

Using Molecular Dynamics Simulations and Genetic Function Approximation to Model Corrosion Inhibition of Iron in Chloride Solutions

K.F. Khaled^{1,2,*} and A.M. El-Sherik³

¹Electrochemistry Research Laboratory, Chemistry Department, Faculty of Education, Ain Shams Univ., Roxy, Cairo, Egypt

²Materials and Corrosion Laboratory, Chemistry Department, Faculty of Science, Taif University, Saudi Arabia

³Research and Development Center, Saudi Aramco, Dhahran, Saudi Arabia 31311

*E-mail: khaledrice2003@yahoo.com

Received: 14 April 2013 / *Accepted:* 30 May 2013 / *Published:* 1 July 2013

Density functional theory (DFT) calculations have been used to investigate the minimum energy structures of asparagine molecules on iron (Fe) (111) surface. Adsorption of the asparagine molecule on a Fe (111) surface has been studied computationally to generate adsorption configurations and to use the force field method to obtain a ranking of the energies for each generated configuration, thereby indicating the preferred adsorption sites. In this article Monte Carlo simulation has been used to find low energy adsorption sites on both adsorbate (asparagines) — substrate (Fe 111) — systems as the temperature of the system is gradually decreased. The results indicated that asparagine could adsorb on a Fe surface through the nitrogen/oxygen atoms with the lone pair of electrons in its molecule. The Quantitative Structure Activity Relationship (QSAR) method is becoming more desirable for predicting corrosion inhibition properties. The inhibition efficiency of organic compounds is dependent on many basic molecular descriptors, such as dipole moments, molecular surface area, molecular volume, electronic parameters as E_{HOMO} (highest occupied molecular orbital energy); E_{LUMO} (lowest unoccupied molecular orbital energy); and energy gap ($E_{\text{LUMO}} - E_{\text{HOMO}}$). A Genetic Function Approximation (GFA) method was used to run the regression analysis and establish correlations between different types of descriptors and the measured corrosion inhibition efficiencies of 28 amino acids and their related compounds. Similarly, a QSAR equation was developed and used to predict the corrosion inhibition efficiencies of 28 amino acids and their related compounds. The prediction of corrosion efficiencies of these compounds nicely matched the experimental measurements.

Keywords: DFT; Monte Carlo simulations; QSAR; GFA

1. INTRODUCTION

Iron and its alloys are widely used in many applications, which have resulted in research into their resistance to corrosion in various environments [1]. In efforts to mitigate metal corrosion, the primary strategy is to isolate the metal from the corrosive agents. Among the different methods available, to mitigate corrosion, is the use of corrosion inhibitors [1]. Quantum chemical calculations have been widely used as a powerful tool for studying the reaction mechanisms of corrosion inhibition [2-4]. The relationships between structural parameters, such as electronic properties of inhibitors, the frontier molecular orbital energy (E_{HOMO} , E_{LUMO}) and the hydrophobic/hydrophilic nature, the charge distribution of the studied inhibitors and their inhibition efficiencies were investigated in these studies.

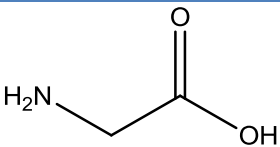
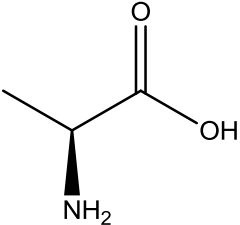
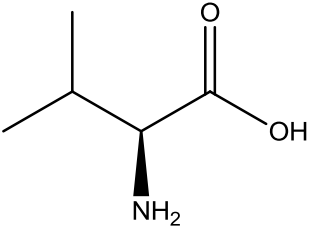
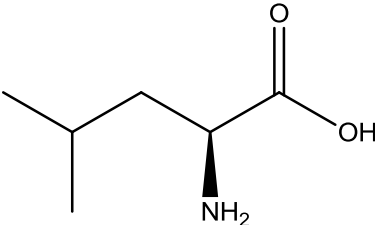
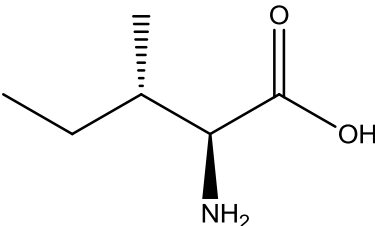
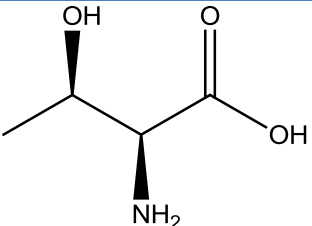
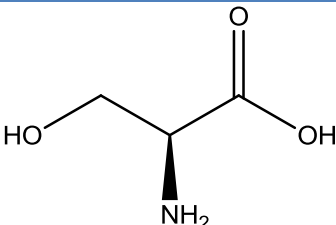
Quantum chemical descriptors have obvious advantages, they are not restricted closely to related compounds, as is often the case with group theoretical, topological and others, and they make interpretation of Quantitative Structure and Activity Relationship (QSAR) equations more straightforward. In addition, they can be obtained without laboratory measurements, thereby saving time and equipment, alleviating safety and disposal concerns [5].

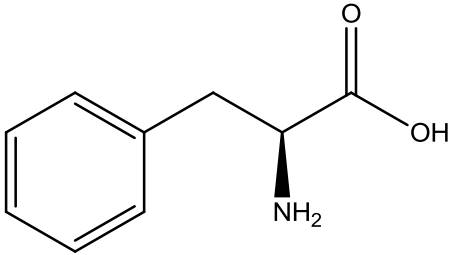
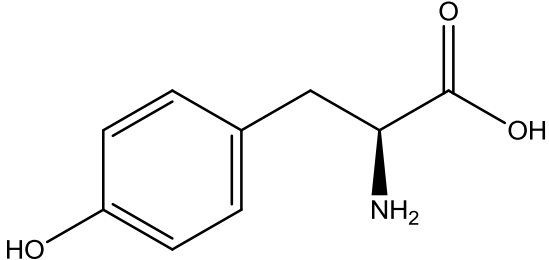
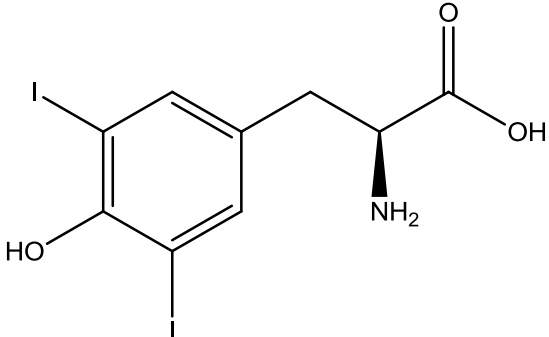
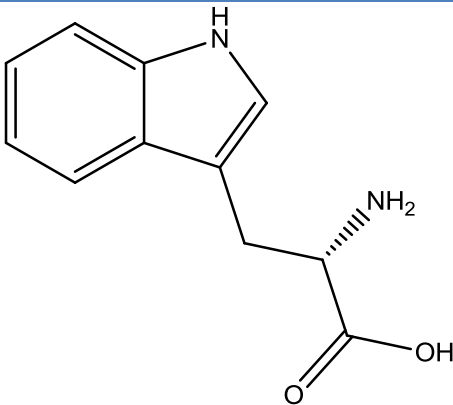
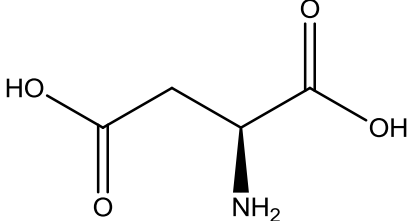
Inhibition of metal corrosion occurs via adsorption of the organic molecules on the metal surface with the polar groups acting as the reactive centers in the inhibiting molecules. The resulting adsorption layer acts as a barrier that isolates the metal surface from the corrosive environment and the protection efficiency depends on the characteristics of the adsorbed layer under the experimental conditions.

