ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

A Generalized Machine Learning Model for Predicting Ionic Conductivity for Ionic Liquids'†

Pratik Dhakal and Jindal K. Shah†

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Ionic liquids are currently being considered as potential electrolyte candidates for next-generation batteries and energy storage devices due to their high thermal and chemical stability. However, high viscosity and low conductivity at lower temperatures have severely hampered their commercial applications. To overcome these challenges, it is necessary to develop structure-property models for ionic liquid transport properties to guide the ionic liquid design. This work expands our previous effort in developing a machine learning model on imidazolium-based ionic liquids to now include ten different cation families, representing structural and chemical diversity. The model dataset contains 2869 ionic conductivity values over a temperature range of 238-472 K collected from the NIST ILThermo database and literature values for 397 unique ionic liquids. The database covers 214 unique cations and 68 unique anions. Three machine learning models, multiple linear regression, random forest, and extreme gradient boosting, are applied to correlate the ionic liquid conductivity data with cation and anion features. Shapely additive analysis is performed to glean insights into cation and anion features with significant impact on ionic conductivity. Finally, the extreme gradient boosting model is used to predict ionic conductivity of ionic liquids from all the possible combinations of unique cations and anions to identify ionic liquids crossing the ionic conductivity threshold of 2.0 S/m.

1 Introduction

The asymmetric cationic structures and articulated nature of anion are responsible for charge delocalization and frustrated crystal packing for a large number of ionic liquids leading many to exist as liquid at ambient conditions. In contrast to conventional solvents, ionic liquids offer several unique and desirable properties such as negligible vapor pressure, low melting point and nonflammability. These attributes are primary reasons ionic liquids are studied extensively for various industrial applications such as solvents in chemical separation/purification[1,2], as catalysts[3,4], use in CO_2 capture[5,6] and potential electrolytes for battery application[7,8].

The use of ionic liquids for battery applications and energy storage medium is primarily due to their high thermal[9] and chemical stability[10] to address tremendous safety concern associated with the current state-of-the-art electrolytes found in Li-ion batteries[11–13]. For example, current electrolytes powering Li-ion batteries are carbonate-based electrolytes mixed with salts such as lithium hexafluorophosphate $LiPF_6$, which are very volatile, flammable, and potentially hazardous during thermal runaway reactions or short-circuit[14,15]. Kalhoff et al. carried out an extensive study on the performance and safety of electrolytes based on organic carbonates (OC) and ionic liquids among others[16]. The authors noted the superiority of OC electrolytes in terms of ionic conductivity; however, the performance of OC was poor for electrochemical and thermal stability. Additionally, these solvents posed safety concerns. On the other hand, ionic liquids received a high rating for electrochemical and thermal stability, and safety consideration, but only medium for ionic conductivity, and suffered from poor low-temperature performance. Thus, for ionic liquids to be considered potential electrolyte candidates, an improvement in low ionic conductivity performance at sub-ambient conditions is needed in the next-generation of ionic liquids.

As is common for almost all applications involving ionic liquids, a systematic improvement in the transport properties of ions can be accomplished by selecting an optimal cation-anion combination using chemical intuition. The approach, however, is likely to be slow and time consuming due to the staggering number of such possible combinations[17] in the range of 10^{14}. The presence of a myriad of interactions such as electrostatic, hydrogen bonding, π-π stacking, anion-π, and van der Waals further complicates choosing cation-anion pairing to deliver anticipated property

School of Chemical Engineering, Oklahoma State University, Stillwater, Oklahoma, 74078, United States; E-mail: jindal.shah@okstate.edu

† Electronic Supplementary Information (ESI) available: [add github repo]. See DOI: 00.0000/00000000.

enhancement. For example, the attempt to alter the hydrogen bonding interactions through alkyl substitution of the most acidic hydrogen site in the imidazolium cation led to an increase in the viscosity of the resulting ionic liquid - a counterintuitive result[18–20]. An experimental high throughput screening approach may also not be feasible due to the requirement for ensuring the purity of the synthesized ionic liquids. Similarly, molecular simulation-based techniques such as molecular dynamics and Monte Carlo simulations can, in principle, accelerate the search of ionic liquids with desired properties[21–23]; however, describing interactions between ionic liquid components continues to be a nontrivial task due to the long simulation time required to calculate ionic conductivity[24] and a large deviation in prediction between experiment and simulation with the current available forcefields for ionic liquids[22,25,26]. Given these challenges and the availability of the ionic liquid property database - ILThermo - maintained by NIST, machine learning-based methods are gaining attention as a pre-screening tool to correlate ionic liquid properties with attributes that describe cations and anions[27–30]. Genetic mutation and generative-based models also allow the accelerated discovery of ionic liquids with properties within desired range[31,32].

In our previous proof-of-concept article,[30] we focused on modeling ionic conductivity using an artificial neural network and support vector regression models for imidazolium-based ionic liquids as these ionic liquids are generally less viscous and possess high ionic conductivity at room temperature - key properties for battery electrolytes[8]. Additionally, a large amount of data is available for imidazolium-based ionic liquids enabling machine learning model development. One of the difficulties of using imidazolium-based ionic liquids is that the electrochemical stability of imidazolium cations is rather low - less than 4.0 V (vs. Li/Li^+) - which is not suitable for high voltage battery application[33]. The primary reason for this behavior is the susceptibility of the cation to reduction at the most acidic proton at the C_2 position. Protecting this position by substituting various functional groups improves the stability but leads to slower dynamics[18] in comparison to that for the parent ionic liquid. The next closest relative to imidazolium cations are the pyridinium-based cations that are more sluggish with high viscosity and low ionic conductivity, which is why there is a limited amount of study done on exploring its application as electrolytes for battery application[34–36]. Beyond the aromatic cations, cyclic cations such as pyrrolidinium and piperidinium cations have generated tremendous interest as they have a high biodegradability rate and low toxicity[37,38]. The pyrrolidinium cation also offers low viscosity and high ionic conductivity, and unlike imidazolium cations, are more electrochemically stable, with a majority of them exhibiting electrochemical window reaching above 4.5 - 5.0 V[8]. Along with faster dynamics, pyrrolidinium cations also have better stability towards lithium metal, making them an ideal candidate for battery application as potential electrolytes[39,40].

