predictions are then made on the test set. It is important to note that the model is re-trained (and re-optimized) ten times altogether, and that each time the test set is unknown to the model. These calculations are repeated for each test set and Q^2 is computed using the test set predictions as

$$Q^2 = 1 - \frac{\sum_i (e_i - p_i)^2}{\sum_i (e_i - \bar{e})^2}. \qquad (8)$$

Finally, to assess the quality of our QSPR extrapolations, we use a confidence measure[31]. We consider extrapolations to be predictions where no experimental data is known. We extrapolate over all possible cation-anion pairs, with at least one of the pair (cation or anion) in the original dataset (i.e. having a measured value). The confidence metric is computed as

$$c_i = 1 - \frac{\min_{j \neq i} \|L_i - L_j\|}{\max_{j,k} \|L_k - L_j\|}, \qquad (9)$$

where L_i is the IL under consideration, L_j ranges over ILs with measured values only, and L_k ranges over all potential ILs (both extrapolated and in the dataset). In all cases, L_* refers to the IL descriptor vectors. The metric c_i tells us how close structure L_i is a structure L_j with a measured value. Values of c_i closer to 1 indicate higher confidence predictions.

Results and Discussion

Cross-Validation. Our first results are obtained on the ILThermo dataset using cross-validation to assess our models. Thus the data consists of ILs with previously obtained experimental measurements. We apply the LS-SVM model to the ILThermo data using both the concatenated and product descriptors to predict conductivity, viscosity, and melting point. The models are compared using cross-validation, and all predictions are made on test sets.

Our first effort was towards predicting conductivity in ILs to replicate the work of Gharagheizi *et al.*[21] using the concatenated descriptors. For this effort, we used both a standard linear kernel and a Gaussian kernel (Gharagheizi *et al.* used only a Gaussian kernel). The LS-SVM Lab package chose $\gamma = 0.2996$ for the Gaussian kernel. For the linear kernel, we obtained a goodness-of-fit measure of $R^2 = 0.9207$ and ten-fold generalization measure of $Q^2 = 0.9036$. For the Gaussian kernel, we obtained $R^2 = 0.9975$ and $Q^2 = 0.9842$. The results are shown in Figure 2. Fit statistics are summarized in Table 1.

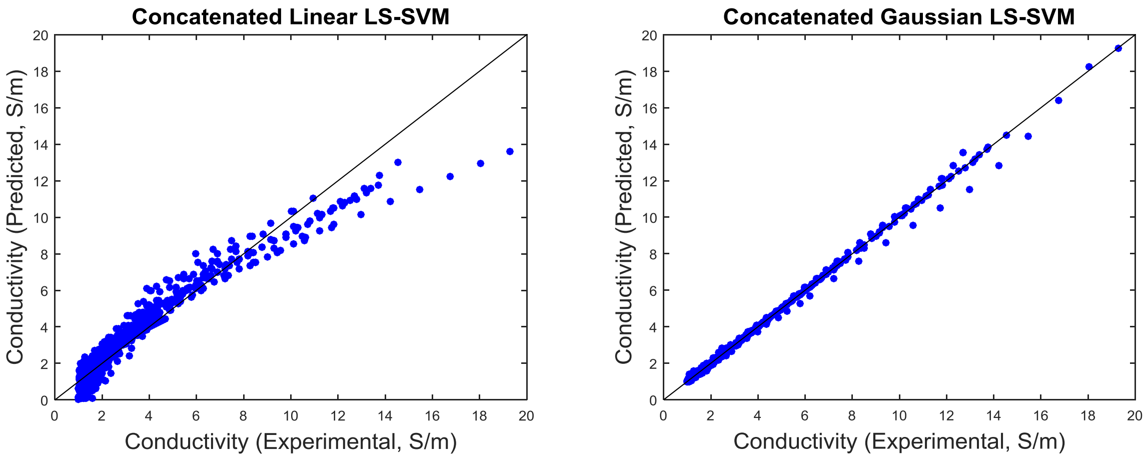


Figure 2. Conductivity Predictions Using Concatenated Descriptors. On the left (a), we compare the experimentally measured conductivity values with the predictions from the linear LS-SVM model using concatenated descriptor vectors. On the right (b), we make the same comparison using the non-linear Gaussian LS-SVM model. This is the model used by Gharagheizi *et al.* for IL conductivity predictions (*21*). Predicted measurements shown are from the R^2 measurement (i.e. training on the full dataset).

We repeated the same computations using the product descriptors, as shown in Figure 3. The LS-SVM Lab package chose $\gamma = 0.5039$. For the linear kernel using the product descriptor we obtained a goodness-of-fit measure of $R^2 = 0.9313$ and ten-fold $Q^2 = 0.9164$. For the Gaussian kernel we obtained $R^2 = 0.9975$ and $Q^2 = 0.9909$. The statistics for the product descriptors offer improvement over the standard concatenated descriptor for both the linear and Gaussian LS-SVM models. Statistics are summarized in Table 1.

