

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

Xuejiao Li (✉ xuejiao.li@hereon.de)

Helmholtz-Zentrum Hereon

Bahram Vaghefinazari

Institute of Surface Science, Helmholtz-Zentrum Hereon

Tim Würger

Helmholtz-Zentrum Hereon

Sviatlana Lamaka

Helmholtz-Zentrum Hereon

Mikhail Zheludkevich

Helmholtz-Zentrum Hereon

Christian Feiler

Helmholtz-Zentrum Hereon <https://orcid.org/0000-0003-4312-7629>

Article

Keywords: Magnesium, Corrosion inhibitors, Chemical space, QSPR, Feature selection

Posted Date: January 13th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-2386421/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: (Not answered)

Version of Record: A version of this preprint was published at npj Materials Degradation on August 9th, 2023. See the published version at <https://doi.org/10.1038/s41529-023-00384-z>.

**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*

Abstract

Selecting effective corrosion inhibitors from the vast chemical space is not a trivial task, as it is essentially infinite. Fortunately, machine learning techniques have shown great potential in generating shortlists of inhibitor candidates prior to large-scale experimental testing. In this work, we used the corrosion responses of 58 small organic molecules on the magnesium alloy AZ91 and utilized molecular descriptors derived from their geometry and density functional theory calculations to encode their molecular information. Statistical methods were applied to select the most relevant features to the target property for support vector regression and kernel ridge regression models, respectively, to predict the behavior of untested compounds. We compared the performance of the two supervised learning approaches and assessed the robustness of our data-driven models by experimental blind testing.

9 Keywords: Magnesium, Corrosion inhibitors, Chemical space, QSPR, Feature selection

10 I. INTRODUCTION

11 Magnesium (Mg), the lightest structural metal, is a promising material in automotive
12 and aeronautic engineering due to its outstanding mechanical properties as well as in med-
13 ical industries due to its biocompatibility. [1–3] However, Mg-based materials have to be
14 protected from corrosion to facilitate their application in advanced engineering applications,
15 as Mg is a highly reactive metal. Surface coatings depict a reliable and effective strategy
16 to realize the corrosion protection of Mg by adding a barrier layer between the substrate
17 and the service environment. [3–5] However, scratches or cracks in the protective coating
18 may lead to severe local corrosion reactions. [6] This can be mitigated by incorporating
19 corrosion inhibitors into the coatings that will be released on demand and inhibit corrosion
20 in the damaged areas. [6–8] It is noteworthy that direct embedding of corrosion inhibitors
21 into a coating matrix [9] may impair their functionality by no or limited release [10, 11]
22 or may release all corrosion inhibitors at once without control once a defect occurs.[12]
23 Application of layered double hydroxides (LDHs) intercalated with corrosion inhibitors is
24 one of the promising routes to achieve a controllable active corrosion protection. [12–14]

* xuejiao.li@hereon.de

† christian.feiler@hereon.de

25 An LDH is an inorganic sheet-like clay with a brucite structure in its pure $\text{Mg}(\text{OH})_2$ form.
26 Thanks to the anion exchange property of the LDH structure, the corrosion inhibitors can
27 be intercalated into this layered structure and their release can be subsequently triggered
28 by exchanging with an aggressive corrosive species (e.g. chloride) to suppress corrosion re-
29 actions. [12] Aside from the inorganic corrosion inhibitors commonly intercalated in the
30 LDHs such as vanadate [12], tungstate [15], and molybdate [16], organic corrosion inhibitors
31 have gained more and more attention recently because a large number of organic compounds
32 have shown promising corrosion inhibition for Mg and its alloys.[7] Furthermore, it has been
33 demonstrated that small organic molecules can be intercalated into LDHs [17–19].

34 However, pure experimental studies on the intercalation of new organic molecules into
35 LDHs can be time-consuming, especially when considering the large number of candidate
36 molecules to choose from. [20] Aside from that, identification of an effective organic corrosion
37 inhibitor to be intercalated into LDHs (see Figure 1) to protect a specific type of Mg alloy
38 can be very challenging due to the large number of organic compounds with potentially
39 useful properties[21]. Luckily, machine learning-based approaches promise to facilitate the
40 screening for useful compounds.

41 Machine learning (ML) has developed rapidly in recent years due to the augmentation
42 of algorithms and technological advances in computing hardware. [22] While influencing
43 our daily life [23, 24], machine learning algorithms have also gained an important role in
44 material science [25, 26]. Different algorithms have been applied in material discovery such as
45 compound prediction [27–29], structure prediction [30, 31] and predicting material properties
46 such as band gap [32], superconductivity [33], bulk and shear moduli [34] and to identify
47 effective corrosion inhibitors based on quantitative structure-property relationships (QSPR)
48 [35, 36]. For the latter, a number of different machine learning algorithms [21, 37, 38] were
49 successfully developed to predict the corrosion inhibiting effect of small organic compounds
50 for different types of Mg and its alloys [7, 21, 37], Aluminum alloys [35, 36, 39] and Copper-
51 based materials [40]. Naturally, a sufficiently large, diverse and reliable training data set and
52 a suitable modeling framework (usually based on one or more machine learning algorithms),
53 are two of the crucial prerequisites for the development of predictive QSPR models. A third
54 key step is the selection of relevant input features which can either be selected by chemical
55 intuition [38] or based on statistical methods [37].

56 In this work, corrosion inhibition responses of 58 small organic molecules on Mg alloy

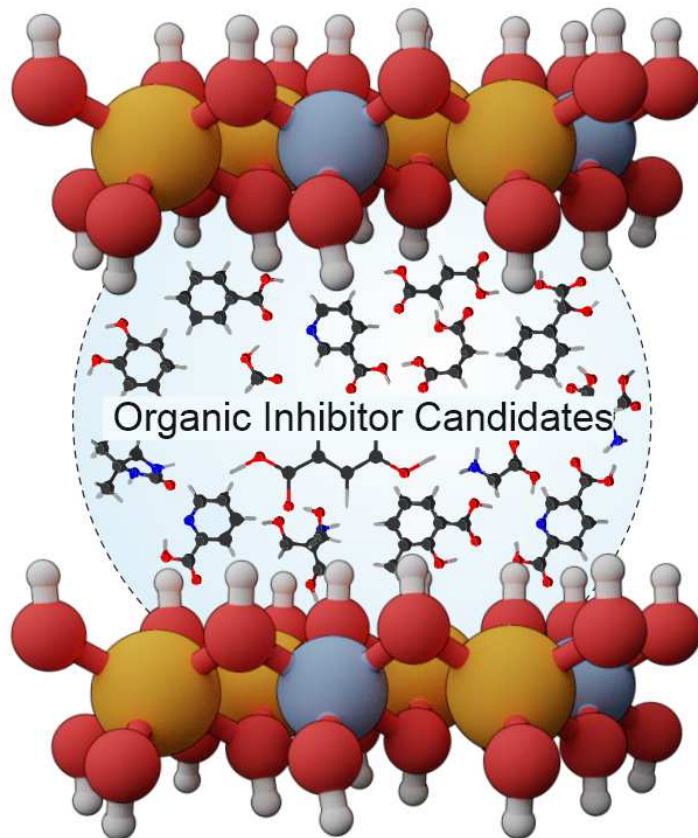


FIG. 1: Schematic representation of a layered double hydroxide system with a large number of organic inhibitor candidates.

57 AZ91 from a previous work [7] were used to train a QSPR model. AZ91 was the selected
 58 substrate in this study because our previous experimental work [41] proved that LDHs can
 59 be directly synthesized at the surface of this alloy as a conversion layer. A potential al-
 60 gorithm that can be employed to establish a QSPR workflow are support vector machines
 61 (SVMs) which represent one of the most powerful, precise and robust supervised learning
 62 methods due to their good theoretical foundations and generalization capacity. [42, 43]
 63 They have been widely applied to solve various complex real-world problems such as: image
 64 classification [44], hand-written character recognition [45] and face detection [46] in the past
 65 twenty years. [42] Applying the same principle as SVMs, support vector regression (SVR)
 66 was developed to solve regression problems with high accuracy.[47–49] Moreover, approaches
 67 based on kernel ridge regression (KRR) [50] have been applied by other researchers to de-
 68 velop reliable models. [48, 49] In the end, the QSPR model developed in this work can
 69 assist the selection of an effective organic corrosion inhibitor from a large number of organic

70 compounds, whose intercalation into the LDHs will be further investigated to achieve the
71 goal of corrosion protection for AZ91.

