Post-print version of the paper by Venkatraman et al. in J Mol. Liq, 264 (2018) 563-570: http://dx.doi.org/10.1016/j.molliq.2018.05.067

In Silico Prediction and Experimental Verification of Ionic Liquid Refractive Indices

Vishwesh Venkatraman[a,*], Jaganathan Joshua Raj[b], Sigvart Evjen[a], Kallidanthiyil Chellappan Lethesh[a], Anne Fiksdahl[a]

[a] *Department of Chemistry, Norwegian University of Science and Technology, 7491, Trondheim, Norway*
[b] *Centre of Research in Ionic Liquids (CORIL), Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 32610, Perak, Malaysia*

Abstract

Ionic liquids (ILs) have seen increasing use as environmentally friendly solvents in a wide array of applications from energy to pharmaceuticals. Among the many properties of interest, the refractive index, is of considerable importance since several related properties can be estimated once the refractive index of a material is known. Furthermore, high refractive index ILs are also used as reference solutions to determine properties of optical materials. However, with a large collection of cation-anion combinations to choose from, the task of finding suitable ionic liquids is far from trivial. In this article, machine learning models have been used to estimate the temperature-dependent refractive index over 450 diverse ILs using cheap to compute semi-empirically derived structure descriptors. In addition to using independent test sets for evaluating the predictive ability of the models, the efficacy of the models was further evaluated using 14 new ionic liquids that were synthesized. Overall, ensemble decision tree-based approaches gave the best results with mean absolute errors < 0.01 and squared correlations > 0.85 across both calibration and test data.

1. Introduction

In recent years, ionic liquids (ILs) have garnered a lot of attention with numerous application areas such as energy storage[1], corrosion inhibitors[2], food and bioproducts[3], recovery of rare earth elements[4], pharmaceuticals[5, 6] and CO2 capture[7]. Furthermore, given their utility as green solvents (non-volatile, non-flammable and recyclable) and the fact that the cations and anions constituting them can be tailored to application requirements, they have emerged as promising alternatives to traditional molecular solvents[8].

*Corresponding author.
Email address: vishwesh.venkatraman@ntnu.no (Vishwesh Venkatraman)

The number of potential ILs is estimated to be around 10^{18}[9], of which only a small fraction have been realized. With ILs being seen as designer solvents, the task of screening such a large collection is indeed formidable. Computational approaches based on density functional theory[10–12] and molecular dynamics[13, 14] have generally been used to provide a mechanistic understanding of the working of ILs but are restricted by the time complexity of the modelling. Faster alternatives have relied on quantitative structure property relationship (QSPR) models wherein physicochemical descriptors derived from the molecular structure are correlated with the property of interest using chemometric and machine learning tools. These methods have been applied to the prediction of a number of IL properties such as melting points[15], thermal decomposition temperatures[16], gas solubility[17], viscosity[18] etc.

Among the many properties investigated, the refractive index, an optical property has received recent interest with applications in the quality control and characterization of ionic liquids[19], and immersion fluids in the optical microscopy studies of minerals[20, 21]. With a view to understanding how the structure of the cation-anion pair influences the refractive index of the IL, several studies have used quantum chemistry, chemometrics or a combination of both to establish predictive models. Since the refractive index is related to the molar polarisability, Bica et al[22] deconstructed the polarisabilities and molar volumes into their individual atomic contributions that are then used to predict the refractive index. Group contribution[23–25] (GC) methods have been quite effective with coefficients of determination ranging between 0.95-0.99 for data sets containing 200-2150 experimental data points. Other efforts have made use of non-linear approaches such as genetic function approximation[26] and artificial neural networks[27–29]. While GC methods make use of group or atom properties, other approaches rely on a range of descriptors that include topological indices, connectivity indices, atom-centered fragments and 3D conformational descriptors. In some studies, quantum chemistry based descriptors derived from the surface-charge distribution such as σ-profiles[29, 30] have also been used. However, the computational cost associated with the descriptor calculation is quite high and may limit large scale application.

In this article, we investigate the utility of descriptors derived from semi-empirical quantum chemistry methods to model temperature-dependent refractive indices of a large and diverse set of ionic liquids. A number of machine learning algorithms for predicting ionic liquid refractive indices have been evaluated using independent calibration and test sets. As further validation of the predictive ability of the created models, 14 new ionic liquids were synthesized and the predicted values were compared with the experimental refractive indices. Overall, decision tree based approaches were found to yield the best performance. We believe the obtained models can be effectively used for high-throughput, predictive screening of application oriented ionic liquids.

2. Methods and Materials

2.1. Data Curation

The refractive indices of 467 ionic liquids at the sodium D line, $n_D = 1.355$ to 1.659 covering $T = 283.15$ to 571.15 K were extracted from the ILThermo database[31, 32] and other articles in the literature. After the removal of duplicates, a total of 3147 experimental data points were obtained. The ILs are composed of 240 cations with major classes such as imidazolium, ammonium, pyrrolidinium and pyridinium, and 86 anions that are dominated by carboxylates, halides and sulfates. The structures of the cations and anions, the experimental n_D values and associated references are listed in Tables S1 and S2 in the supplementary material. A summary of the collected data is presented in Table 1.

Table 1: Summary of the experimental data with respect to the popular cation classes found in the data.

Cation	n_D	**Temperature (K)**	**#Data points**
Imidazolium	1.355-1.659	283-362	1379
Pyridinium	1.405-1.577	283-353	550
Pyrrolidinium	1.395-1.498	283-353	146
Piperidinium	1.412-1.514	288-353	81
Ammonium	1.362-1.545	283-571	572

2.2. Descriptor Calculation

Data extracted from the ILThermo database was parsed and the chemical names of the ionic liquids were converted to 2D format using the chemical name to structure software, OPSIN[33]. A conformational search for both cations and anions was carried out using using OpenBabel[34] (based on the Universal Force Field[35]). The structures were further optimized using the PM6 Hamiltonian in MOPAC[36] with the keywords: "PM6 XYZ PRECISE STATIC POLAR MMOK SUPER ENPART LARGE". The HOMO/LUMO energies, polarizabilities, superdelocalizabilities, charge partial surface areas (CPSA) and geometrical indices were used as descriptors that were calculated using the software KRAKENX[37]. For each ion, 113 descriptors are computed yielding a total of 226 indices for each cation-anion pair while ion-ion interactions are ignored. The temperature at which the refractive index was recorded was included as an additional variable in the data matrix.