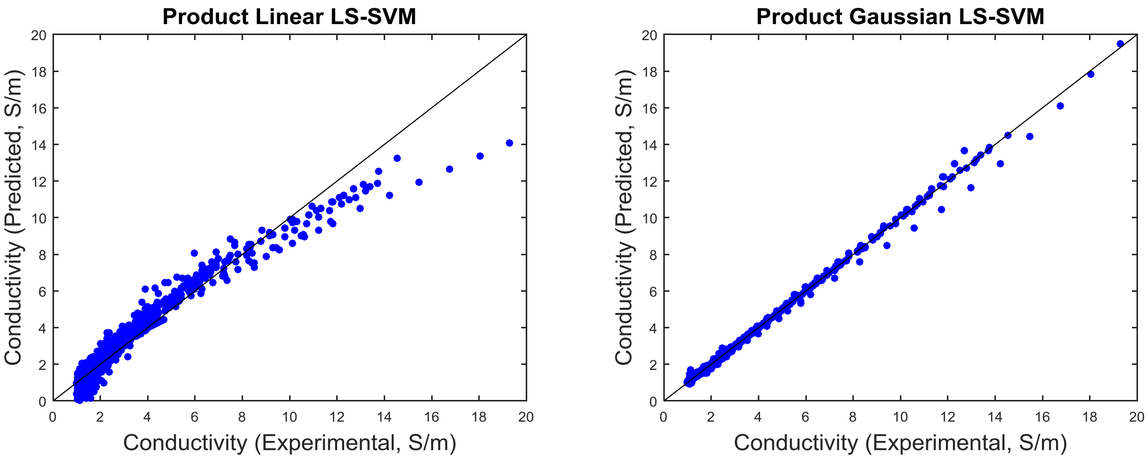


Figure 3. Conductivity Predictions Using Product Descriptors. On the left (a), we compare the experimentally measured conductivity values with the predicted values using a linear LS-SVM model with the product kernel. On the right (b), we make the same comparison using a Gaussian LS-SVM with the product descriptors. Predicted measurements shown are from the R^2 measurement (i.e. trained on the full dataset).

We repeated our analysis for the ILThermo viscosity data. We restricted measurements between 10^{-3} Pa $*$ s and 2 Pa $*$ s and performed our analysis using $\log_{10}$ transformed temperature and viscosity values. We again compared the standard concatenated descriptors with product descriptors for linear and Gaussian LS-SVM models. For the concatenated descriptors, the LS-SVM Lab package chose $\gamma = 0.0248$. For the linear LS-SVM model, we obtained goodness-of-fit $R^2 = 0.9170$ and ten-fold generalization $Q^2 = 0.9020$. For the Gaussian LS-SVM model, we obtained $R^2 = 0.9870$ and $Q^2 = 0.9726$. The results are shown in Figure 4, and statistics are summarized in Table 1.

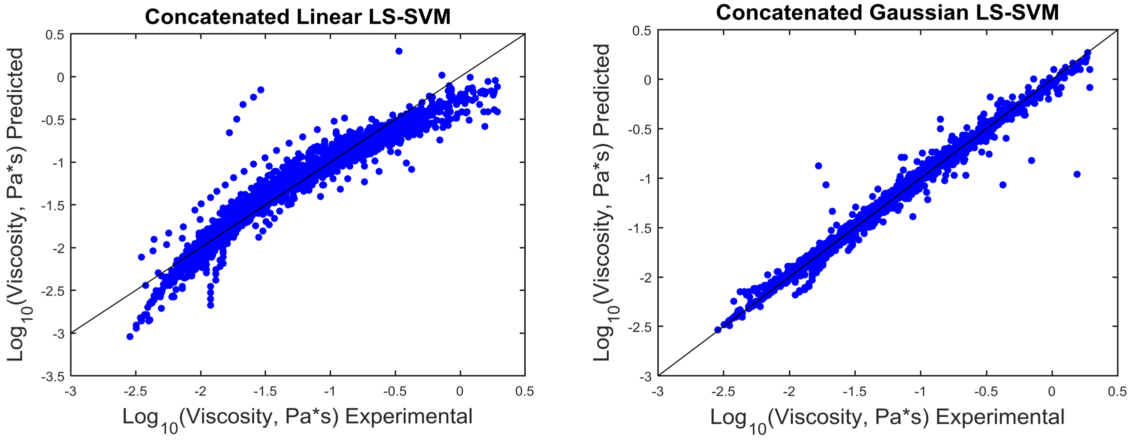


Figure 4. Viscosity Predictions Using Concatenated Descriptor. On the left (a), we compare the experimentally measured viscosity values with the predicted values for the linear LS-SVM model with the standard concatenated descriptors. On the right (b), we compare the viscosity values for the Gaussian LS-SVM model with the concatenated descriptors. Predicted measurements shown are from the R^2 measurement (i.e. trained on the full dataset).

Repeating the computations using the product descriptors, the LS-SVM Lab package chose $\gamma = 0.0212$. For the linear kernel using the product descriptor we obtained a goodness-of-fit measure of $R^2 = 0.9373$ and ten-fold $Q^2 = 0.9246$. For the Gaussian kernel we obtained $R^2 = 0.9866$ and $Q^2 = 0.9700$. The statistics for the product descriptors were an improvement for the linear model, but almost identical for the Gaussian model. The results are quite similar to those shown in Figure 4, so another figure is not provided, but statistics are given in Table 1.

