Data-driven corrosion inhibition efficiency prediction model incorporating 2D–3D molecular graphs and inhibitor concentration

Ma, Jinbo; Dai, Jiaxin; Guo, Xin; Fu, Dongmei; Ma, Lingwei; Keil, Patrick; Mol, Arjan; Zhang, Dawei

DOI
10.1016/j.corsci.2023.111420

Publication date
2023

Document Version
Final published version

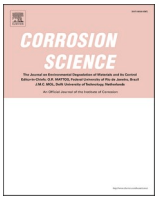
Published in
Corrosion Science

Corrosion Science 222 (2023) 111420

Contents lists available at ScienceDirect

Corrosion Science

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



Data-driven corrosion inhibition efficiency prediction model incorporating 2D–3D molecular graphs and inhibitor concentration

Jinbo Ma [a,c], Jiaxin Dai [a,c], Xin Guo [b,c], Dongmei Fu [a,c,*], Lingwei Ma [b,c,d,**], Patrick Keil [e], Arjan Mol [f], Dawei Zhang [b,c,d,**]

[a] School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China
[b] Beijing Advanced Innovation Center for Materials Genome Engineering, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China
[c] National Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, Beijing 100083, China
[d] Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, China
[e] BASF Coatings GmbH, Münster 48165, Germany
[f] Department of Materials Science and Engineering, Delft University of Technology, Mekelweg 2, Delft 2628CD, the Netherlands

ARTICLE INFO

Keywords:
Corrosion inhibition
Machine learning
Molecular graph
Corrosion prediction

ABSTRACT

Following the construction of a dataset of cross-category corrosion inhibitors at different concentrations based on 1241 data from 184 research papers, a performance prediction model incorporating 2D–3D molecular graph representation and corrosion inhibitor concentration information was established. This model was shown to effectively predict the inhibition efficiency (IE) of different categories of corrosion inhibitors for carbon steel in 1 mol/L HCl solution. The model was also able to predict IEs of corrosion inhibitors at different concentrations. The results demonstrated that 3D features of corrosion inhibitors, especially those of large molecules, had a significant impact on the prediction precision of IEs.

1. Introduction

Corrosion is one of the main causes of damages to metallic materials and structures, causing huge economic losses worldwide [1]. Corrosion inhibitors are an effective method to suppress metal corrosion, with advantages such as low cost, simplicity, and high efficiency. The inhibition efficiency (IE) is an index to evaluate the effectiveness of corrosion inhibitors, which is closely related to molecular structures and concentrations, as well as metal substrates and corrosive environments [2,3]. Traditional methods for experimental assessment of IE, such as weight loss measurements [4], electrochemical tests [5] or spectroscopic analyses [6] can only determine IE of corrosion inhibitors at specific concentrations one by one. To select high-performance corrosion inhibitors and their reasonable concentrations from the entire chemical space, a large number of experimental tests are needed. Therefore, developing a fast and accurate computational method to evaluate IE provides important support for material scientists to screen potential highly efficient corrosion inhibitors.

Computational chemistry and machine learning (ML) have been successfully employed in the study of corrosion inhibitors [7,8]. Computational chemistry methods include density functional theory (DFT) [9,10] and molecular dynamics (MD) simulations [11]. DFT has been frequently used to predict the performance of organic corrosion inhibitors based on electronic/molecular properties and reactivity indices [12]. However, recent literatures have clearly demonstrated that the correlations between the DFT derived parameters and the IEs are misleading or are too weak to be quantitative for a large data set of corrosion inhibitors [13–15]. MD simulation provides useful information regarding the adsorption behavior of corrosion inhibitors on the metal-electrolyte interfaces to promote the development of effective corrosion inhibitors [16]. However, this method can only predict the performance of the same class of corrosion inhibitors one by one. In contrast, data-driven ML methods can utilize molecular structural parameters more efficiently and explore the chemical space more quickly. Artificial neural networks [17,18], unsupervised clustering [19], and other algorithms have been used for predicting IE of corrosion inhibitors

* Corresponding author at: School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China
** Corresponding authors at: Beijing Advanced Innovation Center for Materials Genome Engineering, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China.
E-mail addresses: fdm2003@163.com (D. Fu), mlw1215@ustb.edu.cn (L. Ma), dzhang@ustb.edu.cn (D. Zhang).

https://doi.org/10.1016/j.corsci.2023.111420
Received 15 June 2023; Received in revised form 18 July 2023; Accepted 22 July 2023
Available online 25 July 2023

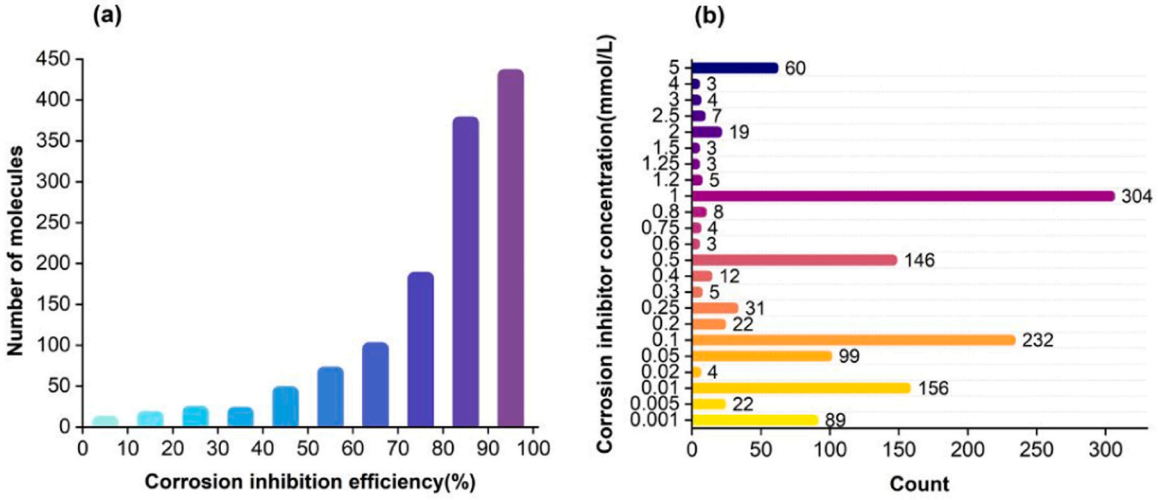


Fig. 1. The statistics of the CoInDataset 1: (a) distribution of IE values, (b) distribution of corrosion inhibitor concentration.

for various metals. By extracting quantitative molecular descriptors and other structured data, a prediction model can be established, which can predict IE in minutes or even shorter time. However, the models obtained by the above algorithms are largely dependent on the selection of molecular structure features, and existing research is limited to predict IEs of a certain class of corrosion inhibitors at specific concentrations.

