

Cross-category prediction of corrosion inhibitor performance based on molecular graph structures via a three-level message passing neural network model

Dai, Jiaxin; Fu, Dongmei; Song, Guangxuan; Ma, Lingwei; Guo, Xin; Mol, Arjan; Cole, Ivan; Zhang, Dawei

DOI

[10.1016/j.corsci.2022.110780](https://doi.org/10.1016/j.corsci.2022.110780)

Publication date

2022

Document Version

Final published version

Published in

Corrosion Science

Citation (APA)

Dai, J., Fu, D., Song, G., Ma, L., Guo, X., Mol, A., Cole, I., & Zhang, D. (2022). Cross-category prediction of corrosion inhibitor performance based on molecular graph structures via a three-level message passing neural network model. *Corrosion Science*, 209, Article 110780. <https://doi.org/10.1016/j.corsci.2022.110780>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

<https://www.openaccess.nl/en/you-share-we-take-care>

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.



Cross-category prediction of corrosion inhibitor performance based on molecular graph structures via a three-level message passing neural network model

Jiaxin Dai^{a,b}, Dongmei Fu^{a,b,*}, Guangxuan Song^{a,b}, Lingwei Ma^{b,c}, Xin Guo^{b,c}, Arjan Mol^d, Ivan Cole^e, Dawei Zhang^{b,c,**}

^a School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

^b National Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, Beijing, China

^c Beijing Advanced Innovation Center for Materials Genome Engineering, Institute for Advanced Materials & Technology, University of Science and Technology Beijing, Beijing, China

^d Department of Materials Science and Engineering, Delft University of Technology, Delft, the Netherlands

^e School of Engineering, RMIT University, Melbourne, Australia

ARTICLE INFO

Keywords:

Corrosion inhibitors
Molecular structure
Machine learning
Message passing neural network
SMILES

ABSTRACT

Current experimental verification, computational modeling, and machine learning methods for predicting corrosion inhibition efficiency (IE) are limited to specific inhibitor categories with high cost and poor generalization. In this study, a cross-category corrosion inhibitor dataset is constructed and a three-level direct message passing neural network (3 L-DMPNN) model using molecular structure information that integrates atomic-level, chemical bond-level, and molecular-level features to predict the IEs of compounds in a specific environment is established. This work demonstrates that the 3 L-DMPNN model can predict IEs of cross-category corrosion inhibitors from other independent literature and experimental dataset effectively and quickly.

1. Introduction

Corrosion is the leading cause of materials damage in industrial applications. According to recent surveys, the annual cost of corrosion is equivalent to 2–5% of GDP of different countries, totaling \$2.5 trillion globally [1,2]. Owing to the high protective efficiency, low cost, simple operational process, and strong adaptability, the application of corrosion inhibitors has become a popular method for combating internal corrosion in various industries. The effectiveness of corrosion inhibitors is closely related to their molecular structure. In general, heterocyclic organic compounds with electronegative atoms such as S, P, N, and O; polar groups such as $-NH_2$, $-NO_2$, $-OC_2H_5$, $-COOH$, and $-CONH_2$; or certain chemical structures such as conjugate bonds and aromatic rings can potentially serve as effective corrosion inhibitors [3]. However, the number of organic molecules with the aforementioned molecular structures is immense, which necessitates the development of fast and efficient screening methods for estimating the corrosion inhibition efficiency (IE). Traditionally, IE is determined experimentally by

performing weight loss measurements, potentiodynamic polarization studies, electrochemical impedance spectroscopy, optical analysis, and analytical spectroscopy analysis [4–6]. Researchers have identified new corrosion inhibitors by either incrementally adjusting the structures of existing inhibitors or testing hundreds of compounds in a laboratory [7]. Nevertheless, these experimental methods are expensive and time-consuming, often taking hours or days.

In addition to experimental approaches, theoretical tools, such as density functional theory (DFT) and molecular dynamics (MD) simulation, have been widely used in the studies of corrosion inhibitors [8–10]. DFT method provides important information about the charge sharing (donor-acceptor) interactions between inhibitor molecules and metallic surfaces [11] and thus describes the effect of the structural properties of inhibitors on the corrosion process [12]. MD computer simulations are performed to model inhibitor/surface systems, visualize the adsorption process, and determine the energy of their interaction to elucidate the corrosion inhibition mechanism at the mesoscopic level [13,14]. Quantitative structure-activity/property relationships (QSAR/QSPR)

* Corresponding author at: School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China.

** Corresponding author at: National Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, Beijing, China.

E-mail addresses: fdm2003@163.com (D. Fu), dzhang@ustb.edu.cn (D. Zhang).

are applied to establish correlations between the IEs and structural parameters (electronegativity, polarizability, van der Waals volume, etc.) of corrosion inhibitors and predict the corrosion inhibition performance of the same series of molecules. Common methods used to construct QSAR/QSPR models include multiple linear regression analysis, neural networks, and support vector machines [15–18]. However, all these techniques are time-consuming, complex, and not suitable for screening candidate inhibitors in a large compound space. In addition, these methods do not properly utilize molecular structure and are unable to obtain either local or global information on the molecular parameters.

Recently, machine learning (ML) methods have been employed to predict the IEs of corrosion inhibitors and design new molecules with improved inhibition performance. For example, an artificial neural network (ANN) was successfully utilized to predict the corrosion inhibition potentials of pyridazine derivatives [19] and IEs of 28 [20,21] and 100 [22] small organic compounds for aluminum alloys, and to quantitatively study the relationship between the molecular features of 38 inhibitor compounds and their experimentally measured electrochemical properties [23]. Galvão et al. compared different ML methods to identify efficient corrosion inhibitors for aluminum alloys commonly used in aeronautical applications [24]. Schiessler et al. predicted the IEs of magnesium dissolution modulators using sparse ML models [25]. Wurger et al. used a multidisciplinary approach combining high-throughput experimental screening, unsupervised clustering-based ML algorithms, and DFT calculation to estimate the IEs of previously untested molecules for magnesium alloys [26]. The models utilized in the above-mentioned studies could predict corrosion inhibitor performance within minutes rather than hours or days. However, the limitation of these conventional ML models lies in the requirement of selecting appropriate molecular features, which is a process depending on considerable domain expertise. Most previously reported ML models employed small datasets with limited homologous molecules and exhibited poor generalization properties that did not allow them to make reliable prediction for any molecules outside the training data domain (e.g., the molecules with functional groups not present in the training set).

Since the advent of big data, deep learning (DL) models with more complex architectures have more powerful feature learning and representation capabilities and have been widely used in image recognition, object detection, and other fields, such as drug discovery and genomics [27]. Traditional deep convolutional neural networks (CNN) and recurrent neural networks (RNN) can only process Euclidean data such as text, audio, images, and video. Unlike images and texts, graph data contain essential structural information. A graph neural network (GNN) is a framework for learning directly from graph data that consists of nodes and edges, where nodes and edges can be well used to represent atoms and bonds in molecular structure [28]. Thus, GNN can be used to process non-Euclidean data such as chemical molecular structures and proteins and predict the properties of molecules from their structures [29,30], including quantum mechanical characteristics such as energetic, electronic, and thermodynamic properties [31–33]; physicochemical properties such as hydrophobicity, hydration free energy in water, and octanol/water distribution coefficients [34,35], and toxicity [36,37]. Message-passing neural networks (MPNNs) are general frameworks for supervised graph learning that simply abstract the commonalities between several of the most promising GNN models [33], which are capable of learning atomic-level and chemical bond-level features from molecular graphs directly and predicting molecular properties. These networks were able to accurately predict geometric, energetic, electronic, and thermodynamic properties in the QM9 quantum chemistry dataset consisting of organic molecules.

The structures of organic compounds play a dominating role in the effective inhibition of metal corrosion. The purpose of this study was to develop an IE prediction model independent of theoretical calculations and expert-crafted features to establish the relationship between the molecular structure and IE with high accuracy and generalization for the

rapid screening of corrosion inhibitors. In addition, the efficiency of a corrosion inhibitor is closely related not only to its internal structural parameters (such as hybridization degree, number of bonds per atom, number of valence electrons per atom, and bond type), but also to global molecular characteristics (such as molecular weight, number of aromatic rings, number of acceptors, and number of donors). Hence, we proposed a three-level direct message passing neural network (3 L-DMPNN) model based on the DMPNN framework [38] for screening corrosion inhibitors by combining atomic-level features, chemical bond-level features, and molecular-level features. Specifically, the simplified molecular-input line-entry system (SMILES) [39] was used as the sole input and considered a molecule structure as a graph, and the atomic and chemical bond features were extracted from SMILES by employing the opensource RDKit package [40]. Subsequently, the new molecular graph vector after the message passing module was combined with the global molecular features to predict the IE of the molecule via feed-forward neural network (FFN). The data used in this study were extracted from 110 publications, including 270 organic inhibitor molecules. The accuracy of the utilized model was compared with those of the support vector machine (SVM) [41], random forest (RF) [42], and DMPNN models [38], and its generalization ability was verified using additional 23 recently published papers and 4 laboratory data.

2. Methods

2.1. Corrosion inhibitor datasets

2.1.1. Dataset for building predictive models

Although a large amount of corrosion inhibitor data is contained in research papers, each individual paper often reports no more than a handful of molecules of the same or similar categories. To predict the properties of cross-category corrosion inhibitors, structured and intelligent dataset containing corrosion inhibitor with diverse molecular structures is needed. For example, Galvão et al. have recently established CORDATA as a public data management platform for corrosion inhibitors [43], which include nearly 400 corrosion inhibitors from more than 120 publications, mainly for aluminum, copper, magnesium, iron and their major alloys.

