The Journal of Chemical Physics

# The effect of descriptor choice in machine learning models for ionic liquid melting point prediction

Cite as: J. Chem. Phys. **153**, 104101 (2020); https://doi.org/10.1063/5.0016289
Submitted: 04 June 2020 • Accepted: 13 August 2020 • Published Online: 08 September 2020

Kaycee Low, Rika Kobayashi and Ekaterina I. Izgorodina

COLLECTIONS

Paper published as part of the special topic on Machine Learning Meets Chemical Physics

View Online Export Citation CrossMark

## ARTICLES YOU MAY BE INTERESTED IN

Machine learning for interatomic potential models
The Journal of Chemical Physics **152**, 050902 (2020); https://doi.org/10.1063/1.5126336

TRAVIS—A free analyzer for trajectories from molecular simulation
The Journal of Chemical Physics **152**, 164105 (2020); https://doi.org/10.1063/5.0005078

SchNet – A deep learning architecture for molecules and materials
The Journal of Chemical Physics **148**, 241722 (2018); https://doi.org/10.1063/1.5019779

AIP Publishing

# The effect of descriptor choice in machine learning models for ionic liquid melting point prediction

Cite as: J. Chem. Phys. **153**, 104101 (2020); doi: 10.1063/5.0016289
Submitted: 4 June 2020 • Accepted: 13 August 2020 •
Published Online: 8 September 2020

View Online Export Citation CrossMark

Kaycee Low,[1] Rika Kobayashi,[2] and Ekaterina I. Izgorodina[1,a)]

## AFFILIATIONS

[1] Monash Computational Chemistry Group, Monash University, 17 Rainforest Walk, Clayton, VIC 3800, Australia
[2] ANU Supercomputer Facility, Leonard Huxley Building 56, Mills Road, Canberra, ACT 2601, Australia

**Note:** This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.

[a)] **Author to whom correspondence should be addressed:** katya.pas@monash.edu.
**URL:** http://www.https://mccg.erc.monash.edu

## ABSTRACT

The characterization of an ionic liquid's properties based on structural information is a longstanding goal of computational chemistry, which has received much focus from *ab initio* and molecular dynamics calculations. This work examines kernel ridge regression models built from an experimental dataset of 2212 ionic liquid melting points consisting of diverse ion types. Structural descriptors, which have been shown to predict quantum mechanical properties of small neutral molecules within chemical accuracy, benefit from the addition of first-principles data related to the target property (molecular orbital energy, charge density profile, and interaction energy based on the geometry of a single ion pair) when predicting the melting point of ionic liquids. Out of the two chosen structural descriptors, ECFP4 circular fingerprints and the Coulomb matrix, the addition of molecular orbital energies and all quantum mechanical data to each descriptor, respectively, increases the accuracy of surrogate models for melting point prediction compared to using the structural descriptors alone. The best model, based on ECFP4 and molecular orbital energies, predicts ionic liquid melting points with an average mean absolute error of 29 K and, unlike group contribution methods, which have achieved similar results, is applicable to any type of ionic liquid.

## I. INTRODUCTION

Ionic liquids, which display a wide range of physico-chemical properties due to their unique structure, have incited much research into their characterization for use as new-generation solvents and catalysts. One of the main focuses is on their potential as high-conductivity, low-volatility electrolytes, which requires properties such as low viscosity and low melting temperature for practical applications.[1,2] Considering the number of possible ionic liquids, estimated to be in the trillions, synthesis of every possible cation and anion combination would be a near endless task.[3] The prediction of thermochemical and physical properties of an ionic liquid based on only its cation and anion structure represents one of the great challenges in computational chemistry, with progress being made in the fields of molecular dynamics, *ab initio* MD, and quantum chemical calculations.[4–6] However, the timescale required for these rigorous methods still exceeds what is desirable for high-throughput screening. As an alternative, machine learning algorithms have been making strides in predicting materials properties using a small set of reference calculations only, requiring inputs such as chemical formula descriptors, or, in the case of graph networks, molecular connectivity information.[7,8]

Recent research into chemical machine learning has resulted in significant successes in predicting materials properties, including but not limited to crystal lattice energies,[9] thermal conductivities of perovskites,[10] and melting temperature of solids.[11,12] Ionic liquids present a particularly unique challenge, being a cocktail of interactions including electrostatic repulsion and attraction between like-charged and oppositely charged ions, hydrogen bonding, and London dispersion forces between cations and anions, and alkyl groups on the cation.[13] Experimentally, the range of properties such as density and conductivity exhibited by ionic liquids is wider when compared with that of standard solvents and electrolytes. Thus, the relationship between the individual cation and anion structure and bulk ionic liquid properties may not be straightforward to elucidate via machine learning. Previous work in this field has focused on relating experimental or calculated properties of the ions, such as molecular volume or HOMO–LUMO energies,[14] to observed properties such as viscosity or melting temperature via quantitative structure–property relationship (QSPR) studies.[15,16] Throughout these studies, there has been little focus on the rationale behind the selection of descriptors and effect of the descriptor on the machine learning model. Descriptor choice varies according to the study at hand, and considerations such as time and computational resources. Numerical vectors built using molecular properties have long been utilized in the quantitative structure–activity and structure–property relationship (QSAR and QSPR) fields,[17] though these descriptors can miss often crucial information about the substructure and connectivity of molecules.[18] While incorporating quantum mechanical (QM) information is fairly common, many studies have kept to lower levels of theory and semi-empirical optimizations using methods such as PM6 or PM7.[19,20] To reduce computational cost, the dataset size is frequently limited to several hundred ionic liquids or smaller.[16,21] Other than numerical vector descriptors built from molecular and electronic properties, descriptors can be based on the one-dimensional, two-dimensional, or three-dimensional structure of a molecule: atomic formula; bonding and connectivity; or 3D geometry, respectively. One-dimensional descriptors constructed from the chemical formula, such as some molecular fingerprints, require little effort when inputting these easily computed features such as the presence or absence of a functional group into a bit vector. Including molecular structure information such as size, shape, cyclicity, and symmetry provides a two-dimensional descriptor. Two-dimensional molecular fingerprints encode graph structural information, including a larger radius over two or more bonds, compared to one-dimensional fingerprints.[17] Since one-dimensional and two-dimensional descriptors do not include stereochemical information, recently developed three-dimensional descriptors encode the structure and bonding, e.g., the Coulomb matrix[22] (CM) or the bag of bonds (BoB) model,[23] though these require optimized geometries of the molecules of interest. Three-dimensional descriptors eliminate the need to create hand-engineered vectors and have achieved impressive accuracy when applied to electronic molecular properties such as atomization energy,[24,25] though they have not yet been applied to ionic liquid systems. The application of 3D descriptors in machine learning has mostly been limited to small, neutral molecules, in learning problems where the exact form of the property is known and the aim is to reduce computational expense via machine learning, for example, in predicting the atomization energy of molecules.[22] In this work, we apply such structural descriptors to assess whether or not they are viable for use in surrogate models where the relationship between the input (structure) and the target property (melting point) is not known.

Beyond the choice of descriptor, there still remains the choice of an appropriate machine learning algorithm. As stated by the *no free lunch* theorem,[26] each method performs differently given a different application. As the purpose of this work is not to compare the performance of different models but rather the effect of descriptor choice on such models, we choose kernel ridge regression (KRR), a regression algorithm that has been widely used for predictions of materials properties, including molecular orbital[27] and atomization energies.[28] The relatively few hyperparameters (essentially the choice of kernel function, the kernel width, and a regularization parameter) make KRR ideal for a study attempting to deconvolute the effect of descriptor choice from the machine learning algorithm. While end-to-end learning using deep neural networks, where the representation is learned from the input data, has exploded in popularity in fields where millions of data points are available—such as image search and text-to-speech—in the scientific literature, such volumes of data are rare. A comparison of KRR with a deep neural network would be ideal to evaluate descriptor performance but is not possible at this stage as deep learning suffers in performance with smaller datasets.[29] Hence, this work focuses on the effect of descriptor choice applied to a small dataset of ~2000 experimental ionic liquid melting points using the KRR algorithm.

With a plethora of descriptors and algorithms available for use, this paper rationalizes the effect of each descriptor type on the performance of regression models applied to predicting the melting temperature of ionic liquids from a relatively small (in machine learning terms) experimental dataset. We evaluate the "accuracy" of given descriptors (here, 1D vector-based descriptors are considered to be lower in accuracy than quantum-chemical ones) with respect to their performance in several machine learning algorithms. The outputs of quantum mechanical calculations are used in addition to the structural descriptors, with the aim of overcoming the errors associated with a limited size dataset in machine learning. Although several studies have shown the accuracy of structure-encoding descriptors for large datasets typically with tens of thousands of samples,[18,27] this work utilizes a small dataset, representative of what can be measured experimentally rather than using theoretical data. By adding molecular information via *ab initio* calculations such as frontier orbital energies and interaction energies to simple 1D-chemical and 2D-chemical formula-based descriptors, we consider whether machine learning models applied to a dataset of limited size can make increasingly accurate predictions aided by the information included in high-level calculations. Few have looked at the combination of 3D descriptors with *ab initio* data, though in the same vein Tchagang and Valdés recently showed that the combination of the Coulomb matrix and atomic composition improved prediction of atomization energy by an average of 0.5 kcal/mol for molecules in the QM7 database.[30] Additionally, we investigate whether 3D descriptors encoding structural features that have achieved near chemical accuracy for predicting quantum mechanical properties—a relatively noise-less property—can perform well for the experimental property of melting temperature, which is likely to have noise in the measured value.

