© International Journal of Combinatorial Optimization Problems and Informatics, Vol. 8, No. 2, May-Aug 2017, pp. 2-9. ISSN: 2007-1558.

Use of Neuroevolution to Estimate the Melting Point of Ionic Liquids

Jorge A. Cerecedo-Cordoba.[1], Juan Javier González Barbosa.[1], J. David Terán-Villanueva[1], Juan Frausto-Solís [1], José A. Martínez Flores[1]
[1]*Tecnológico Nacional de México - Instituto Tecnológico de Ciudad Madero*

E-mail: joalceco@gmail.com, jjgonzalezbarbosa@hotmail.com, david_teran01@yahoo.com.mx, juan.frausto@gmail.com

Abstract. The Physical Properties Estimation Problem of Ionic Liquids (PPEPILs) arises from the need of designing Ionic Liquids (ILs) for specific tasks. It is important to emphasize that the synthesis of ILs is generally expensive and time-consuming. Furthermore, the number of possible ionic liquids that can be synthesized is extremely large. The purpose of PPEPILs is to avoid the experimental synthesis of Ionic Liquids (ILs) estimating their physical properties. Moreover, to estimate the melting temperature is the most difficult task. This problem has attracted the attention of interdisciplinary researchers due to their relevant applications such as their usages as catalysts and solvents. Additionally, the ILs are relevant due to their distinctive characteristics and reduced toxicity. This problem is particularly complex since the behavior of ILs is unconventional and the available information may not be accurate. This paper presents a new approach for the PPEPILs based on neuroevolutionary neural networks using molecular descriptors to predict the melting temperatures of ILs with encouraging results. Neuroevolutionary networks had been previously used in diverse areas of knowledge and present advantages over classic Neural Networks.

Keywords. Melting Point, Ionic Liquids, QSPR, Neuroevolution.

1. Introduction

ILs are salts with low melting points, some of them even have melting points as low as room temperature. ILs are compounds of only ions. Ions are molecules such as anions and cations. The ILs are ideal for various applications for their singular characteristics. Some applications of ILs are as compounds in paint, Lithium-ion batteries, and air products. Plechkova et al. [1] presented a larger list of applications in the chemical industry. However, the principal application of ILs is their usage as solvents and substitutes of volatile organic solvents. The main characteristics of ILs [2]–[4]:

a) They have a high polarity.
b) They remain stable in presence of organic compounds and catalyst.
c) They remain in liquid state in a large temperature range.
d) Their toxicity level is low compared to solvents with similar properties.

However, the number of anion-cation combinations that could produce a IL is as large as 10^{18} [3]. Moreover, there is no known database with all the information related to physical properties of the currently synthetized ILs. Thus, the selection of "the best" IL for a specific application by experimentation or by using a computational algorithm is almost impossible due to:

a) The amount of ILs that can be synthetized is too large.
b) The physical properties of most ILs are unknown.
c) Known physical properties of ILs differ between sources [4], [5].

The synthesis of ILs requires a considerable investment of time and money. Thus, a large-scale study of all ILs is unrealistic. One of the most important properties in the selection of an ideal IL is the melting point. Lower melting points prevent the solidification of the liquid. Therefore, the estimation of ILs estimation points is important because they produce valuable data of ILs without the investment of a large study. Also, this enables the possibility of designing efficient ILs with lower costs.

Received Jun 19, 2016 / Accepted Dec 10, 2016

This paper proposes a neuroevolutionary network model to predict the melting point of ILs. This paper is organized as follows: Section two presents an overview of the related work. Section three presents an introduction to Neuroevolution models. Section four shows our proposal including the general methodology. In section five the obtained results and their discussion are presented. Finally, sections six and seven contain conclusions and future work respectively.

2. Related work

Several works have addressed the problem of determining the melting temperature of ILs. This problem was firstly presented by Katritzky et al. [2] who proposed to use molecular descriptors as inputs on a traditional linear regression model [2], since then many works follow a similar methodology with different datasets [3], [6]–[10]. However, traditional prediction methods do not solve the problem thoroughly, which has led to the usage of machine learning methods for prediction. Carrera et al. [11] proposed a Regression Forest method for melting point estimation; that model lead to a better performance than the previous works due to their non-linear approach.

