

Multitask Learning Framework for Screening of Corrosion Inhibitors for Mild Steel

Abhishek Agarwal^a, Pradeep Rathore^a, Dharmendr Kumar^a, Sri Harsha Nistala^a, S Sanjana^b, Vinay Jain^a, Beena Rai^a

^aTCS Research, Tata Consultancy Services Ltd., Hadapsar, Pune-411013, India

^bDepartment of Materials Science and Metallurgical Engineering, Indian Institute of Technology

Hyderabad, Hyderabad – 502285, India

ABSTRACT

Corrosion inhibitors are useful to mitigate corrosion of metal/alloy components. However, traditional corrosion inhibitors are toxic and need to be replaced by greener alternatives. Efficient screening models are required to find molecules with desired properties from millions of molecules available in public domain. To make these models, database of experimental inhibition efficiency of molecules is essential. In this work, we have developed a computational framework to accelerate the discovery of new corrosion inhibitors. We have used machine learning based algorithms to predict corrosion inhibition efficiency of organic molecules for steel in hydrochloric acid by using structural information of the molecules along with experimental conditions. Our multitask learning based neural network architecture was able to outperform traditional machine learning algorithms such as random forest, lasso and ridge regression. We have also created the largest dataset for predictive modelling of corrosion inhibitors for steel. Besides, we have also used the model to screen molecules from the ZINC15 dataset and found potential inhibitors with high inhibition efficiency.

Key words: Corrosion inhibitors, machine learning, auto-encoders, artificial intelligence, multitask learning

INTRODUCTION

Corrosion of metallic structures is a ubiquitous problem in industries such as power generation, oil and gas, pulp and paper, metals processing etc. which also results in significant financial losses [1]. According to the National Association of Corrosion Engineers (NACE) International report, the global cost of corrosion was ~ 2.5 trillion USD in 2013 - close to 3.4 percent GDP of the entire world [2]. The use of corrosion inhibitors is one of the most effective and economical ways to mitigate corrosion of metal and alloy components [3,4]. Corrosion inhibitors are substances that are added in small quantities in corrosive media to protect metal and alloy components from corrosion [5]. Organic

corrosion inhibitors are typically surfactant-type compounds comprising mainly of two parts: (a) a hydrophilic head group containing heteroatoms such as O, N and S and (b) a hydrophobic tail group comprising hydrocarbon groups [6-8]. These compounds form an effective barrier layer on the metal or alloy surface thereby stopping the transport of ions and corrosive species [3,9], thus, inhibiting corrosion. However, most of these inhibitors and their formulations are toxic in nature, and many organizations such as the OSPAR Commission (Oslo/Paris commission for protection of the marine environment of north-east Atlantic) have restricted their use [10]. Therefore, there is a need to develop novel and effective corrosion inhibitors that meet environmental regulations.

To develop new inhibitors, the scientific community has been extensively using experimental, molecular modelling as well as machine learning based approaches [11-13]. Experiments, however, are usually time-consuming and expensive [14]. Computer simulation based molecular modelling approaches such as density functional theory and molecular dynamics have gained popularity in the recent years due to the increase in available computational power. However, these approaches still take long time and as such cannot be used for screening a large number of molecules for potential corrosion inhibitors. Reducing the time for screening large number of molecules will help experimentalists in accelerating the discovery of new inhibitors. Also, molecular modelling techniques cannot predict new and effective inhibitors with high accuracy as they cannot simulate various materials with realistic chemical composition (e.g. steel). Thus, the use of machine learning techniques for screening potential corrosion inhibitor molecules has been suggested [15].

Development and design of novel molecules via machine learning based approaches for applications in the field of materials science and drug discovery has become very popular [16-19]. Availability of large open-source libraries of organic compounds has provided great opportunity to discover effective and novel corrosion inhibitors [20,21]. Several studies in the past have used machine learning algorithms to predict the corrosion inhibition efficiency of molecules, and new molecules were screened based on the predicted inhibition efficiency [22-24]. However, the datasets used in these studies were very small and usually contained a single class of compounds (i.e. lacked chemical diversity) [25]. Thus, models from these studies are not applicable to new compounds from other classes, and different models are needed to address the issue. In this work, we attempt to address the above-mentioned limitations of previous works. The main contributions of our work are:

- 1) A unique end-to-end computational framework for screening of potential molecules for corrosion inhibition
- 2) A novel 'multi-input multi-output neural network' based model using multitask learning for predicting corrosion inhibition efficiency
- 3) Largest experimental dataset (to the best of our knowledge) of corrosion inhibitors for mild steel

COMPUTATIONAL FRAMEWORK

The computational framework used for screening of molecules in this work consists of two phases viz. model development and screening of molecules. The two phases are depicted in Figures 1(a) and 1(b) respectively.