To investigate the effect of molecular structure on corrosion inhibition efficiency, we constructed a cross-category dataset of corrosion inhibitor molecules based on a large number of publicly available literature studies by performing the following two steps: (I) data collection and data cleaning, and (II) SMILES generation (Fig. 1). We retrieved 116 papers that studied the influences of different corrosion inhibitor molecules on carbon steel in a hydrochloric acid solution using the keyword “corrosion inhibitor” and extracted their textual characteristics (e.g., names of corrosion inhibitors, categories of corrosion inhibitors, materials, and solutions) and molecular structure pictures via crowdsourcing. The experimental IE values listed in the tables were extracted by Tabula [44] and associated with the text data via string matching. After that, we discarded molecules for which IE was not measured under the specified environmental conditions (ambient temperature: 25 °C–30 °C, corrosion inhibitor concentration: 1 mmol/L, and HCl concentration: 1 mol/L), which are commonly reported in literature. Data cleaning and filtering were performed by verifying the accuracy of molecular structures and removing duplicate inhibitor molecules. Finally, data for 270 corrosion inhibitor molecules were obtained, including their names, categories, molecular structures, and experimental IE values. The content of the datasets used in this study can be accessed through the following URL: <https://www.corrdata.org.cn/inhibitor/>.

SMILES is a symbol widely used in chemistry to describe the structures of molecules, in which ASCII symbols are employed to represent atoms, bonds, and structural information (such as rings) readable by computers [39]. In this study, we utilized the ChemSchematic [45], OpenBabel [46], and OSRA [47] software packages to generate SMILES

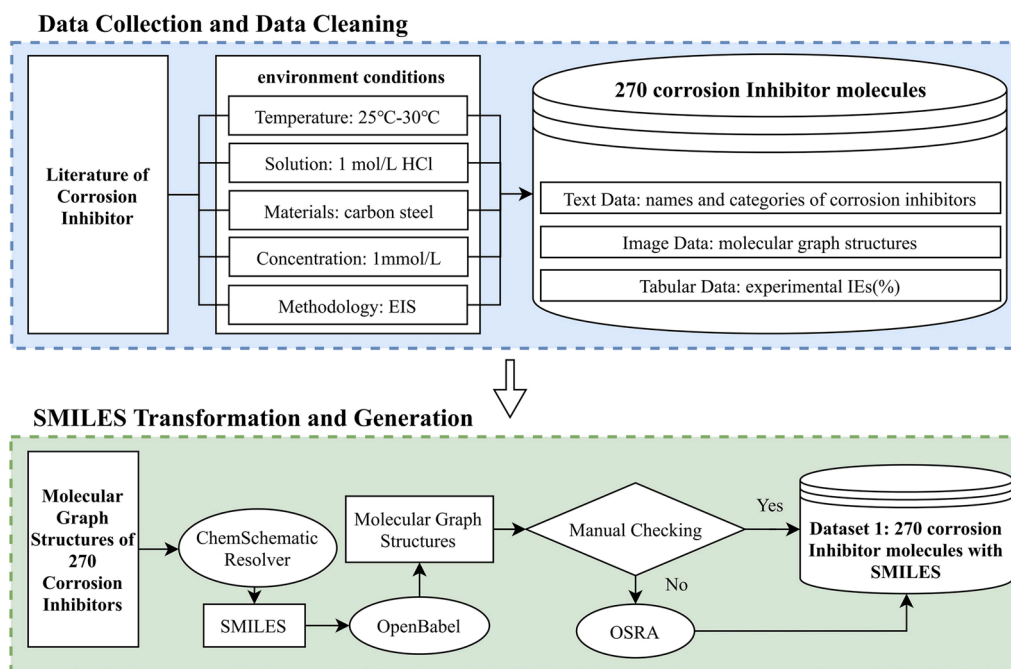


Fig. 1. Process of building the dataset of corrosion inhibitors.

from molecular structure pictures. However, a molecule can have more than one possible SMILES string, leading to the definition of canonical SMILES [48]. We obtained canonical SMILES by a canonical algorithm

to ensure that only one expression was used for each molecule. The resulting **Dataset 1** contained the names of the studied corrosion inhibitors, their categories, IE values determined at the concentration of

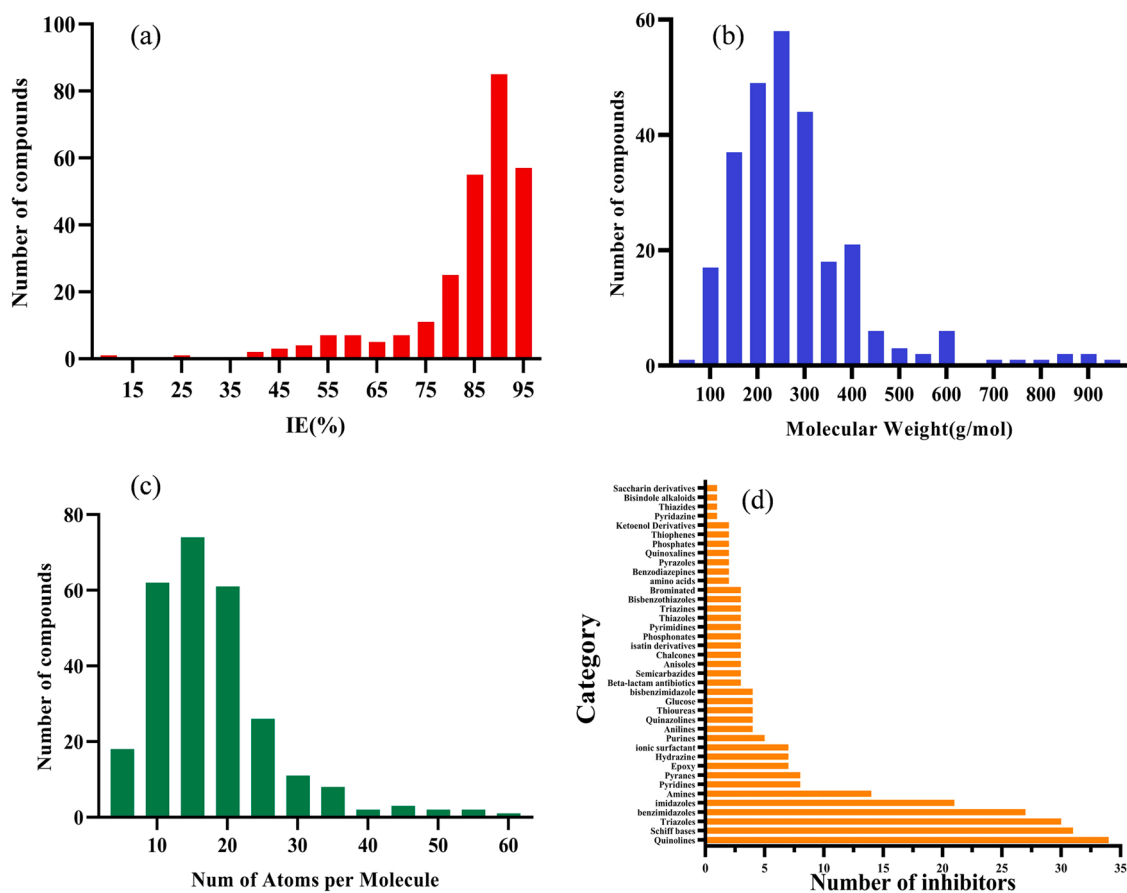


Fig. 2. Statistics of Dataset 1. Distributions of the (a) IE values of corrosion inhibitors, (b) molecular weights of molecules, (c) numbers of atoms per molecule, and (d) categories of molecules.

1 mmol/L in 1 mol/L HCl solution, and canonical SMILES.

Fig. 2a displays the distribution of the obtained IE values. It shows that the number of weak corrosion inhibitors is much lower than that of high-efficiency ones, which makes data learning and correctly predicting the performance of weak inhibitors a challenging process. As shown in Figs. 2b and 2c, the molecular weights of the studied inhibitors range from 96 to 991 g/mol, while their molecule sizes (the number of atoms omitting hydrogen atoms in the molecule) vary from 6 to 63 atoms. Fig. 2d presents a statistical plot of 39 molecular categories, which includes but not limited to triazoles, quinolines, pyrazoles, and Schiff bases. Among these categories, quinolines have the largest number of molecules (34). The molecular categories in the dataset exhibit both diversity and similarity, which is helpful for model learning. In addition, the time of immersion after which the IE was determined ranged from 5 min to 1440 min but mainly concentrated in the range of 30–60 min (70% of the data).