Varnek et al. [5] proposed a collection of linear regression methods and machine learning algorithms applied to the prediction of melting points of ILs that include Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and K-Nearest Neighbors algorithm (KNN). In their work, they validated the advantages of non-linear methods and show that SVMs and ANNs outperform the traditional linear methods.

Bini et al. [12] and Torrecilla et al. [13] used a recursive neural network and a feed-forward neural network for estimating the melting temperature of ILs. In both works, it was shown that the neural networks possess a high level of robustness and quality of the results. However, it is known that ANNs can be stagnated in local optima. Also, these algorithms require manual tuning for obtaining good results. This is a time-consuming process were multiple topologies and parameters must be tested.

Since the ILs data available is known for containing errors, the estimation algorithm should have some tolerance to erroneous data. Also, it is desirable that the algorithm can navigate in the solution space without getting stagnated in local optima. Lastly, a correct tuning of the parameters is required for obtaining good results. However, most of these parameters are not constant along the execution of the algorithm. Thus, an automatic parameter tuning method is desirable. Therefore, the algorithm used for this problem should have the next features:

a) Robustness to erroneous data.
b) An automatic tuning process for the parameters.
c) A local-optima escape mechanism.

Table 1 shows a comparison of our evolutionary approach with previous works. Here is shown that none of these works have the three features that were previously discussed.

Table 1. Comparison of previous works and the neuroevolutionary approach.

Algorithm	Features		
	Robust to erroneous data	Auto-tuning	Local optima escape mechanism
KNN [5]	✗	✗	✗
Regression Trees [11]	✗	✗	✗
SVM [5]	✓	✗	Kernel tuning is needed
ANN [5], [12], [13]	✓	✗	✗
Neuroevolution (This Work)	✓	✓	✓

In this paper, we explore the usage of neuroevolutionary neural networks in the prediction of melting points of ILs. We are interested in this approach since the neuroevolution can predict non-linear behaviors without an extensive study of the ideal topology for a specific problem. The neuroevolutionary neural networks had been successful in various fields of science [14][15]. This type of neural networks has had a low application domain in similar problem despite their multiple benefits, such as their auto-tuning and optima escape qualities. Also, since the training algorithm is a metaheuristic method, is possible to apply numerous techniques such as crossovers, mutations, memory mechanisms and local searches. In this work, we assess the neuroevolutionary performance as a prediction mechanism for the physical properties of ionic liquids.

3. Neuroevolution

Regression methods are predictive modeling methods that seek the inherent relationships between one or several input entities with different outputs. That is, it searches for an unknown function that transforms the independent variables (input) to dependent variables (output). Regression analysis is commonly used in areas such as medicine, finance, chemistry, mechanics, and others. It is an indispensable tool for analyzing and modeling information with an unknown behavior.

Statisticians have developed increasingly advanced methods of regression over the years. Linear regression is still a good choice when a very simple model for a basic prediction task is required. Linear regression also tends to work well in sparse, non-complex data sets. Nevertheless, it has difficulties predicting nonlinear behaviors. Recently, machine learning methods have been used as alternatives to predict nonlinear relationships. Machine Learning is a branch of computer science that studies the development of methods that allow computer programs to have pseudo-learning. Learning is the acquisition of knowledge through study or experience. In the specific case of Machine Learning, the experience usually refers to past information or records available.

The Machine Learning methods have different learning schemes; they determine how the algorithm should learn from a series of input data and how to manipulate them. The selection of the learning method is based on the nature of the data and the goal to achieve with the learning task. The nature of the input data plays an important role in the algorithm training. The possibility that some data are missing, irrelevant or some noise information is undeniable. Therefore, identifying the omitted and irrelevant variables is an important issue.

Currently, the Artificial Neural Networks (ANN) are present in many systems thanks to their learning ability, self-organization, and handling of diffuse, noisy, and inconsistent data. Besides, ANNs, perform efficiently in a wide variety of applications [16], [17]. ANNs excel in the discovery of patterns between inputs and outputs.

