Towards the application of structure-property relationship modeling in material science: Predicting the Seebeck coefficient for ionic liquids/redox couple system

Anita Sosnowska[a], Maciej Barycki[a], Agnieszka Gajewicz[a], Maciej Bobrowski[b], Sylwia Freza[c], Piotr Skurski[c], Stefanie Uhl[d], Edith Laux[d], Tony Journot[d], Laure Jeandupeux[d], Herbert Keppner[d] and Tomasz Puzyn[a,*]

[a]Laboratory of Environmental Chemometrics, Department of Chemistry, University of Gdansk Wita Stwosza 63, 80-308 Gdansk (Poland)

[b]Department of Technical Physics and Applied Mathematics, Gdansk University of Technology, Gdansk (Poland)

[c]Department of Chemistry, University of Gdansk, Gdansk, Wita Stwosza 63, 80-308 Gdansk (Poland)

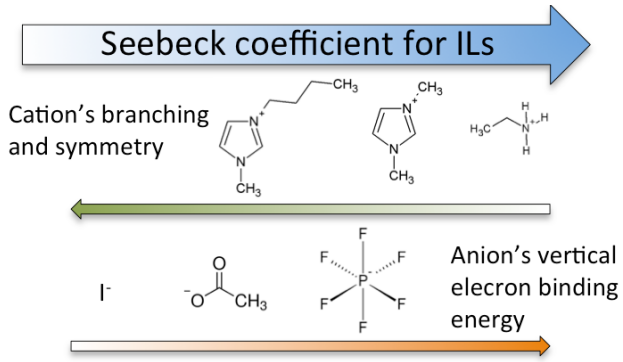
[d]HES-SO Arc, Institut des Microtechnologies Appliquees, La Chaux-de Fonds (Switzerland)

*Corresponding author: Tomasz Puzyn, tel.: (+48 58) 523 5248,

e-mail address: t.puzyn@qsar.eu.org

Table of contents:

Graphical abstract:



Abstract

This work focuses on determining the influence of both ionic liquid type and redox couple concentration on Seebeck coefficient values of such system. Thanks to their properties (especially high thermal stability), ionic liquids are very promising alternative for the electrolytes used in the Thermoelectrochemical cells (TECs). Our work covers the experimental and theoretical approach on Seebeck effect phenomenon, which is one of the main features responsible for performance of TEC devices.

The quantitative structure-property relationship (QSPR) and read-across techniques are proposed as the methods of identifying structural features of ionic liquids (ILs) (mixed with LiI/I_2 redox couple), which influence the Seebeck coefficient (S_e) values. ILs consisted of small, symmetric cations and anions indicating high values of the vertical electron binding energy are recognized as ones having the highest values of S_e.

In addition, the developed QSPR model enables predicting the values of S_e for each IL that belongs to the applicability domain of the model. The influence of the redox couple concentration on the Seebeck coefficient values is quantitatively described as well. Thus, it is possible to calculate, how the value of S_e will change with changing redox couple concentration. The presence of LiI/I_2 redox couple in lower concentrations increases the values of S_e as expected.

Key Words:

Ionic liquids; ILs; read-across analysis; QSPR; Seebeck coefficient

1. Introduction

In many industrial applications (melting furnaces, industrial kilns, incinerators, power plants, etc.) a large amount of heat energy remains unused. Conversion of unemployed heat energy to electricity is nowadays one of the most important tasks in the field of thermoelectricity. Optimization of the efficiency of energy converting devices (like thermoelectrochemical cells - TECs) is a big challenge of twenty-first century. The performance of such devices depends on different factors[1-4]. Recently, researches performing conversion from thermal to electrical energy have been focused on the solid-state devices using thermoelectric semiconductor materials, in which the energy could be converted via Seebeck effect.[5] The magnitude of Seebeck effect is described by the Seebeck coefficient (S_e), which determines the open circle voltage that can be produced by the device at any given temperature differences (Equation 1):

$$S_e = -\Delta V / \Delta T \qquad (1)$$

where: ΔV - is the voltage difference between the hot and cold electrodes.

However the effectiveness of semiconductor devices depends largely on the quality of the thermoelectric material, which they are composed of. What is more, there is a limitation of obtained S_e (≤ 1 mVK^{-1}) for such generators.[6, 7] In this case, the liquid-based thermoelectrical technology is a promising alternative for direct energy conversion (from thermal to electric). The main advantage of this approach is the choice of a solvent, which may influence the kinetics of electron transfer reaction.[8, 9] Recently, aqueous electrolytes[10] and organic solvents[11] were utilized in the thermoelectrochemical devices. However, these types of solvents are limited because of their volatility and limited long-term durability. In this case, ionic liquids (ILs) – molten salts having its melting point lower than 100 °C, seem to be an ideal choice that can replace traditional electrolytes. ILs are characterized by low flammability, low vapor pressure,

enlarged stability at high temperatures and ability to retain the liquid state for a wide range of temperatures. Recent studies proved that ionic liquids in the presence of redox couple (RC) are used with a great success as electrolytes in liquid-based thermoelectrical cells,[9, 12, 13] but still, there is a strong need to look for ILs which can provide higher Seebeck coefficient. Experimental measurement focused on obtaining the Seebeck coefficient are very time consuming and expensive. In this case, the computational methods that enable predicting missing data in relatively short time, without necessity of performing additional experiments are very helpful. Here we applied the technique based on structural relation between substances (read-across) and structure-property relationship modeling (QSPR).

In principle, the read-across approach is based on the assumption that chemicals that are structurally alike, or follow a regular structural pattern, should exhibit similar physicochemical and toxicological properties.[14] Once similar chemicals have been merged together, endpoint information (e.g. physicochemical property) for one, or more, chemical(s) (the so-called "source chemical(s)") can be used to make predictions of the same physicochemical property for another chemical (the "target chemical").[14] It needs to be highlighted that although qualitative read-across is usually applied at the first stage of data exploration, it can in fact lead to many valuable conclusions. It allows identifying structural features responsible for the specific physicochemical properties or classifying the chemicals according to their potential toxic effects.[15]

Second computational method is structure-property relationship modeling (QSPR). QSPR approach is based on the assumption that the variance of physicochemical property in the group of similar compounds can be predicted using the variance of their chemical structures, represented by the range of numerical features, so-called molecular descriptors.[16]

In this work, we focused on determining the effect of the structural features of cation

and anion, which influences the value of the Seebeck coefficient. First, we used the experimental values of S_e for ILs with addition of LiI/I_2 redox couple to defined classes of ionic liquids, based on structure-Seebeck coefficient relationship using read-across approach. Then, we developed a single QSPR model to estimate the Seebeck coefficient for ionic liquids (with 0.01M LiI/I_2 redox couple). Finally, we proposed a simple arithmetic relationship that can be employed to estimate S_e values in different redox couple's concentration.

