doi:10.2533/chimia.2023.625 Chimia 77 (2023) 625–627 © M. Blaise, S. Barras, F. Yerly

FH HES
Universities of Applied Sciences
Fachhochschulen – Hautes Ecoles Spécialisées

Machine Learning Models for Melting Point Prediction of Ionic Liquids: *CatBoost* Approach

Mathias Blaise, Simon Barras, and Florence Yerly*

**Correspondence*: Dr. F. Yerly, E-mail: florence.yerly@hefr.ch
Haute école d'ingénierie et d'architecture Fribourg, HES-SO University of Applied Sciences and Arts Western Switzerland, Pérolles 80, CH-1700 Fribourg, Switzerland

Abstract: Using ionic liquids as phase changing materials is of particular interest in the context of heat storage. As a consequence, predicting accurately the melting point of ionic liquids is of capital importance as it is one of the most important thermophysical properties in this context. In this work we consider a data set composed of 2249 different ionic liquids, with a majority of imidazole or ammonium cation-based molecules. We present a free and easy-to-use melting point predictive algorithm built on the *CatBoost* algorithm, making extensive use of molecular descriptors. Based on LASSO, we select the most relevant descriptors for the task at hand and compare the model with previous ones.

Keywords: Ionic liquids · Melting point · Molecular descriptors · Prediction

1. Introduction

Over the last decades, ionic liquids (ILs) have been of great interest in chemistry. Their low melting point (MP)[1] and other particular thermophysical properties can be utilized in different fields, such as heat transfer, conductivity and storage media.[2–4] Our focus is on their nature as phase-changing materials (PCM). Indeed, having a low MP can be of great interest when studying PCM: less energy is required to force the change in state of matter. Considering the large number of possible ILs, estimated in the range of 10^{18}, machine learning (ML) and statistical models are of great help when it comes to predicting ILs properties such as their MP.

In this work, we discuss a ML method to build a robust and general predictive model for the MP of ILs (Fig. 1). Indeed, having a room-temperature MP is particularly interesting, as the phase change is known to store heat. Here, the temperature range of interest I is given by $I = [25\ °C, 50\ °C] = [298.15\ K, 323.15\ K]$, that is, slightly above room temperature. ILs of interest for further experimental measures, such as heat of fusion and heat capacity, should have an MP within the I range. Therefore, it is important to have a particular evaluation metric, the ratio of false positive (F_p), which gives us the percentage of classification within I that should not have been predicted in I. F_p is to be minimized to help reduce costly experimental measures, both in time and resources. It should be kept in mind that the heat of fusion also plays an important role in the selection of good candidates, but MP prediction is of greater importance, since it reduces the number of candidates (arbitrarily built ILs) to test in the laboratory. Predictive models for heat of fusion and heat capacity already exist,[5–7] but there is only a very limited number of molecules for which the heat of fusion is publicly available, which means the construction of general models is almost impossible as of yet.

We present a quantitative structure–property relationship (QSPR) approach. As has been emphasized by Valderrama[8] and Holbrey and Rogers,[9] there are difficulties in accurately measuring melting properties of ILs. These difficulties, among others, can explain why several values of MP for the same ILs have been reported, with differences up to 200 K. This makes building accurate general predictive models quite challenging. Once the model is built, we end by comparing its performance with two naive predictive models: *random draw* and *all to mean*. The results are then further compared with previous state-of-the-art machine-learning models in the domain.

2. Methods and Data

The data set consists of 2249 ILs retrieved from Low *et al*.[10] as well as some in-house measured MPs. The number of molecules considered here is slightly larger than in previous works[11–15] where the data set size is frequently limited to several hundred ILs or smaller to reduce computational cost[10] and measuring errors.[9] Note that the majority of ionic liquids in this data set are imidazolium- or ammonium cation-based[10] and about a quarter of the ILs are bromides, which could influence the predictive abilities of the models, as the data set is not so chemically diverse. A model trained on over-represented ions is likely to make errors when predicting values for the underrepresented ions.

2.1 *Descriptors Generation*

To build the machine-learning models, we first translate the molecules into their canonical SMILES with the help of the website *Cactus* and generate 860 molecular descriptors from the SMILES