## II. DATASET AND MOLECULAR DESCRIPTORS

The dataset used in this work is sourced from a study of Venkatraman *et al.*, which used PM6-calculated descriptors for predicting ionic liquid melting points using machine learning.[31] 2212 experimental values were extracted from over 300 literature sources, with ionic liquids containing various types of ions and a melting point range from 177 K to 632 K. Any experimental measurement such as melting point characterization may have its result influenced by impurities and the measurement technique, and discrepancies in the experimental data are common. Though steps were taken to reduce discrepancies in the dataset, such as taking the most frequent melting point for ionic liquids with multiple reported values, there is still likely some amount of error present (further details on dataset cleaning and refinement can be found within the source paper, Ref. 31). The difference in experimental melting point temperatures measured for a single ionic liquid can vary greatly: by around 10 K–20 K in a good example, or by up to 70 K for a temperamental ionic liquid type.[32] This is obviously an issue for machine learning models as errors impart noise into the data. Hence, choosing an appropriate descriptor that works even when applied to noisy data is essential for a successful ionic liquid predictive model.

Distribution of the target property, melting point, in this dataset is shown in Fig. 1. The mean melting temperature is 361 K, with a standard deviation of 78 K. Further separation of the dataset by cation type reveals that the majority of ionic liquids are imidazolium or ammonium cation based, which is expected as these are the most widely studied cations in the field of ionic liquids. This may cause problems for the ML model, as the dataset is not so chemically diverse. A model trained on mostly ammonium-based and imidazolium-based ionic liquids is likely to make errors when predicting values for the underrepresented ions such as sulfonium and phosphonium, which we investigate further in Sec. IV C.

The simplest types of descriptors utilize only atom types in a molecule (1D), or atom and bonding information (2D) stored either in a vector indicating the presence/absence of a group or as a fingerprint array. The appeal of these descriptors is that the three-dimensional structure of the molecule is not needed; a line drawing or SMILES string is sufficient to extract relevant information. These fingerprints are popular in cheminformatics as they facilitate simple substructure searches. One of the most prominent examples is the extended connectivity fingerprint (ECFP),[33] which has well-established performance in virtual screening, for example, when identifying compounds with similar bioactivity. In this study, we use ECFP4, which encodes functional groups up to four bonds away from the central atom, as this has performed well in benchmark virtual screening studies.[34] The ECFP4 fingerprint was encoded as a 2048-bit vector as implemented in `rdkit`,[35] as this length was found to give better performance compared to all available shorter lengths. A plot of bit vector length vs cross-validated mean average error showing that the 2048-bit vector length produces the lowest error is included in the supplementary material (Fig. S1).

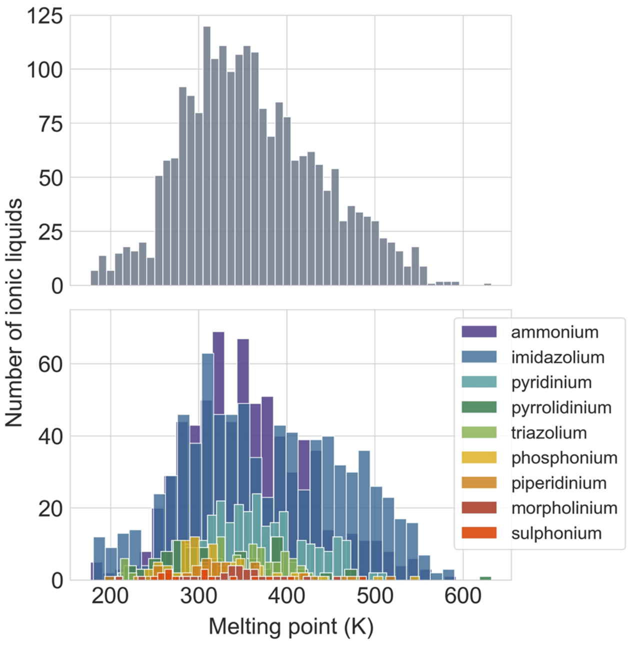

**FIG. 1**. Distribution of the ionic liquid melting point in the entire dataset (top) and separated by cation type (bottom).

3D geometrical descriptors have been shown to lead to good performance for the prediction of thermodynamic and electronic properties.[18,25] The Coulomb matrix[36] (CM) encodes the three-dimensional geometry of the molecule in a square matrix $M$ where the off-diagonal elements correspond to the nuclear repulsion between atoms, whereas diagonal elements represent the electronic potential energy of the free atom,

$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4}, & \mathrm{I} = \mathrm{J} \\ \dfrac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}, & \mathrm{I} \neq \mathrm{J}. \end{cases} \qquad (1)$$

The CM implementation in the `qmmlpack` package[36] was used, with rows of the matrix sorted by their L1-norm. To obtain the geometry required for 3D descriptors as well as *ab initio* values, structures of the cations and anions and their corresponding neutral single ion pairs were optimized at the M06-2X/cc-pVDZ level of theory in Gaussian16[37] using an ethanol solvent, using the conductor-like polarizable continuum (CPCM) solvation model.[38,39] All individual ions were screened for the lowest possible energy conformation. Where more than one interaction site between the cation and the ion was possible, multiple configurations were optimized and the configuration with lowest energy was chosen for further analysis. The supplementary material contains exact details of geometry optimizations. Single point energy calculations to extract interaction energies for use as *ab initio* descriptors were conducted at the SRS-MP2/cc-pVTZ level of theory using counterpoise-corrected HF/cc-pVTZ for the calculation of Hartree–Fock interaction energy (for more details, see Ref. 40). The quantum mechanical descriptors including the interaction energy, density functional theory (DFT)-computed HOMO–LUMO energies and bandgap, and charge density distribution ($\sigma$-profile) were calculated for both the single ions and the ion pair. As the melting point of an ionic liquid is largely the product of inter-molecular and intra-molecular interactions, the quantum mechanical descriptors were chosen to best represent these interactions in numerical form, based on our chemical intuition. The $\sigma$-profile, originally developed by Klamt,[41] is calculated as the distribution $p_i(\sigma) = \frac{A_{i(\sigma)}}{A_i}$, where $A_{i(\sigma)}$ is the area on the surface of the molecule with charge density $\sigma$ and $A_i$ is the total cavity surface area. Sigma profiles have previously been used in a number of ionic liquid machine learning studies as a descriptor and are assumed to have some correlation with properties such as melting point and viscosity.[16,42]

The diversity of the dataset as represented by the CM descriptor for the ionic liquid ion pair is shown using the t-distributed Stochastic Neighbor Embedding (t-SNE) technique in two dimensions. t-SNE is a method of dimensionality reduction that seeks to preserve the distances between neighboring points during projection of a high-dimensional data space to a lower dimensional one.[43] As shown in Fig. 2(a), there are several obvious clusters of ionic liquid, which have been grouped by anion type according to the t-SNE analysis. Interestingly, the t-SNE dimensions seem to correspond mostly to the anion contribution, and the contribution from the cation is not clear (t-SNE plot colored by cation type shows no clear clusters; this plot is included in the supplementary material, Fig. S2). The most common anion is bromide, and as can be seen in Fig. 2(b), these ionic liquids tend to be associated with a higher melting temperature and are grouped around the outer edges of the t-SNE plot. The average melting point for the bromide-based ionic liquids is 423 K, compared to the average melting point of 361 K for the whole dataset. Outside of the clusters belonging to each anion, the inter-cluster distance is relatively large by comparison, indicating that there are structural discontinuities in the dataset as represented by the CM. This is expected for a varied dataset of this size where the majority of cations and anions belong to one or two chemical groups. This may cause problems when predicting the melting point of less frequently found ions as fewer training examples will be present.

For multi-component systems such as ionic liquids, the question of how to best represent the ion pair via descriptor choice arises. Many of the descriptors used in materials science do not include information about which atom belongs to which molecule. One solution is to treat the cation–anion ion pair as a single entity, for example, by constructing a single Coulomb matrix or a single fingerprint based on the single ion pair geometry, and use this as the input to the ML model. Alternatively, separate descriptors can be constructed for the cation and anion individually, resulting in two vectors that are concatenated together to provide a single vector input to the ML model. Both of these options have been investigated in the following.

