

Quantum Modeling and Molecular Dynamic Simulation of some Amino Acids and Related Compounds on their Corrosion Inhibition of Steel in Acidic Media

Bello Abdullahi Umar*, Adamu Uzairu and Gideon Adamu Shallangwa

Department of Chemistry, Ahmadu Bello University, Zaria, Nigeria

Received October 25, 2017; accepted March 13, 2018

Abstract

The inhibition performance of twenty-five amino acids and related compounds was studied by theoretical techniques. The effect of the acidic solution was considered on the molecular dynamics simulation, and the calculated binding energies for most of the inhibitors was $>100 \text{ kcal mol}^{-1}$, suggesting chemisorptive interactions. Density Functional Theory (B3LYP/6-31G*) quantum substance chemical study was utilized to discover the upgraded geometry of the inhibitors. Also, a linear quantitative structure-activity relationship (QSAR) model was built by Genetic Function Approximation (GFA) method, to run the regression analysis and build up connections between various descriptors and the experimental inhibition efficiencies. The prediction of corrosion efficiencies of these inhibitors nicely matched the experimental measurements. The statistical parameters are:

$R_{\text{train}}^2 = 0.963814$, $R_{\text{adjusted}}^2 = 0.95317$, $Q_{\text{LOO}}^2 = 0.921998$ and $R_{\text{test}}^2 = 0.973421$, which indicates that the model was excellent. The proposed model has great dependability, strength, and consistency on checking, with inward and outside approval.

Keywords: amino acids; quantum chemical calculation; molecular dynamics simulation; QSAR; GFA; DFT (B3LYP/6-31G*).

Introduction

Pipelines assume a critical part everywhere throughout the world as equipment for transporting gases and fluids over long distances, from their sources to shoppers. So, corrosion issue exists in the oil business at each phase of creation, from the extraction to refining and storage, preceding use, which requires the utilization of corrosion inhibitors [5]. Numerous strategies, including experimental and theoretical methodologies, have been typically utilized to consider the performance of amino acids as corrosion inhibitors. In spite of the fact that experimental measures, for example, weight reduction technique, potentiodynamic polarization, electrochemical impedance spectroscopy (EIS),

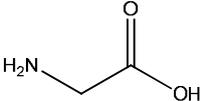
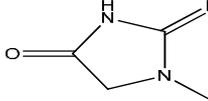
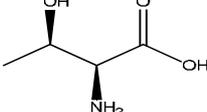
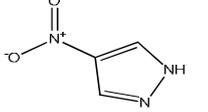
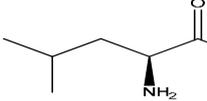
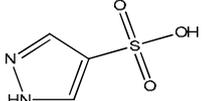
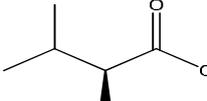
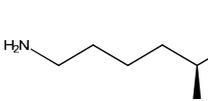
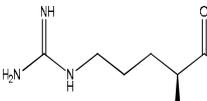
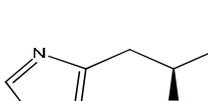
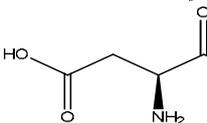
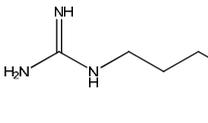
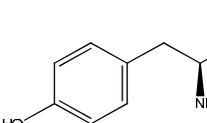
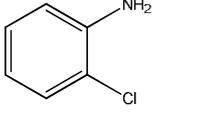
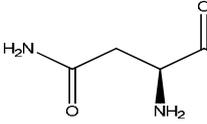
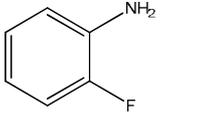
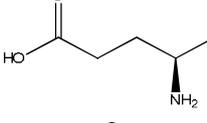
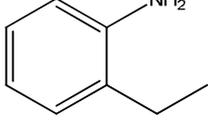
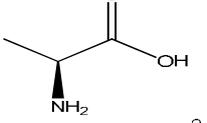
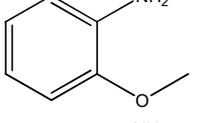
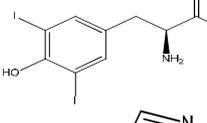
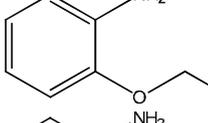
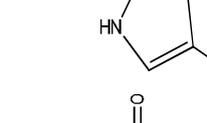
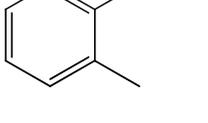
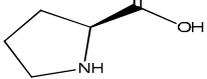
* Corresponding author. E-mail address: m.belkhaouda@uiz.ac.ma

etc. [6], are the most conventional methods to test the inhibition performance, they are costly and tedious, since huge scale trial tests have been completed. Theoretical methods, which can defeat these deficiencies, have gained scientists' incredible consideration as of late. Quantum chemical studies have officially turned out to be extremely helpful in deciding the atomic structure and explaining the electronic structure and reactivity [8]. Consequently, it has turned into a typical practice to complete quantum chemical calculations in corrosion studies. The idea of surveying the productivity of a corrosion inhibitor with the assistance of computational science is to look for compounds with wanted properties utilizing chemical intuition and experience into a mathematically quantified and computerized form. Once a connection between the structure and activity or property is discovered, any number of compounds, including those not yet synthesized, can be promptly screened utilizing computational procedures [9], a set of mathematical equations which are capable of representing accurately the chemical phenomenon under study [10]. Being utilized as a part of science amid the second half of the twentieth century as an expanded measurable examination [11], the quantitative structure-activity relationship (QSAR) technique has recently achieved an uncommon status, formally confirmed by European Union as the fundamental computational apparatus (inside the purported "in silico" approach) for the administrative appraisals of chemicals by methods for non-testing strategies [12]. A structure-activity relationship is generally defined as a mathematical relationship between a property of a chemical (its activity) and a combination of molecular parameters. Normally, the main thrust behind the development of any QSAR is the induction of major conditions which will, somehow, characterize corrosion inhibition efficiency as a function of physical and chemical descriptors characterizing the inhibitor molecules.

