

High Dimensional QSAR Study of Mild Steel Corrosion Inhibition in acidic medium by Furan Derivatives

Abdo M. Al-Fakih^{1,3,*}, Madzlan Aziz^{1,*}, Hassan H. Abdallah⁴, Zakariya Y. Algama^{2,5}, Muhammad H. Lee², Hasmerya Maarof⁴

¹Department of Chemistry, Faculty of Science, University Technology Malaysia, 81310 UTM Skudai, Johor, Malaysia

²Department of Mathematical Sciences, University Technology Malaysia, 81310 UTM Skudai, Johor, Malaysia

³Department of Chemistry, Faculty of Science, Sana'a University, Sana'a, Yemen

⁴Department of Chemistry, College of Education, Salahaddin University, Erbil, Iraq

⁵Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Iraq

*E-mail: aalfakih2011@gmail.com, madzlan@utm.my

Received: 22 December 2014 / Accepted: 18 January 2015 / Published: 24 February 2015

The inhibition of mild steel corrosion in 1 M HCl by 17 furan derivatives was investigated experimentally using potentiodynamic polarization measurements. The furan derivatives inhibit the mild steel corrosion. The experimental inhibition efficiency (IE) was used in a Quantitative Structure-Activity Relationship (QSAR) study. Dragon software was used to calculate the molecular descriptors. Penalized multiple linear regression (PMLR) was applied as a variable selection method using three penalties namely, ridge, LASSO, and elastic net. A number of 8 and 38 significant molecular descriptors were selected by LASSO and elastic net methods, respectively. The most significant descriptors namely, PJI3, P_VSA_s_4, Mor16u, MATS3p, and PDI were selected by both LASSO and elastic net methods. The elastic net results show low mean-squared error of the training set (MSE_{train}) of 0.0004 and test set (MSE_{test}) of 5.332. The results confirm that the penalized multiple linear regression based on elastic net penalty is the most effective method to deal with high dimensional data.

Keywords: Polarization; corrosion inhibitors; furan derivatives, high dimensional QSAR, penalized multiple linear regression (PMLR)

1. INTRODUCTION

Metal corrosion causes a huge loss in resources and industrial equipment especially in acidic medium. Acid solutions are the most corrosive media because of their widely use in industry [1]. The

most reported corrosion inhibitors are organic compounds with heteroatoms such as oxygen, nitrogen, sulfur and phosphorous and compounds containing multiple bonds [2,3]. Experimental and theoretical methods have been used to investigate the corrosion inhibition efficiency of many organic compounds [4]. Computational methods has become more developed and increasingly used in the corrosion inhibition studies [5]. Quantitative structure activity relationship (QSAR) is a computational modeling method which has been applied in many disciplines of chemistry [6,7].

A good QSAR model should possess high prediction power and prediction reliability [8]. In the QSAR modeling area, compounds are treated as observations and descriptors are treated as explanatory variables. Quantum chemical calculations are the traditional methods used to calculate the molecular descriptors. In addition, software such as Molconn-Z, CODESSA and Dragon are used to calculate descriptors based on the molecular structures [9]. Dragon software has considerable applications in QSAR and scientific studies. A number of 4885 descriptors can be calculated using Dragon software version 6.0 [10].

A problem of high dimensionality in QSAR modeling, which the number of molecular descriptors, \mathbf{p} , exceeds the number of compounds, \mathbf{n} , is one of the new challenges [11]. Statistical issues associated with modeling high-dimensional QSAR include model overfitting and multicollinearity [12,13]. Classical statistical methods such as multiple linear regression (MLR) cannot solve overfitting and multicollinearity issues. Several methods have been proposed to deal with high dimensional data problem. For example, dimensional reduction methods act by representing the original explanatory variables with orthogonal components such as principle component analysis (PCA) [14], and partial least squares (PLS) [15]. Other methods such as penalized regression methods act to do simultaneously shrinkage and variable selection.

Variable selection is the main objective in high dimensional data [16]. The aim of selecting optimal subset of molecular descriptors is to reduce the descriptors number to those that contain relevant information, and thereby to improve QSAR modeling. This should be observed in terms of predictive performance (by decreasing the effect of multicollinearity) and interpretability (to prevent overfitting). A procedure called penalization is used for variable selection in high dimensional data. This penalization attaches a penalty term $P_{\lambda}(\boldsymbol{\beta})$ to the ordinary least squares (OLS) to get a better estimate of the prediction error by avoiding overfitting and multicollinearity.

In this study, corrosion inhibition efficiencies of furan derivatives on mild steel in 1 M HCl solutions were evaluated using electrochemical potentiodynamic polarization. Dragon software version 6.0 was used to calculate the structural-based descriptors. A high number of molecular descriptors with high dimensionality were obtained. High dimensional data is more informative to develop better models; however, it is a big challenge to the classical variable selection methods to deal with such data. Therefore, the aim of this paper is to apply new proposed variable selection methods (i.e. Penalized multiple linear regression (PMLR) based on ridge, LASSO, and elastic net penalties) in the QSAR studies. In addition, the study aims to evaluate 17 furan derivatives as corrosion inhibitors for mild steel in 1 M HCl solution.

2. MATERIALS AND METHODS

2.1. Experimental Preparation of Materials and Inhibitors

A number of 17 derivatives of furan were obtained from Sigma-Aldrich and investigated as corrosion inhibitors of mild steel in 1 M HCl (Table 1). The test solution (1 M HCl) was prepared from analytical grade hydrochloric acid (37 wt. %). The composition of mild steel specimens (wt%) was: C-0.036, Mn-0.172, Cu-0.082, Ni-0.108, Cr-0.053, Al-0.035, Zr-0.146 and Fe balance. The surface of the steel was abraded using 240, 320, 400, 600, 800 and 1500 grades of sand papers. The specimens were well cleaned with deionized water and then again by acetone.

