Predicting the Inhibition Efficiencies of Magnesium Dissolution Modulators using Sparse Machine Learning Models

SUPPLEMENTARY MATERIAL

Elisabeth J. Schiessler[1], Tim Würger[2, 3], Sviatlana V. Lamaka[2], Robert H. Meißner[2], Christian J. Cyron[1], Mikhail L. Zheludkevich[2, 4], Christian Feiler✉, [2], and Roland C. Aydin✉, [1]

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

SUPPLEMENTARY NOTES

Probability of Selecting Specific Features

We want to calculate the probability of a random 5-tuple of features containing any specific molecular descriptor. Since the tuple is not ordered, the number of possible combinations of k-tuples is given by the formula for drawing k elements out of a set of N samples via the binomial coefficient

$$\binom{N}{k} = \frac{N!}{k! \cdot (N-k)!}. \tag{1}$$

Thus, the number of possible 5-tuples equals

$$\binom{1260}{5} \approx 2.63 \cdot 10^{13}, \tag{2}$$

whereas the number of 5-tuples containing a specific element is given by

$$\binom{1259}{4} \approx 1.04 \cdot 10^{11}. \tag{3}$$

The probability that a specific molecular descriptor is included in a randomly selected 5-tuple thus equals

$$\frac{1.04 \cdot 10^{11}}{2.63 \cdot 10^{13}} \approx 3.97 \cdot 10^{-3} \approx 0.004 \tag{4}$$

We performed a 5-fold cross-validation on selecting 5-tuples, with 100 runs each, so 500 runs in total. Thus, the expected value for a specific feature to be included would be 1.98 runs. The descriptor *LUMO / eV* was included in 479/500 runs, which is a factor of $2.41 \cdot 10^2$ more often than the expected value.

SUPPLEMENTARY TABLES

Model Hyperparameters

The regression models all contain 3 hidden layers that use `relu` activations. We choose an Adam optimizer and use mean squared error (MSE) as the loss function. Layer sizes vary, as do learning rates, *cf.* Supplementary Table 1. The models all contain a Gaussian noise layer with $\mu = 0$ and $\sigma = 0.1$ that is only active during training, and are trained for 25 epochs each.

The autoencoder contains three hidden layers with `leaky relu` activations, and does not use a Gaussian noise layer. It is compiled using an Adam optimizer with learning rate of 0.01 and MSE as loss function. The model is trained for 100 epochs.

Supplementary Table 1: Chosen hyperparameters for the different model types.

Model category	Layer sizes	Learning rate	Task
tiny	50-20-10	0.01	Regression
small	50-20-10	0.01	Regression
medium	50-20-10	0.05	Regression
large	100-50-10	0.005	Regression
small	10-2-10	0.01	Autoencoder

Top 63 Molecular Descriptors

The top 63 features as identified by analysis of variance (ANOVA) are provided in Supplementary Table 2. The top 63 features as identified by recursive feature elimination (RFE) are provided in Supplementary Table 3.

Mean Model Performance across all 10 Cross Validation Folds

The mean values of all reported statistics over all 10 cross validation folds, feature selection and model types can be found in Supplementary Table 4.

Model Performance for the Representative Validation Set

The average and standard deviations of predictions on the representative validation set from 100 training runs over all feature selection methods and model types can be found in Supplementary Table 5. Summarising statistics are provided in Supplementary Table 6.

Experimental Data

More information on the investigated compounds, including their SMILES strings and experimentally determined inhibition efficiencies (IE) for ZE41, can be found in Supplementary Table 7.

SUPPLEMENTARY REFERENCES

[1] Lamaka, S. V. *et al.* Comprehensive screening of Mg corrosion inhibitors. *Corros. Sci.* **128**, 224–240 (2017).

✉ email: christian.feiler@hereon.de, roland.aydin@hereon.de

Supplementary Table 2: Top 63 most relevant features as determined by ANOVA.

Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature
1	CATS2D_03_AP	17	Mor19m	33	F03[N-O]	49	Mor12p
2	CATS3D_03_AP	18	Mv	34	SsNH2	50	Eta_FL_A
3	CATS3D_02_AP	19	Mor12e	35	CATS3D_04_NL	51	GATS1m
4	LUMO / eV	20	B02[C-N]	36	E3p	52	Ui
5	P_VSA_MR_5	21	Eta_F_A	37	Hlgap / eV	53	Mor12m
6	P_VSA_LogP_2	22	R2e+	38	CATS2D_04_AA	54	SHED_PN
7	CATS2D_02_PN	23	Mor14s	39	Mor08i	55	SHED_DP
8	CATS2D_03_DP	24	NssCH2	40	HOMO / eV	56	GATS1v
9	H3m	25	Mor32m	41	TDB03e	57	MLOGP
10	nRNH2	26	SHED_AP	42	TDB02e	58	MPC07
11	H%	27	P_VSA_ppp_con	43	B03[N-O]	59	nN
12	Mor14u	28	Mor14v	44	CATS3D_02_AN	60	DLS_05
13	CATS2D_03_AN	29	Mor08s	45	Mi	61	nArCOOH
14	Eta_epsi_5	30	E2m	46	Mor27m	62	Mor08p
15	CATS3D_03_DP	31	SPH	47	TDB04v	63	SpPosA_B(e)
16	P_VSA_e_3	32	MATS7m	48	ZM2V		

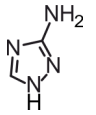
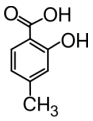
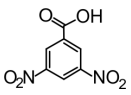
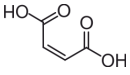
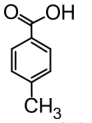
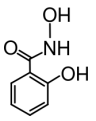
Supplementary Table 3: Top 63 most relevant features as determined by RFE.

Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature
1	P_VSA_MR_5	17	H3m	33	VE2sign_G	49	E3e
2	Mor22s	18	TDB04s	34	SpMAD_RG	50	Mor13u
3	Mor04m	19	E2s	35	R3s+	51	Eig03_AEA(dm)
4	LUMO / eV	20	R5p+	36	R5e+	52	X4Av
5	E1p	21	R2e+	37	E2v	53	P_VSA_e_3
6	HOMO / eV	22	ISH	38	Mor15i	54	Mor29e
7	P_VSA_LogP_2	23	DISPm	39	T(N..O)	55	Mor16m
8	Mor29v	24	R5i+	40	R2u+	56	GATS5m
9	MATS5v	25	Mor04i	41	MATS8p	57	E3p
10	Mor14s	26	Ds	42	Eta_epsi_5	58	E2e
11	Mor14u	27	Mor03s	43	MATS4s	59	X3Av
12	CATS3D_02_AP	28	E2m	44	H0v	60	Mor19u
13	GATS5v	29	Mor28s	45	Hy	61	GATS4s
14	MATS5m	30	Mor11u	46	E1i	62	E3v
15	GATS2s	31	TDB03m	47	VE1sign_G	63	TDB04m
16	Mor32m	32	Mor19m	48	Mor15s		

Supplementary Table 4: Mean values of root mean squared errors (RMSE), coefficients of determination (R^2), correlation coefficients (Pearson's r) and p-values of the full 10-fold cross validation for all trained models by model type and feature selection method (a: ANOVA, b: RFE, c: random selection).

No. of features	3			5			63			1260
Model Type	tiny model			small model			medium model			large model
Selection Method	a	b	c	a	b	c	a	b	c	
RMSE / pp	70	65	78	61	63	75	61	63	70	72
R^2	0.40	0.47	0.42	0.56	0.55	0.35	0.56	0.62	0.47	0.44
Pearson's r	0.53	0.65	0.53	0.69	0.72	0.50	0.73	0.77	0.63	0.61
p-value	0.38	0.20	0.30	0.20	0.14	0.37	0.14	0.10	0.24	0.26

Supplementary Table 5: Predictions for all trained models by type for the representative validation set. The averages and standard deviations are calculated over 100 training runs.

