Anal. Bioanal. Electrochem., Vol. 16, No. 8, 2024, 764-785
https://www.doi.org/10.22034/abec.2024.715433

Analytical & Bioanalytical Electrochemistry

Full Paper

Heavy Metals Potentiometric Sensitivity Prediction by Firefly-Support Vector Machine Modeling Method

Eslam Pourbasheer,* Reza Mahmoudzadeh Laki, and Mohammad Sarafraz Khalifehlou

Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, P.O. Box 179, Ardabil, Iran

*Corresponding Author, Tel.: +98-45-33505204
E-Mail: e.pourbasheer@uma.ac.ir

Received: 3 July 2024 / Received in revised form: 17 August 2024 / Accepted: 18 August 2024 / Published online: 31 August 2024

Abstract- The quantitative structure-property relationship (QSPR) method is an efficient and elegant method for estimating the critical parameters of a wide range of compounds. In this work, the QSPR data set included the structures of 45 modified diphenyl phosphoryl acetamide ionophores along with their sensitivity to Cd^{2+}, Cu^{2+}, and Pb^{2+}. The data set was divided into the training set, including 36 compounds, and the test set, including 9 compounds. The stepwise -multiple linear regressions (SW-MLR), firefly multiple linear regressions (FA-MLR), and firefly-support vector machine (FA-SVM) models were produced on the training set with sensitivity of ionophores for Cd^{2+}, Cu^{2+}, and Pb^{2+} for predicting the potentiometric sensitivity of plastic polymer membrane sensors. The FA-SVM model showed good statistical results for all three cations. Internal and external validation was done to ensure the performance of the model. The results showed acceptable accuracy of the proposed method in identifying important descriptors in QSPR. The results of this study and the interpretation of the descriptors entered in the model can help to design new selective ligands.

Keywords- Ion-selective electrode; Heavy metals; QSPR; FireFly; Support vector machine

1. INTRODUCTION

Potentiometric sensors, specifically ion-selective electrodes, are widely used for measuring ion concentrations in aqueous solutions. Unlike many analytical techniques, they are straightforward, portable, cost-effective, and accurate. These sensors also facilitate in-line and on-line measurements in automated systems, making them highly useful in pharmaceutical,

environmental, and food analysis, among other areas [1,2]. The type of sensitive membrane used (glass, polymeric, or polycrystalline) differentiates various ion-selective electrodes, but most modern research focuses on sensors with plasticized polymeric membranes. This focus is due to the ability to customize the analytical performance of these membranes by altering the active substance, ion-exchanger, and solvent-plasticizer [2-5]. Ionophore-based ion-selective sensors are extensively used to measure ionic compositions in diverse samples [6,7]. Their core component is a plasticized polymeric membrane containing an ionophore, a lipophilic ligand that ensures selectivity towards the target analyte [8]. The variety of ionophores available for detecting different inorganic ions is continually expanding to meet practical needs, as different sample types require distinct selectivity patterns [9].

Studying new potential ligands for ion-selective sensors typically involves synthesizing candidate substances, purifying and characterizing them, preparing sensor membranes with various compositions, and conducting potentiometric measurements to evaluate the sensitivity and selectivity of the new membranes. This process is laborious and time-consuming, with no guarantee of success, as some candidates may lack the necessary selectivity. Therefore, an instrument that could initially screen candidate ionophores and predict their suitability based on their chemical structure would be valuable, saving time and resources [10-14].

Quantitative structure-activity-property relationships (QSAR/QSPR) offer such a tool [15, 16]. QSPR uses statistical and modeling methods to predict the physicochemical and biological properties of molecules, describing how these properties vary with molecular descriptors[17-19]. This approach can replace costly and potentially hazardous experimental tests with calculated descriptors, which predict the properties of new compounds. The primary strategy of QSPR is to establish an optimal quantitative relationship to predict the properties of unmeasured compounds. Recent studies have demonstrated the effectiveness of QSPR for predicting sensor properties based on ionophore structure [20-22]. This study aims to extend the application of QSPR in the field of potentiometric sensors of different ionophores towards three heavy metal ions: copper, cadmium, and lead.

Heavy metals are naturally occurring elements with high atomic weights and densities at least five times that of water. Their extensive use in industrial, domestic, agricultural, medical, and technological fields has led to widespread environmental distribution, raising concerns about their potential health and environmental impacts[4, 23-25]. Their toxicity depends on factors such as dose, exposure route, chemical species, and the age, gender, genetics, and nutritional status of exposed individuals [26]. In this work, for the first time, we used the firefly algorithm as a variable selection and support vector machine (SVM) as the modeling method for the prediction of potentiometric sensitivities of Cd^{2+}, Cu^{2+}, and Pb^{2+}.

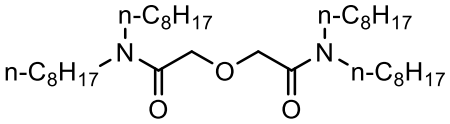
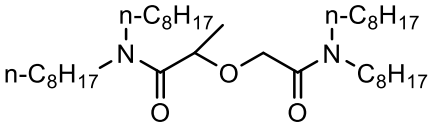
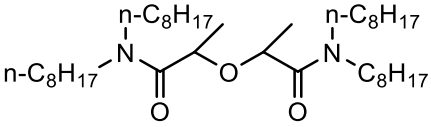
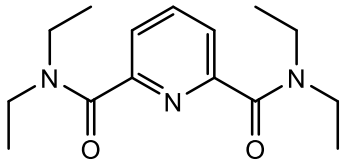
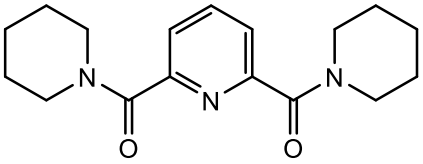
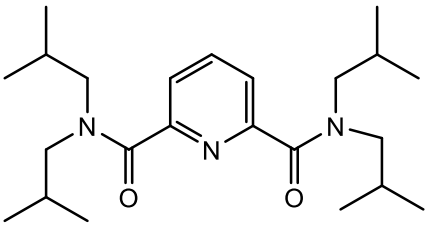
2. EXPERIMENTAL SECTION

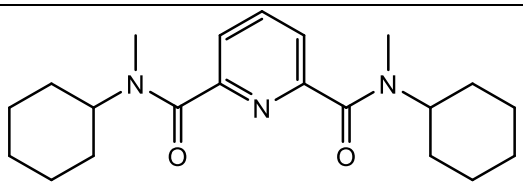
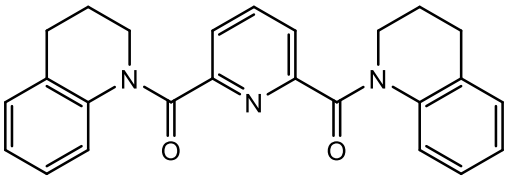
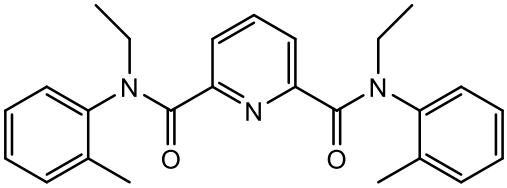
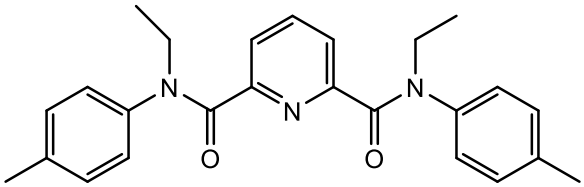
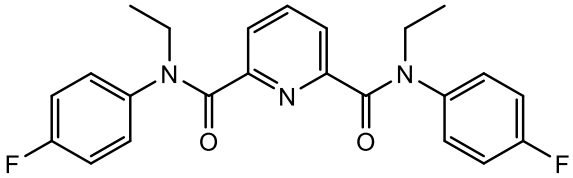
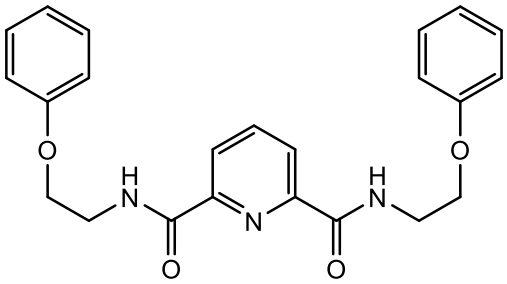
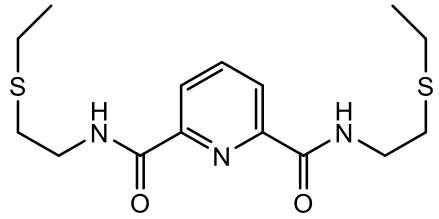
2.1. Dataset and procedure

