Chemometrics and Intelligent Laboratory Systems 123 (2013) 9–14

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

# Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium

Georgia Melagraki *, Antreas Afantitis *

*Department of Chemoinformatics, Novamechanics Ltd, Nicosia, Cyprus*

ARTICLE INFO

*Article history:*
Received 12 October 2012
Received in revised form 28 January 2013
Accepted 2 February 2013
Available online 16 February 2013

*Keywords:*
corrosion inhibition
QSPR
QSAR
KNIME workflow
Enalos KNIME node

ABSTRACT

In this work we explore the corrosion inhibition of steel in acidic medium for a diverse set of organic compounds by developing a KNIME workflow including the newly introduced Enalos KNIME nodes. We have integrated in a single database 186 corrosion inhibition data of steel in acidic medium including 55 organic inhibitors in different concentrations and investigated the structural characteristics that influence the corrosion inhibition effect. We introduce the custom made Enalos KNIME nodes that are made publicly available by Novamechanics Ltd, as key – nodes to develop robust and validated quantitative structure–property models (QSPRs). Tasks such as assessing the structural characteristics of compounds, validating the model and defining the domain of its applicability are easily addressed using the Enalos family nodes. We have concluded in an accurate kNN model that can reliably predict the corrosion inhibition of a given compound.

## 1. Introduction

Among the corrosion control techniques, the development of new corrosion inhibitors have substantially increased in the recent years because it is believed to be one of the most effective and economic methods to protect metal corrosion in acidic media [1,2]. The design of novel effective corrosion inhibitors is very important in various industrial processes [3]. Various types of organic inhibitors that contain heteroatoms such as oxygen, nitrogen and sulphur and multiple bonds have been reported in literature for several corrosion systems, metals and alloys. The inhibiting effect is generally explained by the formation of physical and/or chemical absorption film on the metal surface [4–7]. The planarity of heterocycles and the presence of lone pair of electrons on the heteroatoms are important requirements that determine the absorption of these molecules on the metallic surface.

Most attempts in designing novel corrosion inhibitors follow a trial and error approach which is time-consuming and costly. These attempts have limited potential in identifying novel molecules with desired characteristics. On the other hand, in silico techniques have continuously gained an important role in the modeling and prediction of properties [8–14]. The design of novel corrosion inhibitors is a new challenge for computational chemistry which models material properties as functions of the molecular structures using the so-called Quantitative Structure Property Relationships (QSPR). Quite recently a series of QSPR studies on corrosion inhibitors were presented in the literature [4–7].

In this work we present a KNIME workflow [15] for modeling and predicting corrosion inhibition for steel in acidic medium. Within this workflow we have used the Enalos KNIME nodes [16] that were developed to facilitate QSPR development. For the development of the proposed kNN model we calculated Mold2 molecular descriptors [17] for the organic inhibitors using Enalos Mold2 KNIME node. The proposed predictive model was fully validated using various validation techniques provided by Enalos Model Acceptability Criteria KNIME node. Moreover we calculated the domain of applicability of the model to identify the area of reliable predictions with Enalos Domain KNIME nodes. The results were interpreted so that design of novel effective corrosive inhibitors could be stimulated. The modeling procedure presented has the potential to considerably decrease the time and efforts required to design or improve corrosion inhibitors.

## 2. Material and methods

### 2.1. Enalos KNIME nodes

KNIME (Konstanz Information Miner) [15] is a user friendly and comprehensive open-source data integration, processing, analysis, and exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models. KNIME is a very powerful tool for data analysis which also integrates all analysis modules of the well known Weka data mining [18]. A great variety of machine learning methods have been applied in Quantitative Structure – Property Studies (QSPR) studies and the best approach for a specific problem needs to be explored. In this work we have used KNIME platform in order to simultaneously run

* Corresponding authors. Tel.: +357 99048039; fax: +357 22347772.
*E-mail addresses:* melagraki@novamechanics.com (G. Melagraki), afantitis@novamechanics.com (A. Afantitis).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.02.003

and compare different modeling methodologies and explore which of the available methods (or combination) best suites our data.

Enalos KNIME Nodes are designed and developed by Novamechanics Ltd with the aim to facilitate model development targeting lead identification and design for all KNIME users. The Nodes are freely available via the KNIME Community and the company's website (http://www.novamechanics.com/knime.php). The Enalos family nodes (Fig. 1) contain:

(i) Enalos Mold2 node for the calculation of Mold2 molecular descriptors (ii) Enalos Model Acceptability Criteria node that can be used to validate the Quality of Fit and Predictive Ability of a continuous QSAR Model, (iii) Enalos Domain – Similarity node that can be used to define Applicability Domain (APD) based on the Euclidean distances, (iv) Enalos Domain – Leverages node that can be used to define Applicability Domain based on the Leverages.