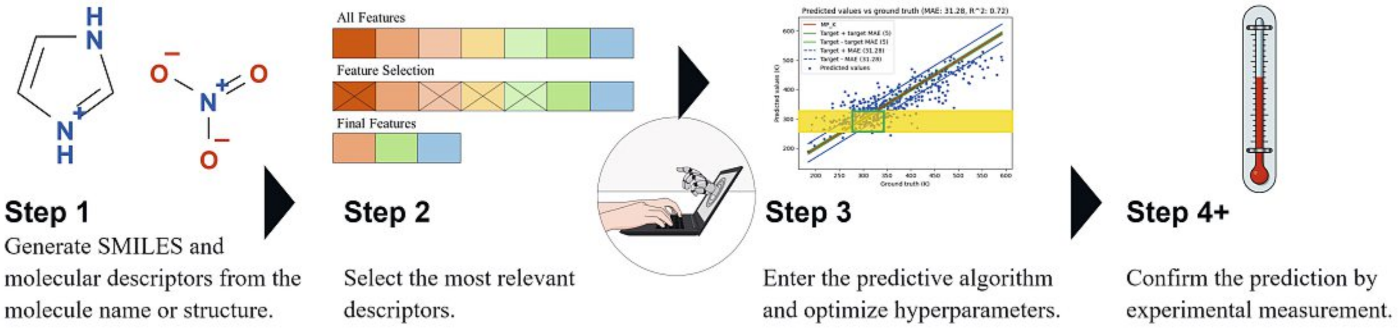
H
N
N
H
+
O
−
N
+
O
O
−
Step 1
Generate SMILES and molecular descriptors from the molecule name or structure.
All Features
Feature Selection
Final Features
Step 2
Select the most relevant descriptors.
Predicted values vs ground truth (MAE: 31.28, R^2: 0.72)
MP_K
Target + target MAE (5)
Target - target MAE (5)
Target + MAE (31.28)
Target - MAE (31.28)
Predicted values
Predicted values (K)
Ground truth (K)
200
250
300
350
400
450
500
550
600
Step 3
Enter the predictive algorithm and optimize hyperparameters.
Step 4+
Confirm the prediction by experimental measurement.

Fig. 1. Machine-learning model for MP prediction. Step 1 to 3 are automated.

using the *mordred* software.[16] It is important to note that the molecular descriptors generated will depend on the software used[17] and discrepancies for higher dimensional descriptors are not negligible. Here, the choice of the *mordred* software is based on its free availability and ease of use. Nevertheless, from the 860 descriptors, we cannot decide *a priori* which are relevant and which add noise in the predicting algorithms. In machine learning, it is known[10,18–20] that reducing the number of features, here molecular descriptors, helps increasing learning speed and predicting accuracy.

2.2 *Preprocessing the Descriptors*

The selection algorithm used is the LASSO method,[21] which stands for Least Absolute Shrinkage and Selection Operator. It can be described as a linear regression with a regularization parameter subject to a constraint, allowing to set some of the coefficients of the standard linear regression to 0. This particular feature makes LASSO a particularly robust and well-suited method for selecting relevant molecular descriptors, reducing their number from 860 down to 156.

2.3 *Algorithm*

After the data preprocessing, it remains to build and train a predictive model for our target, the MPs of ILs. The large number of ML algorithms available in various Python libraries makes it difficult for researchers to choose the best one for the particular task at hand. Here, we chose a *CatBoost* approach. This choice was influenced by the underlying minimization process based on gradient boosting, which seemed better suited for the problem at hand as well as the non-linearity within the data.

CatBoost stands for *Categorical boosting* and is an open-source library for gradient boosting on decision trees, developed by Yandex researchers and engineers.[22] It can easily be implemented in a Python code. The idea behind the algorithm is a simple gradient boosting: if the goal is to teach a model F to predict values of the form $\hat{y} = F(x)$ by minimizing a loss function depending on the metric chosen for the error, the minimization is made by gradient descent. In our model, we chose to minimize the root mean squared error (RMSE), as it was evident that minimizing the mean absolute error (MAE) was causing the model to overfit more. Additionally, maximizing the R^2 score was of great importance. Performing Bayesian Optimization, a sequential optimization strategy widely used in many fields,[23–27] led to the following hyperparameters for the *CatBoost* regression model: iterations = 739, learning_rate = 0.0615, depth = 8, loss_function = 'RMSE'.

2.4 *Metrics*

To evaluate the accuracy of the predictive models, different metrics are used. For more details, see *e.g.* the book by Hastie.[21] In order of importance for this work we have:

- The ratio of false-positive F_p, given by

$$F_p = \frac{f_p}{t_p + f_p} \tag{1}$$

where f_p and t_p are the number of false positives and true positives, respectively.
- The coefficient of determination R^2, with values usually varying between 0 and 1. A value of 1 indicates that the explanatory variables can perfectly explain the variance in the response variable and a value of 0 indicates that the explanatory variables have no ability to explain the variance in the response variable. In the case of negative values, it means that the model is performing worse than the *all to mean* model.

3. Results and Discussion

Once the data set is preprocessed and loaded, we generate 100 random train-test (80% – 20%) splits for training the algorithms. Each train split is again subdivided following a similar split size, to get a validation set. This validation set allows to minimize overfitting. On each split, we compute the ratio R^2 and the R^2 score of the model. On 100 train-test splits, the average results are F_p equal to 0.12 (standard deviation, sd = 0.01) and R^2 equal to 0.75 (sd = 0.02) with the *CatBoost* method. Fig. 2 represents the predicted versus the measured MP of a particular random test-train split.