Modifying the five-ring pyrrolidinium structure to a six-ring structure gives rise to piperidinium cations. Similar to pyrrolidinium cations, piperidinium cations have slower dynamics than pyrrolidinium cations because of the bulky nature of the cation. As such, there are relatively few studies that have explored the possibility of piperidinium cations as electrolytes for battery application[41–43]. Besides cyclic and aromatic cations, other central atom-based cations such as tetralkylphosphonium, tetraalkylammonium, and trialkylsulfonium, are also extensively studied for various applications[44–46]. The ammonium-based cations are characterized by a high electrochemical window compared to imidazolium but suffer from high viscosity and low ionic conductivity[47]. An alternative to nitrogen-based cations is phosphonium-based cations that have similar properties as ammonium ionic liquids, with some of them outperforming ammonium cations[45,48]. The other common cation type is sulfonium-based ionic liquids which have favorable properties compared to phosphonium-based cations because of the small volume occupied by the core sulfur atom leading to lower viscosity and high ionic conductivity[49–51]. In addition to the commonly studied cations, there are several other cation types such as morpholinium[41,43,52], pyrazolium[53,54], oxazolidinium[55] which might offer desirable properties for battery application but there is very limited information on the physicochemical properties of these ionic liquids in the literature.

Given the availability of ionic conductivity data for ionic liquids belonging to a large variety of different cation types, it is conceivable to find an ionic liquid with high ionic conductivity, if an accurate structure-property relationship is uncovered. With this objective, the present article focuses on developing machine learning models capable of predicting ionic conductivity covering various cation families and anions with high accuracy. Additionally, important features contributing to the ionic conductivity have been identified using shapely additive (SHAP) analysis technique. The insight is used to develop a classification model to categorize cations that are likely to yield ionic liquids, with a given anion, into high/low ionic conductivity. Lastly, ionic conductivity for all possible pairings of the cation and anion are predicted to identify ionic liquids possessing high ionic conductivity.

2 Methodology

2.1 Data collection and processing

In this study, we developed machine learning models trained on experimental ionic conductivity data primarily obtained from NIST ILThermo Database[56,57]. We supplemented the data extracted from the ILThermo database with data collected from various sources found in literature[58–80]. This led to a total of 4786 data points covering ten different cation types as seen in Figure S1. Data download, data cleaning, duplicate removal, and conversion of chemical structures to SMILES convention followed a similar approach outlined in our previous study[30]. The state property filter was set between 95-110 kPa, eliminating some of the very high-pressure data, while no restrictions were imposed on the temperature. The temperatures and pressure were selected considering that ionic liquid-based batteries would

be operated over a wide range of temperatures and close to atmospheric pressure. The final dataset contained 2869 data points, 397 unique ionic liquids, 214 unique cations, and 68 unique anions ranging from 238 K to 472 K covering ionic conductivity from 10^{-5} S/m to 19.3 S/m, spanning six orders of magnitude.

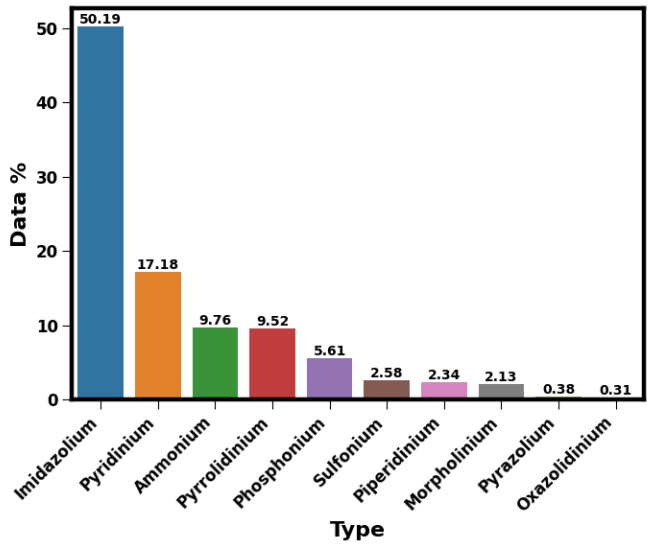

Fig. 1 Experimental ionic conductivity data distribution % by cation type for the model development data set after data cleaning. The percentage for each of the cation family is calculated using the number of data points collected in Table S1.

The percentage distribution of the individual cation type in the model development data set is depicted in Figure 1. As expected, half of the data belongs to the imidazolium family as it is one of the widely studied cations. Additional 40% of the data points are contributed by pyridinium, ammonium, pyrrolidinium, and phosphonium cations. One concern with such skewed data distribution would be the bias in prediction towards the imidazolium data set due to the relative abundance of the ionic conductivity data for this cation family. In a later section, we discuss our approach to systematically evaluate the model's prediction by cation type to evaluate such bias in prediction.

2.2 Feature generation and processing

Features for the cations and anions are generated using open-source cheminformatics RDKit package[81] that produced 196 unique features each for cation and anion. Temperature and pressure were included as additional features which led to a total count of features to 394. Some of the features, however, were not essential for model development as they were assigned a value of zero for all the cations and anions. Besides, a high-dimensional feature space could lead to overfitting of the data, resulting in a poor performance of the model for test data[82]. To avoid such issues, we first reduced the number of features by eliminating features exhibiting high correlations. A further reduction in the dimensionality of the feature space was achieved through the Least absolute shrinkage and selection operator (LASSO)[83,84] algorithm. Lasso is a regularization technique that is used to shrink the dimensionality of the feature space by adding a penalty parameter λ to the minimization function that denotes the amount of feature shrinkage (eq. 1). Larger values of λ parameter lead to the coefficients of features with lower importance to zero, thereby reducing the number of features necessary for a model; the minimization function is recovered for $\lambda = 0$.

$$\text{Obj} = \sum_{i=1}^{n}(y_i - \sum_{j} x_{ij} w_j)^2 + \lambda \sum_{j=1}^{p} |w_j| \quad (1)$$

In eq. 1, y_i denotes the ionic conductivity value for the i^{th} observation, x_{ij} refers to the corresponding value of the j^{th} feature and w_j signifies the weight of the feature. The hyperparameter λ was determined using 5-fold cross-validation (CV) technique by fitting a linear regression model with a $\log\lambda$ in the range of [-6, 50]. Based on the CV, the optimum value ($\log\lambda = -5$) helped reduce the number of features to 51 cations, 47 anion features leading to a total of 100 features including temperature and pressure.

2.3 Model Development

For the model development, the data set was split into 90% training set, while the remaining 10% of the data was set aside as a test case. The input features and the ionic liquid conductivity data were normalized to fall within the range of [0,1] using MinMax scaling implemented in Scikit-learn[85]. The ionic conductivity values were represented on a $\log 10$ scale before scaling as the values spanned six orders of magnitude. Three different models (Figure 2) were developed to correlate the ionic conductivity data: multiple linear regression (MLR), random forest (RF), and extreme gradient boosting (XGBoost). The rationale for choosing these models compared to the widely popular neural network was to offer insights into the importance of individual features.