72 II. RESULTS & DISCUSSION

73 Schiessler et al. [37] applied statistical feature selection methods and selected the rel-
74 evant descriptors to predict the inhibition efficiencies determined for the Mg alloy ZE41.
75 In this work, further investigations of the feature selection were carried out which is a key
76 element in the development of an ML model that predicts the corrosion IEs of small organic
77 molecules. Based on the selected features, two different QSPR models (based on SVR and
78 KRR algorithms) are trained to predict the IEs of small organic molecules on AZ91 and their
79 accuracy is subsequently validated and compared based on experimental blind testing using
80 ten compounds which were not part of the initial data set. An overview of the workflow
81 used in this study is illustrated in Figure 2.

82 A. Feature selection

83 A pool of 2876 distinct molecular descriptors was generated using the cheminformatics
84 software package alvaDesc [51] and using density functional theory (DFT) calculations per-
85 formed at the TPSSh/def2SVP level of theory employing the quantum chemical software
86 package Gaussian 16 [52] as input features for the development of a QSPR model. After
87 omitting all molecular descriptors with constant values, the remaining 876 were exposed to
88 a sparse selection approach based on recursive feature elimination (RFE) as its potential to
89 select relevant features was demonstrated in a recent study. [37] In this work, we added an
90 additional step to the feature selection by gradually decreasing the number of used input
91 features, starting from the 25-tuple features that were selected using RFE (see the feature
92 selection section of Figure 2).

93 In the first step, we selected the group of 25 features out of the initial 876 features with
94 the application of RFE based on random forests. More details on the selection process of
95 the selected 25 features are available in the 'Methods' section. The list of 25 features serves
96 as basis for the following feature importance investigation considering two different ML
97 approaches based on SVR and KRR models. In the second step of the feature importance

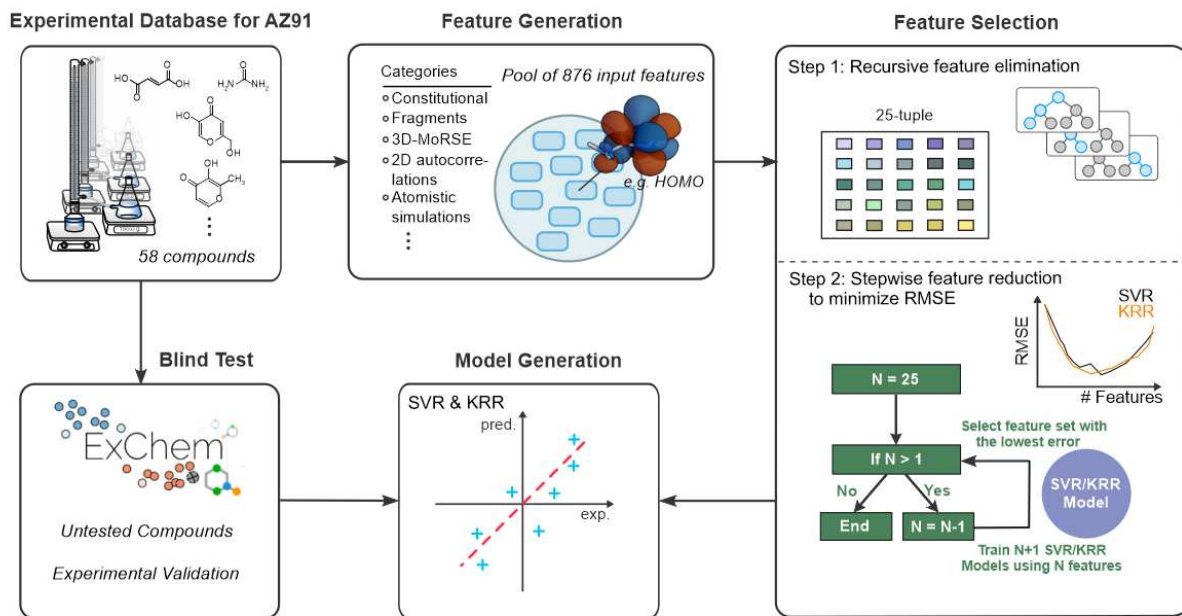


FIG. 2: Schematic representation of the ML workflow used in this study. A database of 58 small organic molecules and their corrosion responses on AZ91 are employed as training database. First a pool of molecular descriptors to encode their molecular structure is generated and exposed to a two-step sparse feature selection approach. The most relevant descriptors are subsequently used to train supervised machine learning models to predict the behaviour of untested chemicals. The small organic molecules for this step are selected following our previously published ExChem[21] approach.

98 investigation, the initially selected 25 features were removed one-by-one in 24 steps. Instead
 99 of applying RFE, the SVR and KRR models were used directly to select features at each
 100 step together with hyperparameter optimization and cross validations. At each step, there is
 101 more than one possibility to remove a feature from the previous step, e.g., there are twenty-
 102 five possibilities to remove one feature from the selected 25 features. Attempts across all
 103 possibilities were conducted and the possibility with the lowest averaged root mean squared
 104 error (RMSE) of the IEs for the test sets in the cross validation was selected at each step
 105 and plotted in Figure 3. The averaged RMSEs for the train sets in the cross validation
 106 corresponding to the plot in Figure 3 were listed in the Supporting Information (Table S1).
 107 For the selected possibility, the removed feature was defined as the least important feature
 108 in the previous step. In the end, the selected 25 features were ordered according to the

previously defined importance, obtaining an order of importance for the features.

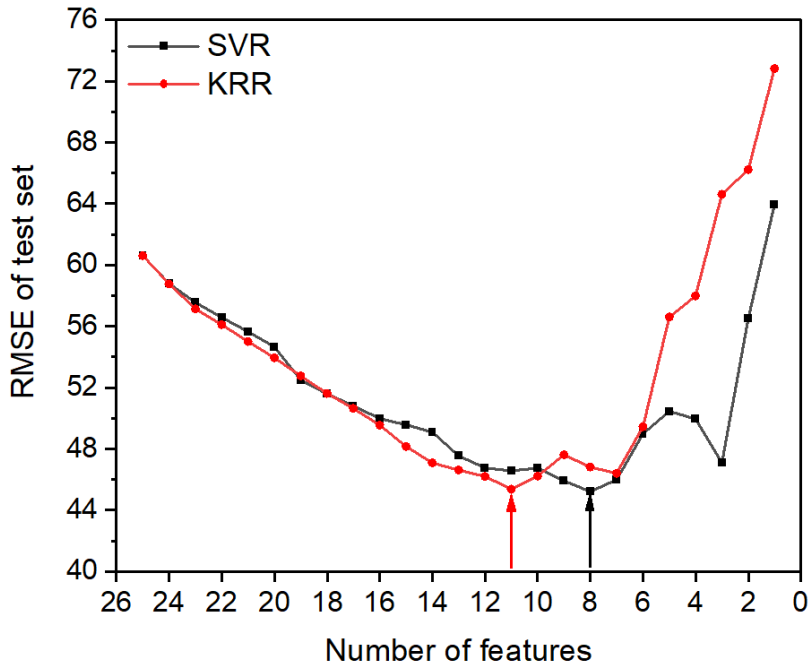


FIG. 3: 25-tuple features selected after the application of RFE based on random forests in Step 1 were removed one-by-one and the minimum averaged RMSE of the test sets in the cross-validations varied with the number of features for SVR (in black line) and KRR (in red line) models.