Molecular fingerprint is an abstract representation of a molecule that transforms (encodes) it into a series of bit vectors containing 1 and 0 [49]. Extended connectivity fingerprinting, also known as Morgan fingerprinting, is one of the most widely used molecular fingerprinting techniques [50]. In this study, we used RDKit to calculate 2,048-bit (i.e., 2048 dimensions) Morgan fingerprints with a radius of two atoms for each molecule in **Dataset 1**. The bits in the fingerprint indicate the presence (1) or absence (0) of certain substructures, such as C=C, C(C)C, C=O, and C-N, in the molecule, which makes it easy to measure molecular similarity. Tanimoto similarity [51] is a measure of the proportion of shared chemical substructures in a molecule, which represents a number between 0 and 1, with 0 indicating the lowest degree of similarity (no substructures are shared) and 1 indicating the highest degree of similarity (all substructures are shared). Furthermore, t-distributed stochastic neighbor embedding (t-SNE) is a popular method for dimensionality reduction [52]. In this work, we used t-SNE

with a Tanimoto distance metric to reduce the data points from 2048 dimensions of the Morgan fingerprints to the two dimensions plotted in Fig. 3 to quantify and visualize the similarity between molecules (Tanimoto distance = 1 – Tanimoto similarity). The distance between two blue points represents the Tanimoto similarity of the corresponding molecules (the larger distance represents lower similarity). For example, the boxes in Fig. 3 are filled with similar molecules of the same category, which are relatively close in space, and the serial numbers in the boxes correspond to those in **Dataset 1**. The boxes that are located farther apart also have fewer similar molecules. These results suggest that the molecules in **Dataset 1** do not exhibit significant aggregation in the chemical space, indicating that their molecular structures are highly diverse and suitable for constructing a generalizable prediction models.

2.1.2. Dataset for validating model generalizability

To further validate the generalization ability of the 3 L-DMPNN model, we constructed an additional independent validation set, **Dataset 2**, which included the experimental parameters of 4 corrosion inhibitors determined in the laboratory and data for 23 corrosion inhibitors retrieved from 14 papers published in 2022. We considered not only strong inhibitors but also weak inhibitors in the model validation, with experimental IEs ranging from 55% to 97% for **Dataset 2**. Fig. 4 shows the t-SNE plot of all molecules derived from the training **Dataset 1** (blue dots) and independent test **Dataset 2** (red dots). Note that **Dataset 2** is uniformly distributed in the molecular space covered by **Dataset 1** and contains not only similar molecules (that are close to each other in Fig. 4 such as Nos. 5 and 6 or Nos. 14 and 19), but also molecules of different categories with lower similarity at a high distance (such as Nos. 4 and 11), which are representative for the evaluation of the generalization performance of the studied model.

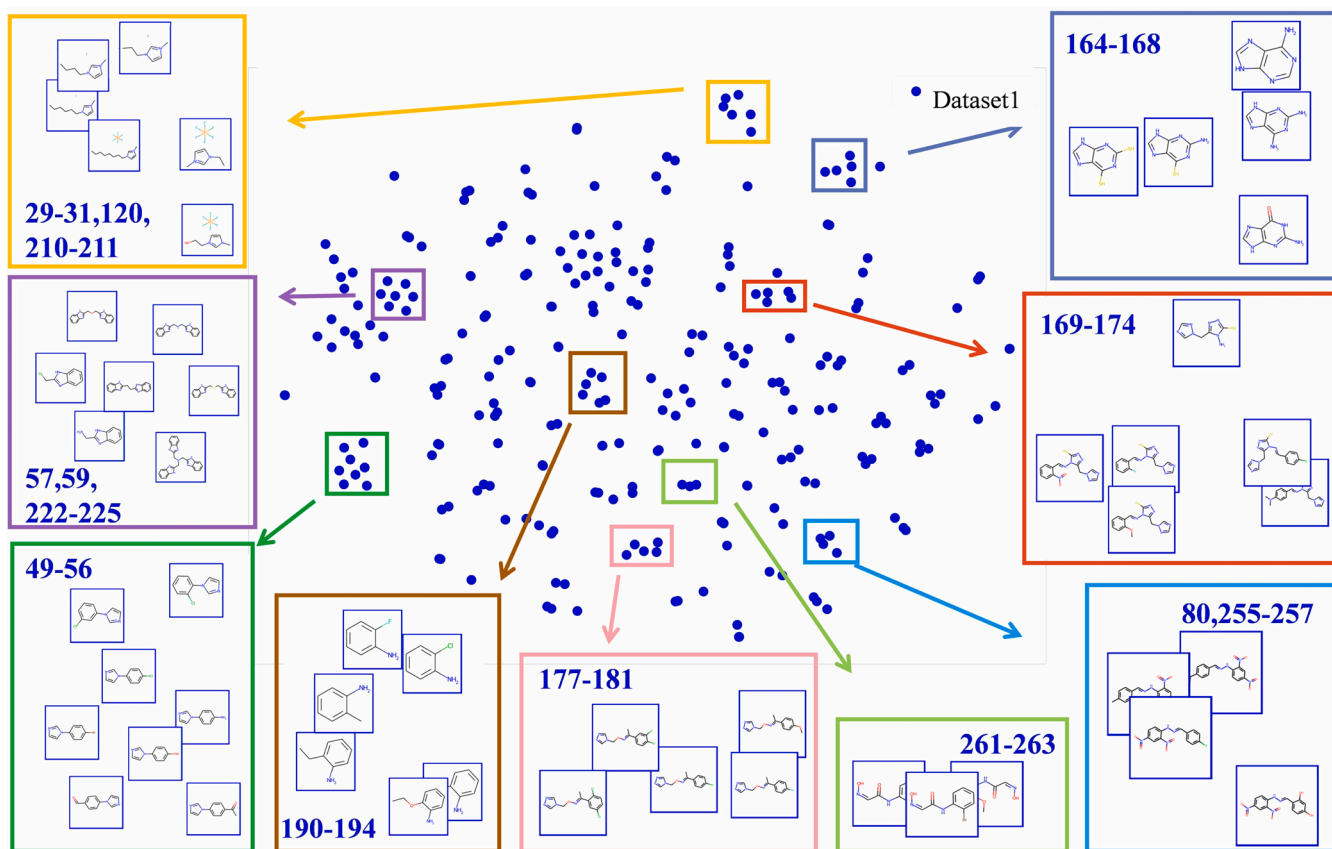


Fig. 3. t-SNE of all molecules from **Dataset 1** (blue dots).

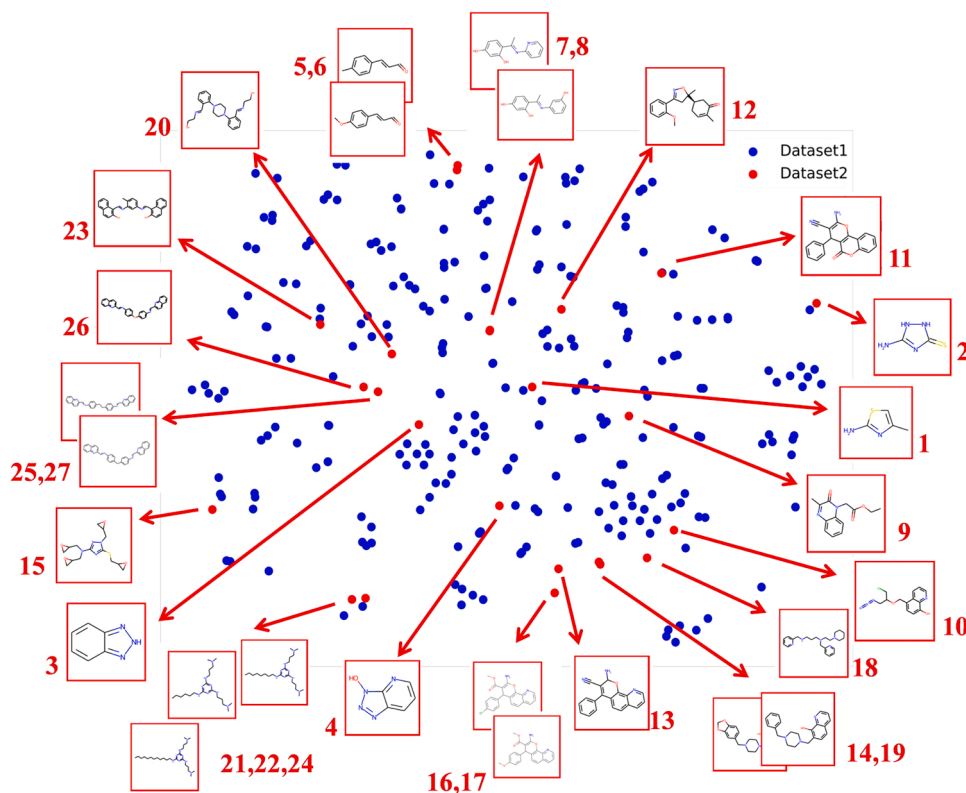


Fig. 4. t-SNE of all molecules from the training Dataset 1 (blue dots) and independent validation Dataset 2 (red dots).

2.2. Models

The 3 L-DMPNN model developed in this study is implemented using the open-source package Chemprop [53], which contains message-passing neural networks for molecular property prediction. Fig. 5 shows the overall framework of the model consisting of a three-level directed message passing network (3 L-DMPN) and a feed-forward neural network (FFN). The input information is processed by the molecular graph representation phase, direct message-passing module, readout phase, and model evaluation phase.

2.2.1. Molecular graphs input representation

A molecular structure was first expressed by a molecular graph G using SMILES as the sole input.

$$G = (V, E) \quad (1)$$

where $v \in V$ is the set of nodes comprising atomic attribute vectors, and $e \in E$ is the set of edges comprising bond attribute vectors.

In the topological graph representation of a molecular structure, the node features correspond to atomic properties such as atomic identity and degree, and the edge features represent bond properties such as bond type and aromaticity. The atomic-level features x_v and chemical bond-level features, which were computed using the open-source package RDKit, were encoded as a one-hot vector and are summarized in Table 1.