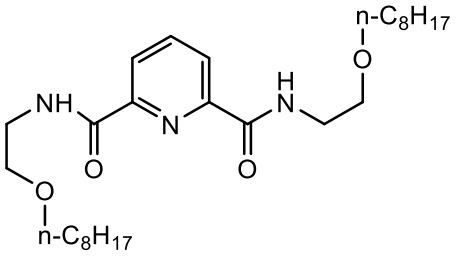
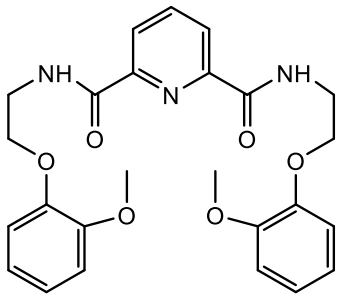
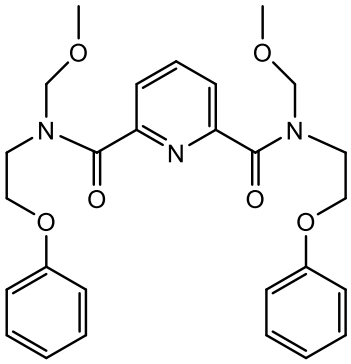
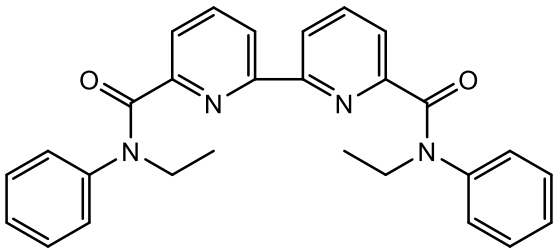
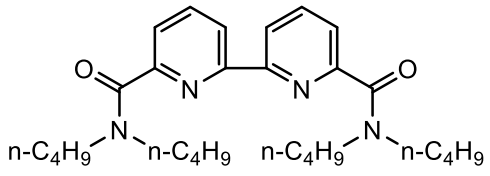
The QSPR data used in this work included diphenyl phosphoryl acetamide derivatives, which were extracted along with their sensitivities from the studies of Vladimirova et al [1]. In

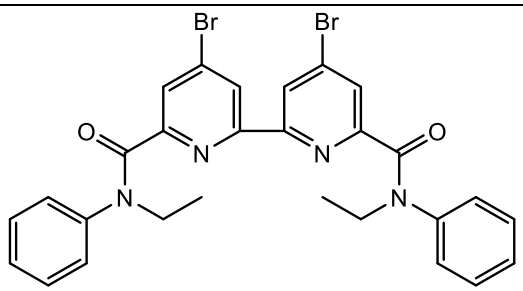
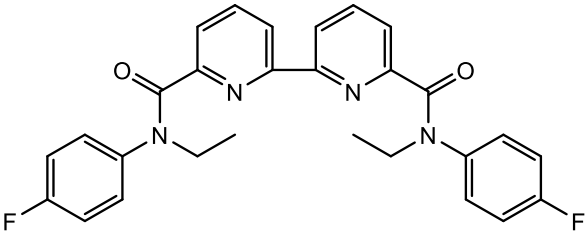
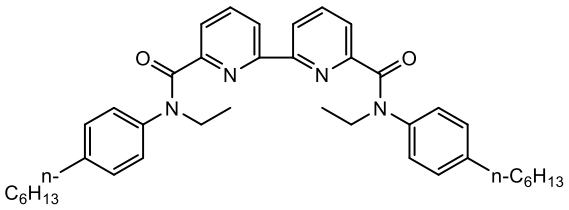
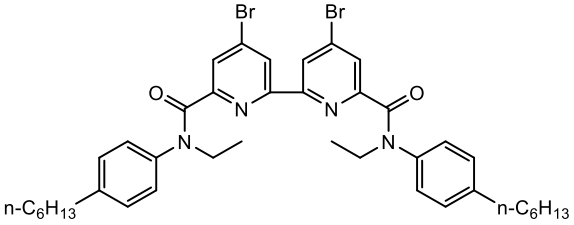
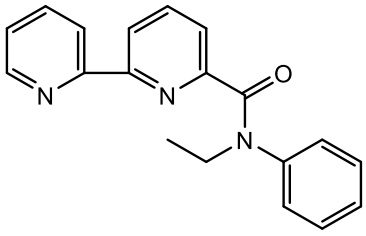
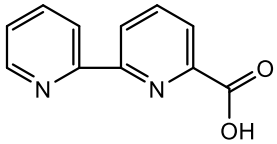
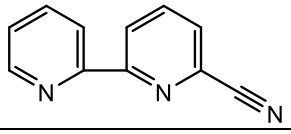
order to use the sensitivity data as the response variables in subsequent QSPR studies, they were first transformed to the logarithmic scale [pS]. All of these molecules were drawn by Hyperchem software and pre-optimized using the MM+ molecular mechanics force field [27]. A more precise optimization was done with the semi-empirical method (AM1) in Hyperchem, and saved with the HIN extension. The data set was divided into a training set and a test set, which contained 36 and 9 chemicals, respectively, to create and assess the validity of a model. The molecular structure and sensitivity values are presented in Table 1. Thereafter, all molecular descriptors were calculated based on the molecular structure of these organic compounds.

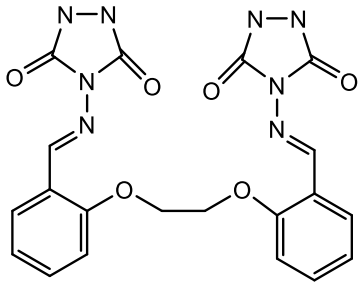
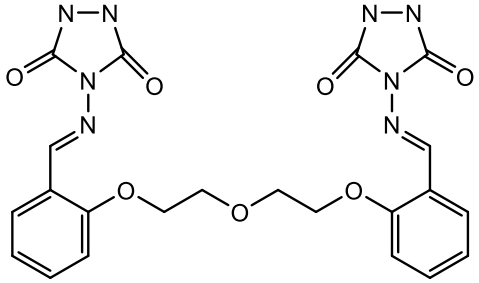
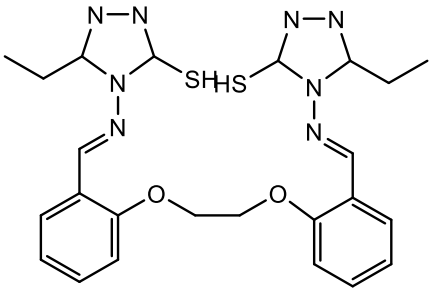
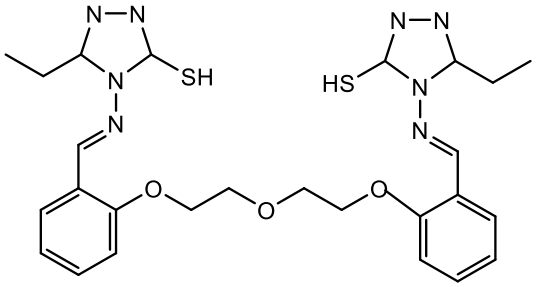
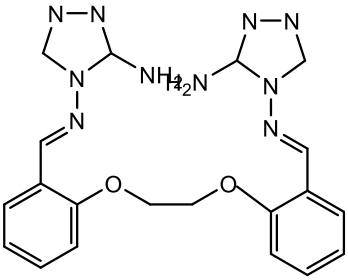
Table 1. Chemical structure and experimental *vs.* predicted sensitivity of ions by FA-SVM method