2. Results and discussion

Since our experimental data for the Seebeck coefficient were obtained within the one protocol (the same experimental conditions – please see Experimental section), we started computational modeling. The molecular models of each ionic liquid were built and then optimized by employing the Density Functional Theory. We used the equilibrium structures of the cations and anions to calculate molecular descriptors. The most optimal combination of the molecular descriptors to be utilized in the read-across and QSPR modeling were selected by employing the genetic algorithm[17] implemented in the QSARINS software.[18] [19] Read-across analyses, as well as the QSPR modeling were performed using the values of Seebeck coefficient for ILs with addition of 0.01M LiI/I_2 redox couple (for more details please see Experimental section).

2.1. Approximating Seebeck coefficient with read-across technique

The main goal of qualitative read-across technique is to identify classes containing chemicals of similar structure and properties, without using pre-established class memberships (so-called unsupervised pattern recognition approach).[15] Since majority of cluster analysis methods are based on the assumption that two samples (here: chemicals) are similar, when are located close (i.e., in terms of a given distance measure) to each other, two-ways hierarchical cluster analysis (t-HCA)[20, 21] was applied to explore similarity between the ionic liquids.

As previously mentioned, molecular descriptors that define the similarity of ILs have been selected with the genetic algorithm. There were $GATSi^{A}$ (Geary autocorrelation of lag 1 weighted by ionization potential for anion's structure)[22] and $piPC04^{C}$ (molecular multiple path count of order 4 calculated for cation)[23]. The correlation coefficients between the descriptors and the Seebeck coefficient were -0.81 and -0.43 $GATSi^{A}$ and $piPC04^{C}$, respectively. This means, $GATSi^{A}$ descriptor explains approximately 66% while, $piPC04^{C}$ descriptor explains 18% of the total variance in the values of S_e in the group of the studied ILs.

Dendrogram **(Figure 1A)** resulted from t-HCA performed on the matrix containing two descriptors ($GATSi^{A}$ and $piPC04^{C}$) and the Seebeck coefficient (in columns) calculated for the training set of ionic liquids (in rows) reveals four main classes of ILs. Compounds within each class are characterized by similar values of the S_e. One can observe that the S_e of ILs systematically decreases when moving from class **I** to class **IV**.

In the next step, the identified classes have been used to predict the values of Seebeck coefficient ILs from the validation set (ILs not previously used for defining the classes). The compounds were assigned into the classes based on their structural similarity to the particular members of the training set. To make it possible, data from the validation set had to be scaled first. The scaling procedure was repeated for every descriptor. We subtracted the mean value calculated for the training set from every descriptor value in the validation set and divided the result by the standard deviation calculated for the training set. We found that IL0011, IL0021 and IL0012 were assigned to the class **II**, IL0134, and IL0009 to the class III, while IL0051 was predicted to be a member of class **IV**. The obtained estimation for the majority of ILs was finally confirmed by two-dimensional hierarchical cluster analysis for all ionic liquids from both: training and validation sets **(Figure 1B)**.

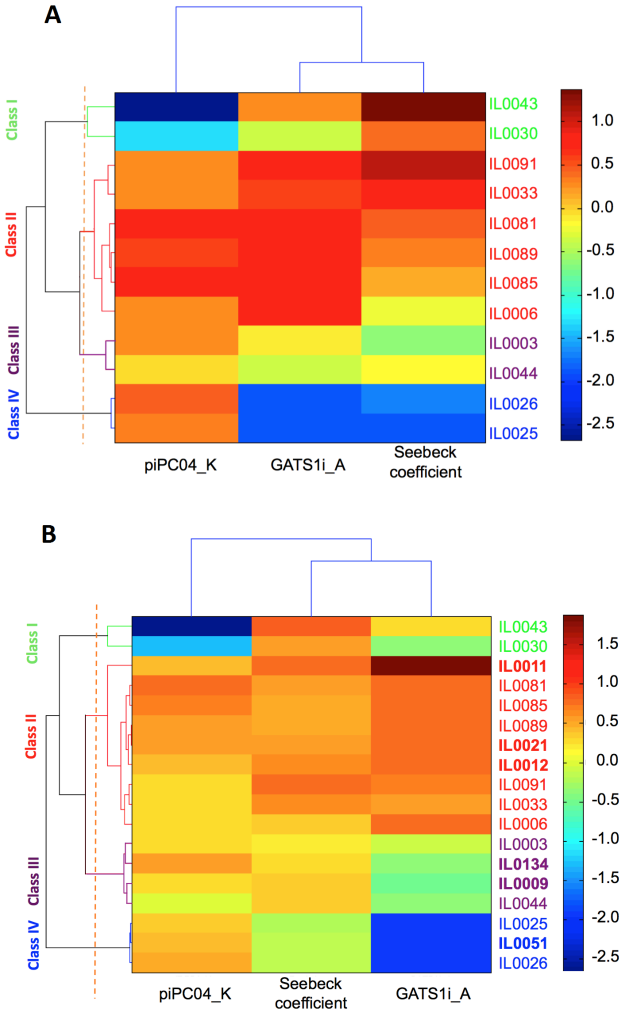


Figure 1. (A) - Two-dimensional cluster analysis. Four identified natural clusters (classes) in the data are presented. Colors represent the auto-scaled values of the selected descriptors and experimentally determined S_e values, (B) - Two-dimensional cluster analysis for all ionic liquids from both: training and validation sets.

By analyzing the results obtained, it can be seen that Seebeck coefficient systematically decreases when moving from class I to class IV. The observed trend is in accordance with the results of QSPR studies. Interestingly, there were false negative predictions for two ILs. The read-across model failed to predict the high Seebeck coefficient for IL0011 and IL0030. A possible explanation is that both ILs are extremely close to the lower limit of the appropriate class and that was the cause of the misclassification (for more details, please refer to Supporting Materials (Figure S1).

To verify goodness-of-fit and externally validated predictive ability of the read-across model for estimating the Seebeck coefficient the accuracy (A) and the error rate (E) were calculated. The results were as follows: $A_{training}$=83.3%, $E_{training}$=16.3%, $A_{validation}$=100%, and $E_{validation}$=0%. Since the value of accuracy was much higher than the error rate, the significance of the model was confirmed.

Moreover, the predictive potential of the proposed approach was additionally confirmed by employing Spearman's rank correlation test. It was found that the values obtained from the read-across technique did not differ significantly from those measured experimentally (ρ_S = 0.648, p = 0.076), as well as from those predicted from the QSPR model (ρ_S = 0.918, p = 0.0001). Therefore, the presented technique is sufficiently accurate to identifying groups of ionic liquids having similar properties as well as to filling data gaps in qualitative manner. For more details, please refer to Supporting Materials (**Table S4** and **Table S5**).