The molecular structure of corrosion inhibitors is the main factor affecting their inhibition mechanism and efficiency [20]. Thus, accurate and comprehensive representation of their molecular structures is the key to establishing IE prediction models. With the rapid development of data-driven ML in various fields, it has been successfully applied to molecular representation learning (MRL). MRL encodes molecules as numerical vectors that retain molecular structure and characteristics as feature vectors for downstream tasks such as molecular property prediction. For example, molecules can be treated as 2D molecular graphs, with atoms as nodes and chemical bonds between atoms as edges [21]. As such, the topological structure and node attribute information of 2D molecular graphs can be directly processed through graph convolutional neural networks (GCN) [22], message passing neural networks (MPNN) [23,24], directed message passing neural networks (D-MPNN) [25], and other graph neural network algorithms to achieve molecular representation learning. In a previous work, we have developed a three-layer directed message passing networks (3 L-DMPNN) [26] model involving atomic, bond and molecular features to predict the IEs of compounds on carbon steel in a specific environment. However, for some corrosion inhibitors with large molecule weight, their spatial configuration is complex, making it difficult to fully characterize the structural features of the molecules with only 2D graphs. Some scholars [27–29] have improved the prediction accuracy of the geometric, energy, electronic, and thermodynamic properties of molecules by mining the 3D structural features of molecules. Furthermore, the 2D and 3D molecular graphs have been combined, such as in the case of the GeomGCL model [30], which showed improved accuracy of downstream prediction tasks of molecular properties including hydrophobicity, toxicity, octanol/water partition coefficient, and hydration free energy. However, GeomGCL only generates rough 3D molecular structures by using the RDKit [31] software package, limiting the ability to extract the fine 3D features of corrosion inhibitors. In addition, all the above models can only predict the IEs at specific inhibitor concentrations, and the generalization performance is poor.

In the present study, a new prediction model for the IEs was constructed by incorporating 2D–3D molecular graphs and corrosion inhibitor concentration (2D3DMol-CIC). The data used in this study were extracted from 184 publications, including 1241 IE values for 414 corrosion inhibitors at concentrations ranging from 0.001 to 5 mmol/L. The accuracy of the proposed model was compared with those of the support vector machine (SVM) [32], random forest (RF) [33], and 3 L-DMPNN[21] models, and the effect of 3D features on the prediction of IE and the generalization ability of the model were verified. Based on this model, the study also provided a selection of corrosion inhibitor concentration for practical application.

2. Methods

2.1. Corrosion inhibitor datasets

Existing literature has reported a large amount of data on corrosion inhibitor performance, but typically each paper only reports IE values of one or several corrosion inhibitors at different concentrations. In order to study the 2D–3D molecular structure features of molecules and the effect of corrosion inhibitor concentration on IEs, we used the dataset containing 116 papers reported in the work by Dai et al.[26], and retrieved additional 68 papers that studied the influences of different corrosion inhibitor concentrations on carbon steel in 1 mol/L

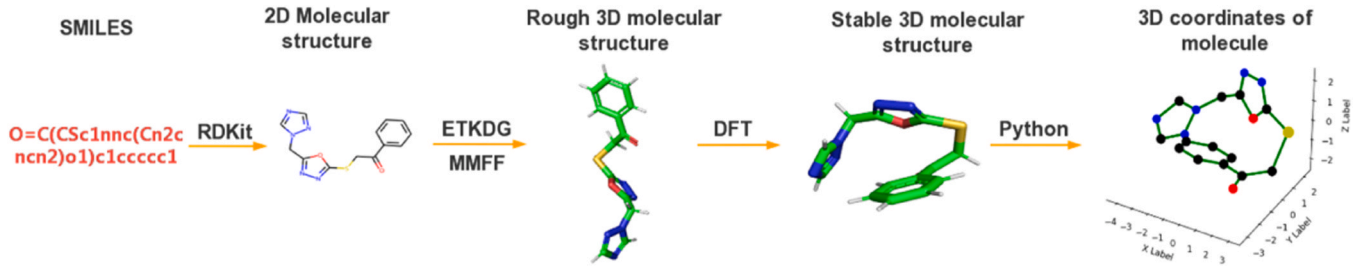


Fig. 2. The process of extracting the 3D coordinates of a molecule.