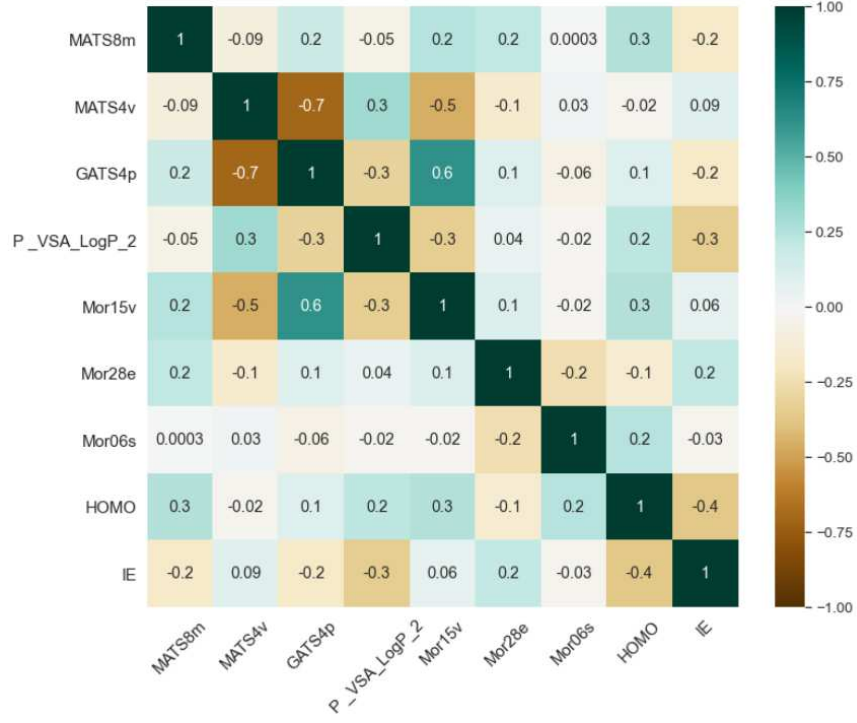
The trend of the black line in Figure 3 shows that the optimal number of features selected for the SVR model equals eight, since the resulting model exhibits the lowest RMSE. The selected molecular descriptors are P_VSA_LogP_2, Mor28e, HOMO, MATS4v, Mor06s, GATS4p, MATS8m, and Mor15v, ordered by their feature importance. Except for the highest occupied molecular orbital (HOMO) which is obtained from DFT calculations, the other seven features are from three descriptor categories (P_VSA-like descriptors, 3D-MoRSE descriptors[53] and 2D autocorrelations) obtained from the chemoinformatic software package alvaDesc [51]. P_VSA-like descriptors are based on the van der Waals surface area of the compounds by summing up all the atomic contributions. 3D-MoRSE descriptors incorporate the whole molecule structure information by summarising the atomic pairwise information related to the scattering parameter based on electron diffraction and then weighted by either of the properties, e.g., mass, Sanderson electronegativity, van der Waals volume and atomic polarizability. The 2D autocorrelations descriptors are calculated to provide the interde-

pendence between atomic properties (analogous to the 3D-MoRSE descriptors), which are connected by a log function.[54] All these three descriptor categories focus on calculating the spatial distribution of a generic molecular property rather than only considering the atomic configurations.

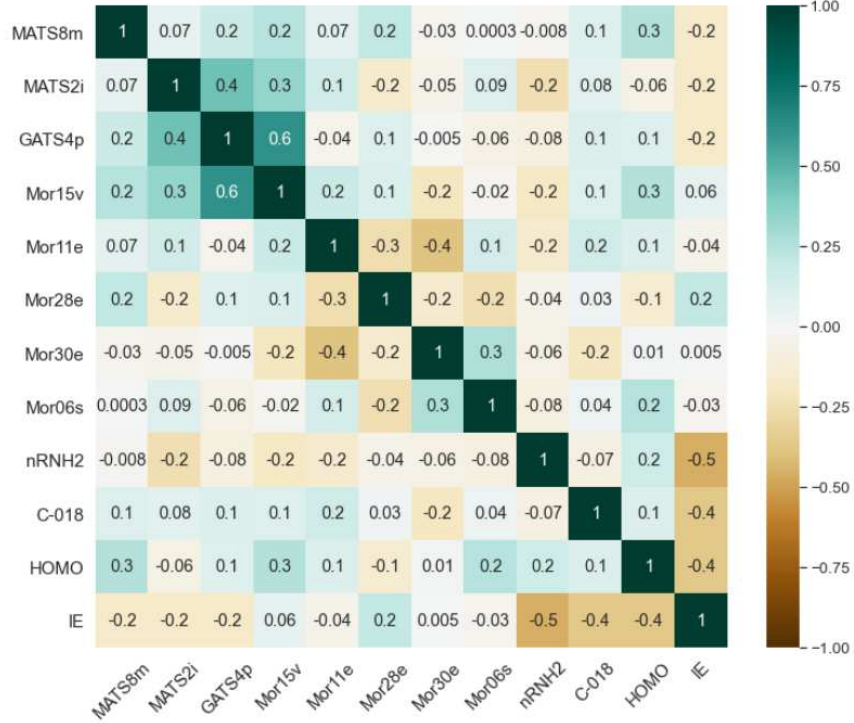
In the KRR model, the optimal number of features resulted in eleven as shown by the red line in Figure 3. The eleven selected features were identified as Mor15v, HOMO, MATS8m, Mor30e, nRNH2, C-018, GATS4p, MATS2i, Mor11e, Mor06s, Mor28e, ordered by their feature importance. It is noteworthy that six out of the eleven features are identical with those selected for the SVR model. The overlapping features are Mor15v, HOMO, MATS8m, GATS4p, Mor06s, Mor28e. This finding implies that the HOMO energies derived from DFT calculations, 3D-MoRSE descriptors and 2D autocorrelations descriptors seem to encode crucial structural information concerning the prediction of the corrosion inhibition efficiency of small organic molecules for AZ91. This observation agrees well with the conclusion from Schiessler et al. [37] where DFT calculated features as well as 3D-MoRSE descriptors were identified as important input features for an artificial neural network using IEs of small organic molecules for the Mg-based alloy ZE41 as a target property. Apart from these three feature groups, a number of features encoding functional group counts and atom-centred fragments were identified for the top eleven features in the KRR model, e.g., nRNH2 which directly encodes the number of aliphatic primary amines. All five compounds that contain nRNH2 moieties in our data set are amino acids (Cysteine, Glutamic acid, Glycine, DL-norleucine and DL-phenylalanine) which exhibit negative inhibition efficiencies. This finding agrees well with the conclusion in ref. 7 that amino acids accelerated corrosion of Mg alloys. The corrosion acceleration behaviour of amino acids can be attributed to the solubility of their corresponding magnesium complex in water.[55, 56] The feature C-018 from the class of atom-centred fragments represents =CHX, where "=" depicts a double bond and X any of the following heteroatoms: O, N, S, P, Se or any halogen. In our training data set, this specific functional group is present in the compounds Kojic acid, Maltol and Uracil whereas all three organic molecules display negative IE values. It has been proven that the complexes formed by these three compounds with magnesium are water-soluble. [56–58] Compared to the capability to form complexes with metal ions, the solubility of these complexes in water appears to be a more decisive factor in determining the efficiency of the organic inhibitors. This observation agrees well with the work from Lamaka et al.[7] and Anjum et al.[19]

155 that organic compounds whose complexes have a low solubility in water exhibited a high
156 inhibiting effect since they delay corrosion by forming a protective barrier layer.