Moreover, to consider the adsorption conduct of amino acids onto the metal surface, molecular dynamics simulation was used to research the adsorption configuration and adsorption strength of amino acids onto the metal surface [13]. For instance, Fu [14] researched the inhibition behavior of four amino acids compounds on a Fe(110) surface in an aqueous solution, and found that they could be absorbed onto the iron surface through the heteroatoms and a heterocyclic ring. Though some useful information has been obtained from these studies, there still exists some disparity between the theoretical adsorption model and realistic inhibition systems. Various factors, such as the adsorption of the solvent molecules, the protonation of the inhibitor molecules, and the affection of the acidic solution, which would greatly influence the adsorption behaviors of the amino acid compounds, should also be considered in the molecular dynamics simulation.

In this work, molecular simulation studies were performed to simulate the adsorption of the amino acids on an iron surface. Also, the goal of this study is to encapsulate knowledge about the selected amino acid which is used as corrosion inhibitor for iron in molar hydrochloric (HCl) acid.

Table 1. Inhibition efficiencies and molecular structures of the studied inhibitor series.

S/N	Compound	%IE	S/N	Compound	%IE
1		50	14		34
2		59	15		43
3		63	16		77.4
4		47	17		75.1
5		80	18		41
6		52	19		71
7		39	20		63.62
8		73	21		71.79
9		53	22		63.24
10		51	23		66.83
11		87	24		49.88
12		75	25		60.09
13		67			

Materials and methods

Materials

Twenty-five amino acids and related molecules were collected from the literature [15-17] and investigated in the present study, and their molecular structures and inhibition efficiencies are shown in Table 1. The inhibition efficiencies of all these molecules were obtained by potentiodynamic polarization curves in 1 mol/L hydrochloric acid with 0.01 mol/L concentration of the amino acids against iron corrosion.

Methods

Computational details

Geometry optimization was performed using density functional theory (DFT). The Becke's Three Parameter Hybrid Functional using the Lee-Yang-Parr correlation functional theory was selected for the calculations. Calculations were done using the 6-31+G(d) basis set.

All optimization calculations were done using the Spartan 14v.1.1.0 software. Schematic structures were drawn using the Chemdraw ultra 12.0. The quantum chemical descriptors were calculated using the Spartan'14 V.1.1.0 quantum chemistry package and Material studios 8.0.

Molecular dynamics simulation

The molecular dynamics (MD) simulation was performed using Forcite module of Materials Studio 8.0 program developed by Accelrys Inc [19]. The whole system was performed at 298 K, controlled by the Andersen thermostat, NVE ensemble, with a time step of 1.0 fs, simulation time of 2000 ps, and 5000 Number of steps using the compass force field. The MD simulation was carried out in a simulation box (24.823752A×24.82752A×45.268509A) with periodic boundary conditions. The box includes a Fe slab, an acid solution layer and an inhibitor molecule. Iron (Fe (110)) was selected as the studied surface, since it was density packed and it was the most stable [18]. The iron crystal contained ten layers, and seven layers near the bottom were frozen. The density of the acidic solution layer was set as 1.0 g/cm⁻³. Non-bond interactions, van der Waals and electrostatic were set as atom-based summation method and Ewald summation method, respectively.

Quantitative structure-activity relationship (QSAR)

Quantitative structure-activity relationship (QSAR) was built by the Genetic Function Approximation to correlate the inhibition efficiencies and the molecular structure characteristics of the amino acids' molecules, which were freely available in Materials Studio 8.0. All calculations were performed using the Microsoft office Excel 2013.

The GFA algorithm approach has a number of important advantages over other standard regression analysis techniques. It builds multiple models rather than a single model [21]. It automatically selects which features are to be used in the models, and it is better at discovering combinations of features that take advantage of correlations between multiple features [20]. GFA incorporates Friedman's lack-of-fit (LOF) error measure, which estimates the most

appropriate number of features, resists over fitting, and allows control over the smoothness of fit. Also, it can use a larger variety of equation term types in the construction of its models and finally, it provides, through the study of the evolving models, additional information not available from standard regression analysis.

Training and test set

The training set is comprised of molecules used in the model development, while the test set is made up of molecules not used in building the model; they were used in the external validation of the model generated by the training set. The data-set for the inhibition efficiency was split into the training set and the test set. 18 of the data-sets were used as a training set, while the remaining 7 were used as a test set in line with the optimum splitting pattern of the data-set in the QSAR study [4], as shown in Table 1. The training set was used to generate the model, while the test set was used to evaluate its predictive abilities.

Model validation

Internal and external validation parameters were used to evaluate the reliability and predictive ability of the models. The validation parameters were compared with the standard for the generally acceptable QSAR model, as reported in Table 2.

Table 2. Minimum recommended value of validated parameters for the generally accepted QSAR.