2.4 Multiple Linear Regression

Correlation of ionic conductivity is first attempted using the multiple linear regression (MLR) model as it is the simplest form of regression method. In an MLR model, the structure-property relationship is expressed as a linear combination of features x_i (eq. 2)

$$y_p = b + w_1 x_1 + w_1 x_2 .... + w_n x_p. \quad (2)$$

where b is the bias in the model, and w_i corresponds to the weight of feature x_i, which are determined by minimizing the least square error between the model prediction and the labels. Note that the model contains p features.

2.5 Random Forest

Random forest (RF) is a supervised machine learning method based on ensemble learning technique similar to decision-tree (DT) method[85]. However, unlike the DT method, for which outputs are generated using a single tree, RF methodology consists of multiple decision trees, which are generated in parallel, in an

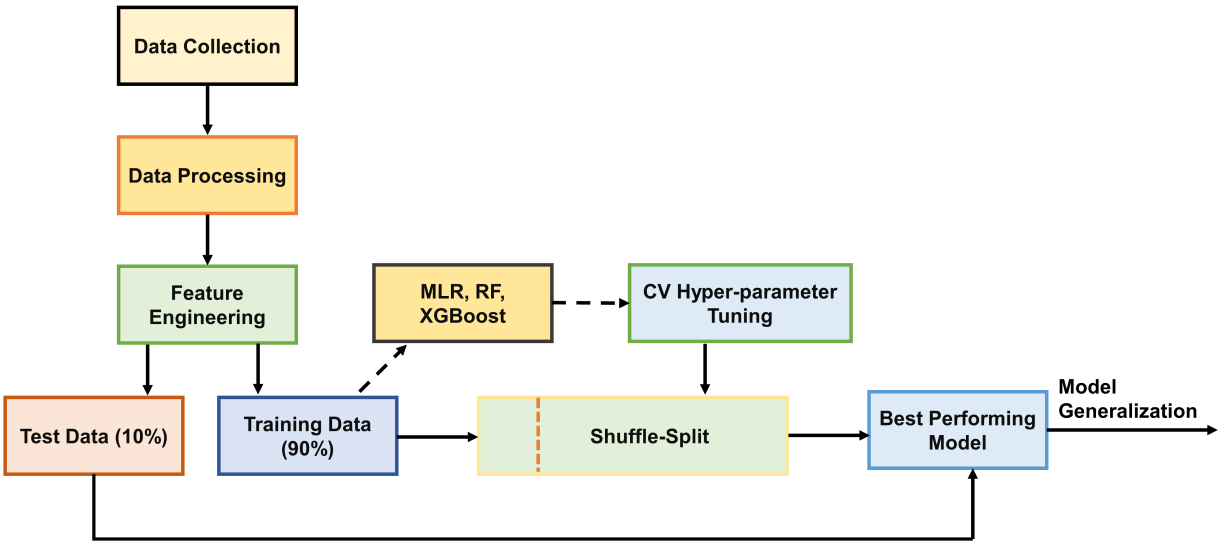

Fig. 2 Workflow for developing machine learning models for correlating ionic conductivity of ionic liquids.

effort to reduce the possibility of overfitting and minimizing any bias towards feature selection. A sample from the training set is drawn at random with replacement to initiate a given tree. The final prediction is the ensemble average of the outputs predicted by individual trees. The number of trees or estimators and the depth of the trees are hyper-parameters of the model, which were determined using randomized cross-validation (RandomizedCV)[85]. In this approach, a random combination of hyperparameters is evaluated using 5-fold CV. Figure S2 provides the grids used for the number of trees and estimators.

2.6 XGBoost

Extreme gradient boosting (XGBoost) is a decision tree-based ensemble method similar to RF that uses a gradient boosting algorithm[86]. However, unlike RF, where individual trees are formed in parallel, XGBoost consists of a series of trees built iteratively. The model starts with weak learners that are intentionally added to make a significant error which gets added to the loss function of the subsequent tree using a gradient descent algorithm. The objective of the XGBoost function is to minimize the loss as each tree is added until the accuracy no longer improves. The hyperparameters for XGBoost are determined using RandomizedCV using 5-fold CV[85]. The final set of hyperparameters for the XGBoost model is listed in the supporting information (Figure S3).

2.7 Cross Validation and Model Evaluation

In most cases, performance evaluation of a model on the test data is enough to assess the ability of the model to generalize on out-of-bag samples. However, it is not always guaranteed that the models will generalize, especially when there is an overabundance of one or more types of data. For instance, in this study, approximately half of the data is represented by the imidazolium family (Figure 1). Even with a proper random shuffle of train and test split, most of the data in the training set and test set could belong to the imidazolium cation, leading to high accuracy on the training set and test set, which might not reflect the model's inherent ability to generalize beyond the imidazolium-based ionic liquids. Another challenge with the present data is that there may be ionic liquids for which there is data at multiple temperatures, while for others, the ionic liquid conductivity is reported at only one temperature; the cation 1-ethyl-3-methylimidazolium or the anion bis(trifluoromethylsulfonyl)imide anion are such examples as they are usually studied over a wide range of temperatures. Despite the train/test split, most of the data could be for the same ionic liquid, simply reflecting the ability of the model to scale ionic conductivity as a function of temperature, not truly reflecting the underlying structure-property relationship.

For this work, we initially started with a 90:10 train/test split, where the test set data was never exposed during the model development. Next, we employed the shuffle-split technique by dividing the training set data into 90:10 for model development and validation data, respectively, repeated 100 times. The extensive sampling is expected to minimize any bias in the split for training and validation data and is likely to provide uncertainty in the prediction based on the data split. Performance metrics for the training and validation set were evaluated at each instance. The model with the best performance on the validation set was selected as the model for test set evaluation.

The predictive capability of the model was also determined for each of the cation types to examine overfitting of the model towards a certain class of the cation type. Finally, the performance of the model was evaluated on a separate external test set consisting of 30 ionic liquids gathered from a literature review[60,62,68,78,80,87–89]. The external test set contained unique ionic liquid combinations that are not present in the model

data set. Here, unique ionic liquid combination refers to ionic liquids with given cation and anion pairs that the model did not encounter during the training phase. However, these cations and anions were present in the dataset paired with a different anion/cation. Furthermore, a few of the ionic liquids in this test case are novel cation family types that did not belong to any of the ten cation families for which models were developed. As the chemical structures of the cations in these ionic liquids resemble those in the model dataset, such performance evaluation would be informative to understand the extent to which these models can be generalized to cations families beyond those studied here.