Model development

Model development consists of the steps of data collection and preparation, data pre-processing, data sampling, data normalization, model building and model selection.

Data Collection and Preparation

We have collected corrosion data from literature [26-80] and information available in the public domain to build the dataset of corrosion inhibitors for mild steel in hydrochloric acid. Our database, to the best of our knowledge, is the largest of its kind till date and has 135 organic corrosion inhibitors across

multiple classes such as pyridine, imidazole, Schiff bases, triazole, thiol, amine, thiazole, hydrazone, etc. Figure 2 lists some of the classes of molecules in the dataset. The data collected includes experimental inhibition efficiencies at different temperatures, inhibitor concentrations and acid concentrations obtained from weight loss, Tafel polarization and electrochemical impedance spectroscopy (EIS) methods. Please refer to Table S1 at the end of the paper to find the entire dataset containing inhibition efficiencies for all the 135 molecules obtained using different experimental techniques (weight loss, Tafel and EIS) at various HCl concentrations and temperatures.

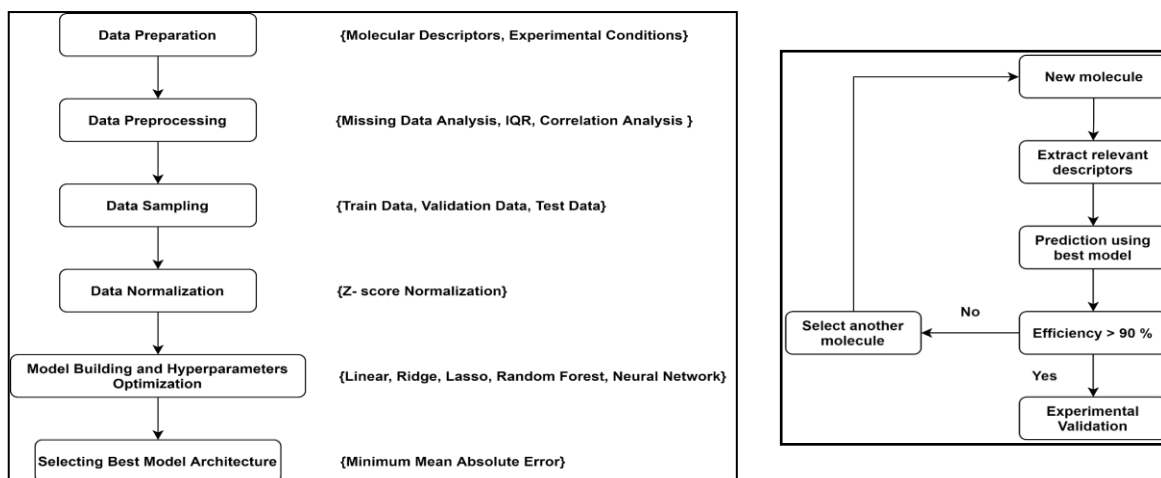


Figure 1 : Computational Framework - (a) Model building(left), (b) Screening of molecules (right)

It should be mentioned that in several publications, the inhibition efficiencies have been calculated using either one or two of the measurement methods. Therefore, we have calculated the average inhibition efficiency (Avg. I.E.) of each molecule corresponding to specific experimental conditions (Equation 1). We refer the average inhibition efficiency, corresponding to a molecule and experimental information as a 'data point'. It was observed from the limited data that efficiency measured using different methods does not vary to a large extent. The average inhibition efficiency is, therefore, used as the target variable for data-driven models. In total, we have 700 data points.