Lastly, we analyzed the ILThermo melting point data. We compared the same four LS-SVM models used to predict conductivity and viscosity. For the concatenated descriptor, the LS-SVM Lab package chose $\gamma = 0.0006$. For the linear LS-SVM model we obtained $R^2 = 0.3869$ and $Q^2 = 0.0977$, and for the Gaussian LS-SVM model we obtained $R^2 = 0.4444$ and $Q^2 = 0.1336$. The results are shown in Figure 5, and statistics are provided in Table 1.

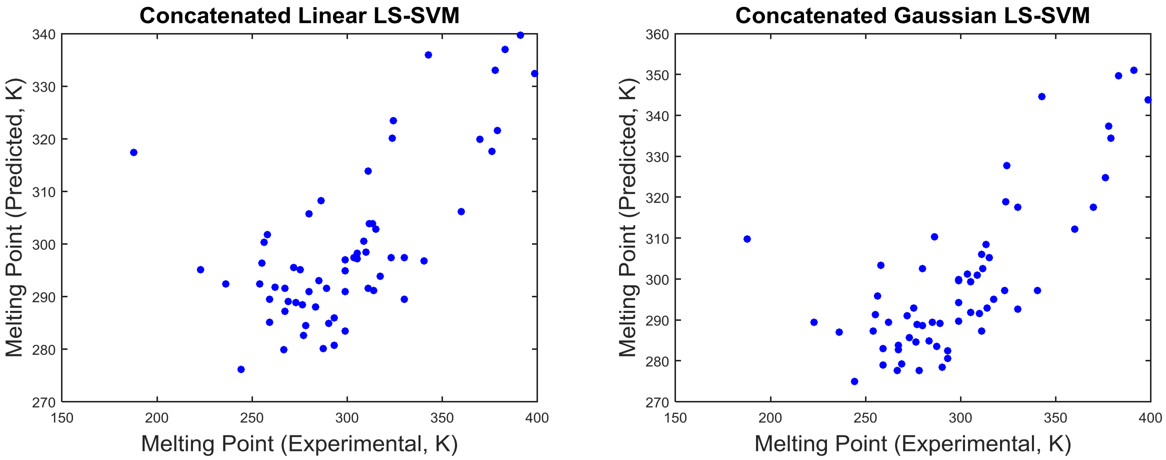


Figure 5. Melting Point Predictions Using Concatenated Descriptor. On the left (a), we compare the measured melting point values with the predicted values using a linear LS-SVM with a concatenated descriptor. On the right (b), we compare the same values using a Gaussian LS-SVM. Predicted measurements shown are from the R^2 measurement (i.e. trained on the full dataset).

Repeating the computations using the product descriptors, the LS-SVM Lab package chose $\gamma = 8.9\ \times 10^{-6}$. For the linear kernel using the product descriptor we obtained $R^2 = 0.3737$ and $Q^2 = 0.0157$. For the Gaussian kernel we obtained $R^2 = 0.6191$ and $Q^2 = 0.1675$. Results are similar to those shown in Figure 5 so another figure is not provided, but statistics are shown in Table 1.

From our analysis, it is evident that all four models fail to predict the melting point data. Although some success has been achieved predicting melting point with different methods and using more focused IL datasets[23,24], a study with a wider variety of methods and a larger dataset showed more modest success[25]. Our results agree with this later study, and given the heterogeneity of the ILThermo dataset, it is perhaps not surprising that our methods failed to predict melting point.

	R^2	Q^2	γ (Gaussian Models Only)
Conductivity			
Linear Concatenated	0.9207	0.9036	
Gaussian Concatenated	0.9975	0.9842	0.2986
Linear Product	0.9313	0.9164	
Gaussian Product	0.9975	0.9909	0.5039
Viscosity			
Linear Concatenated	0.9170	0.9020	
Gaussian Concatenated	0.9870	0.9726	0.0248
Linear Product	0.9373	0.9246	
Gaussian Product	0.9866	0.9700	0.0212
Melting Point			
Linear Concatenated	0.3869	0.0977	
Product Concatenated	0.4444	0.1336	0.0006
Linear Product	0.3737	0.0157	
Gaussian Product	0.6191	0.1675	8.9×10^{-6}

Table 1. QSAR Model Fit Statistics. Here we collect the QSAR model statistics for the various predictions made in this paper. In the first column, we groups the models according to property predicted and model type. In the second column, we give R^2 values; in the third column, we give Q^2 values; and in the last column we give the LS-SVM γ values for Gaussian kernels.

Extrapolation. Next, we consider the extrapolation ability of the four different LS-SVM models. In this situation, we consider ILs where no experimental data is available, but at least one of the cation or anion in the IL is present in the ILThermo dataset, and has a measured conductivity value. Our goal is to determine which model provides the most useful extrapolations outside of the original training data (ILThermo data with experimental measurements). For conductivity, our extrapolations are shown in Figure 6 for the concatenated descriptor and Figure 7 for the product descriptor.