Symbol	Name	Value
R ²	Coefficient of determination	≥ 0.6
P(95%)	Confidence interval at 95% confidence level	> 0.05
Q ²	Cross validation coefficient	> 0.5
R ² _{ext}	Coefficient of determination for external test set	≥ 0.6
R ² -Q ²	Difference between R ² and Q ²	≤ 0.3
N _{ext,tes}	Minimum number of external test set	≥ 5

Internal validation parameters

Lack of fit (LOF)

A “fitness function” or lack of fit (LOF) was used to estimate the quality of the model, so that the best model receives the best fitness score. The error measurement term is determined by equation (1):

$$LOF = \frac{LSE}{\left(1 - \frac{c+d \cdot p}{M}\right)^2} \quad (1)$$

where ‘c’ is the number of basic functions (other than the constant term); ‘d’ is the smoothing parameter (adjustable by the user); ‘M’ is the number of samples in the training set; LSE is the least squares error; and ‘p’ is the total numbers of the features contained in all basic functions [22].

Coefficient of multiple determination (R^2)

To assess the goodness-of-fit, the coefficient of multiple determination is used. R^2 estimates the proportion of the variation in the response that is explained by the predictor:

$$R^2 = 1 - \frac{\sum_{i=1}^I (y_i - \hat{y}_i)^2}{\sum_{i=1}^I (y_i - \bar{y})^2} \quad (2)$$

where y_i is the observed dependent variable, \bar{y} is the mean value of the dependent variable and \hat{y} is the calculated dependent variable.

If there is no linear relationship between the dependent variable and the descriptors, then $R^2 = 0.00$; if there is a perfect fit, then $R^2 = 1.00$. R^2 values higher than 0.5 indicate that the explained variance by the model is higher than the unexplained one [27].

Adjusted R^2 (R^2_{adj})

The value of R^2 can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduce the unexplained variance of the dependent variable. It follows that R^2 should be used with caution. This can be avoided by using another statistical parameter: the so-called adjusted R^2 (R^2_{adj})

$$R^2_{adj} = 1 - (1 - R^2) \left(\frac{I-1}{I-K} \right) \quad (3)$$

R^2_{adj} is interpreted similarly to the R^2 value, except that it takes into consideration the number of degrees of freedom [26].

The value of R^2_{adj} decreases if an added variable to the equation does not reduce the unexplained variable.

Standard error of estimate (SEE)

$$SEE = \sqrt{\frac{\sum_{i=1}^I (y_i - \hat{y}_i)^2}{(I - (K+1))}} \quad (4)$$

The smaller the value of SEE is, the higher the reliability of the prediction. However, it is not recommended to have the standard error of estimate smaller than the experimental error of the corrosion data, because it is an indication of an over fitted model.

F-value

The F-value is determined using equation 5:

$$F = \frac{\sum_{i=0}^I (y_i - \bar{y})^2 / (K-1)}{\sum_{i=1}^I (y_i - \hat{y}_i)^2 / (I-K)} \quad (5)$$

The higher the F-value, the greater the probability that the equation is significant [23].

Cross-validation squared correlation coefficient R^2 (R^2_{cv})

Cross-validation squared correlation coefficient R^2 (LOO- Q^2) is calculated according to the formula:

$$Q^2 = 1 - \frac{\sum(Y_{pred} - Y)^2}{\sum(Y - \bar{Y})^2} \quad (6)$$

where - Y_{pred} and Y indicate predicted and observed activity values, respectively, and \bar{Y} indicates the mean activity value. A model is considered acceptable when the value of Q^2 exceeds 0.5 [24].

In the case of this research, external validation techniques (LMO-Leave Many Out) were applied, in which the 7 compounds of the test set were used for the external validation, and the predicted R^2 for the validation was calculated using equation 2.

External validation parameters*Predicted R^2 (R^2_{pred})*

The predictive R^2 was calculated only based on molecules not included in the training set (test set). Models are generated based on training set compounds, and the predictive capacity of the models was judged based on the predictive R^2 (R^2_{pred}) value which was calculated using equation 7:

$$R^2_{pred} = 1 - \frac{\sum(Y_{pred(test)} - Y_{test})^2}{\sum(Y_{(test)} - \bar{Y}_{training})^2} \quad (7)$$

where $Y_{pred(test)}$ and Y_{test} indicate predicted and observed activity values respectively of the test set compounds and $\bar{Y}_{training}$ indicates the mean activity of the training set. For a QSAR model, the value of R^2_{pred} should be more than 0.5. All the calculated statistical parameters agree with the criteria reported in Table 2.

Applicability domain

The applicability domain (AD) of the QSAR model was used to verify the prediction reliability, identify the problematic compounds and predict the compounds with acceptable activity that fall within this domain. The most common methods used for the determination of the AD of QSAR models have been described by *Gramatica* that used the leverage values for each compound. The leverage approach allows the determination of the position of new chemicals in the QSAR model, *i.e.*, whether a new chemical will lie within the structural model domain or outside of it. The leverage approach along with the Williams Plot is used to determine the applicability domain in all QSAR models. To construct the William Plot, the leverage h_i for each chemical compound – in which QSAR model was used to predict its property – was calculated according to the following equation:

$$h_i = x_i(X^T X)^{-1}x_i^T \quad (8)$$

where x refers to the descriptor vector of the considered compound and X represents the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) was determined as:

$$h^* = \frac{3(p+1)}{N} \quad (9)$$

where N is the number of training compounds and p is the number of descriptors in the model.

Results and discussion

Molecular dynamic simulation study

To get further information about the interaction between the amino acids and related compounds and the Fe surface, molecular dynamics (MD) simulation was performed. In order to build a more reliable model, both water and hydrogen chloride were added to the solution layer for the studied system. In 1 mol/L hydrochloric acid solution, the ratio of water molecules and hydrogen chloride was 500/9. The system was constructed using the amorphous cell module, and the geometry optimization of the system was made. Then, the dynamics process was carried out until equilibrium was reached when both temperature and energy of the system were balanced. Fig. 1 shows the complex molecular dynamic system of inhibitor-14. All other systems for the inhibitors were similarly studied.