$$\text{Avg. I.E.} = \frac{\sum \text{I.E.}}{n} \quad (\text{Eq. 1})$$

where n corresponds to number of methods

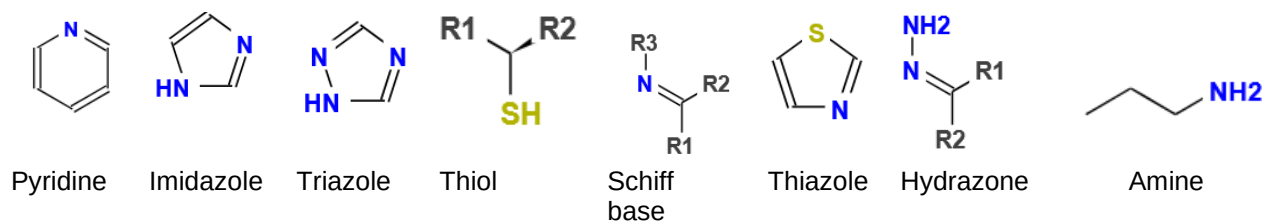


Figure 2: List of few classes of molecules in the dataset.

We then calculated the descriptors for molecules as illustrated in Figure 3. We first created molecular structures in LigParGen* software [81] and obtained smiles notations which were subsequently fed to

* Trade Name

Open babel[†] [82] software to obtain the corresponding 3D structures of the molecules. Structural optimizations were then performed to obtain minimum energy configuration structure for each of these molecules using the Generalized Amber force field (GAFF) [83]. This was followed by generation of 1875 molecular descriptors comprising of constitutional, topological, electronic and geometric features using the open-source PaDEL-descriptor[‡] software [84].

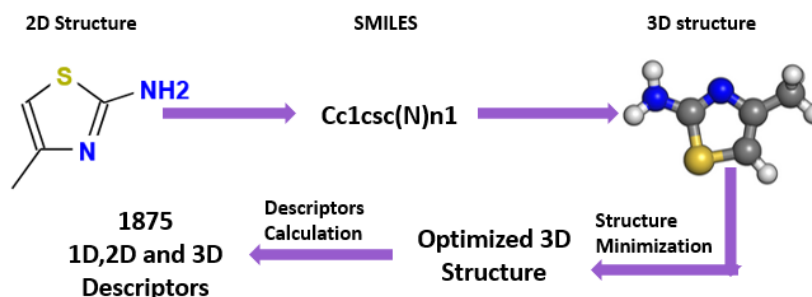


Figure 3: Workflow for molecular descriptors calculations

Data Preprocessing

Molecular descriptor columns containing missing values were removed. The interquartile range (IQR), which is the difference between the 75th and 25th percentiles of each descriptor [85] was then calculated. A descriptor was removed if IQR for that particular descriptor was found to be zero. A zero IQR implies that the descriptor value was constant over the middle 50% entries. Further, correlation analysis was performed to remove highly correlated descriptors as they carry redundant information. We have calculated the pairwise correlation value using the expression:

$$r = \frac{\sum_i^n (\bar{x} - x_i)(\bar{y} - y_i)}{\sqrt{\sum_i^n (\bar{x} - x_i)^2} \sqrt{\sum_i^n (\bar{y} - y_i)^2}} \quad (\text{Eq. 2})$$

where r is Pearson's correlation coefficient, x & y are descriptors and n is the sample size. One of the descriptors was removed if the value of r for a pair of descriptors was greater than 0.95.

Data Sampling and Normalization

The data was arranged in an ascending order according to the target variable i.e. inhibition efficiency. Subsequently, it was divided into training, validation and test datasets such that the probability distribution of target variable in each of them remained similar. This particular way of dividing the dataset is known as 'stratified random sampling' and is considered better than random sampling as it maintains uniform distribution of target variable over training, validation and test datasets [86,87]. We first sampled 20% of the data for testing. From the remaining data, 10% was sampled for validation while the rest was used for training. These training, validation and test datasets were then employed respectively to develop models, tune their hyper-parameters and, finally, evaluate the performance of these models. We normalized the values of descriptors in the training data using the z-normalization formula:

[†] Trade Name

[‡] Trade Name

$$x'_i = \frac{x_i - \mu}{\sigma} \quad (\text{Eq. 3})$$

Where x'_i is a normalized value, x_i is the original value, μ is mean and σ is the standard deviation of a descriptor. Using the mean and standard deviation of a descriptor, we scaled the values in validation and test data.