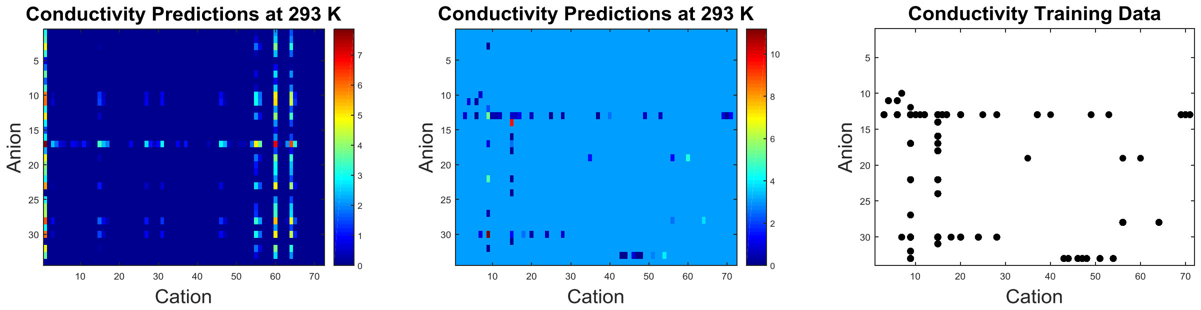


Figure 6. Extrapolated Conductivity Predictions Using Concatenated Descriptor. On the left (a), we show conductivity predictions at 293 K using the concatenated linear LS-SVM model extrapolated to the entire IL dataset. In the middle (b), we show predictions using the concatenated Gaussian LS-SVM model. On the right (c), we show the locations of ILs in the original dataset distributed throughout the extrapolated space. Comparing (b) and (c), we see that the non-linear LS-SVM model is only accurate on or near the training set.

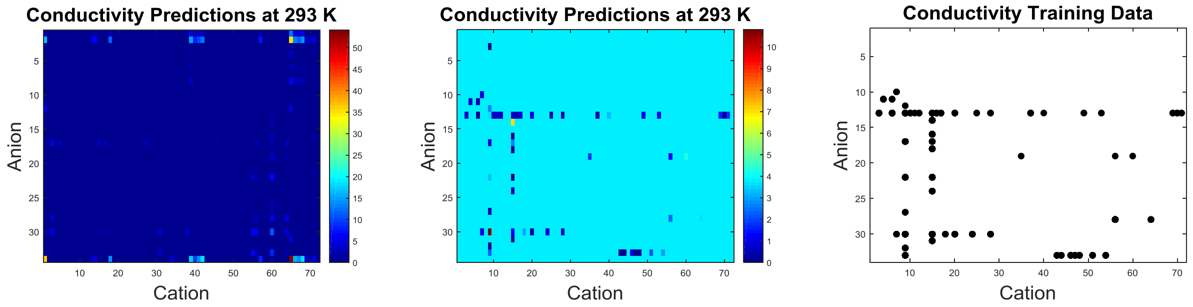


Figure 7. Extrapolated Conductivity Predictions Using Product Descriptor. On the left (a), we show conductivity predictions at 293 K using the product linear LS-SVM model. In the middle (b), we show predictions using the product Gaussian LS-SVM model. On the right (c), we show the locations of the ILs in the original dataset. Comparing (b) and (c), we see that the non-linear LS-SVM model is only accurate very near the training set.

Given the performance of the four models according to the cross-validation analysis, which model is best suited for extrapolation of conductivity predictions to all possible ILs in the dataset? Using only the R^2 and Q^2 statistics, we might conclude that the non-linear Gaussian LS-SVM models are superior to the linear LS-SVM models, and that the product descriptor offers a slight improvement over the standard concatenated descriptor. However, considering the qualitative evidence in Figures 6 and 7, we see that the non-linear models are accurate only on or very near the training set, i.e. they are over-fitting the training data, and are generalizing poorly. In fact, their predictions are constant for any IL even a little different from the ILs in the training set. Therefore, we can eliminate the non-linear models for the purpose of extrapolating our predictions.

Comparing the two linear models, we note that the model based on the concatenated descriptor exhibits striping behavior, as seen in Figure 6(a). For a vertical stripe, the model is identifying certain cations as high performing, regardless of their pairings with different anions. For a horizontal stripe, the situation is reversed. While this same striping is evident in the product descriptor predictions in Figure 7(a), it is not as pronounced. In order for an IL to be highly conductive, it must have both a high performing cation and a high performing anion. Although we cannot know which predictions are better, the behavior of the product descriptor seems to fit better with chemical intuition. Thus, we've chosen the linear product LS-SVM model as the basis for our screening predictions (to follow).

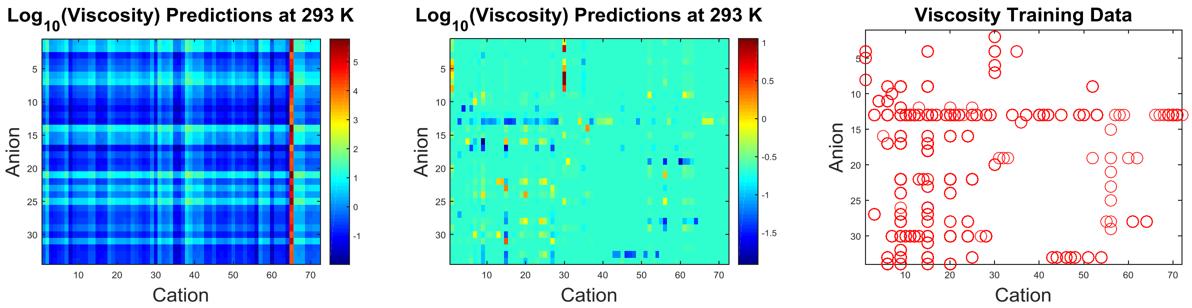


Figure 8. Extrapolated Viscosity Predictions Using Concatenated Descriptor. On the left (a), we show viscosity predictions at 293 K using the concatenated linear LS-SVM model extrapolated to the entire IL dataset. In the middle (b), we show predictions using the concatenated Gaussian LS-SVM model. On the right (c), we show the locations of ILs in the original dataset distributed throughout the extrapolated space. Comparing (b) and (c), we again see that the non-linear LS-SVM model is only accurate on or near the training set.