3 Results and Discussion

3.1 Model Performance Metrics

In this work, we evaluate the correlation of ionic conductivity for ten different cation types, 214 unique cations, 68 unique anions with ionic conductivity data over a temperature range of 238-472 K using linear and non-linear machine learning methods. The accuracy and robustness of these models are evaluated using three different performance metrics: correlation coefficient (R^2), root-mean-squared-error (RMSE) and mean absolute error (MAE). The average and standard deviation of the performance metrics for the 100 shuffle-split is reported in Table 1. The model with the best performance on the validation becomes the model of choice for test set evaluation.

Among the three models, MLR has the lowest accuracy in correlating ionic conductivity compared to the other two models with low R^2 and high RMSE/MAE for training, validation, and test set. Increasing the complexity of the model that takes into account non-linear behavior in the model drastically increases the model performance as seen with the RF method, implying a non-linear correlation between features and ionic conductivity. A further improvement in the performance metrics can be observed with the XGBoost model, as the model is designed to iteratively learn and correct the errors incurred in the previous steps.

Correlations plots for MLR, RF, XGBoost are provided in the supporting information (Figure S4, S5 and S6). Based on the trends in the figures, it is readily apparent that the non-linear models significantly outperform the MLR model, similar to other studies that have examined correlation of ionic liquid properties using linear and non-linear approaches[90,91].

We also examined the overall correlation coefficient (R^2) for the XGBoost and RF model for each of the cation families, the results of which are presented in Figure S7. It can be observed that the performance of the XGBoost model is somewhat independent of the type of the cation family, while the RF model is more sensitive to the type of cation and the corresponding number of data points. For example, the R^2 value drops to 0.55 for morpholinium cation, despite being present in the training set, while the R^2 predicted using the XGBoost method is 0.9. Similarly, the R^2 value obtained with the RF model for oxazolidinium cations drops below 0.90, whereas the XGBoost model again yields R^2 values $\sim$0.9. For all other cations types, the correlation coefficients are nearly perfect as deduced from the XGBoost method, while those calculated from the RF model are lower, which shows that the XGBoost model can be accurate even when the data is limited (Figure 1). Based on the performance metrics by cation type, the XGBoost model is chosen as the choice of model for further prediction as it outperforms the RF model for individual cation types, which is essential for unique ionic liquid prediction discussed in the later section.

We further tested the predictive capability of the XGBoost model for the external data, which were neither part of the training data or test data, consisting of 30 data points collected from the literature. This data set included 27 unique ionic liquids with a temperature range between 293-323.15 K and ionic conductivity range of 0.06-1.68 S/m. The XGBoost model obtained an R^2 of 0.80, RMSE of 0.20 S/m, and MAE of 0.14 S/m for this external test set compared to experimental data. The entire prediction on this external test case using XGBoost along with experimental data, the source, along with schematics of the cations and anions are provided in Table S2 (unique ionic liquid combinations), Table S3 (cations structurally similar to those on which the model was trained) and Table S4 (substituted imidazolium-based cations).

3.2 Model Interpretation

A significant advantage of ensemble-based models over 'black box' models such as neural network is the easy interpretability of the feature importance. This insight can be valuable for developing design heuristics for the search and development of new cations with high ionic conductivity. In this work, we employed the Shapley additive explanations (SHAP)[92] method, which provides a reliable way to explain the importance of features and the model decision making[93,94]. As shown in Figure 3, the SHAP analysis ranks the features in terms of their importance, while the SHAP value indicates how varying a certain feature is likely to affect the output, ionic conductivity in the present case. A negative SHAP value suggests a lowering of conductivity, while a positive value implies an increase in conductivity.

Figure 3 presents nine features deemed most important for ionic liquid predictions. Out of these nine features, six features correspond to the cations, while two features are linked to the anions and one for temperature. The second most important feature in the list is the IPC descriptor for the cation that takes into account the information content of the molecule, such as the number of atoms through a graph representation[95]. Based on the SHAP value, it is clear that a high value of IPC (denoted by the red color) negatively impacts the output. As the IPC descriptor is related to the content of the molecule, higher values of the feature delineates bulky cations, for example those containing long alkyl chains, slowing down the dynamics of the system[96,97]. Furthermore, Figure S11 (a) also shows the relation between the IPC descriptor of the cation and experimental ionic conductivity at 298.15 K, demonstrating the effect of the descriptor on the ionic conductivity; there is a general trend of decreasing ionic

Table 1 Average and standard deviation of the performance metrics for the training and validation set using MLR, RF, XGBoost model. RMSE is the root mean squared error, MAE is the mean absolute error, and R^2 is the correlation coefficient between experiment and predicted data. Shuffle-Split indicates random data shuffle into 100 different training/validation splits. The model with the best performance on the validation set during shuffle-split becomes the final choice of model for test set evaluation. Note: The RMSE and MAE values are for ionic conductivity in the $log10$ scale

		Shuffle-Split			Best Performing		
Model	Data Set	R^2	RMSE	MAE	R^2	RMSE	MAE
MLR	Training	0.877±0.002	0.260±0.003	0.173±0.002	0.873	0.265	0.175
	Validation	0.867±0.023	0.268±0.029	0.180±0.011	0.914	0.213	0.159
	Test	–	–	–	0.853	0.322	0.204
RF	Training	0.994±0.000	0.059±0.001	0.031±0.000	0.994	0.060	0.031
	Validation	0.956±0.013	0.152±0.026	0.083±0.007	0.977	0.116	0.074
	Test	–	–	–	0.963	0.161	0.087
XGBoost	Training	0.999±0.000	0.020±0.001	0.012±0.000	0.999	0.021	0.012
	Validation	0.977±0.011	0.109±0.026	0.050±0.006	0.993	0.061	0.039
	Test	–	–	–	0.987	0.094	0.047

conductivity with IPC as revealed by the SHAP analysis. The next feature Chi0, that contributes to the ionic conductivity is also related to cations, capturing the nature of molecular connectivity[98]. The influence of Chi0 is similar to that identified for IPC in that it is negatively correlated to the ionic conductivity as confirmed in Figure S11 (b). MaxAbsEstateIndex is the absolute maximum of the electronic state index for the cation. The Balabanj descriptor is a topological connectivity descriptor that takes in account the structure complexity of the cations[99]. Given that the model contains a wide range of cation types, it's easy to see why this descriptor is so important for ionic conductivity. The BertzCT descriptor is a topological descriptor that relates to the complexity of the molecule through graph theory[100]. The descriptor is sum of two quantities: complexity of the atoms and complexity of the connectivity. Lastly, VSA_EState8 descriptor is the sum of the electrotopological state index of an atom with van der Waals surface area between 6.45 - 7.0[81]. The electropological descriptor encodes both the electronic and topological state of the cation. More on the relation between the descriptors and experimental ionic conductivity is depicted in Figures S11 (c)-(f).