2.3. Machine Learning

With a view to reducing the dimensionality of the original descriptor matrix, low variance columns and those containing missing values were excluded. In addition, a pairwise correlation of the descriptor columns was performed and only one among the highly correlated pair of variables ($R^2 > 0.95$) was retained[38]. The remaining variables were then autoscaled to zero mean and unit variance.

The data set was further split randomly (50:50) into independent calibration and test sets containing 1646 and 1501 data points respectively while ensuring that no cation-anion pair was common to both sets. Analysis of the splits showed that 98 cations and 23 anions were unique to the test set which provides for a more robust test of the ML methods.

Prediction of ionic liquid properties using non-linear methods has been found to be quite successful and have therefore been employed to carry out the regression analysis. Available routines implemented in the statistical software *R*[39] have been used and include ensemble methods such as generalized boosted regression[40] (GBM), random forests[41] (RF) and Cubist [42]. Ensemble learning methods aggregate the results from individual trees/rules. For both GBM and RF routines, the number of trees was set to 500. For the Cubist approach, wherein iterative model trees (with adjusted weights) are created in sequence, the number of optimal committees was identified using a grid search. For the Cubist model, the final prediction is then calculated as the average of predictions from all committee members. The linear partial least-squares regression[43] (PLSR) was also included for comparison. In all cases, the data were mean-centered and scaled before modelling. For the models created using the calibration data, 5-fold cross-validation (repeated three times to account for the randomness of the data splits) was carried out to assess the predictive ability. Further, in order to establish the reliability of the model predictions, bootstrap-estimated uncertainties[44] were calculated for the test set. Given a set of N objects/samples, a random sample of N members is drawn (with replacement) from the collection. In this study, a total of 100 ML models (for computational expediency) were built using the different bootstrap samples. For a given IL, the uncertainty was then computed as the standard deviation of predictions obtained from the 100 models where small values typically indicate more reliable predictions. All calculations were carried out on a desktop PC with Intel i5-2400 Quad-Core 3.10GHz CPU and 8GB RAM.

2.4. Experimental Details

The refractive indices for ionic liquids: 3-(2-diethylaminoethyl)-1-methylimidazolium bis(trifluoromethylsulfonyl)imide, 1-butyl-3-methylimidazolium 1,2,4-triazolate, 1-butyl-3-methylimidazolium 1,2,3-triazolate and 1-butyl-1-methylpyrrolidinium 1,2,4-triazolate, were measured at 276.5±0.5 K using a PAL-RI refractometer from Atago, with an uncertainty of ±0.0003 (water at 273 K). These ILs were prepared based on previously reported methods[45–47]. Refractive indexes of 1-(2-cyanoethyl)-3-(2-(2-(2-methoxyethoxy)ethoxy)ethyl)imidazolium bis(trifluoromethanesulfonyl)amide, 1-(2-cyanoethyl)-3-(2-ethoxyethyl)-imidazolium, 1-ethyl-3-(2-methoxycarbonyl-ethyl)-3-imidazolium dicyanamide, 1-propyl-3-(2-methoxycarbonyl-ethyl)-3-imidazolium dicyanamide, 1-butyl-3-(2-methoxycarbonyl-ethyl)-3-imidazolium dicyanamide, 1-pentyl-3-(2-methoxycarbonyl-ethyl)-3-imidazolium dicyanamide, 1-hexyl-3-(2-methoxycarbonyl-ethyl)-3-imidazolium dicyanamide were measured using a refractometer (Mettler Toledo, RM40) with a temperature scan ranging from 20°C - 60°C. The apparatus was calibrated by measuring the refractive index of Millipore quality water before

measurements. The uncertainty of the refractometer was 0.01. Triplicate measurements were taken for each sample at each temperature to ensure the effectiveness of the measurement. These ionic liquids were synthesized according to the previously reported procedures[48–50].

3. Results and Discussion

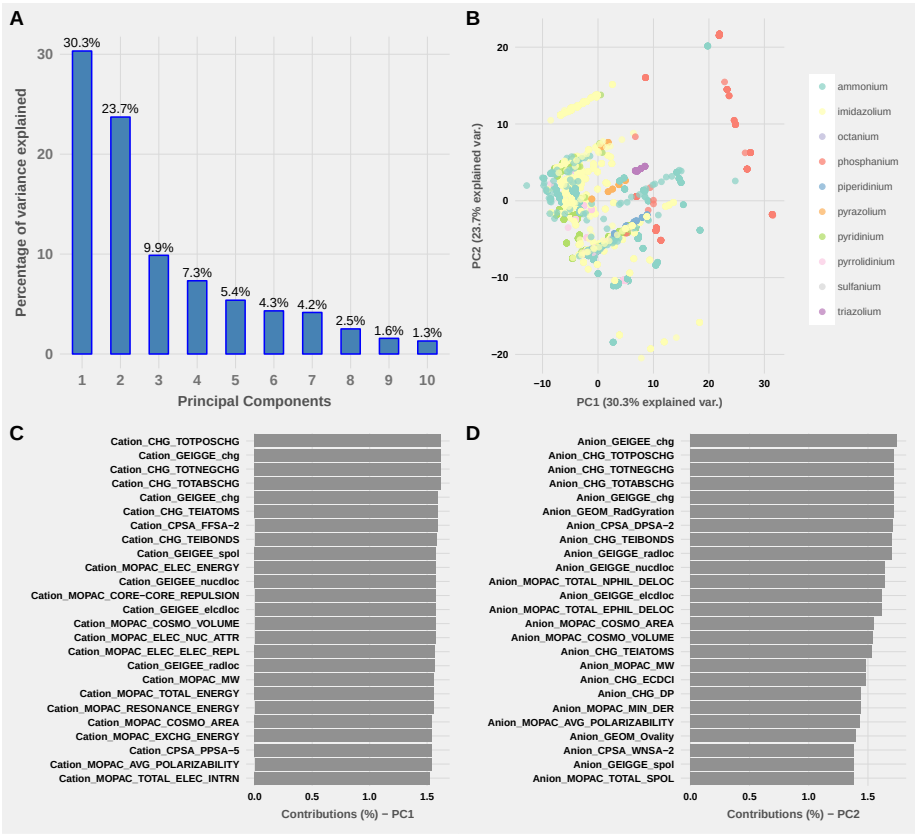


Figure 1: Visual summary of the data based on PCA: (A) Explained variances for the first 10 PCs. (B) Score plot with respect to the first two PCs, shows the locations of the prominent cation families (imidazolium, pyridinium, ammonium, phosphanium), (C) and (D) show the variable contributions with respect to PC1 and PC2. For brevity, only the top 25 variables are shown.