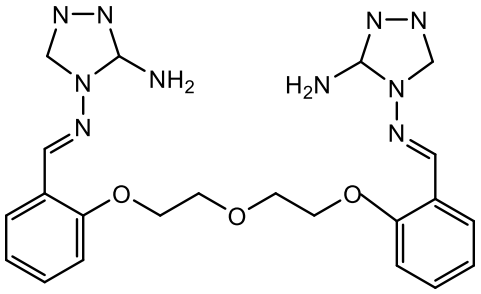
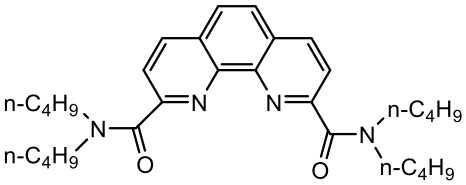
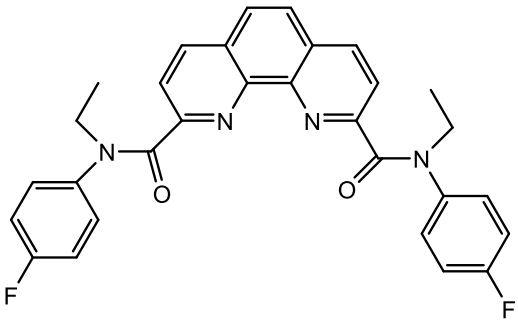
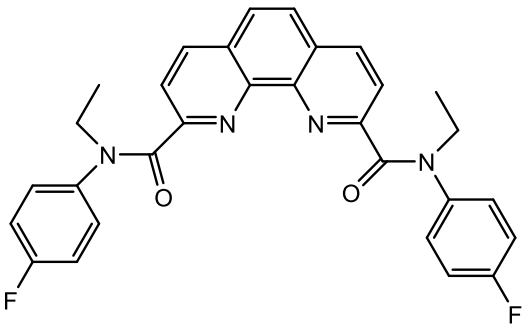
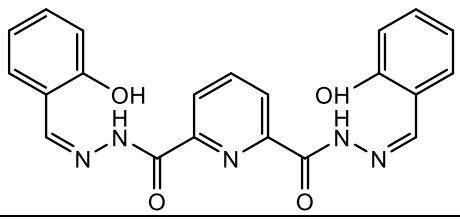
No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
1		9.0	11.2	5.0	8.5	30.0	30.0
2		13.0	13.6	12.0	12.8	24.0	25.6
3 [t]		13.0	15.8	9.0	22.3	18.0	25.9
4		14.0	13.4	25.0	27.9	27.0	27.0
5 [t]		14.0	13.6	23.0	17.0	26.0	27.5
6		27.0	26.4	34.0	34.2	51.0	46.4

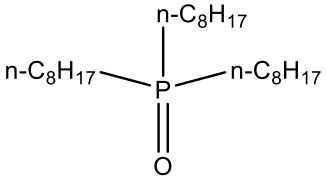
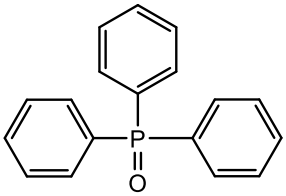
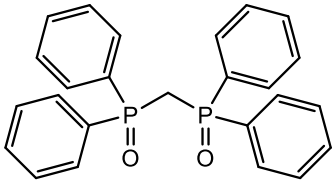
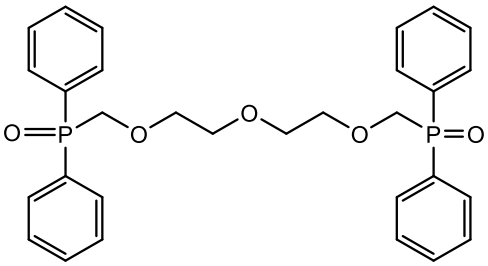
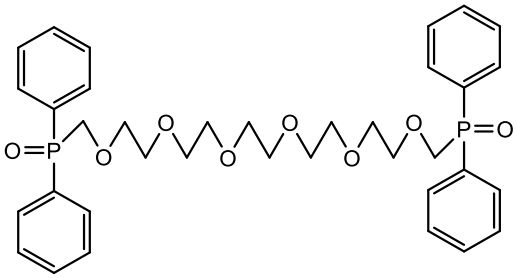
No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
7		21.0	20.3	31.0	30.8	37.0	37.0
8[t]		26.0	21.3	34.0	25.2	34.0	37.9
9[t]		32.0	28.4	43.0	31.2	44.0	43.4
10		32.0	31.4	43.0	36.9	45.0	43.4
11		19.0	19.6	37.0	36.8	37.0	34.9
12		23.0	19.6	28.0	28.2	28.0	28.0
13		22.0	21.4	27.0	26.8	28.0	28.0

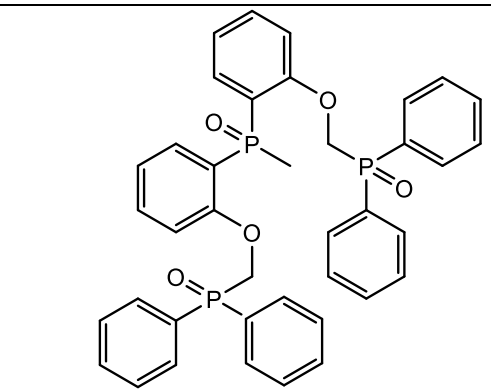
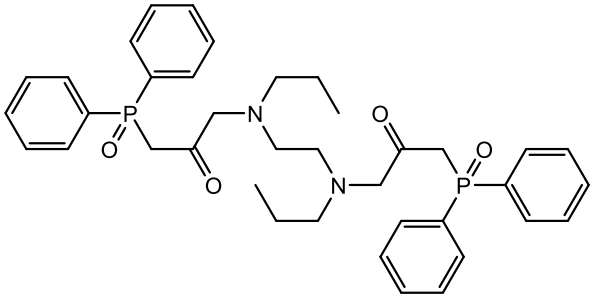
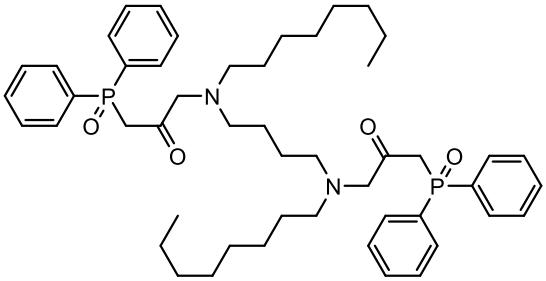
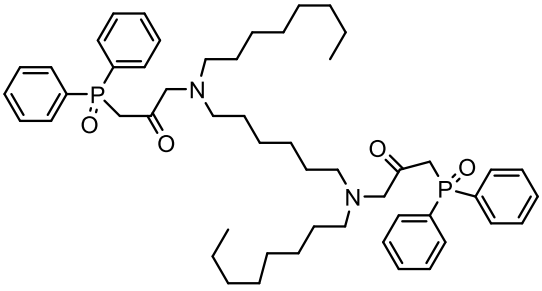
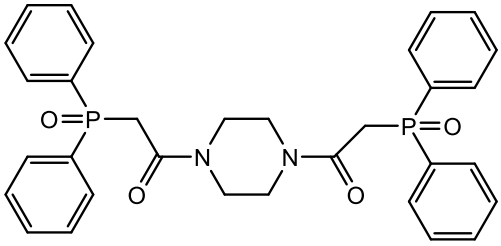
No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
14 [t]		24.0	24.8	31.0	31.2	31.0	26.1
15		25.0	23.4	28.0	27.8	29.0	29.0
16 [t]		22.0	19.6	25.0	29.8	42.0	32.3
17		36.0	29.2	31.0	30.8	24.0	24.0
18 [t]		37.0	32.6	39.0	28.3	26.0	28.3

