Electronic Supplementary Information

A data-driven QSPR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques

Thanh Hai Pham*[abc], Phung K. Le [ab], Do Ngoc Son*[ab]

[a] *Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam.*
[b] *Vietnam National University Ho Chi Minh City, Linh Trung Ward, Ho Chi Minh City, Vietnam.*
[c] *Vietnam Institute for Tropical Technology and Environmental Protection, 57A Truong Quoc Dung Street, Phu Nhuan District, Ho Chi Minh City 70000, Vietnam.*

** Corresponding authors. Email addresses: pthai.sdh21@hcmut.edu.vn (Thanh Hai Pham); dnson@hcmut.edu.vn (Do Ngoc Son)*

Table S1. List of 208 2D descriptors in 2Ddes feature set.

Groups	Number of descriptors	Name of descriptors
Partial charge	4	MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge
Molecular property	4	MolLogP, MolMR, MolWt, ExactMolWt
Topological and connectivity	22	BalabanJ, BertzCT, HallKierAlpha, Ipc, Kappa1, Kappa2 , Kappa3, Chi0, Chi1, Chi0n, Chi1n, Chi2n, Chi3n, Chi4n, Chi0v, Chi1v, Chi2v, Chi3v, Chi4v, FpDensityMorgan1, FpDensityMorgan2, FpDensityMorgan3
Lipinski	21	HeavyAtomCount, HeavyAtomMolWt, NHOHCount, NOCount, NumHAcceptors, NumHDonors, NumHeteroatoms, NumRotatableBonds, NumValenceElectrons, NumRadicalElectrons, NumAromaticRings, NumSaturatedRings, NumAliphaticRings, NumAromaticHeterocycles, NumSaturatedHeterocycles, NumAliphaticHeterocycles, NumAromaticCarbocycles, NumSaturatedCarbocycles, NumAliphaticCarbocycles, RingCount, FractionCSP3
MOE-type	59	TPSA, LabuteASA, PEOE_VSAn (n=1-14), SMR_VSAn (n=1-10), SlogP_VSAn (n=1-12), Estate_VSAn (n=1-11), VSA_EStaten (n=1-10)
BCUT2D	8	'BCUT2D_MWHI', 'BCUT2D_MWLOW', 'BCUT2D_CHGHI', 'BCUT2D_CHGLO', 'BCUT2D_LOGPHI', 'BCUT2D_LOGPLOW', 'BCUT2D_MRHI', 'BCUT2D_MRLOW'
EState	4	MaxEStateIndex, MinEStateIndex, MaxAbsEStateIndex, MinAbsEStateIndex
Constitutional	85	fr_X (X=Al_COO, Al_OH, …)
QED	1	qed

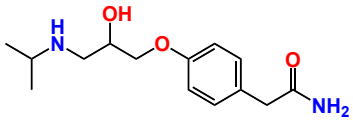
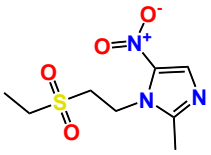
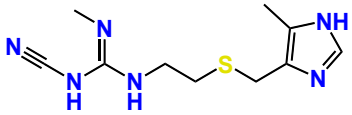
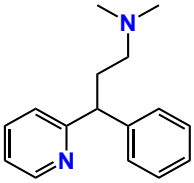
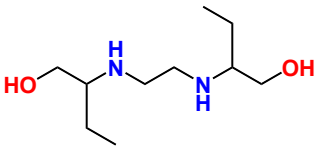
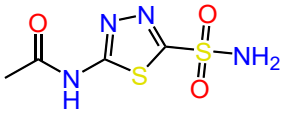
Table S2. The optimal values of hyper-parameters for other considered models.