Model Architecture

We pose the problem of regression as a ‘multitask learning problem’ by forcing the neural network model to learn reconstruction of the input data as well as predicting the target output. We have implemented input reconstruction and regression tasks into one framework via autoencoder and prediction network respectively. Auto-encoders have two parts, namely encoder and decoder, former encodes data into latent space while latter decodes it [88]. They learn an effective low dimensional representation (encoding) for a set of high dimensional data, aiding in dimensionality reduction.

The architecture has two inputs and two outputs. Molecular descriptors and experimental conditions were fed as two separate input branches at two different locations in the Multi-Input Multi-Output (MIMO) architecture. Similarly, prediction of inhibition efficiency and reconstruction of molecular descriptors are two main objectives of our model architecture. Loss function used in the training was the weighted combination of both reconstruction loss and inhibition efficiency prediction loss. We used 4 dense layers in prediction network (64, 32, 16, 1) and 3 dense layers each in encoder (128, 64, 64) and decoder (64, 128, 647). Figure 4 shows the architecture of the MIMO model.

We have introduced dropout layers in the encoder to regularize our model. In all hidden layers, rectified linear units (ReLU) [89] were used. Linear activation was used at the reconstruction output layer whereas sigmoid activation was used at the inhibition efficiency output layer. We concatenated experimental conditions with latent space encodings of molecular descriptors and predicted the inhibition efficiency. This model is also useful in cases where you have different types of inputs and have multi-optimization task to address. The cost function used to train multitask architecture i.e. MIMO is shown in Equation 4

$$\text{cost function} = \lambda_1 \times MSE_{\text{target}} + \lambda_2 \times MSE_{\text{autoencoder}} \quad (\text{Eq. 4})$$

Where,

$$MSE_{\text{target}} = 1/M \sum_{i=1}^M (y_i - \hat{y}_i)^2 \quad (\text{Eq. 5})$$

And,

$$MSE_{\text{autoencoder}} = 1/M \sum_{i=1}^M (X_i - \hat{X}_i)^2 \quad (\text{Eq. 6})$$

Here, y_i is the actual value and $\hat{y}_i$ is the predicted value; X_i is the input descriptor vector and $\hat{X}_i$ is reconstructed descriptor vector; λ_1 and λ_2 are hyperparameters; and M is the number of data points

Dropout rates were varied as 0.1, 0.2, 0.3 and 0.4. The same dropout rate was applied to all dropout layers. We also took $\lambda_1 : \{1\}$, $\lambda_2 : \{1, 10, 50, 100, 200\}$. A batch size of 32 was used for training the model. Adam optimizer [90] was used to train the network. Learning rate scheduler was used with patience of 500 epochs. The initial learning rate value was set to 0.001. The model was trained for maximum 2500 epochs. Best model was chosen based on the minimum value of the cost function.

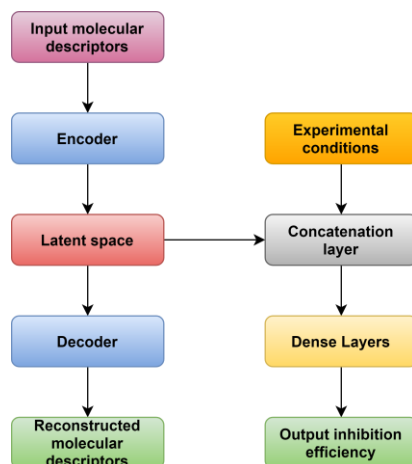


Figure 4: Architecture for Multitask Learning (multi-input multi-output) model. Left block represents autoencoder while right block represents prediction network

Screening of molecules

To screen a new molecule, the relevant descriptors were first extracted and then, the inhibition efficiency was predicted using the best model. If the predicted inhibition efficiency exceeds 90%, experiments can be done to further validate the results. The threshold value of 90% can be adjusted as per requirement.

RESULTS

Evaluating various models

The unprocessed corrosion data consists of 700 rows and 1875 columns of descriptors apart from experimental conditions. After missing data analysis, the molecular descriptors were reduced to 1496. Further, after IQR analysis, the number of descriptors were reduced to 1327 and following removal of highly correlated descriptors, the count was further reduced to 647. Hence, after data preprocessing, we selected 647 descriptors out of the original 1875 descriptors.