2.2.2. Direct message passing

Prior to the first step of message passing, edge hidden h_{vw}^0 states were initialized using Eq. (2):

$$h_{vw}^0 = \tau(W_i \text{cat}(x_v, e_{vw})) \quad (2)$$

where τ is the ReLU, and W_i is the learned matrix.

The messages are subsequently propagated for hidden states h_{vw}^t and messages m_{vw}^t via directed keys based on the graph structure. During each direct message passing step t , the featurization of each bond is updated by summing the featurization of the neighboring bonds. The

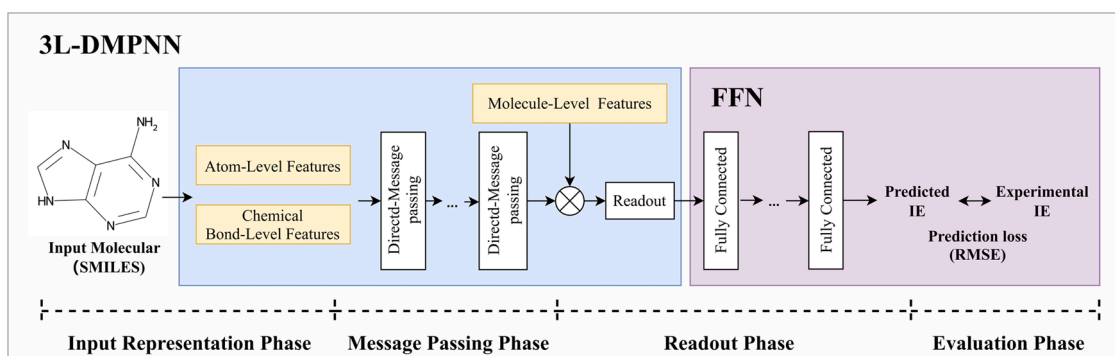


Fig. 5. Neural network architecture of the 3 L-DMPNN model.

Table 1
Atom and chemical bond features used for molecular graph representation.

Features	Description
Atom type	Atomic number
Formal charge	Integer electronic charge assigned to an atom
Hybridization type	sp, sp ² , sp ³ , sp ³ d, or sp ³ d ²
Degree	Number of bonds formed by an atom
Explicit valence	Number of valence electrons
Atomic mass	Mass of an atom (divided by 100)
Aromaticity	Whether an atom is part of an aromatic system
Bond type	Single, double, triple, or aromatic
In ring	Whether a bond is part of a ring
Conjugated	Whether a bond is conjugated
Aromatic	Whether a bond is part of an aromatic system
Stereo	None, any, E/Z, cis/trans

featurization of the current bond is then connected to the sum to construct a neural representation of the molecule. The calculation was performed according to Eqs. (3) and (4):

$$m_{vw}^{t+1} = \sum_{k \in \{N(v)/w\}} h_{kv}^t \quad (3)$$

$$h_{vw}^{t+1} = \tau(h_{kv}^0 + W_m m_{vw}^{t+1}) \quad (4)$$

where $N(v)$ is the set of neighbors of v in graph G , $t \in \{1, \dots, T\}$, T is the total number of steps of the message passing phase, and $W_m \in R^{h \times h}$ is the learned matrix. Next, each atom representation of the molecule was calculated by aggregating the incoming bond features according to Eq. (5).

$$m_v = \sum_{w \in N(v)} h_{vw}^T \quad (5)$$

2.2.3. Readout

After the direct message passing phase, 208 global molecular-level features for each molecule including the number of hydrogen bond donors, number of rotatable bonds, number of aliphatic rings, number of aromatic rings, proportion of sp³-hybridized carbon atoms, lipid-water partition coefficient, and topological polar surface area, etc. were computed using the open-source package RDKit.

The molecular-level features of the 270 corrosion inhibitors comprising **Dataset 3** and the molecular-level features of the 27 corrosion inhibitors comprising **Dataset 4** were incorporated into the model to provide information, where the feature vector h was obtained for the entire molecule by summing the hidden states of all atoms and molecular-level features h_u :

$$h = \sum_{v \in G} h_v + h_u \quad (6)$$

Finally, the feature vector h was fed through a feed-forward neural network for efficiency prediction.

$$\hat{y} = f(h) \quad (7)$$

2.2.4. Evaluation metrics for models

The IE prediction model constructed in this study was a regression model, and the root-mean-square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) values were used as statistical metrics of its performance and were calculated using Eqs. (8–10). The best-performing model from a mathematical point of view was the one with the lowest RMSE and MAE values and highest R^2 magnitude.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (IE_{pred}^i - IE_{exp}^i)^2} \quad (8)$$

$$MAE = \frac{1}{m} \sum_{i=1}^m |IE_{pred}^i - IE_{exp}^i| \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=0}^m (IE_{pred}^i - IE_{exp}^i)^2}{\sum_{i=1}^m (IE_{pred}^i - \overline{IE})^2} \quad (10)$$

$$\overline{IE} = \frac{1}{m} \sum_{i=1}^m IE_{pred}^i \quad (11)$$

where IE_{pred}^i is the predicted IE value for sample i , IE_{exp}^i is the experimental IE value for sample i , m is the total number of samples, and e is the average value of the predicted IE values.

In addition, a cumulative distribution function (CDF) was used to evaluate the absolute error of the test data, and the higher CDF indicated a larger number of molecules with prediction errors less than ΔIE . CDF is obtained by calculating the integral of the probability efficiency function of ΔIE via the Eqs. (12–13).

$$P(\Delta IE) = \int_0^{\Delta IE} \frac{N_t}{N} \times 100 dt \quad (12)$$

$$\Delta IE = 100 \times (IE_{pred} - IE_{exp}) \quad (13)$$

where $P(\Delta IE)$ is the CDF of ΔIE , N is the total number of molecules of the entire testing set, and N_t is the number of molecules with a prediction error below the specified upper limit error t .

2.3. Electrochemical measurements

1 M HCl solution was prepared by analytical reagent-grade 37% hydrochloric acid with distilled water. The electrochemical impedance spectroscopy (EIS) of four corrosion inhibitors (4-Methylthiazol-2-amine, 3-Amino-5-mercapto-1,2,4-triazole, 1 H-benzotriazole and 1-Hydroxy-7-azabenzotriazole) were performed by CHI660E workstation based on a three-electrode system at room temperature (25 °C–30 °C). Q235 carbon steel sample, platinum foil, and saturated calomel electrode were used as the working electrode (WE), the counter electrode and the reference electrode (RE), respectively. Each electrode was immersed for 30 min in the 1 mol/L HCl solutions containing 1 mmol/L corrosion inhibitor before EIS measurements. Thereafter, EIS tests were recorded in the frequency range from 0.01 Hz to 100 kHz with a potential amplitude of ± 10 mV. The corresponding inhibition efficiency (IE) was estimated using Eq. (14).

$$IE(\%) = \frac{R_{ct} - R_{ct}^0}{R_{ct}} \times 100 \quad (14)$$

where R_{ct}^0 and R_{ct} correspond to the charge transfer resistance in the absence and presence of inhibitors, respectively.

3. Results and discussion

3.1. Evaluation of the accuracy of the model

To predict IE values, SMILES was used in **Dataset 1** as input and a rectified linear rectification function (ReLU) as the activation function [54], which increases the nonlinear relationship between the layers of the neural network. To fully utilize the dataset and obtain a reliable model, we employed all 270 molecules from **Dataset 1** but applied a 10-fold cross-validation mode instead of using the pre-split data for training and testing to avoid overfitting and underfitting. During the 10-fold cross-validation procedure, the 270 molecules were divided into ten subsets by random selection. In each run, nine of the ten subsets were selected as the training data, and the remaining subset was used as the testing data for performance evaluation. This process was repeated 10 times until each of the ten subsets has been used as the testing data once. The average test results from the ten runs were utilized to calculate the final score.

Using the Hyper Python package [55], Bayesian optimization [56]

was performed to optimize the hyperparameters and minimize the RMSE metric, including the depth (number of message-passing steps), hidden size (size of bond message vectors), number of feed-forward network layers, and dropout probability of the model. The results of the Bayesian optimization revealed that the recommended hyperparameter set for the current dataset and model were depth = 5, hidden_size = 700, feed-forward network layers = 3, and dropout = 0.0, which were adopted as the hyperparameters in all the subsequent training jobs.

The performance of the 3 L-DMPNN model was compared with those of other ML methods, including SVM, RF, and DMPNN. For a fair comparison, the datasets (including training, valid, and test datasets) and hyperparameter settings were kept the same for each model. Both the SVM and RF models used 2,048-bit Morgan fingerprint vectors as the sole inputs. The RMSE, MAE, and R^2 values determined from the SVM, RF, DMPNN, and 3 L-DMPNN models are summarized in Table 2. The predicted IE values of the 3 L-DMPNN model and the corresponding experimental IE values are shown in Fig. 6. These results revealed that the prediction errors of the four models were ranked as follows: SVM > RF > DMPNN > 3 L-DMPNN. The prediction accuracies of the GNN models (DMPNN and 3 L-DMPNN) were significantly higher than those of the ML models (SVM and RF), which confirmed the high effectiveness of modeling the structure-efficiency relationship based on the molecular structure graph. The 3 L-DMPNN model combining atomic level features, chemical bond-level features, and molecular-level features demonstrated a significantly better performance than the DMPNN model, suggesting that the IE of the corrosion inhibitors was closely related to certain molecular-level features in the studied molecular structures.