However, traditional training methods of ANNs such as backpropagation have the disadvantage of being deterministic and often end up truncated at local optima. In addition, another problem with the use of ANNs comes from the need to find an effective configuration of the number of neurons and their distribution in the layers; this problem is even more critical in the development of deep learning networks.

An alternative to ANNs is Neuroevolution, this method consists of training ANNs through evolutionary approaches. In other words, an evolutionary algorithm modifies the weights and structure of an ANN. These types of networks are also known as TWEANNs (Topology & Weight Evolving Artificial Neural Network algorithms). Historically, TWEANNs have improved their training techniques through the incorporation of new components, examples of such components are minimal topologies and Speciation [18]. TWEANNs principal approach is the training with reinforcement learning to accomplish several tasks. Among the main components available for neuroevolution are:

1) **Genetic parameters.** The parameters of metaheuristics are the main factors to consider when performing training. Parameters such as population size and the number of evaluations have an important effect on the execution time and in the performance.

2) **Genotype and phenotype**. The representation of the neural network and how it manipulates the metaheuristic algorithm, and allows or denies the use of diverse techniques. The coding and manipulation of the neural network significantly escalate the complexity of the algorithm.

3) **Genetic operators.** The use of classical selection, crossover and mutation operators show their insufficiency due to the change of coding used in these algorithms. This justifies the design of new genetic operators.

4) **Initial topology.** Neuroevolution algorithms tend to increase the topology of networks over time, this also increases processing time. Complex solutions in the early stages of training are counterproductive for time wasted on poor quality solutions. Choosing the complexity of the initial networks affects the performance of the population.

5) **Life cycle.** Training poor quality networks is not ideal, not only for processing time but also for the quality of the population in general. Identifying the effectiveness of a topology is a process that involves computational time in which weights can fit the new topology.

4. Neuroevolution Model for ILs

The ILs estimation model proposed is divided into two stages, training, and prediction. The training uses a transformation of the data to molecular descriptors. The molecular descriptors calculated are the input data for our training. At first, a group of randomly generated ANNs is created; this set will function as the population. Then the training of all the population is carried out through a genetic algorithm. After the training is concluded the best ANN is selected from the population and it will be the model used in the estimations. Figure 1 shows the general procedure for training models and predictions. As can be seen there are two main phases:

- Training phase: It involves the preparation of the data and the training of the predictive models.
- Estimation Phase: It uses the best ANN found in training in order to make estimations with the best precision possible.

The different components of these phases are explained in the next sections.