2.2. Predicting Seebeck coefficient with QSPR approach

The developed QSPR model (Equation 2) describing the linear relationship between the molecular structure of ionic liquids and the Seebeck coefficient utilized the same two descriptors used in read-across analysis (piPC04^C and GATSi$^{A)}$. Pairwise correlation coefficient between the selected descriptors was negligible (r = 0.04).

$$S_e = 0.41(\pm0.04) - 0.15(\pm0.04)\ piPC04^{C} + 0.26(\pm0.04)\ GATSi^{A} \quad (2)$$

n = 12 , k = 6, F = 31.05 , p = 9.13×10^{-5}, R^2 = 0.87, $RMSE_C$ = 0.11, Q^2_{CV} = 0.82, $RMSE_{CV}$ = 0.13, Q^2_{EXT} = 0.97, $RMSE_P$ = 0.05, CCC = 0.98

The model was well fitted to the training data (R^2 = 0.87) and was characterized by the low value of the root mean square error ($RMSE_C$ = 0.11). Both measures are calculated from

residuals (i.e. differences between the experimental values and the values predicted by the model) for the training set compounds (for equations please refer to the Supporting material, Table S1). Analogical statistics obtained based on the internal validation (Q^2_{CV} and $RMSE_{CV}$) confirmed the good flexibility of QSPR model. Finally, the prognostic ability of the model was confirmed by calculating of Q^2_{Ext}, CCC and $RMSE_P$ from residuals for the validation compounds. Since there were no significant differences between particular mean square errors ($RMSE_C$, $RMSE_{CV}$ and $RMSE_P$) we can conclude that, the model is not overfitted and predicts S_e correctly, not only in the space of training set, but also for other (new) ILs. Furthermore, the high visual correlation between the observed (experimental) and predicted values of the S_e **(Figure 2A)** additionally supports the conclusions from the validation step.

An integral part of the QSPR modeling is to determine and verify the applicability domain (AD) of the developed model. AD is a space, defined by the descriptors' values ($\mathbf{X_i}$) and the response of the model ($\mathbf{y}$), in which the predictions are reliable. In this study, the applicability domain was verified based on the leverage approach with the Williams plot **(Figure 2B)**. The plot visualizes the differences between the predicted and observed S_e values (standardized residuals) versus the similarity of a given compound to the training set (leverage values). Boarders of the AD are determinate by the critical values of standardized residuals (threshold of three standard deviation units, 3σ) and the threshold leverage h* (h* = 3p'/n, where *p'* is the number of model variables plus one, and *n* is the number of compounds in the training set) [24]. Interestingly, in the developed model, all compounds are situated in the range of residuals differing by ±3 standard deviations from zero. Moreover, there were no compounds, which leverage (similarity) value exceed the critical threshold value h* = 0.75. One compound (IL0043) from the training set is located on the border of the h*. The IL0043 is the only one ammonium ionic liquids contained in the training set, therefore it may differ in terms of structure compared to the other, most imidazolium ILs. However, the leverage value

of IL0043 and hence the similarity of a IL0043 to the training set do not exceed the threshold h*, thus this compound broaden the applicability domain of developed model. Therefore, we demonstrated the reliability of predictions for all ILs, for which the leverage value h value is smaller than h*.

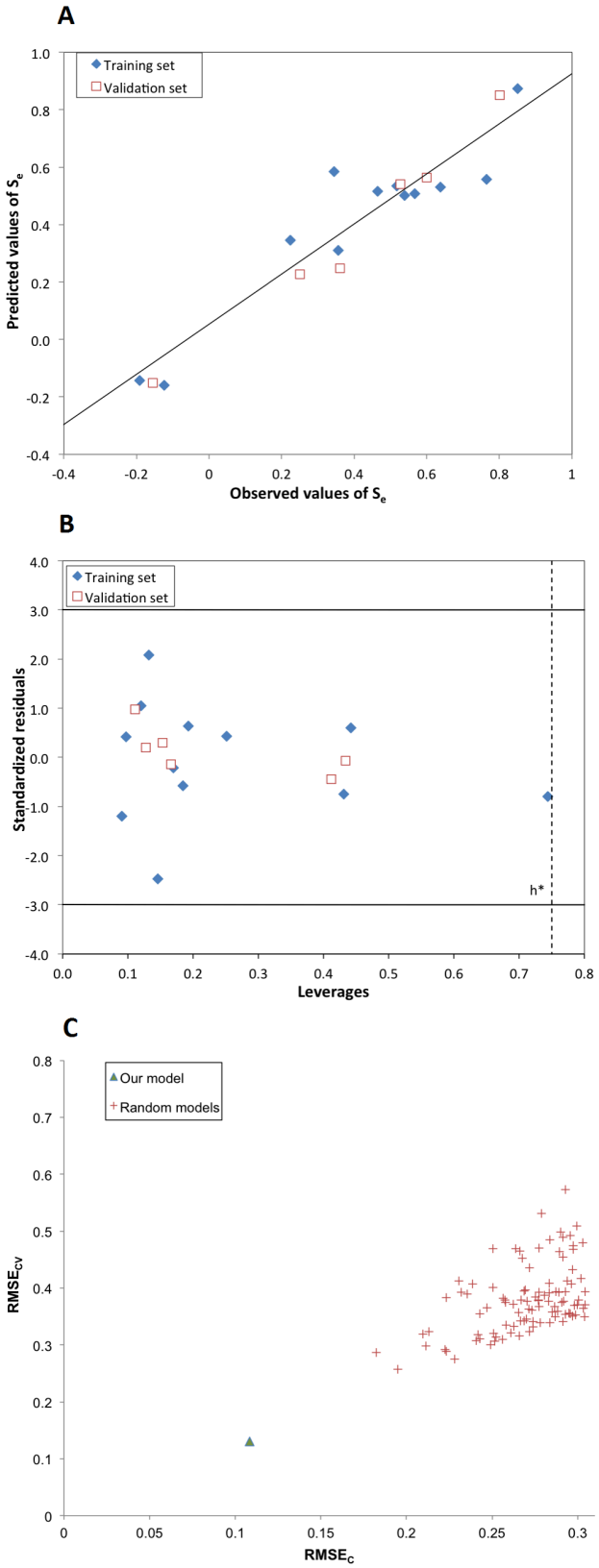


Figure 2. (A) - Observed versus predicted values of S_e.; (B) - Williams plot describing applicability domain of developed model. Dash line indicates threshold value (h* = 0.75), solid

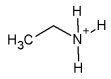
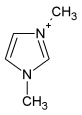
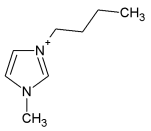
lines represent ±3 SD units; (C) - Y-scrambling results: average values of the square errors of calibration versus cross-validation of the real QSPR model and 100 random models.

In addition, we applied dependent variable scrambling test (Y-scrambling) to reduce the possibility of so-called correlation-by-chance and confirmed the statistical significance of the presented QSPR model.[25] Following the Y-scrambling algorithm we utilized the same two descriptors (*piPC04*^{C}, *GATSi*^{A}) to build 100 random "models" – every time the descriptors were correlated with randomly shuffled values of the S_e for the training compounds. Almost two times higher values of $RMSE_C$ and $RMSE_{CV}$ calculated for the randomly generated models than these of the real QSPR model confirmed the relevance of the QSPR model and the lack of chance correlation **(Figure 2C)**.