### 2.2. Data set

Data for the corrosion inhibition of steel in acidic medium from different organic compounds were collected from the literature [4–7] and compiled in a single database. Corrosion inhibitors include triazole, oxadiazole and thiadiazole derivatives, aromatic hydrazides and Schiff bases, benzimidazole and 2- substituted derivatives and pyridine derivatives. In total 186 inhibition data were integrated for the total of 55 organic compounds in different concentrations as shown in Table S1 of the Supporting Information. Experimental inhibition efficiency is obtained using weight loss of compounds.

### 2.3. Descriptor calculation

Mold2 software was used to assess the structural characteristics of corrosion inhibitors used in this study. Mold2 calculates a large and diverse set of molecular descriptors encoding two-dimensional chemical structure information [17]. Comparative analysis of Mold2 descriptors with those calculated from commercial software on several published datasets demonstrated that Mold2 descriptors convey sufficient structural information and in addition, better models were generated using Mold2 descriptors than the compared commercial software packages. For each compound 777 descriptors were calculated using Mold2 software which account for the topological, geometric and structural characteristics of compounds. As some of the descriptors do not have any discrimination power (i.e. they have no variation) a filter was applied for their removal [19]. In total 320 descriptors remained to be used as possible inputs during the QSPR model development.

NovaMechanics Ltd through Enalos KNIME nodes made Mold2 available as an extension for KNIME platform. Enalos Mold2 KNIME node can be combined with custom made workflows and real time descriptor calculations combined with state of the art modeling techniques.

### 2.4. Variable selection

Before running the modeling methodology the most significant attributes among the 320 available were preselected by using Correlation – based feature subset selection (CfsSubset) variable selection and BestFirst evaluator [20] which are included in Weka [18]. CfsSubset algorithm evaluated the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that were highly correlated with the class while having low inter-correlation were preferred.

### 2.5. Model development

A great variety of machine learning methods have been applied in QSAR studies [8–14] and the best approach for a specific problem needs to be explored. In this work we have used KNIME platform in order to simultaneously run and compare different modeling methodologies and explore which of the available methods (or combination) best suites our data. k-Nearest neighbors (kNN) methodology outperformed all methodologies tested among which Support Vector Machines (with Sequential Minimal Optimization), Linear Regression and Gaussian Processes for regression [18].

kNN algorithm [21] is a method for classifying objects based on closest training examples in the feature space and belongs to instance-based (or lazy) learning. Based on the kNN algorithm an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (where k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbor. In this work we have used the optimal k value. Euclidean distance was used with all descriptors and contributions of neighbors are weighted by the inverse of distance.

### 2.6. Model validation

The internal performance, as represented by goodness-of-fit and robustness, and the predictivity of a model, as determined by external validation, needs to be evaluated. The produced model was validated using external validation and cross validation methods [21]. The model was internally and externally validated paying special attention to the principles of model validation for accepting QSAR models as described by the Organisation for Economic Cooperation and Development (OECD) [22].

External validation was applied, by randomly splitting the dataset into training and validation set in a proportion of 70:30. The separation of the data set was performed using the Partitioning KNIME node by applying the default random seed. The use of random seed provides reproducible results upon re-execution of the node. The 55 compounds that constituted the test set were not involved by any means in the training procedure. The following statistical criteria were used to assess the robustness, reliability and predictive activity of the model: the coefficient of determination between experimental values and model predictions ($R^2$), validation through an external test set, leave-one-out cross validation procedure and Quality of Fit and Predictive Ability of a continuous QSAR Model according to Tropsha's tests [23–25]. Enalos Model Acceptability Criteria node was used for this purpose.