2.2. Experimental Potentiodynamic Polarization Measurements

Potentiodynamic polarization measurements were used to investigate the inhibition efficiency of the inhibitors. Potentiodynamic polarization measurements were carried out at room temperature ($25\pm 1^\circ\text{C}$) using 250 ml of 1 M HCl electrolyte with and without the addition of 0.005 M of the inhibitors. Before the polarization measurements, the system was stabilized within 30 min to reach open circuit potential (OCP) steady state. Polarization curves were recorded at a scan rate of 10 mV/s with a scan range from -0.25 and +0.25 V with respect to OCP. The Autolab Potentiostat/Galvanostat instrument was used to carry out potentiodynamic polarization measurements by recording the Tafel polarization curve. The used cell was a three-electrode cell assembly that contained a 1 cm^2 coupon of a mild steel embedded in a specimen holder. The mild steel specimen acted as working electrode (WE). A platinum electrode was used as a counter electrode (CE). A saturated calomel electrode (SCE) was used as the reference electrode (RE).

2.3. High-Dimensional QSAR Dataset

The dataset consisted of 17 furan derivatives used as corrosion inhibitors. The molecular structures of the dataset compounds were drawn using Chem3D software. The molecular structures were optimized using the molecular mechanics MM2 method and then again by a Molecular Orbital Package (MOPAC) module in Chem3D software. Dragon software Version 6.0 was used to calculate the molecular descriptors based on the optimized molecular structures [10]. A total of 1951 descriptors were calculated. The dataset was randomly split into 70% training set and 30% test set.

2.4. High-Dimensional QSAR Variable Selection

The most informatics descriptors are needed to be selected precisely from the whole dataset molecular descriptors. The problem of variable selection is one of the most prominent problems in QSAR study. The variable selection is to find a subset of significant descriptors to build a QSAR model with better predictive accuracy compared to a model built with whole dataset descriptors. In this work, the obtained dataset was high dimensional data. Unlike classical variable selection methods,

penalization methods can deal with high-dimensional data. In this paper penalized multiple linear regression was applied using three well-known penalties, ridge, LASSO, and elastic net. Although the ridge penalty cannot do variable selection, it is useful to deal with multicollinearity.

In general, classical linear regression assumes that the response variable $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)'$ is a linear combination of p molecular descriptors $\mathbf{x}_1, \dots, \mathbf{x}_p$, an unknown parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)'$, and an additive error term $\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_2)'$. When $n > p$ the usual estimation procedure for the parameter vector $\boldsymbol{\beta}$ is the minimization of the residual sum of squares (RSS) with respect to $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg \min_{\boldsymbol{\beta}} \text{RSS} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1)$$

Then, the OLS estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is obtained by solving Eq. (1). The OLS estimator is optimal within the class of linear unbiased estimators if the molecular descriptors are not correlated. However, multicollinearity occurs if there are highly correlated molecular descriptors in the regression model. This can lead to problems in the computation of the OLS estimator. In the case of high dimensional data, $n < p$, both the design matrix \mathbf{X} and the matrix $\mathbf{X}'\mathbf{X}$ no longer have full rank p . Thus, $(\mathbf{X}'\mathbf{X})^{-1}$ cannot be calculated and the OLS estimator cannot be solved.

The penalization methods are based on penalty terms and should yield unique estimates of the parameter vector $\boldsymbol{\beta}$. An improvement of the prediction accuracy can be achieved by shrinking the coefficients, and an improvement of the interpretability can be done by setting some of the coefficients to zero. Thereby, the obtained QSAR regression models should contain only the relevant molecular descriptors which are easier to interpret. In general, the penalized multiple linear regression (PMLR) is defined as:

$$\text{PMLR} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + P_{\lambda}(\boldsymbol{\beta}). \quad (2)$$

The estimates of the penalized parameter vector are obtained by minimizing Eq. (2) with respect to $\boldsymbol{\beta}$ as shown by Eq. 3:

$$\hat{\boldsymbol{\beta}}_{\text{PMLR}} = \arg \min_{\boldsymbol{\beta}} \text{PMLR}. \quad (3)$$

The penalty term $P_{\lambda}(\boldsymbol{\beta})$ depends on the tuning parameter λ which controls the shrinkage intensity. For the tuning parameter $\lambda = 0$, the obtained result is the OLS estimators. On the contrary, for large values of λ , the influence of the penalty term on the coefficient estimates will increase. Therefore, the penalty region determines the properties of the penalized estimated parameter vector, whereas the desirable molecular descriptors will be the selected variables. Different forms of the penalty terms have been introduced in the literature such as ridge, LASSO, and elastic net penalties.

2.4.1. Ridge Regression

One of the most popular penalties is ridge regression (RR), which introduced by Hoerl and Kennard [17] as an alternative solution to OLS when there is multicollinearity between molecular descriptors. The ridge regression solves the RSS using $P_{\lambda}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p \boldsymbol{\beta}_j^2$. Consequently, the ridge estimate is defined by the Eq. (4):

$$\hat{\beta}_{Ridge} = \arg \min_{\beta} (SSR + \lambda \sum_{j=1}^p \beta_j^2). \quad (4)$$

In RR, the tuning parameter λ controls the amount of shrinkage, but never set molecular descriptor coefficients to be exactly equal zero. Therefore, in high dimensional data when $n < p$, the RR will not perform variable selection. Although RR does not have the variable selection property, it is preferred in high dimensional data since highly correlations between molecular descriptors is expected. Unlike the OLS estimates, the RR is biased. Therefore, this penalized method accepts a little bias to reduce the variance and the mean squared error (MSE). Since the RR cannot perform selection of the variables, further penalization methods were developed such as LASSO and elastic net.