According to the OECD quality standards for QSARs it is highly recommended to search for a mechanistic interpretation of the developed model. This can be made by interpreting the results of the descriptors selection. First one, piPC04^{C} (molecular multiple path count of order 4 calculated for cation) [23], is a member of the walk and path family of descriptors. It is defined as the sum of weights of the paths of length 4 in the cation. In this case, 4 edges of the cation's structure were involved in the path calculations.[26] It delivers information about size, symmetry and branching of the cation. We noticed that piPC04^{C} value increases proportionally with increasing size, branching, and length of carbon chains in cations (**Table 1, Table S6**). For instance, piPC04^{C} increases from 0 for ethyloammonium (small and relatively symmetric cation), thought piPC04^{C} = 2.833 for 1,3-dimethylimidazolium (bigger imidazolium cation with two methyl substituents) up to high value of piPC04^{C} = 3.296 for 1-methyl-3-octylimidazolium (unsymmetrical cation with a long carbon side chain). The values of piPC04^{C} are inversely correlated with the Seebeck coefficient, i.e. smaller cations exhibiting low values of piPC04^{C} are generally characterized by the highest Seebeck coefficients.

The origin of the voltage across the thermoelectric cell at ΔT-application is entirely based on the thermally-induced asymmetry of electrical double layers that are formed at electrode surfaces. The first layer at the electrode-surface charges comprises ions adsorbed onto the object due to chemical interactions. Towards the hot electrode the ions become more mobile and overcoming more and more molecular/electrostatic interaction. This leads to an increase of collision frequency on the hot electrode and therefore higher adsorption rate of ions can occur.

Table 1. Values of piPC04^{C} for different structure of the IL' cations

Cation Structure			
	ethylammonium	1,3-dimethylimidazolium	1-buthyl-3-methylimidazolium
piPC04^{C}	0	2.833	3.135

Second descriptor, GATSi^{A} (Geary autocorrelation of lag 1 weighted by ionization potential for anion's structures) comes from 2D-autocorrelation group of descriptors [22]. These descriptors are independent on the original numbering of atom. Moreover, size of the molecule does not affect the calculated length of correlation vector. Autocorrelation descriptors are calculated by summing up certain properties of two atoms, located at a given topological distance.[26] In our model, the vertical electron binding energy of the anion was used as a weighting property. We noticed that anions with higher vertical electron binding energy of the individual atoms in the anion have higher descriptor values (**Table 2, Table S6**). The GATSi^{A} is simply proportional to Seebeck's values, i.e. high values of the descriptor GATSi^{A} increase the values of Seebeck coefficient.

Furthermore, adsorption capabilities of molecules at electrode interfaces are linked to polarization of the molecules and so far to the ionization potential. It is assumed that high vertical electronic stability of anions caused less adsorption rate. Less adsorption of anions at

the electrode is supposed to increase the positive potential at the hot electrode and therefore the Seebeck coefficient.

Table 2. Values of $GATSi^{A}$ for different structure of the IL' anions

Anion Structure	iodide	 acetate	 hexafluorophospate
$GATSi^{A}$	0	1.75	3.50

In the next step, we used the developed QSPR model to predict the unavailable values of Seebeck coefficient for 13 ILs (see Supplementary Materials, **Table S2**). In order to assess wheatear ILs from the prediction set are within the domain of model's applicability we used Insubria graph[27] (**Figure 3**). The graph provides information about the leverage values for the prediction set and also about the relationship between the predicted values for the training set and the prediction set. For this purpose we plotted the leverages, calculated for the prediction set, versus the predicted values of the S_e. Because all compound (except one) from the prediction set were located inside the square defined by the minimum ($Y_t$min) and maximum ($Y_t$max) Seebeck value from training set and by critical h* value, we concluded that the predictions for tested compounds are within the model's applicability domain. There is only one ionic liquid (Trihexyltetradecylphosphonium chlorid), which has the predicted S_e lower than the lowest experimental value in the training set. This ionic liquid consist of cation, which has relatively long alkyl side chains – such ionic liquids were not include in the training set, so in this case, the prediction should be treated with greater care as less reliable. The S_e value for this IL has been extrapolated (not interpolated as for the remaining ILs). One should remember as well that the model is valid only for the system containing ILs with addition of the redox couple (LiI/I_2) in the concentration of 0.01 M.

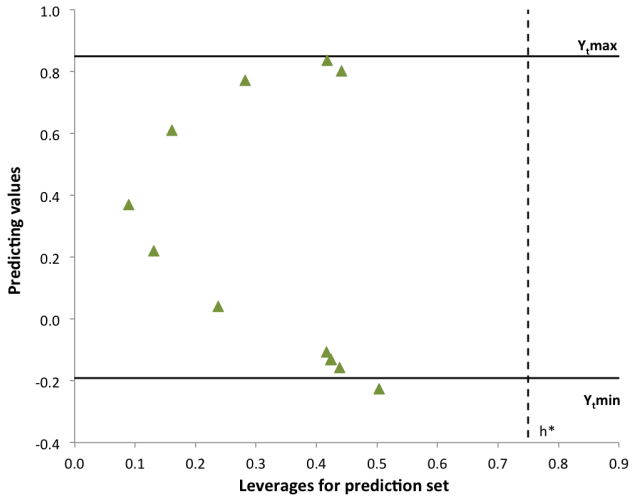


Figure 3. Insubria plot: Leverage values for the prediction set versus predicted values of S_e. Dash line indicates the critical leverage value (h*=0.75), solid lines represent minimum and maximum values of experimental values of S_e for the training set.

2.3. How much the Seebeck coefficient of ILs depends on the concentration of redox couple?

In our work we also intended to extend the predictions onto systems containing ILs and the redox couple (LiI/I_2) in other concentrations than 0.01 M. Therefore, we additionally measured the S_e for each IL in different concentration of the redox couple added to the system. In the previous paragraph, we pointed out, how we attempted to build a QSPR model for Seebeck coefficient prediction, basing only on IL's molecular descriptors. It was possible to achieve, because in every case, both redox couple type as well as its concentration were the same. Therefore, we treated them as constant conditions, assuming the only parameter that could have any affection on the S_e value was the varying structure of IL.

In case of different redox concentration, we applied different approach that can be considered as the complementary to the first one. At this stage, we had a constant structure of IL as well as the constant type of a redox couple. The only factor varying in the system was the

concentration of the redox couple. Therefore, we attempted to describe the relationship between the concentration and the S_e by means of an appropriate mathematic formula.

The idea was to create an additional computational tool (correction equation) that could be added to the previously developed QSPR model, in order to extend its predicting performance onto different concentrations of the redox couple.