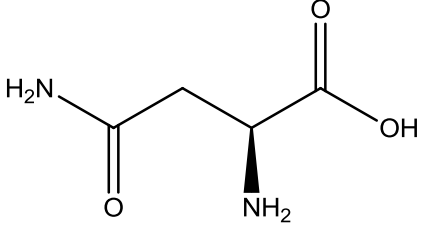
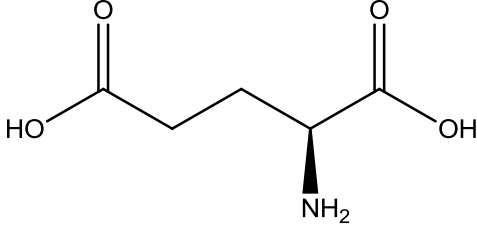
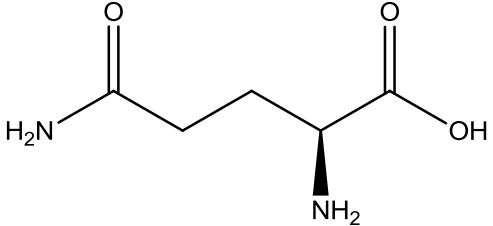
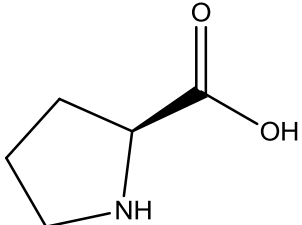
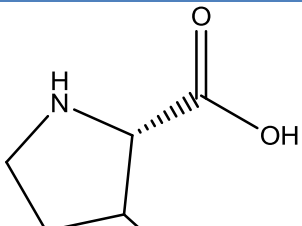
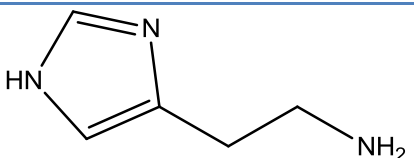
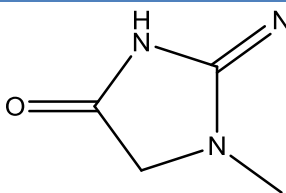
Despite several experimental and computational tools that have been designed to study the structural characteristics of the inhibitor molecules, little is known about the interaction between the adsorbed inhibitor molecules and the corroding metal surface. A practical route to study these complex processes is computer simulations using suitable models.

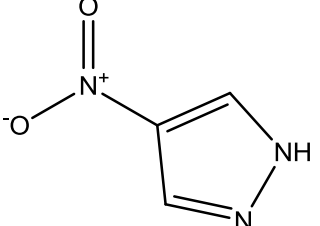
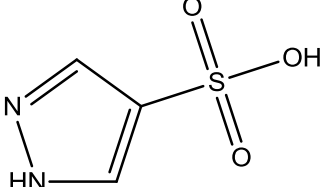
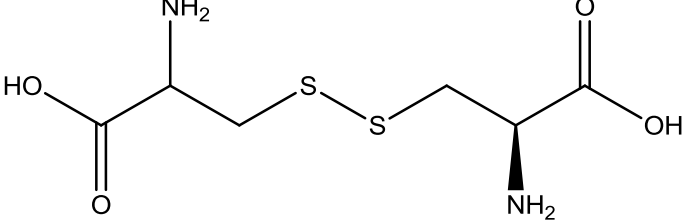
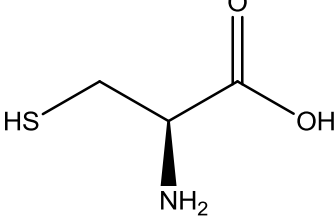
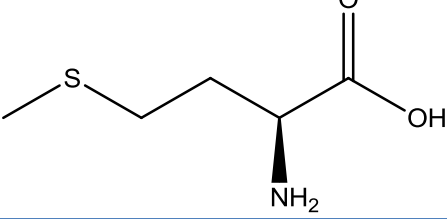
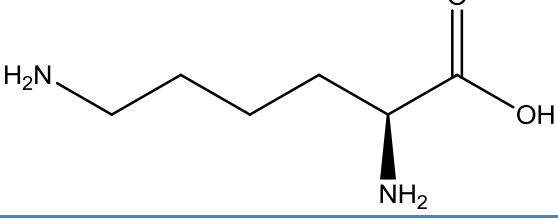
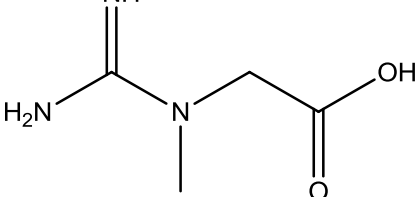
Computer simulations are suitable for more complex systems, such as those involving a relatively large number of molecules where the quantum chemistry computing method is no longer suitable. We have concluded [6] that the quantum mechanical approach may well be able to predict certain molecule structures that are better for corrosion inhibition purposes. This ability to predict is only possible by making the following assumptions: (i) the effect depends only on the inhibitor molecule properties, and (ii) everything else in the inhibitor vicinity is uninvolved either with respect to competition for the surface or with respect to itself. Also, it is clear that there is no general approach for predicting compound usefulness to be a potentially effective corrosion inhibitor or find some universal type of correlation. A number of excluded parameters that should be involved include the effect of solvent molecules, surface nature, and adsorption sites of the metal atoms or oxide sites or vacancies, competitive adsorption with other chemical species in the fluid phase and solubility. In this circumstance, a molecular simulation method is the best choice in an attempt to take into account the effect some of these excluded parameters [7].

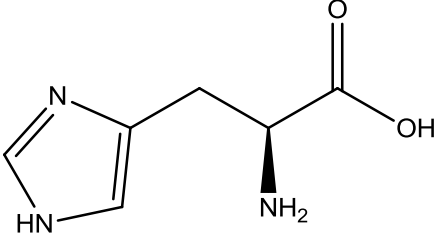
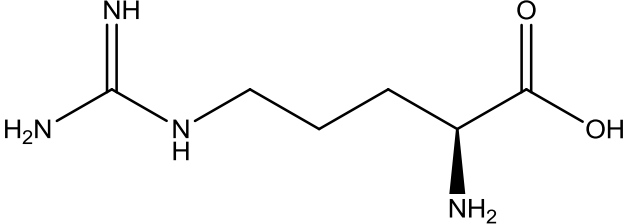
Table 1. Inhibition efficiencies and molecular structures of the studied inhibitor series

Inhibitor name	Structure	Inhibition Efficiency [20, 21]
1 Glycine		50
2 Alanine		51
3 Valine		47
4 Leucine		63
5 Isoleucine		59
6 Threonine		59
7 Serine		63

Inhibitor name	Structure	Inhibition Efficiency [20, 21]
8 Phenylalanine		0
9 Tyrosine		39
10 -3,5 Diodotyrosine		87
11 Tryptophane		80
12 Aspartic acid		52

Inhibitor name	Structure	Inhibition Efficiency [20, 21]
13 Asparagine		73
14 Glutamic acid		53
15 Glutamine		75
16 Proline		34
17 Hydroxyproline		-140
18 Histamine		67
19 Creatinine		43

	Inhibitor name	Structure	Inhibition Efficiency [20, 21]
20	4-nitropyrazole		77.4
21	4--Sulfo-pyrazole		75.1
22	Cystine		-55
23	Cysteine		-179
24	Methionine		59
25	Lysine		71
26	Creatine		-10

Inhibitor name	Structure	Inhibition Efficiency [20, 21]
27	Histidine 	41
28	Arginine 	16

QSAR correlates and predicts physical and chemical properties of chemicals and plays an important role in effective assessment of organic inhibitors. Detailed description of QSAR has previously been presented elsewhere [4, 8-14].

In this work, molecular simulation studies were performed to simulate the adsorption of the asparagine amino acid on an iron surface. Also, the goal of this study is to encapsulate knowledge about how the selected amino acid, which is used as corrosion inhibitors for iron in molar hydrochloric (HCl) acid [15, 16], perform in a structure-activity relationship (SAR) model using the Genetic Function Approximation (GFA) algorithm.

2. COMPUTATIONAL DETAILS

There is no doubt that the recent progress in density functional theory (DFT) has provided a very useful tool for understanding molecular properties and for describing the behavior of atoms in molecules. DFT methods have become very popular in the last decade due to their accuracy, which is similar to other methods (ab initio method), but DFT requires less computational cost from the computational point of view. In agreement with DFT, the energy of the fundamental state of polyelectronic systems can be expressed through the total electronic density, and in fact the use of the electronic density instead of the wave function for the calculation of the energy constitutes the fundamental basis of DFT [17, 18]. Monte Carlo simulations help in finding the most stable adsorption sites on metal surfaces through finding the low-energy adsorption sites on both periodic and nonperiodic substrates or to investigate the preferential adsorption of mixtures of adsorbate components [19].