2.4.2. Least Absolute Shrinkage and Selection Operator (LASSO)

Tibshirani [18] proposed the least absolute shrinkage and selection operator (LASSO) as a penalty to perform the variable selection by setting some variable coefficients to zero. It does both continuous shrinkage and automatic variable selection simultaneously. Similar to the RR, the LASSO estimates are obtained by adding the penalty of: $P_{\lambda}(\beta) = \lambda \sum_{j=1}^p |\beta_j|$ to the RSS. The PMLR estimates using LASSO is given by Eq. (5):

$$\hat{\beta}_{LASSO} = \arg \min_{\beta} (SSR + \lambda \sum_{j=1}^p |\beta_j|). \quad (5)$$

Depending on the property of the LASSO penalty, some coefficients will be exactly equal to zero. Hence, LASSO performs the variable selection. Although LASSO is widely used in many applications, it has some drawbacks. One of the drawbacks, it is not robust to high correlation among molecular descriptors and will randomly choose one of these descriptors and ignores the rest. Another drawback of LASSO in high dimensional data is that the maximum number of selected descriptors is equal to n even if there is more descriptors with non-zero coefficients in the final model. Therefore, elastic net penalized method was developed to overcome the drawbacks of the LASSO.

2.4.3. Elastic Net

Elastic net is a penalized method for variable selection. It was introduced by Zou and Hastie [19] to deal with the drawbacks of LASSO. Elastic net tries to merge both LASSO and ridge penalties, by using ridge regression penalty to deal with high correlation problem and taking the advantage of LASSO penalty of variable selection property. The elastic net estimates for PMLR are defined by Eq. (6):

$$\hat{\beta}_{elastic} = \arg \min_{\beta} (SSR + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2). \quad (6)$$

As it can be observed by Eq. (6), elastic net depends on non-negative two tuning parameters λ_1, λ_2 . According to lemma 1 in Zou and Hastie [19], to find the estimates of $\hat{\beta}_{elastic}$ in Eq. (6), the given data set (\mathbf{y}, \mathbf{X}) is extended to an augmented data $(\mathbf{y}^*, \mathbf{X}^*)$ and defined by Eq. (7):

$$\mathbf{X}_{(n+p,p)}^* = (1 + \lambda_2)^{-\frac{1}{2}} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2} \mathbf{I} \end{pmatrix}, \quad \mathbf{y}_{(n+p,1)}^* = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \quad (7)$$

As a result of this augmentation, the elastic net can be written and solved as a LASSO penalty. Hence, the elastic net can select all p molecular descriptors in the high dimensional when $n < p$ and not only n molecular descriptors since \mathbf{X}^* has rank p .

3. RESULTS AND DISCUSSION

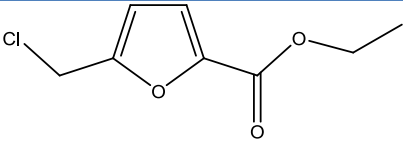
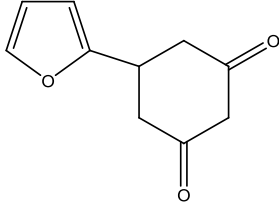
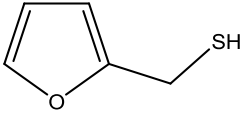
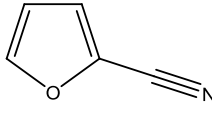
3.1. Experimental Potentiodynamic Polarization Measurements

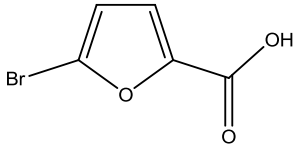
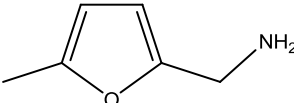
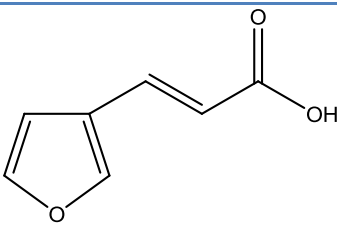
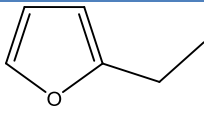
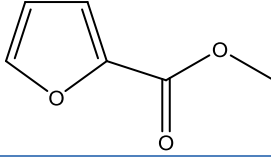
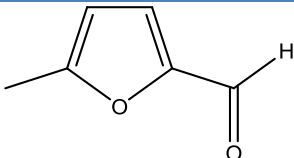
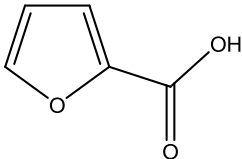
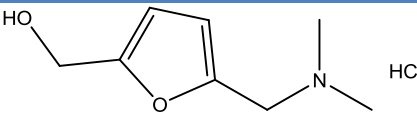
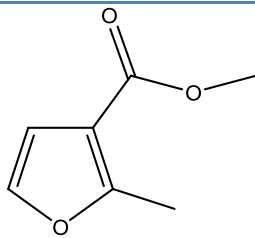
Tafel polarization curves were recorded for mild steel corrosion in 1 M HCl in the presence and absence of the inhibitors (Figure 1). Tafel curves were analyzed and the values of the electrochemical parameters are given in Table 2. The IE was calculated using the Eq. (8):