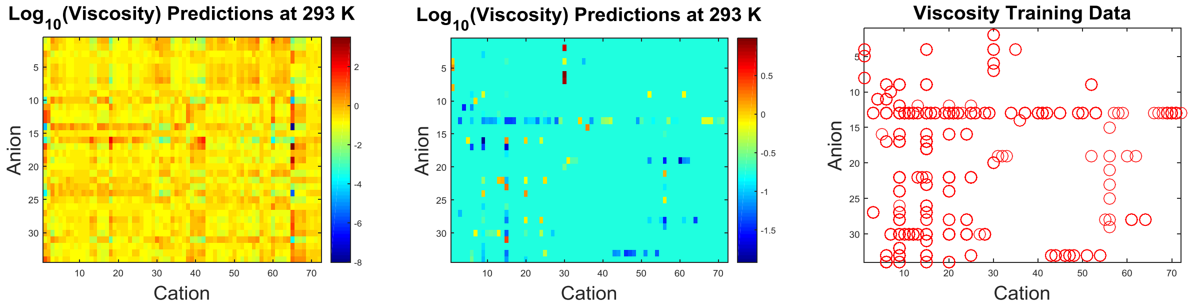


Figure 9. Extrapolated Viscosity Predictions Using Product Descriptor. On the left (a), we show viscosity predictions at 293 K using the product linear LS-SVM model. In the middle (b), we show predictions using the product Gaussian LS-SVM model. On the right (c), we show the locations of the ILs in the original dataset. Again comparing (b) and (c), we see that the non-linear LS-SVM model is only accurate very near the training set.

For viscosity, we again made extrapolation on every possible cation-anion pair in the database. These results are shown in Figure 8 for the concatenated descriptor and Figure 9 for the product descriptor. Given the cross-validation performance of the four LS-SVM models for predicting viscosity, we can make the same conclusions that we made in the case of conductivity. Specifically, the non-linear Gaussian models are again accurate only very near or on the training set, making constant predictions outside otherwise, and that the linear concatenated LS-SVM model exhibits the same striping behavior. We again conclude that the linear LS-SVM model with the product descriptor is best suited for extrapolating viscosity predictions to the entire set of cation-anion pairs.

Screening. Finally, we screen for high-conductivity/low-viscosity ILs by using our extrapolated conductivity and viscosity predictions on the all possible cation-anion pairs with at least one cation or anion in the ILThermo dataset. The melting point predictions are not considered due to the questionable accuracy of the melting point QSPR. Sorting by conductivity and thresholding for predictions above 4 S/m we obtained the list in Table 2. A list thresholded above 1 S/m is given in Supplement 3.