We had the measurements performed for three LiI/I_2 concentrations as follows: 0.01M (the one we developed QSPR model for), 0.1M and 0.2M. Since we were interested in the relation between 0.01M concentration with other concentrations, we constructed two scatter plots, in order to analyze their mutual relations (**Figure 4A and 4B**). We noticed that those relations are logarithmic. In order to simplify the calculations, we transformed the data by adding a constant value of 0.2 to every result. Using nonlinear regression method we calculated the proper equation's parameters (Equation 3 and 4).[28]

$$S_e 0.1M\ LiI/I_2 = 0.1269 * e^{1.9249*x} \quad (3)$$

$$S_e 0.2M\ LiI/I_2 = 0.1358 * e^{1.702*x} \quad (4)$$

where x is a value of Seebeck coefficient for 0.01M RC system.

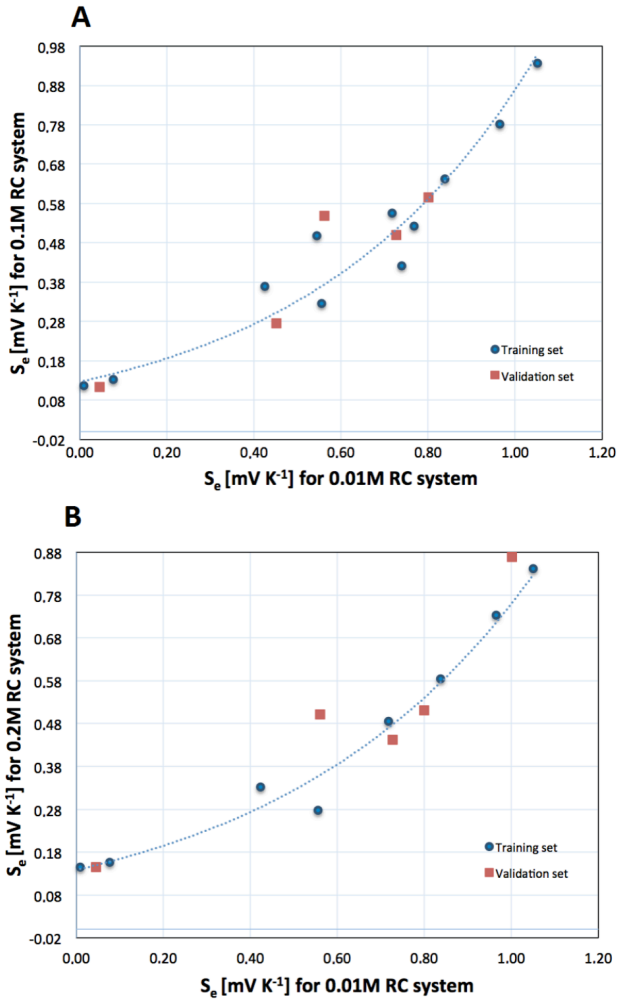


Figure 4. Experimental values of Seebeck coefficient measured for different ionic liquids containing 0.01M redox couple vs. (A) 0.1M and (B) 0.2M redox couple.

For parameters calculations we used only the compounds from the previously developed training set (see Experimental section), therefore at the end we were able to determine their predictive abilities as well using validation set. For the goodness-of-fit we calculated R^2 values and, in order to determine whether our equations are statistically significant, we calculated F statistics for both of them. We received R^2 = 0.928 and F =12.762 for correction equation Equation 3, and R^2 = 0.981 and F = 44.307 for correction equation Equation 4 Those values proved that both equations are correct. The values of Q^2_{EXT} calculated for both equations (Q^2_{EXT}

= 0.898 for the first one and Q^2_{EXT} = 0.873 for the second one) proved good performance of both models and correctness of our approach.

Thereafter, we used the joined QSPR + correction equations approach to verify, weather our holistic idea of predicting the S_e is valid. For this purpose, we first calculated S_e for different ILs with 0.01M redox concentration using the QSPR model, and used those predictions as x values in the correction equations. At the end, we compared the results obtained from the complex QSPR – correction equation calculations with the experimental values (**Figure 5**) and, once again, calculated R^2 and F parameters, in order to validate the method. The results (R^2 = 0.855 and F = 6.470 for the first equation, R^2 = 0.847 and F = 5.985 for the second) proved high performance of our method.

Trends of Seebeck coefficient changes occurring in the IL/redox couple system which we described in this work stay in a good accordance with previous reports available in the literature. Mixtures of ILs and the redox couple used as a thermoelectric energy source were already proven to indicate descending S_e values with increasing redox couple concentration.[9, 12, 29] This effect was observed not only for I^-/I_2 system but also for other redox couples tested.[12] It is believed that this fact is connected to the effect of redox ions solvation by the particles of the solvent. Smaller concentration of the redox couple grants more "independent" particles available to solvate the ions.[9] However, the exact interpretation of this phenomenon requires additional studies aimed at investigating the intermolecular interactions within the system and was beyond the scope of this work.

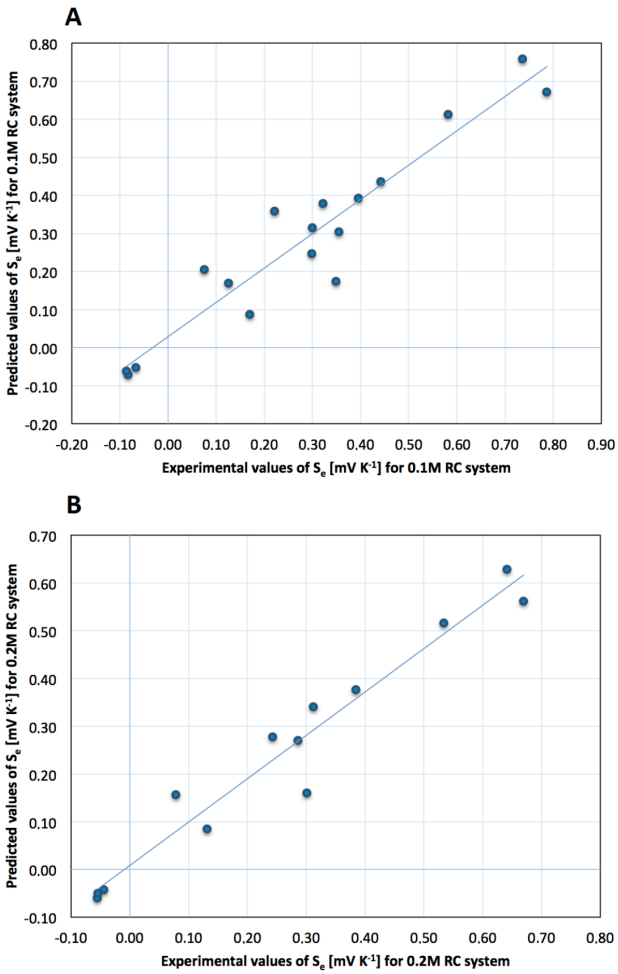


Figure 5. Observed values of S_e versus predicted values of S_e for ionic liquids containing (A) 0.1M redox couple and (B) 0.2M redox couple, obtained by QSPR – correction equation holistic approach.