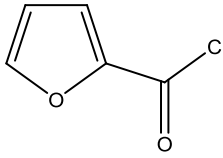
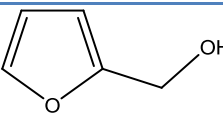
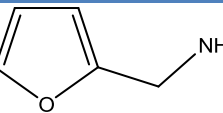
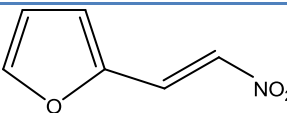
$$IE \% = \frac{i_{corr}^\circ - i_{corr}}{i_{corr}^\circ} \times 100 \quad (8)$$

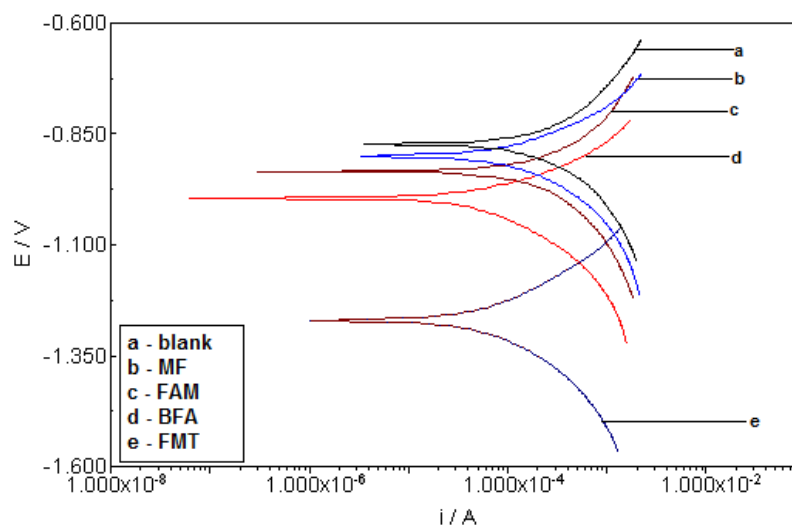
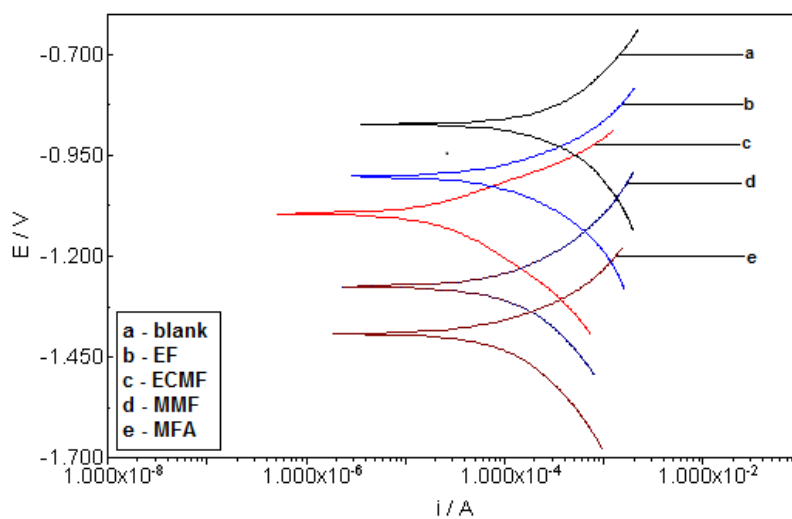
where i_{corr}° and i_{corr} are corrosion current densities in the absence and presence of the inhibitors, respectively. The IE values are given in Table 2.

Table 1. Names and structures of the furan derivatives used as corrosion inhibitors

No.	Compound	Abbreviation	Structure
1	Ethyl 5-(chloromethyl)-2-furoate	ECMF	
2	5-(2-Furyl)-1,3-cyclohexanedione	FCH	
3	2-Furanmethanethiol	FMT	
4	2-Furonitrile	FN	

5	5-Bromo-2-furoic acid	BFA	
6	5-Methylfurfurylamine	MFA	
7	trans-3-Furanacrylic acid	FAA	
8	2-Ethylfuran	EF	
9	Methyl 2-furoate	MF	
10	5-Methylfurfural	MFF	
11	2-Furoic acid	FA	
12	5-(Dimethylaminomethyl)furfuryl alcohol hydrochloride	DMFA	
13	Methyl 2-methyl-3-furoate	MMF	

14	2-Furoyl chloride	FC	
15	Furfuryl alcohol	FFA	
16	Furfurylamine	FAM	
17	2-(2-Nitrovinyl)furan	NVF	