No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
19		36.0	34.2	26.0	26.6	28.0	28.0
20		31.0	30.4	30.0	29.8	23.0	23.0
21		36.0	35.4	34.0	25.7	23.0	25.1
22		41.0	33.6	31.0	29.0	27.0	31.2
23		3.0	3.0	-12.0	-8.3	0.0	13.5
24		6.0	6.6	1.0	0.8	0.0	13.3
25[t]		5.0	3.8	-10.0	-1.0	3.0	6.3

No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
26		16.0	18.6	15.0	14.8	24.0	24.0
27		15.0	14.4	15.0	15.2	24.0	24.0
28		17.0	18.4	19.0	21.1	25.0	25.0
29		24.0	23.2	34.0	24.4	24.0	24.6
30		18.0	18.6	28.0	27.7	27.0	27.0

No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
31[t]		18.0	14.5	20.0	18.7	28.0	19.5
32		26.0	19.8	24.0	23.8	31.0	31.0
33		27.0	27.9	23.0	24.5	26.0	26.0
34[*]		7	-	0	-	0	-
35		5.0	5.6	0.0	5.1	4.0	15.8

No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
36		18.0	17.4	9.0	14.1	9.0	13.3
37		-10.0	-8.9	5.0	5.2	-20.0	2.1
38		23.0	22.4	30.0	30.2	28.0	28.0
39		25.0	24.4	30.0	29.8	33.0	33.0
40		27.0	25.8	26.0	25.8	40.0	40.0

No.	Structure of the ionophore	Sensitivity, (mV/dec)					
		Cd^{2+}		Cu^{2+}		Pb^{2+}	
		Exp.	Pred.	Exp.	Pred.	Exp.	Pred.
41		24.0	18.4	20.0	20.2	16.0	16.9
42		24.6	24.3	20.3	20.5	32.9	32.9
43		22.1	21.5	18.3	18.5	31.7	31.7
44		24.7	24.5	21.1	20.9	34.6	32.3
45		23.0	22.6	23.4	23.2	34.9	34.9

[t] Test set
[*] Outlier

2.2. Molecular Descriptors

Considering that each molecular descriptor takes into account a small part of the total chemical information in the main molecule, the selection of these descriptors can be mentioned among the steps that have the greatest impact on the QSPR modeling process. For this purpose, a total of 3224 molecular descriptors were calculated by Dragon software for the data set [28]. However, since the high number of descriptors (as independent variables) is one of the important problems in QSPR modeling, by removing descriptors with constant and relatively constant values (more than 90% constant) as well as descriptors with a correlation more than 0.9, their number was reduced. After removing the mentioned descriptors for ionophores along with their sensitivity to Cd^{2+}, Cu^{2+}, and Pb^{2+}, there were 336 descriptors for Cd^{2+}, 341 descriptors for Cu^{2+}, and 355 descriptors for Pb^{2+}.

2.3. Variable selection by firefly

The firefly algorithm (FA) presents a compelling approach for variable selection within the domain of complex data modeling. Drawing inspiration from the bioluminescent communication of fireflies, FA leverages a light intensity-based attraction mechanism. Each firefly represents a candidate solution, with its position in a multidimensional space corresponding to the chosen variables from the dataset [29]. The light intensity of a firefly directly correlates with the "fitness" of its variable selection, as measured by a relevant performance metric associated with the constructed model (e.g., mean squared error in regression). The core principle of FA lies in the attraction of less bright fireflies (representing inferior variable sets) towards brighter ones (indicating superior variable selections). This collaborative exploration guides the entire swarm towards increasingly optimal combinations of variables. By strategically selecting the most informative variables, FA offers a two-pronged benefit. Firstly, it promotes model parsimony, potentially leading to enhanced generalization and mitigating overfitting. Secondly, it refines the model's ability to capture the underlying relationships within the data by focusing on the most informative variables [30]. Furthermore, FA's inherent swarm intelligence fosters an escape from local minima during the search process, ultimately culminating in a more robust solution for variable selection. Using the firefly algorithm, independent variables were selected among the descriptors according to the objective function. In this work, a firefly algorithm code was created in MATLAB and used for variable selection.

2.4. Support vector machines

In the realm of chemometrics, support vector machines have emerged as a powerful tool for tackling both classification and regression tasks [31-33]. A key strength of SVMs lies in their ability to handle non-linearity through the implementation of kernel functions. These

functions implicitly project the data points into a higher-dimensional space where linear relationships become more readily apparent. This empowers SVMs to effectively model the often-complex relationships between molecular structure and chemical properties. Furthermore, the focus on maximizing the margin between the hyperplane and the most influential data points (support vectors) inherently reduces the risk of overfitting during model training [34]. This leads to the development of robust models with superior generalizability, capable of accurately predicting properties or classifying new samples not included in the training dataset. In conclusion, SVMs offer a valuable approach for various chemometric tasks, particularly when dealing with non-linear data and limited samples.

3. RESULTS AND DISCUSSION

3.1. SW-MLR model

The variable selection was done by the stepwise (SW) method for the studied ions. Based on stepwise algorithm, five descriptors were selected for modeling the potentiometric sensitivity of Cd^{2+}, Cu^{2+}, and Pb^{2+} with structural descriptors of ionophores. Using these descriptors, the MLR models were applied to the training set. The obtained results were validated by the test series and the sensitivity of studied ions with selected descriptors was predicted. The linear equations and the statistical results obtained for each ion were also determined as follows.

$$pS\ (Cd^{2+}) = -49.833 + 130.343(MSD) + 11.04(GATS5e) + 24.448(GGI5) - 9.946(Mor27m) - 11.12(B07[O\text{-}O]) \quad \text{(eq.1)}$$

$N_{train}=36$, $R^2_{train}=0.808$, $R^2_{test}=0.176$, $R^2_{adj}=0.776$, $F_{train}=25.233$, $F_{test}=0.335$, $Q^2_{LOO}=0.725$, $Q_{LGO}{}^2=0.706$

$$pS\ (Cu^{2+}) = 39.953 - 53.168(MATS1e) + 3202.448(JGI9) - 11.205(Mor22m) - 9.946(G2m) - 265.534(B07[O\text{-}O]) \quad \text{(eq.2)}$$

$N_{train}=36$, $R^2_{train}=0.809$, $R^2_{test}=0.306$, $R^2_{adj}=0.776$, $F_{train}=25.253$, $F_{test}=0.381$, $Q^2_{LOO}=0.711$, $Q^2_{LGO}=0.680$

$$pS\ (Pb^{2+}) = 23.476 + 51.886(MATS2m) + 2721.684(JGI9) + 2.417(Mor23u) - 105.525(Gu) + 3.423(C\text{-}006) \quad \text{(eq.3)}$$

$N_{train}=36$, $R^2_{train}=0.792$, $R^2_{test}=0.454$, $R^2_{adj}=0.757$, $F_{train}=22.793$, $F_{test}=1.097$, $Q^2_{LOO}=0.670$, $Q^2_{LGO}=0.693$

N is the number of molecules in the training set, Q^2_{LOO} and Q^2_{LGO} represent the cross-validation coefficients for the leave-one-out and leave-group methods, respectively. The Q^2_{LOO} value obtained for the model corresponding to each of the ions (Cd^{2+}= 0.725, Cu^{2+}= 0.711, Pb^{2+}= 0.670) shows high reliability. The squared correlation coefficient (R^2), adjusted

correlation coefficient (R^2_{adj}), and Fisher's F statistic (F) are also calculated. The statistical parameters of the SW-MLR models for the studied ions are listed in Table 2. The lower root mean squared error values (RMSE) and higher R^2 and F values show the predictive ability of the models.