Further analyses are illustrated by the histograms of prediction errors plotted in Fig. 7 and CDF curves presented in Fig. 8. Fig. 7 shows that the error distribution of the SVM model is between -16% and 15%, while that of the RF model is between -11% and 25%, which are both relatively discrete. In comparison, the prediction errors of DMPNN are mainly distributed between -13% and 16%, and those of 3 L-DMPNN vary between -8% and 9%, which are concentrated near zero and consistent with the normal distribution. The prediction errors of 3 L-DMPNN are the smallest, which is in good agreement with the results presented in Table 2.

CDF is an integral of the probability density function. When the upper limit error ΔIE is set to 5%, the red dotted line in Fig. 8 marks the proportion of molecules with prediction errors less than 5%, i.e., P(5%). The P(5%) values were 27.0% for SVM, 73.3% for DMPNN, 83.3% for RF, and 94.8% for 3 L-DMPNN. The areas under the curves for different models correspond to the boxes in Figs. 7a-7d, following the order 3 L-DMPNN > RF > DMPNN > SVM. The larger the area, the larger P(5%) and the greater the number of molecules with prediction errors less than 5%. As the upper limit error increased, the values of P(10%) determined for the four models were 65.9% for SVM, 94.8% for RF, 93.7% for DMPNN, and 100% for 3 L-DMPNN (see the blue dotted lines in Fig. 8), while the values of P(15%) determined for the four models were 98.5% for SVM, 97.7% for RF, 99.6% for DMPNN, and 100% for 3 L-DMPNN (see the green dotted lines in Fig. 8). Figs. 6-8 all indicate that the accuracy and effectiveness of the 3 L-DMPNN model are better

Table 2
10-fold cross-validation results obtained for Dataset 1.

Model	RMSE	MAE	R^2
SVM	0.112133 ± 0.016459	0.092632 ± 0.005552	0.225193 ± 0.328819
RF	0.107009 ± 0.029765	0.068037 ± 0.013303	0.339981 ± 0.260449
DMPNN	0.086089 ± 0.019266	0.060416 ± 0.010578	0.458843 ± 0.255685
3 L-DMPNN	0.078170 ± 0.021574	0.053039 ± 0.010515	0.460557 ± 0.584432

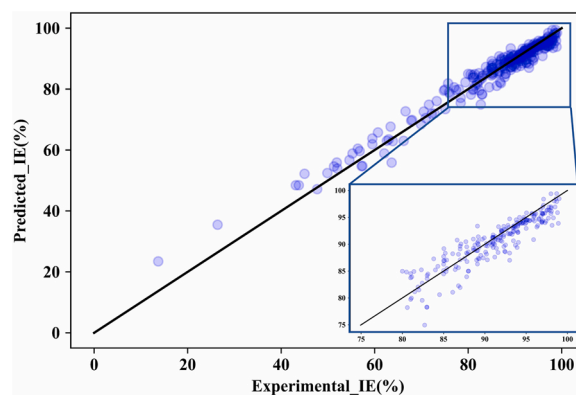


Fig. 6. Model performance evaluation. Predicted and experimental IE values plotted for Dataset 1 using the 3 L-DMPNN model. Each molecule is presented as a blue circle.

than those of the other models. Therefore, this model is most suitable for predicting IE values.

3.2. Validation of the generalizability of the model

The 3 L-DMPNN model trained in Section 3.1 for efficiency prediction was tested using the independent Dataset 2. The immersion time of 23 corrosion inhibitors from the literature in Dataset 2 ranged from 500 s to 60 min, and more than 70% of IEs were measured after 30 min of immersion. Therefore, we performed EIS measurements for 4 molecules (4-Methylthiazol-2-amine, 3-Amino-5-mercapto-1,2,4-triazole, 1 H-benzotriazole, and 1-Hydroxy-7-azabenzotriazole) after 30 min of immersion at the inhibitor concentration of 1 mM. Fig. S1 in Supporting Information shows the Nyquist plots of the EIS results and the equivalent electrical circuit to fit the EIS data. In the circuit, R_s is the solution resistance, R_{ct} and CPE_{dl} correspond to charge transfer resistance and double layer capacitance, respectively. R_L and $L(H)$ represent inductive resistance and inductance, respectively. Constant phase elements (CPEs) are used instead of ideal capacitors because of the inhomogeneity of the surface. The corresponding EIS parameters obtained by fitting the experimental data, such as R_s , R_{ct} , CPE_{dl} , $L(H)$ and R_L , and the IE calculated using Eq. (14), are listed in Table S1 in Supporting Information.

Fig. 9 summarizes the IE values of the corrosion inhibitors measured in the laboratory (green), the IE values obtained from the latest literature studies (red), and the values predicted using the 3 L-DMPNN model (blue). The molecular structures and IE values of the 27 corrosion inhibitors are summarized in Table 3, where the experimental IEs of corrosion inhibitors Nos. 1-4 in Fig. 9 and Table 3 were measured after 30 min of immersion. Unlike the existing efficiency prediction models for corrosion inhibitors, which were trained based on the datasets of molecules under specific categories [19,57,58], the 3 L-DMPNN model was able to learn both the intra-class features of similar molecules and inter-class features of all molecules. Among the 27 molecules, Nos. 5 and 6 (aldehydes), No. 9 (ethyl acetate), No. 12 (isoxazole) and No. 20 (piperazine) are not included in the training Dataset 1. Table 3 shows that the IEs of these five molecules were predicted accurately with absolute errors not exceeding 6%, indicating that the model exhibits good generalization ability and can make a reliable prediction for molecular categories outside the training data domain.

3.3. Significance and limitations of the model

The IE of a corrosion inhibitor is related to its molecular-level features according to expertise, and the dataset utilized in this study was small and contained only hundreds of molecules. Using the DMPNN model with only atomic-level features and chemical bond-level features

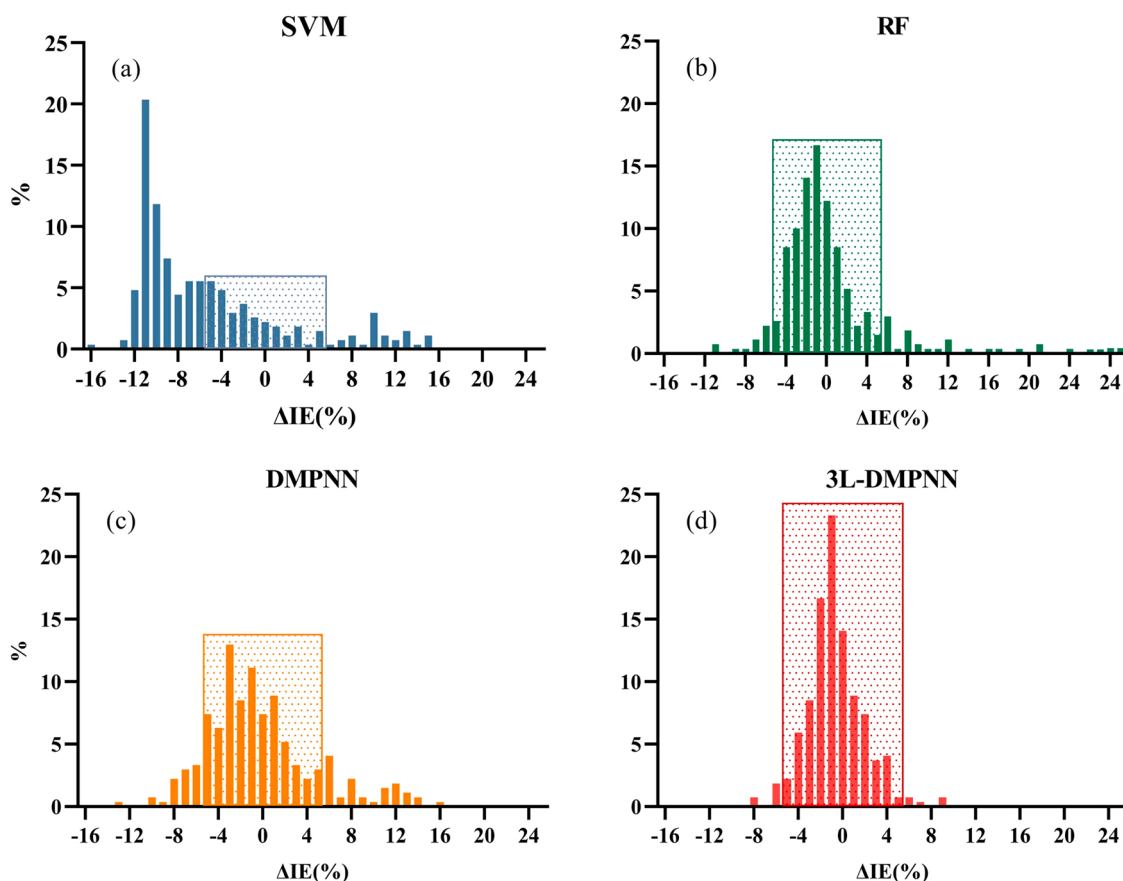


Fig. 7. Histograms of prediction errors obtained for (a) SVM, (b) RF, (c) DMPNN, and (d) 3 L-DMPNN.

