Supporting Information: Predicting corrosion inhibition efficiencies of small organic molecules using data-driven techniques

Xuejiao Li,[1,*] Bahram Vaghefinazari,[1] Tim Würger,[1,2] Sviatlana V. Lamaka,[1] Mikhail L. Zheludkevich,[1,3] and Christian Feiler[1,†]

[1] *Institute of Surface Science, Helmholtz-Zentrum Hereon, Geesthacht, Germany*
[2] *Institute of Polymers and Composites, Hamburg University of Technology, Hamburg, Germany*
[3] *Institute for Materials Science, Faculty of Engineering, Kiel University, Kiel, Germany*

I. PREDICTION ERROR AND NUMBER OF FEATURES

TABLE I. At each step, the averaged RMSEs of the test and train sets corresponding to the probability with the lowest RMSE of the test set in the 5-fold cross validations are listed in the table. The values marked in red are the averaged RMSEs of the optimized SVR and KRR models respectively.

Num. features	SVR_test	SVR_train	KRR_test	KRR_train
25	60.61	7.75	60.61	7.72
24	58.79	7.73	58.76	7.71
23	57.56	7.61	57.13	7.72
22	56.57	7.60	56.11	7.98
21	55.64	7.62	55.00	7.96
20	54.66	7.63	53.94	8.01
19	52.49	7.74	52.75	8.38
18	51.58	7.77	51.61	8.42
17	50.81	7.68	50.63	8.58
16	49.98	7.71	49.54	9.01
15	49.57	7.74	48.15	9.71
14	49.08	7.74	47.08	9.86
13	47.54	7.84	46.62	10.57
12	46.76	7.88	46.19	10.75
11	46.56	7.84	45.36	10.89
10	46.74	7.89	46.21	12.54
9	45.90	8.47	47.61	14.22
8	45.20	9.35	46.80	18.55
7	45.98	10.02	46.40	22.15
6	48.97	21.01	49.44	30.72
5	50.45	23.31	56.61	23.34
4	49.94	26.53	57.98	30.23
3	47.08	29.59	64.62	38.09
2	56.50	50.11	66.24	44.06
1	63.94	63.12	72.83	70.32