As for the two anions descriptors seen from Figure 3, VSA_EState2_a is the sum of the electrotopological state index of an atom with van der Waals surface area between 4.78 - 5.0[81]. The electropological descriptor encodes both the electronic and topological state of the anion[101]. The electronic state here refers to the electron distribution of the atoms in a molecule. Next, MaxAbsPartialCharge_a descriptor stands for the maximum absolute partial charge of the molecule calculated using the Gasteiger partial charge method based on electronegativity of the atoms in the molecule[102]. In the experimental data set, the highest value for this descriptor is for the halogen anions with a maximum partial charge of 1.0, followed by anions based on the phosphorous atom, oxygen-based anions, and cyano-based anions. The descriptor relation to ionic conductivity can be explained through the SHAP feature importance insight as the cyano group has the lowest MaxAbsPartialCharge_a, which results in higher ionic conductivity. In contrast, the halogen, phosphorous, and oxygen-based anions have the highest MaxAbsPartialCharge_a reducing ionic conductivity. Relation between these descriptors and ionic conductivity can be seen from Figure S11 (g) and (h).

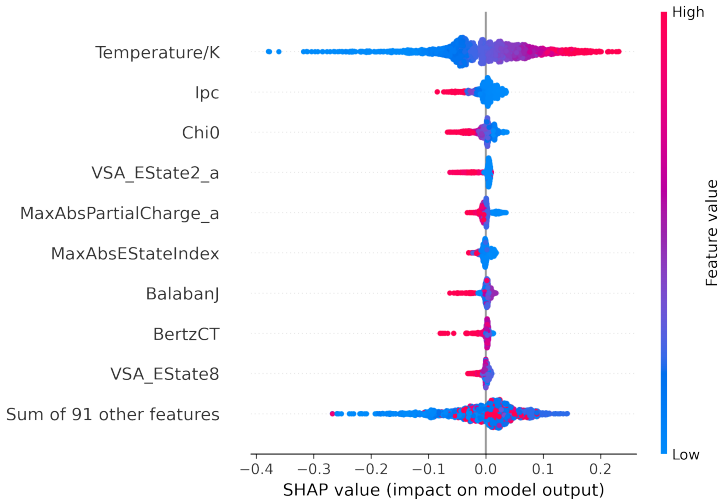

Fig. 3 SHAP feature importance for the training set data. Features ending with '_a ' indicates features for anions.

Based on the SHAP analysis, it is clear that some of the features have a very high influence on the ionic conductivity compared to the rest. To examine how important these features are with respect to ionic conductivity, we also attempted to build a small-scale decision tree-based classification model to leverage insights generated from the SHAP analysis. The primary objective here is to determine the accuracy of such a model by using a few selected features as inputs for the model's development. Further details on the classification model development and results can be found in the supporting information. The classification model is able to classify ionic liquids in the high/low ionic conductivity categories with 98% accuracy for the training set, 92% for the test set, and 63% for the external test set. The accuracy is very high considering that the model is built only with six descriptors.

3.3 Unique Ionic Liquids

As the XGBoost model is rigorously cross-validated for the test case and an external test case, we proceeded to combine all the unique cations (214) and anions (68), significantly expanding the pool of ionic liquids from mere 337 ionic liquids to a staggering 14,552 unique ionic liquids. Although impressive, it is important to note that not all the ionic liquids generated from the combination may exist in a liquid state at 298.15 K, necessitating a separate model to estimate the melting point of these ionic liquids. The XGBoost predictions for 14,552 unique ionic liquids and experimental measurement for 337 ionic liquids at 298 K are provided in Figure 4 for different cation families. As can be observed from the figure, the pyrrolidinium cation type has the highest experimental ionic conductivity followed by imidazolium, ammonium, pyrazolium, and pyridinium. The model predictions accurately capture this trend.

The current conventional electrolyte LP30 found in Li-ion batteries consists of 1 M $LiPF_6$ in 1:1 ethyl carbonate (EC) and dimethylcarbonate (DMC) mixture that is known to have an ionic conductivity of 1.26 at 298.15 K[103,104]. Thus for ionic liquids to be considered as a potential electrolyte additive to replace LP30, the ionic conductivity target should at least be close to 2.0 S/m as the addition of Li salts dramatically reduces the ionic conductivity and increases viscosity by 30-50%[88,105,106]. In our model development database, there are only five ionic liquids with ionic conductivity greater than 2.0 S/m, which is now expanded to 21 ionic liquids using the unique ionic liquid combination. This is possible as some of the cations for the ionic liquids with experimental data higher than 2.0 S/m when combined with other cyano-based anions present in the data set lead to more ionic liquids with high ionic conductivity. Breaking the unique ionic liquid combination analysis by cation type, there are just two cations beyond 2.0 S/m present in the experimental data set for the pyrrolidinium cations. Using the model, this region of space is now expanded to seven unique pyrrolidinium ionic liquids. Similarly, the number of ionic liquids beyond 2.0 S/m has grown from four to nine ionic liquids for imidazolium cations. For the piperidinium experimentally, there is no data beyond 0.5 S/m at 298.15 K. That now has expanded to a large number of them crossing the 1.0 S/m as the piperidinium cations are paired with some other anions, mainly cyano based anions, as they can push ionic liquids to have ionic conductivity. Based on the cations and anion combinations resulting from available experimental data, our analysis suggests that there are no high ionic conductivity ionic liquids that can be formed using oxazolidinium, phosphonium, or morpholinium cations.

4 Conclusion

In search of ionic liquids with high ionic conductivity for battery application, we developed three different machine learning models to correlate ionic conductivity of ten different cation types covering a temperature range of 238-472 K. It was found the multiple linear regression model was least accurate, while the non-linear model XGBoost performed the best. Although the accuracy of the model developed using RF methodology was similar to that for the XGBoost model, a degradation in its predictive capability was noted for cation families that represented a very small portion of the overall data set. On the other hand, the XGBoost model retained its high accuracy across all the cation families.

Feature importance based on SHAP analysis showed temperature, six cation features, and two anions features to have the most influence on ionic conductivity output. The insight gained from the SHAP analysis was used to develop a decision tree-based model containing only six cation features to classify ionic liquids containing $[NTf_2]^-$ anion into two categories: high ionic conductivity and low ionic conductivity. The model showed a high accuracy, successfully classifying 92% of the ionic liquids from the test set, demonstrating the usefulness of the SHAP analysis.

Lastly, all the unique cations and anions in the database were combined to dramatically expand the chemical space of ionic liquids as demonstrated by the increase in the number of ionic liquids from 337 to 14,552 unique ionic liquids. The model predictions hint at 21 ionic liquids possessing ionic conductivity greater than 2.0 S/m at 298.15 K. We envision that the large database of ionic liquid conductivity predictions can serve as a roadmap for future computational and experimental efforts in search for ionic liquids with very high ionic conductivity suitable for battery application as electrolytes.