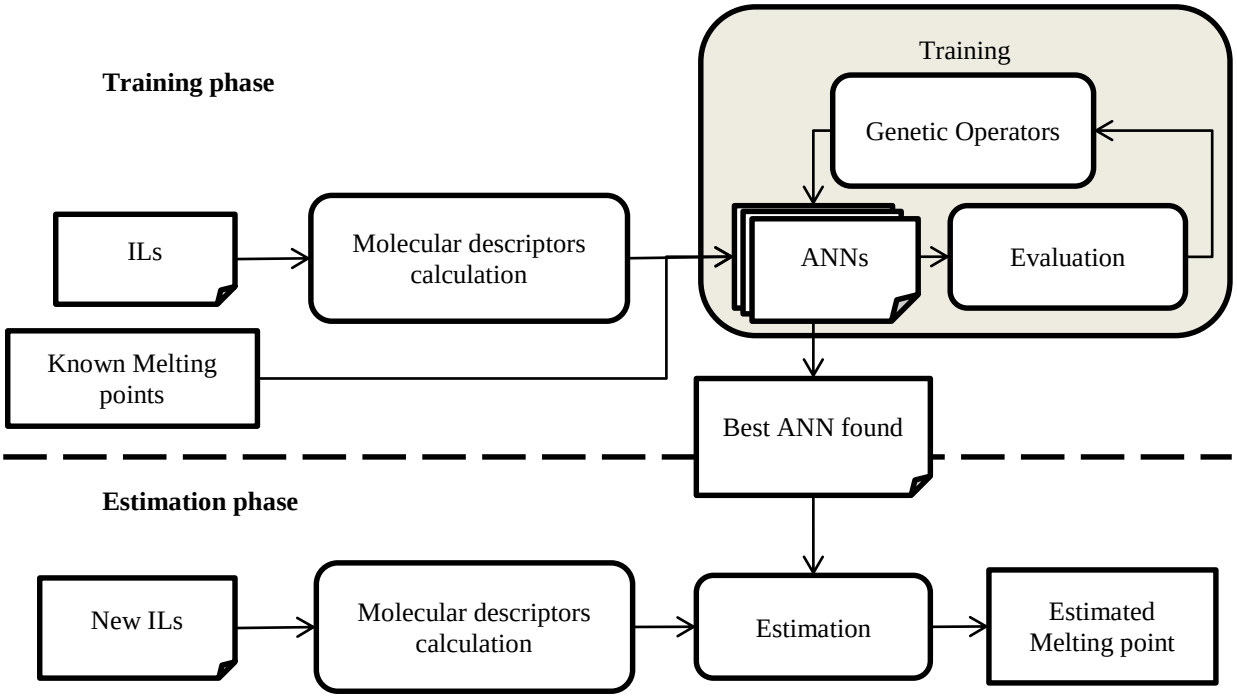

Figure 1. Methodology for estimation and prediction.

4.1 Molecular descriptors calculation

The data collected belongs to the literature [19]. The data is composed of 43 Ionic Liquids. We transformed these Ionic Liquids to a digital format using the software JChem v16.8.8.0 [20]. The software allows drawing and designing molecules through a friendly user interface, also the software can calculate some molecular descriptors. JChem also offers several tools for visualizing chemical structures.

Quantitative structure-activity relationships (QSARs) are mathematical models that attempt to represent the relation between independent and dependent variables, QSAR is useful to find the relationship formed between the characteristics derived from a compound with respect to its biological activity, physicochemical properties, or its toxicity [21].

Molecular descriptors are numerical values that describe the structure and other characteristics of a molecule; such data is obtained through a well-defined algorithm which takes the information from a molecular representation. The field of molecular descriptors is interdisciplinary, as it involves algebra, graph theory, computational chemistry and physical chemistry [22].

There is a variety of software in the literature that can perform the calculation of descriptors. In particular, for this paper, we used the free software PaDEL [23]. The software PaDEL uses “mol" or "smiles" files, such files contain a digital representation of the ionic liquids, and are used to calculate various descriptors and generates a file which concentrates all the information. PaDEL generates 1445 different descriptors with their native configurations. Then, due to the many descriptors produced by this software, it is convenient to select a smaller set of descriptors to train the artificial neural network. Some of the descriptors are easily discarded because the software is not able to calculate them or the descriptors are constant through the set of ILs. A reduced set of descriptors of size four was generated by selecting those with the highest coefficient of correlation between the descriptor and the melting point temperature.

4.2 Training

For the neuroevolution model proposed in this work the neural network scheme know as NEAT (NeuroEvolution of Augmenting Topologies) [18] was selected. In addition, we implement it with the Encog framework [24].

NEAT uses an encoding that contains the structure and weights of the neural network. Figure 2 contains an example of the NEAT encoding. A single solution (a neural network) is a two vectors compound, the first one corresponds to the nodes of the network, and the second one is related to the connections between nodes.