157 Some of the molecular descriptors obtained from chemoinformatics tools like alvaDesc
158 are arcane and cannot be easily linked to physical properties since they are derived from
159 extensive mathematical manipulations of the chemical graph. To provide a better under-
160 standing of the correlation between the used input features and IEs, all respective correlation
161 coefficients were determined based on Pearson correlation tests. The Pearson correlation co-
162 efficient measures the linear relationship between two sets of data, which varies between -1
163 and 1 with 0 implying no correlation while -1 and 1 implying exact negative and positive cor-
164 relations, respectively.[59] For both models, the correlation between the individual features
165 and the IEs is moderate to weak since the values of the determined correlation coefficients
166 in Figure 4a and b are not higher/lower than ± 0.5 , where the most pronounced negative
167 and positive correlations are -0.5 and 0.2 , respectively. This observation agrees well with
168 the findings of Guyon et al.[60] that the selected features are on its own not necessarily the
169 most relevant with respect to the target property. For the correlation between the selected
170 features, neither of the correlation is considered as a strong relationship (> 0.9) and most of
171 the correlations (over 90%) are interpreted as weak relationships ($0.1 - 0.39$) or are negligible
172 (< 0.1) according to the definitions in the work of Schober et al.[59]. Moreover, the p-value
173 between the used input features and IEs was calculated and illustrated in the Supporting
174 Information Figure S1, where the p-value is an indicative measure whether the correlation is
175 statistically significant. The weak correlations between most of the selected features largely
176 ensure that there is no redundant feature selected as input for the models. Although most
177 of the selected features are only weakly correlated with the target property itself, our results
178 indicate that they can still be used to build a predictive model when used as a group due to
179 underlying synergistic effects which is in good agreement with our previous works.[37, 38]



(a) SVR model



(b) KRR model

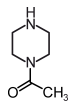
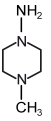
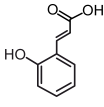


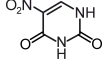
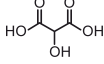
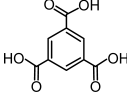
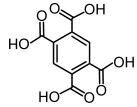

FIG. 4: Pearson correlation coefficients for the two models. (a) Pearson correlation among the selected 8-tuple features for the SVR model and IEs. (b) Pearson correlation among the selected 11-tuple features for the KRR model and IEs.

180 In summary, the feature selection method proposed in this work is able to increase the
181 accuracy of the predictions in the cross-validation stage by applying the step-wise reduction
182 to the group of features which was selected based on RFE in a first step. Moreover, the
183 proposed method can be employed to perform RFE for SVR with an radial basis function
184 (RBF) kernel, since only the linear kernel is currently supported in scikit-learn [61]. Another
185 advantage of this proposed method is that there is no prerequisite on the number of features
186 to be selected, therefore all possible combinations of feature groups are explored in the
187 feature selection and a comprehensive exploration can be guaranteed.

188 B. Model validation

189 Hyperparameters for the SVR and KRR models were optimized in a grid search with
190 5-fold cross validations together with the feature importance investigation. As a result,
191 the set of hyperparameters for the SVR (`random_state`=10, $C=17$, $\gamma=0.1$) and the KRR
192 (`random_state`=10, $\alpha=0.1$, $\gamma=0.1$) were selected respectively. For both models, the value
193 selected for the random state parameter (`random_state`) is identical which indicates the
194 same split of the data set into train and test sets in the cross validations. After the selection
195 of the hyperparameters, the full initial data set was used to fit the two models and then these
196 models were applied to predict the behavior of the blind test compounds to evaluate their
197 robustness. For the selection of the compounds in the blind tests, Trimesic acid and Py-
198 romellitic acid were suggested by our experimentalists based on chemical intuition, whereas
199 the remaining candidates were selected by following the ExChem approach described in our
200 previous work[21], using a database of 7094 commercially available compounds provided by
201 Thermo Fisher Scientific. The experimental and predicted values for the 10 compounds in
202 the blind tests are listed in Table I where values in red indicate overestimated and in blue
203 underestimated predictions. The predicted values for the piperazine derivatives **1** and **2** are
204 marked in brown for both models as their predicted acceleration efficiencies are significantly
205 less negative than the corresponding experimental values, which are beyond the inhibition
206 efficiency range of the chemicals used as initial data set in this work. However, it is note-
207 worthy that both compounds were correctly predicted to accelerate the dissolution of AZ91.
208 These two compounds were excluded in the following analysis since they are outside of the
209 domain of applicability of the used initial data set.

TABLE I: Experimental and predicted values (IEs in %) for the blind test compounds.

Index	Compound	IE (exp.)	SVR (pred.)	KRR (pred.)
1	 1-Acetylpiperazine	-563	-172	-108
2	 1-Amino-4-methylpiperazine	-517	-195	-109
3	 2-Hydroxycinnamic acid	-24	51	2
4	 2-Hydroxyphenylacetic acid	-14	-21	-53
5	 3-Methylpyrazole	16	9	-17
6	 5-Nitrouracil	79	-68	-82
7	 Tartronic acid	30	37	60
8	 Trimesic acid	67	-89	23
9	 Pyromellitic acid	52	34	23
10	 Trimethylolpropane	20	-42	-49

The SVR and KRR models performed similarly well for the full initial data set, the blue points in Figure 5a and b, where the predicted and experimental values correlated well with an RMSE around 10%. The performance of some of the blind test compounds that were under- or overestimated, circled by red and blue dashed circles or ellipses in Figure 5 (the colors correspond to the ones in Table I), results in a relatively high RMSE value for both employed models. From Figure 5, it can be seen that both SVR and KRR models underestimated 5-Nitrouracil (**6**, $IE_{\text{pred,KRR}} = -82\%$, $IE_{\text{pred,SVR}} = -68\%$) and Trimethylolpropane (**10**, $IE_{\text{pred,KRR}} = -49\%$, $IE_{\text{pred,SVR}} = -48\%$) in a similar way. There are other two outliers (2-

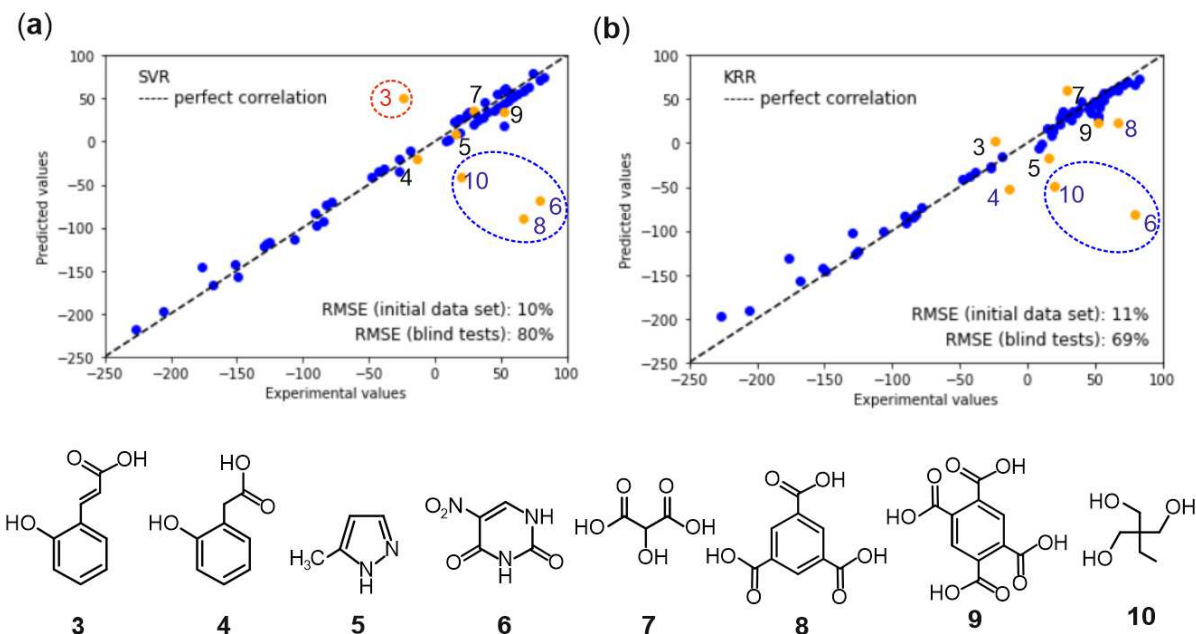


FIG. 5: Performance assessment of the (a) SVR model and (b) KRR model. Both figures show the correlation between the predicted values and the measured values from experiments (IEs in %). The blue points represent the full initial data set (58 compounds, the names and IEs were listed in the Supporting Information (Table S2)). The orange points depict the blind test compounds. Please note that 1-Acetylpiperazine (**1**) and 1-Amino-4-methylpiperazine (**2**) were excluded from the plot. Although their estimates were qualitatively correct (**1**: $IE_{\text{pred,SVR}} = -172\%$, $IE_{\text{pred,KRR}} = -108\%$, $IE_{\text{exp}} = -563\%$; **2**: $IE_{\text{pred,SVR}} = -195\%$, $IE_{\text{pred,KRR}} = -109\%$, $IE_{\text{exp}} = -517\%$), their measured values were far outside the models domain. For the sake of clarity, the corresponding structures of the plotted blind test compounds are shown at the bottom of the figure. Red and blue dashed circles or ellipses mark the over- and underestimated compounds, respectively.