## III. MACHINE LEARNING MODEL

Due to its computational efficiency and good performance in previous materials prediction studies, we base all machine learning models on the kernel ridge regression (KRR) algorithm.[7,44] In essence, KRR establishes a mapping between the input features and the target property by projecting the inputs and outputs into a high-dimensional space where a relationship is learned. It is assumed that similar compounds, defined by a distance measure in the high-dimensional space, should exhibit similar properties. An estimate for the target property $p$ of a compound $x_{\text{test}}$ not in the training set is obtained as the weighted sum of the $N$ kernel functions placed on each training compound $x_i$,

$$p(x_{\text{test}}) = \sum_{i=1}^{N} \alpha_i K(x_{\text{test}}, x_i). \tag{2}$$

Solutions for the coefficients $a_i$ are obtained from the training process that minimizes the expression

$$\sum_{i=1}^{N} (p^{\text{pred}}(x_i) - p^{\text{true}}(x_i))^2 + \lambda \sum_{i=1}^{N} \alpha_i^2, \tag{3}$$

where $\lambda$ is a regularization parameter that penalizes larger regression coefficients and complex models.[45] We use KRR as implemented in the `scikit-learn` package[46] with a Laplacian kernel, which is defined by

$$K_{\text{Laplacian}}(x_i, x_j) = \exp \frac{-\|x_i - x_j\|_1}{2\sigma^2}, \tag{4}$$

where the distance measure in the exponential term is the L1-norm. $\sigma$ is the kernel width, and the second optimizable hyperparameter in KRR. The hyperparameters $\lambda$ and $\sigma$ were optimized over a square grid from $10^{-10}$ to $10^{-1}$ using fivefold cross-validation. The available data were split into a training set (80%) with the remaining 20% as a test set. The hyperparameters with best performance as determined via cross-validation were then applied to a model, which was trained on the entire training set, and then the test set was applied to make predictions. The final test error values as presented in Sec. IV correspond to these test set predictions. No test data were used in the generation or tuning of the models. To examine the robustness of

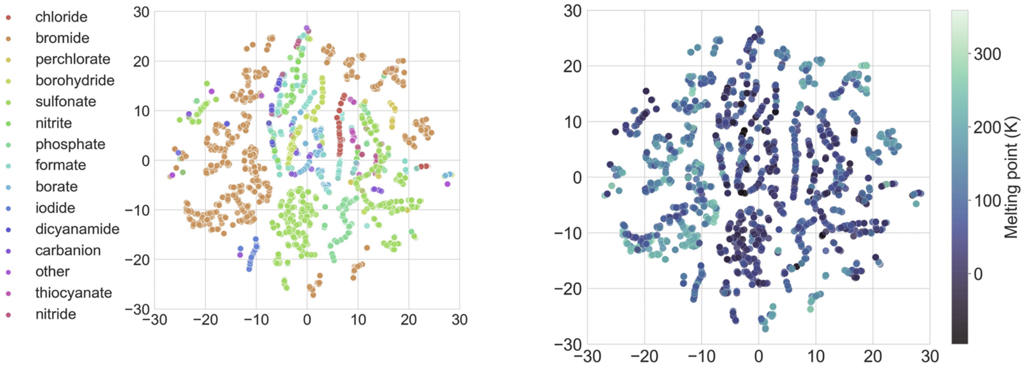

**FIG. 2**. t-SNE projection of the ionic liquid database into two dimensions as represented by the Coulomb matrix descriptor. Points are colored by anion type [(a) "other" corresponds to infrequent anions such as Al⁻ and As⁻] and melting point value (b).

each model and account for variations in prediction associated with randomly splitting the data into training and test sets, the results for each model have been averaged over 100 different randomly selected training/test splits. Hyperparameters for each model can be found in the supplementary material.

## IV. RESULTS AND DISCUSSION

This section presents the results of each KRR model beginning with the structural descriptors (ECFP4 and CM) alone, followed by the results of the model when quantum mechanical information is added. The standard metrics presented are the test mean absolute error (MAE), test root mean square error (RMSE) as defined in Eq. (5), and test $R^2$ score. The standard deviation associated with the mean of each metric over 100 random training/test splits is also provided.

$$RMSE = \sqrt{\frac{1}{N}\sum_{p=1}^{N}(p_{\text{true}} - p_{\text{pred}})}. \tag{5}$$

The notation in the following tables and figures is as follows: the **CA** suffix added to the descriptor indicates that the cation and anion of the ionic liquid have been treated separately (e.g., optimized as separate ions), and the **IL** suffix indicates that the ionic liquid was treated as a single ion pair.

Of the structural descriptors, the descriptor with the best performance for predicting ionic liquid melting points was the ECFP4 fingerprint. As shown in Table I, by comparison, the performance of the KRR model decreases by up to 6 K on average when using the CM descriptor. The CM is based on interatomic distances and charge and was originally developed for predicting atomization energies, a property based on exactly those values.[22] Hence, it is not so surprising that there is a weak relationship between the CM and the melting point. By comparison, the ECFP4 fingerprint performs considerably better. Fingerprints, which were developed for substructure searching, indicate the presence or absence of functional groups and bonding information in molecules. For ionic liquids, the relationship between the melting point and functionality is evident, as shown by the $R^2$ value of 0.74 ± 0.02 for ECFP4-**CA**. This is likely one of the reasons for the failure of the CM: since all interatomic distances in the molecule are treated equally [as seen in Eq. (1)], the KRR model is not able to learn relationships based on similar functional groups, whereas functionality is a key feature of the ECFP4 descriptor that focuses more on the local environment and bonding. Interestingly, for both descriptors, a greater predictive accuracy is associated with the use of the **CA** formulation compared to **IL** results. For the ECFP4 fingerprint, this is probably due to the fingerprinting algorithm that removes duplicate substructures present in both the cation and the anion; the overall result is the loss of important structural information for one of the ions. When the cation and anion are described with separate fingerprints, substructures present in both the cation and the anion would not be considered duplicates and would thus be present in both vectors. The CM has a similar result, showing an improvement in the prediction accuracy when individual ion geometries are used. This could be due to the inflexibility of an attempt to represent all possible ionic liquid interactions through the geometry of a single ion pair, when multiple ion configurations are possible due to the importance of induction and dispersion forces.[47] The use of individual ion geometries does not constrain the ionic liquid to being described through a single ion pair configuration of the multiple possible, resulting in improved accuracy.

TABLE I. Performance of structural descriptors for out-of-sample ionic liquid melting point predictions. Errors and $R^2$ value are reported as the average over 100 repetitions, accompanied by the standard deviation of the mean.

| Descriptor | MAE (K) | RMSE (K) | $R^2$ |
|---|---|---|---|
| ECFP4-**IL** | 33.20 ± 1.40 | 43.77 ± 2.00 | 0.68 ± 0.02 |
| ECFP4-**CA** | 29.78 ± 1.20 | 39.84 ± 1.81 | 0.74 ± 0.02 |
| CM-**IL** | 39.92 ± 1.29 | 52.11 ± 2.01 | 0.55 ± 0.03 |
| CM-**CA** | 35.92 ± 1.44 | 47.30 ± 2.22 | 0.63 ± 0.03 |

Perhaps, the biggest issue with the CM descriptor is in the use of standard atomic charges to describe ionic liquids, a topic that has been widely considered both experimentally and computationally in the field.[48,49] It is well known that atomic charges in ionic liquids fluctuate significantly, leading to fractional atomic charges.[50,51] Hence, the use of standard whole charges as in the CM formulation would not be representative of the true atomic charges in the ionic liquid. The fluctuation of atomic charges in ionic liquids occurs on both an inter-molecular and an intra-molecular scale. To include a charge-based descriptor, we have considered the $\sigma$-profile for the individual ions and ion pair. These, and the other chosen quantum mechanical descriptors, are discussed in Sec. IV A.

### A. Quantum mechanical features

In theory, the $\sigma$-profile of an ionic liquid should be the sum of the $\sigma$-profiles of the individual ions.[42] However, as significant charge transfer may occur between the single ion pair to a greater extent than present in bulk ionic liquids, we have decided to compare the sigma profile of the individual ions with that of the ion pair, analogous to the **IL** and the **CA** formulation of the structural descriptors. 51 numeric values describing the ionic liquid or ion's $\sigma$-profile were included, within a range from −0.025 to 0.025 e/Å$^2$, which is the commonly used range as based on the $\sigma$-profile of water.[52] Molecular orbital data refer to the energy of the HOMO, LUMO, and band gap in eV for the single ion pair, calculated using M06-2X/cc-pVDZ. Interaction energy data contain a value representing the strength and nature of interaction between the cation and the anion: the ratio of the correlation (i.e., dispersion) interaction energy to the total interaction energy $\frac{\mathrm{IE_{total}}}{\mathrm{IE_{Corr}}}$, where the total interaction energy and correlation interaction energy are defined as follows:

$$\mathrm{IE_{total}} = \mathrm{E_{ion\ pair}} - (\mathrm{E_{anion}} + \mathrm{E_{cation}}), \tag{6}$$

$$\mathrm{IE_{total}} = \mathrm{IE_{HartreeâĂŞFock}} + \mathrm{IE_{Correlation}}. \tag{7}$$

We have previously shown that the ratio $\frac{\mathrm{IE_{total}}}{\mathrm{IE_{Corr}}}$ is correlated with the melting point for clusters of ionic liquids containing two ion pairs.[53] Results of the structural descriptors combined with quantum mechanical data are shown in Table II. As the **CA** descriptors

**TABLE II**. Performance of structural descriptors with additional quantum mechanical features for out-of-sample ionic liquid melting point predictions. Errors and $R^2$ value are reported as the average over 100 repetitions, accompanied by the standard deviation of the mean.

| | QM features | MAE (K) | RMSE (K) | $R^2$ |
|---|---|---|---|---|
| ECFP4-**CA** | $\sigma$-profile-**IL** | 31.64 ± 0.95 | 41.79 ± 1.75 | 0.71 ± 0.02 |
| | $\sigma$-profile-**CA** | 30.72 ± 1.11 | 41.02 ± 1.77 | 0.72 ± 0.03 |
| | MO energy | 29.15 ± 1.06 | 39.50 ± 1.82 | 0.74 ± 0.02 |
| | Int. energy | 29.62 ± 0.99 | 39.88 ± 1.72 | 0.74 ± 0.02 |
| | All QM | 30.73 ± 1.08 | 41.07 ± 1.72 | 0.74 ± 0.02 |
| CM-**CA** | $\sigma$-profile-**IL** | 35.03 ± 1.28 | 47.96 ± 2.61 | 0.62 ± 0.04 |
| | $\sigma$-profile-**CA** | 35.00 ± 1.30 | 47.84 ± 2.64 | 0.62 ± 0.04 |
| | MO energy | 36.05 ± 1.39 | 49.02 ± 2.60 | 0.60 ± 0.04 |
| | Int. energy | 36.09 ± 1.39 | 49.07 ± 2.60 | 0.60 ± 0.04 |
| | All QM | 34.82 ± 1.33 | 45.85 ± 2.19 | 0.65 ± 0.03 |

showed better prediction accuracy than **IL** descriptors based on the results in Table I, and this pattern was also observed upon addition of QM descriptors, only results for ECFP4-**CA** and CM-**CA** are shown in Table II. Results from ECFP4-**IL** and CM-**IL** models with QM descriptors can be found in the supplementary material. In this section, each QM descriptor has been examined individually, and the combination of all QM descriptors ($\sigma$-profile-**CA**, molecular orbital energies, and interaction energy ratio) is also included. Parity plots for the predicted vs true values corresponding to the first run out of 100 random splits are shown for the best performing ECFP4 and CM models in Fig. 3.