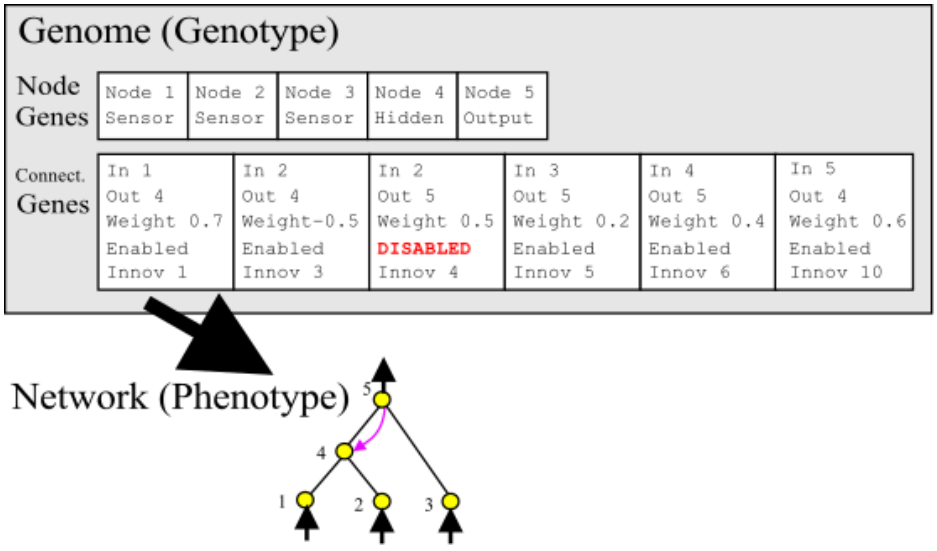

Figure 2. NEAT Encoding [18].

NEAT performs a perceptron-like feed-forward network with a minimalistic set of input neurons and output neurons. While the evolution process improves, the complexity of the networks may grow. The topology of the networks can be modified either by inserting a new neuron into a connection path or by creating a new connection between neurons.

The encoding of NEAT requires special crossover and mutation operators to ensure the feasibility of the offspring produced [18]. The crossover procedure aligns both selected parents. The common nodes and connections are passed directly to the offspring, also disjoints and excesses. If a connection is disabled in any of the parents, that connection will be disabled in the offspring. Figure 3 shows an example of the crossover procedure previous explained. The mutation and crossover operator allows the genetic evolution of a group of ANNs. It is intended that different types of ANNs evolve at the same time in order to discover the best topology for this problem.

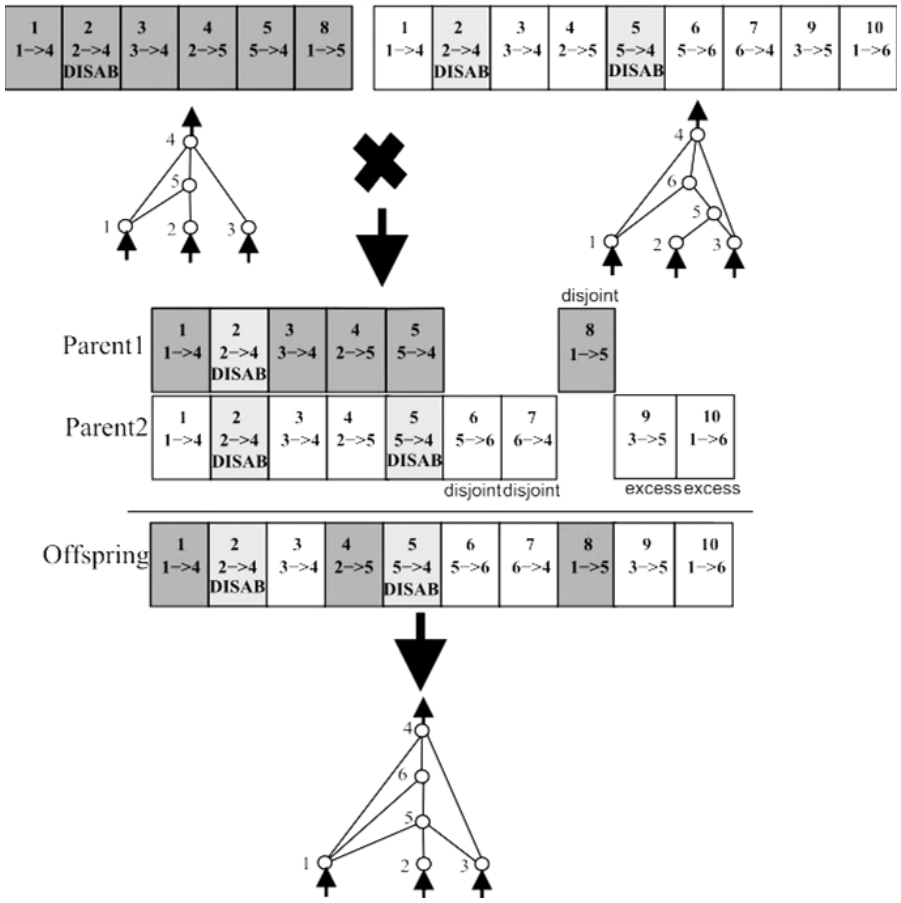

Figure 3. Crossover for NEAT Encoding [18]**.**

5. Results and Discussion

The purpose of a prediction model is to allow estimations of a previously unknown behavior with relative accuracy. This behavior is implicit in the available training data set. Traditionally, in the validation of machine learning models, the data is divided into training and validation. The division of the data into training and validation helps to evade overfitting since it is possible to obtain measurements of error of the model with an independent set of training; effectively capturing the generality of the model.