Model	Validation inhibition efficiency / %					
Index	0	5	13	36	45	54
						
True Value	-157.00	39.00	38.00	12.00	-6.00	-17.00
M3a	-105.01	-34.98	148.79	17.64	-69.53	-29.09
	± 35.58	± 22.34	± 50.81	± 24.70	± 22.45	± 23.01
M5a	-100.74	-35.98	154.62	15.88	-69.72	-30.13
	± 39.29	± 20.27	± 56.17	± 23.81	± 22.98	± 21.06
M63a	-104.19	-33.72	149.63	14.24	-69.31	-28.54
	± 35.85	± 22.21	± 57.00	± 21.76	± 24.12	± 23.43
M3b	-140.43	63.30	147.96	34.83	-22.28	19.11
	± 33.15	± 18.21	± 52.35	± 28.93	± 20.72	± 27.66
M5b	-138.90	65.77	144.03	32.88	-19.88	23.68
	± 33.19	± 19.36	± 49.41	± 29.44	± 22.36	± 31.75
M63b	-142.42	64.90	146.93	30.14	-20.63	18.86
	± 32.52	± 18.40	± 53.77	± 27.20	± 19.22	± 27.80
M3c	-36.90	-25.75	8.70	-33.71	-33.64	-29.09
	± 68.92	± 28.62	± 85.66	± 31.67	± 27.23	± 28.77
M5c	-41.96	-16.33	28.68	-23.96	-26.99	-27.66
	± 48.06	± 30.42	± 105.86	± 43.23	± 32.82	± 37.10
M63c	-73.99	17.67	127.29	10.09	-21.94	-12.83
	± 48.54	± 37.49	± 96.01	± 52.30	± 41.77	± 44.33
M1260	-77.54	14.46	64.30	8.63	-19.81	-4.01
	± 48.67	± 41.44	± 97.76	± 46.00	± 37.87	± 39.35

Supplementary Table 6: Coefficients of determination (R^2), root mean squared errors (RMSE), correlation coefficients (Pearson's r) and p-values of the full validation set predictions for all trained models by model type and feature selection method (a: ANOVA, b: RFE, c: random selection).

No. of features	3			5			63			1260
Model Type	tiny model			small model			medium model			large model
Selection Method	a	b	c	a	b	c	a	b	c	
RMSE / pp	71	50	61	67	49	55	64	49	51	36
R^2	0.26	0.83	0.28	0.38	0.84	0.42	0.41	0.84	0.56	0.81
Pearson's r	0.51	0.91	0.53	0.62	0.91	0.65	0.63	0.91	0.75	0.90
p-value	0.30	0.01	0.28	0.19	0.01	0.17	0.17	0.01	0.09	0.02

Supplementary Table 7: Investigated compounds with their SMILES strings and experimentally determined inhibition efficiencies (IE) for ZE41. Presented values of IE were determined by comparing the amount of evolved hydrogen during immersion of ZE41 alloy chips in pure 0.5% NaCl solution with the amount of hydrogen that evolves in presence of one of the listed dissolution modulators at an operation concentration of 0.05 M. The elemental composition of the alloy is provided within the published database of inhibition efficiencies [1].