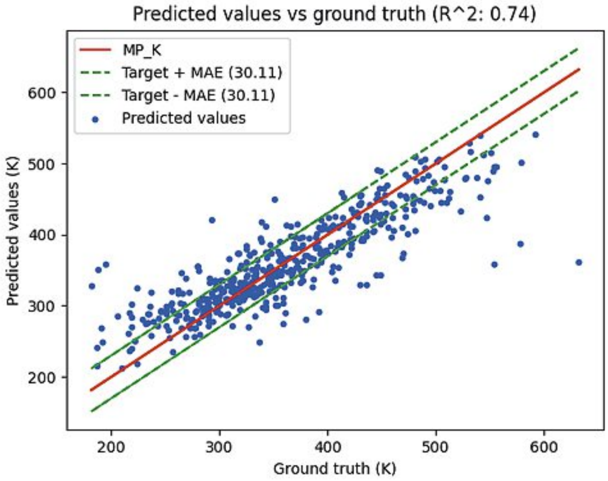

Fig. 2. Plot of predictions vs ground truth of MPs on a random test set. The prediction from the *CatBoost* model yields an R^2 score of 0.74.

We compare the built model with the naïve ones: *random draw* and *all to mean*. The first one is constructed by assigning an empirical distribution to the MP values of a random training set and assigning to each IL of the associated test set a random draw from the distribution. The second one is quite intuitive: we compute the mean value of the MPs in the training set and assign this value to each IL of the test set. The naïve models obtained R^2 scores of 0.02 and 0, respectively. We see that the constructed model ($R^2 = 0.75$) improves drastically the accuracy of the naïve ones.

We emphasize the fact that the ratio of false positive is of special interest in this work, since we are interested in ILs as PCM, and it must be taken with caution. Indeed, changing the interval of interest I would give totally different results depending on the width and position of the interval. In this work, the false-positive ratio of 12% will be of great help for further experimental processes, as it can drastically reduce the number of candidates. Indeed, it means that the model is capable of classifying the molecules with respect to the interval of interest I with an average accuracy of 88%. It will allow researchers to better preselect the ILs before experimentally confirm their MPs and search for the molecules with highest heat of fusion in the particular context of heat storage.

As a MP predictive model, this work falls in the range of the previous models in the literature, even though models with better R^2 values have been constructed, as can be seen in Table 1. Those models, however, are built on half as many molecules to predict from and are thus less general. In Low *et al.*,[10] the authors constructed a Kernel Ridge Regression (KRR) model yielding an R^2 of 0.76 on the same data set to the one considered here, up to 37 molecules. The results in our work are very similar, but the generation and selection process of descriptors is greatly eased for direct use in laboratory. In Acar *et al.*,[12] a deep analysis shows that the model does not perform well for all types of ILs, with R^2 scores of around 0.6 for certain clusters in their dataset.

Table 1. Comparison of previous prediction models for the MP of large sets of ILs.

Reference	Method	Molecules	R^2
Our work	*CatBoost*	**2249**	0.75
[10]	KRR	2212	0.76
[11]	GC	799	0.82
[12]	DL	1253	**0.90**
[13]	GB	2212	0.66
[14]	ANN	799	0.54

The model presented here has two clear advantages over the previous ones presented in the literature, especially that from Low *et al.*:[10] one only needs the molecules' names or SMILES and the descriptor generation software, *mordred*, is freely accessible, as opposed to other chemoinformatic software such as DRAGON[28] or AlvaDesc.[29] We also see from Table 1 that increasing the number of molecules in the data set does not always help increase the accuracy of the model. Indeed, as it has been said previously, measuring MPs of ILs is quite difficult and errors occur often. Thus, increasing the number of ILs can increase the error of measurement.

4. Conclusion

In this work, we proposed a complete MP prediction model for ILs for which only the molecule names are needed, with various feature selection methods. We optimized the ratio of false positive F_p of the predictive model in the context of MP of interest in I = [298.15 K, 323.15 K] and arrived at $F_p = 12\%$.

We are convinced that ILs will be of great use in the energy crisis we are currently in. In this context, the MP is one of the most important thermochemical properties to know. That is why the presented model can be of great help for researchers as it will reduce the time spent in the laboratory for candidate selection.

Acknowledgments

We thank Denis Marti (University of Fribourg), Roger Marti and Beat Wolf (School of Engineering and Architecture of Fribourg) for very useful discussions.

Received: July 18, 2023

[1] T. Endo, K. Sunada, H. Sumida, Y. Kimura, *Chem. Sci.* **2022**, *13*, 7560, https://doi.org/10.1039/D2SC02342C.