Table 2. Statistical parameters of SW-MLR, FA-MLR and FA-SVM models for the studied ions along with its comparison with other studies

Ion	Method	Training set			Test set			Ref.
		R^2	RMSE	F	R^2	RMSE	F	
Cd^{2+}	SW-MLR	0.808	4.475	25.233	0.176	9.086	0.335	This work
	FA-MLR	0.844	4.027	31.398	0.669	5.932	1.640	
	FA-SVM	0.957	2.463	79.983	0.943	3.029	4.610	
	The study of Vladimirova et al.	0.91	2.850	-	0.81	4.220	-	[1]
Cu^{2+}	SW-MLR	0.809	5.256	25.354	0.306	13.195	0.381	This work
	FA-MLR	0.766	5.649	18.969	0.629	9.422	0.858	
	FA-SVM	0.949	2.917	69.987	0.760	8.523	0.784	
	The study of Vladimirova et al.	0.86	4.290	-	0.66	6.880	-	[1]
Pb^{2+}	SW-MLR	0.792	6.124	22.793	0.454	14.178	1.098	This work
	FA-MLR	0.839	5.312	30.246	0.744	8.357	2.110	
	FA-SVM	0.942	5.514	15.216	0.769	5.646	1.893	
	The study of Vladimirova et al.	0.950	2.600	-	0.640	7.490	-	[1]

3.2. Identification of outliers

A William's plot was used to show the outliers in the data set. This plot for the studied ions is presented in Figure 1 (a-c). The warning leverage (h^*) is a threshold value used in William plots to identify influential compounds. It is defined by the following formula:

$$h^* = 3(p+1)/N$$

where p is the number of predictor variables (descriptors) in the model, and N is the number of compounds (observations) in the training set. This threshold helps to determine the boundary within which the compounds are considered to be within the applicability domain of the model.

Compounds with leverage values greater than h* are considered to be influential points and may need closer examination. In this graph, the X-axis (leverage values) measures the effect of each combination on the model, and the Y-axis (standardized residuals) shows the difference between the experimental and predicted values. Typically, standardized residuals within the range of -3 to +3 are considered acceptable. Horizontal lines at +3 and -3 can be drawn to mark these boundaries. Additionally, compounds with standardized residuals outside the range of -3 to +3 are considered outliers. These compounds are poorly predicted by the model and may indicate potential issues with the model or the data for these compounds. In the Williams diagram, compounds that have an h value greater than h* are considered structural outliers. However, these compounds can be retained in the model if they have a low standardized residual value. Based on Figure 1a, for Cd^{2+}, compound 34 showed a larger standardized residual value and thus this compound was detected as an outlier. Also, in Figure 1b, which shows the William diagram for Cu^{2+}, compounds 3 and 34 were recognized as outliers. In Figure 1c, outlier data for Pb^{2+} can be identified as compounds 25 and 34. Fortunately, all outliers were in the test set and removing these outliers from the test set did not change the MLR equations. Since compound 34 was outlier in all the models, it was removed from the data set. In the next steps, the selected compound (Compound 34) was removed from this series and modeling was done.

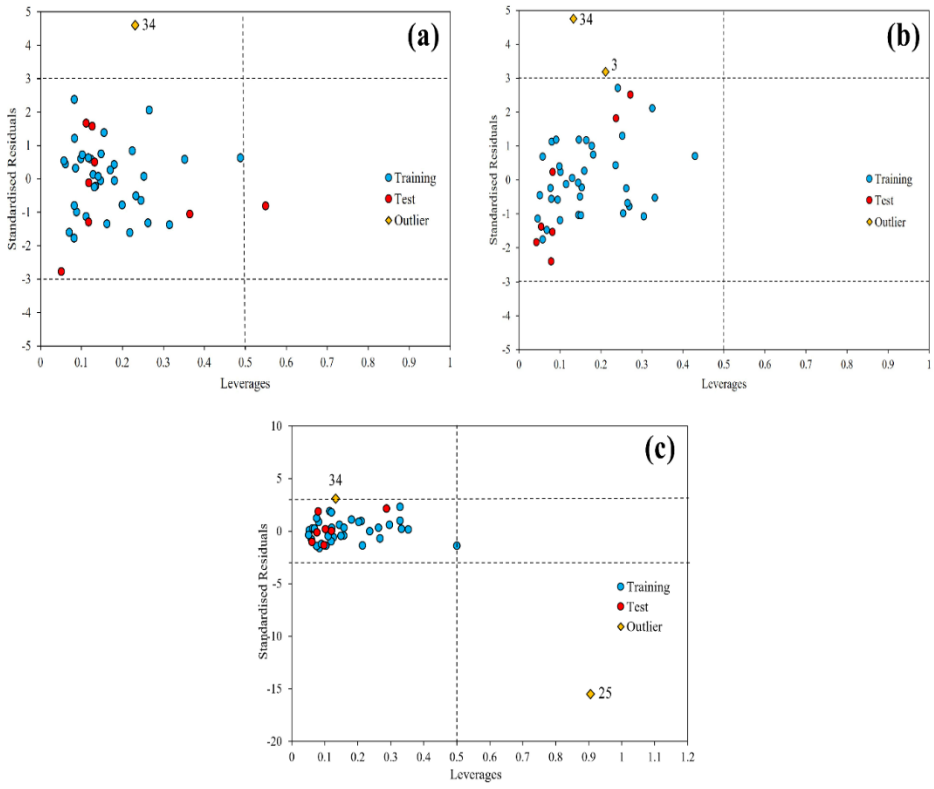


Figure 1. SW-MLR design of Williams model for training and test set a) Cd^{2+}, b) Cu^{2+}, and c) Pb^{2+}

3.3. FA-MLR model

After removing the outlier, the FA-MLR models were created based on the selected descriptors by firefly algorithm (five descriptors) for the studied ions, whose linear equations and statistical results are as follows:

$pS\ (Cd^{2+}) = 106.123 + 1340.625(JGI9) – 268.484(G2m) – 243.672(G2s) – 7.416(Inflammat\text{-}80) –1.303(F09[C\text{-}O])$ (eq.4)

N_{train}=35, R^2_{train}=0.844, R^2_{test}= 0.669, R^2_{adj}= 0.817 F_{train}=31.398, F_{test}=1.640, Q^2_{LOO}=0.784, Q^2_{LGO}=0.763

$pS\ (Cu^{2+}) = 75.491+ 2924.617(JGI9) – 0.792(RDF095v) – 7.973(Mor27u) – 436.463(G2m) –15.406(B07[O\text{-}O])$ (eq.5)

N_{train}=35, R^2_{train}=0.766, R^2_{test}= 0.629, R^2_{adj}= 0.725, F_{train}=18.969, F_{test}=0.858, Q^2_{LOO}=0.674, Q^2_{LGO}=0.678

$pS\ (Pb^{2+}) = – 37.309 + 54.438\ (MATS2m) + 1494.746(JGI4) + 4.459(C\text{-}006) + 3.283(H\text{-}048) + 8.116(B09[C–O])$ (eq.6)

N_{train}=35, R^2_{train}=0.839, R^2_{test}= 0.744, R^2_{adj}= 0.811, F_{train}=30.246, F_{test}=2.110, Q^2_{LOO}=0.759, Q^2_{LGO}=0.724

The statistical results related to the FA-MLR model are reported in Table 2. The graph of the predicted values for the sensitivity of cadmium, copper and lead ions of the compounds in the training and test sets was drawn against the experimental values, respectively, in Figure 2a to c. As can be seen from Table 2, and figure 2, the FA-MLR had relatively better statistical results compared to the SW-MLR models.