In the current study, selected amino acids have been simulated as adsorbate on an iron (Fe) (111) surface substrate to find the low energy adsorption sites and to investigate the preferential adsorption of the studied inhibitors. To calculate the adsorption density as well as the binding energy

of the studied inhibitor, the Monte Carlo method has been used. In this computational work, possible adsorption configurations have been identified by carrying out Monte Carlo searches of the configurational space of the iron/asparagine inhibitor system as the temperature is gradually decreased. Asparagine is constructed and its energy was optimized using Forcite classical simulation engine [20, 21]. The geometry optimization process is carried out using an iterative process, in which the atomic coordinates are adjusted until the total energy of a structure is minimized, i.e., it corresponds to a local minimum in the potential energy surface. Geometry optimization is based on reducing the magnitude of the calculated forces until they become smaller than defined convergence tolerances [22]. The forces on the atoms in the studied inhibitors are calculated from the potential energy expression and will, therefore, depend on the force field that is selected [22].

The molecular dynamic (MD) simulations were performed using the software, Materials Studio [23]. The MD simulation of the interaction between the asparagine inhibitor molecule and Fe (111) surface was carried out in a simulation box ($17.38 \text{ \AA} \times 17.38 \text{ \AA} \times 44.57 \text{ \AA}$) with periodic boundary conditions to model a representative part of the interface devoid of any arbitrary boundary effects [22]. The Fe (111) was first built and relaxed by minimizing its energy using molecular mechanics, then the surface area of Fe (111) was increased and its periodicity is changed by constructing a super cell, and then a vacuum slab with 15 \AA thicknesses was built on the Fe (111) surface [22]. The number of layers in the structure was chosen so that the depth of the surface is greater than the non-bond cutoff used in calculation. Using six layers of iron atoms gives a sufficient depth that the inhibitor molecules will only be involved in non-bond interactions with iron atoms in the layers of the surface, without increasing the calculation time unreasonably. This structure is then converted to exhibit 3D periodicity. As 3D periodic boundary conditions are used, it is important that the size of the vacuum slab is enough (15 \AA) that the non-bond calculation for the adsorbate does not interact with the periodic image of the bottom layer of atoms in the surface. After minimizing the Fe (111) surface and the amino acids molecules, the corrosion system will be built by layer builder to place the inhibitor molecules on the Fe (111) surface, and the behaviors of these molecules on the Fe (111) surface were simulated using the COMPASS (condensed phase optimized molecular potentials for atomistic simulation studies) force field. The adsorption locator module in Materials Studio 6.0 [22, 24] has been used to model the adsorption of the inhibitor molecules onto the Fe (111) surface and therefore provides access to the energy of the adsorption and its effects on the inhibition efficiencies of the studied amino acid [19, 25-30]. The binding energy between the asparagine inhibitor and the Fe (111) surface were calculated using the following equation [1, 31]:

$$E_{\text{binding}} = E_{\text{total}} - (E_{\text{surface}} + E_{\text{inhibitor}}) \quad (1)$$

Where E_{total} is the total energy of the surface and inhibitor, E_{surface} is the energy of the surface without the inhibitor, and $E_{\text{inhibitor}}$ is the energy of the inhibitor without the surface.

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 [22, 32]. It automatically selects which features are to be used in the models and is better at discovering combinations of features that take advantage of correlations between multiple features [22]. 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 construction of its models and finally, it provides, through study of the evolving models, additional information not available from standard regression analysis [22, 32].

3. INHIBITORS

Corrosion inhibition experiments for the 28 amino acids and their related compounds presented in Table 1 were conducted in a laboratory [15, 16]. The experimental data was collected from literature [15, 16], and the experimental details are presented briefly in this paper. Measurements were performed with a Gamry Instrument Potentiostat/Galvanostat/ZRA with a Gamry Framework system based on the ESA400 and the VFP600 and Gamry applications, namely DC105 corrosion and EIS300 electrochemical impedance spectroscopy measurements. A computer collected the data, and Echem Analyst 4.0 software was used for plotting, graphing, and fitting data. Tafel curves were obtained by changing the electrode potential automatically from -250 to +250 mV vs. open circuit potential (E_{oc}) at a scan rate of 1 mV/s. The inhibitor concentration was 10^{-2} M. Corrosion tests have been carried out on electrodes cut from iron (Puratronic 99.9999%, from Johnson Matthey Ltd.). Iron rods were mounted in Teflon (surface area 0.28 cm^2). Corrosion inhibition efficiency of the studied amino acids was measured in HCl acid (1 M) solutions in the presence of the 28 amino acids and related compounds at 10^{-2} M concentration. The temperature of the solutions was maintained at $25 \text{ }^\circ\text{C}$. The corrosion rate was determined using the Tafel polarization method [15, 16].

4. RESULTS AND DISCUSSION

4.1. Molecular dynamics simulation study

Before performing the Monte Carlo simulation, molecular dynamics techniques are applied on a system comprising an asparagine amino acid, solvent molecules and iron surface. The selected amino acid is placed on the iron surface, optimized and then run quench molecular dynamics. Figure 1 shows the optimization energy curves for the asparagine amino acid before putting it on the iron surface. It can be seen from Fig. 1 that asparagine is energy optimized as well as the total energy; average total energy; van der Waals energy, electrostatic energy and intramolecular energy for asparagine/solvent/iron surface are calculated by optimizing the whole system and are presented in Fig. 2.

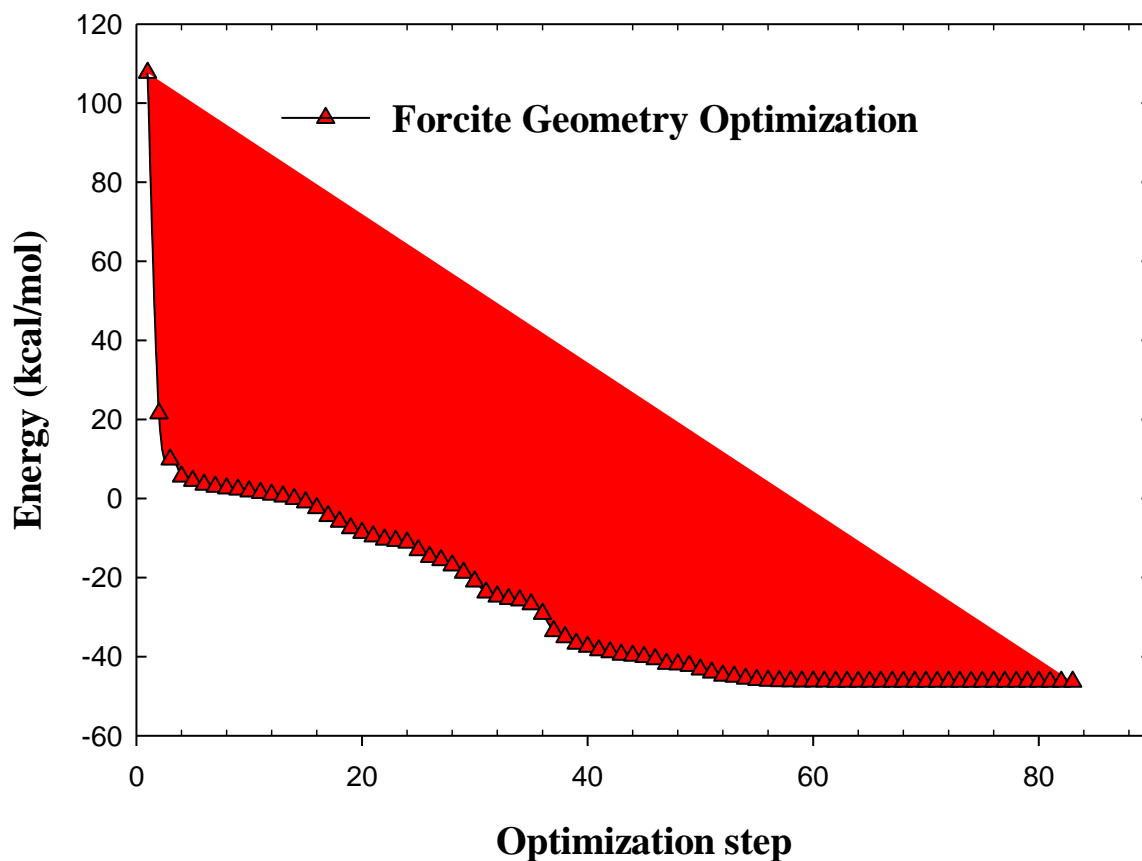


Figure 1. Optimization energy curves for the asparagine molecule before putting it on the iron surface.

The Monte Carlo simulation process tries to find the lowest energy for the whole system. The structures of the adsorbate (asparagine) are minimized until it satisfies certain specified criteria. The Metropolis Monte Carlo method used in this simulation, samples the configurations in an ensemble by generating a chain of configurations, for example m, n, \dots . The step that transforms configuration m to n is a two-stage process [32].