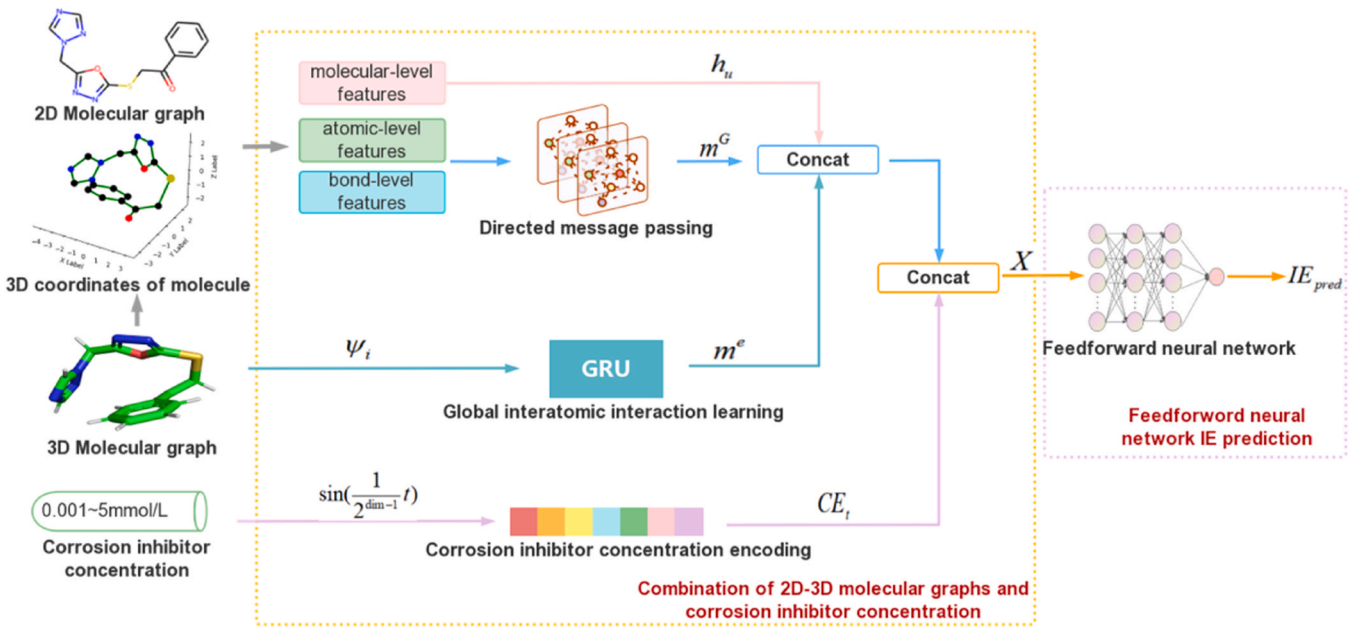


Fig. 3. Neural network architecture of the 2D3DMol-CIC.

hydrochloric acid (HCl) solution at room temperature published from February 2005 to December 2021. From the above 184 papers, we collected 1241 IE values for 414 corrosion inhibitors applied to carbon steel surfaces with the inhibitor concentrations ranging from 0.001 to 5 mmol/L. This dataset, named as CoInDataset 1, includes information such as the name of corrosion inhibitor, the SimplifiedMolecular Input Line Entry System (SMILES) [34] of the molecule, the atomic coordinates in the molecule, the corrosion inhibitor concentration, and IE value. The dataset used in this study can be accessed via the following URL: https://www.corrdata.org.cn/inhibitor/.

Fig. 1a shows the distribution of IEs in CoInDataset 1, the number of inhibitors with low IEs is much lower than that of the ones with high IEs. The IE values in CoInDataset 1 range from 0.98 % to 99.3 % and the uneven data distribution may make learning and prediction of the model complex and challenging [35]. Fig. 1b shows the distribution of corrosion inhibitor concentrations in CoInDataset 1, covering a concentration range of 0.001–5 mmol/L with 23 different concentrations, reflecting a good diversity.

In order to more effectively express the 3D structural features of molecules, we adopted the process shown in Fig. 2 to extract the 3D coordinates of corrosion inhibitors: (1) an open-source software package RDKit is used to convert the SMILES of corrosion inhibitors into 2D structures of the molecules; (2) A rough 3D molecular structure was generated based on the Experimental-Torsion Distance Geometry (ETKDG) [36] algorithm and the Merck Molecular Force Field [37]; (3) A stable 3D molecular structure was optimized based on the DFT method with the B3LYP functional and the 6–31 G* * basis set; (4) The 3D coordinates of the molecules were extracted using Python for subsequent modeling and analysis, providing data such as bond lengths and bond angles. The code are available on GitHub at https://github.com/jinbo0906/2D3DMol-CIC/blob/main/three-dimensional_dataset_production.ipynb.

To verify the generalization ability of the 2D3DMol-CIC model, an independent validation dataset (CoInDataset 2) is constructed including 12 IE values obtained from laboratory experiments and 115 IE values retrieved from 35 papers published from January 2022 to January 2023. The IE values in CoInDataset 2 range from 33.0 % to 99.6 %, and the concentrations of the corrosion inhibitors range from 0.001 to 5 mmol/L. CoInDataset 2 can be accessed via the following URL: https://www.corrdata.org.cn/inhibitor/.

2.2. Models

The 2D3DMol-CIC model proposed in this study is implemented using the open-source software package Chemprop [38], and its overall network structure is shown in Fig. 3. The model can be divided into two stages: combination of 2D–3D molecular graphs and corrosion inhibitor concentration and feedforward neural network IE prediction.

2.2.1. Combination of 2D–3D molecular graphs and corrosion inhibitor concentration

This module mainly includes four parts: representation of 2D–3D molecular features, directed message passing, global atomic interaction learning and corrosion inhibitor concentration encoding.

a) Representation of 2D–3D molecular features. The model takes 2D molecular graphs $G^{2D} = (V, E)$ and 3D molecular graphs $G^{3D} = (Z, R)$ as inputs, where $v \in V$ is the set of nodes; $e \in E$ is the set of edges; $z \in Z$ is the set of atomic number; $r \in R$ represents the set of 3D coordinates of atoms; $d_{vw} = ||r_v - r_w||_2$ is the distance between two atoms and $\alpha(kv, vw) = \angle z_k z_v z_w$ is the bond angle. Atomic-level features x_v correspond to atomic properties. Bond-level features e_{vw} represent bond properties. Molecular-level features h_u represent global molecular properties. They are calculated using RDKit and encoded as numerical vectors. For specific feature descriptions, please refer to the Supporting Information.

b) Directed message passing. Firstly, the hidden state of edges is initialized $h^0_{vw} = \tau(W_i cat(x_v, e_{vw}))$, where τ is the ReLU [39] function, W_i is a learnable matrix, and *Concat* is a connection function.Then, in each directed message passing step t, the features of adjacent bonds are updated by adding them to their bond angle features according to Eq. (1):

$$h^{t+1}_{vw} = \tau\left(h^0_{kv} + W_m\left(\sum_{k\in\{N(v)/w\}} h^t_{kv} + \sum_{k\in\{N(v)/w\}} a_{(kw,wv)}\right)\right) \quad (1)$$

Where $N(v)$ represents the neighboring nodes of node v in the graph G^{2D}; $t \in \{1, \ldots, T\}$ and T is the total number of message passing steps; and W_m is a learnable matrix. Lastly, by aggregating all the atomic representations that have received the passed-in bond features, the molecular topology representation is obtained through Eq. (2).