* xuejiao.li@hereon.de
† christian.feiler@hereon.de

II. P-VALUE BETWEEN FEATURES AND TARGET

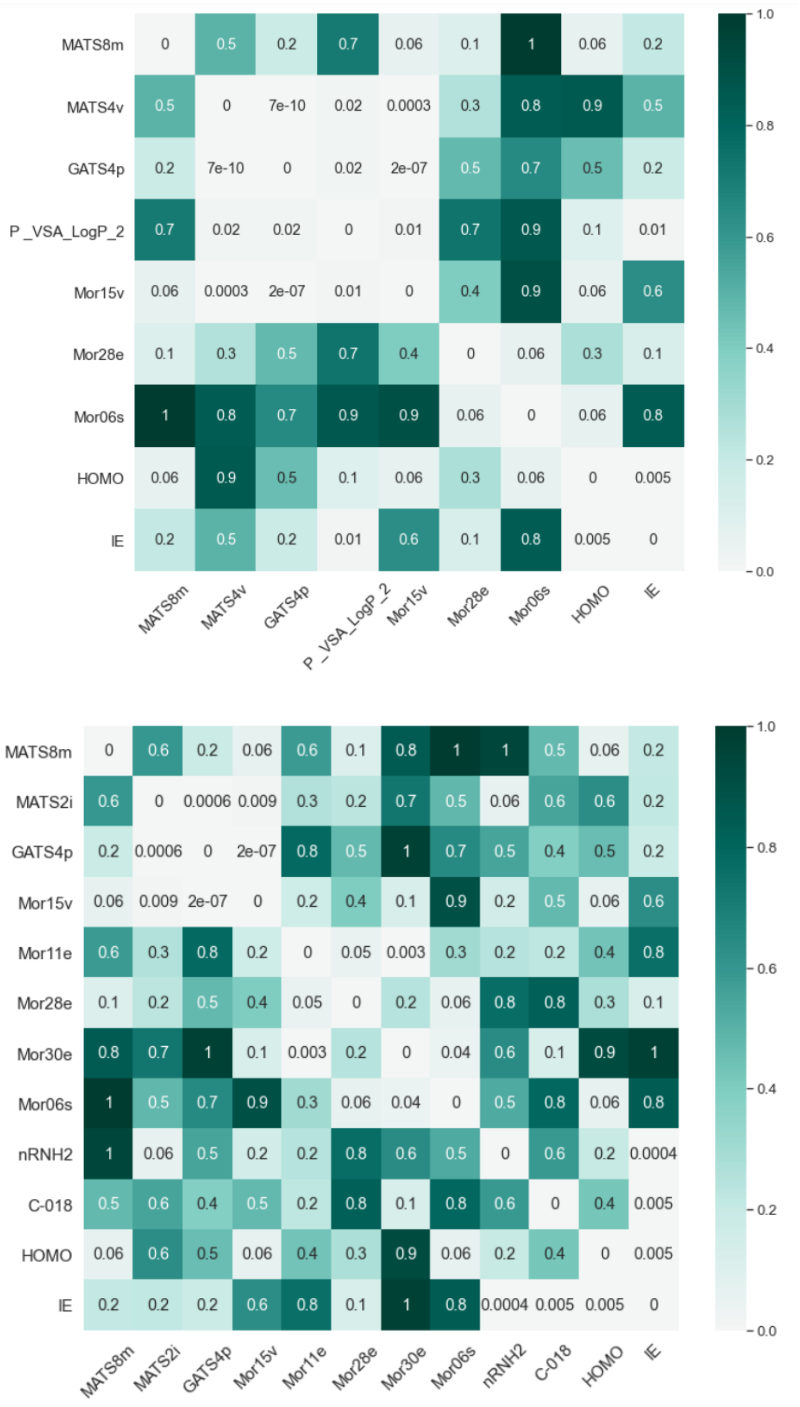

FIG. 1. p-value calculation for the two models. (**top**) p-value among the selected 8-tuple features for the SVR model and IEs. (**bottom**) p-value among the selected 11-tuple features for the KRR model and IEs.

III. EXPERIMENTAL DATABASE

TABLE II: Names and IEs of the 58 chemicals in the data set extracted from ref. 7

Name	IE(%)
5-Amino-salicylic acid	80
5-Methyl-salicylic acid	56
5-Nitrobarbiturc acid	71
5-Sulfosalicylic acid	-125
2,2'Bipyridyl	47
1,2,4-Triazole	-43
Acetic acid	60
Adipic acid	-48
Ascorbic acid	38
Benzoic acid	40
Bicine	-227
Citric acid	-78
L-Cysteine	-206
Diglycolic acid	54
Dodecylbenzene sulfonic acid	67
Formic acid	19
Fumaric acid	83
Glucose	29
L-Glutamic acid	-84
Glyceric acid	57
Glycerol	24
Glycine	-82
Glycolic acid	49
Kojic acid	-90
Maleic acid	26
Maltol	-168
Mandelic acid	11
Nicotinic acid	24
DL-Norleucine	-127
Oxalic acid	63
D-Panthenol	-27
DL-Phenylalanine	-149
Phthalic acid	33
Picolinic acid	45
Piperazine	-106
Propionic acid	34
Para-Toluic acid	-39
Pyrazole	31
2,3-Dipicolinic acid	47
2,5-Dipicolinic acid	68
Quinaldic acid	74
Quinic acid	37
Salicylaldehide	18
Salicylhydroxamic acid	-177
Salicylic acid	26
Tartaric acid	52
Thiourea	57
Uracil	-151

Urea	-19
3,5-Dimethylpyrazole	-27
3,5-Dinitrosalicylic acid	15
3-Amino-1,2,4-triazole	19
3-Methyl-2-nitrobenzoic acid	25
3-Methylcatechol	-129
3-Methylsalicylic acid	52
4-Amino-salicylic acid	52
4-Methylsalicylic acid	9
4-Tert-butylbenzoic acid	-91