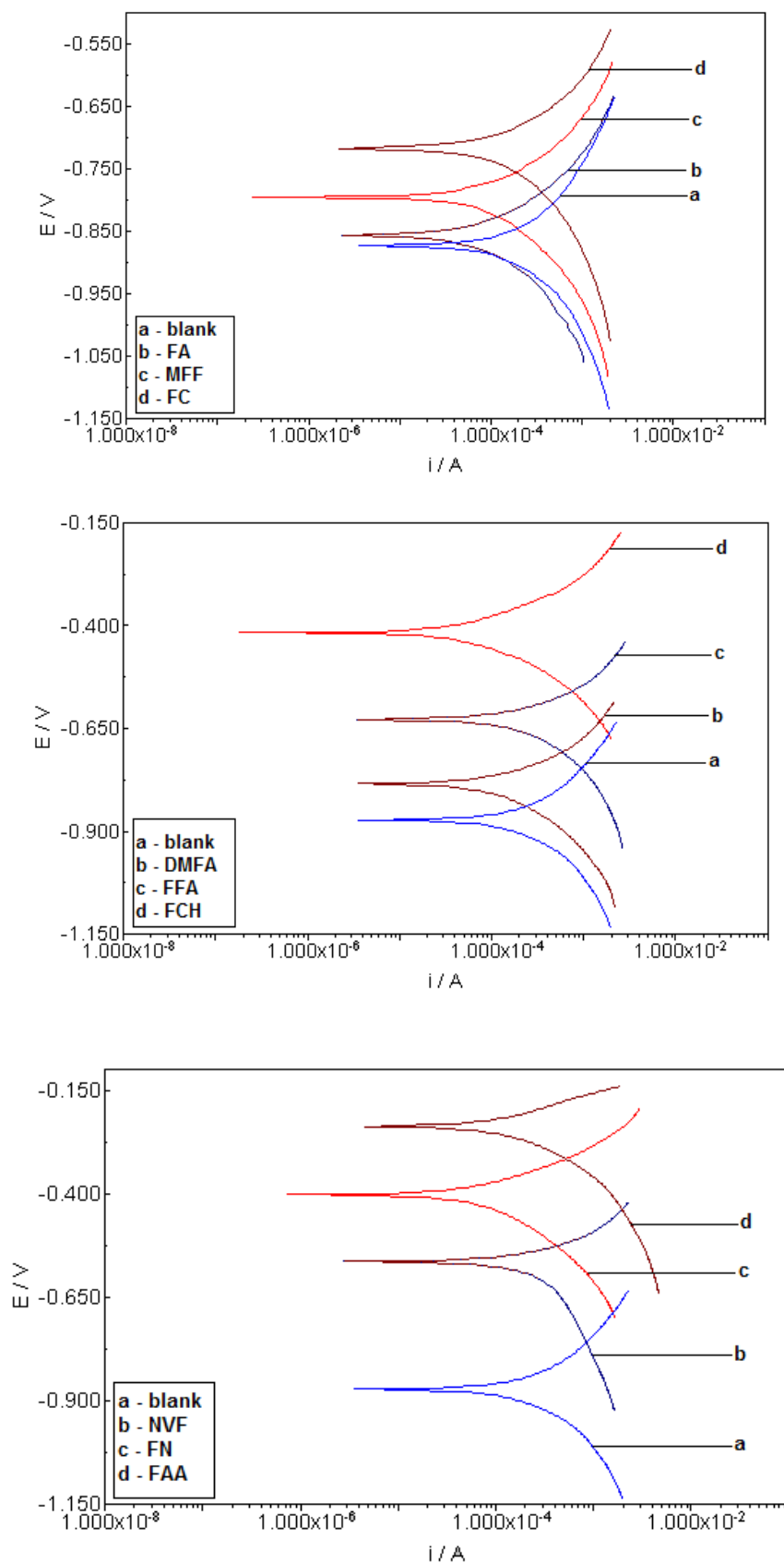


Figure 1. Tafel polarization curves of mild steel in 1 M HCl without and with 0.005 M of furan derivatives

It is clear from the Tafel curves, Figure 1, that the addition of the furan derivatives affected both the cathodic and anodic parts of the curves, indicating that the furan derivatives act as cathodic and anodic inhibitors. In addition, the values of both anodic (b_a) and cathodic (b_c) Tafel constants, Table 2, are changed in the presence of the inhibitors. This also confirms the mixed mode inhibition mechanism of the inhibitors. The corrosion current density (i_{corr}) was calculated by extrapolating anodic and cathodic Tafel lines. The results reveal that in the presence of the inhibitors, the values of corrosion current densities (i_{corr}) are decreased. This behavior reflects the ability of furan derivatives to inhibit the corrosion of mild steel in 1 M HCl solution. The corrosion inhibition efficiencies of the furan derivatives ranged from (96.54% for ECMF) to (35.96% for NVF). The differences in the inhibition efficiencies are due to the electronic structures of inhibiting molecules, steric factor, aromaticity, electron density at the donor site, molecular area and molecular weight of the inhibitor. Corrosion inhibitors act by the adsorption on the metal's surface via lone pair and π -electrons donated by the heteroatoms and multiple bonds. The higher number of lone pair and π -electrons increases the electron density on the molecule and causes a strong interaction with the metal's surface. For example, the presence of chloride atom ($-\text{Cl}$) on the ECMF molecule together with other three heteroatoms increases the electron density on the molecule and contributes to enhance the inhibition efficiency to attain 96.54%. In addition, the high inhibition efficiencies of the FCH, FMT, FN and BFA derivatives of 89.93, 89.44, 89.03 and 88.60%, respectively could be attributed to the presence of different heteroatoms and functional groups that can donate electrons. Furthermore, the obtained inhibition efficiency of MFA and FAM inhibitors is 84.77 and 41.75%, respectively. The MFA molecule gives more inhibition efficiency compared to FAM. This enhanced efficiency could be attributed to the replacement of hydrogen atom in furan ring by an alkyl group ($-\text{CH}_3$). Such group with an inductive effect (+I), would assist to increase electron density and cause an enhancement in the inhibition efficiency. Moreover, the lower inhibition for NVF of 35.96% can be attributed to the presence of the nitro group ($-\text{NO}_2$) as an electron-withdrawing group which maybe causes electron deficiency on the furan ring.