With 647 molecular descriptors, models were constructed with both traditional algorithms such as Linear regression, lasso regression [91], ridge regression [92], random forest [93] as well as our novel MIMO architecture. For traditional algorithms, experimental conditions were incorporated as input features along with molecular descriptors. In MIMO architecture, the dimensionality of molecular descriptors was reduced from 647 to 64 using autoencoder and learned encodings in the latent space were fed to prediction network. Table 1 and Table 2 show the optimized hyper-parameters and the performance of various machine learning models. Linear regression performed the worst. The MAE on training data was found to be 7.77, but errors on validation and test data were extremely high (order of 10^{10}) proving to be a clear case of over-fitting. Ridge and lasso regressions model showed improvement over linear regression with MAE values on test data of 10.78 and 10.75. They are regularized regression techniques and in general, regularization tries to find the optimum model by avoiding both under-fitting and over-fitting. However, tree based nonlinear random forest model outperformed the linear models with MAE value of 7.06 on test data. The reason for this may be non-linear relationship of descriptors and experimental variables with experimental inhibition efficiency. However, our multi-tasking network, MIMO, performed the best. The MAE value on test data was found to be 6.46. Figure 5 shows the parity plot for prediction of inhibition efficiency using the MIMO model. From the parity plot, it can be observed that the majority of points lie close to the iso-line (within $\pm 10\%$ error) indicating the excellent ability of the MIMO model to predict the corrosion inhibition efficiencies.

Table 1
Optimized parameters for various machine-learning based algorithms

Model	Optimal values of hyper-parameters
Linear	No hyper-parameter
Ridge	$\lambda = 1.12\text{E-}06$
Lasso	$\lambda = 0.003069$
Random Forest	N_estimators=50, max_features='auto', min_samples_split=2, min_samples_leaf=1, bootstrap='True'
MIMO	batch size=32, dropout rate=0.1, $\lambda_1=1$, $\lambda_2=50$, epochs=2500

Table 2
Mean absolute error (MAE) of different machine learning algorithms on train, validation and test datasets using stratified random sampling of data points

Model	Train MAE	Validation MAE	Test MAE
Linear	7.77	4.8E+10	2.1E+11
Ridge	7.77	10.92	10.78
Lasso	7.79	10.85	10.75
Random forest	2.82	6.74	7.06
MIMO	1.96	5.03	6.46

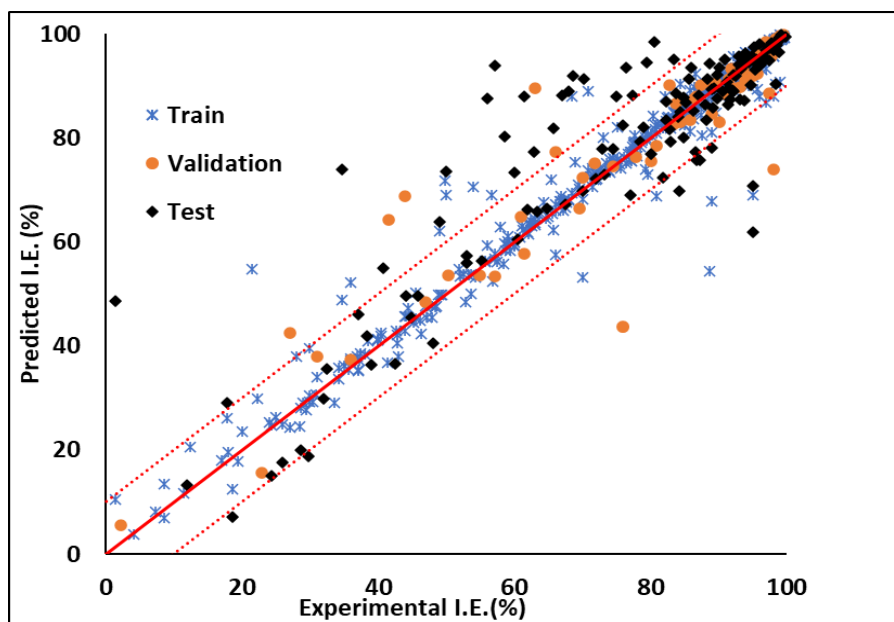


Figure 5: Parity plot of predicted I.E Vs experimental I.E. Red line represents iso-line where predicted I.E. is equal to experimental IE. IE represents Inhibition efficiency. Dotted lines parallel to iso-line represents the error bound of $\pm 10\%$.