Although the $\sigma$-profile is a popular descriptor in the ionic liquid QSAR community for predicting physicochemical properties,[16,42] in ECFP4 models, there is no appreciable increase in prediction accuracy when adding the $\sigma$-profile to structural descriptors: the test MAE increases slightly, by an average of 0.94 K. Using $\sigma$-profiles alone as inputs to KRR results in MAE values between 34 K and 52 K (see the supplementary material for KRR models using only *ab initio* data), with poorer results for the $\sigma$-profile-**IL** descriptor. The use of $\sigma$-profile-**CA** however gives the best QM-only model, reinforcing the notion that an apt description of ion charge distribution is necessary to predict the ionic liquid melting temperature. From a chemical point of view, the charge density of each molecule is likely to play an important role in determining the melting point of a liquid. In the case of ionic liquids however, which have regions of inhomogeneity throughout the bulk structure,[54] different nanostructural organization within the ionic liquid would produce slightly different $\sigma$-profiles for each ion pair depending on its surroundings and configuration, a complex effect that cannot be captured through a $\sigma$-profile calculation of a single ion pair geometry. Hence, we observed that the $\sigma$-profile for the individual ions gave better results than that of the single ion pair. The CM descriptor in combination with the $\sigma$-profile produces a slightly better model compared to using the CM alone, with the test MAE decreasing by an average of 0.92 K. This indicates that incorporating information about charge distribution within an ionic liquid is important for this particular descriptor, which otherwise contains an overly simplistic illustration of the atomic charges for ionic liquid systems.

A small increase in prediction accuracy is afforded to the ECFP4 descriptor by addition of molecular orbital energies. On average, the ECFP4-**CA** test MAE decreased by 0.63 K. The CM showed an insignificant increase in accuracy with the addition of molecular orbitals, by 0.13 K on average. At first glance, there is no clear correlation between the melting points in the dataset and the energy of the frontier orbitals of the ion pair, and a KRR model based on only these three values produces weak test scores (MAE = 45.26 ± 1.65 K, see the supplementary material for full metrics). The energy of the bandgap is related to the stability of the ionic liquid, and HOMO/LUMO energies are related to the overall molecular structure. While there is no linear correlation between the bandgap and the melting point as shown in Fig. 4, there may be some non-linear correlation between the combination of frontier orbital energies and the ECFP4 descriptor, which is uncovered through the kernel regression. The use of the ECFP4 descriptor, which encodes functionality and local bonding, plus the frontier orbital energies providing some information about the overall molecular structure, suggests that the combination of this information increases predictive power in comparison with using each descriptor alone.

Including the correlation interaction energy ratio has a surprisingly small effect on performance metrics. Examining the relationship between the melting point and $\frac{\mathrm{IE_{tot}}}{\mathrm{IE_{corr}}}$, as shown in Fig. 5, provides some insight as to why. There does not appear to be a strong correlation between the two values, linear or otherwise. The majority of ionic liquids are clustered around a $\frac{\mathrm{IE_{tot}}}{\mathrm{IE_{corr}}}$ ratio of 10 but display a wide range of melting points from 177 K to 632 K. This type

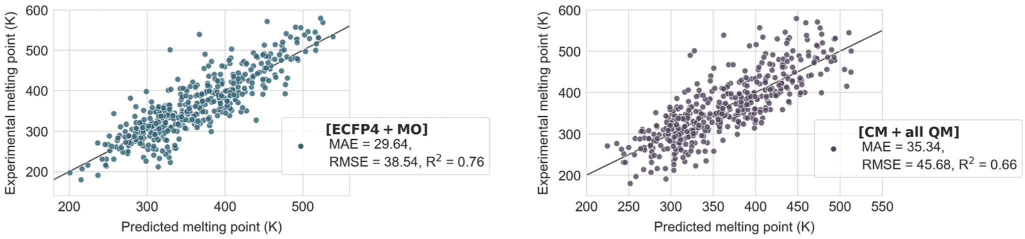

**FIG. 3**. Parity plots comparing the predicted melting point values with the corresponding experimental data for the best performing ECFP4 (a) and CM (b) models. Plots correspond to predictions for the first run out of 100 repetitions.

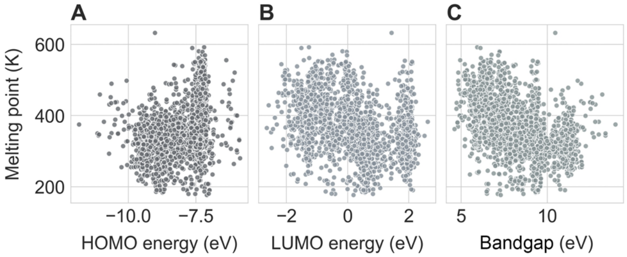

**FIG. 4**. Plots of the frontier orbital energy [(a) highest occupied molecular orbital (HOMO) and (b) lowest unoccupied molecular orbital (LUMO)] and (c) bandgap energy (eV) vs melting point (K) for the ion pairs in the dataset.

of relationship is unlikely to be a successful input for kernel ridge regression, which is reinforced by the metrics of the models built from the interaction energy ratio alone: MAE = 57.00 ± 1.91 K, $R^2$ = 0.14 ± 0.02. Searching over several types of kernels (linear, Gaussian, and Laplacian) fails to improve these results. The plots of molecular orbital energies show a similar trend with weak correlation, but the addition of MO data improves model performance. Of course, it is given that—similar to the results of the $\sigma$-profile—a description of interaction energy using a single ion pair conformation is an overly simplistic approximation. In contrast, frontier orbital energies and bandgap remain the same regardless of the number of ions present. Therefore, when calculating interaction energies in the future work for use in these models, it would be worthwhile to consider the possible range of interaction energies and furthermore use a two ion pair cluster rather than a single ion pair, as this would likely display a stronger relationship with the melting point.

The difference in the approach of the two structural descriptors ECFP4 and CM is highlighted by prediction metrics when all quantum mechanical data are included in the models. In the case of the ECFP4 descriptor, accuracy decreases slightly, whereas for the CM, the addition of the QM data results in the best performing model of all: the MAE decreases by 1.20 K on average and the RMSE by 2.55 K compared to those in the CM model alone. This indicates that newer 3D structural descriptors may benefit predictive accuracy by adding simple but high-quality quantum mechanical data that are related to the property of interest. Of course, the biggest problem lies in filtering through data that are relevant and discarding those that add unhelpful noise to the model. As we have shown, some first-principles data that intuitively correspond to the target property from a chemical point of view—for example, interaction energy and melting point—may prove to not have shown any correlation when incorporated in a machine learning model. The performance of the two structural descriptors reflects their history and intended use, with one (ECFP4) having been developed for use in the QSAR/QSPR field to correlate the molecular substructure with an experimental value such as drug potency, which as with any experiment measurement would have a high degree of noise in the measurement. On the other hand, the CM was developed for predicting results of quantum mechanical calculations, which are essentially noiseless. For this ionic liquid melting point dataset, where there is likely some amount of noise due to variation in the melting point characterization method as well as the presence of impurities, ECFP4 proves its worth as a structural descriptor on its own with excellent performance without any QM data added to this dataset. There is a certainty merit in the addition of QM data to the CM however, as this provides the best performing CM model.

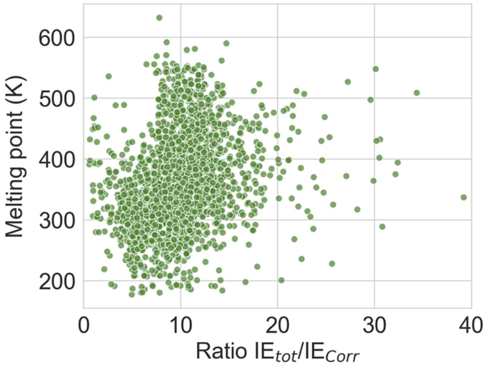

**FIG. 5**. Plot of the SRS-MP2/cc-pVTZ dispersion interaction energy ratio vs melting temperature (K) as calculated for single ion pairs in the dataset.