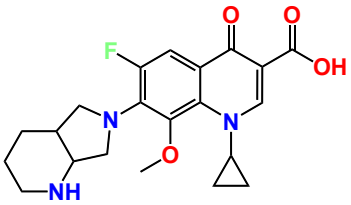
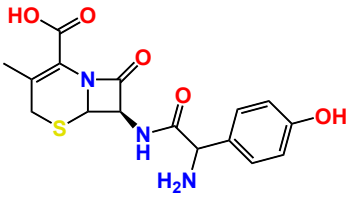
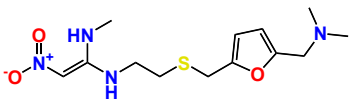
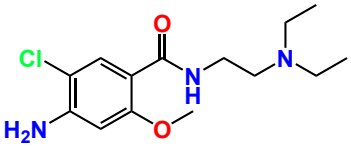
Model	n_estimators	min_samples_split	min_samples_leaf	max_depth
GB/rdkit2ddes_ 5	100	4	1	3
GB/rdkit2ddes_ 10	400	2	3	5
GB/rdkit2ddes_ 15	100	2	5	4
GB/rdkit2ddes_ 20	200	2	4	3
GB/rdkit2ddes_ 25	100	2	4	5
GB/rdkit2ddes_ 30	600	2	4	3
GB/rdkit2ddes_ 35	200	2	5	4
GB/rdkit2ddes_ 45	600	2	5	3
GB/rdkit2ddes_ 50	600	2	5	4
GB/rdkit2ddes_ 55	600	2	5	3
GB/rdkit2ddes_ 60	200	2	4	5
GB/rdkit2ddes_ 65	600	2	5	4
GB/rdkit2ddes_ 70	600	2	4	3
GB/rdkit2ddes_ 80	600	2	4	3
GB/rdkit2ddes_ 90	400	2	4	3
GB/rdkit2ddes_ 100	600	2	5	4
GB/rdkit2ddes_ 110	600	2	5	3
GB/rdkit2ddes_ 120	600	2	5	4
GB/rdkit2ddes_ 130	600	2	5	4

Table S3. Comparison of the characteristics and predictive performance of models on published datasets.

Dataset	Model	Input features	Validation method	RMSE (%)	Ref.
PQ-41 (41 pyridines and quinolines)	GA-NN	Molecular descriptors and adsorption energies	5-fold CV	16.74	Ser et al. [1]
	GB/2Ddes	208 2D descriptors	5-fold CV	19.67	This work
	GB/2Ddes10	Top 10 2D descriptors	5-fold CV	14.44	This work
P-20 (20 pyridazines)	NN	5 selected descriptors	5-fold CV	14.69	Quadri et al. [2]
	GB/2Ddes	208 2D descriptors	5-fold CV	9.80	This work
	GB/2Ddes5	Top 5 2D descriptors	5-fold CV	8.89	This work
IL-30 (30 ionic liquids)	NN	5 selected descriptors	5-fold CV	10.01	Quadri et al. [3]
	GB/2Ddes	208 2D descriptors	5-fold CV	6.92	This work
	GB/2Ddes10	Top 10 2D descriptors	5-fold CV	6.08	This work
CO-270 (270 cross-category organic compounds)	3L–DMPNN	Atomic, bond, and molecular descriptors	10-fold CV	7.82	Dai et al. [4]
	GB/2Ddes	208 2D descriptors	10-fold CV	6.72	This work
	GB/2Ddes50	Top 50 2D descriptors	10-fold CV	6.15	This work

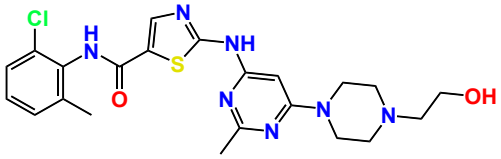
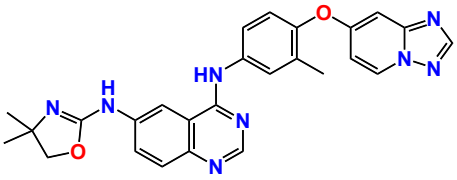
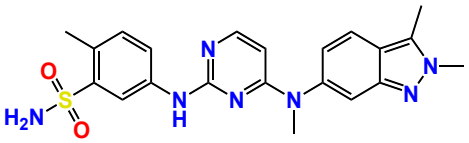
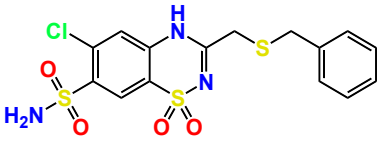
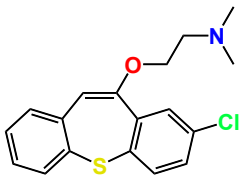
Table S4. Comparison of predicted IE with the published experimental values for ten drug compounds.