$$m^G = \sum_{v \in G^{2D}} \sum_{w \in N(v)} h^l_{vw} \tag{2}$$

c) Global atomic interaction learning. Firstly, the distance relationship $\psi_i = \sum_{j \in \{U/i\}} RBF\left(\frac{1}{M^2_{dij}}\right) W_d$ between atoms is encoded based on their 3D coordinates, where M_{dij} is the distance matrix between atoms; U is the node set of graph G^{3D}; W_d is a learnable matrix. Through gated recurrent unit (GRU), the interaction between each atom and other atoms in graph G^{3D} is simulated $m^e_i = \sum_{j \in (U/i)} GRU(\psi_i, \psi_j)$, and all atoms are aggregated to obtain the global representation vector for atomic interactions $m^e = \sum_{i \in U} m^e_i$.

d) Corrosion inhibitor concentration encoding. Inspired by the position encoding mechanism in Transformer [40], each concentration value is represented by a series of sine functions $CE_t = \left[\sin\left(\frac{1}{2^0}t\right), \sin\left(\frac{1}{2^1}t\right), \ldots, \sin\left(\frac{1}{2^{dim-1}}t\right)\right]$, where *dim* is the dimension of the encoded vector.

After directed message passing and global atomic interaction encoding, the molecular-level features h_u and inhibitor concentration vector CE_t are fused to obtain the representation vector of the entire molecule $X = Concat(m^G, m^e, h_u, CE_t)$.

2.2.2. Feedforward neural network IE prediction

With the molecular representation vector X obtained by incorporating 2D–3D molecular graphs and corrosion inhibitor concentration as input and IE as output, IE prediction model is established based on four feedforward neural network layers, as shown in Eq. (3).

$$IE_{pred} = f(X) \tag{3}$$

This model did not take consideration of the deviation of the IE values, which was usually very low in the referred literature.

2.2.3. Evaluation metrics for models

The model uses root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) as performance metrics which are defined in Eqs. (4–6). From a mathematical perspective, the best performing model will have the lowest RMSE and MAE values, with the highest R^2.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(IE^i_{pred} - IE^i_{\exp}\right)^2} \tag{4}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|IE^i_{pred} - IE^i_{\exp}\right| \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(IE^i_{pred} - IE^i_{\exp}\right)^2}{\sum_{i=1}^{m}\left(IE^i_{pred} - \overline{IE}\right)^2} \tag{6}$$

where IE^i_{pred} is the predicted IE value for sample *i*, $IE^i_{\exp}$ is the experimental *IE* value for sample *i*,m is the total number of samples, and $\overline{IE} = \frac{1}{m}\sum_{i=1}^{m} IE^i_{pred}$ is the average of the predicted IE values.

3. Results and discussion

3.1. Evaluation of the accuracy of the model

The 2D3DMol-CIC model using ReLU as the activation function is trained on CoInDataset 1, and evaluated based on 10-fold cross-validation with 1241 molecules being randomly divided into 10 subsets. In this data-driven model, nine subsets were used for training while the remaining one was used for testing, and this process was repeated 10 times until each subset was used as testing data once. The average of the test results from 10 runs was taken as the final evaluation metric.

Table 1
Comparison of prediction results of different models on the CoInDataset 1 (10-fold cross-validation).

Model	RMSE	MAE	R^2
SVM	0.115233 + / − 0.006346	0.091222 + /- 0.004662	0.558671 + /- 0.087741
RF	0.100045 + /- 0.011266	0.069236 + /- 0.005220	0.668630 + /- 0.071529
3 L-DMPNN	0.085584 + /- 0.008397	0.058129 + /- 0.005442	0.724137 + /- 0.043756
2D3DMol-CIC	**0.076810 + /- 0.008541**	**0.047714 + /- 0.003650**	**0.803525 + /- 0.049372**

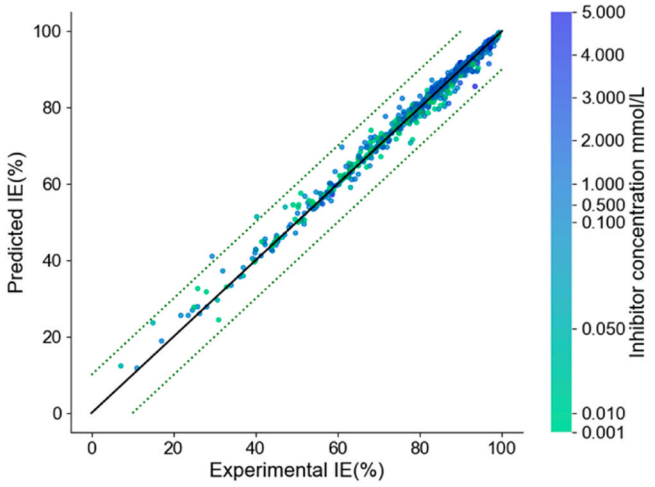


Fig. 4. Comparison of the prediction efficiency of the 2D3DMol-CIC model with experimental IEs on the CoInDataset 1.

The hyperparameters for the 2D3DMol-CIC model are set to depth = 4, hidden_size = 400, layer_num = 4, and dropout = 0.0. The performance of the model was compared with SVM, RF, and 3 L-DMPNN models, as shown in Table 1. Fig. 4 shows IE values predicted by the 2D3DMol-CIC model and the corresponding experimental values, in which, the labeled points between the green lines represent inhibitors with a prediction error within 10%, accounting for 99.8 % of all inhibitors. These results suggest that the 2D3DMol-CIC model can accurately predict IEs of different molecular compounds, and performs better than SVM and RF models established using only structural data (such as electronegativity, polarizability, van der Waals volume, etc.), and 3 L-DMPNN model using only 2D molecular graph structure.