## B. Error analysis of best performing models

As the range of ions in the investigated dataset is disproportionate in its representation of certain ions, we investigate in this section whether the above models perform better on certain subsections of the dataset. For example, 822 of the 2212 ionic liquids have a bromine anion; hence, it is possible that the model would have a lower test error for bromine-based ionic liquids. To test this hypothesis, we have plotted the histogram of errors for the best performing ECFP4 and CM models for the first model from the 100 training/test split runs. In Fig. 6, the orange ticks on each histogram represent the error for a bromide-based ionic liquid in the test set. While the range reduces somewhat, the errors are still for the most part distributed equally within the possible range of errors, which is [−113, 172] for the ECFP4 model and [−92, 152] for the CM model. Further narrowing down the ionic liquid type to be a bromide anion with an imidazolium cation (one of the most common cation types in the dataset) is shown by the red ticks on the right-hand side of Fig. 6. As can be seen in the two histograms, the range of residual errors narrows down slightly, from [−113, 117] to [−84, 117], for the ECFP4 model but does not change for the CM model. As both ranges still encompass most of the entire residual error range, we can assume that the models are not biased toward a certain type of ionic liquid, despite the prevalence of both bromide and imidazolium ions in the dataset. This lack of predictive accuracy is likely due to descriptor choice.

For further analysis, we examine the identity of the worst-performing ionic liquids for each model, as this can provide some insight into the shortfalls of the model in terms of what features are not being captured by the chosen descriptors. The structure of these four ionic liquids is shown in Fig. 7. For the ECFP4 model, the ionic liquid with the highest residual error (+172 K) is structurally not complex: an ammonium cation with a pyridinylmethanolate anion

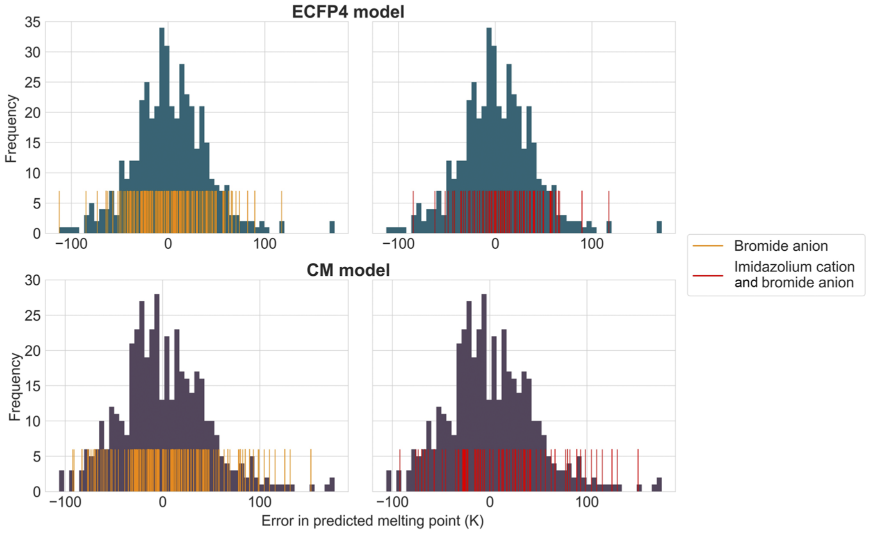

**FIG. 6**. Histogram of residual errors ($y_{true} - y_{pred}$) for the best performing ECFP4 model [ECFP4 + MO] (top) and the best performing CM model [CM + all QM] (bottom). Errors corresponding to bromide-based ionic liquids and imidazolium–bromide based ionic liquids are highlighted in the plots using orange and red ticks, respectively.

[Fig. 7(a)]. Incidentally, this is also the worst ionic liquid for the CM model, with a slightly higher residual error of +177 K. The worst negative residual error for ECFP4 is associated with a pyridinium bromide ionic liquid [Fig. 7(b), −106 K]. For the CM model, the largest negative residual error −106 K occurs for triazolium nitrite [Fig. 7(d)]. There appear to be some similarities and differences in the types of molecules that each model predicts poorly. For the ECFP4 model, a pyridine ring appears in both ionic liquids, whereas the CM appears to perform worse with systems with diffuse charges: both the formate and nitrate anions as well as the triazolium cation are all highly charge-delocalized structures. Figure 6 shows that the CM model has several outliers grouped on the outer right of the histogram with residual errors greater than 170 K: two are ammonium formate ionic liquids—we describe issues with representing ammonium systems in the following paragraph—and one is sulfonium tetrafluoroborate, another highly charge-diffused system. This could reflect the shortfalls of the CM in describing these ionic systems, which as discussed previously require a non-static description of charges. The future work will investigate the use of newer 3D descriptors, which include the atomic charge but lack the fitted exponent term present in the CM [Eq. (1)], which allows for the substitution of fixed atomic charges with quantum-mechanically determined ones. For ionic liquids, we have previously shown that the Geodesic charge scheme can best capture the extent of charge transfer and delocalization present in ionic liquids, out of several available partial charge schemes.[55] Substituting a fixed, standard charge description with Geodesic charges is likely to have a positive effect on prediction accuracy based on assessment of the worst-performing ionic liquids for the CM model.

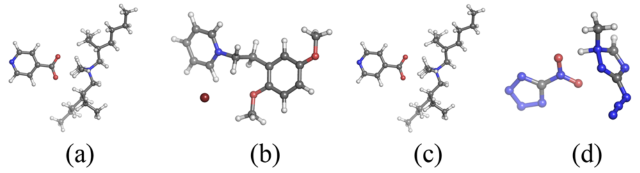

**FIG. 7**. Ionic liquids with the largest positive and negative residual errors for the ECFP4 model: (a) ammonium cation and formate anion (+172 K) and (b) pyridinium bromide (−106 K), and the CM model: (c) the same ammonium formate ionic liquid as for ECFP4 (+177 K) and (d) triazolium cation with nitrate anion (−106 K). Atom coloring: carbon = gray, hydrogen = white, oxygen = red, nitrogen = blue, and bromine = burgundy.

Considering the worst-performing ionic liquids for the ECFP4 model, the presence of elements and functional groups (pyridine, carboxylate, bromide) in both ionic liquids that are relatively common throughout the entire dataset indicates that something in the descriptor, in either the cation–anion ECFP4 fingerprint or the added frontier orbital energies, lacks the detailed structural encoding required to learn the relationship between the descriptor and the melting point. Examining plots of the residual error vs frontier orbital energy (HOMO, LUMO, and bandgap) for each ionic liquid in the test set, Fig. 8, it can be seen that different ionic liquid types with similar frontier orbital energy values to the ionic liquids in Figs. 7(a) and 7(b) have much smaller residual errors. Narrowing these down to focus only on ionic liquids similar to those in Fig. 7(a), with ammonium formate ions, colored blue in Fig. 8, shows that these types of ions have a wide distribution of residual errors: several are scattered about the $y = 0$ residual error line, but equally three ammonium formate ionic liquids have some of the highest residual errors of all test set examples. Similarly, the errors associated with pyridinium bromide ionic (orange dots in Fig. 8) are randomly scattered around the $y = 0$ line, though the range of errors is not as large compared to that of ammonium formate ionic liquids. As similar ionic liquid types with similar frontier orbital energies can be associated with lower residual errors, we speculate that it is either the description provided by the ECFP4 fingerprint or the use of the single cation–anion representation—or, likely, a combination of the two—which lacks key structural information needed for prediction of the target property, melting point. In terms of the cation–anion representation, molecular dynamics simulations have shown that

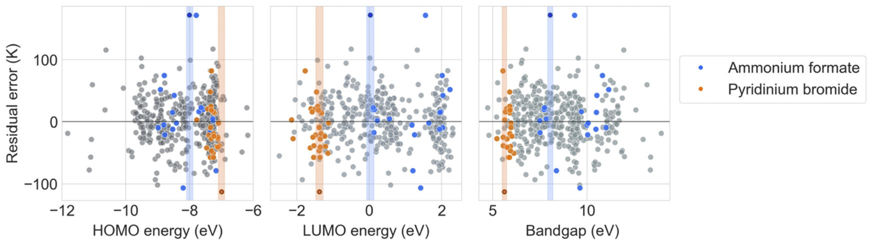

**FIG. 8**. Residual errors (K) for all test set ionic liquids and their frontier orbital energies (eV). Frontier orbital values corresponding to the ionic liquid with the largest positive error [Fig. 7(a)] and largest negative residual error [Fig. 7(a)] for the ECFP4-MO model are highlighted in blue and orange, respectively. Similar ionic liquid types are shown by the same colored markers, and other ionic liquid types are shown in gray.

in ammonium ionic liquids such as that in Fig. 7(a), alkyl–alkyl chain aggregation interactions can overtake cation–anion interactions, especially for chain lengths with $n = 14$ or larger.[54,56] Equally, ionic liquids with a pyridinium cation such as that in Fig. 7(b) are known to have a large dispersion component in the interaction energies of bulk clusters.[47] For these systems, many-body effects are particularly important in their contribution to the melting point. Hence, the representation of the bulk structure using a single ion results in significant errors, which is emphasized through the fact that pyridinium ionic liquids delete have high residual errors in both ECFP4 and CM models.