By developing the correction equations, we gained an opportunity to predict the S_e for ILs with higher concentration of redox couple. The important conclusion coming from this observation is that we were able to confirm previous findings (S_e decreasing with increasing redox couple concentration),[9, 12, 29] support them with a mathematical description, and take into account a broader set of ILs. It means that this trend is a global trend, valid for every ionic liquid.

3. Conclusions

In our contribution, we carried out a qualitative and quantitative analysis of Seebeck coefficient values of the set of different ionic liquids in order to determine their potential application in the liquid-based thermoelectrical device. Based on the structural similarity, we have estimated the Seebeck coefficient values for ionic liquids with 0.01M redox couple, using read-across analysis and QSPR modeling.

What is more, we found that the structural features of particular ions, consisting the IL could describe the S_e. We have noticed that the size, symmetry and branching of cation and vertical electron binding energy of anion have a huge impact on Seebeck coefficient's value. Based on the read-across analysis and developed QSPR model, we concluded that the highest S_e values are observed for ionic liquids consisted of small, not so branching and relatively symmetric cations, and anions with a high vertical electron binding energy.

Finally, on the basis on the simple correction equations, we conclude that the low redox couple concentration combined with the ILs is much more efficient in the energy conversion.

4. Materials and methods

4.1. Materials

All ionic liquids were purchased from IoLiTec (Ionic Liquids Technologies, Germany). A total concentration of impurities was less than 2% and they were used as obtained. The iodine and lithium iodide powder (99.9%) was purchased from Sigma-Aldrich.

4.2. Experimental data

The crucial condition to obtain good results in modeling (QSPR and read-across analysis) is a high-quality of the experimental data.[30] This means that data used in modeling

should be obtained in a systematic way, by following the standardized protocol. Such procedure minimizes the risk of introducing additional sources of variance to the QSPR model.

Measurements of S_e have been performed in a specially designed test cell that is sketched in **Figure 6** and represented in **Figure 7.** The IL container (volume 0.8 ml) was sealed using a PDMS (Polydimethylsiloxane) ring that was squeezed (for tightness and electrical and thermal contacts between the electrodes and the IL) between two metallized sapphire disks and clamped by two aluminium blocs. The active electrode surface was 1.8cm^2. The distance between the electrodes was 4mm. The aluminium parts were connected to a cooling circuit (Thermostat Frigiterm, J.P. Selecta S.A, Barcelona, Spain), and a heating system (thermal resistor, Vitelec, France). On both sides the temperature could be monitored by an closly to the surface inserted thermocouple (Type K, Jumo-Regulations, Metz, France). Sapphire was used due to its good thermal conductivity; the Rhodium-metallized side was in contact with the liquid.

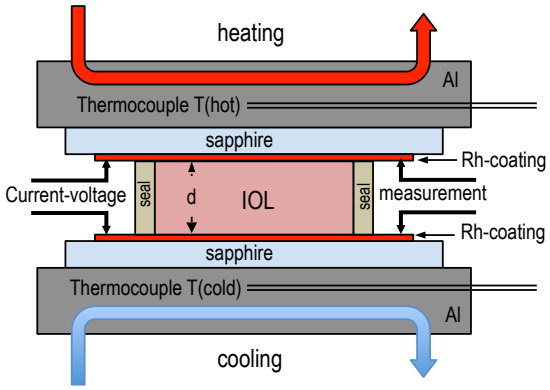


Figure 6. Setup of the test cell, used as the measurement system for characterization of the IOL-based TEGs with a sealed container.

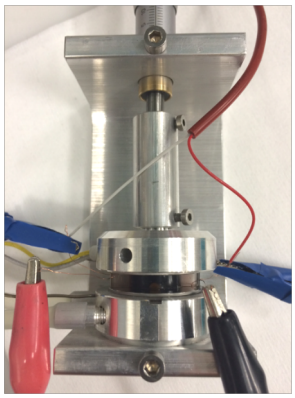


Figure 7. Representation of the TEG characterization setup. The mounting of the two aluminium-bodies allows precise gap adjustment of the chamber

The liquid has been filled into the sealed IL container and measured. The system was fully automatized and allowed measuring the potential difference, current and the heating/cooling dynamics, as well as the stabilization of the measured data. The maximum errors from the measurement system were less than 1 μV and ±1.5 K (Data Acquisition System 34970A and 34901A, Agilent, Santa Clara, CA, USA; thermocouple Type K, Jumo-Regulations, Metz, France). The drop of temperature over the distance between electrode surface and thermocouples causes a addition, however, negligible measurement error. and a concentration inaccuracy of the added redox couple of less than ±10 %. The inaccuracies in temperature measurements and of preparations of IL with redox-couple do not allow to determine an absolute Seebeck-coefficient. However it allows to observe tendencies for the chosen IL as they are characterized under same conditions.

The determination of Seebeck-coefficient is based on a linear dependency between temperature and voltage and has been used subsequently to determine the relative Seebeck-coefficient. The electrode was heated up by a heating-resistor and every 10 seconds the temperature and voltage were registered. The maximal and minimal cell potential as well as the maximal and minimal temperature difference between hot and cold electrode were read out. In our system we observed that during the heating-up and cooling-down the correlation of some

ILs are not linear. After two temperature cycles the linear behavior was established for all ILs. For all the measured Seebeck Coefficients, the difference between heating-up and cooling-down was less than 60μV/K). The published values are taken from the cooling-down curves.

The relative Seebeck coefficient was determined from the slope of the measured voltage-temperature curve as:

$$\alpha = (U_{max} - U_{min}) / (\Delta T_{max} - \Delta T_{min}) \quad (5)$$

4.3. Molecular descriptors

The equilibrium structures of the cations and anions (that the investigated ILs consist of) were obtained by employing the Density Functional Theory (DFT) with Becke's Three Parameter Hybrid Method with the LYP (Lee-Yang-Parr) correlation functional (B3LYP)[31, 32] as well as with the second-order Møller-Plesset (MP2) perturbational method. In both types of calculations we applied the 6-311++G(d,p)[33, 34] Pople's style, one-electron basis set, whose usefulness has been proven in the previous studies of structurally similar ionic liquids.[35]

All calculations were performed with the *Gaussian09 (Rev.A.02)* software package.[36] In order to avoid erroneous results from the default direct SCF calculations, the keyword SCF=NoVarAcc was used and the two-electron integrals were evaluated (without prescreening) to a tolerance of 10^{-20} a.u. The optimizations of the geometries were performed using relatively tight convergence thresholds (i.e., 10^{-5} hartree/bohr (or radian) for the root mean square first derivative).

In the next step, we imported the optimized molecular structures into the DRAGON software[37] for calculating different groups of molecular descriptors. We obtained 2920 descriptors: 1460 of cations' and 1460 of anions' structure, respectively. For selected group of calculated descriptors please refer to Table S3 in the electronic Supplementary Material).