Hydroxycinnamic acid (**3**) and Trimesic acid (**8**)) in the SVR model as shown in Figure 5a. Even though there are two more outliers in the SVR model, it is important to note that the predicted values for the other four compounds in the blind test set correlated well with the corresponding experimental values for the acetic acid **4** ($IE_{\text{pred,SVR}} = -21\%$, $IE_{\text{exp.}} = -14\%$), the pyrazole **5** ($IE_{\text{pred,SVR}} = 9\%$, $IE_{\text{exp.}} = 16\%$) as well as the aliphatic (**7** ($IE_{\text{pred,SVR}} = 37\%$, $IE_{\text{exp.}} = 30\%$)) and aromatic (**9** ($IE_{\text{pred,SVR}} = 34\%$, $IE_{\text{exp.}} = 52\%$)) carboxylic acids with an RMSE 22% in the SVR model. The RMSE calculated for the non-outlier compounds **3**, **4**, **5**, **7**, **8** and **9** in the KRR model results to 34%. These observations indicate that both

the SVR and KRR models are able to provide good estimates for majority of the blind test compounds. The difference between these two models is that the SVR model can provide a higher accuracy of predictions for the non-outlier compounds while there are less outliers in the KRR model.

One out of the 10 compounds (5-Nitrouracil (**6**)) in the blind test set contains a =CHX fragment, suggesting that it has a negative IE value. However, in contrast to the predicted negative inhibition efficiency, the experimental result showed that 5-Nitrouracil gave adequate inhibition performance. This could be attributed to the nitro compounds of 5-Nitrouracil which have been proven to be able to assist the corrosion protection of a variety of alloys. [62–64] Moreover, a number of organic compounds with nitro groups were experimentally tested to have positive inhibition efficiencies for commercial magnesium and the Mg ZE41 alloy.[7, 21] This observation is, however, not captured by neither of the employed models because of the limited information on the effect of a nitro functionality in our data set as there is only one compound (5-Nitrobarbituric acid) that exhibits this functional moiety. This strongly indicates that future experimental data set need to include more compounds with a nitro moiety to enable the model to recuperate the impact of this group on the corrosion inhibiting effect.

To gain more insights of the compounds which are outliers, we calculated the pairwise distances (see the 'Methods' section) based on the input features between the compounds in the blind test and the initial data set used in building the models to evaluate the highly similar structures for each blind test compound. Two compounds with high similarity end up with a value close to 1 in the similarity matrix. With the similarity decreasing, the value in the similarity matrix decreases approaching 0. This yields a similarity value between 0 (no similarity) and 1 (identical). Figure 6a and b show the similarity matrix for the eight blind test compounds and the initial data set for the SVR and KRR models, respectively. The top 5 similar structures (containing the names and the inhibition efficiencies) for 5-Nitrouracil (**6**) are shown in Figure 6 for both models. Based on the color shown in the matrix, a similarity order from high to low can be extracted for these 5 structures in SVR (Uracil, Glycine, 5-Nitrobarbituric acid, DL-Phenylalanine, Glutamic acid) and KRR (Uracil, Maltol, Kojic acid, Fumaric acid, Urea). It is noteworthy that there are obvious color differences for some of the top 5 similar structures such as the difference between Uracil and Urea in the KRR model as shown in Figure 6 b. This indicates the limitation of the data set used in this

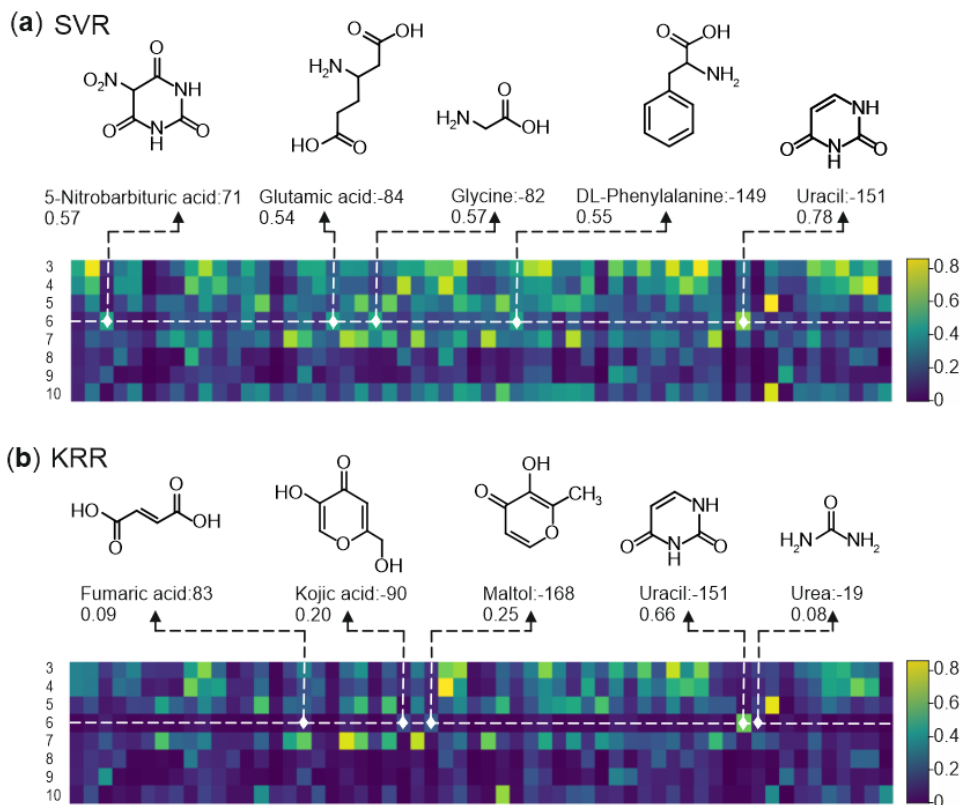


FIG. 6: Similarity matrix of the 8 blind test compounds and the 58 compounds in the data set for the (a) SVR model and (b) KRR model. The top 5 similar structures containing the names and the inhibition efficiencies for 5-Nitrouracil (6) are plotted in the figure as an example. The values below the names are the similarity values. The color scale corresponds to the values in the matrix where dark blue indicates low / no, green moderate and yellow high similarity values.

work where there are only 58 data points in total. As a consequence, there are not enough structures in the data set with similarities comparable to the similarity between Uracil and the blind test compound 5-Nitrouracil (6). The IEs of these 5 similar structures are ordered by similarity in Table II. In the last column of Table II, the average of the IEs for these 5 similar structures is listed. We followed the same process to extract the top 5 similar structures and list their IEs and the average IE in Table II for all the other outliers. The average values of the IEs exhibit a decent agreement with the predicted values as shown in Table II. This indicates that our models are able to capture the similarity connections existing in the data set and make according predictions. The similarity connections are

however limited by the small size of our data set, resulting in the appearance of these outliers. The learning curves for the SVR and KRR models (as illustrated in the Supporting Information Figure S2) show that the averaged RMSEs for the test sets in the cross validation decrease as the size of the training set increases, although the averaged RMSEs of the test sets for both models are higher relative to that of the train sets. One possible remedy is to expand the data set, so the averaged RMSEs of the test sets can consistently decrease by adding additional training data.