In terms of the ECFP4 descriptor itself, it is possible that not including information about molecular 3D geometry contributes to these errors, as the ECFP4 descriptor is only based on bonding information. This would result in erroneous predictions, for example, for chiral ionic liquids, which would have the same ECFP4 representation but display different properties.[57] Given this information based on analysis of each model's errors, we can suppose that an ideal structural descriptor for predicting ionic liquid melting points should contain information including 3D geometry, as in the CM, in addition to 2D bonding information over four or more bonds, as in ECFP4. For interest's sake, the results of a model combining the ECFP4-CA and CM-CA descriptors were trained, but the performance of this model was slightly worse than that of ECFP4 models alone (though better than that of CM only models), with a test MAE of 34.19 ± 1.37, RMSE of 45.1 ± 2.19, and $R^2$ equal to 0.66 ± 0.02. Realistically, the solution is not so simple as combining the two structural descriptors chosen in this study: based on our results, we hypothesize that the inclusion of *all* "bonds" within a molecule, regardless of distance between atoms, in the CM imparts a fair amount of irrelevant information into the model. An ideal descriptor for these ionic liquid systems should strike a balance between incorporating the overall three-dimensional structure and shape of the ions while including local bonding information—encapsulating bond order and functional group patterns—within some cutoff for each atom. While such a descriptor may not exist yet, much progress is being made in the field of developing new representations to describe local environments. Several examples from the past few years include the many-body tensor representation[58] and symmetry functions,[59,60] which were not used in the present work but will be considered in the future studies. The Sherrill group very recently extended the application of weighted symmetry functions to intermolecular systems to predict the symmetry-adapted perturbation theory (SAPT) interaction energy between neutral dimers;[61] investigating the performance of these descriptors with charged systems such as ionic liquids would be enlightening. Finally, it would be worthwhile to consider for certain types of ionic liquids where aggregation between cations is known to occur—e.g., pyridinium and long-chain ammonium cations—a cluster approach involving two or more ion pairs when computing *ab initio* descriptors, as single ion interactions clearly do not suffice for these cases.

## C. Model generalizability

To evaluate generalizability and performance of the model when applied to unseen ion types, we evaluate our best performing ECFP4-MO model in several cases where ion types in the test set are not present in the training set. The results are shown in Table III for four anions: $NTf_2^-$, $PF_6^-$, $Cl^-$, and $BF_4^-$, and three cation types: pyridinium, pyrrolidinium, and triazolium. These anions and cations are the second-most, third-most, and fourth-most prevalent ion types in the whole dataset. The most common ions, bromide anion and imidazolium/ammonium cations, were not considered—since each of these ions makes up over one-third of the original dataset, studying these types would have resulted in an insufficiently small training set.

The anion with lowest errors from this study is the bistriflimide anion $NTf_2^-$. Both the test MAE (30.51 K) and the test RMSE (38.58 K) are close to the original test metrics reported in Sec. IV, suggesting that the model is able to extrapolate to anion types that were not present in the training dataset. Parity plots for the $NTf_2^-$

**TABLE III**. Performance of the ECFP4-MO model when various ion types in the test set are excluded from the training set. $\mathbf{N}_{\text{train}}$ and $\mathbf{N}_{\text{test}}$ refer to the number of ionic liquids in the training set and test set, respectively.

| Anion/cation type | $\mathbf{N}_{\text{train}}$ | $\mathbf{N}_{\text{test}}$ | MAE (K) | RMSE (K) |
|---|---|---|---|---|
| $NTf_2^-$ | 1975 | 237 | 30.51 | 38.58 |
| $PF_6^-$ | 2097 | 115 | 36.50 | 49.11 |
| $BF_4^-$ | 2100 | 112 | 38.25 | 50.28 |
| $Cl^-$ | 2132 | 80 | 34.47 | 44.89 |
| Pyridinium | 1919 | 293 | 41.53 | 51.42 |
| Pyrrolidinium | 2112 | 100 | 47.59 | 61.83 |
| Triazolium | 2114 | 98 | 40.73 | 54.51 |

model (included in the supplementary material, Fig. S3) show that although many of the predicted melting points lie on or close to the $y = x$ line, there are also several outliers that have predicted melting points within ±100 K of the true value. For all anions studied, the presented results are similar to or worse than the metrics from the original (randomly split) dataset. However, as this study used a much smaller test set in comparison with the original test set, outliers have a larger effect on test metrics and particularly the $R^2$ value. This was also seen in the $BF_4^-$ results; the anion with the largest test MAE and RMSE had the highest $R^2$ value since points were clustered between an equal positive and negative distance about the $y = x$ line. Hence, we consider test MAE and RMSE values only to evaluate model performance in this section. Table III shows that the ECFP4-MO model has acceptable predictive accuracy when applied to ion types outside of the training set. This is particularly true of the results for the $Cl^-$ anion: considering only 70 cations and 8 anions in the training set that contain a chlorine atom after all chloride-containing ionic liquids were removed, the test MAE and test RMSE of 34.47 K and 44.80 K for chloride-based ionic liquids are indeed promising.

Errors for the three studied cations are slightly higher in comparison with those for the anions, which is expected as pyridinium, pyrrolidinium, and triazolium cation groups encompass multiple molecules with a shared structural backbone. Prediction accuracy is best for the pyridinium and triazolium cations (MAE = 41.53 K and 40.73 K, respectively) and worst for the pyrrolidinium cation (MAE = 47.59 K). This is likely due to the structure of the training set: pyrrolidinium and triazolium cations both share a conjugated backbone, similar to the imidazolium cation that makes up over 35% of the dataset. Hence, these two ion types can inherit learned features from the imidazolium cations. In contrast, fewer cations in the dataset share similar features to pyrrolidinium, a cation type that encompasses a wide range of structures due to numerous possible variations in the two ammonium R-groups including alkyl chains or ring motifs. Naturally, errors are higher for predictions made for ionic liquids containing this cation.

By examining models trained on different subsets of the dataset that exclude various anion and cation types, we have shown that the applicability domain of the ECFP4-MO model extends to ion types outside of which the model was trained on. It is possible to achieve good prediction accuracy for previously unseen anion types, ranging from small halogen anions to a larger anion such as $NTf_2^-$. Furthermore, examining applicability of the model to various unseen cation structures shows that while the model performs better if similar ion types to the test samples are present in the training set (e.g., imidazolium ions in the training set extend well to other types of planar, conjugated cations such as triazolium), the prediction accuracy is still reasonable for cation types that differ in the backbone structure. Ultimately, the most important factor in achieving a robust model with reliable predictions for any new type of ion is to ensure the diversity of the training dataset, dataset quality being one of the key factors in successful application of machine learning models to predict experimental properties.

### D. Comparison with literature models

This section compares the results of our best performing model, ECFP4-MO, with those of selected other models in the literature. However, as there are still relatively few studies on the application of machine learning for ionic liquids (most literature studies have targeted a specific property for an intended purpose, such as $CO_2$ solubility[21]), results in the literature are mostly based on QSAR or group contribution methods. The previous work in the literature has reported similar or higher MAE values for melting point prediction, such as in the source paper for the database: the random forest-based models used by Venkatraman *et al.*[31] achieved a test MAE of 33 K and test RMSE of 45 K, compared to our values of 29 K and 40 K, respectively.

The test $R^2$ score and test RMSE as reported by several other authors are shown in Table IV, along with the type of model, number of features used as the input (we consider each ECFP4 descriptor as a single feature in our case, though argument could be made against this as it is a 2048-length bit vector), model type, and size of the dataset. As machine learning for ionic liquid property predictions is still a relatively new field, the majority of existing papers using QSAR or group contribution approaches were limited to a small subset of ionic liquids such as only imidazolium-based ones.[62] To keep comparisons on an equal footing, we have included only literature sources that studied a varied ionic liquid dataset containing several hundred ionic liquids or more.

Compared to the models in Table IV, our model uses a fewer number of features yet achieves comparable accuracy to the other models. The best performing model is based on a group contribution method, which involves enumerating each possible substructure in the ionic liquid dataset and assigning it a value. In order to make predictions on new data with chemical substructures outside of the original dataset, the group contribution scheme would need to be re-devised. By comparison, our model makes use of readily available features: the ECFP4 fingerprint that is generated from the SMILES string and the molecular orbital energies that can be computed from DFT-optimized geometries. Hence, any type of ionic liquid can theoretically be included, though performance will depend on the composition of the training set, as explored in Sec. IV C. This reflects the larger size of our dataset in comparison with that in Ref. 63.

In terms of other computational methods for calculating ionic liquid melting points, molecular dynamics has been a popular choice in the field. However, it is accepted that ionic liquids require specialized force fields and additional treatment of polarization effects for the most accurate description of intermolecular interactions.[66,67] Simulations using these force fields have shown a similar range of errors to the models in Table IV,[68] though some MD

**TABLE IV**. Test metrics for predictions of the ionic liquid melting temperature as reported in the literature. *N* refers to the size of the entire dataset. GC = group contribution; ANN = artificial neural network; RF = random forest.

| References | *N* | Features | Model | $R^2_{test}$ | RMSE (K) |
|---|---|---|---|---|---|
| 63 | 799 | 80 | GC | 0.89 | 24.86 |
| 64 | 799 | 40 | ANN | 0.54 | 33.33 |
| 65 | 808 | 12 | QSPR | 0.72 | 26.85 |
| 19 | 2212 | 226 | RF | 0.66 | 45.00 |
| This work | 2212 | 5 | KRR | 0.76 | 38.54 |

simulations have come within ±5 K of the experimental value depending on the type of ionic liquid and methodology used.[69] For example, imidazolium-based ionic liquids (especially $[C_{1-4}mim]^+$) have been widely studied, and the use of specialized force fields allows molecular dynamics simulations to accurately reproduce some transport properties for these cation types.[70,71] Molecular dynamics simulations of other cation types—such as morpholinium and sulfonium, which are both present in our dataset—are unlikely to give accurate results, since parameterized force fields for these cations do not yet exist. Considering the time and computational resources required for MD simulations, by comparison, the machine learning models such as those presented herein offer a suitable alternative that provides comparable accuracy for ionic liquid property prediction. Touching on the time requirements for each descriptor examined in this work, ECFP4 fingerprints are naturally much faster to compute than the CM and all QM descriptors, which require an optimized geometry. Geometry optimizations of the 141 anions required an average of 6.8 central processing unit (CPU) hours per anion, roughly doubling in time for the cations, requiring 12.5 CPU hours on average for each of the 1369 cations. However, based on these results presented when adding all QM descriptors, it is likely that a QM descriptor outside of those considered in this work could have a more significant effect on performance, such that the gain in accuracy outweighs the increased computational demand. The great advantage of machine learning is that it requires far less computational time than simulation, as once training is completed, predictions for new compounds can be made effectively in an instant. Machine learning models are therefore ideal for high-throughput screening of chemical databases and can efficiently filter out possible candidates, which fall outside of the desired property range. The addition of QM properties to machine learning models, while increasing computational costs, will ultimately save experimental hours by further filtering the pool of candidates once an initial screen using semi-empirical methods has taken place.