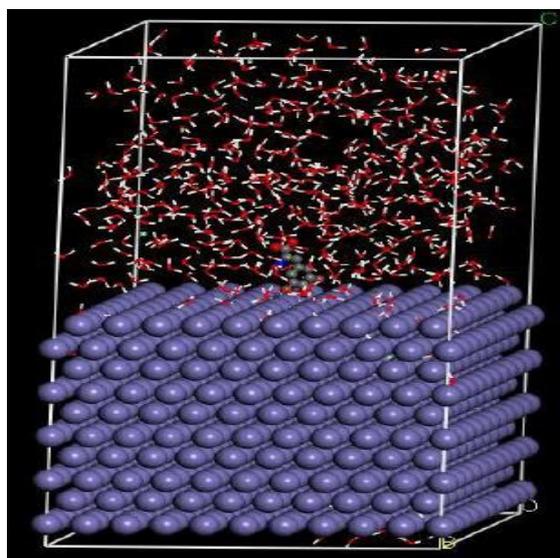


Figure 1. Molecular dynamic system of Fe, inhibitor molecule and acidic layer.

The strength of corrosion inhibitors absorbed onto the iron surface can be expressed by the binding energy, so it will be very interesting to study the binding energies of amino acids absorbed onto the iron surface. The binding energy in the solution can be calculated by the following equations [28]:

$$E_{adsorption} = E_{Total} - (E_{Fe_surface+solution} + E_{Inhibitor+solution}) + E_{Solution} \quad (10)$$

$$E_{binding} = -E_{adsorption} \quad (11)$$

where E_{Total} was the total energy of the system, which includes iron crystal, the adsorbed inhibitor molecule and solution; $E_{\text{Fe-surface+solution}}$ and $E_{\text{inhibitor+solution}}$ were the energies of the system without the inhibitor and the system without the iron crystal, respectively; and E_{Solution} was the energy of the solution in kcal/mol. The calculated adsorption energies and binding energies were listed in Table 3.

Table 3. Adsorption energies and binding energies of the inhibitors.

Compound	Adsorption energy (Kcal/mol)	Binding energy (Kcal/mol)
1	42.826	-42.826
2	-43.164	43.164
3	-356.726	356.726
4	-214.82	214.82
5	399.126	-399.126
6	-719.231	719.231
7	-244.218	244.218
8	-297.5	297.5
9	-284.587	284.587
10	-1051.7	1051.704
11	-117.71	117.71
12	-290.94	290.94
13	-227.216	227.216
14	-281.67	281.67
15	-297.001	297.001
16	-280.415	280.415
17	-771.218	771.218
18	-128.242	128.242
19	-127.376	127.376
20	-685.058	685.058
21	-126.585	126.585
22	-467.954	467.954
23	-953.353	953.353
24	-380.736	380.736
25	-646.243	646.243

In Table 3, adsorption and binding energies calculated between Fe (110) surface and twenty-five amino acids and related compounds, using equation 10 and 11, utilizing Molecular dynamics simulations approach, are given.

Adsorption energy is characterized as the energy released when the inhibitor molecule was adsorbed onto the metal surface. As said in equation 11, the binding is the negative value of the adsorption energy. The most stable low energy configurations obtained for the adsorption of Inhibitor-14 on Fe (110) in 1 M HCl are exhibited in Fig. 2. All different systems for the inhibitors were similarly examined.

It is apparent from the molecular structures of the examined Azoles derivatives that these molecules contain various lone pair electrons on N and S atoms, as well as π -aromatic frameworks.

Therefore, giving the lone pair electrons on heteroatoms to the empty d orbitals of iron, specified inhibitors can form a stable coordination bonding. It can be noticed from Fig. 2 that the inhibitor is adsorbed nearly parallel to the Fe (110) surface, with the assistance of the donation of π electrons of the rings appearing in the structures of the particles and the lone pair of the heteroatoms.

It was accounted for in many investigations that the primary mechanism of the interaction between corrosion inhibitors and iron is by adsorption. In this way, the adsorption energies calculated via molecular dynamics simulations approach can give us an immediate understanding to compare the anticorrosive performances of the inhibitor molecules.

It is seen from Table 3 that the calculated adsorption energies of the examined inhibitors on the iron surface are generally negative values, with the exception of six of the inhibitors (1 and 5) that appear to be positive, which may be expected due to the solvent impact. These negative values denote that the adsorption happening amongst metal and inhibitors could happen spontaneously.

The largest negative adsorption energy indicates that the system is most stable and that adsorption is exceptionally strong.

Then again, a positive and larger value of the binding energy implies that the corrosion inhibitor binds with the Fe (110) surface more easily and firmly [3].

Quantitative structure-activity relationship (QSAR)

Usually, quantitative structure and activity relationships using the GFA method are done in three stages. The first stage is represented in Table 4. The second and third stages, correlation matrix, and regression parameters are presented in Tables 6 and 5, respectively.

Table 4. Study table of descriptors for the studied 25 inhibitor molecules.