4.4. Data preparation for modeling

Read-across analyses, as well as the QSPR modeling were performed using the values of Seebeck coefficient for ILs with addition of 0.01M LiI/I_2 redox couple. In the first step of modeling, we divided the studied ILs into two separate parts called "training set" and "validation set". The first one (training set) was used (i) to identify groups of structurally similar ILs in read-across analysis and (ii) to develop the QSPR model. Then, the second one (validation set) was used to evaluate the predictive ability of both approaches the read-across estimation and QSPR model.

Data splitting has been performed as follows. Ionic liquids were sorted in the order of the increasing values of S_e. Then, every third compound was moved to the validation set, whereas the remaining compounds formed training set. The second and the penultimate ILs were arbitrarily included in the validation set. In this way, two sets accurately represented the whole range of S_e.[38, 39] Training set contained 12 ILs (67% of all studied ILs), whereas the validation set 6 compounds (33%). A table containing the results of splitting can be found in the electronic Supplementary Materials (Table S2).

In order to select the most optimal combination of the molecular descriptors to be utilized in the read-across and QSPR modeling, we employed the genetic algorithm[17] implemented in the QSARINS software.[18, 19] The algorithm is controlled by the set of steering parameters. In our work we applied the following combination of the parameters: the size of the population: 100 and the mutation rate: 20%. All descriptors used in the study have been auto-scaled by subtracting the mean values and dividing by the standard deviations.

4.5. Read-across

Similarity between the studied ionic liquids was analyzed in the multidimensional space defined by *p* structural features (molecular descriptors) augmented by the experimentally

measured Seebeck coefficient. As such, every ionic liquid is described by p+1 coordinates, where p is the number of molecular descriptors. Ionic liquids located close each other in the p+1 dimensional space are similar, therefore belong to the same class. Such classes may be further used to predict Seebeck coefficient of other, untested ILs. They can be assigned to the same class, based on only its structural similarity to the other class members (the similarity in p dimensions).

T-HCA was performed using Euclidean distance as the similarity measure and Ward's method of linkage. The Euclidean distance between the samples was calculated according to the Equation 6.

$$d_{(ij)}=\sqrt{\sum_{k=1}^{p}(x_{(i)k}-x_{(j)k})^2} \quad (6)$$

where: $x_{(i)k}$, $x_{(j)k}$ are k-coordinate values for i and j object respectively.

Ward's method merges together these clusters, for which the minimum increase in the total within-cluster variance is observed. This increase is a weighted squared distance between cluster centers.[40]

To confirm goodness-of-fit as well as predictive ability of the developed read-across model the accuracy (A) reflecting the total number of compounds correctly classified (eq. 7) and the error rate (E) describing the total number of misclassified compounds (eq. 8) were applied.

$$A=\frac{\sum_g c_g}{n} \times 100\% \quad (7)$$

$$E=\frac{n-\sum_g c_g}{n} \times 100\% \quad (8)$$

where:

g - denotes the number of class,

c_g - is the number of compounds correctly classified to the given class g,

n - is the total number of compounds.

In order to verify, whether the predictions from the qualitative read-across technique differ significantly from the experimentally measured Seebeck coefficient, as well as from the predictions obtained by QSPR modeling, the non-parametric Spearman rank correlation test was applied.[41] The Spearman's rank correlation coefficient ρ_S was calculated as an alternative to Pearson's correlation coefficient according to the Equation 9.

$$\rho_S = 1 - \frac{6\Sigma d_i^2}{N(N^2 - 1)} \quad (9)$$

where: ρ_S – is Spearman's rank correlation coefficient; d^2 - is the square of the difference between ranks; N - is the number of data pairs.

4.6. QSPR modeling

The high-quality of experimental data and optimal combination of the molecular descriptors led us to develop a QSPR model consistent with the recommendations of the Organization for Economic Cooperation and Development (OECD).[42] According to the five golden standards, correctly developed and validated model should be associated with: (i) a defined endpoint, (ii) an unambiguous algorithm, (iii) a defined domain of applicability (iv) appropriate measures of goodness-of-fit, robustness and predictivity and (v) a mechanistic interpretation, if possible[14, 24, 43].

We have applied multiple linear regression technique (MLR) assuming that there is a linear relationship between S_e and the molecular descriptors ($\mathbf{x_1}$, $\mathbf{x_2}$) according to the Equation 10.

$$S_e = b_0 + b_1\mathbf{x_1} + b_2\mathbf{x_2} \quad (10)$$

where: b_1, b_2 are regression coefficient and b_0 is the intercept.

In order to verify the model's goodness-of-fit we calculated the determination coefficient (R^2) and the root mean square error of calibration ($RMSE_C$) (Supplementary Materials, Table S1). Well-fitted model is characterized by R^2 value close to unity and simultaneously, by low $RMSE_C$ value. The evaluation of the robustness of the model was performed during the internal validation. In this step we applied leave-one-out cross-validation method (LOO).[14, 43] As such, the model's flexibility was assessed by using the cross-validation coefficient (Q^2_{CV}) and root mean square error of cross-validation ($RMSE_{CV}$). Finally, we carried out the external validation (the prediction of S_e using the data from the validation set) to determine the predictive ability of the developed model. The external validation coefficient Q^2_{EXT} and root mean square error of prediction $RMSE_P$ were calculated based on the predictions for chemicals from the validation set.[24, 44-47] Q^2_{EXT} value close to unity as well as possibly low $RMSE_P$ value indicates that the model predicts correctly for new, untested chemical. In addition, we have calculated concordance correlation coefficient (CCC) as a complementary, more prudent measure of the external predictivity of the model.

An integral part of QSPR modeling is a proper definition of the group of chemicals, for which the model can be successfully applied (applicability domain, AD). In our case, AD was verified by the use of two approaches called Williams[43] plot and Insubria graph[27]. Finally, after detailed validation, we applied the developed QSPR model to predict the values of Seebeck coefficient for ionic liquids, for which the experimental data have been unavailable.

Acknowledgments

We would like to thank Prof. Paola Gramatica from the University of Insubria for giving access to QSARINS software.

This work was supported by Switzerland through the Swiss Contribution to the enlarged European Union [grant number PSPB-051//2010] and partially by the Polish Ministry of Science and Higher Education [grant number – DS530-8637-D510-14].