ID	Medicine name	Structure	Exp. condition*	Exp. IE (%)	Pre. IE (%)
T1	Atenolol[5]		HCl 1 M, 0.94 mM, EIS	90.9	91.4
T2	Tinidazole[6]		HCl 1 M, 1.21 mM, EIS	87.8	92.4
T3	Cimetidine[7]		HCl 1 M, 0.79 mM, EIS	95.5	91.2
T4	Pheniramine[8]		HCl 0,5 M, 0.83 mM, EIS	86.4	86.5
T5	Ethambutol[9]		HCl 2 M, 0,98 mM, EIS	86.4	82.9
T6	Acetazolamide[10]		HCl 1 M, 0.90 mM, EIS	88.8	85.3

T7	Moxifloxacin[11]		HCl 1 M, 0.75 mM, EIS	92.0	87.5
T8	Cefadroxil[12]		HCl 1 M, 1.00 mM, Tafel	84.1	84.0
T9	Ranitidine[13]		HCl 1M, 1.27 mM, EIS	92.0	93.0
T10	Metoclopramide[14]		HCl 1 M, 1.00 mM, EIS	94.0	94.1

*The experimental conditions include solution, inhibitor concentration, and method for determining IE, respectively.

Table S5. Predicted IEs of the top 5 drug compounds with the highest predicted IEs

ID	Medicine name	Molecular structure	Predicted IE (%)
B1	Dasatinib		98.5
B2	Tucatinib		98.0
B3	Pazopanib		97.5
B4	Benzthiazide		97.4
B5	Zotepine		97.3

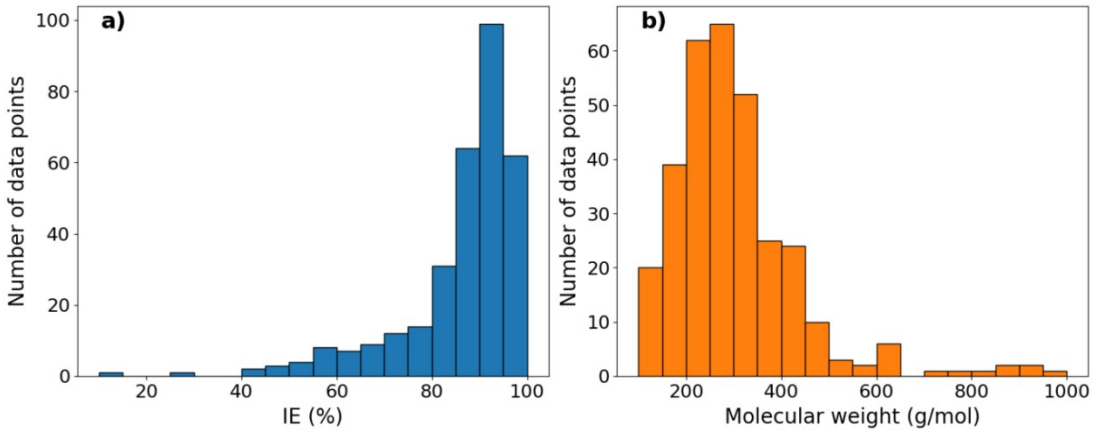


Figure S1. Distributions of the (a) IE values and (b) molecular weights of corrosion inhibitors in the Fe-HCl-317 dataset.