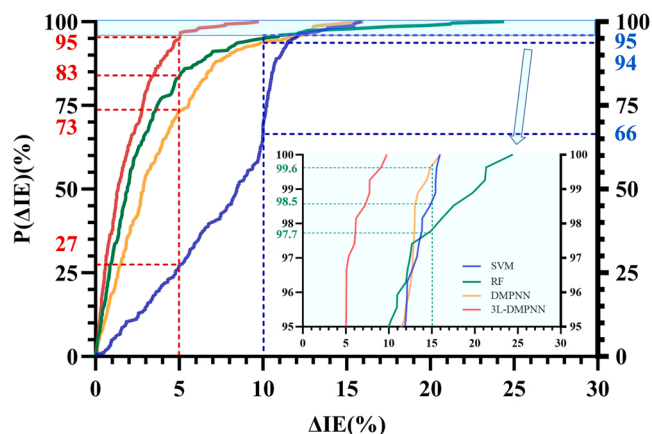


Fig. 8. Plots of CDF versus prediction error constructed for SVM, RF, DMPNN, and 3 L-DMPNN models. The horizontal axis represents the absolute error ΔIE , and the vertical axis denotes the fraction of molecules $P(\Delta IE)$ with errors less than ΔIE .

is poor to identify and extract all features of molecules that may be relevant to IE prediction and are susceptible to overfitting artifacts in the dataset because most original DMPNNs models use fewer message-passing steps than the diameter of the molecular graph [36]. As a result, the atoms with distances larger than the chemical bonds never receive messages from each other.

The novelty of the present work lies in the application of the GNN modelling tool to predict IE values from molecular structure. The 3 L-DMPNN model using SMILES as the input not only extracts atomic-level features and chemical bond-level features in the molecular

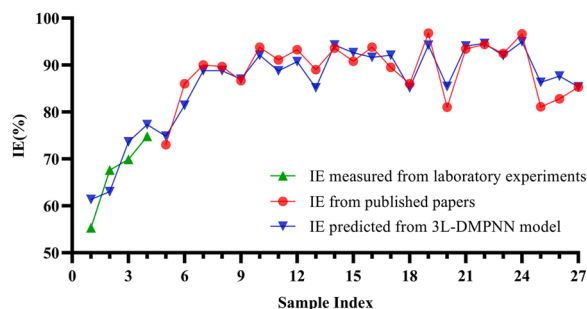


Fig. 9. Experimental efficiencies of the molecules from Dataset 2 and their predicted values obtained by the 3 L-DMPNN model.

structure to disrupt the molecular structure but also calculates molecular-level features to take into account the correlation between molecules to achieve effective prediction of IE for molecules of different molecular weights and categories. The 3 L-DMPNN model exhibited superior performance in predicting IE magnitudes as compared with those of the other models. Unlike the existing models that are limited to the prediction of homologous molecules, the 3 L-DMPNN model enables the cross-category prediction of IE. This model can be applied to virtually any corrosion inhibition dataset in the same corrosive environment, which saves time and facilitates batch processing. The 3 L-DMPNN model can also be utilized as a powerful screening tool before electrochemical measurement to reduce the high cost of experiments and accelerate the corrosion inhibitor development processes.

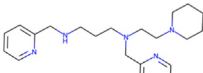
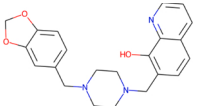
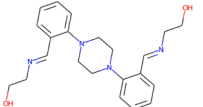
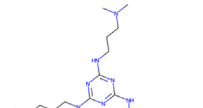
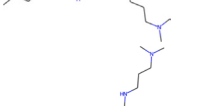
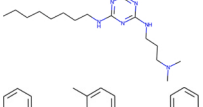
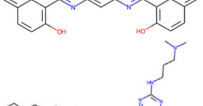
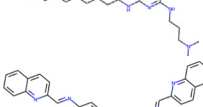
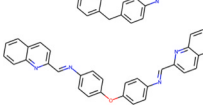
The current model is limited to making accurate prediction for corrosion inhibitor molecules used for the same type of metal (i.e. carbon steels) in the same corrosive environment (1 mol/L HCl). In future

Table 3
Chemical structures of 27 corrosion inhibitors and their IE values.

No.	IUPAC nomenclature	Molecular structure	Category	Experimental IE	Predicted IE	Ref
1	4-Methylthiazol-2-amine		Thiazoles	55.3%	61.3%	laboratory
2	3-Amino-5-mercapto-1,2,4-triazole		Triazoles	67.6%	63.0%	laboratory
3	1 H-Benzotriazole		Triazoles	69.9%	73.7%	laboratory
4	1-Hydroxy-7-azabenzotriazole		Triazoles	74.8%	77.3%	laboratory
5	(2E)- 3-(4-methylphenyl)prop-2-enal		Aldehydes	73.0%	74.9%	[59]
6	(2E)- 3-(4-methoxyphenyl)prop-2-enal		Aldehydes	86.0%	81.5%	[59]
7	(E)- 4-(1-(pyridin-2-ylimino)ethyl)benzene-1,3-diol		Schiff bases	90.0%	88.8%	[60]
8	(E)- 4-(1-((3-hydroxyphenyl)imino)ethyl) benzene-1,3-diol		Schiff bases	89.7%	88.7%	[60]
9	ethyl 2-(3-methyl-2-oxo-1,2-dihydroquinixalin-1-yl)acetate		Ethyl acetate	86.7%	87.0%	[61]
10	5-(((1-azido-3-chloropropan-2-yl)oxy)methyl)quinolin-8-ol		Quinolines	93.8%	92.1%	[62]
11	2-Amino-4-phenyl-2 H-pyrano[3,2-h]quinolin-3-carbonitrile(QP-H)		Quinolines	91.1%	88.8%	[63]
12	3-(2-methoxyphenyl)-isoxazole-carvone		Isoxazoles	93.3%	90.8%	[64]
13	2-amino-5-oxo-4-phenyl-4 H,5 H-pyrano [3,2-c]chromene-3-carbonitrile		Pyranes	89.0%	85.1%	[65]
14	7-((4-benzylpiperazin-1-yl)methyl)quinolin-8-ol		Quinolines	93.6%	94.3%	[66]
15	N,N,1-tri(oxiran-2-ylmethoxy)- 5-((oxiran-2-ylmethoxy)thio)- 1 H-1,2,4-triazol-3-amine (TTA)		Epoxy	90.8%	92.7%	[67]
16	methyl 2-amino-4-(4-methoxyphenyl)- 4 H-pyrano[3,2-h]quinoline-3-carboxylate (P-2)		Quinolines	93.8%	91.7%	[68]
17	methyl 2-amino-4-(4-chlorophenyl)- 4 H-pyrano[3,2-h]quinoline-3-carboxylate (P-1)		Quinolines	89.5%	92.1%	[68]
18			Pyridines	86.0%	85.2%	[69]

(continued on next page)

Table 3 (continued)

No.	IUPAC nomenclature	Molecular structure	Category	Experimental IE	Predicted IE	Ref
19	N1-(2-morpholinoethyl)-N1,N3-bis(pyridine-2-ylmethyl)propane-1,3-diamine		Quinolines	96.8%	94.2%	[66]
20	7-((4-(benzo[d][1,3]dioxol-5-ylmethyl)piperazin-1-yl)methyl)quinolin-8-ol		Piperazine	81.0%	85.5%	[70]
21	1,4-bis(2-(2-hydroxyethyliminomethyl)phenyl)piperazine		Triazines	93.5%	94.1%	[71]
22	2-(n-Hexylamino)-4,6-bis(3-N, N-dimethylaminopropyl) amino-1,3,5-triazine		Triazines	94.4%	94.6%	[71]
23	2,4-Bis(2-hydroxy naphthaldehyde) diiminotoluene (L)		Schiff bases	92.5%	92.0%	[72]
24	2-(n-n-Dodecylamino)-4,6-bis(3-N, N-dimethylaminopropyl) amino-1,3,5-triazine		Triazines	96.7%	95.0%	[71]
25	44,4'-Methylenebis{N-[(E)-quinoléine-2-ylmethylidene] aniline}		Schiff bases	81.1%	86.3%	[73]
26	4,4'-Oxybis{N-[(E)-quinoléine-2-ylmethylidene] aniline}		Schiff bases	82.8%	87.7%	[73]
27	4,4'-Ethane bis{N-[(E)-quinoléine-2-ylmethylidene] aniline}		Schiff bases	85.3%	85.4%	[73]

work, the environmental parameters can be added to the model input to extend the molecular data. In addition, the evolution of IE with time of immersion and the influence of surface finish of the metal can be considered on the basis of an extended dataset to further generalize the model. Quantum chemical parameters of corrosion inhibitor molecules, such as dipole moment, orbital energy, and the number of transferred electrons, are also commonly used to establish molecular structure–efficiency relationships. Adding these parameters as [supplementary information](#) to the learning model can provide a more comprehensive description of the molecular properties from a quantum chemical perspective and increase the prediction accuracy.

4. Conclusion

In this study, a cross-category corrosion inhibitor efficiency dataset was constructed from published research works. As a result, a molecular structure–efficiency prediction model, 3 L–DMPNN, for corrosion inhibitors based on the topological structures of molecular graphs was established. The 3 L–DMPNN model uses the identified molecular descriptors (SMILES) as the sole input while combining atomic-level features, chemical bond-level features, and molecular-level features. The results demonstrated that the 3 L–DMPNN exhibited high prediction accuracy as compared with those of the SVM, RF, and DMPNN models.