Author Contributions

We strongly encourage authors to include author contributions and recommend using CRediT for standardised contribution descriptions. Please refer to our general author guidelines for more information about authorship.

Conflicts of interest

In accordance with our policy on Conflicts of interest please ensure that a conflicts of interest statement is included in your manuscript here. Please note that this statement is required for all submitted manuscripts. If no conflicts exist, please state that "There are no conflicts to declare".

Acknowledgements

Authors acknowledge funding for this work from the National Science Foundation grant CBET-1706978.

Notes and references

1 A. Berthod, M. Ruiz-Angel and S. Carda-Broch, *Journal of Chromatography A*, 2008, **1184**, 6–18.
2 A. S. Paluch and P. Dhakal, *ChemEngineering*, 2018, **2**, 54.
3 J. Peng and Y. Deng, *New Journal of Chemistry*, 2001, **25**, 639–641.
4 D. Zhao, M. Wu, Y. Kou and E. Min, *Catalysis today*, 2002, **74**, 157–189.
5 E. D. Bates, R. D. Mayton, I. Ntai and J. H. Davis, *Journal of the American Chemical Society*, 2002, **124**, 926–927.

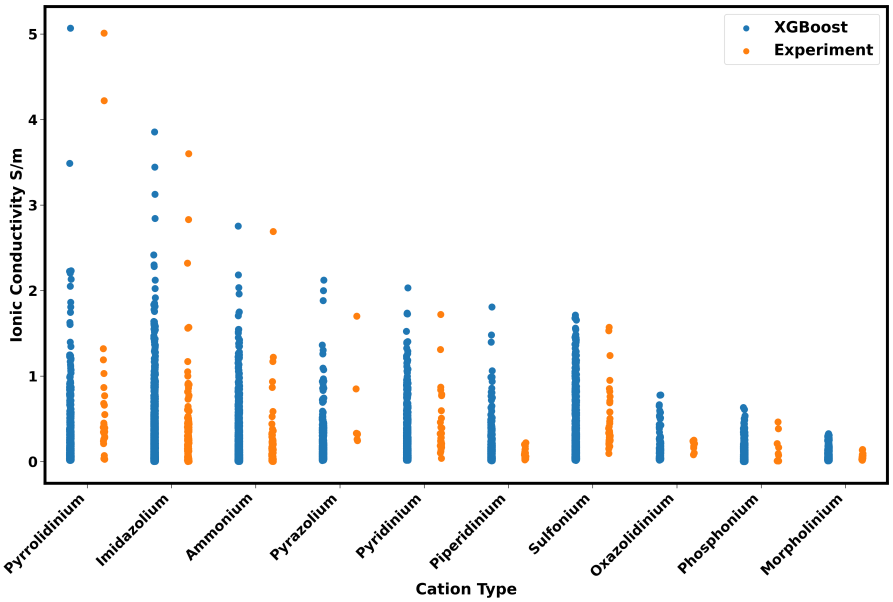

Fig. 4 Categorical data of unique ionic liquids based on cation family type at 298.15 K.

6 J. D. Figueroa, T. Fout, S. Plasynski, H. McIlvried and R. D. Srivastava, *International journal of greenhouse gas control*, 2008, **2**, 9–20.

7 A. Lewandowski and A. Świderska-Mocek, *Journal of Power Sources*, 2009, **194**, 601–609.

8 M. Galiński, A. Lewandowski and I. Stępniak, *Electrochimica acta*, 2006, **51**, 5567–5580.

9 C. Xu and Z. Cheng, *Processes*, 2021, **9**, 337.

10 D. Chand, M. Wilk-Kozubek, V. Smetana and A.-V. Mudring, *ACS sustainable chemistry & engineering*, 2019, **7**, 15995–16006.

11 L. Hu and K. Xu, *Proceedings of the National Academy of Sciences*, 2014, **111**, 3205–3206.

12 T. Yim, M.-S. Kwon, J. Mun and K. T. Lee, *Israel Journal of Chemistry*, 2015, **55**, 586–598.

13 C. Arbizzani, G. Gabrielli and M. Mastragostino, *Journal of Power Sources*, 2011, **196**, 4801–4805.

14 A. Andersson and K. Edström, *Journal of the Electrochemical Society*, 2001, **148**, A1100.

15 S. E. Sloop, J. K. Pugh, S. Wang, J. Kerr and K. Kinoshita, *Electrochemical and Solid State Letters*, 2001, **4**, A42.

16 J. Kalhoff, G. G. Eshetu, D. Bresser and S. Passerini, *ChemSusChem*, 2015, **8**, 2154–2175.

17 N. V. Plechkova and K. R. Seddon, *Chemical Society Reviews*, 2008, **37**, 123–150.

18 E. I. Izgorodina, R. Maganti, V. Armel, P. M. Dean, J. M. Pringle, K. R. Seddon and D. R. MacFarlane, *The Journal of Physical Chemistry B*, 2011, **115**, 14688–14697.

19 Y. Zhang and E. J. Maginn, *The journal of physical chemistry letters*, 2015, **6**, 700–705.

20 H. Liu and E. Maginn, *The Journal of chemical physics*, 2011, **135**, 124507.

21 O. Borodin and G. D. Smith, *The Journal of Physical Chemistry B*, 2006, **110**, 11481–11490.

22 S. U. Lee, J. Jung and Y.-K. Han, *Chemical physics letters*, 2005, **406**, 332–340.

23 Z. Li, G. D. Smith and D. Bedrov, *The Journal of Physical Chemistry B*, 2012, **116**, 12801–12809.

24 V. Zeindlhofer, L. Zehetner, W. Paschinger, A. Bismarck and C. Schroeder, *Journal of Molecular Liquids*, 2019, **288**, 110993.

25 A. T. Nasrabadi and L. D. Gelb, *The Journal of Physical Chemistry B*, 2017, **121**, 1908–1921.

26 T. D. N. Reddy and B. S. Mallik, *The Journal of Physical Chemistry B*, 2020, **124**, 4960–4974.

27 S. Koutsoukos, F. Philippi, F. Malaret and T. Welton, *Chemical science*, 2021, **12**, 6820–6843.

28 Z. Song, H. Shi, X. Zhang and T. Zhou, *Chemical Engineering Science*, 2020, **223**, 115752.

29 A. Shafiei, M. A. Ahmadi, S. H. Zaheri, A. Baghban, A. Amirfakhrian and R. Soleimani, *The Journal of Supercritical Fluids*, 2014, **95**, 525–534.