Table 2. Corrosion inhibition efficiency (IE) and electrochemical parameters obtained from Tafel polarization curves

Inhibitors	b_a (mV dec ⁻¹)	b_c (mV dec ⁻¹)	E_{corr} (V)	I_{corr} ($\mu\text{A cm}^{-2}$)	IE (%)
Blank (1 M HCl)	424	474	-0.873	655.10	-
ECMF	121	168	-1.094	22.63	96.54
FCH	111	129	-0.417	65.94	89.93
FMT	149	163	-1.271	69.14	89.44
FN	98	158	-0.402	71.86	89.03
BFA	106	172	-0.995	74.64	88.60
MFA	138	244	-1.392	99.77	84.77
FAA	104	134	-0.238	142.50	78.24
EF	160	217	-1.002	148.40	77.34

MF	134	175	-0.900	152.30	76.75
MFF	151	198	-0.796	156.30	76.14
FA	156	221	-0.857	156.80	76.06
DMFA	142	216	-0.785	183.50	71.99
MMF	256	367	-1.275	209.30	68.05
FC	165	249	-0.717	234.20	64.25
FFA	145	235	-0.628	301.80	53.93
FAM	245	357	-0.935	381.60	41.75
NVF	137	695	-0.564	419.50	35.96

3.2. Variable Selection

In order to select the most informatics descriptors with PMLR, the 17 compounds were randomly divide into a training set of 70% and a test set of 30%. The training set was used to select the descriptors by finding the optimal value for the tuning parameter. The test set was then used to validate the quality of the selected descriptors. The partition process of selecting the training and test observations was repeated 100 times. To find the optimal values of the tuning parameters (λ) for both ridge and LASSO, and the pair of two tuning parameters (λ_1, λ_2) for elastic net, K-fold cross-validation method was used with $K=5$. The tuning parameter for ridge and LASSO was 3.319 and 1.286, respectively. For the tuning parameters of elastic net, the solution was different because this method required prior value of λ_2 to transform the original training dataset to the new augmented training dataset. A sequence of values for λ_2 was given, where $0 \leq \lambda_2 \leq 100$. For each value of λ_2 , a 5-fold cross-validation was performed to select the remaining tuning parameters. The best value for the pair of both tuning parameters was (0.174, 0.01).

Two statistical criteria were used to evaluate PMLR in variable selection, the mean-squared error of the training set ($MSE_{train} = \sum_{i=1}^{n_{train}} (y_{i,train} - \hat{y}_{i,train})^2 / n_{train}$) and the number of the selected molecular descriptors. Figure 2 displays the corresponding boxplots of the training error for the three used PMLR methods. It is clear that elastic net has less variability among the three penalized methods.

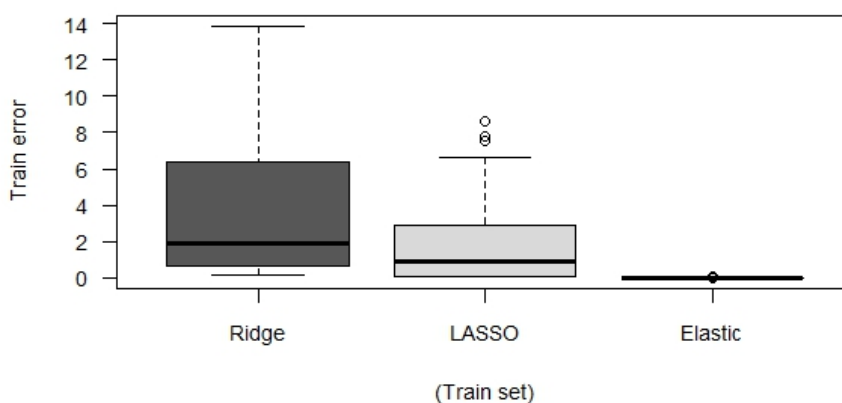


Figure 2. Training error boxplots of the PMLR methods for 100 random partitions

The two evaluation criteria for PMLR are summarized in Table 3. Concerning the MSE_{train} criterion, the MSE_{train} value for the elastic net penalty of 0.0004 was smaller than the MSE_{train} values of 0.9138 for LASSO and 1.8942 for ridge.

Table 3. Median values of the evaluation criteria for the PMLR methods

	MSE_{train}	No. of selected molecular descriptors
Ridge	1.8942	1951
LASSO	0.9138	8
Elastic net	0.0004	38

In terms of the number of selected molecular descriptors, elastic net selected 38 descriptors, whereas LASSO selected 8 descriptors. In ridge penalty, there was no variable selection; therefore, whole variables (1951 descriptors) were selected. It can be observed that elastic net penalty selected descriptors larger than the number of compound. In contrast to elastic net, the number of selected descriptors by LASSO was lower than the number of compounds. Elastic net selected more descriptors compared to LASSO because of the existence of several correlations among the descriptors. Elastic net has the ability to deal with the grouping effect by selecting the correlated molecular descriptors together or to leave them out together. On the contrary, LASSO can deal with grouping effect by selecting only one descriptor randomly from the group of correlated descriptors. For example, elastic net selected Mor16i, Mor25p, Mor16p, Mor16u, Mor11m, Mor19m, Mor23m, Mor31m, Mor16e, and Mor30e descriptors that belong to 3D-MoRSE group. The correlations between most of these descriptors were statistically significant and ranged from 0.55 to 0.96. Figure 3 shows the frequency of the most selected molecular descriptors for both LASSO and elastic net penalties over 100 splits. A number of 28 descriptors, which possessed frequencies higher than 75%, were presented in Figure 3.

Significant descriptors with higher than 95% frequency were selected using elastic net such as MATS3v, MATS3p, P_VSA_s_4, PJI3, Mor16u, and Mor16e. The most frequent descriptors selected using LASSO were PJI3, Mor16u, and P_VSA_s_4 with frequency higher than 55%. The PJI3, P_VSA_s_4, Mor16u, MATS3p, and PDI descriptors were the most significant descriptors since they were selected by both LASSO and elastic net methods.