First, a trial configuration is generated with probability α_{mn} . Then, either the proposed configuration, n , is accepted with a probability P_{mn} or the original configuration, m , is retained with a probability $1 - P_{mn}$. The overall transition probability, π_{mn} , is thereby obtained from Eq. 2 [32]:

$$\pi_{mn} = \alpha_{mn} \cdot P_{mn} \quad (2)$$

Using the Adsorption locator simulation module distributed by Accelrys [33], the asparagine molecule – Fe (111) configuration are sampled from a canonical ensemble. In the canonical ensemble, the loading of all asparagine molecules on the Fe (111) substrate, as well as the temperature, are fixed.

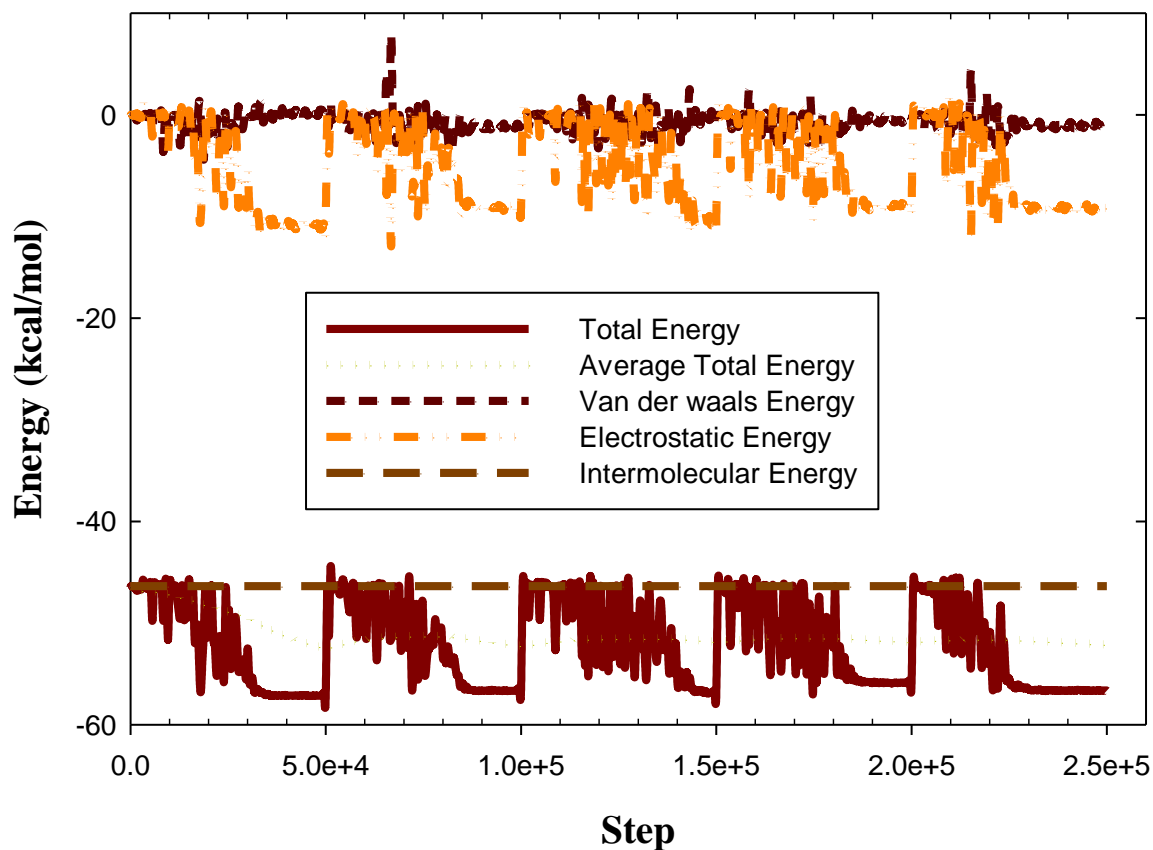


Figure 2. Total energy distributions for asparagine/solvent/iron system.

The probability of a configuration, m , in the canonical ensemble is given by Eq. 3 [34]:

$$P_m = C e^{-\beta E_m} \tag{3}$$

where C is an arbitrary normalization constant, β is the reciprocal temperature, and E_m is the total energy of configuration m .

The reciprocal temperature is given by:

$$\beta = \frac{1}{k_B T} \tag{4}$$

where k_B is the Boltzmann constant and T is the absolute temperature.

The total energy of configuration m is calculated according to the following sum [28, 32]:

$$E_m = E_m^{AA} + E_m^{AS} + U_m^A \tag{5}$$

where E_m^{AA} is the intermolecular energy between the asparagine molecules, E_m^{AS} is the interaction energy between the asparagine molecules and the Fe (111), and U_m^A is the total intramolecular energy of the asparagine molecules. The intramolecular energy of the asparagine is not included as its structure is fixed throughout the simulation; therefore, this energy contribution is fixed and vanishes, since only energy differences play a role in Adsorption Locator calculations.

The total intramolecular energy, U^A , is the sum of the intramolecular energy of all adsorbates of all components [28, 32]:

$$U^A = \sum_{\{N\}_m} u_{\text{intra}} \quad (6)$$

Where $\{N\}_m$ denotes the set of adsorbate loadings of all components in configuration m .

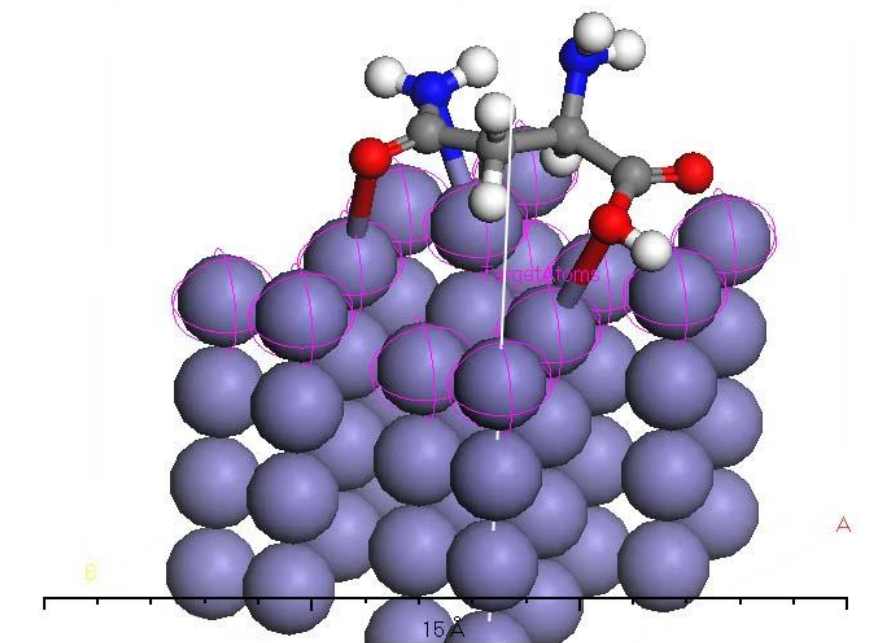


Figure 3. Equilibrium adsorption configurations of asparagines on Fe (111) surfaces obtained by molecular dynamics simulations.

As the simulation starts with a clean Fe (111) substrate, the first stage is to adsorb the specified number of asparagine molecules. This is accomplished by a random series of insertion steps and equilibration moves (only moves that do not change the loading are permitted) until the specified loading has been reached. During this stage, only insertion steps that do not create structures with intermolecular close contacts and that pass all adsorbate location constraints are accepted [28, 32].

The starting configuration will take several steps to adjust to the current temperature. A simulation is, therefore, separated into an equilibration and a production stage. The properties returned at the end of the run are based on the production stage only [32].

In the equilibration and production stages of an Adsorption Locator simulation, each step starts with the selection of a step type using the weights set at the start of the run. The step type can be either a translation or a rotation. After a step type is selected, a random component is chosen and the step type is applied to a random adsorbate of that component [32]. The Metropolis Monte Carlo method is then used to decide whether to accept or reject the change [32].