In addition, a 10-fold cross-validation approach was utilized to determine the proportions of compounds with prediction errors less than 5% in the overall dataset; the values obtained for the SVM, RF, DMPNN, and 3 L–DMPNN models were 27.0%, 83.3%, 73.3%, and 94.8%, respectively. The generalization ability of the developed model was also validated using 23 molecules from the latest literature studies and 4 molecules tested in laboratory, making prediction for the categories of molecules outside the training data domain. The obtained results indicated that the 3 L–DMPNN model could accurately predict IE values for both strong and weak corrosion inhibitors, allowing rapid screening of corrosion inhibitor molecules at low costs.

CRediT authorship contribution statement

Jiixin Dai: Conceptualization, Methodology, Investigation, Analysis, Writing – original manuscript. **Dongmei Fu:** Supervision, Conceptualization, Methodology, Analysis, Writing – review & editing. **Guangxuan Song:** Methodology, Analysis. **Lingwei Ma:** Analysis, Writing – review & editing. **Xin Guo:** Investigation. **Arjan Mol:** Writing – review & editing. **Ivan Cole:** Writing – review & editing. **Dawei Zhang:** Supervision, Conceptualization, Methodology, Analysis, Writing – review & editing. All authors contributed to the discussion of the results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

All datasets used in this paper can be assessed at <https://www.corrdata.org.cn/inhibitor/>.

Acknowledgements

This work was supported by the Science and Technology Basic Resources Investigation Project (No. 2019FY101404).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.corsci.2022.110780](https://doi.org/10.1016/j.corsci.2022.110780).

References

- G. Koch, J. Varney, N. Thompson, O. Moghissi, M. Gould, J. Payer, International measures of prevention, application, and economics of corrosion technologies study, *NACE Int.* 216 (2016) 2–3.
- 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 Mat. Degrad.* 1 (2017) 1–10.
- C. Verma, Handbook of science & engineering of green corrosion inhibitors: modern theory, fundamentals & Practical Applications, Elsevier, 2021, pp. 41–48.
- H.M. Abd El-Lateef, A.M. Abu-Dief, M.A.A. Mohamed, Corrosion inhibition of carbon steel pipelines by some novel Schiff base compounds during acidizing treatment of oil wells studied by electrochemical and quantum chemical methods, *J. Mol. Struct.* 1130 (2017) 522–542.
- T.H. Muster, A.E. Hughes, S.A. Furman, T. Harvey, N. Sherman, S. Hardin, P. Corrigan, D. Lau, F.H. Scholes, P.A. White, M. Glenn, J. Mardel, S.J. Garcia, J.M.C. Mol, A rapid screening multi-electrode method for the evaluation of corrosion inhibitors, *Electrochim. Acta* 54 (2009) 3402–3411.
- T.H. Muster, H. Sullivan, D. Lau, D.L.J. Alexander, N. Sherman, S.J. Garcia, T. G. Harvey, T.A. Markley, A.E. Hughes, P.A. White, M. Glenn, P.A. White, S. G. Hardin, J. Mardel, J.M.C. Mol, A combinatorial matrix of rare earth chloride mixtures as corrosion inhibitors of AA2024-T3: optimisation using potentiodynamic polarisation and EIS, *Electrochim. Acta* 67 (2012) 95–103.
- I.B. Obot, D.D. Macdonald, Z.M. 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.
- G. Bahlakeh, B. Ramezanzadeh, M. Ramezanzadeh, Cerium oxide nanoparticles influences on the binding and corrosion protection characteristics of a melamine-cured polyester resin on mild steel: An experimental, density functional theory and molecular dynamics simulation study, *Corros. Sci.* 118 (2017) 69–83.
- L. Boucherit, M. Al-Noaimi, D. Daoud, T. Douadi, N. Chafai, S. Chafaa, Synthesis, characterization and the inhibition activity of 3-(4-cyanophenylazo)-2,4-pentanedione (L) on the corrosion of carbon steel, synergistic effect with other halide ions in 0.5 M H₂SO₄, *J. Mol. Struct.* 1177 (2019) 371–380.
- G. Gece, The use of quantum chemical methods in corrosion inhibitor studies, *Corros. Sci.* 50 (2008) 2981–2992.
- D.K. Verma, R. Aslam, J. Aslam, M.A. Quraishi, E.E. Ebenso, C. Verma, Computational modeling: theoretical predictive tools for designing of potential organic corrosion inhibitors, *J. Mol. Struct.* 1236 (2021), 130294.
- I.B. Obot, D.D. Macdonald, Z.M. 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.
- I.B. Obot, Z.M. Gasem, Theoretical evaluation of corrosion inhibition performance of some pyrazine derivatives, *Corros. Sci.* 83 (2014) 359–366.
- Y. Tang, X. Yang, W. Yang, R. Wan, Y. Chen, X. Yin, A preliminary investigation of corrosion inhibition of mild steel in 0.5M H₂SO₄ by 2-amino-5-(n-pyridyl)-1,3,4-thiadiazole: Polarization, EIS and molecular dynamics simulations, *Corros. Sci.* 52 (2010) 1801–1808.
- H. Zhao, X. Zhang, L. Ji, H. Hu, Q. Li, Quantitative structure–activity relationship model for amino acids as corrosion inhibitors based on the support vector machine and molecular design, *Corros. Sci.* 83 (2014) 261–271.
- L. Li, X. Zhang, S. Gong, H. Zhao, Y. Bai, Q. Li, L. Ji, The discussion of descriptors for the QSAR model and molecular dynamics simulation of benzimidazole derivatives as corrosion inhibitors, *Corros. Sci.* 99 (2015) 76–88.
- I. Cole, C. Chu, M. Breedon, F. Chen, D. Winkler, E. Sapper, Computational design of inhibited primers, in: *TechConnect World Innovation 2015: Conference & Expo*, Taylor & Francis, 2015, pp. 99–102.
- M. Fernandez, M. Breedon, I.S. Cole, A.S. Barnard, Modeling corrosion inhibition efficacy of small organic molecules as non-toxic chromate alternatives using comparative molecular surface analysis (CoMSA), *Chemosphere* 160 (2016) 80–88.
- T.W. Quadri, L.O. Olasunkanmi, E.D. Akpan, O.E. Fayemi, H.-S. Lee, H. Lgaz, C. Verma, L. Guo, S. Kaya, E.E. Ebenso, Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors, *Mater. Today Commun.* 30 (2022), 103163.
- T.G. Harvey, S.G. Hardin, A.E. Hughes, T.H. Muster, P.A. White, T.A. Markley, P. A. Corrigan, J. Mardel, S.J. Garcia, J.M.C. Mol, A.M. Glenn, The effect of inhibitor structure on the corrosion of AA2024 and AA7075, *Corros. Sci.* 53 (2011) 2184–2190.
- D.A. Winkler, M. Breedon, A.E. Hughes, F.R. Burden, A.S. Barnard, T.G. Harvey, I. Cole, Towards chromate-free corrosion inhibitors: structure–property models for organic alternatives, *Green. Chem.* 16 (2014) 3349–3357.
- D.A. Winkler, M. Breedon, P. White, A.E. Hughes, E.D. 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.
- F.F. Chen, M. Breedon, P. White, C. Chu, D. Mallick, S. Thomas, E. Sapper, I. Cole, Correlation between molecular features and electrochemical properties using an artificial neural network, *Mater. Des.* 112 (2016) 410–418.
- T.L.P. Galvão, G. Novell-Leruth, A. Kuznetsova, J. Tedim, J.R.B. Gomes, Elucidating Structure–Property Relationships in Aluminum Alloy Corrosion Inhibitors by Machine Learning, *J. Phys. Chem. C* 124 (2020) 5624–5635.
- E.J. Schiessler, T. Würger, S.V. Lamaka, R.H. Meißner, C.J. Cyron, M. L. Zheludkevich, C. Feiler, R.C. Aydin, Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning models, *npj Comput. Mater.* 7 (2021) 1–9.
- T. Würger, C. Feiler, F. Musil, G.B.V. Feldbauer, D. Höche, S.V. Lamaka, M. L. Zheludkevich, R.H. Meißner, Data science based Mg corrosion engineering, *Front. Mater.* 6 (2019) 53.
- Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (2015) 436–444.
- Z. Hao, C. Lu, Z. Huang, H. Wang, Z. Hu, Q. Liu, E. Chen, C. Lee, ASGN: An active semi-supervised graph neural network for molecular property prediction, *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* (2020) 731–752.
- C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin, L. He, Molecular property prediction: A multilevel quantum interactions modeling perspective, *Proc. AAAI Conf. Artif. Intell.* (2019) 1052–1060.
- N.A. Asif, Y. Sarker, R.K. Chakraborty, M.J. Ryan, M.H. Ahamed, D.K. Saha, F. R. Badal, S.K. Das, M.F. Ali, S.I. Moyeen, M.R. Islam, Z. Tasneem, Graph neural network: A comprehensive review on non-euclidean space, *IEEE Access* 9 (2021) 60588–60606.
- Z. Wu, B. Ramsundar, Evan N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: a benchmark for molecular machine learning, *Chem. Sci.* 9 (2018) 513–530.
- K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing learned molecular representations for property prediction, *J. Chem. Inf. Model* 59 (2019) 3370–3388.
- J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: *International conference on machine learning*, PMLR, 2017, pp. 1263–1272.
- X. Wang, Z. Li, M. Jiang, S. Wang, S. Zhang, Z. Wei, Molecule property prediction based on spatial graph embedding, *J. Chem. Inf. Model* 59 (2019) 3817–3828.
- B. Chen, G. Bécigneul, O.-E. Ganea, R. Barzilay, T. Jaakkola, Optimal transport graph neural networks, *arXiv preprint arXiv: 2006.04804*, (2020).
- Y. Xu, J. Pei, L. Lai, Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction, *J. Chem. Inf. Model* 57 (2017) 2672–2685.
- M. Withnall, E. Lindelof, O. Engkvist, H. Chen, Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction, *J. Cheminform.* 12 (2020) 1–18.
- K. Yang, K. Swanson, W. Jin, C. Coley, H. Gao, A. Guzman-Perez, T. Hopper, B.P. Kelley, A. Palmer, V. Settels, Are learned molecular representations ready for prime time? *ChemRxiv* (2019).
- D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Model* 28 (1988) 31–36.
- RDKit: Open-Source Cheminformatics. <https://rdkit.org/docs/index.html>, 2021 (accessed 20 Oct 2021).
- N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press, 2000.
- M. Belgiu, L. Drăguț, Random forest in remote sensing: a review of applications and future directions, *ISPRS J. Photogramm.* 114 (2016) 24–31.
- T.L.P. Galvão, I. Ferreira, A. Kuznetsova, G. Novell-Leruth, C. Song, C. Feiler, S. V. Lamaka, C. Rocha, F. Maia, M.L. Zheludkevich, J.R.B. Gomes, J. Tedim, CORDATA: an open data management web application to select corrosion inhibitors, *npj Mat, Degrad* 6 (2022) 1–4.
- M. Arístarán, M. Tigas, J.B. Merrill, Tabula (Version 1.2.1), June 4, 2018. <https://tabula.technology/>.
- E.J. Beard, J.M. Cole, ChemSchematicResolver: a toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities, *J. Chem. Inf. Model* 60 (2020) 2059–2072.
- N.M. O’Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open babel: an open chemical toolbox, *J. Cheminform.* 3 (2011) 1–14.