As an initial step, principal component analysis (PCA) was carried out on the autoscaled descriptor matrix. Figure 1 shows a graphical summary of the analysis. The first 10 principal components (PCs) explain almost 90% of the variance with the a little more than 50% concentrated in the first 2 PCs. While, the dominant groups of imidazolium and ammonium cations are somewhat scattered, the other families such as pyrazoliums, pyridiniums and to some

extent, phosphaniums are clustered around the center. The loading contributions for the first 2 PCs suggest that the first PC is significantly influenced by the cation-based descriptors while the second is largely dominated by the anion related variables. In both cases, the top ranking descriptors highlight the charge distribution that implicitly account for the interactions between the functional moieties. Interestingly, temperature is seen as a less influential descriptor with respect to the PCs (see the loadings plot in Figure F1 in the Supplementary material).

The model performances are summarised in terms of standard metrics: the squared coefficient of correlation (R^2), root mean square error ($RMSE$) and the mean absolute error (MAE). Following the removal of near-constant columns and highly correlated descriptors, each model was built with a descriptor matrix containing 103 variables. Table 2 summarises the results for the different machine learning models. For PLSR, a 10-component model was obtained which performs well for both training and test sets with R^2 of 0.80 and 0.73 respectively. Non-linear methods however show much improved statistics with $R^2_{cv} > 0.95$ and $R^2 > 0.80$ with only marginal differences in the calculated metrics. A single regression tree based model was also evaluated. However, while the performance for this model on the training data was found to be comparable with the rest of the methods, the test set prediction ($R^2 = 0.65$, $RMSE = 0.025$) were less impressive. Overall, the Cubist approach was seen to produce the best performance followed by the GBM and RF methods. For all models, $Y-$randomization tests[16] repeated 500 times, yielded p-values < 0.001 that suggest that the possibility of overfitting is low.

Table 2: Table summarises the machine learning performances for different regression methods applied to independent calibration and test sets. R^2_{cv} is the 5-fold cross-validated squared correlation coefficient. Numbers in brackets in the RMSE column are the corresponding mean absolute errors.

Method	**Training**		**Testing**	
	$RMSE$ (MAE)	R^2_{cv}	$RMSE$ (MAE)	R^2
PLSR	0.018 (0.012)	0.80	0.021 (0.015)	0.74
GBM	0.006 (0.002)	0.97	0.017 (0.011)	0.82
CUBIST	0.006 (0.0004)	0.97	0.016 (0.010)	0.84
RF	0.009 (0.004)	0.96	0.018 (0.013)	0.82
CART	0.009 (0.005)	0.95	0.025 (0.017)	0.65

The results were further analysed with respect to the prominent cation classes present in the data set, a summary of which is provided in Table 3. Although small fluctuations in performance are seen for the different models, the Cubist approach consistently performs well across all classes with $R^2 > 0.85$. For the imidazoliums which are the dominant cation group, all models yield fairly consistent predictions with $R^2 > 0.90$ with a slightly lower performance for the PLSR model. Good performance trends are also observed for the piperidinium based ILs.

In an attempt to improve the results where possible using variable selec-

Table 3: Table summarises the predictive performances for different regression methods applied to the different cation groups across the entire data.

Cation	Method	$RMSE$	MAE	R^2
Imidazolium	PLSR	0.019	0.013	0.83
	GBM	0.011	0.005	0.94
	CUBIST	0.01	0.005	0.95
	RF	0.012	0.008	0.93
Pyridinium	PLSR	0.016	0.012	0.77
	GBM	0.013	0.004	0.83
	CUBIST	0.011	0.004	0.87
	RF	0.011	0.006	0.88
Pyrrolidinium	PLSR	0.014	0.009	0.70
	GBM	0.01	0.006	0.87
	CUBIST	0.009	0.004	0.86
	RF	0.012	0.008	0.82
Piperidinium	PLSR	0.01	0.008	0.99
	GBM	0.007	0.006	0.98
	CUBIST	0.006	0.005	0.98
	RF	0.006	0.005	0.98
Ammonium	PLSR	0.023	0.018	0.73
	GBM	0.013	0.008	0.92
	CUBIST	0.015	0.008	0.88
	RF	0.017	0.011	0.87

tion, the recursive feature elimination algorithm (implemented in the caret[51] package in R) was applied to the RF, GBM and Cubist models. The algorithm selects features by recursively considering smaller and smaller sets of features. For PLSR, the variable importance in projection (VIP)[52] was used. The VIP score is useful in analysing the predictor variables that best explain the variance in the response. However, despite using variable selection no discernible improvement in the model performance was observed. Variable importance plots highlighting the top 20 influential descriptors in each model are shown in Figure 2. Additional figures showing the rankings for more variables are shown in Figures F2 and F3 in the Supplementary material. While the VIP score is used to highlight prominent variables for PLSR, in the case of RF, GBM and Cubist approaches, the importance was calculated with respect to the reduction in error when the predictor of interest is permuted. For the PLSR model, the top ranking variables are dominated by anion-specific descriptors, while for the RF model, cation-specific predictors are ranked at the top. Examination of other lower ranked descriptors extends these trends to a large extent, although some cationic (such as the HOMO energy for PLSR) and anionic (such as the heat of formation) variables start to become relevant. In comparison, both Cubist and GBM models exhibit a mixture of cationic and anionic variables. Many of the top ranking descriptors are also seen to be crucial in determining the refractive