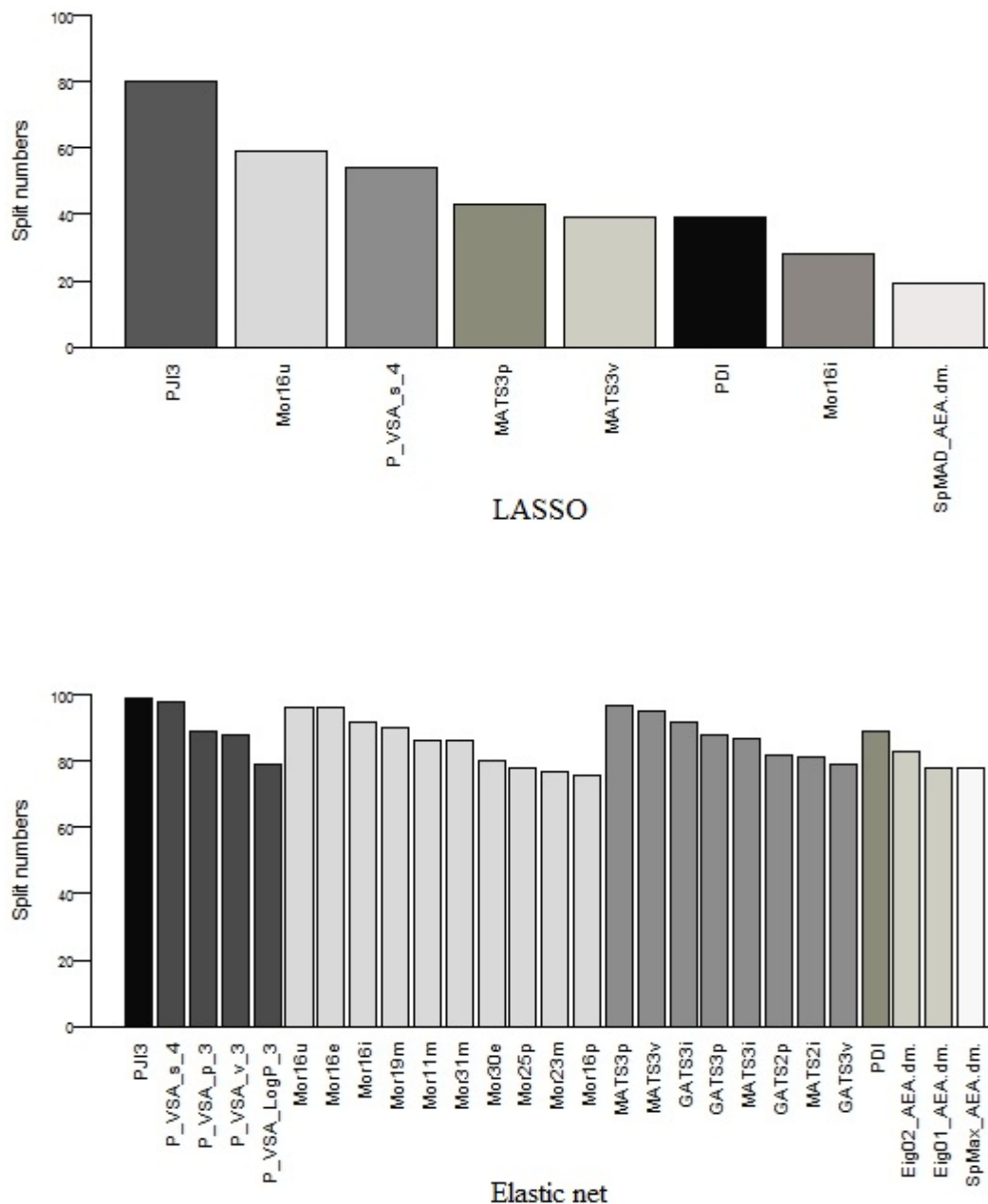


Figure 3. Frequency of the most selected molecular descriptors

3.3. Validation of the PMLR

Test set was used to validate the PMLR using two statistical criteria namely, the mean-squared error of the test set ($MSE_{test} = \sum_{i=1}^{n_{test}} (y_{i, test} - \hat{y}_{i, test})^2 / n_{test}$) and the Pearson correlation between the predicted and the experimental IE in the test set. The validation results are reported in Table 4. The lowest value for MSE_{test} of 5.332 was obtained for elastic net penalty. The Pearson correlation is defined as a correlation between the true IE values and the predicted IE of the test set. The higher Pearson Correlation value of 0.972 was achieved by elastic net. The higher value of the Pearson

correlation, the closer fitted of the predicted IE. It is clear that elastic net was the best in terms of the validity of PMLR followed by LASSO.

Table 4. Median values of the validation criteria for the PMLR methods

	MSE_{test}	Pearson Correlation
Ridge	9.288	0.641
LASSO	7.086	0.738
Elastic net	5.332	0.972

3.4. Interpretation of Descriptors

The interpretation of the descriptors gives insight into the related factors to the corrosion inhibition efficiencies of the inhibitors. The most significant descriptors which were selected by both LASSO and elastic net methods were PJI3, P_VSA_s_4, Mor16u, MATS3p, and PDI. The PJI3 (3D Petitjean shape index) is one of the geometrical descriptors. It was derived based on the molecular shape using the information of the geometry matrix [20]. The descriptor P_VSA_s_4 (P_VSA-like on I-state, bin 4) belongs to P_VSA-like descriptors (the amount of van der Waals surface area (VSA) having a property in a certain range). P_VSA descriptors describe several properties such as electrostatic, lipophilic, steric and pharmacophoric in terms of molecular surface [21]. The P_VSA_s_4 descriptor has been proposed based on the intrinsic state (I-state) property. The intrinsic state (I-state) refers to the ratio of π and lone-pair electrons over the count of σ bonds in the molecular graph. The descriptor Mor16u (signal 16 / unweighted) is one of 3D-MoRSE (3D-Molecule representation of structures based on electron diffraction) descriptors. The 3D-MoRSE descriptors were proposed based on electron diffraction studies which used to prepare theoretical scattering curves [22]. The MATS3p (Moran autocorrelation of lag 3 weighted by polarizability) descriptor is type of 2D autocorrelations descriptors that derived based on the molecular topology with the consideration of chemical information by specified weights of the molecule atoms. The MATS3p descriptor is related to polarizability property of molecule atoms [23]. The descriptor PDI (Para-Delocalization Index) is one of the molecular properties descriptors. Delocalization index quantified the π -delocalization between two atoms. The PDI descriptor has been derived based on electron delocalization as a criterion of aromaticity [20].