[2] G. Kaur, H. Kumar, M. Singla, *J. Mol. Liq.* **2022**, *351*, 118556, https://doi.org/10.1016/j.molliq.2022.118556.

[3] V. D. Bhatt, K. Gohil, A. Mishra, *Int. J. Chem. Tech. Res.* **2010**, *2*, 1771.

[4] R. Datta, R. Ramprasad, S. Venkatram, *J. Chem. Phys.* **2022**, 156, https://doi.org/10.1063/5.0089568.

[5] L. Bai, J. Zhu, B. Chen, *Fluid Phase Equilib.* **2011**, *312*, 7, https://doi.org/10.1016/j.fluid.2011.09.005.

[6] Z. Dai, Y. Chen, C. Liu, X. Lu, Y. Liu, X. Ji, Chin. *J. Chem. Eng.* **2021**, *31*, 169, https://doi.org/10.1016/j.cjche.2020.10.040.

[7] J. Han, M. Li, N. Tian, C. Liu, Y. Zhang, Z. Ji, X. Sun, *Fluid Phase Equilib.* **2023**, *565*, 113675, https://doi.org/10.1016/j.fluid.2022.113675.

[8] J. O. Valderrama, L. F. Cardona, *J. Ion. Liq.* **2021**, *1*, 100002, https://doi.org/10.1016/j.jil.2021.100002.

[9] P. Wasserscheid, T. Welton, 'Ionic liquids in synthesis', Wiley Online Library, **2008**.

[10] K. Low, R. Kobayashi, E. I. Izgorodina, *J. Chem. Phys.* **2020**, 153, https://doi.org/10.1063/5.0016289.

[11] F. Gharagheizi, P. Ilani-Kashkouli, A. H. Mohammadi, *Fluid Phase Equilib.* **2012**, *329*, 1, https://doi.org/10.1016/j.fluid.2012.05.017.

[12] Z. Acar, P. Nguyen, K. C. Lau, *Appl. Sci.* **2022**, *12*, 2408, https://doi.org/10.3390/app12052408.

[13] V. Venkatraman, S. Evjen, H. K. Knuutila, A. Fiksdahl, B. K. Alsberg, *J. Mol. Liq.* **2018**, *264*, 318, https://doi.org/10.1016/j.molliq.2018.03.090.

[14] J. O. Valderrama, C. A. Faundez, V. J. Vicencio, *Ind. Eng. Chem. Res.* **2014**, *53*, 10504, https://doi.org/10.1021/ie5010459.

[15] A. R. Katritzky, A. Lomaka, R. Petrukhin, R. Jain, M. Karelson, A. E. Visser, R. D. Rogers, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71, https://doi.org/10.1021/ci0100503.

[16] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, *J. Cheminf.* **2018**, *10*, 1,://doi.org/10.1186/s13321-018-0258-y.

[17] R. Guha, E. Willighagen, *Curr. Top. Med. Chem.* **2012**, *12*, 1946, https://doi.org/10.2174/156802612804910278.

[18] M. Kuhn, K. Johnson, 'Applied predictive modeling', Springer, **2013**.

[19] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, *ACM computing surveys (CSUR)* **2017**, *50*, 1, https://doi.org/10.1145/3136625.

[20] K. Kira, L. A. Rendell, in 'Machine Learning Proceedings 1992', Elsevier, **1992**, pp. 249.

[21] T. Hastie, R. Tibshirani, J. H. Friedman, 'The elements of statistical learning: data mining, inference, and prediction', Springer, **2009**.

[22] J. T. Hancock, T. M. Khoshgoftaar, *J. Big Data* **2020**, *7*, 1, https://doi.org/10.1186/s40537-020-00369-8.

[23] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268, https://doi.org/10.1021/acscentsci.7b00572.

[24] L. Chan, G. R. Hutchison, G. M. Morris, *J. Cheminf.* **2019**, *11*, 1, https://doi.org/10.1186/s13321-019-0354-7.

[25] D. Ulmasov, C. Baroukh, B. Chachuat, M. P. Deisenroth, R. Misener, in 'Computer Aided Chemical Engineering', Vol. 38, Eds. Z. Kravanja, M. Bogataj, Elsevier, **2016**, pp. 1051, 10.1016/B978-0-444-63428-3.50180-6.

[26] C. Banchhor, N. Srinivasu, *J. Big Data* **2021**, *8*, 81, https://doi.org/10.1186/s40537-021-00464-4.

[27] J. Guo, B. Rankovic, P. Schwaller, *CHIMIA* **2023**, *77*, 31, https://doi.org/10.2533/chimia.2023.31.

[28] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, *Match* **2006**, *56*, 237.

[29] A. Mauri, in 'Ecotoxicological QSARs', Ed. K. Roy, Springer US, New York, NY, **2020**, pp. 801, https://doi.org/10.1007/978-1-0716-0150-1_32.