Comp.	%IE	Energy (kJ/mol)	Energy (aq)kJ/mol	Dipole moment	Acc. Area	H	G	Chi (3): path	Inform. content (IC)
1	50	-284.423	-284.439	1.26	78.95	-284.337	-284.372	0.816497	2.321928
2	59	-398.946	-398.964	3.66	91.13	-398.824	-398.863	1.782022	2.807355
3	63	-438.262	-438.278	2.79	99.7	-438.111	-438.153	2.103134	2.405639
4	47	-441.676	-441.684	1.78	115.81	-441.472	-441.517	1.981261	2.641604
5	80	-512.3	-512.322	1.59	102.14	-512.167	-512.209	1.981261	2.725481
6	52	-492.435	-492.46	5.29	108.06	-492.29	-492.333	1.981261	2.725481
7	39	-651.569	-651.587	4.65	193.56	-651.382	-651.437	5.048563	3.106891
8	73	-551.617	-551.636	2.07	115.65	-551.455	-551.5	2.33743	2.921928
9	53	-531.746	-531.773	3.7	118.22	-531.571	-531.617	2.33743	2.921928
10	51	-323.737	-323.75	1.71	86.41	-323.621	-323.659	1.333333	2.251629
11	87	-686.356	-686.373	3.4	166.7	-686.124	-686.174	5.294224	3.456565
12	75	-401.155	-401.166	5.63	103.45	-401.001	-401.04	2.342639	2.5
13	67	-396.126	-396.146	3.78	106.08	-395.997	-396.036	2.519712	2.75
14	34	-360.19	-360.216	4.21	115.64	-360.037	-360.076	2.171669	2.5
15	43	-430.697	-430.713	4.73	102.28	-430.616	-430.653	2.342639	2.5
16	77.4	-849.988	-850.023	3.71	111.24	-849.894	-849.935	2.579917	2.503258
17	75.1	-800.548	-800.563	2.7	123.64	-800.371	-800.417	2.151091	3.169925
18	41	-606.522	-606.552	4.27	142	-606.287	-606.336	2.829011	3.084963
19	71	-548.755	-548.783	3.45	130.3	-548.585	-548.629	3.317934	3.095795
20	63.62	-422.979	-422.992	1.75	86.31	-422.871	-422.91	1.732051	1.664498
21	71.79	-438.254	-438.266	2.46	101.57	-438.104	-438.146	2.47418	2.405639
22	63.24	-477.573	-477.584	2.29	111.57	-477.394	-477.438	2.457286	2.641604
23	66.83	-402.364	-402.374	2.47	103.05	-402.189	-402.232	2.47418	2.405639
24	49.88	-363.052	-363.064	2.03	94.28	-362.908	-362.947	1.732051	1.664498
25	60.09	-402.362	-402.373	2.25	103.94	-402.187	-402.229	2.103134	2.405639

A univariate analysis is performed on the inhibition efficiency data from table (Table 4), and the result of the univariate analysis is presented in Table 5. The univariate analysis is a tool that assesses the quality of the data available and its suitability for next statistical analysis. The data in Table 5 show acceptable normal distribution. The normal distribution behavior of the studied data was confirmed by the values of standard deviation, mean absolute deviation, variance, skewness and kurtosis presented in Table 5. A description of these parameters has been reported elsewhere [25].

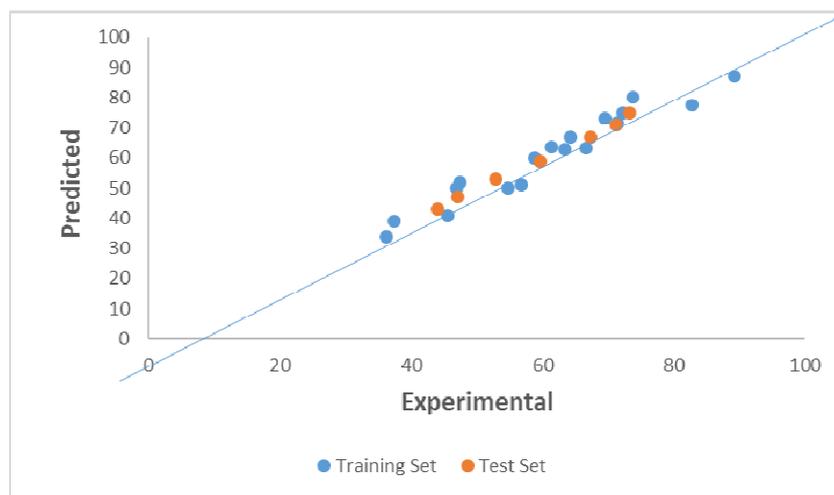


Figure 2. Plot of predicted versus actual inhibition efficiency (%IE) value for model 1.

Table 5. Univariate analysis of the inhibition data.

Statistical parameters	
Number of samples points	25
Range	53
Maximum	87
Minimum	34
Mean	60.51800000
Median	63
Variance	190.75000000
Standard deviation	14.09600000
Mean absolute deviation	11.77970000
Skewness	-0.08762560
Kurtosis	-1.08957000

Table 6 contains a correlation matrix, which gives the correlation coefficients between each pair of columns included in the analysis. Correlation coefficients between a pair of columns approaching +1.0 or -1.0 suggest that the two columns of data are not independent of each other [7]. The correlation matrix can help to identify highly correlated pairs of variables, and thereby identify redundancy in the data set. After constructing the correlation matrix in Table 6, four (4) QSAR generated GFA models for %IE of the compounds are presented below. Out of the 4-models, model-1 was selected as the best for predicting the inhibition efficiency of the studied inhibitors, based on the fact that it has the best statistical

parameters. The validation parameters of the models agree with the standard reported in Table 2.