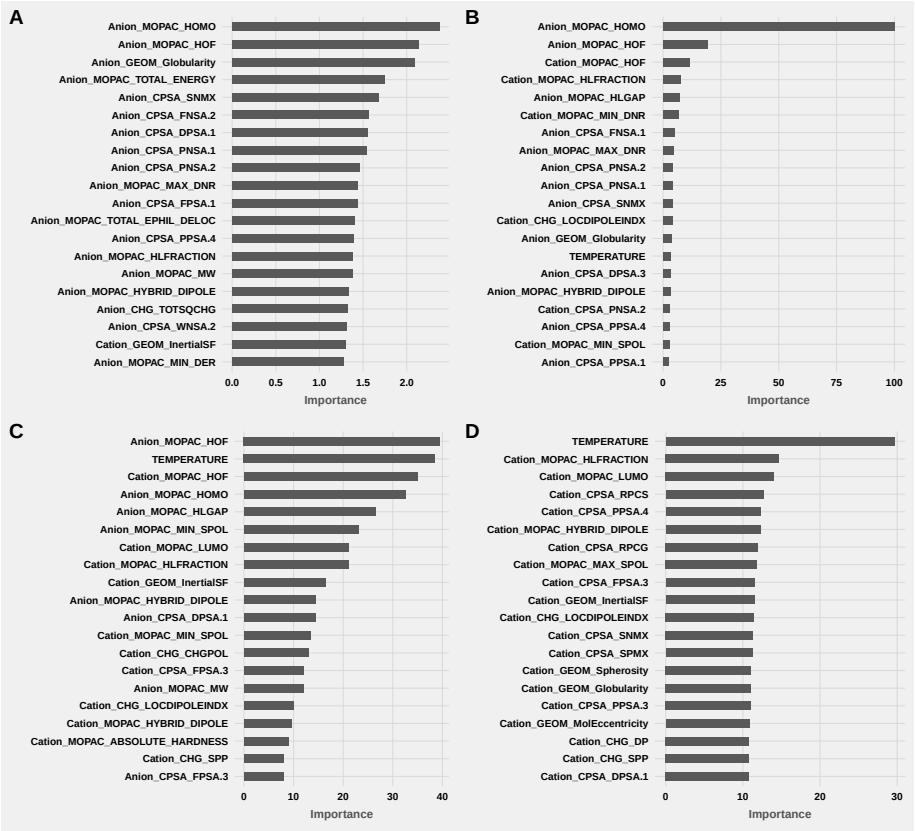


Figure 2: For each model, the 20 most prominent variables influencing the n_D predictions are shown (A) PLSR, (B) GBM (C) Cubist and (D) RF.

index of polymers[53, 54]. The refractive index is considered as a response to electronic polarization with large ions typically having large polarizabilities[55]. A descriptor that expresses the polarization of the molecule is given by the local dipole index [56] which is the average of the charge differences over all bonded atom pairs. The charged partial surface area descriptors that summarize the charge distribution in the ion, are related to the molecular size. Another indicator of size is the inertial shape factor calculated for the cation is based on the principal moments of inertia[57]. The self polarisabilities are dynamic reactivity indices and reflect the interactions between the cation and anion[58]. Other variables such as the HOMO-LUMO energy gap, are also related to the polarisability where a small value can indicate that the structure is easily polarised. The molecular weight of the anion is also seen to impact the refractive index where an increase in weight leads to a decrease in n_D[59]. The heat of formation (HOF) can be taken as a measure of the thermodynamic stability of the IL[58]. While temperature features as a low ranking variable (based on the VIP scores)

for PLSR (not seen in the top 20 variables), both Cubist and random forests attach a high importance to the same.

Reliability of Predictions

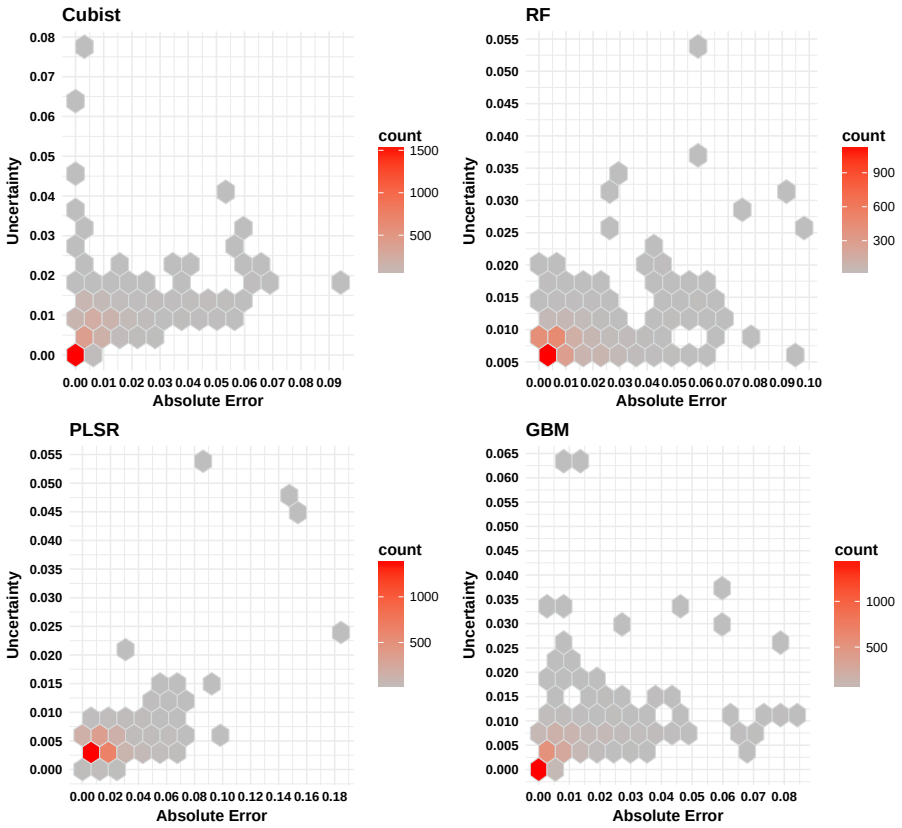


Figure 3: For each model, absolute errors (absolute value of each error) vs the corresponding bootstrap uncertainties are shown as a 2D histogram. The counts reflect the number of instances where the two values/bins overlap.