## V. CONCLUSION

In the ionic liquid field, the QSAR model with hundreds of hand-engineered features has long been the best and accepted approach for predicting physicochemical properties such as viscosity and melting point.[32] While impressive results can be achieved using these methods, this study strives to move away from this "shotgun" approach and toward structural descriptors plus a focus on a small number of quantum mechanical descriptors, which are related to the target property. Our approach reduces the need to calculate large feature vectors followed by the process of downward feature selection, in addition to being able to apply such models to any type of ionic liquid and ion types outside of the training dataset with acceptable accuracy. The application of newer structural descriptors such as the Coulomb matrix (which have so far only been applied to small, organic, neutral molecules) to ionic liquid ions has shown that they can feasibly be used in such surrogate models where the exact form of the relationship between the structure and the property is unknown. However, in the case of ionic liquid melting points, the correlation between the CM alone and the melting point is weak, achieving an $R^2$ score of 0.63 ± 0.03. Nonetheless, the three-dimensional geometric information contained within the CM proved to be important for the prediction of the ionic liquid melting point, and better predictive performance was achieved upon combining the CM with all QM descriptors ($R^2$ = 0.65 ± 0.03). The *ab initio* information and particularly the ion $\sigma$-profiles likely compensate for the use of fixed charges in the CM formulation by providing a more flexible description of charge. Overall, the best results were achieved using the ECFP4 fingerprint ($R^2$ = 0.74 ± 0.02), which offers a faster computation in comparison with the CM as it does not need 3D geometries and contains only 2D bonding information. ECFP4 has been a favorite among the many possible QSAR descriptors for years. With the addition of DFT molecular orbital energies, a very small increase in accuracy was afforded compared to using the ECFP4 descriptor alone (the MAE decreases by 0.63 K on average). Although the time requirements for computing molecular orbitals in comparison with computing the ECFP4 alone are not likely to justify such a small increase in predictive accuracy, it is likely that there are other quantum mechanical features outside of those explored in this work, which could lower the error more significantly, as was observed for the CM.

Given the number of possible ionic liquids that could be considered for an application, the models developed in this manuscript present a viable alternative to molecular simulation for rapid screening of ionic liquids with a desired melting point range. The use of a relatively expensive DFT method, M06-2X/cc-pVDZ for optimizations, was chosen based on our previous studies.[72,73] This combination contrasts with that in the literature where semi-empirical methods are commonly employed for rapid exploratory searches of large datasets.[19,20] However, we believe that these two approaches would work best in tandem: the use of higher levels of theory is more suited for medium-sized datasets due to increased accuracy of the predicted geometries and should follow an initial screening of a large dataset using a semi-empirical method. Using a higher level of theory could further refine semi-empirical predictions and narrow the potential pool of suitable ionic liquids to be synthesized, potentially eliminating several ionic liquids that were incorrectly predicted using semi-empirical methods or adding ones where effects could not be captured properly using a lower level of theory. Several questions remain to be answered in future work: what role the cation plays as these models seem to favor contribution from the anion, in addition to finding a descriptor for these systems that can achieve better prediction accuracy. Based on the few descriptors studied herein and their errors, the ideal descriptor for predicting ionic liquid melting points should combine bonding information with some 3D geometry, an accurate description of the charge environment. Determining the exact form of the structural descriptor for the chosen target property remains the challenge at hand. With the right combination of the structural descriptor and appropriate high-level *ab initio* data, we are confident that a level of accuracy is possible such that machine learning becomes the new norm for property prediction in the ionic liquid field, as it has become so in many other areas of materials science.

## SUPPLEMENTARY MATERIAL

The supplementary material includes the following: SMILES strings and experimental melting point values, details of geometry optimizations, evaluation of the ECFP4 vector length, t-SNE analysis colored by cation type, KRR models using *ab initio* data only, ECFP4-IL and CM-**IL** models with QM descriptors' results, all KRR model parameters, model generalizability parity plots, and *ab initio* descriptor data. Optimized geometries of the ionic liquids or individual ions are available upon request.

## AUTHORS' CONTRIBUTIONS

All authors contributed equally to this work.

## ACKNOWLEDGMENTS

The authors acknowledge a generous allocation of computational resources from the Monash *e*Research Centre and the National Computational Infrastructure. This work was supported by the Australian Government through a Research Training Program scholarship to K.L. The authors would like to thank Fernando Clemente at Gaussian, Inc., for his helpful advice on sigma profile generation.

## DATA AVAILABILITY

The data that support the findings of this study are available within this article and its supplementary material.

## REFERENCES

[1] M. Watanabe, M. L. Thomas, S. Zhang, K. Ueno, T. Yasuda, and K. Dokko, "Application of ionic liquids to energy storage and conversion materials and devices," Chem. Rev. **117**, 7190–7239 (2017).

[2] A. Basile, M. Hilder, F. Makhlooghiazad, C. Pozo-Gonzalo, D. R. MacFarlane, P. C. Howlett, and M. Forsyth, "Ionic liquids and organic ionic plastic crystals: Advanced electrolytes for safer high performance sodium energy storage technologies," Adv. Energy Mater. **8**, 1703491 (2018).

[3] J. P. Hallett and T. Welton, "Room-temperature ionic liquids: Solvents for synthesis and catalysis. 2," Chem. Rev. **111**, 3508–3576 (2011).

[4] E. I. Izgorodina, Z. L. Seeger, D. L. A. Scarborough, and S. Y. S. Tan, "Quantum chemical methods for the prediction of energetic, physical, and spectroscopic properties of ionic liquids," Chem. Rev. **117**, 6696–6754 (2017).

[5] B. Kirchner, O. Hollóczki, J. N. Canongia Lopes, and A. A. H. Pádua, "Multiresolution calculation of ionic liquids," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **5**, 202–214 (2015).

[6] S. Zahn, J. Thar, and B. Kirchner, "Structure and dynamics of the protic ionic liquid monomethylammonium nitrate ([$CH_3NH_3$][$NO_3$]) from *ab initio* molecular dynamics simulations," J. Chem. Phys. **132**, 124506 (2010).

[7] B. Huang, N. O. Symonds, and O. A. v. Lilienfeld, "Quantum machine learning in chemistry and materials," in *Handbook of Materials Modeling*, Methods: Theory and Modeling, edited by W. Andreoni and S. Yip (Springer International Publishing, Cham, 2018), pp. 1–27.

[8] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," Chem. Sci. **9**, 513–530 (2018).

[9] F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day, and M. Ceriotti, "Machine learning for the structure–energy–property landscapes of molecular crystals," Chem. Sci. **9**, 1289–1300 (2018).

[10] A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo, and N. Mingo, "High-throughput computation of thermal conductivity of high-temperature solid phases: The case of oxide and fluoride perovskites," Phys. Rev. X **6**, 041061 (2016).

[11] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, "Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single- and binary-component solids," Phys. Rev. B **89**, 054303 (2014).

[12] J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Comput. Mater. **5**, 83 (2019).

[13] E. I. Izgorodina, D. Golze, R. Maganti, V. Armel, M. Taige, T. J. S. Schubert, and D. R. MacFarlane, "Importance of dispersion forces for prediction of thermodynamic and transport properties of some common ionic liquids," Phys. Chem. Chem. Phys. **16**, 7209–7221 (2014).

[14] K. Paduszyński and U. Domańska, "Viscosity of ionic liquids: An extensive database and a new group contribution model based on a feed-forward artificial neural network," J. Chem. Inf. Model. **54**, 1311–1324 (2014).

[15] W. Beckner, C. M. Mao, and J. Pfaendtner, "Statistical models are able to predict ionic liquid viscosity across a wide range of chemical functionalities and experimental conditions," Mol. Syst. Des. Eng. **3**, 253–263 (2018).

[16] Y. Zhao, Y. Huang, X. Zhang, and S. Zhang, "A quantitative prediction of the viscosity of ionic liquids using $S_{\sigma\text{-profile}}$ molecular descriptors," Phys. Chem. Chem. Phys. **17**, 3761–3767 (2015).

[17] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (John Wiley & Sons, 2008), Vol. 11.

[18] E. T. Swann, M. Fernandez, M. L. Coote, and A. S. Barnard, "Bias-free chemically diverse test sets from machine learning," ACS Comb. Sci. **19**, 544–554 (2017).

[19] V. Venkatraman, S. Evjen, K. C. Lethesh, J. J. Raj, H. K. Knuutila, and A. Fiksdahl, "Rapid, comprehensive screening of ionic liquids towards sustainable applications," Sustainable Energy Fuels **3**, 2798–2808 (2019).

[20] E. Wyrzykowska, A. Rybińska-Fryca, A. Sosnowska, and T. Puzyn, "Virtual screening in the design of ionic liquids as environmentally safe bactericides," Green Chem. **21**, 1965–1973 (2019).