Four (4) generated models by GFA

Model-1

$$\%IE = -4.874661832 * \text{Dipole Moment} - 1.174299367 * \text{Acc. Area} - 0.059953986 * G + 23.097495708 * \text{Chi (3): path} + 17.990708266 * \text{Information content (IC)} + 75.8665$$

Model-2

$$\%IE = -0.059957497 * \text{Energy} \left(\frac{\text{kJ}}{\text{mol}} \right) - 4.874480061 * \text{Dipole Moment} - 1.174347231 * \text{Acc. Area} + 23.097513744 * \text{Chi (3): path} + 17.988882184 * \text{Information content (IC)} + 75.86772$$

Model-3

$$\%IE = -4.874686058 * \text{Dipole Moment} - 1.174287173 * \text{Acc. Area} - 0.059954016 * H + 23.097440354 * \text{Chi (3): path} + 17.990720692 * \text{Information content (IC)} + 75.86783$$

Model-4

$$\%IE = -0.059955920 * \text{Energy(aq)} \left(\frac{\text{kJ}}{\text{mol}} \right) - 4.874723401 * \text{Dipole Moment} - 1.174343509 * \text{Acc. Area} + 23.097637573 * \text{Chi (3): path} + 17.988854310 * \text{Information content (IC)} + 75.86747$$

Table 6. Correlation matrix of the studied variables.

	%IE	Energy (kJ/mol)	Energy (aq) kJ/mol	Dipole Moment	Acc. Area	H	G	Chi (3): path	Inf. content (IC)
%IE	1	-0.39052	-0.39051	-0.1909	-0.03712	-0.39055	-0.39054	0.189523	0.209786
Energy (kJ/mol)	-0.39052	1	1	-0.24184	-0.62922	1	1	-0.56676	-0.60679
Energy(aq) kJ/mol	-0.39051	1	1	-0.24185	-0.62922	1	1	-0.56675	-0.6068
Dipole Moment	-0.1909	-0.24184	-0.24185	1	0.412699	-0.24183	-0.24183	0.413499	0.384878
Acc.Area	-0.03712	-0.62922	-0.62922	0.412699	1	-0.62909	-0.62911	0.91897	0.725406
H	-0.39055	1	1	-0.24183	-0.62909	1	1	-0.56665	-0.60668
G	-0.39054	1	1	-0.24183	-0.62911	1	1	-0.56666	-0.60669
Chi (3): path	0.189523	-0.56676	-0.56675	0.413499	0.91897	-0.56665	-0.56666	1	0.642714
Inform. content	0.209786	-0.60679	-0.6068	0.384878	0.725406	-0.60668	-0.60669	0.642714	1

Statistical/Validation parameters for the generated models

Statistical parameters for the internal validation of all the 4 models were calculated and presented in Table 8. There is a good agreement of the validation parameters with the standard reported in Table 2.

Comparison of the observed and predicted %IE

The comparison of the predicted inhibition efficiency of the models with the experimental values for the training and test sets is presented in Table 7.

Table 7. Observed versus predicted inhibition efficiency values.

Comps	Observed (%IE)	Predicted values			
		Equation 1	Equation 2	Equation 3	Equation 4
1	50	54.69488	54.69241	54.69506	54.69269
2*	59	59.66207	59.66088	59.66236	59.65964
3	63	63.31383	63.31453	63.31375	63.3144
4*	47	47.05575	47.05809	47.05587	47.05737
5	80	73.6773	73.67623	73.67726	73.67666
6	52	47.49757	47.49751	47.49745	47.49722
7	39	37.46253	37.4599	37.46256	37.45984
8	73	69.58916	69.58893	69.5891	69.58907
9*	53	52.81395	52.81451	52.81413	52.81323
10	51	56.7693	56.76846	56.76938	56.76844
11	87	89.14515	89.14613	89.14524	89.1462
12*	75	73.16818	73.16926	73.16864	73.16781
13	67	64.28804	64.28666	64.28813	64.28668
14	34	36.27307	36.27311	36.27327	36.27346
15*	43	43.98839	43.9852	43.98897	43.98372
16	77.4	82.73443	82.73265	82.73449	82.73294
17	75.1	72.21672	72.21747	72.21672	72.21682
18	41	45.49718	45.50008	45.49707	45.50044
19*	71	71.26173	71.26124	71.26206	71.26019
20	63.62	61.28874	61.28878	61.2887	61.28869
21	71.79	71.29635	71.29681	71.29625	71.29661
22	63.24	66.59273	66.59401	66.59262	66.59376
23	66.83	67.35643	67.35815	67.35653	67.35736
24	49.88	46.96967	46.97131	46.96965	46.97123
25	60.09	58.81336	58.81504	58.8133	58.81485

* = test set

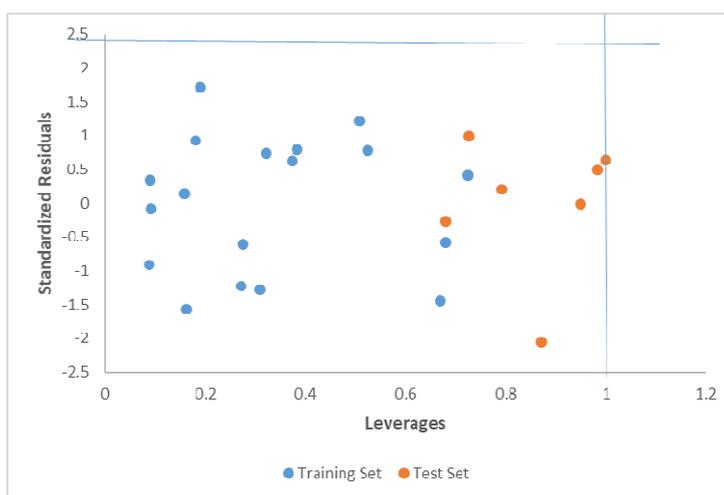


Figure 3. The Williams plot, the plot of standardized residuals versus the leverage value for all the data set.