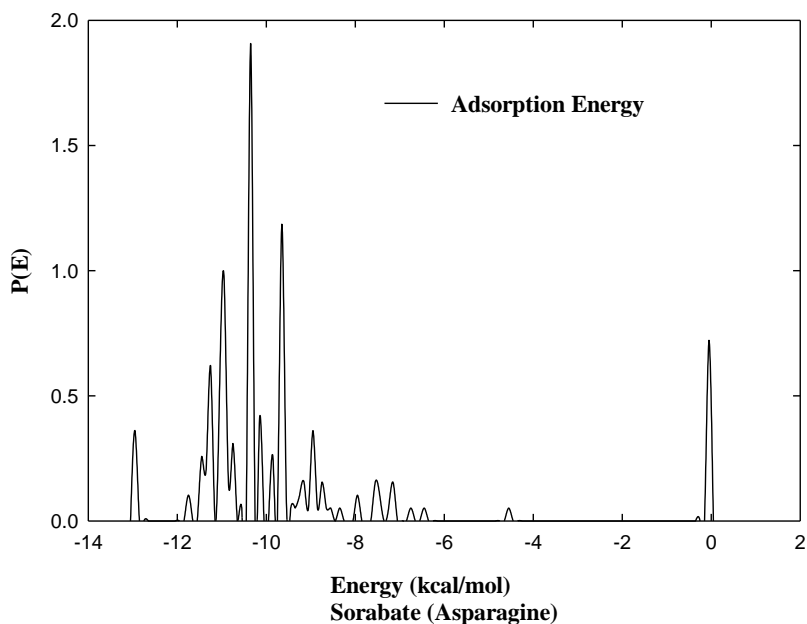


Figure 4. The adsorption energy distribution of the adsorbate (asparagine molecules) on Fe (111) surface.

The Metropolis Monte Carlo method in Adsorption Locator provides four step types for a canonical ensemble: conformer, rotation, translation and regrowth [35]. Figure 3 shows the most suitable asparagine conformation adsorbed on Fe (111) substrate obtained by adsorption locator module [36-38]. The adsorption density of asparagine on the Fe (111) substrate is presented in Fig. 4. As can be seen from Figs. 3 and 4 the asparagine molecule shows the ability to adsorb on a Fe (111) surface. Also, it has high binding energy to the Fe surface as seen in Table 2.

The outputs and descriptors calculated by the Monte Carlo simulation are presented in Table 2.

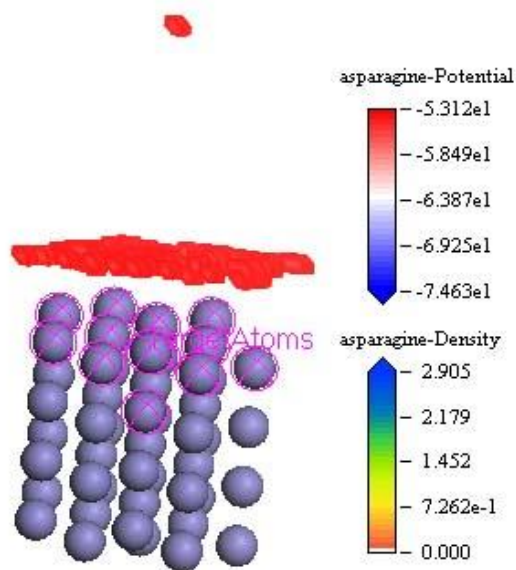


Figure 5. Adsorption density field of asparagine on the Fe (111) substrate.

Table 2. Different adsorption structures and the corresponding adsorption energy

Structures	Total Energy	Adsorption Energy	Rigid Adsorption Energy	Deformation Energy	Asparagine (2): dE_{ad}/dN_i
Substrate	0				
asparagine (2)	-46.3684				
asparagine - 1	-59.3531	-12.9846	-13.0882	0.103588	-12.9846
asparagine - 2	-58.0856	-11.7172	-12.5598	0.842595	-11.7172
asparagine - 3	-57.7907	-11.4222	-11.5475	0.125256	-11.4222
asparagine - 4	-57.653	-11.2846	-12.6871	1.402535	-11.2846
asparagine - 5	-57.6458	-11.2773	-12.0287	0.751371	-11.2773
asparagine - 6	-57.4457	-11.0773	-11.2408	0.163543	-11.0773
asparagine - 7	-57.4327	-11.0643	-11.3543	0.290057	-11.0643
asparagine - 8	-57.431	-11.0625	-11.2698	0.207297	-11.0625
asparagine - 9	-57.4242	-11.0558	-11.3005	0.244678	-11.0558
asparagine - 10	-57.2306	-10.8622	-11.0224	0.160206	-10.8622

The parameters presented in Table 2 include total energy, in kJ mol^{-1} , of the substrate-adsorbate configuration. The total energy is defined as the sum of the energies of the adsorbate components, the rigid adsorption energy and the deformation energy. In this study, the substrate energy (iron surface) is taken as zero. In addition, adsorption energy in kJ mol^{-1} , reports energy released (or required) when the relaxed adsorbate components (asparagine in H_2O) are adsorbed on the substrate. The adsorption energy is defined as the sum of the rigid adsorption energy and the deformation energy for the adsorbate components. The rigid adsorption energy reports the energy, in kJ mol^{-1} , released (or required) when the unrelaxed adsorbate components (i.e., before the geometry optimization step) are adsorbed on the substrate. The deformation energy reports the energy, in kJ mol^{-1} , released when the adsorbed adsorbate components are relaxed on the substrate surface. Table 2 shows also that (dE_{ads}/dN_i), which reports the energy, in kJ mol^{-1} , of substrate-adsorbate configurations where one of the adsorbate components has been removed.

Figure 5 shows the adsorption energy distribution of the asparagine molecules on Fe (111). As can be seen in Fig. 5, the adsorption energy of asparagine reaches (-75 KJ mole^{-1}), which shows the adsorption power for asparagine molecules on Fe (111) surface.

4.2. QSAR

The main problem for QSAR resides not in performing the correlation itself but setting the variable selection for it [22, 39]; the mathematical counterpart for such a problem is known as the “factor indeterminacy” [40, 41] and affirms that the same degree of correlation may be reached with in principle an infinity of latent variable combinations [39]. Fortunately, in chemical-physics there are limited (although sufficient) indicators to be considered with a clear-cut meaning in molecular

structure that allows for rationale of reactivity and bindings [42, 43]. Subsequently, the main point is that given a set of N-molecules, one can choose to correlate their observed activities $A_{i=1,N}$ with M-selected structural indicators in as many combinations as [39]:

$$C = \sum_{k=1}^M C_M^k, C_M^k = \frac{M!}{k!(M-k)!} \quad (7)$$

linked by different endpoint paths, as many as [39]:

$$K = \prod_{k=1}^M C_M^k \quad (8)$$

indexing the numbers of paths built from connected distinct models with orders (dimension of correlation) from k=1 to k=M [39].

In the present study we developed the best QSAR model to explain the correlations between the physicochemical parameters and corrosion inhibition efficiency for 28 amino acids and some related organic compounds as corrosion inhibitors extracted from literature [15, 16].

Table 1 shows the molecular structures for the studied inhibitors with their inhibition efficiencies as presented in the literature [15, 16].

A univariate analysis is performed on the inhibition efficiency data in Table 3 as a tool to assess the quality of the data available and its suitability for next statistical analysis. The data in Table 3 shows 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 3. A description of these parameters have been reported elsewhere [44].

Table 3. Univariate analysis of the inhibition data

Statistical Parameters	
Number of sample points	28
Range	266
Maximum	87
Minimum	-179
Mean	33.95071
Median	52.5
Variance	3.80E+03
Standard deviation	62.7343
Mean absolute deviation	40.836
Skewness	-2.17234
Kurtosis	4.07874

Table 4 shows the structural descriptors for the 28 amino acids and its related compounds. It also records their inhibition efficiencies. Unless otherwise specified, the following units are used for quantities calculated by QSAR descriptors and properties; area (\AA^2), volume (\AA^3), dipole moment (e

Å), HOMO and LUMO (Hartree). The atom volumes and surfaces model calculates surface areas and volumes of surfaces around atomistic structures using the atom volumes and surfaces functionality of the Materials Studio software [45, 46].

The molecular area (vdW area) in Table 4 describes the volume inside the van der waals area of the molecular surface area and determines the extent to which a molecule is exposed to the external environment [22].