There are several studies in the domain of virtual drug screening and computer vision that have also found multitask learning based framework outperforming single task based traditional machine learning algorithms [94-100]. For instance, Wenzel et al. used the multitask learning to predict drug related properties and observed an improvement in the r-squared value to 0.7 from 0.6 compared to single task learning [96]. The additional benefit of using multitask learning over single task learning was that it prevents the network to focus on particular task and avoid over-fitting [98]. The random-forest based classifiers have been most commonly used for QSAR predictions as they generally avoid over-fitting

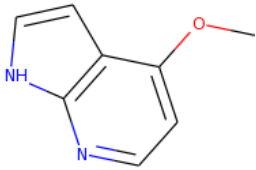
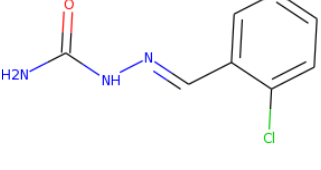
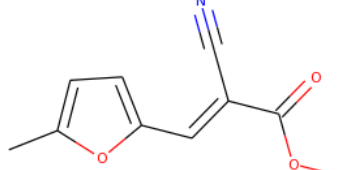
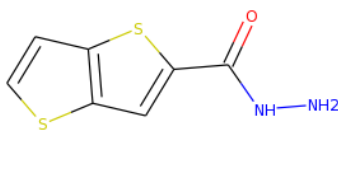
data even in case of data scarcity [98]. However, Dahl et al. found neural nets with multitasking perform better than random forest classifier [98].

The approach used in this framework considers effective contribution from the entire molecule encompassing both head and tail groups obtained through the calculation of molecular descriptors. The head group provides protection from anodic dissolution by bonding with metal surface while the tail group provides protection by restricting species that can affect cathodic reaction [7]. Inclusion of alkyl tail group effect can be verified from the example of isoxazolidines [73]. Experimentally at 100 ppm in 1M HCl, this compound derivative without alkyl tail (tail R = H) showed I.E. = 53.6 %, while derivatives with tail of higher chain lengths (tail R = (CH₂)₉CH₃) showed I.E. = 98.2%. This trend was captured by the model with predicted I.E. of 50.9% and 96.7% for tails with R=H and R=(CH₂)₉CH₃ respectively. Inclusion of effect of head group can be identified from the example of benzimidazole derivatives [74]. The experimental I.E. for benzimidazole (BI) and 2-mercaptobenzimidazole (MBI) at 250 ppm at 298K in 1 M HCl were 63 % and 89.6% respectively while corresponding predicted values were 63.6% and 88.1% respectively. Sulfur atom in MBI provides extra bonding effect to metal surface, and same has been captured in the modelling.

Screening of new molecules using MIMO

Since, the MIMO model had the least MAE on test data, we used the model for screening new molecules for their corrosion inhibition efficiency. We chose a subset of molecules from ZINC15 dataset [100,102] for screening, with the experimental conditions as 1M HCl, 25°C temperature and 100-ppm inhibitor concentration. Table 3 shows a few molecules from the dataset along with their predicted inhibition efficiency values. Among the screened molecules, we found that the molecule with ZINC id – ZINC000001532777 (cinnamaldehyde) is already a well-known corrosion inhibitor (Figure 6) [103]. For this molecule, our model successfully predicted a high inhibition efficiency of 92.80 %, in accordance with experimental findings. Apart from cinnamaldehyde, our model also predicts promising inhibition efficiency for several other molecules from the ZINC15 dataset (as shown in Table 3) viz. 4-methoxy-1H-pyrrolo[2,3-b]pyridine, 2-chlorobenzaldehyde semicarbazone, methyl 2-cyano-3-(5-methylfuran-2-yl) propenoate, and thieno[3,2-b]thiophene-2-carbohydrazide. Of course, these need to be further tested experimentally to validate our predictions.