Plot of predicted versus actual inhibition efficiency (%IE) values

The plot of the predicted versus actual (%IE) values for model-1 is presented in Fig. 2.

Fig. 3 shows the Williams plot of standardized residuals against calculated leverages for both the training and test set.

The leverages for every compound in the dataset were plotted against their standardized residuals, leading to the discovery of outliers and influential chemicals in the models. The applicability domain is established inside a squared area within $\pm 3d$ bound for residuals and a leverage threshold h^* is equal to 1.0 ($N = 18$ and $p=5$)n [1-2]. From our result, it is evident that all the compounds of the training set and test set for the dataset were within the square area (Table 8).

Table 8. Statistical/validation parameters for the generated models.

	Eq. 1	Eq. 2	Eq. 3	Eq. 4
Friedman LOF	94.943	94.94308	94.94501	94.94736
R-squared	0.940152	0.940152	0.940151	0.940149
Adjusted R-squared	0.915215	0.915215	0.915213	0.915211
Cross validated R-squared	0.806823	0.806864	0.806815	0.806839
Significant regression	Yes	Yes	Yes	Yes
Significance of regression F-value	37.70153	37.7015	37.70068	37.69969
Critical SOR F-value (95%)	3.124817	3.124817	3.124817	3.124817
Replicate points	0	0	0	0
Computed experimental error	0	0	0	0
Lack-of-fit points	12	12	12	12
Min expt. error for non-significant LOF (95%)	3.307713	3.307714	3.307748	3.307789

Table 9 gives a list of all the descriptors used to develop the models used in the study.

The result of the GFA QSAR model is in conformity with the standard shown in Table 3, as seen in equation 3. The closeness of coefficient of determination (R^2) to its absolute value of 1.0 is an indication that the model explained a very high percentage of the response variable (descriptor) variation, high enough for a robust QSAR model. The high adjusted R^2 (R^2_{adj}) value and its closeness in value to the value of R^2 implies that the model has excellent explanatory power to the descriptors in it.

Table 9. List of descriptors.

S.No	Name Descriptors
1	Energy(kJ/mol)
2	Energy(aq)kJ/mol
3	Dipole Moment
4	Acc.Area
5	Enthalpy(H)
6	Gibb's Free Energy(G)
7	Simple Path order Chi index(Chi(3)path
8	Information content(IC)

Also, the high Q^2 value and its closeness to R^2 revealed that the model was not over fitted. The high $R^2_{pred.}$ is an indication that the model is capable of providing valid predictions for new molecules that fall within its applicability domain.

F value judges the overall significance of the regression coefficients. The high F value of the model is an indication that the regression coefficients are significant.

Furthermore, the equation contains five descriptors and each descriptor has a positive or negative coefficient attached to it.

These coefficients, along with the value of descriptor, have a significant role in deciding the overall inhibition efficiency of the inhibitor molecules.

Examination of equation 4 shows that the coefficients of each descriptor play an important role in deriving the inhibition efficiency.

From the point of view of inhibition of the molecules in terms of %IE values, the weight of a positive coefficient is very significant because it contributes towards an increased value of %IE.

Table 10 shows the external validation of model 1, and Table 11 shows the calculated R_{ext}^2 .

Table 10. External validation of Model 1.

Comps	Dipole moment	Acc.Area	G	Chi (3): path	Information content (IC)	Actual (%IE)	Predicted (%IE)	Residuals
2	3.66	91.13	-398.863	1.482022	2.807355	59	59.66207	0.662075
4	1.78	115.81	-441.517	1.81261	2.641604	47	47.05575	0.055746
9	3.7	118.22	-531.617	2.13743	2.921928	53	52.81395	-0.18605
12	5.63	103.45	-401.04	3.342639	2.5	75	73.16818	-1.83182
15	4.73	102.28	-430.653	1.753014	2.5	43	43.98839	0.988389
19	3.45	130.3	-548.629	3.317934	3.095795	71	71.26173	0.26173
23	2.47	103.05	-402.232	2.47418	2.405639	66.83	67.35643	0.52643

Table 11. Calculation of Predictive R^2 of model 1.

Comps.	Yp(test)	Ytest	Ym(trn)	[Yp(test) - Ytest] ²	(Ytest - Ym(trn)) ²
2	59.66207	59	61.00666667	0.438343	1.807927
4	47.05575	47	61.00666667	0.003108	194.6282
9	52.81395	53	61.00666667	0.034615	67.12064
12	73.16818	75	61.00666667	3.355583	147.9023
15	43.98839	43	61.00666667	0.976912	289.6218
19	71.26173	71	61.00666667	0.068503	105.1663
23	67.35643	66.83	61.00666667	0.277129	40.3195
				$\Sigma = 5.154192$	$\Sigma = 918.1515779$

$$\text{Pred-}R^2 = 1 - (5.154192/918.1515779) = 0.994386$$

i.e., using the formulae in equation 7.

So, the descriptors with high weight positive coefficients are the most important, followed by descriptors with a low weight negative coefficient and, lastly, the descriptors with high weight negative coefficients.

On the basis of the coefficient values on the model, the associated descriptors are arranged in a sequence pertaining to their contribution towards overall inhibition

efficiency of the inhibitors, in the following increasing order of inhibition efficiency towards steel corrosion.