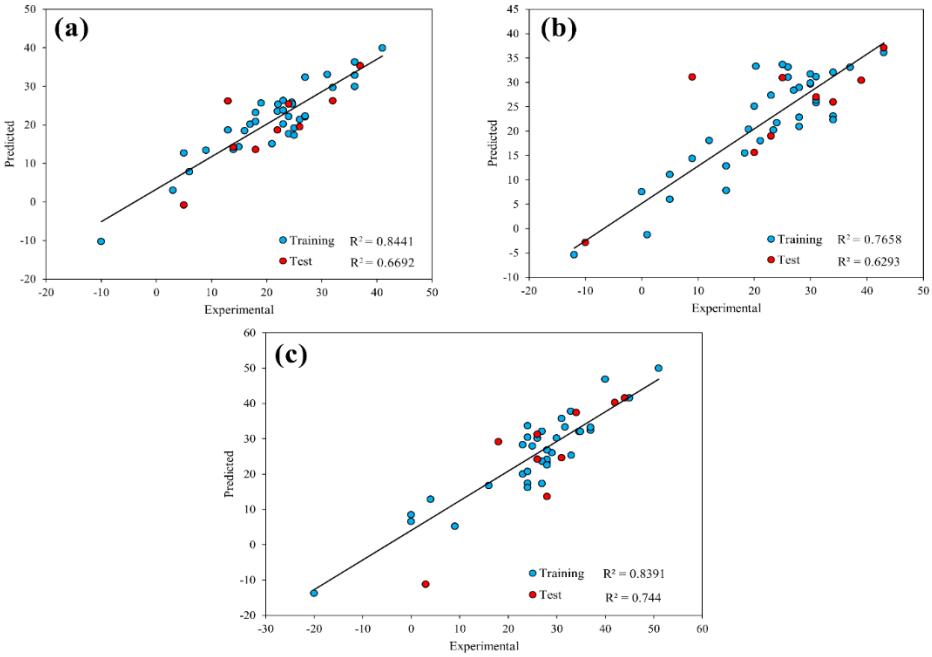


Figure 2. The graph of predicted values versus experimental values of a) Cd^{2+} b) Cu^{2+} c) Pb^{2+} sensitivity based on the FA-MLR model

3.4. Y randomization test

A Y randomization test was conducted with the aim of evaluating the robustness of the constructed models. The process involves training an initial model with the original dataset and evaluating its performance. Then, the target variable (Y) is randomly shuffled while keeping the input features (X) unchanged, and a new model is trained and evaluated. This randomization and evaluation process is repeated multiple times to generate a distribution of performance metrics for the randomized models. By comparing the performance of the original model to this distribution, one can determine if the model's performance is significantly better than random chance. If the original model's performance is much higher than that of the randomized models, it indicates the model has captured a genuine relationship between X and Y; otherwise, it suggests potential overfitting or indicates that the model may be capturing random noise. Lower R^2 and Q^2_{LOO} values in the randomized models indicate that the original model has captured a genuine relationship between the input features and the target variable, rather than fitting random noise. This validation step is crucial to ensure the reliability and robustness of the predictive model, confirming that its performance is due to underlying data patterns and not random chance [35]. Table 3, present the results of Y-randomization tests base on FA-MLR model for Cd^{2+}, Cu^{2+}, and Pb^{2+}. According to the results, it can be seen that the values are mostly less than 0.2 and it can be said that the results of the models are not random.

Table 3. Q^2_{LOO} and R^2_{train} values after several Y-randomization test by FA-MLR models

Iteration	Cd^{2+}		Cu^{2+}		Pb^{2+}	
	Q^2_{LOO}	R^2_{train}	Q^2_{LOO}	R^2_{train}	Q^2_{LOO}	R^2_{train}
1	0.060	0.302	0.016	0.198	0.000	0.163
2	0.057	0.024	0.026	0.087	0.037	0.117
3	0.013	0.113	0.122	0.072	0.077	0.057
4	0.095	0.314	0.008	0.126	0.297	0.506
5	0.222	0.038	0.108	0.010	0.057	0.278
6	0.134	0.047	0.005	0.158	0.001	0.209
7	0.005	0.121	0.035	0.236	0.005	0.140
8	0.001	0.127	0.026	0.246	0.001	0.186
9	0.143	0.062	0.003	0.196	0.016	0.165
10	0.073	0.105	0.002	0.146	0.069	0.088

3.5. Firefly -support vector machine model

The SVM approach, which is a non-linear model, was also used based on the same descriptors selected by the firefly algorithm. Then its performance was compared with the FA-MLR technique, which is a linear model. Our previous works provide information on support vector machines [32,36]. Support vector machine (SVM) regression, aims to find a function that deviates from actual target values by no more than a specified margin of tolerance, known as epsilon (ε), while also maintaining the flattest possible function. Only the data points on the edges of this margin, called support vectors, influence the model, making SVM robust and efficient. The method can handle non-linear relationships through kernel functions like linear, polynomial, and radial basis functions (RBF). Additionally, the regularization parameter (C) controls the trade-off between the model's complexity and the tolerance for errors, allowing for a balance between underfitting and overfitting. RBF is reported in the following formula:

$$\exp(-\gamma * |u - \upsilon|^2)$$

In this formula, u and υ are independent variables, and γ is a kernel parameter that influences SVM performance and training duration. To optimize the gamma parameter, cross-validation of root mean square error (RMSE-cross validation) related to the studied ions was calculated and gamma values from 0.01 to 5 were checked (Figure 3 (a-c)). Next, the graph of γ values against RMSE-cross validation showed that the optimal γ value for cadmium, copper, and lead ions were equal to 0.6, 1.9, and 2.5 respectively.

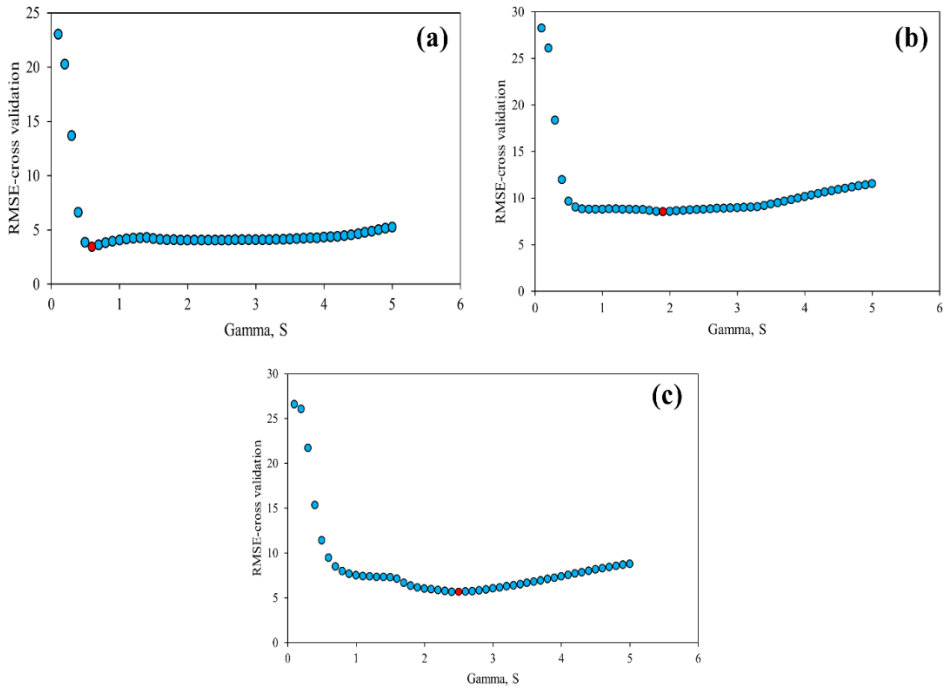


Figure 3. The graph of γ values against RMSE-cross validation for a) Cd^{2+}, b) Cu^{2+} and c) Pb^{2+}

Due to the sensitive parameter ε, the entire training set cannot satisfy the boundary conditions, which allows for dispersion in the solution of the dual formula. The ideal value for this parameter depends on the type of noise in the data. Figure 4 (a-c) shows the ε-insensitive values as a function of RMSE obtained from cross-validation for Cd^{2+}, Cu^{2+}, and Pb^{2+}, respectively. A low ε value corresponds to a lower RMSE, indicating better model performance in terms of fitting the data. According to the results of these graphs, the optimal epsilon values for Cd^{2+}, Cu^{2+} and Pb^{2+}equals to 0.6, 0.2, and 0.01 respectively.