TABLE II: The IEs in % of the extracted top 5 similar structures from ref. 7 are listed in the similarity order from high to low (from 1st to 5th, please note that 1st, 2nd, 3rd, 4th, 5th do not indicate the same structures but refer to those that are most similar to the ones that were tested in this work.) and the average of these 5 IEs for the outliers in the SVR and KRR models.

		IE _{exp}	IE _{pred}	1 st	2 nd	3 rd	4 th	5 th	Average
SVR	3	-24	51	26	56	18	45	24	34
	6	79	-68	-151	-82	71	-149	-84	-79
	8	67	-89	38	11	-90	-125	52	-23
	10	20	-42	-27	-27	24	34	-106	-20
KRR	6	79	-82	-151	-168	-90	83	-19	-69
	10	20	-49	-27	-91	54	-27	-129	-44

In this work, the performance of two supervised machine learning approaches (SVR and KRR) were assessed concerning their robustness to predict the corrosion inhibition of small organic compounds for AZ91. The blind tests for the models were carried out to assess the reliability of each model. With the data set expanding in size and diversity in the future, similarity connections can be improved to increase the domain of applicability of the model. Either of the described model approaches can then be applied to predict the corrosion inhibition behaviors of a large amount of organic compounds with higher confidence and select promising inhibitors for AZ91, thus significantly decreasing material costs and environmental impact of experiments while accelerating the discovery of effective corrosion inhibitors.

In summary, small organic molecules exhibit great potential to control the corrosion

behavior of magnesium-based materials. Selecting effective organic corrosion inhibitors from the vast space of available compounds is not a trivial task and it cannot be solved by time- and resource-consuming experimental investigations alone. Quantitative structure-property relationship models based on supervised learning techniques such as SVR and KRR create great efficiencies in screening for effective agents for corrosion control.

In this work, the RBF kernel was used to develop two predictive data-driven models based on the available experimental inhibition efficiencies (IEs) of organic compounds for AZ91 from a previous work.[7] A pool of 876 input features derived from the cheminformatics software package and density functional theory calculations (DFT) were generated and exposed to an initial feature selection based on recursive feature elimination to identify the feature group consisting of 25 features with the highest relevance for the target property. These 25 features were subsequently gradually reduced to find the optimal number of features for the respective method and the results indicate that lowest RMSE is obtained for 8 features in the SVR and for 11 features in the KRR approach. There is a considerable overlap between the two groups of selected features as the energy levels of the highest occupied molecular orbital (HOMO) derived from DFT, 3D-MoRSE descriptors and 2D autocorrelations descriptors ended up in the final model for both cases, which agrees well with the findings in our previous work.[37]

Blind tests were carried out to assess the performance of the two model frameworks that were investigated in this work. Both models provide robust estimates for the IE of the untested chemicals. Of the ten compounds in the blind tests, 1-Acetylpiperazine (**1**) and 1-Amino-4-methylpiperazine (**2**) were predicted correctly to be strong accelerators with IE values more negative than -100% by both models. However, the predicted values were not quantitatively correct and neither of the models was able to predict the real values of the two compounds (**1**, $IE_{\text{exp}} = -563\%$ and **2**, $IE_{\text{exp}} = -517\%$) since the IE range of the compounds in the initial data set is limited to -227% IE. For the other eight compounds, 2-Hydroxyphenylacetic acid, 3-Methylpyrazole, Tartronic acid and Pyromellitic acid were correctly predicted by both models, where the values predicted by the SVR model are closer to the real values compared to the KRR model. In addition, both models identified 5-Nitouracil and Trimethylolpropane as outliers, although there are two more outliers for the SVR model. For each of the outliers, we observed that there is a distinct variation for the IEs of its top 5 highly similar structures extracted from the data set, which might ultimately

317 cause the false prediction of the IE value. This indicates that the similarity connection of
318 the structures is limited by available data.

319 Moreover, modulators exhibiting an aliphatic primary amine ($n\text{RNH}_2$), e.g. in an amino
320 acid, or fragments with the general formula $\text{R}=\text{CHX}$ (an alkene with a terminal function-
321 ality X that can either be O, N, S, P, Se or a halogene) cause elevated corrosion rates in
322 experimental studies.[7] The results indicate that small organic molecules that exhibit either
323 of the above mentioned functional moiety can most likely be excluded from the screening
324 for effective corrosion inhibitors. However, they might have beneficial properties for other
325 applications such as battery electrolyte additives where a controlled dissolution of the Mg-
326 based anode material is required.[65] Additionally, organic compounds with nitro groups
327 should be investigated in more detail from experiments since they exhibit rather positive in-
328 hibition efficiencies. After that, these experimentally investigated compounds can be added
329 to the data set to characterize this moiety and as a consequence to improve the accuracy of
330 the predictions for untested structures. Nevertheless, it was shown in this work that data-
331 driven models based on SVR and KRR approaches can be applied to predict the corrosion
332 inhibition efficiencies. Feeding more training samples to the model will facilitate an active
333 design of experiments thereby accelerating the selection of potent inhibitors for AZ91 and
334 other materials. Next, the selected inhibitors can be investigated for intercalation in LDH to
335 achieve an active corrosion protection of AZ91. Finally, the machine-learning based strate-
336 gies developed in this work can also be adapted to explore quantitative structure-property
337 relationships in different application fields given sufficient training data is available to train
338 the respective models.

339 III. METHODS

340 58 organic compounds were extracted from the work of Lamaka et al.[7] for AZ91 and
341 used as data base in this work. These 58 organic compounds were selected based on the
342 following three requirements: the concentration of the tested inhibitor was 0.05 M in 0.5
343 wt.% sodium chloride electrolyte (NaCl) pH neutral aqueous solution, molecular weight ($<$
344 350 Da) and inhibition efficiencies ranging from -250% to 100%. The concentration was
345 selected to be 0.05 M due to the fact that the majority of organic compounds were measured
346 in this concentration for AZ91 and other concentrations influenced the inhibition efficiency

of a chemical compound.[7] The chemical space was explored in a limited range of molecular weight since we are interested in seeking for small molecular organic inhibitors. The selection of the inhibition efficiency range is a balance between the large number of compounds, which is beneficial to build a model, and the small range from the side of the accelerators since the exploration of strong accelerators is out of interest in this work.

A. Feature generation and selection

After the data extraction, the molecular structures of these 58 compounds were built and optimized in the DFT calculations at the TPSSh/def2SVP level of theory using the quantum chemical software package Gaussian 16 [52]. DFT-calculated features, especially the highest occupied (HOMO) and the lowest unoccupied molecular orbital (LUMO), have been shown to be correlated to the corrosion inhibition efficiencies of small organic molecules for some Mg-based materials.[38, 66, 67] The optimized structures from DFT were subsequently used as input in the cheminformatics software package alvaDesc 1.0.22 [51] to generate more features, which were then combined with the HOMO and LUMO features to the initial feature set. There are over 800 features for each compound in the initial feature set, which significantly exceeds the number of compounds in the initial data set. At first, we applied RFE based on random forests to select the 25-tuple features, thus initially reducing the feature space. These selected 25 features can be different if we repeat the selection procedure due to the random initialization in the random forests. We repeated the selection procedure 50 times and obtained 50 different groups of selected top 25 features. The feature group with the lowest averaged test RMSE of the 5-fold cross validation (as illustrated in the Supporting Information Figure S3) in the SVR model was picked out of the 50 feature groups and is the basis for searching the most relevant features for the SVR and KRR models, respectively. We reduced the 25 features in a stepwise manner (one feature per step) to remove insignificant features in the model training. In each step, there is more than one possibility to remove one of the total features and all possibilities were investigated. The option which yielded the lowest averaged test RMSE was selected at each step and the preserved features were used for the next step. The number of considered features ranged from 25 to 1. Applying this method, we were able to select the most relevant features which obtained the lowest averaged test RMSE for the SVR and KRR models, respectively. After the selection of the

377 optimal features for each model, the continued stepwise procedure resulted in an order of
378 importance for the selected features, depending on their removed order.