Table 4. Descriptors for the studied 28 inhibitor molecules calculated using quantum chemical and molecular dynamics simulation methods

Structure	Inhibitor Efficiency[20,21]	E(HOMO) (Ha)	E (LUMO)(Ha)	[E(LUMO)-E(LUMO)](Ha)	Binding Energy (Kcal/mol)	Adsorption Energy (Kcal/mol)	[Total Energy](Kcal/mol)	Total dipole (VAMP Electrostatics)	Dipole x (VAMP Electrostatics)	Dipole y (VAMP Electrostatics)	Dipole z (VAMP Electrostatics)	Molecular area (vdW area) (Spatial Descriptors)	Molecular volume (vdW volume) (Spatial Descriptors)	GFA equation 1: Inhibition Efficiency
glycine	50	-0.2211	-0.0655	0.1556	144.1	-43.85	27.16403	7.99	0.319	-2.017	1.161	107.4517	73.08615	46.65726
4-nitropyrazole	77.48	-0.1954	-0.0563	0.1391	254.2	-47.82	34.90011	3.12	6.93	-1.345	-0.006	132.2922	96.31805	76.4667
alanine	51	-0.2202	-0.0618	0.1584	148.67	-51.311	20.30506	7.91	0.214	-1.837	0.498	128.4906	89.83167	47.56522
serine	63	-0.2086	-0.0547	0.1539	193.2	-59.028	35.88107	6.43	-0.562	-1.283	2.187	138.6798	99.62781	63.43687
threonine	59	-0.212	-0.0528	0.1591	186.3	-62.847	22.80966	7.3	-1.586	-3.518	4.065	157.0477	115.7812	59.62213
proline	34	-0.2158	-0.054	0.1618	117.2	-63.87	10.4539	8.2	-1.006	-2.792	2.249	153.3667	114.0387	42.51106
valine	47	-0.2174	-0.0582	0.1591	141.12	-64.239	26.4814	6.99	-1.313	-2.341	3.074	161.9411	122.4088	51.65046
cysteine	-179	-0.2052	-0.5656	-0.3604	9.87	-64.89	30.29514	9.4	-0.408	-1.058	1.882	147.0621	108.3996	-178.999
creatinine	43	-0.2179	-0.0547	0.1632	137.67	-65.49	-68.3172	7.97	1.659	-1.444	0.388	148.3478	109.073	43.0377
histamine	67	-0.215	-0.0602	0.1548	203.12	-66.62	-10.7584	6.01	-3.433	-0.335	-1.444	157.6746	116.8251	65.03308
leucine	63	-0.2086	-0.0542	0.1545	194.56	-67.46	11.47275	6.32	-1.314	-3.273	0.487	185.1023	139.8994	62.23819
creatine	-10	-0.2134	-0.029	0.1844	29.6	-69.25	-23.7776	8.41	-0.084	0.729	1.092	171.6891	126.3599	-15.8115
isoleucine	59	-0.2088	-0.0526	0.1562	188.76	-71.47	25.71603	6.51	-1.174	-2.39	3.256	180.5596	138.8267	62.53865
hydroxyproline	-140	-0.2677	-0.0654	0.2023	15.8	-71.73	12.79202	9	-0.948	-1.224	-3.271	159.7079	122.4847	-139.014
asparagine	73	-0.2082	-0.0589	0.1493	241.23	-75.06	-46.3684	5.98	-1.389	2.138	2.141	165.4502	123.3463	75.90911
aspartic acid	52	-0.2063	-0.0516	0.1547	158.68	-75.149	-23.7776	7.84	1.739	0.304	0.809	165.3701	120.8097	57.28794
histidine	41	-0.22	-0.0532	0.1668	128.98	-75.33	21.99655	7.98	-2.211	-5.769	3.547	187.1373	146.3167	34.9999
phenylalanine	0	-0.2215	-0.0486	0.1729	50.3	-75.41	25.26654	8.32	0.785	-1.999	0.027	213.33	167.868	1.359289
lysine	71	-0.2015	-0.0567	0.1447	221.56	-76.17	9.344533	5.99	-0.893	-0.892	1.053	207.6746	153.7672	66.95037
glutamic acid	53	-0.2124	-0.0545	0.1579	161.21	-77.05	-7.44533	7.863	0.044	-5.325	5.785	185.522	137.8673	52.63036
4-sulfoxypyrazole	75.14	-0.1943	-0.0572	0.1371	250.1	-77.78	-9.91027	4.34	-2.347	-0.504	-4.926	159.8533	116.9821	75.57902
3,5-diiodotyrosine	87	-0.2639	-0.1377	0.1262	287.45	-77.86	27.23897	2.12	3.424	-0.528	1.419	280.1272	226.3173	88.17121
methionine	59	-0.216	-0.0602	0.1558	187.17	-77.968	14.87045	7.12	-1.013	-4.581	2.911	191.3474	142.7116	57.58461
glutamine	75	-0.2038	-0.0609	0.1428	249.7	-78.91	-27.3518	5.43	-1.093	-7.204	3.419	189.6844	140.8272	76.5997

This descriptor is related to binding, transport and solubility. The molecular volume (vdW volume) in Table 4, describes the volume inside the van der waals area of a molecule [22]. Total

molecular dipole moment, this descriptor calculates the molecule dipole moments from partial charges defined on the atoms of the molecule [22]. If no partial charges are defined, the molecular dipole moment will be zero. Total energy, HOMO and LUMO energy have been described in our previous studies in detail [44].

For understanding the quantitative structure and activity relationships, statistical analysis using the GFA method, first a study table was belted and presented in Table 4. Second, a correlation matrix was derived, and then regression parameters were obtained. Table 4 shows the structural descriptors for the 28 amino acids and its related compounds used in this study (as a training set). The structure descriptors presented in Table 4 include total energy, HOMO and LUMO energy as well as the area and volume of the studied molecules. Also, the adsorption energy and binding energy have been used for the first time in QSAR studies. Adsorption energy and binding energy have been calculated using adsorption locator and discover modules included in materials studio software, are used in understanding the QSAR.

Table 5. Correlation matrix of the studied variables

	B: Inhibition Efficiency [20, 21]	C: E(HOMO) (Ha)	D: E (LUMO) (Ha)	E: [E(LUMO) - E(LUMO)] (Ha)	F: Binding Energy (Kcal/mol)	G: Adsorption Energy (Kcal/mol)	H: [Total Energy] (Kcal/mol)	I: Total dipole (VAMP Electrostatics)	M: Molecular area (vdW area) (Spatial Descriptors)	N: Molecular volume (vdW volume) (Spatial Descriptors)
B: Inhibition Efficiency	1	0.260799	0.594229	0.544128	0.837994	0.080437	0.015057	-0.62235	0.034239	0.037099
C: E(HOMO) (Ha)	0.260799	1	-0.00351	-0.16446	0.173008	0.054463	-0.12415	-0.00149	-0.24957	-0.29717
D: E (LUMO)(Ha)	0.594229	-0.00351	1	0.986954	0.245998	-0.1067	-0.15711	-0.12138	0.101447	0.084384
E: [E(LUMO)-E(LUMO)](Ha)	0.544128	-0.16446	0.986954	1	0.214764	-0.11397	-0.13499	-0.11949	0.140207	0.13104
F: Binding Energy (Kcal/mol)	0.837994	0.173008	0.245998	0.214764	1	0.066709	0.177904	-0.86751	0.041178	0.058703
G: Adsorption Energy (Kcal/mol)	0.080437	0.054463	-0.1067	-0.11397	0.066709	1	0.103563	0.023798	-0.83649	-0.81538
H: [Total Energy](Kcal/mol)	1.51E-02	-0.12415	-0.15711	-0.13499	0.177904	0.103563	1	-0.38487	0.137823	0.198367
I: Total dipole (VAMP Electrostatics)	-0.62235	-0.00149	-0.12138	-0.11949	-0.86751	0.023798	-0.38487	1	-0.24984	-0.28421
M: Molecular area (vdW area) (Spatial Descriptors)	0.034239	-0.24957	0.101447	0.140207	0.041178	-0.83649	0.137823	-0.24984	1	0.994539
N: Molecular volume (vdW volume) (Spatial Descriptors)	0.037099	-0.29717	0.084384	0.13104	0.058703	-0.81538	0.198367	-0.28421	0.994539	1

Table 5 contains a correlation matrix, which gives the correlation coefficients between each pair of columns included in the analysis in Table 4. 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. Correlation matrix can help to identify highly correlated pairs of variables, and thereby identify redundancy in the data set. A correlation coefficient close to 0.0 indicates very little correlation between the two columns. The diagonal of the matrix always has the value of 1.0. To aid in visualizing the results, the cells in the correlation matrix grid are colored according to the correlation value in each cell. A standard color scheme is used when the correlation matrix is generated: $+0.9 \leq X \leq +1.0$ (orange), $+0.7 \leq X < +0.9$ (yellow), $-0.7 < X < +0.7$ (white), $-0.9 < X < -0.7$ (yellow) and $-1.0 \leq X \leq -0.9$ (orange) [20]. Inspection of Table 5 shows that the descriptors most highly correlated with corrosion inhibition efficiency include: E_{LUMO} , E_{HOMO} and energy gap, binding energy and dipole moment. After

constructing the correlation matrix both the GFA algorithm and neural network analysis will be used to perform a regression analysis.