The first indication on the success of a QSPR model is to measure the quality of fit on the available training data. The most common

**Fig. 1.** Enalos family of KNIME nodes.

objective criteria used for this purpose are the squared correlation coefficient $R^2$ and the root mean squared error (RMSE) statistic which are defined next:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}_{tr}\right)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-k-1)}} \tag{2}$$

According to Tropsha et al. [23] the predictive ability of a QSAR model should be tested on an external set of data that has not been taken into account during the process of developing the model. In particular, the following statistical indices have been proposed [23] to assess the predictive power of QSAR models, besides the popular squared correlation coefficient $R^2{}_{pred}$.:

$$R^2_{ext} = 1 - \frac{\sum_{i=1}^{ntest}\left(y_i - \tilde{y}_i\right)^2}{\sum_{i=1}^{ntest}\left(y_i - \overline{y}_{tr}\right)^2} \tag{3}$$

$$k = \frac{\sum_{i=1}^{ntest} y_i \tilde{y}_i}{\sum_{i=1}^{ntest} \tilde{y}_i^2} \tag{4}$$

$$R^2_o = 1 - \frac{\sum_{i=1}^{ntest}\left(\tilde{y}_i - \tilde{y}_i^{ro}\right)^2}{\sum_{i=1}^{ntest}\left(\tilde{y}_i - \overline{\tilde{y}}\right)^2}, \text{where } \tilde{y}_i^{ro} = ky_i, \quad i = 1, \ldots, ntest \tag{5}$$

In the above equation *ntest* is the number of compounds that constitute the validation data set, $\overline{y}_{tr}$ is the averaged value for the dependent variable for the training set, $y_i, \tilde{y}, \quad i = 1, \ldots, ntest$ are the measured values and the QSAR model predictions of the dependent variable over the available validation set and $\overline{\tilde{y}}$ is the average over all $\tilde{y}, \quad i = 1, \ldots, ntest$.

Tropsha et al. [23] considered a QSAR model predictive, if the following conditions are satisfied:

$$R^2{}_{ext} > 0.5 \tag{6}$$

$$R^2{}_{pred} > 0.6 \tag{7}$$

$$\frac{\left(R^2{}_{pred} - R^2_o\right)}{R^2{}_{pred}} < 0.1 \tag{8}$$

$$0.85 \le k \le 1.15 \tag{9}$$

### 2.7. Domain of applicability

The need to define an applicability domain expresses the fact that QSPRs are models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions. Domain of Applicability [26–29] was defined using both the leverages and similarity measurements. Enalos Domain – Similarity and Enalos Domain – Leverages node were both used to assess domain of applicability of the proposed model.

Firstly similarity measurements were used to define the domain of applicability of the models based on the Euclidean distances among all training compounds and the test compounds [26,27]. The distance of a test compound to its nearest neighbor in the training set was compared to the predefined applicability domain (APD) threshold. The prediction was considered unreliable when the distance was higher than APD. APD was calculated as follows:

$$\mathrm{APD} = <\mathrm{d}> + Z\sigma \tag{10}$$

Calculation of<d>and σ was perfomed as follows: First, the average of Euclidean distances between all pairs of training compounds was calculated. Next, the set of distances that were lower than the average was formulated. <d>and σ were finally calculated as the average and standard deviation of all distances included in this set. Z was an empirical cutoff value and for this work, it was chosen equal to 0.5 [26].

The second approach to define applicability of the domain was the *Extent of Extrapolation* [27,30] It is based on the calculation of the leverage $h_i$ [31] for each chemical, where the QSAR model is used to predict its activity:

$$h_i = x_i^T \left(X^T X\right) x_i \tag{11}$$

In Eq. (11) $x_i$ is the descriptor-row vector of the query compound and $X$ is the $k \times n$ matrix containing the $k$ descriptor values for each one of the $n$ training compounds. A leverage value greater than 3 $k/n$ is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may not reliable.

In order for a QSAR model to be used for screening new compounds, its domain of application [31,32] must be defined and predictions for only those compounds that fall into this domain may be considered reliable.

### 2.8. Y-randomization

Y-randomization test also ensures the robustness and the statistical significance of a QSAR/ QSPR model. The dependent variable vector (Eexp%) is randomly shuffled and a new model is developed using the original independent variable matrix. The derived models after several repetitions are expected to have less significant correlation coefficient values than the ones of the original model. This method is usually performed to eliminate the possibility of chance correlation. If the opposite happens then an acceptable QSAR/ QSPR model cannot be obtained for the specific modeling method and data [23,30].

## 3. Results and discussion

For preprocessing, cleansing, attribute selection, modeling and validation of our data we have created a KNIME workflow suitable to run step by step all the aforementioned tasks simultaneously for each of the described methodologies. Enalos KNIME nodes were used to perform each of the corresponding steps. We have created a KNIME workflow that implements the development of a predictive kNN model following the sequence described as following: Compounds and corrosion inhibition were imported and preprocessed, descriptors were calculated and selected, the kNN algorithm was implemented, the produced model was validated and the domain of applicability was defined.