Rank	Cation (ID)	Anion (ID)	Cond., S/m	Conf.	Log$_{10}$(Vis., Pa*s)	Conf.	In Concat. List
1	tributyl(hexadecyl)phosphonium (65)	tris(pentafluoroethyl)trifluorophosphate (34)	54.24	0	0.62	0.34	
2	(3-aminopropyl)tributylphosphonium (1)	tris(pentafluoroethyl)trifluorophosphate (34)	36.09	0.12	0.98	0.62	
3	tributyl(hexadecyl)phosphonium (65)	.beta.-alaninate (2)	35.44	0.04	-1.57	0.6	
4	N,N,N-triethyltetradecan-1-aminium (42)	tris(pentafluoroethyl)trifluorophosphate (34)	18.28	0.79	-0.33	0.78	
5	tributyloctylammonium (68)	tris(pentafluoroethyl)trifluorophosphate (34)	18.12	0.83	0.4	0.82	
7	(3-aminopropyl)tributylphosphonium (1)	.beta.-alaninate (2)	16.21	0.14	-0.94	0.88	
8	tributylheptylammonium (66)	tris(pentafluoroethyl)trifluorophosphate (34)	15.92	0.84	0.37	0.83	
9	N,N,N-tributyloctan-1-aminium (39)	.beta.-alaninate (2)	15.1	0.84	-1.3	0.87	
10	tributyloctylammonium (68)	.beta.-alaninate (2)	14.36	0.84	-1.4	0.87	
11	1-hexadecyl-3-methylimidazolium (18)	tris(pentafluoroethyl)trifluorophosphate (34)	14.32	0.8	-0.49	0.78	
12	tributylhexylammonium (67)	tris(pentafluoroethyl)trifluorophosphate (34)	14.24	0.86	0.32	0.84	
13	N,N,N-triethyldodecan-1-aminium (41)	tris(pentafluoroethyl)trifluorophosphate (34)	14.02	0.84	-0.28	0.83	
14	tributylheptylammonium (66)	.beta.-alaninate (2)	13.33	0.85	-1.26	0.88	
15	N,N,N-triethyltetradecan-1-aminium (42)	.beta.-alaninate (2)	12.35	0.8	-1.61	0.84	
16	tributyl(hexadecyl)phosphonium (65)	(S)-2-amino-3-carboxypropanoate (1)	12.22	0.77	-2.49	0.58	
17	tributylhexylammonium (67)	.beta.-alaninate (2)	11.59	0.86	-1.24	0.88	
18	1-hexadecyl-3-methylimidazolium (18)	.beta.-alaninate (2)	11.41	0.81	-1.89	0.84	
19	methanaminium (60)	tetrafluoroborate (30)	10.25	0.93	-1.52	0.72	***
20	N,N,N-triethyldodecan-1-aminium (41)	.beta.-alaninate (2)	10.07	0.85	-1.39	0.87	
21	N,N,N-triethyl-1-decanaminium (40)	tris(pentafluoroethyl)trifluorophosphate (34)	9.32	0.88	-0.44	0.87	
22	methanaminium (60)	trifluoromethanesulfonate (33)	8.44	0.95	0.04	0.81	***
23	N,N,N-triethyl-1-decanaminium (40)	.beta.-alaninate (2)	7.92	0.89	-1.06	0.89	
24	ethylheptyl-di-(1-methylethyl)ammonium (57)	tris(pentafluoroethyl)trifluorophosphate (34)	7.82	0.9	0	0.89	
25	ethylheptyl-di-(1-methylethyl)ammonium (57)	.beta.-alaninate (2)	7.28	0.91	-1.05	0.91	
26	methanaminium (60)	hexafluorophosphate (22)	6.29	0.92	-1.6	0.74	***
27	1-dodecyl-3-methylimidazolium (14)	.beta.-alaninate (2)	6.21	0.9	-1.29	0.91	
28	N,N,N-tributyloctan-1-aminium (39)	(S)-2-amino-3-carboxypropanoate (1)	5.84	0.95	-1.78	0.88	
29	ethanolammonium (55)	trifluoromethanesulfonate (33)	5.82	0.96	-1.34	0.88	
30	ethanolammonium (55)	tetrafluoroborate (30)	5.71	0.96	-1.05	0.86	***

31	tributyl(hexadecyl)phosphonium (65)	L-valinate (8)	5.56	0.84	-0.2	0.72	
32	(3-aminopropyl)tributylphosphonium (1)	bis(perfluoroethylsulfonyl)imide (12)	5.49	0.84	1.2	0.7	
33	triethyloctylammonium (71)	tris(pentafluoroethyl)trifluorophosphate (34)	5.45	0.91	-0.35	0.91	
34	tributyloctylammonium (68)	(S)-2-amino-3-carboxypropanoate (1)	5.39	0.95	-1.82	0.88	
35	triethyloctylammonium (71)	.beta.-alaninate (2)	4.99	0.92	-0.92	0.92	
36	tributylheptylammonium (66)	(S)-2-amino-3-carboxypropanoate (1)	4.91	0.95	-1.65	0.88	
37	ethylammonium (56)	tetrafluoroborate (30)	4.9	0.97	-1.45	0.9	***
38	methanaminium (60)	nitrate (28)	4.74	0.95	-1.61	0.8	***
39	pyrrolidinium (64)	nitrate (28)	4.7	1	-1.27	1	***
40	1,3-dimethylimidazolium (3)	dicyanamide (17)	4.69	0.98	-1.65	0.93	***
41	1,3-dimethylimidazolium (3)	tetrafluoroborate (30)	4.63	0.99	-1.94	0.94	
42	1-dodecyl-3-methylimidazolium (14)	tris(pentafluoroethyl)trifluorophosphate (34)	4.16	0.9	-0.51	0.89	
43	tributylhexylammonium (67)	(S)-2-amino-3-carboxypropanoate (1)	4.12	0.96	-1.57	0.89	
44	tributyl(hexadecyl)phosphonium (65)	bis(perfluoroethylsulfonyl)imide (12)	4.07	0.77	0.9	0.49	

Table 2. List of Predicted High-Conductivity/Low-Viscosity ILs. Here we give a list of the most promising high-conductivity/low-viscosity ILs according to the extrapolated predictions of our linear product LS-SVM model to all possible cation-anion pairs, with at least one of the cation or anion from the ILThermo dataset. The first column contains the rank. The second and third columns give the cation-anion pair. The cations and anion are named according to the ILThermo database, and tagged with an ID (in parenthesis) which corresponds to the structures given in Supplement 2. The fourth and fifth columns give the predicted conductivity and associated confidence. The sixth and seventh columns give the predicted viscosity and associated confidence. In the last column, we provide an indicator (***) if the cation-anion pair was also predicted to have high-conductivity/low-viscosity by the standard concatenated descriptor with the linear LS-SVM model.

Conclusions

The combinatorial nature of IL chemistry is both a blessing and a curse. While we can in principle engineer a custom IL for a particular application, the number of potential ILs can easily overwhelm an experimentalist looking for that custom IL. Ideally, a computational screening approach could help guide experimental work in ILs, making the initial design decisions more manageable. In this paper we have proposed such an approach based on the use of QSPRs.