Based on the data of corrosion inhibitors collected in this article, the molecular weights are generally between 100 and 1000 g/mol. Fig. 5a shows the distribution of the molecular weights of the corrosion inhibitors in CoInDataset 1. Fig. 5b shows the comparison of the average prediction errors of IEs by 3 L-DMPNN and 2D3DMol-CIC models for corrosion inhibitors with molecular weights in different ranges. The blue line represents the difference between the average prediction error of the 2D3DMol-CIC and that of the 3 L-DMPNN in each range of the molecular weights. Interestingly, it can be seen that as the molecular weight increases, the prediction error of the 2D3DMol-CIC model is generally lower and the difference between the prediction errors of the two models is significantly increased, demonstrating an advantage of 2D3DMol-CIC model over the 3 L-DMPPN model in predicting high-molecular-weight corrosion inhibitors. The higher the molecular weight, the more complex the molecular structure. Therefore, it is of great importance to provide the 3D spatial structural information of corrosion inhibitor molecules during IE prediction.

Fig. 6a shows the distribution proportions of four models with

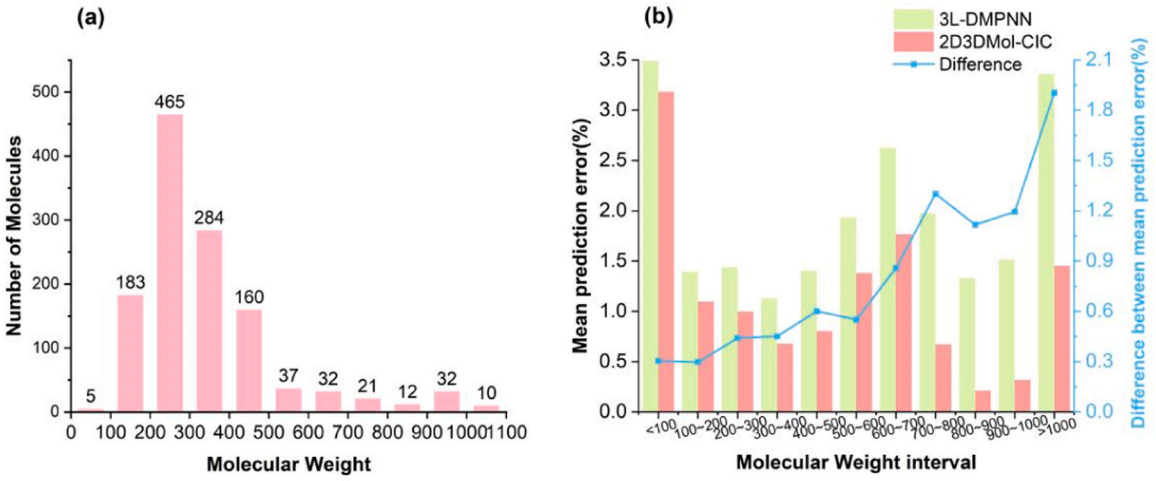


Fig. 5. CoInDataset 1. (a) Distribution of molecular weight, (b) comparison of the average prediction errors of IE by 3 L-DMPNN and 2D3DMol-CIC models for corrosion inhibitors with molecular weights in different ranges.

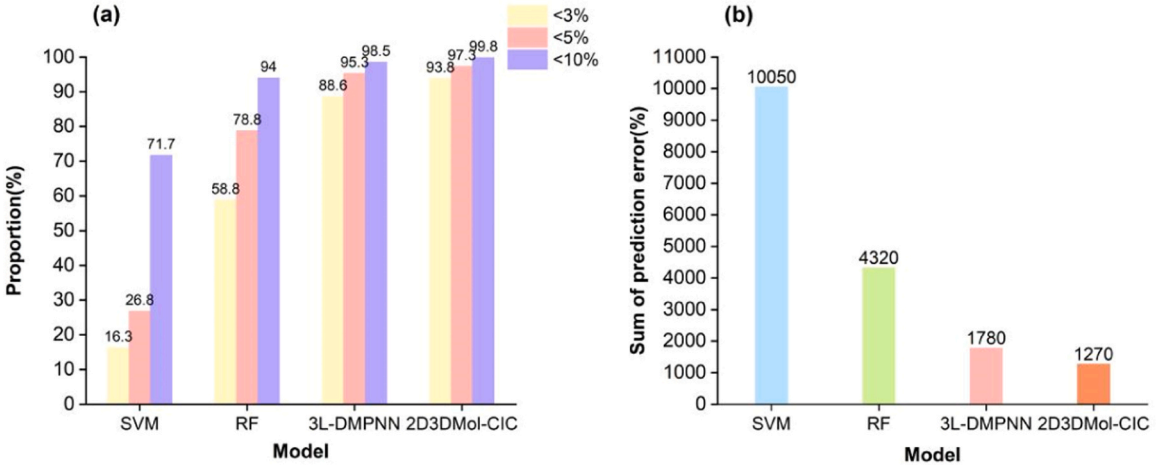


Fig. 6. SVM, RF, 3 L-DMPNN, and 2D3DMol-CIC models (a) proportions of prediction errors less than 10 %, 5 %, and 3 %, (b) the total prediction errors.

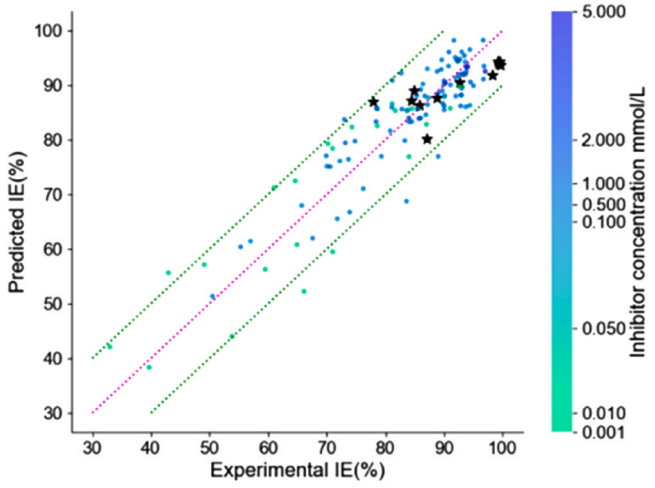


Fig. 7. Comparison of the prediction efficiency of the 2D3DMol-CIC model with experimental efficiency on the CoInDataset 2.

prediction errors less than 10%, 5%, and 3%, the results of 2D3DMol-CIC model are much better than those of other models. Fig. 6b shows the sum of predicted errors of four models. The 2D3DMol-CIC model has the lowest prediction error sum. Therefore, this model has the highest prediction accuracy and effectiveness, and is the most suitable for predicting IE values.