The original dataset of 186 corrosion inhibitors was randomly partitioned into training and validation set in a ratio of 70:30 consisting of 131 and 55 compounds respectively. The training set was used to develop the QSPR models as described below whereas the test set was not involved by any means in the model development. For each compound

777 descriptors were calculated using Mold2 software which account for the topological, geometric and structural characteristics of compounds. As some of the descriptors do not have any discrimination power (i.e. they have no variation) a filter was applied for their removal. In total 320 descriptors remained to be used as possible inputs during the QSPR model development.

The CfsSubset variable selection with BestFirst evaluator method was then applied on the training data to select the most significant, among the 320 available descriptors. Seven descriptors and the concentration were selected as the most important for the development of the model. The selected descriptors are number of Oxygen (D026), structural information content order-1 index (D282), Geary topological structure autocorrelation length-8 weighted by atomic Sanderson electronegativities (D470), Moran topological structure autocorrelation length-7 weighted by atomic polarizabilities (D509), Lowest eigenvalue from Burden matrix weighted by van der Walls order-6 (D545), Hightest eignevalue from Burden matrix weighted by van der Walls order-4 (D575), number of group Ar-CH = X (D741) and the concentration (C in mM).

The chemical meaning of the molecular descriptors used in the development of each model is briefly discussed below [33,34]. The combination of these descriptors has several advantages such as unique representation of the compound and high discriminating power. Descriptors D026 and D741 are indicators that account for the presence or absence of a specific atom or structural group. More specifically D026 is the number of oxygens that are included in the compound. Descriptor D741 is the number of the Ar-CH = X group which might be present in the compound. Descriptor D282 encodes the structural information content order-1 index. This descriptor belongs to the family of topological information indices of a graph based on neighbor degrees and edge multiplicity. Topological information indices are graph theoretical invariants that view the molecular graph as a source of different probability distributions to which the information theory is applied [33]. Descriptor D470 encodes information as described by Geary topological structure autocorrelation length-8 weighted by atomic Sanderson electronegativities. Geary index is a general index of spatial autocorrelation and is a distance-type function varying from zero to infinite. This index is weighted by atomic Sanderson electronegativities. Atom electronegativity is among the most important atomic properties and is defined as the power of an atom in a molecule to attract electrons to itself. The classical definition of atomic electronegativity is due to Mulliken and is defined as the arithmetic mean of the ionization potential and the electronic affinity of the atom. Sanderson electronegativity is based on covalent radius [33]. Descriptor D509 encodes information related to atomic polarizabilities combined with Moran topological structure autocorrelation length-7. Moran coefficient is a general index of spatial autocorrelation and is related to atomic properties, the number of atoms and the topological distance between specific atoms. Moran coefficient usually takes values in the interval of [−1,+1]. The atomic polarizability is the charge dependent effective atomic polarizability calculated by an empirical method as a linear function of the net atomic charge [33]. Descriptors D545 and D575 are the lowest eigenvalue from Burden matrix weighted by van der Walls order-6 and the hightest eignevalue from Burden matrix weighted by van der Walls order-4. Both descriptors belong to the Burden eigenvalue descriptors introduced by Burden [35]. They are derived from the highest and the lowest eigenvalues of the modified adjacency matrix for the molecules [36]. Burden eigenvalue descriptors weighted by different properties (e.g. van der Walls) have been shown to be very discriminating descriptors. The highest and the lowest eigenvalues obtained from the matrices, have been demonstrated to reflect relevant aspects of molecular structure, and are therefore useful for similarity searching [33].

**Table 1**
Statistical parameters of the QSPR model.

| | |
|---|---|
| $R^2$*training* (*n* = 55) | 0.96 |
| *RMSEtraining* | 4.90 |
| $R^2_{LOO}$ | 0.73 |
| $R^2$*pred* (*n* = 131) | 0.84 |
| *RMSEpred* | 9.83 |