Specifically, we have proposed a QSPR approach that exploits the cation-anion pairwise nature of ILs. We have benchmarked our approach on the problem of predicting conductivity and viscosity for ILs, and have shown that our method is competitive with the best known QSPR method for the task of predicting conductivity in ILs[21,22]. Finally, we have produced a list of potentially high-conductivity/low-viscosity ILs using our model.

Based on the data summarized in Table 2, and considering logistical and safety related aspects of our chemical laboratory facility, we have chosen to synthesize tributylheptylammonium β-alaninate (number 14 in Table 2, shown in Figure 10). This compound is prepared by a metathesis reaction of tributylheptylammonium bromide and sodium β-alaninate in polar aprotic solvent mixtures. We are currently optimizing the synthesis in order to achieve a high purity material that will be used as both solvent and supporting electrolyte in a redox flow battery. If successful, this is expected to lead to lower viscosity and a wider electrochemical window than our current system based on imidazolium triflimide salts. This work is presently underway and the results will be reported in due course.

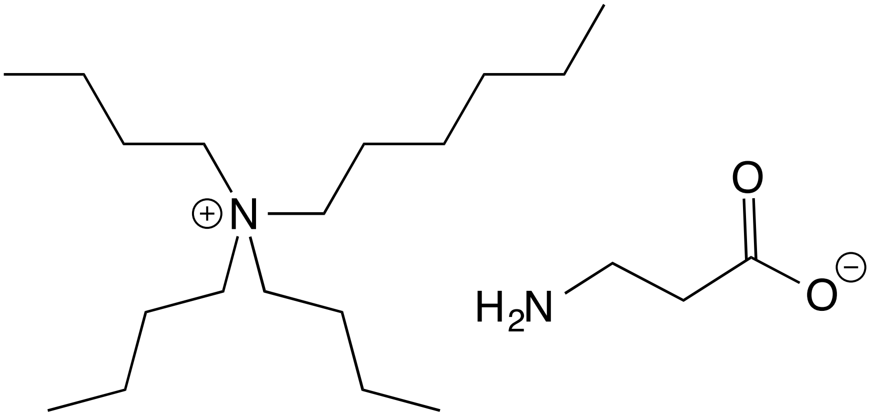


Figure 10. Chemical structure of tributylheptylammonium β-alaninate

In the future, we plan to couple our approach with more detail computational methods such as Quantum Density Function Theory and Molecular Dynamics to provide additional confidence in the most promising QSPR predictions.

Acknowledgements

This work was partially funded by the U.S. Department of Energy, Office of Electricity Delivery and Energy Reliability (Dr. Imre Gyuk, Energy Storage Program). Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Supporting Information for Publication

- Supplement 1 contains the ILThermo dataset used in the analysis.
- Supplement 2 contains the IL names and chemical structures.
- Supplement 3 contains the complete list of best extrapolated ILs (longer version of Table 2).

References

1. Welton, T. Room-Temperature Ionic Liquids. Solvents for Synthesis and Catalysis. *Chem. Rev.* **1999,** *99*, 2071-2084.

2. Earle, M. J.; Seddon, K. R. Ionic Liquids. Green Solvents for the Future. *Pure App. Chem.* **2000,** *72*, 1391-1398.

3. Hubbard, C. D.; Illner, P.; Eldik, R. v. Understanding Chemical Reaction Mechanisms in Ionic Liquids: Successes and Challenges. *Chem. Soc. Rev.* **2011,** *40*, 272-290.

4. Sawant, A. D.; Raut, D. G.; Darvatkar, N. B.; Salunkhe, M. M. Recent Developments of Task-Specific Ionic Liquids in Organic Synthesis. *Green Chem. Lett. Rev.* **2011,** *4*, 41-54.

5. Seddon, K. R. Ionic Liquids for Clean Technology. *J. Chem. Technol. Biotechnol.* **1999,** *68*, 351-356.

6. Sheldon, R. Catalytic Reactions in Ionic Liquids. *Chem. Commun.* **2001,** No. 23, 2399-2407.

7. Olivier-Bourbigou, H.; Magna, L. Ionic Liquids: Perspectives for Organic and Catalytic Reactions. *J. Mol. Catal. A: Chem.* **2002,** *182-183,* 419-437.

8. Tsuda, T.; Hussey, L. C. Electrochemical Applications of Room-Temperature Ionic Liquids. *Interface Sci.* **2007,** *16*, 42-49.

9. Hapiot, P.; Lagrost, C. Electrochemical Reactivity in Room-Temperature Ionic Liquids. *Chem. Rev.* **2008,** *108*, 2238-2264.

10. Ma, Z.; Yu, J.; Dai, S. Preparation of Inorganic Materials Using Ionic Liquids. *Adv. Mater.* **2010,** *22,* 261-284.

11. Das, R. N.; Roy, K. Advances in QSPR/QSTR Models of Ionic Liquids for the Design of Greener Solvents of the Future. *Mol. Diversity* **2013,** *17*, 151-196.

12. Coutinho, J. A. P.; Carvalho, P. J.; Oliveira, N. M. C. Predictive Methods for the Estimation of Thermophysical Properties of Ionic Liquids. *RSC Adv.* **2012,** *2,* 7322-7346.

13. 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* **2010,** *115*, 93-98.