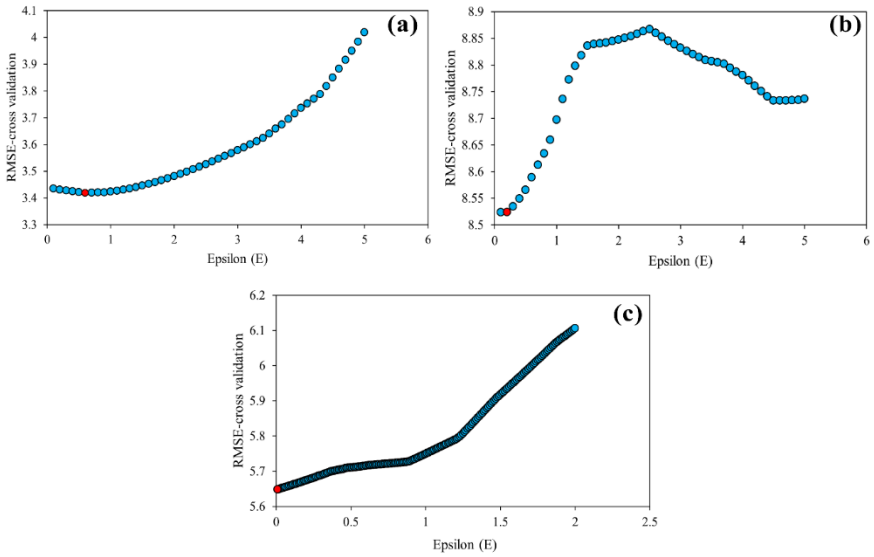


Figure 4. The graph of *ε* values against RMSE-cross-validation for a) Cd^{2+}, b) Cu^{2+} and c) Pb^{2+}

Examining the capacity parameter (C) is the final part of SVM modeling, which controls the trade-off between maximizing margins and minimizing training errors. According to Figure 5(a-c), which shows the optimization graphs of the C parameter related to the studied ions, which were checked from 1 to 300, the optimal values for cadmium, copper, and lead ions were obtained as 12, 100, and 102, respectively.

The predicted sensitivity values for the three studied ions based on the FA-SVM model are presented in Table 1. Figure 6 shows graphs of predicted values versus experimental values for these ions. The statistical results related to the FA-SVM model are listed in Table 2. Also, the obtained results were compared with Vladimir et al.'s [1] study as a source of primary data selection. The comparison results showed that the FA-SVM model with higher R^2 values and lower RMSE had a better performance than the mentioned study. For the model built for Cd^{2+}, both the training set (R^2=0.957, RMSE=2.463, F=79.983) and the test set (R^2=0.943, RMSE=3.029, F=4.610) indicate strong predictive performance. Similarly, the model built for Cu^{2+} shows high predictive ability with training set results (R^2=0.949, RMSE=2.917, F=69.987) and test set results (R^2=0.760, RMSE=8.523, F=0.784). Lastly, the model for Pb^{2+}

also demonstrates excellent predictive capability with training set results (R^2=0.942, RMSE=5.514, F=15.216) and test set results (R^2=0.769, RMSE=5.646, F=1.893).

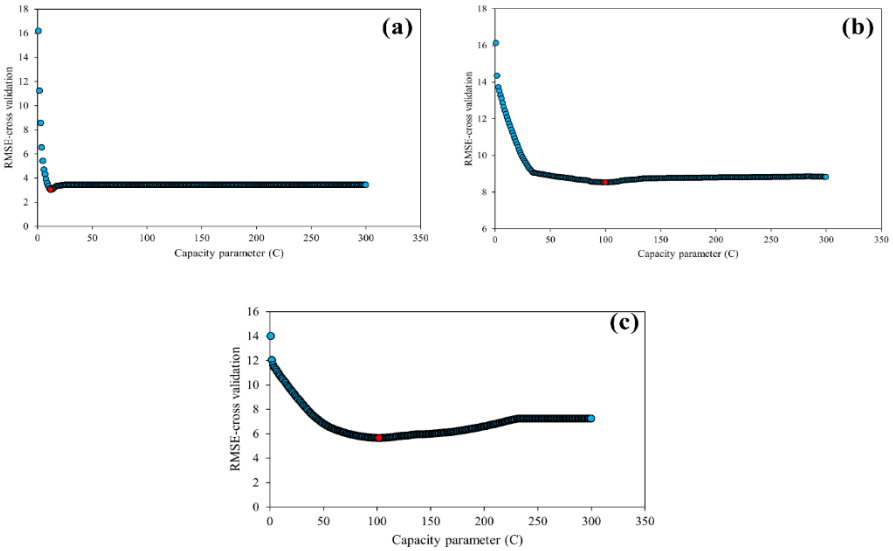


Figure 5. The graph of capacity parameter values against RMSE-cross-validation for a) Cd^{2+}, b) $Cu^{2+,}$ and c) Pb^{2+}

These consistent statistical outcomes across all three ions underline the robustness and accuracy of the FA-SVM model. Compared to the SW-MLR, and FA-MLR models, the FA-SVM models predicted better for all three cations Cd^{2+}, Cu^{2+}, and Pb^{2+} in both training and test sets. Also, the FA-SVM model shows better performance than the SW-MLR and FA-MLR due to lower RMSE and higher F value (Table 2).

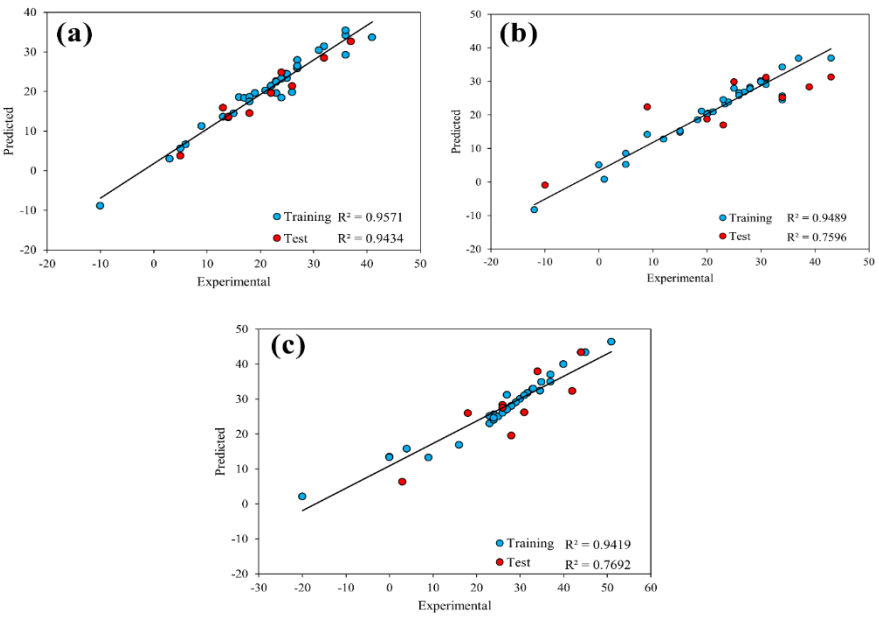


Figure 6. The graph of predicted values versus experimental values of a) Cd^{2+} b) Cu^{2+} c) Pb^{2+} sensitivity based on the FA-SVM models