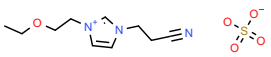
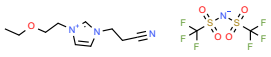
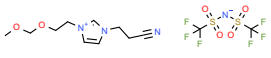
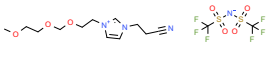
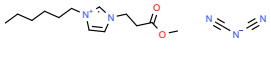
Model predictions and their corresponding bootstrap uncertainties (standard deviation of the bootstrap model predictions) are listed in Table S2 in the Supplementary material. Analysis of the uncertainties calculated for the different ML model predictions, shows that for both the Cubist and random forests approaches, nearly 67% of the absolute value of the prediction errors (calculated for the entire data) fall within one standard deviation. In comparison, for the GBM and PLSR models, only 50% and 17% of the cases have prediction errors within one standard deviation. Figure 3 shows the 2D histogram of the absolute prediction errors against the corresponding uncertainties which enables the analysis of the relationship between the two numerical variables. Within the

2D histogram, the number of data points at a certain value of the error and uncertainty are counted and shown as a colour map, where grey represents low counts and red represent high counts. Assuming a uncertainty cutoff of 0.01 (predictions with greater uncertainties should be treated with caution), it was seen that for the non-linear methods (GBM, Cubist and RF), more than 70% of the predictions with associated uncertainties below the cutoff value have a corresponding absolute error less than 0.01. For the linear PLSR method, however, this number falls to 46%. For uncertainties greater than the cutoff value, less than 1% of the predictions are seen to have absolute errors less than 0.01. The results suggest that predictions with low uncertainties do in general have high accuracies. For uncertainties greater than 0.01, the size of the prediction errors are not consistently captured, and thus poor predictions are not always easily identified (see Figure F4 in the Supplementary material).

Experimental validation

In order to further evaluate the predictive ability of the ML models, a new set of 14 ionic liquids were synthesized and their refractive indices recorded (see Experimental Details). The refractive indices for the newly synthesized ILs range between 1.44 and 1.54. The new ILs include imidazoliums and piperidiniums combined with anions such as bis(trifluoromethylsulfonyl)imide, thiocyanate, hydrogen sulphate, dicyanamide and triazolides. The structures of the cations and anions, their experimental and ML predicted refractive indices are listed in Table 4.

Table 4: Table lists the model predictions for new ILs synthesized. n_D is the experimental refractive index measured at temperature T (K) and $\widehat{n_D}$ the ML predicted values along with the bootstrap uncertainties.

Ionic Liquid	T	n_D	$\widehat{n_D}$			
			PLSR	GBM	CUBIST	RF
	293	1.522	1.485±0.007	1.478±0.007	1.502±0.011	1.487±0.007
	303	1.518	1.483±0.007	1.476±0.007	1.501±0.011	1.487±0.007
	313	1.513	1.481±0.007	1.473±0.007	1.499±0.011	1.486±0.007
	323	1.508	1.480±0.007	1.470±0.007	1.497±0.011	1.486±0.007
	333	1.504	1.478±0.007	1.467±0.007	1.496±0.011	1.486±0.007
	293	1.441	1.447±0.006	1.439±0.005	1.445±0.007	1.435±0.008
	303	1.427	1.445±0.006	1.436±0.005	1.442±0.007	1.435±0.008
	313	1.424	1.443±0.006	1.433±0.005	1.439±0.008	1.435±0.008
	323	1.420	1.441±0.006	1.430±0.005	1.436±0.008	1.434±0.008
	333	1.415	1.439±0.006	1.426±0.005	1.434±0.008	1.434±0.008
	293	1.446	1.447±0.006	1.458±0.010	1.458±0.010	1.443±0.006
	303	1.442	1.445±0.006	1.453±0.010	1.455±0.010	1.443±0.006
	313	1.438	1.443±0.006	1.449±0.011	1.452±0.010	1.442±0.006
	323	1.433	1.441±0.006	1.447±0.011	1.449±0.010	1.442±0.006
	333	1.429	1.439±0.006	1.443±0.011	1.446±0.010	1.442±0.006
	293	1.448	1.440±0.007	1.484±0.013	1.449±0.012	1.445±0.007
	303	1.445	1.438±0.007	1.479±0.013	1.447±0.013	1.444±0.007
	313	1.441	1.436±0.007	1.475±0.013	1.444±0.013	1.444±0.007
	323	1.438	1.434±0.008	1.473±0.013	1.442±0.014	1.444±0.007
	333	1.434	1.432±0.008	1.469±0.013	1.440±0.014	1.444±0.007
	293	1.509	1.507±0.005	1.507±0.005	1.500±0.009	1.509±0.007
	303	1.505	1.505±0.005	1.503±0.005	1.496±0.010	1.509±0.007
	313	1.501	1.503±0.005	1.501±0.005	1.494±0.010	1.509±0.007
	323	1.498	1.501±0.005	1.496±0.005	1.491±0.010	1.509±0.008
	333	1.494	1.499±0.005	1.490±0.005	1.488±0.010	1.508±0.008