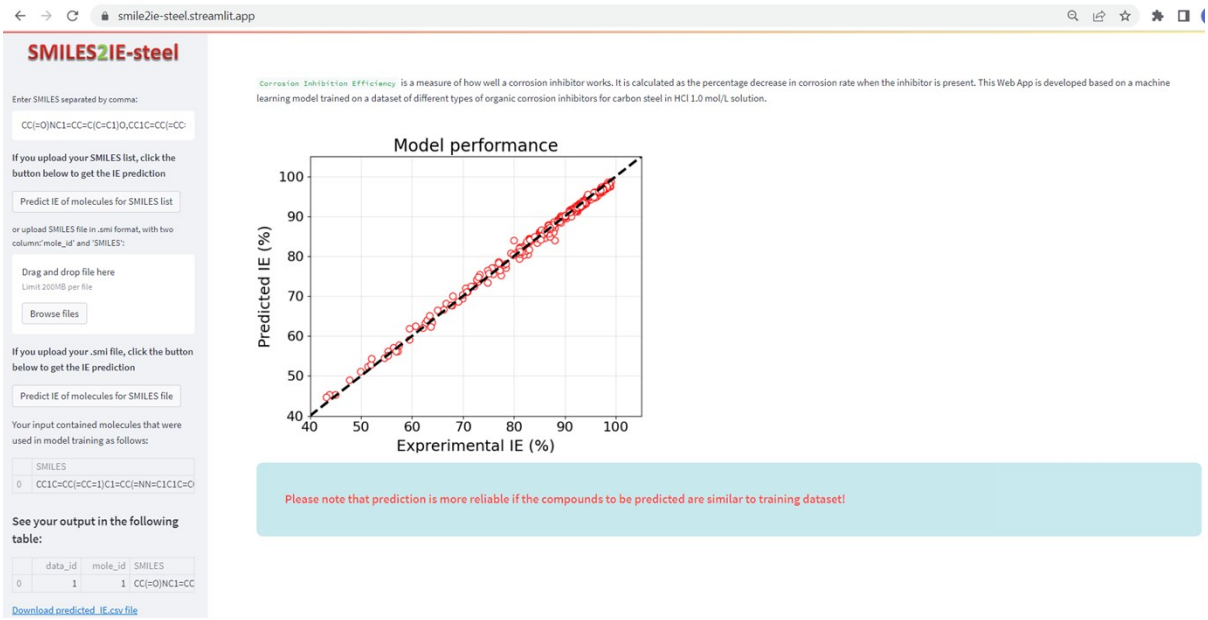


Figure S2. The interface of SMILES2IE-steel web tool

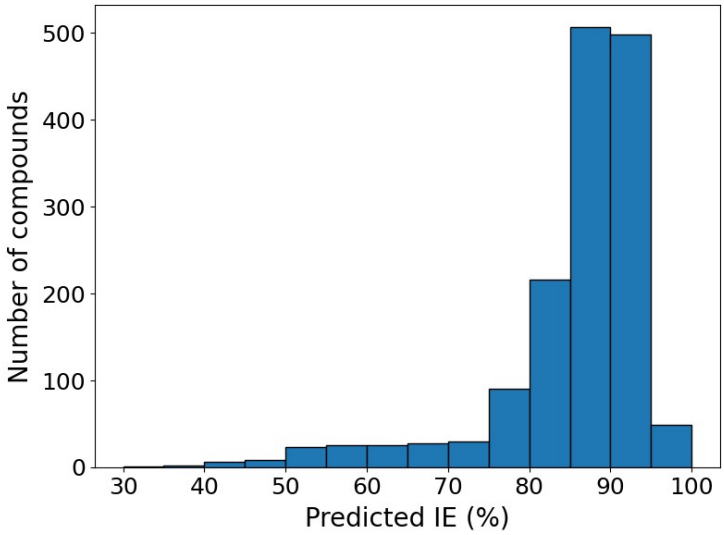


Figure S3. Distribution of number of compounds according to the predicted IE

REFERENCES

1 C. T. Ser, P. Žuvela and M. W. Wong, *Appl. Surf. Sci.*, 2020, **512**, 145612.

2 T. W. Quadri, L. O. Olasunkanmi, E. D. Akpan, O. E. Fayemi, H.-S. Lee, H. Lgaz, C. Verma, L. Guo, S. Kaya and E. E. Ebenso, *Mater. Today Commun.*, 2022, **30**, 103163.

3 T. W. Quadri, L. O. Olasunkanmi, O. E. Fayemi, E. D. Akpan, H.-S. Lee, H. Lgaz, C. Verma, L. Guo, S. Kaya and E. E. Ebenso, *Comput. Mater. Sci.*, 2022, **214**, 111753.

4 J. Dai, D. Fu, G. Song, L. Ma, X. Guo, A. Mol, I. Cole and D. Zhang, *Corros. Sci.*, 2022, **209**, 110780.

5 G. Karthik and M. Sundaravadivelu, *Egypt. J. Pet.*, 2016, **25**, 183–191.

6 I. Reza, A. Saleemi and S. Naveed, *PJCT*, 2011, **13**, 67–71.

7 A. Singh, A. Gupta, A. K. Rawat, K. R. Ansari, M. A. Quraishi and E. E. Ebenso, *Int. J. Electrochem. Sci.*, 2014, **9**, 7614–7628.

8 I. Ahamad, R. Prasad and M. A. Quraishi, *Corros. Sci.*, 2010, **52**, 3033–3041.

9 S. Dahiya, N. Saini, N. Dahiya, H. Lgaz, R. Salghi, S. Jodeh and S. Lata, *Port. Electrochim. Acta*, 2018, **36**, 213–230.

10 L. P. Chaudhari and S. N. Patel, *J. Bio- Tribo-Corrosion*, 2019, **5**, 20.

11 A. S. Fouda, K. Shalabi and A. E-Hossiany, *J. Bio- Tribo-Corrosion*, 2016, **2**, 18.

12 S. K. Shukla and M. A. Quraishi, *Mater. Chem. Phys.*, 2010, **120**, 142–147.

13 R. S. A. Hameed, *Port. Electrochim. Acta*, 2011, **29**, 273–285.

14 Z. Golshani, S. M. A. Hosseini, M. Shahidizandi and M. J. Bahrami, *Mater. Corros.*, 2019, **70**, 1862–1871.