379 B. Support Vector Regression & Kernel Ridge Regression

380 SVMs were initially developed as a supervised algorithm for classification. They are con-
381 structed as an optimization problem by finding the separation hyperplane with the maximum
382 margin, while maintaining correct predictions for most of the training points. [47, 68, 69] The
383 concept of SVMs can be adapted and be applicable to regression problems, which evolved
384 into SVR. In general, SVR builds a connection between a high-dimensional input vector and
385 one-dimensional target values. [48] This connection can be linear, as shown in Figure 7a,
386 and nonlinear with the assist of the kernel trick [70]. For SVR, a kernel function can map
387 the nonlinear distribution data in the input space to a higher-dimensional space where the
388 regression can be in a linear form. The RBF kernel has been used in SVMs for both classifi-
389 cation [42, 71, 72] and regression [73, 74] with considerable success. Apart from SVR, KRR
390 [75] is known as a similar model form as SVR [48] with the application of a different error
391 loss function. While KRR applies a squared error loss, SVR employs an ε -insensitive loss
392 as illustrated in Figure 7a and b. In this work, these two methods were applied in solving
393 the nonlinear connection between the input features and the target inhibition efficiency of
394 small organic compounds with the assist of a RBF kernel.

395 We applied the same feature selection process for both the SVR and KRR models and
396 obtained the most relevant features for each model, respectively. In this work, the high-
397 dimensional input vector is composed of the previously identified most relevant features
398 and the target values are the experimental inhibition efficiency extracted from the work of
399 Lamaka et al.[7]. The regression is achieved by ε -SVR [76] and KRR, and the results obtained
400 from these two methods are compared and discussed in this work. Hyperparameters such as
401 γ of the RBF kernel (as seen in the Supporting Information Figure S4), the regularization
402 parameter C, which manages the trade-off between the smoothness and overfitting of the ε -
403 SVR, and the regularization parameter α for a similar trade-off function in the KRR model,
404 are tuned in a 5-fold grid search to find optimal values with respect to the target property.
405 Except for these three mentioned parameters, the random state parameter (`random_state`)
406 which controls the split of the train and test sets was also tuned in this work to avoid the

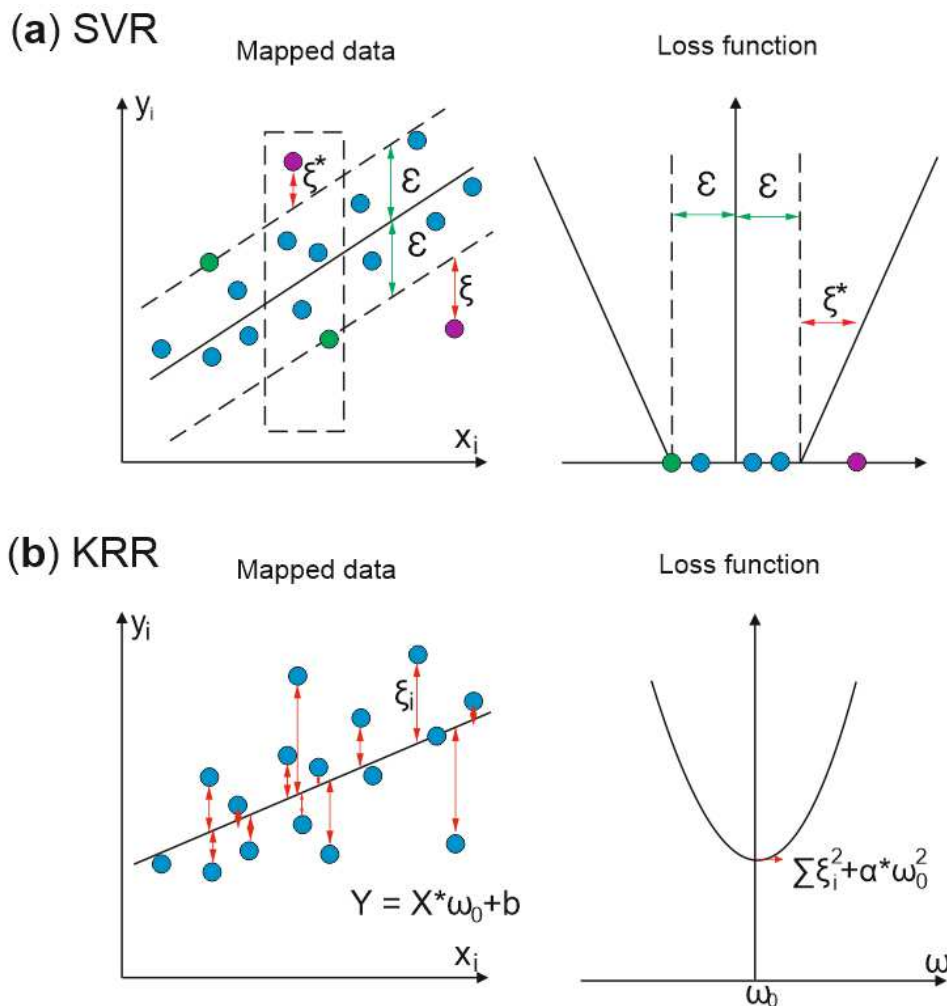


FIG. 7: Schematic diagrams of (a) a SVR with the ϵ -insensitive loss function and (b) a KRR with the squared error loss function.

biased split because of the relatively small data set (58 compounds) and large inhibition efficiency range (from -250% to 100%). The distribution of the inhibition efficiencies is provided in the Supporting Information (Figure S5).

C. Similarity calculation

The similarity calculation used in this work is based on a distance metric where the selected input features are the coordinates of each compound in the corresponding high dimensional feature space. The RBF kernel used in the SVR and KRR model was applied

in the similarity calculation, which is defined as

$$k(x, y) = \exp(-\gamma \|x - y\|^2), \quad (1)$$

where x and y are the vectors of the selected input features for two compounds, respectively.

D. Corrosion Experiments

The data set used in building the SVR and KRR models was extracted from the work of Lamaka et al.[7] and therefore the validation for these two models (blind tests) has been carried out with the same experimental setup and under the same conditions. The IE of compounds was calculated based on a hydrogen evolution test, in which the amount of evolved hydrogen due to the corrosion of magnesium is measured during immersion in a NaCl solution. 0.5 g of AZ91 Mg chips with the surface area of $430 \pm 29 \text{ cm}^2/\text{g}$ from the same batch used in work of Lamaka et al was immersed in 0.5 wt.% NaCl solution without (reference solution) and with the untested compounds. The concentration of compounds was 0.05 M and the pH of solutions was adjusted to 7 ± 0.1 by NaOH/HCl. The hydrogen evolution measurements were repeated three times for each solution and the average of calculated IEs was used for the corresponding blind test data point. The IE was defined by the following equation

$$\text{IE} = \frac{V_{\text{H}_2}^0 - V_{\text{H}_2}^{\text{Inh}}}{V_{\text{H}_2}^0} 100\%, \quad (2)$$

where $V_{\text{H}_2}^0$ and $V_{\text{H}_2}^{\text{Inh}}$ are the volumes of H_2 evolved after 20 h of immersion in the reference NaCl solution and the NaCl solution containing the investigated chemical compound, respectively. More details on the hydrogen evolution tests are available in the original publication[7].

IV. DATA AVAILABILITY

The authors declare that the primary data supporting the results of this study can be found in the paper and its supporting information. Other relevant data are available upon reasonable request.

V. CODE AVAILABILITY

The code used for this study will be uploaded to a suitable repository upon acceptance of the manuscript.

VI. ACKNOWLEDGEMENTS

Funding by the Helmholtz-Zentrum Hereon I2B project MUFFin is gratefully acknowledged. The authors thank Thermo Fisher Scientific for providing a chemical database used for the blind test selection.

VII. AUTHOR CONTRIBUTIONS

X.L., B.V., T.W., S.V.L., M.L.Z and C.F: contributed to the conception and design of the study. B.V. and S.V.L.: provided experimental data. X.L., T.W., and C.F: built the two machine learning models. X.L. and C.F.: wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read and approved the submitted version.