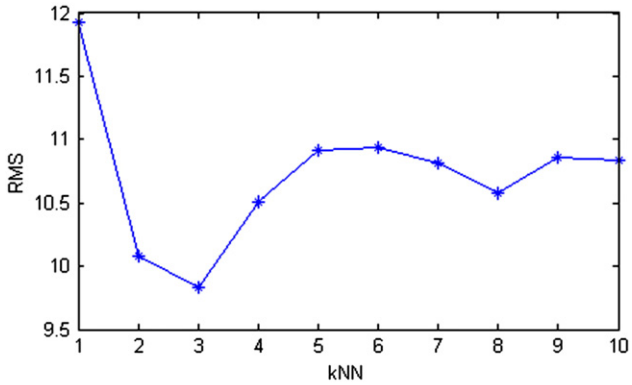


**Fig. 2.** Influence of k value (kNN) to RMSE of the Test Set.

The aforementioned descriptors have different weights that influence the increase or decrease of corrosion inhibition among different compounds. Based on the previous discussion and the positive or negative influence of each descriptor, new derivatives with desired properties can be designed. We have used a KNIME workflow in order to compare different methodologies and explore which of the available methods best suites our data. As described above kNN methodology was selected. By applying on our training data, kNN methodology with an optimized value of k equal to 3 was selected. The influence of k value to RMSE is shown in Fig. 2. Euclidean distance was used with all eight descriptors and contributions of neighbors weighted by the inverse of distance.

Validation of the model was performed using the techniques mentioned in the previous section. The statistics are presented in Table 1, illustrating the accuracy, significance and robustness of the produced model. For comparison reasons in Table 2 we also present the results for three more methodologies tested, namely Support Vector Machines (with Sequential Minimal Optimization), Linear Regression and Gaussian Processes for regression. As can be seen from Tables 1 and 2, kNN methodology results in a robust and accurate model that could be reliably used to predict corrosion inhibition efficiency. We can conclude that the selected descriptors selected by CfsSubset and BestFirst algorithm can encode the structural features of the compounds related to corrosion inhibition.

Fig. 3 presents a plot of experimental versus predicted values of corrosion inhibition ($E_{exp}$% vs $E_{pred}$%) for compounds in the training and test set. The possibility of having included outliers in our dataset was investigated by calculating the standard residuals. Standardized residuals greater than 2 or less than −2 are considered large and are possible outliers. We have indicated outliers with red color (Supporting Information Table S1).

**Table 2**
Comparison of the different modeling methodologies.

| Methodology | *RMSEpred* | $R^2$*pred* |
|---|---|---|
| kNN | 9.83 | 0.84 |
| SVM(SMO) | 13.07 | 0.70 |
| Linear Regression | 16.12 | 0.63 |
| Gaussian Processes | 15.50 | 0.58 |

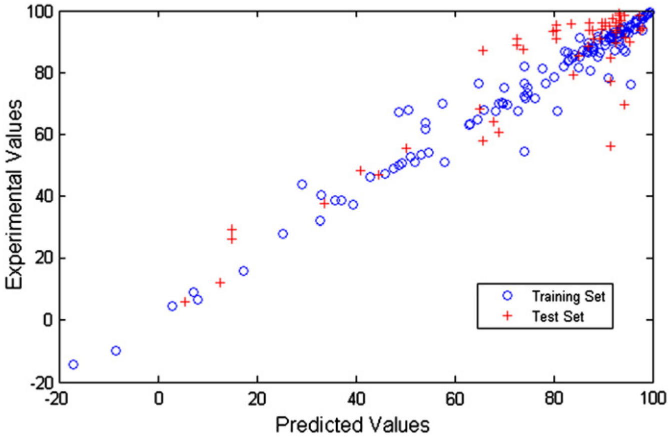


**Fig. 3.** Experimental vs Predicted values for the Training and Test Set.

The Enalos Model Acceptability Criteria KNIME node has been applied to the data (Fig. 4). The model passed Tropsha's [23,31] recommended tests for predictive ability (Eqs. (6)–(9)):

$$R^2_{pred} = 0.84 > 0.5$$

$$R^2_{ext} = 0.83 > 0.6$$

$$\frac{\left(R^2_{pred} - R^2_o\right)}{R^2_{pred}} = -0.001 < 0.1$$

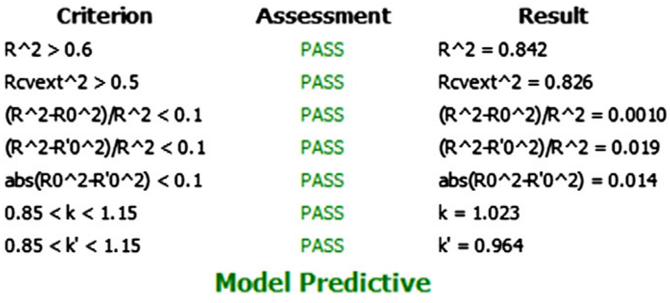
$$k = 1.02 \approx 1$$

In Fig. 4 $R^2$ is the coefficient of determination between experimental values and model prediction on the test set ($R^2_{\text{pred}}$). Mathematical calculations of $R^2_o$, $R'^2_o$ , k, and k'are based on regression of the observed activities against the predicted activities and vice versa using the equations described in the materials methods section (Eqs. (1), (3)–(5)).