30 P. Dhakal and J. K. Shah, *Fluid Phase Equilibria*, 2021, **549**, 113208.

31 W. Beckner and J. Pfaendtner, *Journal of chemical information and modeling*, 2019, **59**, 2617–2625.

32 W. Beckner, C. Ashraf, J. Lee, D. A. Beck and J. Pfaendtner, *The Journal of Physical Chemistry B*, 2020, **124**, 8347–8357.

33 S. Lerch and T. Strassner, *Chemistry–A European Journal*, 2019, **25**, 16251–16256.

34 I. Bandrés, D. F. Montaño, I. Gascón, P. Cea and C. Lafuente, *Electrochimica acta*, 2010, **55**, 2252–2257.

35 Q.-S. Liu, P.-P. Li, U. Welz-Biermann, J. Chen and X.-X. Liu, *The Journal of Chemical Thermodynamics*, 2013, **66**, 88–94.

36 Q.-G. Zhang, Y. Wei, S.-S. Sun, C. Wang, M. Yang, Q.-S. Liu and Y.-A. Gao, *Journal of Chemical & Engineering Data*, 2012,

57, 2185–2190.

37 S. P. Ventura, A. M. Gonçalves, T. Sintra, J. L. Pereira, F. Gonçalves and J. A. Coutinho, *Ecotoxicology*, 2013, **22**, 1–12.

38 I. Mena, E. Diaz, J. Palomar, J. Rodriguez and A. Mohedano, *Chemosphere*, 2020, **240**, 124947.

39 N. Plylahan, M. Kerner, D.-H. Lim, A. Matic and P. Johansson, *Electrochimica Acta*, 2016, **216**, 24–34.

40 S. Fang, Z. Zhang, Y. Jin, L. Yang, S.-i. Hirano, K. Tachibana and S. Katayama, *Journal of Power Sources*, 2011, **196**, 5637–5644.

41 M. H. Ibrahim, M. Hayyan, M. A. Hashim, A. Hayyan and M. K. Hadj-Kali, *Fluid Phase Equilibria*, 2016, **427**, 18–26.

42 N. Sanchez-Ramirez, B. D. Assresahegn, D. Belanger and R. M. Torresi, *Journal of Chemical & Engineering Data*, 2017, **62**, 3437–3444.

43 R. Zarrougui, N. Raouafi and D. Lemordant, *Journal of Chemical & Engineering Data*, 2014, **59**, 1193–1201.

44 J. Weng, C. Wang, H. Li and Y. Wang, *Green Chemistry*, 2006, **8**, 96–99.

45 K. J. Fraser and D. R. MacFarlane, *Australian journal of chemistry*, 2009, **62**, 309–321.

46 A. J. Rennie, V. L. Martins, R. M. Torresi and P. J. Hall, *The Journal of Physical Chemistry C*, 2015, **119**, 23865–23874.

47 M. P. Mousavi, B. E. Wilson, S. Kashefolgheta, E. L. Anderson, S. He, P. Buíhlmann and A. Stein, *ACS applied materials & interfaces*, 2016, **8**, 3396–3406.

48 P. J. Carvalho, S. P. Ventura, M. L. Batista, B. Schröder, F. Gonçalves, J. Esperança, F. Mutelet and J. A. Coutinho, *The Journal of chemical physics*, 2014, **140**, 064505.

49 A. Bhattacharjee, A. Luís, J. H. Santos, J. A. Lopes-da Silva, M. G. Freire, P. J. Carvalho and J. A. Coutinho, *Fluid phase equilibria*, 2014, **381**, 36–45.

50 C.-P. Lee, J.-D. Peng, D. Velayutham, J. Chang, P.-W. Chen, V. Suryanarayanan and K.-C. Ho, *Electrochimica Acta*, 2013, **114**, 303–308.

51 A. M. Sampaio, G. F. L. Pereira, M. Salanne and L. J. A. Siqueira, *Electrochimica Acta*, 2020, **364**, 137181.

52 M. A. Navarra, K. Fujimura, M. Sgambetterra, A. Tsurumaki, S. Panero, N. Nakamura, H. Ohno and B. Scrosati, *ChemSusChem*, 2017, **10**, 2496–2504.

53 S. Shen, S. Fang, L. Qu, D. Luo, L. Yang and S.-i. Hirano, *RSC advances*, 2015, **5**, 93888–93899.

54 M. Chai, Y. Jin, S. Fang, L. Yang, S.-i. Hirano and K. Tachibana, *Electrochimica acta*, 2012, **66**, 67–74.

55 C. S. Kang, R. Yunis, H. Zhu, C. M. Doherty, O. E. Hutt and J. M. Pringle, *Materials Chemistry Frontiers*, 2021, **5**, 6014–6026.

56 Q. Dong, C. D. Muzny, A. Kazakov, V. Diky, J. W. Magee, J. A. Widegren, R. D. Chirico, K. N. Marsh and M. Frenkel, *Journal of Chemical & Engineering Data*, 2007, **52**, 1151–1159.

57 Q. Dong, A. F. Kazakov, C. D. Muzny, R. D. Chirico, J. A. Widegren, V. Diky, J. W. Magee, K. N. Marsh and M. D. Frenkel, *Ionic Liquids Database (ILThermo)*, technical report, 2006.

58 R. R. Hawker, R. S. Haines and J. B. Harper, *Targets Heterocycl. Syst. Prop*, 2015, **18**, 141–213.

59 Y. Yoshida, O. Baba and G. Saito, *The Journal of Physical Chemistry B*, 2007, **111**, 4742–4749.

60 J. Zhang, S. Fang, L. Qu, Y. Jin, L. Yang and S.-i. Hirano, *Industrial & Engineering Chemistry Research*, 2014, **53**, 16633–16643.

61 T.-Y. Wu, S.-G. Su, H. P. Wang, Y.-C. Lin, S.-T. Gung, M.-W. Lin and I.-W. Sun, *Electrochimica Acta*, 2011, **56**, 3209–3218.

62 B. Baek, S. Lee and C. Jung, *Int. J. Electrochem. Sci*, 2011, **6**, 6220–6234.

63 F. Alloin, P. Strobel, J.-C. Leprêtre, L. Cointeaux, C. P. del Valle *et al.*, *Ionics*, 2012, **18**, 817–827.

64 M. L. P. Le, F. Alloin, P. Strobel, J.-C. Leprêtre, C. Peérez del Valle and P. Judeinstein, *The Journal of Physical Chemistry B*, 2010, **114**, 894–903.