Compound	SMILES	IE / %
3-Amino-1,2,4-triazole	c1[nH]nc(n1)N	-157
3-Methylcatechol	c1(c(c(ccc1)O)O)C	-31
3-Methylsalicylic acid	c1(c(c(ccc1)C(=O)O)O)C	75
4-Aminosalicylic acid	c1c(c(ccc1N)C(=O)O)O	57
4-Hydroxybenzoic acid	c1c(ccc(c1)O)C(=O)O	-170
4-Methylsalicylic acid	c1c(c(ccc1C)C(=O)O)O	39
5-Aminosalicylic acid	c1c(c(cc(c1)N)C(=O)O)O	66
5-Methylsalicylic acid	c1c(c(cc(c1)C)C(=O)O)O	55
5-Nitrobarbituric acid	C1(=O)C(C(=O)NC(=O)N1)N(=O)=O	51
2,3-Pyridinedicarboxylic acid	c1(c(cccn1)C(=O)O)C(=O)O	32
2,5-Pyridinedicarboxylic acid	c1(ccc(C(=O)O)cn1)C(=O)O	52
3,4-Dihydroxybenzoic acid	c1c(ccc(c1O)O)C(=O)O	-270
3,5-Dimethylpyrazole	c1(cc(n[nH]1)C)C	-67
3,5-Dinitrosalicylic acid	c1(c(c(cc(c1)N(=O)=O)C(=O)O)O)N(=O)=O	38
5,5-Dimethylhydation	C1(CNC(=O)N1)(C)C	-172
Ascorbic acid	[C@@H]1(OC(=O)C(=C1O)O)[C@H](CO)O	6
Asparagine	[C@H](CC(=O)N)(C(=O)O)N	-188
Aspartic acid	[C@H](CC(=O)O)(C(=O)O)N	-54
Benzoic acid	c1cc(ccc1)C(=O)O	34
Benzotriazole	c1ccc2c(c1)nn[nH]2	-108
Bicine	N(CC(=O)O)(CCO)CCO	-111
Bipyridine	c1(ccccn1)c1ccccn1	6
Bismuthiol	c1(nnc(s1)S)S	-17
Citric acid	C(C(CC(=O)O)(C(=O)O)O)C(=O)O	-47
Cystein	[C@H](CS)(C(=O)O)N	-104
Diethylentriamine	C(CNCCN)N	-18
Diglycolic acid	C(=O)(O)COCC(=O)O	60
Formic acid	C(=O)O	-5
Fumaric acid	C(=C\\C(=O)O)/C(=O)O	17
Gluconic acid	[C@H]([C@@H]([C@H]([C@H](CO)O)O)O)(C(=O)O)O	48
Glutamic acid	[C@H](CCC(=O)O)(C(=O)O)N	-139
Glycerol	C(C(CO)O)O	-13
Glycine	C(C(=O)O)N	-215
Glycolic acid	C(=O)(C)CO	59
Kojic acid	c1(cc(=O)c(co1)O)CO	45
Lysine	[C@H](CCCCN)(C(=O)O)N	-113
Maleic acid	C(=C\\C(=O)O)\\C(=O)O	12
Maltol	c1(c(c(=O)cco1)O)C	9
Mandelic acid	c1ccc(cc1)[C@@H](C(=O)O)O	-6
Nicotinic acid	c1c(cccn1)C(=O)O	3
Norleucine	CCC[C@@H](C(=O)O)N	-138
NTA	N(CC(=O)O)(CC(=O)O)CC(=O)O	-124
Oxalic acid	C(=O)(O)C(=O)O	52
Panthenol	C(=O)(NCCCO)[C@H](C(C)(CO)C)O	-98
p-^{t}Bu-Benzoic acid	c1c(ccc(c1)C(C)(C)C)C(=O)O	31
p-Toluic acid	c1c(ccc(c1)C)C(=O)O	-6
Phenylalanine	[C@H](Cc1ccccc1)(C(=O)O)N	-146
Phthalic acid	c1c(c(ccc1)C(=O)O)C(=O)O	31
Picolinic acid	c1(ccccn1)C(=O)O	54
Piperazine	C1CNCCN1	-145
Pyrazole	c1ccn[nH]1	-38
Quinaldic acid	c1(ccc2c(n1)cccc2)C(=O)O	21
Quinic acid	[C@@]1(CC[C@H]([C@@H]([C@@H]1O)O)O)(O)C(=O)O	19
Salicylaldoxime	c1c(c(ccc1)/C=N/O)O	-89
Salicylhydroxamic acid	c1c(c(ccc1)C(=O)NO)O	-17
Salicylic acid	c1c(c(ccc1)C(=O)O)O	37
Tartaric acid	[C@H]([C@H](C(=O)O)O)(C(=O)O)O	46
Tris	C(CO)(CO)(CO)N	-194
Uracil	[nH]1ccc(=O)[nH]c1=O	-108