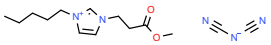
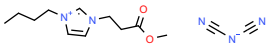
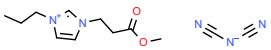
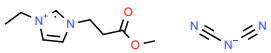
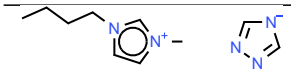
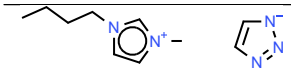
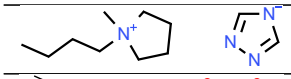
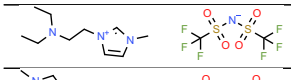
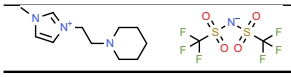
	293	1.514	1.509±0.006	1.511±0.007	1.504±0.011	1.510±0.006
	303	1.510	1.507±0.006	1.505±0.007	1.499±0.011	1.510±0.007
	313	1.507	1.505±0.006	1.501±0.007	1.497±0.011	1.510±0.007
	323	1.503	1.503±0.006	1.497±0.007	1.493±0.011	1.509±0.007
	333	1.500	1.501±0.006	1.491±0.007	1.490±0.011	1.509±0.006
	293	1.518	1.504±0.005	1.506±0.005	1.505±0.009	1.509±0.008
	303	1.515	1.502±0.005	1.502±0.005	1.500±0.010	1.509±0.008
	313	1.511	1.500±0.005	1.499±0.005	1.497±0.010	1.508±0.008
	323	1.508	1.499±0.005	1.496±0.006	1.494±0.009	1.508±0.008
	333	1.505	1.497±0.005	1.489±0.005	1.491±0.009	1.508±0.008
	293	1.522	1.503±0.005	1.511±0.005	1.504±0.009	1.509±0.008
	303	1.518	1.501±0.005	1.507±0.005	1.500±0.009	1.509±0.008
	313	1.513	1.499±0.005	1.504±0.005	1.497±0.009	1.508±0.008
	323	1.507	1.497±0.005	1.501±0.005	1.494±0.009	1.508±0.008
	333	1.504	1.496±0.005	1.495±0.005	1.491±0.009	1.508±0.008
	293	1.528	1.505±0.005	1.507±0.006	1.504±0.010	1.509±0.008
	303	1.524	1.503±0.005	1.503±0.006	1.499±0.010	1.509±0.008
	313	1.521	1.502±0.005	1.501±0.006	1.497±0.010	1.508±0.008
	323	1.518	1.500±0.005	1.497±0.006	1.493±0.010	1.508±0.008
	333	1.514	1.498±0.005	1.491±0.006	1.490±0.010	1.508±0.008
	297	1.521	1.499±0.005	1.514±0.014	1.498±0.024	1.486±0.007
	296	1.521	1.509±0.006	1.532±0.011	1.520±0.014	1.491±0.006
	296	1.506	1.479±0.005	1.489±0.013	1.494±0.022	1.474±0.017
	297	1.436	1.452±0.004	1.446±0.004	1.443±0.013	1.441±0.012
	296	1.449	1.451±0.004	1.475±0.009	1.455±0.015	1.442±0.007

Table 5 provides a statistical summary of the ML predictions for the experimental data. The correlation trends seen for the independent test set is to a large extent reproduced for the second set of test data. While the Cubist, random forests and PLSR models show good predictive power, the GBM model shows a small decrease in performance. Refractive index values for the triazolide anion-based ILs which present a completely unseen chemistry for the models are in general well predicted by the Cubist model.

Table 5: Table summarises the machine learning performances for different regression methods applied to experimentally synthesized ILs.

	PLSR	GBM	CUBIST	RF
R^2	0.89	0.74	0.97	0.87
$RMSE$	0.016	0.020	0.014	0.014
MAE	0.012	0.016	0.013	0.010

4. Conclusions

In this article, we have investigated the efficacy of various machine learning models in predicting temperature dependent refractive indices for a large number of ionic liquids. Predictive ability of the models was evaluated using and independent test set and an additional set containing 14 novel ILs obtained from experiments. Non-linear ensemble approaches are seen to produce significantly better results compared with single tree and linear partial least squares regression methods. The models obtained have broad applicability and should be particularly useful for fast screening of IL compounds.

Acknowledgements

The Norwegian Research Council (NFR) is acknowledged for financial support from the CLIMIT (Grant No. 233776).

References

[1] M. Watanabe, M. L. Thomas, S. Zhang, K. Ueno, T. Yasuda, K. Dokko, Application of ionic liquids to energy storage and conversion materials and devices, Chem. Rev. 117 (10) (2017) 7190–7239. doi:10.1021/acs.chemrev.6b00504.

[2] C. Verma, E. E. Ebenso, M. Quraishi, Ionic liquids as green and sustainable corrosion inhibitors for metals and alloys: An overview, J. Mol. Liq. 233 (2017) 403–414. doi:10.1016/j.molliq.2017.02.111.

[3] A. A. C. T. Hijo, G. J. Maximo, M. C. Costa, E. A. C. Batista, A. J. A. Meirelles, Applications of ionic liquids in the food and bioproducts industries, ACS Sustainable Chem. Eng. 4 (10) (2016) 5347–5369. doi:10.1021/acssuschemeng.6b00560.

[4] K. Wang, H. Adidharma, M. Radosz, P. Wan, X. Xu, C. K. Russell, H. Tian, M. Fan, J. Yu, Recovery of rare earth elements with ionic liquids, Green Chem. 19 (19) (2017) 4469–4493. doi:10.1039/c7gc02141k.

[5] W. L. Hough, M. Smiglak, H. Rodríguez, R. P. Swatloski, S. K. Spear, D. T. Daly, J. Pernak, J. E. Grisel, R. D. Carliss, M. D. Soutullo, J. James H. Davis, R. D. Rogers, The third evolution of ionic liquids: active pharmaceutical ingredients, New J. Chem. 31 (8) (2007) 1429. doi:10.1039/b706677p.

[6] Y. Sahbaz, H. D. Williams, T.-H. Nguyen, J. Saunders, L. Ford, S. A. Charman, P. J. Scammells, C. J. H. Porter, Transformation of poorly water-soluble drugs into lipophilic ionic liquids enhances oral drug exposure from lipid based formulations, Mol. Pharm. 12 (6) (2015) 1980–1991. doi:10.1021/mp500790t.

[7] Z. Lei, C. Dai, B. Chen, Gas solubility in ionic liquids, Chem. Rev. 114 (2) (2014) 1289–1326.

[8] J. Hulsbosch, D. E. D. Vos, K. Binnemans, R. Ameloot, Biobased ionic liquids: Solvents for a green processing industry?, ACS Sustainable Chem. Eng. 4 (6) (2016) 2917–2931. doi:10.1021/acssuschemeng.6b00553.

[9] H. Niedermeyer, J. P. Hallett, I. J. Villar-Garcia, P. A. Hunt, T. Welton, Mixtures of ionic liquids, Chem. Soc. Rev. 41 (23) (2012) 7780. doi:10.1039/c2cs35177c.

[10] D. S. Firaha, O. Hollóczki, B. Kirchner, Computer-aided design of ionic liquids as CO2absorbents, Angewandte Chemie 54 (27) (2015) 7805–7809. doi:10.1002/anie.201502296.