65 R. Zarrougui, R. Hachicha, R. Rjab, S. Messaoudi and O. Ghodbane, *RSC advances*, 2018, **8**, 31213–31223.

66 O. Russina, R. Caminiti, A. Triolo, S. Rajamani, B. Melai, A. Bertoli and C. Chiappe, *Journal of molecular liquids*, 2013, **187**, 252–259.

67 S. Seki, T. Kobayashi, N. Serizawa, Y. Kobayashi, K. Takei, H. Miyashiro, K. Hayamizu, S. Tsuzuki, T. Mitsugi, Y. Umebayashi *et al.*, *Journal of Power Sources*, 2010, **195**, 6207–6211.

68 M. Chai, Y. Jin, S. Fang, L. Yang, S.-i. Hirano and K. Tachibana, *Journal of Power Sources*, 2012, **216**, 323–329.

69 C. Chiappe, A. Sanzone, D. Mendola, F. Castiglione, A. Famulari, G. Raos and A. Mele, *The Journal of Physical Chemistry B*, 2013, **117**, 668–676.

70 L. Guo, X. Pan, C. Zhang, M. Wang, M. Cai, X. Fang and S. Dai, *Journal of Molecular Liquids*, 2011, **158**, 75–79.

71 L. Yang, Z. Zhang, X. Gao, H. Zhang and K. Mashita, *Journal of power sources*, 2006, **162**, 614–619.

72 S. Murphy, F. Ivol, A. R. Neale, P. Goodrich, F. Ghamouss, C. Hardacre and J. Jacquemin, *ChemPhysChem*, 2018, **19**, 3226–3236.

73 J. Landmann, J. A. Sprenger, P. T. Hennig, R. Bertermann, M. Grüne, F. Würthner, N. V. Ignat'ev and M. Finze, *Chemistry–A European Journal*, 2018, **24**, 608–623.

74 L. A. Bischoff, M. Drisch, C. Kerpen, P. T. Hennig, J. Landmann, J. A. Sprenger, R. Bertermann, M. Grüne, Q. Yuan, J. Warneke *et al.*, *Chemistry–A European Journal*, 2019, **25**, 3560–3574.

75 R. Zarrougui, R. Hachicha, R. Rjab and O. Ghodbane, *Journal of Molecular Liquids*, 2018, **249**, 795–804.

76 G.-H. Min, T.-e. Yim, H.-Y. Lee, D.-H. Huh, E.-j. Lee, J.-y. Mun, S. M. Oh and Y.-G. Kim, *Bulletin of the Korean Chemical Society*, 2006, **27**, 847–852.

77 M. Hilder, G. Girard, K. Whitbread, S. Zavorine, M. Moser, D. Nucciarone, M. Forsyth, D. MacFarlane and P. Howlett, *Electrochimica Acta*, 2016, **202**, 100–109.

78 A. Tot, Č. Podlipnik, M. Bešter-Rogač, S. Gadžurić and M. Vraneš, *Arabian Journal of Chemistry*, 2020, **13**, 1598–1611.

79 K. Tsunashima, E. Niwa, S. Kodama, M. Sugiya and Y. Ono, *The Journal of Physical Chemistry B*, 2009, **113**, 15870–15874.

80 Z.-B. Zhou, H. Matsumoto and K. Tatsumi, *Chemistry–A European Journal*, 2006, **12**, 2196–2212.

81 `http://www.rdkit.org/`, Accessed: 2020-03-28.

82 D. M. Hawkins, *Journal of chemical information and computer sciences*, 2004, **44**, 1–12.

83 R. Tibshirani, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, **58**, 267–288.

84 W. Beckner, C. M. Mao and J. Pfaendtner, *Molecular Systems Design & Engineering*, 2018, **3**, 253–263.

85 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

86 T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen *et al.*, *R package version 0.4-2*, 2015, **1**, 1–4.

87 T. Kakibe, N. Yoshimoto, M. Egashira and M. Morita, *Electrochemistry communications*, 2010, **12**, 1630–1633.

88 H.-T. Kim, O. M. Kwon, J. Mun, S. M. Oh, T. Yim and Y. G. Kim, *Electrochimica Acta*, 2017, **240**, 267–276.

89 S. Pohlmann, T. Olyschläger, P. Goodrich, J. A. Vicente, J. Jacquemin and A. Balducci, *Journal of Power Sources*, 2015, **273**, 931–936.

90 M. H. Fatemi and P. Izadiyan, *Chemosphere*, 2011, **84**, 553–563.

91 D. Yalcin, T. C. Le, C. J. Drummond and T. L. Greaves, *The Journal of Physical Chemistry B*, 2019, **123**, 4085–4097.

92 S. M. Lundberg and S.-I. Lee, *Advances in neural information processing systems*, 2017, **30**,.

93 Y. Sun, M. Chen, Y. Zhao, Z. Zhu, H. Xing, P. Zhang, X. Zhang and Y. Ding, *Journal of Molecular Liquids*, 2021, **333**, 115970.

94 Y. Ding, M. Chen, C. Guo, P. Zhang and J. Wang, *Journal of Molecular Liquids*, 2021, **326**, 115212.

95 D. Bonchev and N. Trinajstić, *The Journal of Chemical Physics*, 1977, **67**, 4517–4533.

96 L. F. Zubeir, M. A. Rocha, N. Vergadou, W. M. Weggemans, L. D. Peristeras, P. S. Schulz, I. G. Economou and M. C. Kroon, *Physical Chemistry Chemical Physics*, 2016, **18**, 23121–23138.

97 D. Rooney, J. Jacquemin and R. Gardas, *Ionic liquids*, 2009, 185–212.

98 L. H. Hall and L. B. Kier, *Reviews in computational chemistry*, 1991, 367–422.

99 A. T. Balaban, *Chemical physics letters*, 1982, **89**, 399–404.

100 S. H. Bertz, *Journal of the American Chemical Society*, 1981, **103**, 3599–3601.

101 L. B. Kier and L. H. Hall, *Pharmaceutical research*, 1990, **7**, 801–807.

102 J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219–3228.

103 A. Tsurumaki, M. Agostini, R. Poiana, L. Lombardo, E. Lufrano, C. Simari, A. Matic, I. Nicotera, S. Panero and M. A. Navarra, *Electrochimica Acta*, 2019, **316**, 1–7.

104 J. Tarascon and D. Guyomard, *Solid State Ionics*, 1994, **69**, 293–305.

105 J. Asenbauer, N. B. Hassen, B. D. McCloskey and J. M. Prausnitz, *Electrochimica Acta*, 2017, **247**, 1038–1043.

106 M. J. Monteiro, F. F. Bazito, L. J. Siqueira, M. C. Ribeiro and R. M. Torresi, *The Journal of Physical Chemistry B*, 2008, **112**, 2102–2109.