4. CONCLUSION

Experimental study was carried out to evaluate 17 furan derivatives as corrosion inhibitors for mild steel in 1 M HCl using potentiodynamic polarization measurements. The experimental results showed the effective performance of the furan derivatives as corrosion inhibitors. The corrosion

inhibition efficiencies of the studied inhibitors ranged from 96.54 to 35.96% for ECMF and NVF, respectively. Theoretical high dimensional QSAR modeling study was conducted using the obtained experimental data. Dragon software was used to calculate the molecular descriptors based on the molecular structures of the inhibitors. Penalized multiple linear regression (PMLR) based on ridge, LASSO, and elastic net were applied. Elastic net penalty show low mean-squared error of the training set and test set of 0.0004 and 5.332, respectively. The results show that the elastic net penalty was the best method to deal with high dimensional data followed by LASSO. Five significant descriptors (i.e. PJI3, P_VSA_s_4, Mor16u, MATS3p, and PDI) were selected by both LASSO and elastic net methods. Therefore, Dragon software is a useful tool that can generate more molecular descriptors that are informative to describe the corrosion inhibition properties of the inhibitors. Penalized multiple linear regression (PMLR) based on different forms of the penalty terms such as LASSO and elastic net can successfully deal with high dimensional data to select the most significant descriptors.

ACKNOWLEDGEMENT

The authors acknowledge the Ministry of Higher Education of Malaysia (MOHE), the Research Management Center (RMC) at the University Technology Malaysia (UTM), and the grant with VOT No. 4F257. The authors are grateful to the Surface and Electrochemical laboratory at Faculty of Science, and Material Laboratories at Faculty of Mechanical Engineering, UTM. The authors acknowledge Dr. Mohamed Noor Hasan for his kind permission of using Dragon software and computational laboratory, Faculty of Science, UTM. We also acknowledge the financial support given by Sana'a University, Sana'a, Yemen.

References

1. M.A. Amin, K.F. Khaled and S.A. Fadel-Allah, *Corros. Sci.* 52 (2010) 140-151.
2. K.F. Khaled and N.A. Al-Mobarak, *Int. J. Electrochem. Sci.* 7 (2012) 1045-1059.
3. N.O. Eddy, F.E. Awe, C.E. Gimba, N.O. Ibisi and E.E. Ebenso, *Int. J. Electrochem. Sci.* 6 (2011) 931-957.
4. M. Mousavi, H. Safarizadeh and A. Khosravan, *Corros. Sci.* 65 (2012) 249-258.
5. J. Zhang, J. Liu, W. Yu, Y. Yan, L. You and L. Liu, *Corros. Sci.* 52 (2010) 2059-2065.
6. F. Bentiss, M. Lebrini, M. Lagrenée, M. Traisnel, A. Elfarouk and H. Vezin, *Electrochim. Acta*, 52 (2007) 6865-6872.
7. N.K. Allam, *Appl. Surf. Sci.* 253 (2007) 4570-4577.
8. L. He and P.C. Jurs, *J. Mol. Graph. Model.* 23 (2005) 503-523.
9. B.M. Bababdani and M. Mousavi, *Chemom. Intell. Lab. Syst.* 122 (2013) 1-11.
10. DRAGON (software for molecular descriptor calculation), version 6.0, Talete srl, Milano, Italy (2010). <http://www.taletemi.it/>
11. P. Filzmoser, M. Gschwandtner and V. Todorov, *J. Chemom.* 26 (2012) 42-51.
12. J. Huang and X. Fan, *Mol. Divers.* 17 (2013) 63-73.
13. W. Zhou, Z. Dai, Y. Chen, H. Wang and Z. Yuan, *Int. J. Mol. Sci.* 13 (2012) 1161-1172.
14. P. Liu and W. Long, *Int. J. Mol. Sci.* 10 (2009) 1978-1998.
15. M.C. Sharma, S. Sharma, N.K. Sahu and D.V. Kohli, *J. Saudi Chem. Soc.* 17 (2013) 219-225.
16. M. Pourahmadi, *High-Dimensional Covariance Estimation: With High-Dimensional Data*, John Wiley & Sons: Hoboken, New Jersey, United States of America (2013).
17. A.E. Hoerl and R.W. Kennard, *Technometrics*, 12 (1970) 55-67.

18. R. Tibshirani, *J. R. Statist. Soc. B*, 58 (1996) 267-288.
19. H. Zou and T. Hastie, *J. R. Statist. Soc. B*, 67 (2005) 301-320.
20. R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics*, 2nd ed.; Wiley-VCH: Weinheim, Germany (2009).
21. S. Balaji, C. Karthikeyan, N.S.H.N. Moorthy and P. Trivedi, *Bioorg. Med. Chem. Lett.* 14 (2004) 6089-6094.
22. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer and V. Steinhauer, *J. Chem. Inf. Comput. Sci.* 36 (1996) 1030-1037.
23. V. Consonni, R. Todeschini and M. Pavan, *J. Chem. Inf. Comput. Sci.* 42 (2002) 682-692.

© 2015 The Authors. Published by ESG (www.electrochemsci.org). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).