The available dataset was divided into the training set (70%) and test set (30%). Considering the size of the training set available, a cross-validation of four folds was chosen. A set of 100 ANNs was trained and the optimization goal was the minimization of the training error.

This training was made in a cross-validation manner. The final model was then validated with the test set to assure the generalization of the model. The experiment was executed 30 times to obtain a reliable measure. The root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE) where calculated for the training and validation sets, in addition, a combined set of all ILs in this study was used. The averaged results are shown in Table 2. The RMSE and MAE are in kelvin while MAPE represents the perceptual error.

A direct comparison with the state of art would be unfair due to the different data sets and our limited sample size. Another impediment is the differences in tuning parameters which are not replicable between works. Nevertheless, a metric comparison shows that the results are alike. Proof of this is that Katritzky [2] obtained a slightly better prediction with a MAE of 23.1833K in training error a 25.56K in validation error. Varnek [5] reported a RMSE of all his experiments in a range of 37.5 to 46.4K. Bini [12] obtained a MAE of 23.07K and an RMSE of 29.97K in his validation set. A neuroevolutionary approach to the PPEPILs was never done before and the results show that the proposed methodology is satisfactory. However, future work is needed to improve the quality of the predictions, and an expansion of the dataset is also needed.

Table 2. Experimental results.

Metrics	Data set		
	Training	Validation	Combined
RMSE	32.9870K	42.8398K	36.1498K
MAE	27.2837K	37.0646K	30.1464K
MAPE	6.27%	6.69%	6.38%

6. Conclusions

In this paper, a neuroevolution model for prediction the melting points of ILs is presented. This model is a part of the physical Properties Estimation Problem of Ionic Liquids (PPEPILs). This prediction was made using the NEAT technique. The use of neuroevolution networks is a novel proposal in the area and the results obtained show that it is possible to make estimations with an acceptable precision degree. It is expected that increasing the number of ILs in the training set will achieve a better generalization of the models. Thus, it is possible to achieve a better prediction of melting temperature of ILs despite possible erroneous or missing data. Nevertheless, we consider that the PPEPILs is not still a solved problem and it represents and open area.

7. Future Work

A better selection of molecular descriptors is needed; this can help to obtain lower prediction errors. We are working on this topic which we hope to publish in near future. Besides, the method presented in this article opens the possibility of using other evolutionary algorithms combined with Neuroevolution strategies. We are currently working to improve the genetic algorithm for obtaining better results. Another future work is to define different genotypes of neural networks to analyze their impact on the prediction. Another opportunity area is to modify the genetic operators for obtaining a wide selection of techniques to determine the best for this problem. Also, different components as memory and local searches will be further explored in future work.

Acknowledgements

The authors would like to acknowledge with appreciation and gratitude to CONACYT, TECNM and PRODEP. Also, acknowledge to Laboratorio Nacional de Tecnologías de la Información in the Instituto Tecnológico de Ciudad Madero for the access to the cluster. This work has been partial supported by CONACYT Project 254498. Jorge A. Cerecedo-Cordoba and J. David Terán-Villanueva would like to thank the supports 434694 and 177007.