After constructing the correlation matrix in Table 5, it is now ready to perform a regression analysis of the descriptor variables compared against the measured corrosion inhibition values. There are two separate issues to consider. First, there are many more descriptor variables than measured inhibition values, so we should reduce the number of descriptors. Typically, a ratio between two and five measured values for every descriptor should be sought to prevent over fitting. Second, we are planning on obtaining a parametric representation of the regression, producing a simple equation, which can be validated against our scientific knowledge [32].

The GFA algorithm works with a set of strings, called a population [32]. This population is evolved in a manner that leads it toward the objective of the search [32]. Following this, three operations are performed iteratively in succession: selection, crossover, and mutation. Newly added members are scored according to a fitness criterion. In the GFA, the scoring criteria for models are all related to the quality of the regression fit to the data. The selection probabilities must be re-evaluated each time a new member is added to the population [32]. The procedure continues for a user-specified number of generations, unless convergence occurs in the interim. Convergence is triggered by lack of progress in the highest and average scores of the population [32].

Various statistical measures can be adapted to measure the fitness of a GFA model during the evolution process. Use of the Friedman LOF measure has several advantages over the regular least square error measure. In Materials Studio [45, 46], LOF is measured using a slight variation of the original Friedman formula [47]. The revised formula is:

$$LOF = \frac{SSE}{(1 - \frac{c + dp}{M})^2} \tag{9}$$

Where SSE is the sum of squares of errors, *c* is the number of terms in the model, other than the constant term, *d* is a user-defined smoothing parameter, *p* is the total number of descriptors contained in all model terms (again ignoring the constant term) and *M* is the number of samples in the training set [47].

Table 6. Validation table of the GFA

	Equation 1
Friedman LOF	103.2001
R-squared	0.996943
Adjusted R-squared	0.995656
Cross validated R-squared	-20.8175
Significant Regression	Yes
Significance-of-regression F-value	774.532
Critical SOR F-value (95%)	2.48E+00
Replicate points	0
Computed experimental error	0
LOF points	19
Min expt. error for non-significant LOF (95%)	3.279389

Unlike the commonly used least squares measure, the LOF measure cannot always be reduced by adding more terms to the regression model. While the new term may reduce the SSE, it also increases the values of c and p , which tends to increase the LOF score. Therefore, adding a new term may reduce the SSE, but it actually increases the LOF score. By limiting the tendency to simply add more terms, the LOF measure resists over fitting better than the SSE measure [32, 48].

Table 6 shows the GFA analysis, which gives summary of the input parameters used for the calculation. Also, it reports whether the GFA algorithm converged in a specified number of generations. The GFA algorithm is assumed to have converged when no improvement is seen in the score of the population over a significant length of time, either that of the best model in each population or the average of all the models in each population. When this criterion has been satisfied, no further generations are calculated [44].

Table 7. Equation used to calculate the predicted inhibition efficiency

Equation	Definitions
$Y = -17067.664384301 * X1$	X1: C : E(HOMO) (Ha)
$+ 9194.528783443 * X2$	X2: D : E (LUMO)(Ha)
$+ 0.416465461 * X4$	X4: F : Binding Energy (Kcal/mol)
$+ 0.997912242 * X12$	X12: N : Molecular volume (vdW volume) (Spatial Descriptors)
$+ 49922.123879648 * X14$	X14: (C : E(HOMO) (Ha)) * (E : [E(LUMO)-E(LUMO)](Ha))
$- 3433.185745850 * X24$	X24: (D : E (LUMO)(Ha)) * (E : [E(LUMO)-E(LUMO)](Ha))
$+ 4.385840138 * X25$	X25: (D : E (LUMO)(Ha)) * (F : Binding Energy (Kcal/mol))
$- 0.003332091 * X78$	X78: (M : Molecular area (vdW area) (Spatial Descriptors)) * (N : Molecular volume (vdW volume) (Spatial Descriptors))
-1507.66	

The Friedman’s LOF score in Table 6 evaluates the QSAR model [44]. The lower the LOF, the less likely it is that the GFA model will fit the data. The significant regression is given by F-test, and the higher the value, the better the model.

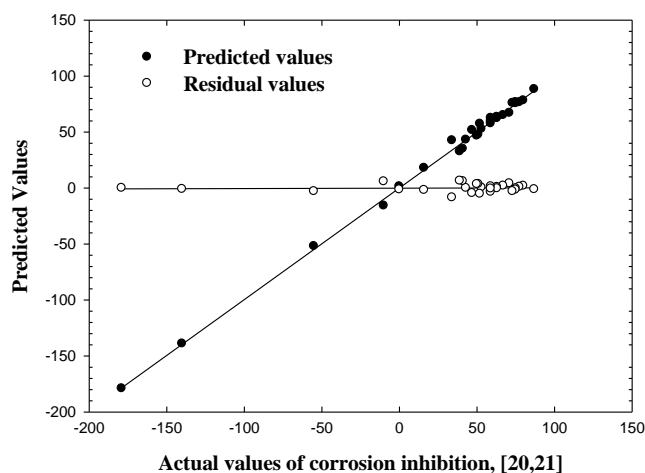


Figure 6. Plot of predicted inhibition and residuals vs. measured corrosion inhibition [20, 21] using GFA.

Figure 6 shows the relationship between the measured corrosion inhibition efficiencies of the studied inhibitors presented in Table 4 and the predicted efficiencies calculated by the equation model presented in Table 7.

The distribution of the residual values against the measured corrosion inhibition efficiencies values are presented in Fig. 6. The residual values can be defined as the difference between the predicted value generated by the model and the measured values of corrosion inhibition efficiencies. An analysis of Fig. 6 shows excellent correlation behavior, with most of the molecular system and showing acceptable deviations. The key feature of Fig. 6 is the distribution of the residual values against the measured corrosion inhibition values. An acceptable variation is observed, which should be present in a valid model. Inspection of Table 6 and Fig. 6 shows that the suggested model gives excellent correlation between the measured and predicted corrosion inhibition values. It is important to point out that the identification of related inhibitors showing very good behavior and this behavior has not been reported previously although the relatively big number of molecules employed in this study.

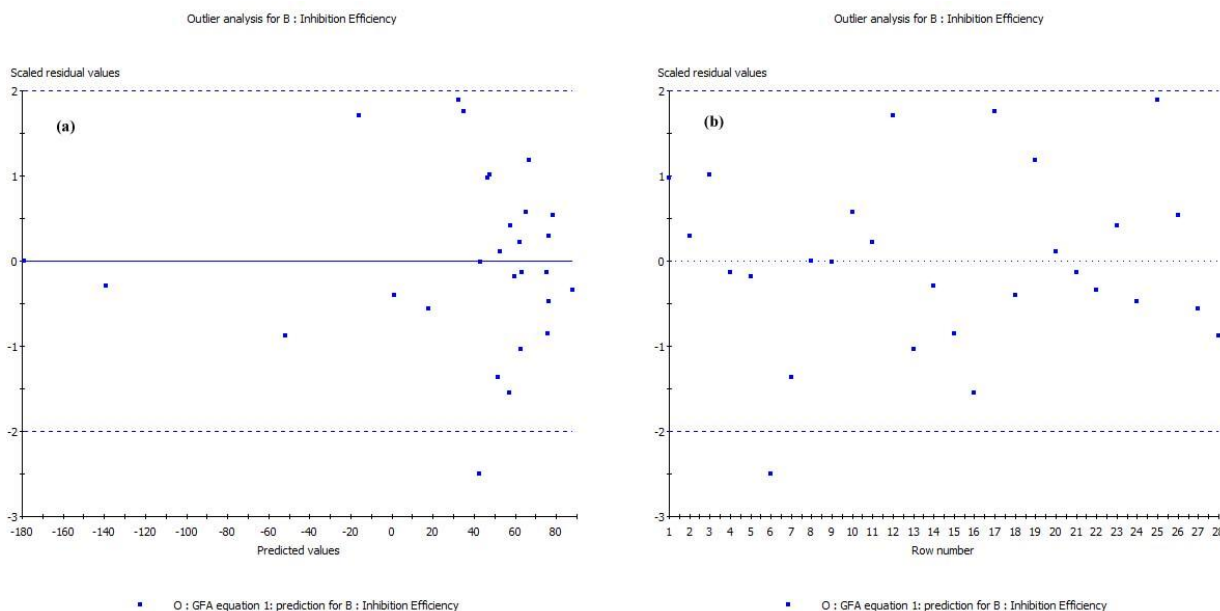


Figure 7. Outlier analysis for inhibition efficiency.