14. Han, C.; Yu, G.; Wen, L.; Zhao, D.; Asumana, C.; Chen, X. Data and QSPR Study for Viscosity of Imidazolium-based Ionic Liquids. *Fluid Phase Equilib.* **2011,** *300,* 95-104.

15. Yu, G.; Zhao, D.; Wen, L.; Yang, S.; Chen, X. Viscosity of Ionic Liquids: Database, Observation, and Quantitative Structure-Property Relationship Analysis. *Am. Inst. Chem. Eng. J.* **2012,** *58*, 2885-2899.

16. Mirkhani, S. A.; Gharagheizi, F. Predictive Quantitative Structure-Property Relationship Model for the Estimation of Ionic Liquid Viscosity. *Ind. Eng. Chem. Res.* **2012,** *51,* 2470-2477.

17. Wang, X.; Chi, Y.; Mu, T. A Review on the Transport Properties of Ionic Liquids. *J. Mol. Liq.* **2014,** *193,* 262-266.

18. Matsuda, H.; Yamamoto, H.; Kurihara, K.; Tochigi, K. Computer-Aided Reverse Design for Ionic Liquids by QSPR using Descriptors of Group Contribution Type for Ionic Conductivites and Viscosities. *Fluid Phase Equilib.* **2007,** *261,* 434-443.

19. Tochigi, K.; Yamamoto, H. Estimation of Ionic Conducitivity and Viscosity of Ionic Liquids Using a QSPR Model. *J. Phys. Chem. C* **2007,** *111,* 15989-15994.

20. Bini, R.; Malvaldi, M.; Pitner, W.; Chiappe, C. QSPR Correlation for Conductivities and Viscosities of Low-Temperature Melting Ionic Liquids. *J. Phys. Org. Chem.* **2008,** *21,* 622-629.

21. Gharagheizi, F.; Sattari, M.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D.; Richon, D. A "Non-Linear" Quantitative Structure-Property Relationship for the Prediction of Electric Conducitivity of Ionic Liquids. *Chem. Eng. Sci.* **2013,** *101,* 478-485.

22. Gharagheizi, F.; Ilani-Kashkouli, P.; Sattari, M.; Mohammadi, A.; Ramjugernath, D.; Richon, D. Development of a LSSVM-GC Model for Estimating the Electrical Conductivity of Ionic Liquids. *Chem. Eng. Res. Des.* **2014,** *92*, 66-79.

23. Katrizky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. QSPR Correlation of the Melting Point for Pyridinium Bromides, Potential Ionic Liquids. *J. Chem. Inf. Model.* **2002,** *42*, 71-74.

24. Trohalaki, S.; Pachter, R.; Drake, G. W.; Hawkins, T. Quantitative Structure-Property Relationships for Melting Points and Densities of Ionic Liquids. *Energy Fuels* **2005,** *19,* 279-284.

25. Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solovev, V. P. Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J.*

Chem. Inf. Model. **2007,** *47,* 1111-1122.

26. Mohammad, F. H.; Izadian, P. In Silico Prediction of Melting Points of Ionic Liquids by Using Multilayer Perceptron Neural Networks. *J. Theor. and Comput. Chem.* **2012,** *11*, 127-141.

27. Martin, S.; Roe, D.; Faulon, J.-L. Predicting Protein-Protein Interactions Using Signature Products. *Bioinf.* **2005,** *21*, 218-226.

28. Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome Scale Enzyme-Metabolite and Drug-Target Interaction Predictions Using the Signature Molecular Descriptor. *Bioinf.* **2008,** *24*, 225-233.

29. Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines;* World Scientific: Singapore, 2002.

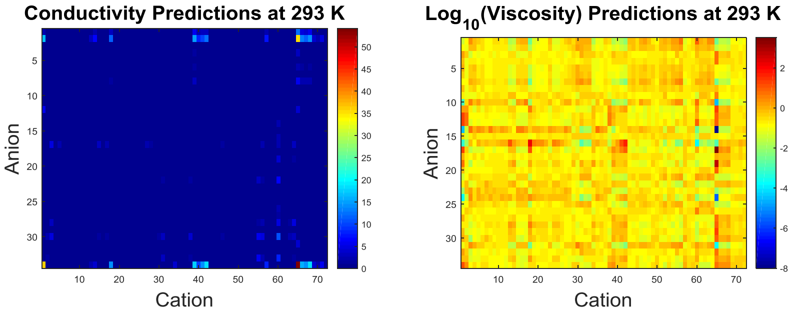
30. Suykens, J. A. K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. *Neural Proc. Lett.* **1999,** *9,* 293-300.

31. Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1912-1928.

32. Roy, K.; Indrani, M.; Kar, S.; Ojha, P. K.; Das, R. N.; Kabir, H. Comparative Studies on Some Metrics for External Validation of QSPR Models. *J. Chem. Inf. Model.* **2011,** *52,* 396-408.

33. Roy, K.; Chakraborty, P.; Indrani, M.; Ojha, P. K.; Kar, S.; Das, R. N. Some Case Studies on Application of "rm2" Metrics for Judging Quality of Quantitative Structure-Activity Relationship Predictions: Emphasis on Scaling of Response Data. *J. Comput. Chem.* **2013,** *34,* 1071-1082.

Table of Contents Graphic


Conductivity Predictions at 293 K
Anion
Cation
Log_{10}(Viscosity) Predictions at 293 K
Anion
Cation