- [47] I.V. Filippov, M.C. Nicklaus, Optical structure recognition software to recover chemical information: OSRA, an open source solution, *J. Chem. Inf. Model* 49 (2009) 740–743.
- [48] E.J. Bjerrum, SMILES enumeration as data augmentation for neural network modeling of molecules, arXiv preprint arXiv: 1703.07076, (2017).
- [49] 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).
- [50] D.S. Wigh, J.M. Goodman, A.A. Lapkin, A review of molecular representation in the age of machine learning, *WIREs Comput. Mol. Sci.* (2022), e1603.
- [51] D. Bajusz, A. Racz, K. Heberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Chemin.-.* 7 (2015) 20.
- [52] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (2008) 2579–2605.
- [53] chemprop/chemprop. <https://github.com/chemprop/Chemprop/>, 2021 (accessed 20 Oct 2021).
- [54] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: *Icml*, 2010.
- [55] Distributed Asynchronous Hyperparameter Optimization in Python. <https://github.com/hyperopt/hyperopt/>, 2021 (accessed 20 Oct 2021).
- [56] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. de Freitas, Taking the human out of the loop: A review of bayesian optimization, *Proc. IEEE* 104 (2016) 148–175.
- [57] A. Kokalj, M. Lozinsek, B. Kapun, P. Taheri, S. Neupane, P. Losada-Pérez, C. Xie, S. Stavber, D. Crespo, F.U. Renner, A. Mol, I. Milošev, Simplistic correlations between molecular electronic properties and inhibition efficiencies: do they really exist? *Corros. Sci.* 179 (2021), 108856.
- [58] C.T. Ser, P. Žuvela, M.W. Wong, Prediction of corrosion inhibition efficiency of pyridines and quinolines on an iron surface using machine learning-powered quantitative structure-property relationships, *Appl. Surf. Sci.* 512 (2020), 145612.
- [59] J. Lazrak, E. Ech-chihbi, B. El Ibrahim, F. El Hajjaji, Z. Rais, M. Tachihante, M. Taleb, Detailed DFT/MD simulation, surface analysis and electrochemical computer explorations of aldehyde derivatives for mild steel in 1.0 M HCl, *Colloid Surf. A* 632 (2022), 127822.
- [60] A. Salhi, H. Amhamdi, M. El Massaoudi, I. Azghay, S. El Barkany, A. Elyoussfi, M. Ahari, A. Bouyanzer, S. Radi, A. Zarrouk, Preventive behavior of phenol Schiff bases on mild steel corrosion in acidic medium part A: experimental and molecular modeling approach, *Chem. Data Collect.* 39 (2022), 100864.
- [61] M. Missioui, M. Bouziani Idrissi, F. Benhiba, Z. Atiöglu, M. Akkurt, H. Oudda, J. T. Mague, E.M. Essassi, A. Zarrouk, Y. Ramli, Synthesis, structural characterization, Hirshfeld surface analysis and anti-corrosion on mild steel in 1M HCl of ethyl 2-(3-methyl-2-oxo-1,2-dihydroquinoxaline-1-yl)acetate, *J. Mol. Struct.* 1251 (2022), 132047.
- [62] M. El Faydy, F. Benhiba, I. Warad, H. About, S. Saoiabi, A. Guenbour, F. Bentiss, B. Lakhrissi, A. Zarrouk, Experimental and theoretical investigations of two quinolin-8-ol derivatives as inhibitors for carbon steel in 1 M HCl solution, *J. Phys. Chem. Solids* 165 (2022), 110699.
- [63] M. Oubaaq, M. Ouakki, M. Rbaa, F. Benhiba, M. Galai, R. Idouhli, M. Maatallah, A. Jarid, I. Warad, B. Lakhrissi, A. Zarrouk, M. Ebn, Touhami, Experimental and theoretical investigation of corrosion inhibition effect of two 8-hydroxyquinoline carbonitrile derivatives on mild steel in 1 M HCl solution, *J. Phys. Chem. Solids* 169 (2022), 110866.
- [64] E. Elqars, A. Oubella, M. Eddine Hachim, S. Byadi, A. Auhmani, M. Guennoun, A. Essadki, A. Riahi, A. Robert, M. Youssef Ait Itto, T. Nbigui, New 3-(2-methoxyphenyl)-isoxazole-carvone: synthesis, spectroscopic characterization, and prevention of carbon steel corrosion in hydrochloric acid, *J. Mol. Liq.* 347 (2022), 118311.
- [65] M. Ouakki, M. Galai, Z. Aribou, Z. Benzekri, E.H.E. Assiri, K. Dahmani, E. Ech-chihbi, A.S. Abousalem, S. Boukhris, M. Cherkaoui, Detailed experimental and computational explorations of pyran derivatives as corrosion inhibitors for mild steel in 1.0 M HCl: electrochemical/surface studies, DFT modeling, and MC simulation, *J. Mol. Struct.* 1261 (2022), 132784.
- [66] M. El Faydy, F. Benhiba, N. Timoudan, B. Lakhrissi, I. Warad, S. Saoiabi, A. Guenbour, F. Bentiss, A. Zarrouk, Experimental and theoretical examinations of two quinolin-8-ol-piperazine derivatives as organic corrosion inhibitors for C35E steel in hydrochloric acid, *Journal of Molecular Liquids* 354 (2022).
- [67] M. Damej, R. Hsissou, A. Berisha, K. Azgaou, M. Sadiku, M. Benmessaoud, N. Labjar, S. El, hajjaji, New epoxy resin as a corrosion inhibitor for the protection of carbon steel C38 in 1M HCl. experimental and theoretical studies (DFT, MC, and MD), *J. Mol. Struct.* 1254 (2022), 132425.
- [68] M. Abouchane, N. Dkhireche, M. Rbaa, F. Benhiba, M. Ouakki, M. Galai, B. Lakhrissi, A. Zarrouk, M. Ebn Touhami, Insight into the corrosion inhibition performance of two quinoline-3-carboxylate derivatives as highly efficient inhibitors for mild steel in acidic medium: Experimental and theoretical evaluations, *J. Mol. Liq.* 360 (2022), 119470.
- [69] M. Rezaeivala, S. Karimi, B. Tuzun, K. Sayin, Anti-corrosion behavior of 2-((3-(2-morpholino ethylamino)-N3-(pyridine-2-yl)methyl)propylimino)methylpyridine and its reduced form on carbon steel in hydrochloric acid solution: Experimental and theoretical studies, *Thin Solid Films* 741 (2022), 139036.
- [70] M. Rezaeivala, S. Karimi, K. Sayin, B. Tüzün, Experimental and theoretical investigation of corrosion inhibition effect of two piperazine-based ligands on carbon steel in acidic media, *Colloid Surf. A* 641 (2022), 128538.
- [71] 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.
- [72] S. Boukazoula, D. Haffar, R. Bourzami, L. Toukal, V. Dorcet, Synthesis, characterizations, crystal structure, inhibition effects and theoretical study of novel Schiff base on the corrosion of carbon steel in 1 M HCl, *J. Mol. Struct.* 1261 (2022), 132852.
- [73] H. Hamani, D. Daoud, S. Benabid, T. Douadi, Electrochemical, density functional theory (DFT) and molecular dynamic (MD) simulations studies of synthesized three new Schiff bases as corrosion inhibitors on mild steel in the acidic environment, *J. Indian Chem. Soc.* 99 (2022), 100492.