3.2. *Generalization testing of the model*

The 2D3DMol-CIC model trained in Section 3.1 is tested using CoInDataset 2, and the results are shown in Fig. 7. The annotation points between the green lines represent molecules with a prediction error within 10 %, accounting for 94.2 % of all inhibitors, indicating that the model can accurately predict IE of compounds outside the training data. In addition, the corrosion inhibitor concentration values marked by black stars are 0.0015 mmol/L, 0.03 mmol/L, 0.09 mmol/L, 0.12 mmol/L, 0.14 mmol/L, 0.18 mmol/L, 0.27 mmol/L, 0.29 mmol/L, 0.48 mmol/L, 1.4 mmol/L, and 1.5 mmol/L, which are not included in CoInDataset 1, but still demonstrate accurate prediction results. In conclusion, the model has good generalization ability and can make a reliable prediction for corrosion inhibitors outside the training data domain.

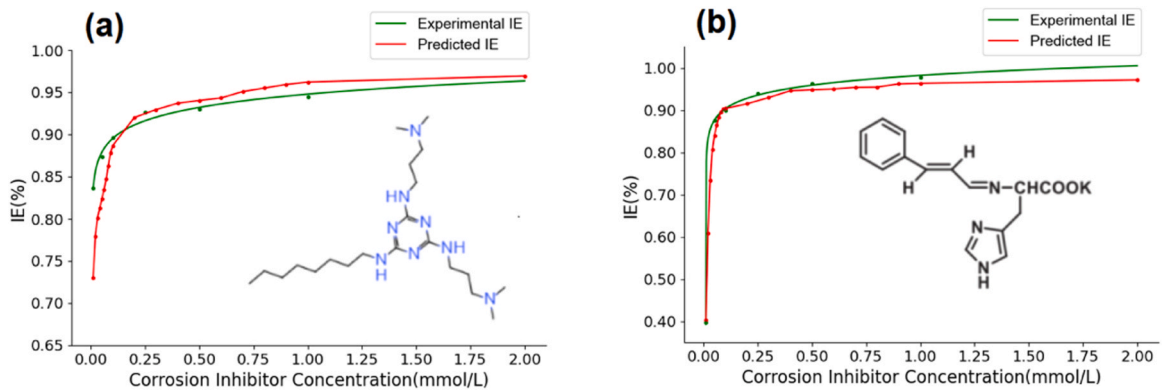


Fig. 8. Corrosion inhibition efficiency-concentration curves (a) 2-(n-Octylamino)− 4,6-bis(3-N,N-dimethylaminopropyl)amino-1,3,5-triazine, (b) Potassium 3-(1 H-imidazol-4-yl)− 2-(((E)− 3-phenylallylidene)amino)propanoate.

3.3. Recommendation of corrosion inhibitor concentration

IE is generally positively correlated with the concentration of the corrosion inhibitor, but IE almost does not change significantly with the increase of concentration after reaching a certain value [41]. This phenomenon is generally attributed to the saturation adsorption of corrosion inhibitors on metal surfaces [42,43], the solubility limit of molecules [44,45] or the micelle formation above critical concentrations [46,47]. With consideration of effectiveness and economic aspects, the concentration of corrosion inhibitors should be controlled at specific values that not only ensures good corrosion inhibition effect but also avoids excessive addition. Therefore, providing a choice of corrosion inhibitor concentration while rapidly predicting the IE has important engineering significance.

To investigate the ability of 2D3DMol-CIC model to recommend the concentration of corrosion inhibitors, this article analyzes the relationship between concentration and IE for two molecules (2-(n-Octylamino)− 4,6-bis(3-N,N-dimethylaminopropyl)amino-1,3,5-triazine [48] and Potassium 3-(1 H-imidazol-4-yl)− 2-(((E)− 3-phenylallylidene)amino)propanoate) [49] with the maximum experimental data on IE under different concentrations in CoInDataset 2. Fig. 8a and Fig. 8b show the experimental curves of corrosion inhibition efficiency-concentration obtained by logarithmic interpolation fitting of the experimental data for the two molecules and the predicted curves of corrosion inhibition efficiency-concentration generated by the 2D3DMol-CIC model, respectively. Both curves show a consistent trend and similar results, indicating that the model can provide a relatively accurate relationship between the concentrations of corrosion inhibitors and corresponding IE values for materials scientists to choose the appropriate addition amount of inhibitions. Meanwhile, the solubility limit and the critical micelle concentrations of molecules should be considered before determining the optimal inhibitor concentration.

3.4. Significance and limitations of the model

The present work proposed a data-driven prediction model of the IEs for corrosion inhibitors based on 2D and 3D structure of molecules and considering the concentration of corrosion inhibitors. Compared with existing prediction models, the constructed model can not only more accurately predict IEs of cross-category corrosion inhibitors at specific concentrations but can also predict IEs of corrosion inhibitors at different concentrations. Using this model, the minimal concentration that yields to high IE (>90%) could be identified, which has significant engineering implications.

Currently, the model is limited to predicting IEs of compounds on carbon steel in 1 mol/L HCl at room temperature, and the training dataset for this model is relatively small. In future work, we plan to construct a larger dataset of corrosion inhibitors that includes different metals, temperatures, corrosive environments, and establish a more generic prediction model. Additionally, in this article, a stable 3D molecular structure is obtained through optimization using Gaussian software. However, this method is time-consuming, often taking minutes to hours to optimize a molecule. In the next step, we will consider using more efficient deep learning methods to generate stable molecular conformations, construct a 3D molecular dataset, and establish an end-to-end model from dataset construction to efficiency prediction.