[11] S. Zahn, D. R. MacFarlane, E. I. Izgorodina, Assessment of kohn–sham density functional theory and møller–plesset perturbation theory for ionic liquids, Phys. Chem. Chem. Phys. 15 (32) (2013) 13664. doi:10.1039/c3cp51682b.

[12] E. I. Izgorodina, Z. L. Seeger, D. L. A. Scarborough, S. Y. S. Tan, Quantum chemical methods for the prediction of energetic, physical, and spectroscopic properties of ionic liquids, Chem. Rev. 117 (10) (2017) 6696–6754. doi:10.1021/acs.chemrev.6b00528.

[13] C. Herrera, G. García, M. Atilhan, S. Aparicio, A molecular dynamics study on aminoacid-based ionic liquids, J. Mol. Liq. 213 (2016) 201–212. doi:10.1016/j.molliq.2015.10.056.

[14] M. Fakhraee, M. R. Gholami, Biodegradable ionic liquids: Effects of temperature, alkyl side-chain length, and anion on the thermodynamic properties and interaction energies as determined by molecular dynamics simulations coupled with ab initio calculations, Ind. Eng. Chem. Res. 54 (46) (2015) 11678–11700. doi:10.1021/acs.iecr.5b03199.

[15] A. Varnek, N. Kireeva, I. V. Tetko, I. I. Baskin, V. P. Solov'ev, Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points?, J Chem. Inf. Model. 47 (3) (2007) 1111–1122. doi:10.1021/ci600493x.

[16] V. Venkatraman, B. K. Alsberg, Quantitative structure-property relationship modelling of thermal decomposition temperatures of ionic liquids, J. Mol. Liquids 223 (2016) 60–67.

[17] V. Venkatraman, B. K. Alsberg, Predicting CO 2 capture of ionic liquids using machine learning, J. CO2 Util. 21 (2017) 162–168. doi:10.1016/j.jcou.2017.06.012.

[18] Y. Zhao, Y. Huang, X. Zhang, S. Zhang, A quantitative prediction of the viscosity of ionic liquids using sσ-profile molecular descriptors, Phys. Chem. Chem. Phys. 17 (5) (2015) 3761–3767. doi:10.1039/c4cp04712e.

[19] M. A. Kareem, F. S. Mjalli, M. A. Hashim, I. M. AlNashef, Phosphonium-based ionic liquids analogues and their physical properties, J. Chem. Eng. Data 55 (11) (2010) 4632–4637. doi:10.1021/je100104v.

[20] M. Deetlefs, K. R. Seddon, M. Shara, Neoteric optical media for refractive index determination of gems and minerals, New J. Chem. 30 (3) (2006) 317. doi:10.1039/b513451j.

[21] Y. Kayama, T. Ichikawa, H. Ohno, Transparent and colourless room temperature ionic liquids having high refractive index over 1.60, Chem. Commun. 50 (94) (2014) 14790–14792. doi:10.1039/c4cc06145d.

[22] K. Bica, M. Deetlefs, C. Schrder, K. R. Seddon, Polarisabilities of alkylimidazolium ionic liquids, Phys. Chem. Chem. Phys. 15 (8) (2013) 2703. doi:10.1039/c3cp43867h.

[23] R. L. Gardas, J. A. P. Coutinho, Group contribution methods for the prediction of thermophysical and transport properties of ionic liquids, AIChE Journal 55 (5) (2009) 1274–1290. doi:10.1002/aic.11737.

[24] M. Sattari, A. Kamari, A. H. Mohammadi, D. Ramjugernath, A group contribution method for estimating the refractive indices of ionic liquids, J. Mol. Liq. 200 (2014) 410–415. doi:10.1016/j.molliq.2014.11.005.

[25] X. Wang, X. Lu, Q. Zhou, Y. Zhao, X. Li, S. Zhang, Database and new models based on a group contribution method to predict the refractive index of ionic liquids, Phys. Chem. Chem. Phys. 19 (30) (2017) 19967–19974. doi:10.1039/c7cp03214e.

[26] M. Sattari, A. Kamari, A. H. Mohammadi, D. Ramjugernath, Prediction of refractive indices of ionic liquids – a quantitative structure-property relationship based model, J. Taiwan Inst. Chem. Eng. 52 (2015) 165–180. doi:10.1016/j.jtice.2015.02.003.

[27] P. Díaz-Rodríguez, J. C. Cancilla, N. V. Plechkova, G. Matute, K. R. Seddon, J. S. Torrecilla, Estimation of the refractive indices of imidazolium-based ionic liquids using their polarisability values, Phys. Chem. Chem. Phys. 16 (1) (2014) 128–134. doi:10.1039/c3cp53685h.

[28] P. Díaz-Rodríguez, J. C. Cancilla, G. Matute, D. Chicharro, J. S. Torrecilla, Inputting molecular weights into a multilayer perceptron to estimate refractive indices of dialkylimidazolium-based ionic liquids-a purity evaluation, Appl. Soft Comp. 28 (2015) 394–399. doi:10.1016/j.asoc.2014.12.004.

[29] X. Kang, Y. Zhao, J. Li, Predicting refractive index of ionic liquids based on the extreme learning machine (ELM) intelligence algorithm, J. Mol. Liq. 250 (2018) 44–49. doi:10.1016/j.molliq.2017.11.166.

[30] E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandler, C.-C. Chen, M. Zwolak, K. C. Seavey, Sigma-profile database for using COSMO-based thermodynamic methods, Ind. Eng. Chem. Res. 45 (12) (2006) 4389–4415. doi:10.1021/ie060370h.

[31] Q. Dong, C. D. Muzny, A. Kazakov, V. Diky, J. W. Magee, J. A. Widegren, R. D. Chirico, K. N. Marsh, M. Frenkel, ILThermo: a free-access web database for thermodynamic properties of ionic liquids†, J. Chem. Eng. Data 52 (4) (2007) 1151–1159. doi:10.1021/je700171f.