References

[1] T. J. Kang, S. L. Fang, M. E. Kozlov, C. S. Haines, N. Li, Y. H. Kim, Y. S. Chen, R. H. Baughman *Adv Funct Mater*. **2012**, 22, 477-489.
[2] B. Burrows *J Electrochem Soc*. **1976**, 123, 154-159.
[3] M. A. Lazar, D. Al-Masri, D. R. MacFarlane, J. M. Pringle *Physical chemistry chemical physics : PCCP*. **2016**, 18, 1404-1410.
[4] P. F. Salazar, S. Kumar, B. A. Cola *J Appl Electrochem*. **2014**, 44, 325-336.
[5] R. Hu, B. A. Cola, N. Haram, J. N. Barisci, S. Lee, S. Stoughton, G. Wallace, C. Too, M. Thomas, A. Gestos, M. E. Cruz, J. P. Ferraris, A. A. Zakhidov, R. H. Baughman *Nano letters*. **2010**, 10, 838-846.
[6] T. O. Poehler, H. E. Katz *Energ Environ Sci*. **2012**, 5, 8110-8115.
[7] M. He, F. Qiu, Z. Q. Lin *Energ Environ Sci*. **2013**, 6, 1352-1361.
[8] E. L. Yee, R. J. Cave, K. L. Guyer, P. D. Tyma, M. J. Weaver *Journal of the American Chemical Society*. **1979**, 101, 1131-1137.
[9] T. J. Abraham, D. R. MacFarlane, J. M. Pringle *Chem Commun*. **2011**, 47, 6260-6262.
[10] T. I. Quickenden, Y. Mua *J Electrochem Soc*. **1995**, 142, 3985-3994.
[11] M. Bonetti, S. Nakamae, M. Roger, P. Guenoun *J Chem Phys*. **2011**, 134.
[12] T. J. Abraham, D. R. MacFarlane, J. M. Pringle *Energ Environ Sci*. **2013**, 6, 2639-2645.
[13] T. Migita, N. Tachikawa, Y. Katayama, T. Miura *Electrochemistry*. **2009**, 77, 639-641.
[14] OECD in Guidance Document on the Validation of (Quantitative) StructureeActivity Relationships [QSAR] Models. Organisation for Economic Co-operation and Development, Vol. (Ed.^Eds.: Editor), City, **2007**.
[15] A. Gajewicz, M. T. D. Cronin, B. Rasulev, J. Leszczynski, T. Puzyn *Nanotechnology*. **2015**, 26.
[16] M. T. D. Cronin, T. W. Schultz *J Mol Struc-Theochem*. **2003**, 622, 39-51.
[17] C. M. Andersen, R. Bro *J Chemometr*. **2010**, 24, 728-737.
[18] P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich *J Comput Chem*. **2013**, 34, 2121-2132.
[19] P. Gramatica, S. Cassani, N. Chirico *J Comput Chem*. **2014**, 35, 1036-1044.
[20] W. N. Vogt, D.; Sator, H., Cluster Analysis in Clinical Chemistry: A Model, Wiley, New York, **1987**.
[21] D. L. K. Massart, L.;, The interpretation of analytical data by use of cluster analysis, Wiley, New York, **1983**.
[22] R. C. Geary *The Incorporated Statistician*. **1954**, 5, 115-145.
[23] H. L. Morgan *J Chem Doc*. **1965**, 5, 107-113.

[24] P. Gramatica *QSAR and Combinatorial Science*. **2007**, 26, 694-701.
[25] S. E. Wold, L.;, Statistical validation of QSAR Results, VCH, Weinheim, **1995**.
[26] R. Todeschini, V. Consonni, Handbook of molecular descriptors, Wiley-VCH, Weinheim ; New York, **2000**.
[27] P. Gramatica, S. Cassani, P. P. Roy, S. Kovarich, Y. C. Wei, E. Papa *Molecular Informatics*. **2012**, 31, 817-835.
[28] D. M. Bates, D. G. Watts, Nonlinear Regression Analysis and Its Applications, John Wiley & Sons, INC., Weinheim.
[29] T. J. Abraham, D. R. MacFarlane, R. H. Baughman, L. Y. Jin, N. Li, J. M. Pringle *Electrochim Acta*. **2013**, 113, 87-93.
[30] J. C. Dearden, M. T. Cronin, K. L. Kaiser *SAR QSAR Environ Res*. **2009**, 20, 241-266.
[31] C. Lee, W. Yang, R. G. Parr *Physical review. B, Condensed matter*. **1988**, 37, 785-789.
[32] A. D. Becke *Physical review. A*. **1988**, 38, 3098-3100.
[33] A. D. Mclean, G. S. Chandler *J Chem Phys*. **1980**, 72, 5639-5648.
[34] R. Krishnan, J. S. Binkley, R. Seeger, J. A. Pople *J Chem Phys*. **1980**, 72, 650-654.
[35] D. A. Wileńska, I.; Freza, S.; Bobrowski, M.; Laux, E.; Uhl, S.; Keppner, H.; Skurski, P. *Molecular Physics*. **2015**, 113, 630-639.
[36] M. J. T. Frisch, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H.P.; Izmaylov, A.F.; Bloino, J.; Zheng, G.; Sonnenberg, J.L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A. Jr.; Peralta, J.E.; Ogliaro, F.; Bearpark, M.; Heyd, J.J.; Brothers, E.; Kudin, K.N.; Staroverov, V.N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.C.; Iyengar, S.S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J.M.; Klene, M.; Knox, J.E.; Cross, J.B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.E.; Yazyev, O.; Austin, A.J.; Cammi, R.; Pomelli, C.; Ochterski, J.W.; Martin, R.L.; Morokuma, K.; Zakrzewski, V.G.; Voth, G.A.; Salvador, P.; Dannenberg, J.J.; Dapprich, S.; Daniels, A.D.; Farkas, Ö.; Foresman, J.B.; Ortiz, J.V.; Cioslowski, J.; Fox, D.J.; in Gaussian 09, revision A.02, Vol. (Ed.^Eds.: Editor), City, **2009**.
[37] Talete in Dragon (Software for Molecular Descriptor Calculation http://www.talete.mi.it/, Vol. (Ed.^Eds.: Editor), City, **2014**.
[38] T. Puzyn, A. Mostrąg-Szlichtyng, A. Gajewicz, M. Skrzyński, P. A. Worth *Structural Chemistry*. **2011**, 22, 795-804.
[39] M. Hewitt, M. T. Cronin, J. C. Madden, P. H. Rowe, C. Johnson, A. Obi, S. J. Enoch *J Chem Inf Model*. **2007**, 47, 1460-1468.
[40] J. H. Ward, Jr. *Journal of the American Satistical Association*. **1963**, 58, 236-244.
[41] C. E. Spearman *British Journal of Psychology*. **1910**, 3, 271-295.
[42] OECD in OECD Principles for the validation, for regulatory purposes, of (Quantittative) Structure Activity Relationship models, Vol. (Ed.^Eds.: Editor), Organisation for Economic Co-Operation and Development, City, **2004**.
[43] A. Tropsha, P. Gramatica, V. K. Gombar *Qsar Comb Sci*. **2003**, 22, 69-77.
[44] N. Chirico, P. Gramatica *J Chem Inf Model*. **2011**, 51, 2320-2335.
[45] R. Bro, K. Kjeldahl, A. K. Smilde, H. A. Kiers *Analytical and bioanalytical chemistry*. **2008**, 390, 1241-1251.
[46] N. Chirico, P. Gramatica *Journal of chemical information and modeling*. **2012**, 52, 2044-2058.
[47] P. Gramatica, E. Giani, E. Papa *J Mol Graph Model*. **2007**, 25, 755-766.