4. Conclusions

This work reported the development of the 2D3DMol-CIC model for predicting IEs based on 2D–3D molecular graph features, under varying concentrations of corrosion inhibitors ranging from 0.005 to 5 mmol/L. The 10-fold cross-validation approach was utilized to determine the proportions of compounds with prediction error of the model less than 3 %, 5 % and 10% in CoInDataset1, the values of 2D3DMol-CIC were 93.8 %, 97.3 % and 99.8 %, respectively, which were better than SVM, RF and 3 L-DMPNN. In addition, 2D3DMol-CIC fully considers and extracts the 3D structural characteristics of molecules, which makes the model capable of accurately predicting the IEs especially of larger molecules. Additionally, the generalization capability of the developed model is verified with 127 independent testing datasets. We also investigated the ability of the model to recommend the concentration of corrosion inhibitors, and the model could predict the corrosion inhibition efficiency-concentration curves very close to that verified experimentally. The obtained results indicated that the 2D3DMol-CIC model can accurately predict the IEs, providing a low-cost and fast screening method for corrosion inhibitors and their concentrations.

CRediT authorship contribution statement

Jinbo Ma: Conceptualization, Methodology, Investigation, Analysis, Writing – original manuscript. **Jiaxin Dai:** Methodology, Analysis. **Xin Guo:** Investigation. **Dongmei Fu:** Supervision, Conceptualization, Methodology, Analysis, Writing – review & editing. **Lingwei Ma:** Methodology, Analysis, Writing – review & editing. **Patrick Keil:** Writing – review & editing. **Arjan Mol:** Writing – review & editing. **Dawei Zhang:** Supervision, Conceptualization, Methodology, Writing – review & editing. All authors contributed to the discussion of the results.

Declaration of Competing Interest

The authors declare no competing financial interest.

Data availability

All datasets used in this paper can be assessed at https://www.

corrdata.org.cn/inhibitor/.

Acknowledgments

This work is supported by National Key R&D Program of China (2022YFB3808803).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.corsci.2023.111420.