[21] V. Venkatraman and B. K. Alsberg, "Predicting $CO_2$ capture of ionic liquids using machine learning," J. $CO_2$ Util. **21**, 162–168 (2017).

[22] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 058301 (2012).

[23] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," J. Phys. Chem. Lett. **6**, 2326–2331 (2015).

[24] K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, "Machine learning of molecular properties: Locality and active learning," J. Chem. Phys. **148**, 241727 (2018).

[25] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, "Constant size descriptors for accurate machine learning models of molecular properties," J. Chem. Phys. **148**, 241718 (2018).

[26] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," IEEE Trans. Evol. Comput. **1**, 67–82 (1997).

[27] A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen, and P. Rinke, "Chemical diversity in molecular orbital energy predictions with kernel ridge regression," J. Chem. Phys. **150**, 204121 (2019).

[28] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. V. Lilienfeld, and K.-R. Müller, "Learning invariant representations of molecules for atomization energy prediction," in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., 2012), pp. 440–448.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[30] A. B. Tchagang and J. J. Valdés, "Prediction of the atomization energy of molecules using Coulomb matrix and atomic composition in a Bayesian regularized neural networks," in *International Conference on Artificial Neural Networks* (Springer, 2019), pp. 793–803.

[31] V. Venkatraman, S. Evjen, H. K. Knuutila, A. Fiksdahl, and B. K. Alsberg, "Predicting ionic liquid melting points using machine learning," J. Mol. Liq. **264**, 318–326 (2018).

[32] J. O. Valderrama, "Myths and realities about existing methods for calculating the melting temperatures of ionic liquids," Ind. Eng. Chem. Res. **53**, 1004–1014 (2014).

[33] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," J. Chem. Inf. Model. **50**, 742–754 (2010).

[34] S. Riniker and G. A. Landrum, "Open-source platform to benchmark fingerprints for ligand-based virtual screening," J. Cheminf. **5**, 26 (2013).

[35] G. Landrum, rdkit: Open-source cheminformatics, http://www.rdkit.org, 2020.
[36] M. Rupp, "Machine learning for quantum mechanics in a nutshell," Int. J. Quantum Chem. **115**, 1058–1073 (2015).
[37] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian ~16, Revision C.01, Gaussian, Inc., Wallingford, CT, 2016.
[38] A. Klamt and G. Schüürmann, "COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient," J. Chem. Soc., Perkin Trans. 2 **1**, 799–805 (1993).
[39] V. Barone and M. Cossi, "Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model," J. Phys. Chem. A **102**, 1995–2001 (1998).
[40] S. Tan, S. Barrera Acevedo, and E. I. Izgorodina, "Generalized spin-ratio scaled MP2 method for accurate prediction of intermolecular interactions for neutral and ionic species," J. Chem. Phys. **146**, 064108 (2017).
[41] A. Klamt, "Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena," J. Phys. Chem. **99**, 2224–2235 (1995).
[42] J. Palomar, J. S. Torrecilla, J. Lemus, V. R. Ferro, and F. Rodríguez, "A COSMO-RS based guide to analyze/quantify the polarity of ionic liquids and their mixtures with organic cosolvents," Phys. Chem. Chem. Phys. **12**, 1991–2000 (2010).
[43] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res. **9**, 2579–2605 (2008).
[44] G. Pilania, K. J. McClellan, C. R. Stanek, and B. P. Uberuaga, "Physics-informed machine learning for inorganic scintillator discovery," J. Chem. Phys. **148**, 241729 (2018).
[45] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics (Springer, 2009).
[46] 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, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[47] P. Halat, Z. L. Seeger, S. Barrera Acevedo, and E. I. Izgorodina, "Trends in two- and three-body effects in multiscale clusters of ionic liquids," J. Phys. Chem. B **121**, 577–588 (2017).
[48] R. M. Fogarty, R. P. Matthews, C. R. Ashworth, A. Brandt-Talbot, R. G. Palgrave, R. A. Bourne, T. Vander Hoogerstraete, P. A. Hunt, and K. R. J. Lovelock, "Experimental validation of calculated atomic charges in ionic liquids," J. Chem. Phys. **148**, 193817 (2018).
[49] O. Hollóczki, F. Malberg, T. Welton, and B. Kirchner, "On the origin of ionicity in ionic liquids. ion pairing versus charge transfer," Phys. Chem. Chem. Phys. **16**, 16880–16890 (2014).
[50] K. Wendler, S. Zahn, F. Dommert, R. Berger, C. Holm, B. Kirchner, and L. Delle Site, "Locality and fluctuations: Trends in imidazolium-based ionic liquids and beyond," J. Chem. Theory Comput. **7**, 3040–3044 (2011).
[51] C. Schröder, "Comparing reduced partial charge models with polarizable simulations of ionic liquids," Phys. Chem. Chem. Phys. **14**, 3089–3102 (2012).
[52] E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandler, C.-C. Chen, M. Zwolak, and K. C. Seavey, "Sigma-profile database for using COSMO-based thermodynamic methods," Ind. Eng. Chem. Res. **45**, 4389–4415 (2006).
[53] Z. L. Seeger, R. Kobayashi, and E. I. Izgorodina, "Cluster approach to the prediction of thermodynamic and transport properties of ionic liquids," J. Chem. Phys. **148**, 193832 (2018).
[54] J. N. Canongia Lopes and A. A. H. Pádua, "Nanostructural organization in ionic liquids," J. Phys. Chem. B **110**, 3330–3335 (2006).
[55] J. Rigby and E. I. Izgorodina, "Assessment of atomic partial charge schemes for polarisation and charge transfer effects in ionic liquids," Phys. Chem. Chem. Phys. **15**, 1632–1646 (2013).
[56] Y. Ji, R. Shi, Y. Wang, and G. Saielli, "Effect of the chain length on the structure of ionic liquids: From spatial heterogeneity to ionic liquid crystals," J. Phys. Chem. B **117**, 1104–1109 (2013).
[57] S. Yu, S. Lindeman, and C. D. Tran, "Chiral ionic liquids: Synthesis, properties, and enantiomeric recognition," J. Org. Chem. **73**, 2576–2591 (2008).
[58] H. Huo and M. Rupp, "Unified representation of molecules and crystals for machine learning," arXiv:1704.06439 (2017).
[59] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).
[60] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials," J. Chem. Phys. **148**, 241709 (2018).
[61] D. P. Metcalf, A. Koutsoukas, S. A. Spronk, B. L. Claus, D. A. Loughney, S. R. Johnson, D. L. Cheney, and C. D. Sherrill, "Approaches for machine learning intermolecular interaction energies and application to energy components from symmetry adapted perturbation theory," J. Chem. Phys. **152**, 074103 (2020).
[62] J. S. Torrecilla, F. Rodríguez, J. L. Bravo, G. Rothenberg, K. R. Seddon, and I. López-Martin, "Optimising an artificial neural network for predicting the melting point of ionic liquids," Phys. Chem. Chem. Phys. **10**, 5826–5831 (2008).
[63] F. Gharagheizi, P. Ilani-Kashkouli, and A. H. Mohammadi, "Computation of normal melting temperature of ionic liquids using a group contribution method," Fluid Phase Equilib. **329**, 1–7 (2012).
[64] J. O. Valderrama, C. A. Faúndez, and V. J. Vicencio, "Artificial neural networks and the melting temperature of ionic liquids," Ind. Eng. Chem. Res. **53**, 10504–10511 (2014).
[65] N. Farahani, F. Gharagheizi, S. A. Mirkhani, and K. Tumba, "Ionic liquids: Prediction of melting point by molecular-based model," Thermochim. Acta **549**, 17–34 (2012).
[66] P. A. Hunt, "The simulation of imidazolium-based ionic liquids," Mol. Simul. **32**, 1–10 (2006).
[67] K. Goloviznina, J. N. Canongia Lopes, M. Costa Gomes, and A. A. H. Pádua, "Transferable, polarizable force field for ionic liquids," J. Chem. Theory Comput. **15**, 5858–5871 (2019).
[68] M. L. S. Batista, J. A. P. Coutinho, and J. R. B. Gomes, "Prediction of ionic liquids properties through molecular dynamics simulations," Curr. Phys. Chem. **4**, 151–172 (2014).
[69] Y. Zhang and E. J. Maginn, "The effect of $C_2$ substitution on melting point and liquid phase dynamics of imidazolium based-ionic liquids: Insights from molecular dynamics simulations," Phys. Chem. Chem. Phys. **14**, 12157–12164 (2012).
[70] J. N. Canongia Lopes and A. A. H. Pádua, "Molecular force field for ionic liquids III: Imidazolium, pyridinium, and phosphonium cations; chloride, bromide, and dicyanamide anions," J. Phys. Chem. B **110**, 19586–19592 (2006).
[71] M. H. Kowsari, S. Alavi, M. Ashrafizaadeh, and B. Najafi, "Molecular dynamics simulation of imidazolium-based ionic liquids. I. Dynamics and diffusion coefficient," J. Chem. Phys. **129**, 224508 (2008).
[72] E. I. Izgorodina, U. L. Bernard, and D. R. MacFarlane, "Ion-pair binding energies of ionic liquids: Can DFT compete with *ab initio*-based methods?," J. Phys. Chem. A **113**, 7064–7072 (2009).
[73] S. Zahn, D. R. MacFarlane, and E. I. Izgorodina, "Assessment of Kohn–Sham density functional theory and Møller–Plesset perturbation theory for ionic liquids," Phys. Chem. Chem. Phys. **15**, 13664–13675 (2013).