The model was also quite stable to the inclusion–exclusion of compounds measured by the LOO (leave one out) cross validation procedure. This is indicated by the following statistic: $R^2_{\text{LOO}} = 0.73$.

The proposed method passed also Y-randomization test which is a method also for testing the robustness and the statistical significance of a QSAR / QSPR model. This method was performed to eliminate the possibility of chance correlation. In particular, 10 random shuffles of the Y vector ($E_{\text{exp}}$%) correlation coefficient values in the ranges of 0.04 to 0.355. The low values of the correlation coefficient indicate that the results from the proposed model were not due to chance correlation.

An additional validation test has been carried out in order to further assess the predictive potential of the applied approach independently of the data set partitioning. The available data were randomly divided 100 times in a ratio of 70:30 for training and test set, respectively. All the random splits results passed Tropsha's recommended tests for modeling validation [23]. The distribution of the RMSE values is presented in Fig. 5. The detailed results are presented in Table S2 of the Supporting Information.



| Criterion | Assessment | Result |
|---|---|---|
| R^2 > 0.6 | PASS | R^2 = 0.842 |
| Rcvext^2 > 0.5 | PASS | Rcvext^2 = 0.826 |
| (R^2-R0^2)/R^2 < 0.1 | PASS | (R^2-R0^2)/R^2 = 0.0010 |
| (R^2-R'0^2)/R^2 < 0.1 | PASS | (R^2-R'0^2)/R^2 = 0.019 |
| abs(R0^2-R'0^2) < 0.1 | PASS | abs(R0^2-R'0^2) = 0.014 |
| 0.85 < k < 1.15 | PASS | k = 1.023 |
| 0.85 < k' < 1.15 | PASS | k' = 0.964 |

Model Predictive

**Fig. 4.** Enalos Model Acceptability Criteria KNIME node screenshot.

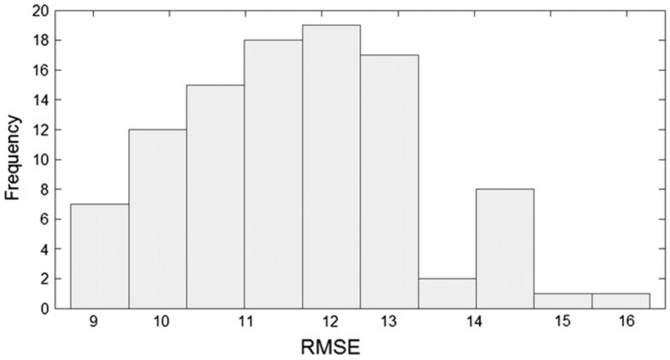


**Fig. 5.** Distribution of the RMSE values (100 Random Splits).

The applicability domain was defined for all compounds that constituted the training set as described in the Materials and Methods section. The applicability domain limit value was equal to 3.774 and 0.183 for similarity [37] and leverage measurements [38] respectively. In the case of similarity measurements all compounds in the test set had values in the range of 0.015-1.23. In the case of leverages the predicted response of one compound (with leverage 0.378) is the result of a substantial extrapolation of the model and the prediction may not be reliable (Table S1). This compound is a simple pyridine and lies outside the domain of applicability due to the fact that the majority of the compounds in the training set are much more complex. Since all validation compounds fell inside the domain of applicability, with only one exception for the domain calculated for leverages, all the other model predictions for the external test set can be considered reliable (Table S1).

The proposed method, due to the high predictive ability and the fact that it requires information related only to the 2D structure of a compound, could be a useful aid to the costly and time consuming experiments for determining the corrosion inhibition. The method can also be used to screen existing databases or virtual chemical structures to identify organic compounds with desired properties. In this case, the applicability domain will serve as a valuable tool to filter out “dissimilar” chemical structures.