Table 3
Organic compounds screened by our computational framework as novel corrosion inhibitors
(Zinc ID, Predicted I.E., Chemical Formula)

			
ZINC00000870035 8 I.E. – 96.68% {4-Methoxy-1H-pyrrolo[2,3-b]pyridine}	ZINC000001041730 I.E– 96.11% {2-chlorobenzaldehyde semicarbazone}	ZINC000000043644 I.E. – 95.03% {Methyl 2-cyano-3-(5-methylfuran-2-yl) propenoate}	ZINC000000108027 I.E. - 95.32% {Thieno[3,2-b]thiophene-2-carbohydrazide}

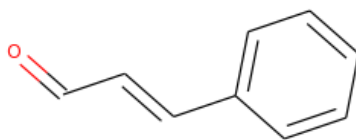


Figure 6: ZINC000001532777 or cinnamaldehyde I.E. =92.80%

A machine-learning based modelling and screening framework was proposed and demonstrated in this work. Currently this model has limitation with respect to the solvent used in the corrosion inhibitor formulation to only HCl and does not consider other constituents such as acidifying agents etc. which might have a considerable effect on its efficacy. However, it can be a powerful tool to screen potential corrosion inhibitors and can significantly help researchers and industries in effectively combating corrosion. The effectiveness of our framework relies on good amount of reliable data. Hence, data obtained from high-throughput experiments for model building would be beneficial. After the experiments on screened compounds, the new data obtained can be added to the database to enrich it. Having increased amount of data will certainly help in making robust models. A coupling of hardware (i.e. experimental setup) and software (i.e. computational framework) to do experiments and model building in an iterative automatic manner is the way forward for material discovery.

CONCLUSIONS

In this work, we have proposed an end-to-end computational framework based on multi-task learning architecture for accelerating the screening of potential corrosion inhibitors for steels. Our work shows that a single model is able to predict the corrosion inhibition efficiency of multiple classes of molecules. Our dataset is the largest known for mild steel inhibitors in HCl. Our deep learning architecture based on multi-task learning solves two tasks simultaneously - reconstructing the inputs and predicting the inhibition efficiency. We have also identified a few novel corrosion inhibitors from ZINC15 datasets. One of the proposed molecules i.e. ZINC000001532777 or cinnamaldehyde, predicted by our model is known to be an excellent corrosion inhibitor for steels in HCl. However, additional experiments need to be done to cross validate the effectiveness of the other predicted molecules for corrosion inhibition application. In future, data for many more molecules can be added to augment our dataset for improving the robustness of the model.

ACKNOWLEDGEMENTS

The authors thank Dr Venkataramana Runkana – Chief Scientist at Tata Consultancy Services (TCS) and Mr. K. Ananth Krishnan - Chief Technology Officer at Tata Consultancy Services (TCS), for the funding and encouragement towards this program.

REFERENCES

1. Y. M. Panchenko, A. I. Marshakov, T. N. Igonin, V. V. Kovtanyuk, L. A. Nikolaeva, Long-term forecast of corrosion mass losses of technically important metals in various world regions using a power function, *Corrosion Science* 88 (2014) 306-316. <https://doi.org/10.1016/j.corsci.2014.07.049>
2. NACE International Technical Report No OAPUS310GK0CH (AP110272), "International Measures of Prevention, Application, and Economics of Corrosion Technologies Study", (Houston, TX: NACE International, 2016)
3. B. E. A. Rani, B. B. J. Basu, Green Inhibitors for Corrosion Protection of Metals and Alloys: An Overview, *International Journal of Corrosion* (2012) 1-15. <https://doi.org/10.1155/2012/380217>
4. William S. Tait, *Handbook of Environmental Degradation of Materials* (Third Edition), (2008), pp. 583-600