When a prediction model is generated to predict response data from predictor data, the prediction model will not normally give an exact fit to the response data. Unless the response data is genuinely an exact linear function of the predictor data, this should not be the case and an exact fit is indicative of over fitting (where there are as many independent observations as there are degrees of freedom in the algorithm from which the model is generated) [48]. With data that are randomly distributed within normal distribution, when a linear prediction model is generated using a least squares analysis technique, the residual values should also have a normal distribution with a mean value of zero. It is then expected that 95% of the values should lie within two standard deviations of the mean value [48]. Figure 7a and 7b represents the potential outlier that was used to test the

constructed QSAR model. An outlier can be defined as a data point whose residual value is not within two standard deviations of the mean of the residual values. Figure 7a represents the residual values plotted against the measured corrosion inhibition efficiencies. Figure 7b shows the residual values plotted against Table 4 row number. Figures 7a and 7b contains a dotted line that indicates the critical threshold of two standard deviations beyond which a value may be considered to an outlier. Inspection of Figs. 7a and 7b shows that there is only one point that appeared outside the dotted lines, which make the QSAR model acceptable.

5. CONCLUSIONS

Computational studies help to find the most stable inhibitor conformation and adsorption sites for a broad range of materials. This information can help to gain further insight into the corrosion system, such as the most likely point of attack for corrosion on a surface, the most stable site for inhibitor adsorption, adsorption density of the inhibitor and the binding energy of the adsorbed layer.

A GFA method was used to run the regression analysis and establish correlations between different types of descriptors and measured corrosion inhibition efficiency of 28 amino acids and their related compounds. A QSAR equation was developed and used to predict the corrosion inhibition efficiency of 28 amino acids and their related compounds. The prediction of corrosion efficiencies of these compounds nicely matched the experimental measurements. The studied amino acids inhibit iron corrosion by forming a molecular layer and decreased the iron dissolution. It is clear from these complex processes that occur on the metal surface that the inhibition process cannot fully be explained by using the QSAR approach. Despite these limitations, the QSAR approach is still an effective method that can be coupled with experimental measurements to predict inhibitor candidates for corrosion process.

ACKNOWLEDGMENTS

The authors are grateful for the financial support provided by Saudi Aramco, contract #6600027957.

References

1. K. Khaled, *Electrochim. Acta*, 53 (2008) 3484-3492.
2. F.B. Growcock, *Corrosion*, 45 (1989) 1003-1007.
3. F.B. Growcock, W.W. Frenier, P.A. Andreozzi, *Corrosion*, 45 (1989) 1007-1015.
4. P.G. Abdul-Ahad, S.H.F. Al-Madfai, *Corrosion*, 45 (1989) 978-980.
5. S.G. Zhang, W. Lei, M.Z. Xia, F.Y. Wang, *Journal of Molecular Structure: THEOCHEM*, 732 (2005) 173-182.
6. K.F. Khaled, K. Babić-Samardžija, N. Hackerman, *Electrochim. Acta*, 50 (2005) 2515-2520.
7. K.F. Khaled, *Electrochim. Acta*, 53 (2008) 3484-3492.
8. P. Dupin, D.A. Vilovia-Vera, A. de Savignac, A. Latta, P. Haicour, Proc. Proceedings of Fifth European Symposium on Corrosion Inhibitors, 1980.
9. W.W. Frenier, F.B. Growcock, V.R. Lopp, *SPE Production Engineering*, 3 (1988) 584-590.
10. I. Lukovits, A. Shaban, E. Kalman, *Electrochim. Acta*, 50 (2005) 4128-4133.
11. I. Lukovits, E. Kalman, F. Zucchi, *Corrosion*, 57 (2001) 3-3.

12. I. Lukovits, I. Bakó, A. Shaban, E. Kálmán, *Electrochim. Acta*, 43 (1998) 131-136.
13. I. Lukovits, K. Pálfi, I. Bakó, E. Kálmán, *Corrosion (Houston)*, 53 (1997) 915-919.
14. J.M. Costa, J.M. Lluch, *Corros. Sci.*, 24 (1984) 929-933.
15. V. Hluchan, L. Wheeler, N. Hackerman, *Werkstoffe und Korrosion*, 39 (1988) 512-517.
16. K. Babić-Samardžija, C. Lupu, N. Hackerman, A.R. Barron, A. Luttge, *Langmuir*, 21 (2003) 12187-12196.
17. K. Khaled, S.A. Fadel-Allah, B. Hammouti, *Mater. Chem. Phys.*, 117 (2009) 148-155.
18. J. Andrés, J. Beltran, in: *Química Teórica y Computacional*, Universitat Jaume Química Teórica y Computacional, Universitat Jaume, Castellón de la Plana, Espana, 2000.
19. K.F. Khaled, *J. Solid State Electrochem.*, 13 (2009) 1743-1756.
20. K.F. Khaled, N.S. Abdel-Shafi, *Int. J. Electrochem. Sci.*, 6 (2011) 4077-4094.
21. O. Ermer, Calculation of molecular properties using force fields. Applications in organic chemistry, in: *Bonding forces*, vol. 27, Springer Berlin Heidelberg, 1976, pp. 161-211.
22. K. Khaled, N. Abdel-Shafi, *Int. J. Electrochem. Sci.*, 6 (2011) 4077-4094.
23. Accelrys to Release Enhanced Suite of Chemicals and Materials Modeling and Simulation Tools with Materials Studio(R) 4.1, in: *Business Wire*, New York, United States, New York, 2006, pp. 0-n/a.
24. J. Barriga, B. Coto, B. Fernandez, *Tribology International*, 40 (2007) 960-966.
25. K. Khaled, N. Al-Mobarak, *Int. J. Electrochem. Sci.*, 7 (2012) 1045-1059.
26. K. Khaled, N. Abdel-Shafi, N. Al-Mobarak, *Int. J. Electrochem. Sci.*, 7 (2012) 1027-1044.
27. K.F. Khaled, N.A. Al-Mobarak, *Int. J. Electrochem. Sci.*, 7 (2012) 1045-1059.
28. K.F. Khaled, N.S. Abdelshafi, A.A. Elmaghraby, A. Aouniti, N.A. Almobarak, B. Hammouti, *Int. J. Electrochem. Sci.*, 7 (2012) in press.
29. K.F. Khaled, *J. Appl. Electrochem.*, 41 (2011) 423-433.
30. K.F. Khaled, *J. Electrochem. Soc.*, 157 (2010) C116-C124.
31. K. Khaled, *Appl. Surf. Sci.*, 255 (2008) 1811-1818.
32. Accelrys Materials Studio 6.0 Manual, (2011).
33. Accelrys European Science Symposium Highlights Advanced Data Analysis and Predictive Science, in: *PR Newswire*, New York, United States, New York, 2012.
34. Z. Shi, A. Atrens, *Corros. Sci.*, 53 (2011) 226-246.
35. C. Guedes Soares, Y. Garbatov, A. Zayed, G. Wang, *Corros. Sci.*, 50 (2008) 3095-3106.
36. H. Tamura, *Corros. Sci.*, 50 (2008) 1872-1883.
37. Y. Tan, *Corros. Sci.*, 53 (2011) 1845-1864.
38. M.T. Gudze, R.E. Melchers, *Corros. Sci.*, 50 (2008) 3296-3307.
39. M.V. Putz, A.-M. Putz, M. Lazea, L. Ienciu, A. Chiriac *International journal of molecular science*, 10 (2009) 1193-1214.
40. J.H. Steiger, P.H. Schonemann, A history of factor indeterminacy. In *Theory Construction and Data Analysis in the Behavioural Science*, San Francisco, CA, USA, 1978., 1978.
41. C. Spearman, *The Abilities of Man*, MacMillan, London, UK, , 1927.
42. J.G. Topliss, R.J. Costello, *J. Med. Chem.*, 15 (1972) 1066-1068.
43. J.G. Topliss, R.P. Edwards, *J. Med. Chem.*, 22 (1979) 1238-1244.
44. K.F. Khaled, *Corros. Sci.*, 53 (2011) 3457-3465.
45. B. Delley, *J. Chem. Phys.*, 92 (1990) 508.
46. B. Delley, *J. Chem. Phys.*, 113 (2000) 7756.
47. J.H. Friedman, Multivariate Adaptive Regression Splines, in: *Technical Report No. 102*, Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford 1988.
48. Materials Studio 6.0 Manual, Accelrys, (2009).