IV. LEARNING CURVES

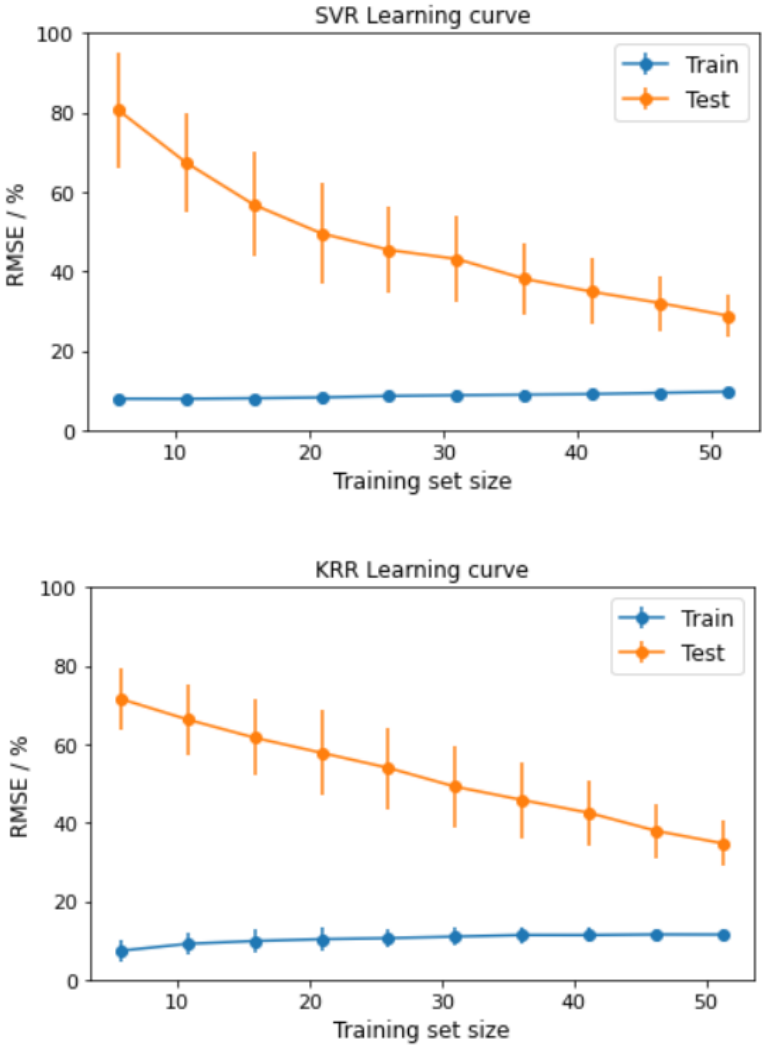

FIG. 2. Learning curves for the two models show the averaged RMSEs of the train and test sets with the variation of the training set size. (**top**) SVR model. (**bottom**) KRR model. Blue depicts the averaged RMSEs of the train sets and orange for the averaged RMSEs of the test sets in both figures.

V. SCHEMATIC DIAGRAM OF A 5-FOLD CROSS VALIDATION

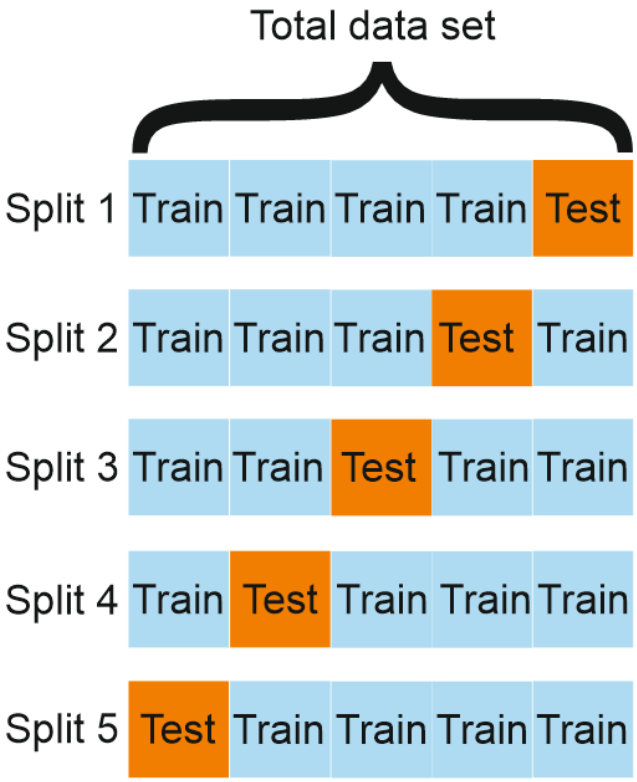

FIG. 3. Schematic diagram of a 5-fold cross validation. The total data set was divided into 5 folds and each fold was used as the test set to validate the model. The average behavior of the 5 test sets (averaged test RMSE) was used to verify the performance.