5. J. de Damborenea, A. Conde, M.A. Arenas, Rare Earth-Based Corrosion Inhibitors, Woodhead Publishing Series in Metals and Surface Engineering 2014, pp. 84-116
6. E. S. H. E. Ashry, A. E. Nemra, S. A. Esawy, S. Ragab, Corrosion inhibitors Part II: Quantum chemical studies on the corrosion inhibitions of steel in acidic medium by some triazole, oxadiazole and thiadiazole derivatives, *Electrochimica Acta* 51 (2006) 3957–3968. <https://doi.org/10.1016/j.electacta.2005.11.010>
7. J. M. D. Olivo, S. Cao, D. Yong, B. Brown, S. Nestic, Effect of Corrosion Inhibitor Head Group on the Electrochemical Processes Governing CO₂ Corrosion, *CORROSION* 2020, paper no-14925, Houston, TX: NACE, 2020)
8. E. McCafferty, *Introduction to corrosion science*, (New York, NY: Springer, 2010), pp. 359-400
9. B. N. Popov, Corrosion Engineering, Principles and Solved Problems, 2015, pp. 581-597
10. M. Finšgar, J. Jackson, Application of corrosion inhibitors for steels in acidic media for the oil and gas industry: A review, *Corrosion Science* 86 (2014) 17-41. <https://doi.org/10.1016/j.corsci.2014.04.044>
11. A. Lgamri, H. A. E. Makarim, A. Guenbour, A. B. Bachir, L. Aries, S. E. Hajjaji, Electrochemical study of the corrosion behaviour of iron in presence of new inhibitor in 1M HCl, *Progress in Organic Coatings* 48 (2003) 63–70. [https://doi.org/10.1016/S0300-9440\(03\)00116-4](https://doi.org/10.1016/S0300-9440(03)00116-4)
12. Z. Hajiahmadi, Z. Tavangar, Extensive theoretical study of corrosion inhibition efficiency of some pyrimidine derivatives on iron and the proposal of new inhibitor, *Journal of Molecular Liquids* 284 (2019) 225-231. <https://doi.org/10.1016/j.molliq.2019.04.009>
13. Y. Liu, Y. Guo, W. Wu, Y. Xiaong, C. Sun, L. Yuan, M. Li, A Machine Learning-Based QSAR Model for Benzimidazole Derivatives as Corrosion Inhibitors by Incorporating Comprehensive Feature Selection, *Interdisciplinary Sciences: Computational Life Sciences* 11 (2019) 738–747. <https://doi.org/10.1007/s12539-019-00346-7>
14. 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, *Corrosion Science* 99 (2015) 76-88. <https://doi.org/10.1016/j.corsci.2015.06.003>
15. A. Cook, G. Frankel, A. Davenport, T. Hughes, S. Gibbon, D. Williams, H. Bluhm, V. Maurice, S. Lyth, P. Marcus, D. Shoesmith, C. Wren, J. Wharton, G. Hunt, S. Lyon, T. Majchrowski, R. Lindsay, G. Williams, B. R. Oller, M. Todorova, S. Nixon, S. T. Cheng, J. Scully, A. Wilson, F. Renner, Y. H. Chen, C. Taylor, H. Habazaki, A. Michaelides, S. Morsch, P. Visser, L. Kyhl, A. Kokalj, Corrosion control: general discussion, *Faraday Discussion*, 180 (2015) 543–576, <https://doi.org/10.1039/C5FD90047F>
16. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, *Nature* 559 (2018) 547–555. <https://doi.org/10.1038/s41586-018-0337-2>
17. O. A. V. Lilienfeld, Quantum Machine Learning in Chemical Compound Space. *Angewandte Chemie International Edition* 57 (2018) 4164–4169. <https://doi.org/10.1002/anie.201709686>
18. L. Zhang, J. Tan, D. Han, H. Zhu, From machine learning to deep learning: progress in machine intelligence for rational drug discovery, *Drug Discovery Today* 22 (2017) 1680-1685. <https://doi.org/10.1016/j.drudis.2017.08.010>
19. J. M. Granda, L. Donina, V. Dragone, D. L. Long, L. Cronin, Controlling an organic synthesis robot with machine learning to search for new reactivity, *Nature* 559 (2018) 377–381. <https://doi.org/10.1038/s41586-018-0307-8>
20. M. G. Medina, J. J. Naveja, N. S. Cruz, J. L. M. Franco, Open chemoinformatic resources to explore the structure, properties and chemical space of molecules, *RSC Advances* 7 (2017) 54153-54163. <https://doi.org/10.1039/C7RA11831G>
21. J. L. Reymond, M. Awale, Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database, *ACS Chemical Neuroscience* 3 (2012) 649-657. <https://doi.org/10.1021/cn3000422>
22. 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,” *Corrosion Science* 99 (2015) 76-88. <https://doi.org/10.1016/j.corsci.2015.06.003>