References

[1] B. Hou , X. Li , X. Ma , C. Du , D. Zhang , M. Zheng , W. Xu , D. Lu , F. Ma , The cost of corrosion in China, npj Materials Degradation, 1 2017. 4.
[2] B.R. Fazal, T. Becker, B. Kinsella, K. Lepkova, A review of plant extracts as green corrosion inhibitors for CO_2 corrosion of carbon steel, npj Mater. Degrad. 6 (2022) 5.
[3] D.S. Kharitonov, M. Zimowska, J. Ryl, A. Zieliński, M.A. Osipenko, J. Adamiec, A. Wrzesińska, P.M. Claesson, I.I. Kurilo, Aqueous molybdate provides effective corrosion inhibition of WE43 magnesium alloy in sodium chloride solutions, Corros. Sci. 190 (2021), 109664.
[4] F. Mansfeld, S. Tsai, Laboratory studies of atmospheric corrosion—I. Weight loss and electrochemical measurements, Corros. Sci. 20 (1980) 853–872.
[5] Y. Zou, J. Wang, Y. Zheng, Electrochemical techniques for determining corrosion rate of rusted steel in seawater, Corros. Sci. 53 (2011) 208–216.
[6] V. Deodeshmukh, A. Venugopal, D. Chandra, A. Yilmaz, J. Daemen, D. Jones, S. Lea, M. Engelhard, X-ray photoelectron spectroscopic analyses of corrosion products formed on rock bolt carbon steel in chloride media with bicarbonate and silicate ions, Corros. Sci. 46 (2004) 2629–2649.
[7] E.E. Oguzie, D.I. Njoku, M.A. Chidebere, C.E. Ogukwe, G.N. Onuoha, K.L. Oguzie, N. Ibisi, Characterization and experimental and computational assessment of Kola nitida extract for corrosion inhibiting efficacy, Ind. Eng. Chem. Res. 53 (2014) 5886–5894.
[8] D.M. Gurudatt, K.N. Mohana, Synthesis of new pyridine based 1, 3, 4-oxadiazole derivatives and their corrosion inhibition performance on mild steel in 0.5 M hydrochloric acid, Ind. Eng. Chem. Res. 53 (2014) 2092–2105.
[9] C.M. Goulart, A. Esteves-Souza, C.A. Martinez-Huitle, C.J.F. Rodrigues, M.A. M. Maciel, A. Echevarria, Experimental and theoretical evaluation of semicarbazones and thiosemicarbazones as organic corrosion inhibitors, Corros. Sci. 67 (2013) 281–291.
[10] Z. El Adnani, M. Mcharfi, M. Sfaira, M. Benzakour, A. Benjelloun, M.E. Touhami, DFT theoretical study of 7-R-3methylquinoxalin-2 (1H)-thiones (RH; CH3; Cl) as corrosion inhibitors in hydrochloric acid, Corros. Sci. 68 (2013) 223–230.
[11] I. Obot, Z. Gasem, Theoretical evaluation of corrosion inhibition performance of some pyrazine derivatives, Corros. Sci. 83 (2014) 359–366.
[12] I. Obot, D. Macdonald, Z. Gasem, Density functional theory (DFT) as a powerful tool for designing new organic corrosion inhibitors. Part 1: an overview, Corros. Sci. 99 (2015) 1–30.
[13] A. Kokalj, On the alleged importance of the molecular electron-donating ability and the HOMO–LUMO gap in corrosion inhibition studies, Corros. Sci. 180 (2021), 109016.
[14] A. Kokalj, M. Lozinšek, B. Kapun, P. Taheri, S. Neupane, P. Losada-Pérez, C. Xie, S. Stavber, D. Crespo, F.U. Renner, Simplistic correlations between molecular electronic properties and inhibition efficiencies: do they really exist? Corros. Sci. 179 (2021), 108856.
[15] A. Kokalj, Molecular modeling of organic corrosion inhibitors: calculations, pitfalls, and conceptualization of molecule–surface bonding, Corros. Sci. 193 (2021), 109650.
[16] C. Verma, H. Lgaz, D. Verma, E.E. Ebenso, I. Bahadur, M. Quraishi, Molecular dynamics and Monte Carlo simulations as powerful tools for study of interfacial adsorption behavior of corrosion inhibitors in aqueous phase: a review, J. Mol. Liq. 260 (2018) 99–120.
[17] T. Harvey, S. Hardin, A. Hughes, T. Muster, P. White, T. Markley, P. Corrigan, J. Mardel, S. Garcia, J. Mol, The effect of inhibitor structure on the corrosion of AA2024 and AA7075, Corros. Sci. 53 (2011) 2184–2190.
[18] D.A. Winkler, M. Breedon, P. White, A. Hughes, E. Sapper, I. Cole, Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors, Corros. Sci. 106 (2016) 229–235.
[19] T. Würger, C. Feiler, F. Musil, G.B. Feldbauer, D. Höche, S.V. Lamaka, M. L. Zheludkevich, R.H. Meißner, Data science based Mg corrosion engineering, Frontiers in, Materials 6 (2019) 53.
[20] H. Assad, A. Kumar, Understanding functional group effect on corrosion inhibition efficiency of selected organic compounds, J. Mol. Liq. 344 (2021), 117755.
[21] H.-C. Yi, Z.-H. You, D.-S. Huang, C.K. Kwoh, Graph representation learning in bioinformatics: trends, methods and applications, Brief. Bioinform. 23 (2022) (bbab340).
[22] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, Adv. Neural Inf. Process. Syst. 28 (2015).
[23] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, Int. Conf. Mach. Learn. (2017) 1263–1272.
[24] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism, J. Med. Chem. 63 (2019) 8749–8760.
[25] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, Analyzing learned molecular representations for property prediction, J. Chem. Inf. Model. 59 (2019) 3370–3388.
[26] J. Dai, D. Fu, G. Song, L. Ma, X. Guo, A. Mol, I. Cole, D. Zhang, Cross-category prediction of corrosion inhibitor performance based on molecular graph structures via a three-level message passing neural network model, Corros. Sci. 209 (2022), 110780.
[27] J. Gasteiger, J. Groß, S. Günnemann, Directional message passing for molecular graphs, Int. Conf. Learn. Represent. (2019).
[28] Y. Liu, L. Wang, M. Liu, Y. Lin, X. Zhang, B. Oztekin, S. Ji, Spherical message passing for 3d molecular graphs, Int. Conf. Learn. Represent (2022).
[29] K. Schütt, P.-J. Kindermans, H.E. Sauceda Felix, S. Chmiela, A. Tkatchenko, K.-R. Müller, Schnet: a continuous-filter convolutional neural network for modeling quantum interactions, Adv. Neural Inf. Process. Syst. 30 (2017).
[30] S. Li, J. Zhou, T. Xu, D. Dou, H. Xiong, Geomgcl: Geometric graph contrastive learning for molecular property prediction, Proc. AAAI Conf. Artif. Intell. (2022) 4541–4549.
[31] RDKit:Open-source cheminformatics. www.rdkit.org. [accessed 11-April-2013.
[32] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge university press,, 2000.
[33] M. Belgiu, L. Drăguţ, Random forest in remote sensing: a review of applications and future directions, ISPRS J. Photogramm. Remote Sens. 114 (2016) 24–31.
[34] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[35] P. Kumar, R. Bhatnagar, K. Gaur, A. Bhatnagar, Classification of imbalanced data: review of methods and applications, in: IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2021, 012077.
[36] S. Riniker, G.A. Landrum, Better informed distance geometry: using what we know to improve conformation generation, J. Chem. Inf. Model. 55 (2015) 2562–2574.
[37] T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, J. Comput. Chem. 17 (1996) 490–519.
[38] chemprop/chemprop.https://github.com/chemprop/Chemprop/, 2021 (accessed 20 Oct 2021).
[39] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I Polosukhin, Attention is all you need, Advances in neural information processing systems, 30, 2017.
[41] Y. Abboud, A. Abourriche, T. Saffaj, M. Berrada, M. Charrouf, A. Bennamara, N. Al Himidi, H. Hannache, 2, 3-Quinoxalinedione as a novel corrosion inhibitor for mild steel in 1 M HCl, Mater. Chem. Phys. 105 (2007) 1–5.
[42] G. Palumbo, M. Gorny, J. Banaś, Corrosion inhibition of pipeline carbon steel (N80) in CO_2-saturated chloride (0.5 M of KCl) solution using gum arabic as a possible environmentally friendly corrosion inhibitor for shale gas industry, J. Mater. Eng. Perform. 28 (2019) 6458–6470.
[43] M. Desimone, G. Gordillo, S.N. Simison, The effect of temperature and concentration on the corrosion inhibition mechanism of an amphiphilic amido-amine in CO_2 saturated solution, Corros. Sci. 53 (2011) 4033–4043.
[44] M. Van Soestbergen, S. Erich, H. Huinink, O. Adan, Dissolution properties of cerium dibutylphosphate corrosion inhibitors, Corros. Eng., Sci. Technol. 48 (2013) 234–240.
[45] L. Ma, J. Wang, Y. Wang, et al., Enhanced active corrosion protection coatings for aluminum alloys with two corrosion inhibitors co-incorporated in nanocontainers, Corros. Sci. 208 (2022), 110663.
[46] K. Kousar, M. Walczak, T. Ljungdahl, A. Wetzel, H. Oskarsson, P. Restuccia, E. Ahmad, N. Harrison, R. Lindsay, Corrosion inhibition of carbon steel in hydrochloric acid: elucidating the performance of an imidazoline-based surfactant, Corros. Sci. 180 (2021), 109195.
[47] J. Wang, J. Jing, L. Feng, H. Zhu, Z. Hu, X. Ma, Study on corrosion inhibition behavior and adsorption mechanism of novel synthetic surfactants for carbon steel in 1 M HCl solution, Sustain. Chem. Pharm. 23 (2021), 100500.
[48] X. Jin, J. Wang, S. Zheng, J. Li, X. Ma, L. Feng, H. Zhu, Z. Hu, The study of surface activity and anti-corrosion of novel surfactants for carbon steel in 1 M HCl, J. Mol. Liq. 353 (2022), 118747.
[49] S. Satpati, A. Suhasaria, S. Ghosal, A. Saha, S. Dey, D. Sukul, Amino acid and cinnamaldehyde conjugated Schiff bases as proficient corrosion inhibitors for mild steel in 1 M HCl at higher temperature and prolonged exposure: detailed electrochemical, adsorption and theoretical study, J. Mol. Liq. 324 (2021), 115077.