[32] A. Kazakov, J. Magee, R. Chirico, E. Paulechka, V. Diky, C. Muzny, K. Kroenlein, M. Frenkel, Nist standard reference database 147: Nist ionic liquids database - (ilthermo) (2017).
URL http://ilthermo.boulder.nist.gov

[33] D. M. Lowe, P. T. Corbett, P. Murray-Rust, R. C. Glen, Chemical name to structure: OPSIN, an open source solution, J Chem. Inf. Model. 51 (3) (2011) 739–753. doi:10.1021/ci100384d.

[34] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open babel: An open chemical toolbox, J. Cheminf. 3 (1) (2011) 33.

[35] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. G. III, W. M. Skiff, Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations, J. Am. Chem. Soc. 114 (25) (1992) 10024–10035.

[36] J. J. P. Stewart, Mopac2016, stewart Computational Chemistry, Colorado Springs, CO, USA, (http://OpenMOPAC.net) (2016).

[37] V. Venkatraman, B. K. Alsberg, Krakenx: software for the generation of alignment-independent 3d descriptors, J. Mol. Model. 22 (4) (2016) 1–8.

[38] M. Shen, Y. Xiao, A. Golbraikh, V. K. Gombar, A. Tropsha, Development and validation ofk-nearest-neighbor QSPR models of metabolic stability of drug candidates, J. Med. Chem. 46 (14) (2003) 3013–3020.

[39] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2017).
URL https://www.R-project.org/

[40] G. R. with contributions from others, gbm: Generalized Boosted Regression Models, r package version 2.1.1 (2015).
URL https://CRAN.R-project.org/package=gbm

[41] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (3) (2002) 18–22.

[42] M. Kuhn, S. Weston, C. Keefer, N. C. C. code for Cubist by Ross Quinlan, Cubist: Rule- And Instance-Based Regression Modeling, r package version 0.0.19 (2016).
URL https://CRAN.R-project.org/package=Cubist

[43] B.-H. Mevik, R. Wehrens, The pls package: Principal component and partial least squares regression in r, J. Stat. Soft. 18 (2) (2007) 1–24.
URL http://www.jstatsoft.org/v18/i02

[44] M. Toplak, R. Močnik, M. Polajnar, Z. Bosnić, L. Carlsson, C. Hasselgren, J. Demšar, S. Boyer, B. Zupan, J. Stålring, Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models, J. Chem. Inf. Model. 54 (2) (2014) 431–441. doi:10.1021/ci4006595.

[45] D. Yang, M. Hou, H. Ning, J. Ma, X. Kang, J. Zhang, B. Han, Reversible capture of SO2through functionalized ionic liquids, ChemSusChem 6 (7) (2013) 1191–1195. doi:10.1002/cssc.201300224.

[46] K. K. Laali, A. Jamalian, G. L. Borosky, Piperidine-appended imidazolium ionic liquids as task-specific catalysts: computational study, synthesis, and multinuclear NMR, J. Phys. Org. Chem. 29 (7) (2016) 346–351. doi:10.1002/poc.3541.

[47] S. Seo, M. A. DeSilva, J. F. Brennecke, Physical properties and CO2reaction pathway of 1-ethyl-3-methylimidazolium ionic liquids with aprotic heterocyclic anions, J. Phys. Chem. B 118 (51) (2014) 14870–14879. doi:10.1021/jp509583c.

[48] J. J. Raj, C. D. Wilfred, S. N. Shah, M. Pranesh, M. A. Mutalib, K. C. Lethesh, Physicochemical and thermodynamic properties of imidazolium ionic liquids with nitrile and ether dual functional groups, J. Mol. Liq. 225 (2017) 281–289. doi:10.1016/j.molliq.2016.11.049.

[49] J. J. Raj, S. Magaret, M. Pranesh, K. C. Lethesh, W. C. Devi, M. A. Mutalib, Extractive desulfurization of model fuel oil using ester functionalized imidazolium ionic liquids, Sep. Purif. Technol.doi:10.1016/j.seppur.2017.08.050.

[50] M. Montanino, M. Carewska, F. Alessandrini, S. Passerini, G. B. Appetecchi, The role of the cation aliphatic side chain length in piperidinium bis(trifluoromethansulfonyl)imide ionic liquids, Electrochim. Acta 57 (2011) 153–159. doi:10.1016/j.electacta.2011.03.089.

[51] M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty,

R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt., caret: Classification and Regression Training, r package version 6.0-73 (2016).
URL https://CRAN.R-project.org/package=caret

[52] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (10) (2015) 528–536. doi:10.1002/cem.2736.

[53] A. R. Katritzky, S. Sild, M. Karelson, Correlation and prediction of the refractive indices of polymers by QSPR, J. Chem. Inf. Model. 38 (6) (1998) 1171–1176. doi:10.1021/ci980087w.

[54] V. Venkatraman, B. Alsberg, Designing high-refractive index polymers using materials informatics, Polymers 10 (2) (2018) 103. doi:10.3390/polym10010103.
URL https://doi.org/10.3390%2Fpolym10010103

[55] M. G. Montalbán, C. L. Bolívar, F. G. D. Baños, G. Víllora, Effect of temperature, anion, and alkyl chain length on the density and refractive index of 1-alkyl-3-methylimidazolium-based ionic liquids, J. Chem. Eng. Data 60 (7) (2015) 1986–1996. doi:10.1021/je501091q.
URL https://doi.org/10.1021%2Fje501091q

[56] O. Kikuchi, Systematic QSAR procedures with quantum chemical descriptors, Mol. Inf. 6 (4) (1987) 179–184. doi:10.1002/qsar.19870060406.
URL https://doi.org/10.1002%2Fqsar.19870060406

[57] R. Todeschini, V. Consonni, Descriptors from Molecular Geometry, Wiley-VCH Verlag GmbH, 2008, pp. 1004–1033.

[58] M. Karelson, V. S. Lobanov, A. R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (3) (1996) 1027–1044. doi:10.1021/cr950202r.

[59] S. Seki, S. Tsuzuki, K. Hayamizu, Y. Umebayashi, N. Serizawa, K. Takei, H. Miyashiro, Comprehensive refractive index property for room-temperature ionic liquids, J. Chem. Eng. Data 57 (8) (2012) 2211–2216. doi:10.1021/je201289w.
URL https://doi.org/10.1021%2Fje201289w