Chi (3): path > Information content (IC) > **G** > **H** >

Energy $\left(\frac{\text{kJ}}{\text{mol}}\right)$ > Energy(aq) $\left(\frac{\text{kJ}}{\text{mol}}\right)$ > Acc. Area > **Dipole Moment**

Conclusion

This research addresses the QSAR between a set of amino acids and related compounds, and their inhibition efficiency against steel corrosion. Our study developed four GFA-derived models, out of which the optimal model was selected on the basis of its superior statistical significance. The prediction of corrosion efficiencies of these compounds nicely matched the experimental measurements. The molecular surface interactions, estimated using molecular dynamics, suggest that inhibitors bind more strongly (chemisorption) in the presence of an aqueous acidic medium through the heteroatoms, carboxylic group, halogen atoms and through the aromatic ring. Therefore, this will provide a guide on designing more efficient corrosion inhibitors.

References

1. OECD. Guidance document on the validation of (quantitative) structure–activity relationships [(Q)SAR] models. Organization for Economic Co-Operation and Development; 2007.
2. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. Chem Intell Lab Syst. 2015;145:22-9.
3. Obot IB, Kaya S, Kaya C, et al. Density Functional Theory (DFT) modeling and Monte Carlo simulation assessment of inhibition performance of some carbohydrazide Schiff bases for steel corrosion. Physica E: Low-dimensional Syst Nanostruc. 2016;80:82–90.
4. Patil SS. A least square approach to analyze usage data for effective web personalization. Int J Comp Eng Res. 2011;2:68-74.
5. Migahed MA. Corrosion inhibition of steel pipelines in oil fields by N,N-di(poly oxy ethylene) amino propyl lauryl amide. Prog Org Coat. 2005;54:91-98.
6. Zhang D, Cai Q, He X, et al. Inhibition effect of some amino acids on copper corrosion in HCl solution. Mater Chem Phys. 2008;112:353-358.
7. Khaled KF, El-Sherik AM. " Using Molecular Dynamics Simulations and Genetic Function Approximation to Model Corrosion Inhibition of Iron in Chloride Solutions". Int J Electrochem Sci. 2013;10022-10043.
8. Kraka E, Cremer D. Computer design of anticancer drugs. J Am Chem Soc. 2000;122:8245–8264.
9. Karelson M, Lobanov V. Quantum chemical descriptors in QSAR/QSPR studies. Chem Rev. 1996;96:1027-1043.

10. Hinchliffe A. Chemical Modelling From Atoms to Liquids. New York: John Wiley & Sons; 1999.
11. Chatterjee SA, Hadi AS, Price B. Regression analysis by examples. 3rd Ed. New York: John-Wiley; 2000.
12. Benigni R, Bossa C, Netzeva TI, et al. Collection and evaluation of QSAR Models for mutagenicity and carcinogenicity. European Commission-Join Reseach Center: Ispra, Italy. 2007; available online: <http://ecb.jrc.it/qsar/publication/>, accessed January, 2009.
13. Khaled KF. Corrosion control of copper in nitric acid solutions using some amino acids-A combined experimental and theoretical study. Corros Sci. 2010;52:3225-3234.
14. Fu J, Li S, Wang YL, et al. Computational and electrochemical studies of some amino acid compounds as corrosion inhibitors for mild steel in hydrochloric acid solution. J Mater Sci. 2010;45:6255-6265.
15. Khaled KF, Hackerman N. Investigation of the inhibitive effect of orthosubstituted anilines on corrosion of iron in 1 M HCl solutions. Electrochim Acta. 2003;48:2715-2723.
16. Hluchan V, Wheeler BL, Hackerman N. Amino acids as corrosion inhibitors in hydrochloric acid solutions. Mater Corros. 1988;39:512-517.
17. Babić-Samardžija K, Lupu C, Hackerman N, et al. Langmuir. 2003;2:12187-12196.
18. Khaled KF. Molecular simulation, quantum chemical calculations and electrochemical studies for inhibition of mild steel by triazoles. Electrochim Acta. 2008;53:3484-3492.
19. Musa A, Jalgham R, Mohamad A. Molecular dynamic and quantum chemical calculations for phthalazine derivatives as corrosion inhibitors of mild steel in 1M HCl. Corros Sci. 2012;56:176-183.
20. Khaled K, Abdel-Shafi N. Int J Electrochem Sci. 2011;6:4077-4094.
21. Accelrys to Release Enhanced Suite of Chemicals and Materials Modeling and Simulation Tools with Materials Studio(R) 4.1, in: Business Wire, New York; 2006.
22. Kunal R, Roy PP, Paul S, et al. Molecules. 2009;14:1660-1701.
23. Sofie Van Damme. "Quantum Chemistry in QSAR, Quantum Chemical Descriptors, use, benefits and drawback". Thesis. Department of inorganic and physical chemistry, Faculty of Science, Universiteit Gent; 2009. p 39.
24. Wold S, Eriksson L. In Chemometrics Methods in Molecular Design; van de Waterbeemd H Ed. Weinheim: VCH; 1995. pp 309-318.
25. Khaled KF. Corros Sci. 2011;53:3457-3465.
26. Jalali-Heravi MJ, Kyani A. Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: A PCA-MLR-ANN approach. J Chem Inf Model. 2004;44:1328-1335.
27. Brandon-Vaughn OKA. Comprehensive R archive network (CRAN). 2015. Retrieved from <http://CRAN.R-project.org>
28. Pradip BR, Sathish P. Rational design of dispersants by molecular modeling for advanced ceramics processing applications. KONA 2004;22:151-158.