VI. SCHEMATIC DIAGRAM OF A RADIAL BASIS FUNCTION (RBF) KERNEL

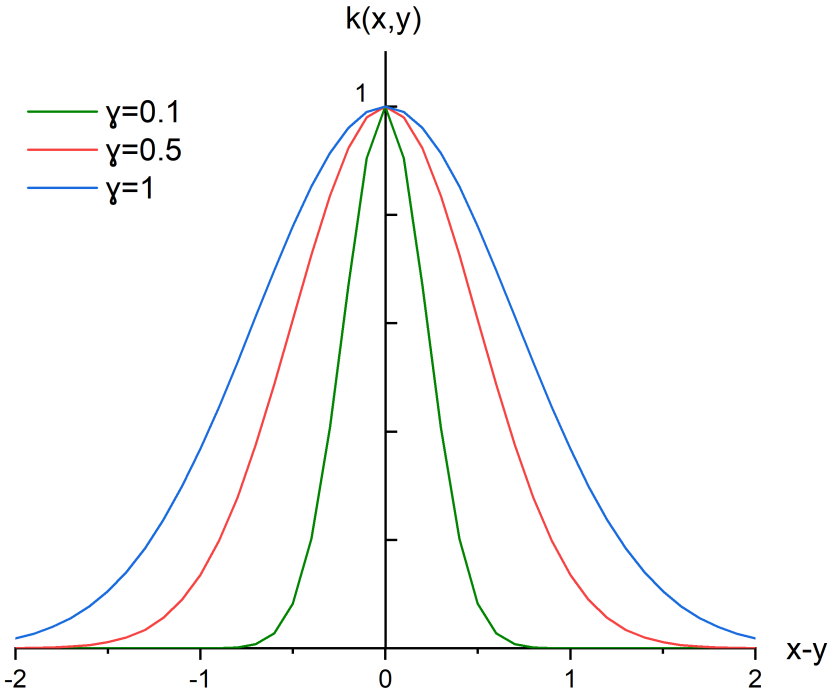

FIG. 4. Schematic diagram of a radial basis function (RBF) kernel with three γ values.

VII. THE DISTRIBUTION OF INHIBITION EFFICIENCIES

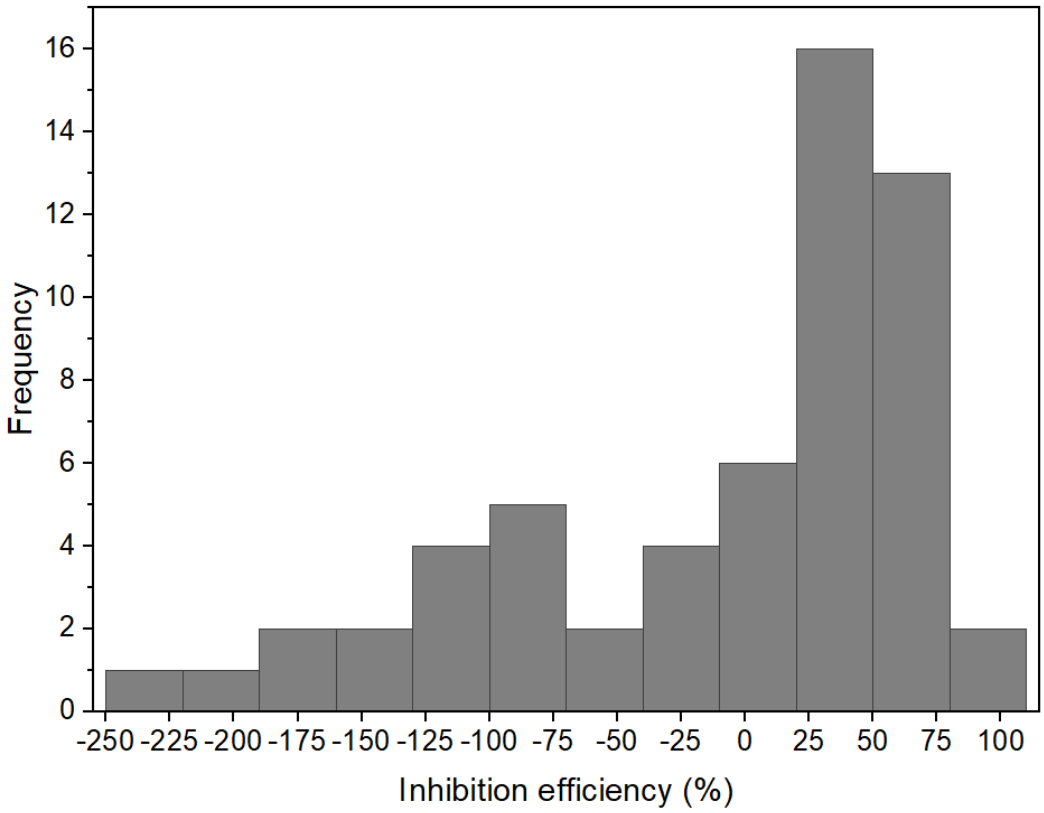

FIG. 5. The distribution of inhibition efficiencies for the 58 chemical compounds in the initial full data set.