## 4. Conclusions

In this paper we present a KNIME workflow that successfully builds an accurate model for the prediction of corrosion inhibition of steel in acidic medium based on a large dataset of 186 organic compounds in different concentrations. Enalos KNIME nodes were included in the workflow to facilitate descriptor calculation, model validation and domain of applicability determination. The molecular descriptors used encode information about the structure, branching, electronic effects, chains and rings of the modules and thus implicitly account for cooperative effects between functional groups. The proposed kNN model was fully validated and was proven accurate and reliable model for the prediction of corrosion inhibition for steel in acidic medium. Applicability domain was defined to identify the reliable predictions. The developed model can accurately predict corrosion inhibition and help the design of novel molecules with desired characteristics.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2013.02.003.

## Acknowledgement

This work was funded under «EUREKA Cyprus Programme» of the Research Promotion Foundation's (RPF) «Framework Programme for Research, Technological Development and Innovation 2009-2010»

(DESMI 2009–2010), Grant Number: EUREKA/EUYΦI/0609/01. DESMI 2009–2010 is co-funded by the Republic of Cyprus and the European Regional Development Fund of the EU.

## References

[1] E.E. Ebenso, M.M. Kabanda, L.C. Murulana, A.K. Singh, S.K. Shukla, Electrochemical and quantum chemical investigation of some azine and thiazine dyes as potential corrosion inhibitors for mild steel in hydrochloric acid solution, Industrial and Engineering Chemistry Research 51 (2012) 12940–12958.

[2] M.M. Kabanda, L.C. Murulana, E.E. Ebenso, Theoretical studies on phenazine and related compounds as corrosion inhibitors for mild steel in sulphuric acid medium, International Journal of Electrochemical Science 7 (2012) 7179–7205.

[3] N.O. Eddy, B.I. Ita, Theoretical and experimental studies on the inhibition potentials of aromatic oxaldehydes for the corrosion of mild steel in 0.1 M HCl, Journal of Molecular Modeling 17 (2011) 633–647.

[4] E.S.H. El Ashry, A. El Nemr, S.A. Esawy, S. Ragab, Corrosion inhibitors, Part II: Quantum chemical studies on the corrosion inhibitions of steel in acidic medium by some triazole, oxadiazole and thiadiazole derivatives, Electrochimica Acta 51 (2006) 3957–3968.

[5] E.S.H. El Ashry, A. El Nemr, S.A. Esawy, S. Ragab, Corrosion inhibitors part 31: Quantum chemical studies on the efficiencies of some aromatic hydrazides and Schiff bases as corrosion inhibitors of steel in acidic medium, Arkivoc 11 (2006) 205–220.

[6] E.S.H. El Ashry, A. El Nemr, S.A. Esawy, S. Ragab, Corrosion inhibitors. Part V: QSAR of benzimidiazole and 2- substituted derivatives as corrosion inhibitors by using the quantum chemical parameters, Progress in Organic Coatings 61 (2008) 11–20.

[7] E.S.H. El Ashry, A. El Nemr, S. Ragab, Quantitative structure activity relationships of some pyridine derivatives as corrosion inhibitors of steel in acidic medium, Journal of Molecular Modeling 18 (2012) 1173–1188.

[8] A. Lee, A.G. Mercader, P.R. Duchowicz, E.A. Castro, A.B. Pomilio, QSAR study of the DPPH radical scavenging activity of di(hetero)arylamines derivatives of benzo[b] thiophenes, halophenols and caffeic acid analogues, Chemometrics and Intelligent Laboratory Systems 116 (2012) 33–40.

[9] X. Xu, X.-G. Li, S.-W. Sun, A QSAR study on the biodegradation activity of PAHs in aged contaminated sediments, Chemometrics and Intelligent Laboratory Systems 114 (2012) 50–55.

[10] A.A. Toropov, A.P. Toropova, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: Predictions of rate constants of hydroxyl radical reaction using representation of the molecular structure obtained by combination of SMILES and Graph approaches, Chemometrics and Intelligent Laboratory Systems 112 (2012) 65–70.

[11] A.A. Toropov, A.P. Toropova, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSAR modeling of toxicity of organic chemicals towards Daphnia magna, Chemometrics and Intelligent Laboratory Systems 110 (2012) 177–181.

[12] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring rm 2 metrics for validation of QSPR models, Chemometrics and Intelligent Laboratory Systems 107 (2011) 194–205.

[13] J. Xu, H. Liang, B. Chen, W. Xu, X. Shen, H. Liu, Linear and nonlinear QSPR models to predict refractive indices of polymers from cyclic dimer structures, Chemometrics and Intelligent Laboratory Systems 92 (2008) 152–156.

[14] E.B. de Melo, A new quantitative structure–property relationship model to predict bioconcentration factors of polychlorinated biphenyls (PCBs) in fishes using E-state index and topological descriptors, Ecotoxicology and Environmental Safety 75 (2012) 213–222.