8. References

1. Plechkova, N. V., Seddon, K. R.: Applications of ionic liquids in the chemical industry. Chemical Society reviews. 37, 123–150, (2008)
2. Katritzky, A. R., Lomaka, A., Petrukhin, R., Jain, R., Karelson, M., Visser, A. E., Rogers, R. D.: QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids. Journal of Chemical Information and Computer Sciences. 42(1), 71-74, (2002)
3. Katritzky, A. R., Jain, R., Lomaka, A., Petrukhin, R., Karelson, M., Visser, A. E., Rogers, R. D.: Correlation of the Melting Points of Potential Ionic Liquids (Imidazolium Bromides and Benzimidazolium Bromides) Using the CODESSA Program. Journal of Chemical Information and Computer Sciences. 42(2), 225–231, (2002)
4. Wasserscheid, P., Welton, T.: Ionic liquids in synthesis (Vol. 1). Weinheim: Wiley-Vch, (2008)
5. Varnek, A., Kireeva, N., Tetko, I. V., Baskin, I. I., Solov'ev, V. P.: Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points?. Journal of chemical information and modeling, 47(3), 1111-1122, (2007)
6. Trohalaki, S., Pachter, R.: Prediction of melting points for ionic liquids. QSAR & Combinatorial Science, 24(4), 485-490, (2005)
7. Sun, N., He, X., Dong, K., Zhang, X., Lu, X., He, H., Zhang, S.: Prediction of the melting points for two kinds of room temperature ionic liquids. Fluid phase equilibria, 246(1), 137-142, (2006)
8. López-Martin, I., Burello, E., Davey, P. N., Seddon, K. R., Rothenberg, G. Anion and cation effects on imidazolium salt melting points: a descriptor modelling study. ChemPhysChem, 8(5), 690-695, (2007)
9. Farahani, N., Gharagheizi, F., Mirkhani, S. A., Tumba, K.: Ionic liquids: Prediction of melting point by molecular-based model. Thermochimica acta, 549, 17-34, (2012)

10. Eike, D. M., Brennecke, J. F., Maginn, E. J.: Predicting melting points of quaternary ammonium ionic liquids. Green Chemistry, 5(3), 323-328, (2003)
11. Carrera, G., Aires-de-Sousa, J.: Estimation of melting points of pyridinium bromide ionic liquids with decision trees and neural networks. Green Chemistry, 7(1), 20-27, (2005)
12. Bini, R., Chiappe, C., Duce, C., Micheli, A., Solaro, R., Starita, A.,Tiné, M. R.: Ionic liquids: prediction of their melting points by a recursive neural network model. Green Chemistry, 10(3), 306-309, (2008)
13. Torrecilla, J. S., Rodríguez, F., Bravo, J. L., Rothenberg, G., Seddon, K. R., Lopez-Martin, I.: Optimising an artificial neural network for predicting the melting point of ionic liquids. Physical Chemistry Chemical Physics, 10(38), 5826-5831, (2008)
14. Stanley, K. O., Bryant, B. D., Miikkulainen, R.: Real-time neuroevolution in the NERO video game. IEEE transactions on evolutionary computation, 9(6), 653-668, (2005)
15. Koppejan, R., & Whiteson, S. (2011). Neuroevolutionary reinforcement learning for generalized control of simulated helicopters. Evolutionary intelligence, 4(4), 219-241.
16. Altamiranda, J., Aguilar, J., Hernández, L.: Sistema de reconocimiento de patrones de sustancias químicas cerebrales basado en minería de datos. Computación y Sistemas, 19(1), 89-107, (2015)
17. Guzmán, D. A. I., Alarcón, J. R. C., Torres, A. A., Bárcenas, M. A. M.: Design of an Artificial Neural Network to Detect Obstacles on Highways through the Flight of an UAV, Computación y Sistemas, 19(1), 31-40, (2015)
18. Stanley, K. O., Miikkulainen, R.: Evolving neural networks through augmenting topologies. Evolutionary computation, 10(2), 99-127, (2002).
19. Zhang, S., Lu, X., Zhou, Q., Li, X., Zhang, X., Li, S.: Ionic Liquids: Physicochemical Properties. Elsevier, (2009)
20. JChem, version v16.8.8.0, http://www.chemaxon.com
21. Roy, K., Kar, S., & Das, R. N.: A Primer on QSAR/QSPR Modeling: Fundamental Concepts. Springer, (2015)
22. Todeschini, R., Consonni, V., Mannhold, R., Kubinyi, H., Timmerman, H.: Handbook of Molecular Descriptors, (2000)
23. Yap, C. W.: PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of computational chemistry, 32(7), 1466-1474, (2011)
24. Heaton, J.: Encog: Library of interchangeable machine learning models for java and c#. Journal of Machine Learning Research, 16, 1243-1247, (2015)