3.6. Interpreting of molecular descriptors

By interpreting the descriptors in the model, it is possible to obtain information about the factors related to the sensitivity of the studied cations. Molecular descriptors are numerical values that describe various properties of molecules, aiding in the prediction of chemical behavior. The mean square distance index (MSD) captures the overall size and shape of a molecule by averaging the squared distances between atoms. Geary autocorrelation (GATS5e) and moran autocorrelation (MATS2m) descriptors, weighted by electronegativity and mass, respectively reveal how these properties are distributed at specific topological distances within the molecule. GGI5 measures electronic interactions at a five-bond distance, while the B07 [O-O] and B09[C-O] descriptors indicate the presence of oxygen-oxygen and carbon-oxygen pairs separated by seven and nine bonds, respectively. 3D-MoRSE descriptors such as Mor27m and Mor23u analyze molecular structure based on electron diffraction signals, either weighted by atomic mass or unweighted. RDF095v (Radial Distribution Function) describes the distribution of van der Waals volume at a 0.95 Å radius, providing spatial distribution data. The gould index (Gu) measures molecular branching, and C-006 indicates the number of six-membered rings. JGI4 captures the mean topological charge distribution at a four-bond distance, while H-048 specifies the presence of sp3 hybridized oxygen atoms. Lastly, the descriptor inflammat-80 relates specifically to the molecule's potential inflammatory properties and F09[C-O] captures the frequency of carbon-oxygen pairs at a nine-bond distance [37,38]. Together, these descriptors form a comprehensive profile of a molecule's structural and electronic characteristics.

4. CONCLUSION

In this study, a QSPR model based on firefly-support vector machine algorithm was developed to predict the potentiometric sensitivity of Cd^{2+}, Cu^{2+}, and Pb^{2+}. Three different models were created for the studied cations. The SW-MLR and FA-MLR models were obtained as linear models and the FA-SVM as a non-linear model with five different descriptors. The obtained results showed that the FA-SVM model was able to establish a satisfactory relationship between the molecular descriptors and the potentiometric sensitivity of different ionophores to the three mentioned cations. Good results with high statistical quality and low prediction errors were obtained. In comparison, the FA-SVM method predicted both training and test sets more accurately than the FA-MLR method as well as the SW-MLR method. The QSPR model developed in this study can provide a useful tool for predicting the potentiometric sensitivity of various ionophores for Cd^{2+}, Cu^{2+}, and Pb^{2+}.

Acknowledgments

We are honestly thankful regarding the financial support from University of Mohaghegh Ardabili.

Declarations of interest

The authors declare no conflict of interest in this reported work.

REFERENCES

[1] N. Vladimirova, E. Puchkova, D. Dar'in, A. Turanov, V. Babain, and D. Kirsanov, Membranes. 12 (2022) 953.
[2] E. Pourbasheer, A. Rashidi, S. S. Hasani, and M. Rezapour, Anal. Bioanal. Electrochem. 14 (2022) 806.
[3] J. Bobacka, Electroanalysis 18 (2006) 7.
[4] M.R. Ganjali, T. Alizadeh, B. Larijani, M. Aghazadeh, E. Pourbasheer, and P. Norouzi, Curr. Anal. Chem. 13 (2017) 62.
[5] M.R. Ganjali, I. Alahdadi, M. Aghazadeh, and E. Pourbasheer, Int. J. Electrochem. Sci. 10 (2015) 6913.
[6] R. Sharma, M. Geranpayehvaghei, F. Ejeian, A. Razmjou, and M. Asadnia, Talanta 235 (2021) 122815.
[7] M. Jozanović, N. Sakač, M. Karnaš, and M. Medvidović-Kosanović, Crit. Rev. Anal. Chem. 51 (2021) 115.
[8] U. Schaller, E. Bakker, U. E. Spichiger, and E. Pretsch, Anal. Chem. 66 (1994) 391.
[9] P.D. Beer, and P.A. Gale, Angew. Chem. Int. Ed. Engl. 40 (2001) 486.
[10] R.D. Johnson, and L.G. Bachas, Anal. Bioanal. Chem. 376 (2003) 328.
[11] E. Pourbasheer, and R. Aalizadeh, SAR. QSAR. Environ. Res. 27 (2016) 385-407.
[12] A. Beheshti, E. Pourbasheer, M. Nekoei, and S. Vahdani, Arab. J. Chem. 20 (2016) 282-290.
[13] E. Pourbasheer, S. Vahdani, R. Aalizadeh, A. Banaei, and M.R. Ganjali, J. Chem. Sci. 127 (2015) 1243.
[14] E. Pourbasheer, S.S. Tabar, V. H. Masand, R. Aalizadeh, and M.R. Ganjali, SAR. QSAR. Environ. Res. 26 (2015) 461.
[15] R. Mahmoudzadeh Laki, and E. Pourbasheer, ACS Omega 9 (2024) 24707.
[16] E. Pourbasheer, R. Aalizadeh, J. S. Ardabili, and M. R. Ganjali, J. Mol. Liq. 204 (2015) 162.
[17] E. Pourbasheer, Anal. Bioanal. Electrochem. 15 (2023) 150.
[18] E. Pourbasheer, R. Aalizadeh, and M. R. Ganjali, Eurasian Chem. Commun. 5 (2022) 154.
[19] E. Pourbasheer, R. Aalizadeh, J. S. Ardabili, and M. R. Ganjali, J. Mol. Liq. 204 (2015) 162.
[20] A. Legin, V. Babain, D. Kirsanov, and O. Mednova, Sens. Actuators B 131 (2008) 29.
[21] M. Alyapyshev, V. Babain, N. Borisova, I. Eliseev, D. Kirsanov, A. Kostin, A. Legin, M. Reshetova, and Z. Smirnova, Polyhedron 29 (2010) 1998.

[22] M. Alyapyshev, J. Ashina, D. Dar'In, E. Kenf, D. Kirsanov, L. Tkachenko, A. Legin, G. Starova, and V. Babain, RSC Adv. 6 (2016) 68642.
[23] E. Pourbasheer, S. Morsali, Z. Azari, M. A. Karimi, and M. R. Ganjali, Appl. Organomet. Chem. 32 (2018) e4110.
[24] A. Banaei, A. Saadat, M. M. Goli, P. McArdle, E. Pourbasheer, and P. Pargolghasemi, Heteroat. Chem. 27 (2016) 353.
[25] A. Banaei, H. Vojoudi, S. Karimi, S. Bahar, and E. Pourbasheer, RSC Adv. 5 (2015) 83304.
[26] P.B. Tchounwou, C.G. Yedjou, A.K. Patlolla, and D.J. Sutton, Exper. Suppl. 101 (2012) 133.
[27] HyperChem, Molecular modeling system, 7.03rd edn. Hypercube, Gainesville (2002).
[28] R. Todeschini, V. Consonni, A. Mauri, and M. Pavan, DRAGON software for the calculation of molecular descriptors, 5.3rd edn. Talete SRL, Milan (2005).
[29] I. Fister, I. Fister Jr, X.S. Yang, and J. Brest, Swarm. Evol. Comput. 13 (2013) 34.
[30] X.S. Yang, and X. He, Int. J. Swarm. Intell. Res. 1 (2013) 36.
[31] A. Beheshti, E. Pourbasheer, and M.R. Ganjali, J. Mol. Model. 29 (2023) 32.
[32] E. Pourbasheer, R. Aalizadeh, and M. R. Ganjali, Arabian J. Chem. 12 (2019) 2141.
[33] E. Pourbasheer, R. Aalizadeh, M. R. Ganjali, and P. Norouzi, Struct. Chem. 25 (2014) 355.
[34] C. J. Burges, Data Min. Knowl. Discovery. 2 (1998) 121.
[35] C. Rücker, G. Rücker, and M. Meringer, J. Chem. Inf. Model. 47 (2007) 2345.
[36] E. Pourbasheer, R. Aalizadeh, M. R. Ganjali, and P. Norouzi, Struct. Chem. 25 (2014) 355.
[37] S. Gosav, M. Praisler, and D. Dorohoi, J. Mol. Struct. 834 (2007) 188.
[38] V. Consonni, and R. Todeschini, Recent advances in QSAR studies: methods and applications. (2010) 29.