[15] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kotter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, in: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.), KNIME: The Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization, GfKl: Springer, 2007, pp. 319–326.

[16] http://www.novamechanics.com/knime.php, (accessed 12 Oct 2012).

[17] H. Hong, Q. Xie, W. Ge, F. Qian, F. Fang, L. Shi, Z. Su, R. Perkins, W. Tong, Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics, Journal of Chemical Information and Modeling 48 (2008) 1337–1344.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explorations Newsletter 11 (2009) 10–18.

[19] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection, Chemometrics and Intelligent Laboratory Systems 109 (2011) 146–161.

[20] I.H. Witten, E. Frank, Data mining, practical machine learning tools and techniques Microsoft Research, in: Jim Gray (Ed.), The Morgan Kaufmann Series in Data Management Systems, second ed., Elsevier, 2005.

[21] H. Franco-Lopez, A.R. Ek, M.E. Bauer, Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method, Remote Sensing of Environment 77 (2001) 251–274.

[22] OECD Principles for the validation, for regulatory purposes of (Quantitative) Structure Activity Relationship Models, www.oecd.org, (accessed 12 Oct 2012).

[23] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Molecular Informatics 29 (2010) 476–488.

[24] G. Melagraki, A. Afantitis, H. Sarimveis, P.A. Koutentis, G. Kollias, O. Igglessi-Markopoulou, Predictive QSAR workflow for the in silico identification and screening of novel HDAC inhibitors, Molecular Diversity 13 (2009) 301–311.

[25] G. Melagraki, A. Afantitis, H. Sarimveis, O. Igglessi-Markopoulou, P.A. Koutentis, G. Kollias, In silico exploration for identifying structure-activity relationship of MEK inhibition and oral bioavailability for isothiazole derivatives, Chemical Biology & Drug Design 76 (2010) 397–406.

[26] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, A. Tropsha, A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models, Journal of Chemical Information and Modeling 46 (2006) 1984–1995.

[27] E. Papa, S. Kovarich, P. Gramatica, Development, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers, QSAR and Combinatorial Science 28 (2009) 790–796.

[28] H. Liu, X. Yao, P. Gramatica, The applications of machine learning algorithms in the modeling of estrogen-like chemicals, Combinatorial Chemistry & High Throughput Screening 12 (2009) 490–496.

[29] A. Afantitis, G. Melagraki, P.A. Koutentis, H. Sarimveis, G. Kollias, Ligand - Based virtual screening procedure for the prediction and the identification of novel β-amyloid ggregation inhibitors using Kohonen maps and Counterpropagation Artificial Neural Networks, European Journal of Medicinal Chemistry 46 (2011) 497–508.

[30] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, O. Igglessi-Markopoulou, G. Kollias, A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs, Molecular Diversity 14 (2010) 225–235.

[31] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR and Combinatorial Science 22 (2003) 69–77.

[32] G. Melagraki, A. Afantitis, Ligand and structure based virtual screening strategies for hit-finding and optimization of Hepatitis C virus (HCV) inhibitors, Current Medicinal Chemistry 18 (2011) 2612–2619.

[33] R. Todeschini, V. Consonni, in: R. Mannhold, H. Kubinyi, G. Folkers (Eds.), Molecular Descriptors for Chemoinformatics, Wiley - VCH, Weinheim, 2009.

[34] J. Devillers, A.T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach Science Publishers, The Netherlands, 1999.

[35] F. Burden, Molecular identification number for substructure searches, Journal of Chemical Information and Computer Sciences 29 (1989) 225–227.

[36] F. Burden, M. Polley, D. Winkler, Toward novel universal descriptors: Charge fingerprints, Journal of Chemical Information and Modeling 49 (2009) 710–715.

[37] V.D. Mouchlis, G. Melagraki, T. Mavromoustakos, G. Kollias, A. Afantitis, Molecular modeling on pyrimidine-urea inhibitors of TNF-α production: An integrated approach using a combination of molecular docking, classification techniques, and 3D-QSAR CoMSIA, Journal of Chemical Information and Modeling 52 (2012) 711–723.

[38] A. Afantitis, G. Melagraki, H. Sarimveis, P.A. Koutentis, J. Markopoulos, O. Igglessi-Markopoulou, Development and evaluation of a QSPR model for the prediction of diamagnetic susceptibility, QSAR and Combinatorial Science 27 (2008) 432–436.