VIII. COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

- [1] Tan, J. & Ramakrishna, S. Applications of magnesium and its alloys: a review. *Applied Sciences* **11**, 6861 (2021).
- [2] Landkof, B. Magnesium applications in aerospace and electronic industries. *Magnesium Alloys and their Applications* 168–172 (2000).
- [3] Luan, B., Yang, D., Liu, X. & Song, G.-L. Corrosion protection of magnesium (mg) alloys using conversion and electrophoretic coatings. In *Corrosion of Magnesium Alloys*, 541–564 (Elsevier, 2011).
- [4] Chen, X.-B., Easton, M., Birbilis, N., Yang, H.-Y. & Abbott, T. Corrosion-resistant coatings for magnesium (mg) alloys. *Corrosion prevention of magnesium alloys* 282–312 (2013).

- [5] Pommiers, S., Frayret, J., Castetbon, A. & Potin-Gautier, M. Alternative conversion coatings to chromate for the protection of magnesium alloys. *Corrosion Science* **84**, 135–146 (2014).
- [6] Zhang, G. *et al.* Corrosion protection properties of different inhibitors containing peo/ldhs composite coating on magnesium alloy az31. *Scientific Reports* **11**, 1–14 (2021).
- [7] Lamaka, S. *et al.* Comprehensive screening of mg corrosion inhibitors. *Corrosion Science* **128**, 224–240 (2017).
- [8] Hu, H., Nie, X. & Ma, Y. Corrosion and surface treatment of magnesium alloys. *Magnesium Alloys-Properties in Solid and Liquid States* 67–108 (2014).
- [9] Latnikova, A. *Polymeric capsules for self-healing anticorrosion coatings*. Ph.D. thesis, Universität Potsdam (2012).
- [10] Denissen, P. J., Shkirskiy, V., Volovitch, P. & Garcia, S. J. Corrosion inhibition at scribed locations in coated aa2024-t3 by cerium-and dmt-d-loaded natural silica microparticles under continuous immersion and wet/dry cyclic exposure. *ACS applied materials & interfaces* **12**, 23417–23431 (2020).
- [11] Yin, Y., Prabhakar, M., Ebbinghaus, P., da Silva, C. C. & Rohwerder, M. Neutral inhibitor molecules entrapped into polypyrrole network for corrosion protection. *Chemical Engineering Journal* **440**, 135739 (2022).
- [12] Zheludkevich, M. *et al.* Active protection coatings with layered double hydroxide nanocontainers of corrosion inhibitor. *Corrosion Science* **52**, 602–611 (2010).
- [13] Zhang, X. *et al.* Active corrosion protection of mg–al layered double hydroxide for magnesium alloys: A short review. *Coatings* **11**, 1316 (2021).
- [14] Jing, C., Dong, B., Raza, A., Zhang, T. & Zhang, Y. Corrosion inhibition of layered double hydroxides for metal-based systems. *Nano Materials Science* **3**, 47–67 (2021).
- [15] Li, D. *et al.* Anticorrosion organic coating with layered double hydroxide loaded with corrosion inhibitor of tungstate. *Progress in organic coatings* **71**, 302–309 (2011).
- [16] Yu, X. *et al.* One-step synthesis of lamellar molybdate pillared hydrotalcite and its application for az31 mg alloy protection. *Solid State Sciences* **11**, 376–381 (2009).
- [17] Poznyak, S. *et al.* Novel inorganic host layered double hydroxides intercalated with guest organic inhibitors for anticorrosion applications. *ACS Applied Materials & Interfaces* **1**, 2353–2362 (2009).

- [18] Zhang, F. *et al.* Corrosion resistance of superhydrophobic layered double hydroxide films on aluminum. *Angewandte Chemie* **120**, 2500–2503 (2008).
- [19] Anjum, M. J. *et al.* Green corrosion inhibitors intercalated mg: Al layered double hydroxide coatings to protect mg alloy. *Rare Metals* **40**, 2254–2265 (2021).
- [20] Tabish, M. *et al.* Reviewing the current status of layered double hydroxide-based smart nanocontainers for corrosion inhibiting applications. *Journal of Materials Research and Technology* **10**, 390–421 (2021).
- [21] Würger, T. *et al.* Exploring structure-property relationships in magnesium dissolution modulators. *npj Materials degradation* **5**, 1–10 (2021).
- [22] Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5**, 1–36 (2019).
- [23] Popel, M. *et al.* Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications* **11**, 1–15 (2020).
- [24] Sharma, S., Bhatt, M. & Sharma, P. Face recognition system using machine learning algorithm. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 1162–1168 (IEEE, 2020).
- [25] Gormley, A. J. & Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nature Reviews Materials* **6**, 642–644 (2021).
- [26] Hart, G. L., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nature Reviews Materials* **6**, 730–755 (2021).
- [27] Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals. *Physical review letters* **117**, 135502 (2016).
- [28] Schmidt, J., Chen, L., Botti, S. & Marques, M. A. Predicting the stability of ternary intermetallics with density functional theory and machine learning. *The Journal of chemical physics* **148**, 241728 (2018).
- [29] Kim, K. *et al.* Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary heusler compounds. *Physical Review Materials* **2**, 123801 (2018).
- [30] Graser, J., Kauwe, S. K. & Sparks, T. D. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. *Chemistry of Materials* **30**, 3601–3612 (2018).

- [31] Oliynyk, A. O., Adutwum, L. A., Harynuk, J. J. & Mar, A. Classifying crystal structures of binary compounds ab through cluster resolution feature selection and support vector machine analysis. *Chemistry of Materials* **28**, 6672–6681 (2016).
- [32] Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters* **9**, 1668–1673 (2018).
- [33] Isayev, O. *et al.* Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials* **27**, 735–743 (2015).
- [34] De Jong, M. *et al.* A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Scientific reports* **6**, 1–11 (2016).
- [35] Winkler, D. A. *et al.* Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors. *Corrosion Science* **106**, 229–235 (2016).
- [36] Galvão, T. L., Novell-Leruth, G., Kuznetsova, A., Tedim, J. & Gomes, J. R. Elucidating structure–property relationships in aluminum alloy corrosion inhibitors by machine learning. *The Journal of Physical Chemistry C* **124**, 5624–5635 (2020).
- [37] Schiessler, E. J. *et al.* Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning models. *npj Computational Materials* **7**, 1–9 (2021).
- [38] Feiler, C. *et al.* In silico screening of modulators of magnesium dissolution. *Corrosion science* **163**, 108245 (2020).
- [39] White, P. A. *et al.* Towards materials discovery: assays for screening and study of chemical interactions of novel corrosion inhibitors in solution and coatings. *New Journal of Chemistry* **44**, 7647–7658 (2020).
- [40] Kokalj, A. Molecular modeling of organic corrosion inhibitors: Calculations, pitfalls, and conceptualization of molecule–surface bonding. *Corrosion Science* **193**, 109650 (2021).
- [41] Shulha, T. *et al.* In situ formation of ldh-based nanocontainers on the surface of az91 magnesium alloy and detailed investigation of their crystal structure. *Journal of Magnesium and Alloys* (2021).
- [42] Thurnhofer-Hemsi, K., López-Rubio, E., Molina-Cabello, M. A. & Najarian, K. Radial basis function kernel optimization for support vector machine classifiers. *arXiv preprint arXiv:2007.08233* (2020).
- [43] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. & Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*

552 **408**, 189–215 (2020).

553 [44] Kaur, P., Pannu, H. S. & Malhi, A. K. Plant disease recognition using fractional-order zernike
554 moments and svm classifier. *Neural Computing and Applications* **31**, 8749–8768 (2019).

555 [45] Bhowmik, T. K., Ghanty, P., Roy, A. & Parui, S. K. Svm-based hierarchical architectures for
556 handwritten bangla character recognition. *International Journal on Document Analysis and*
557 *Recognition (IJDAR)* **12**, 97–108 (2009).

558 [46] Je, H.-M., Kim, D. & Bang, S. Y. Human face detection in digital video using svmensemble.
559 *Neural processing letters* **17**, 239–252 (2003).

560 [47] Awad, M. & Khanna, R. Support vector regression. In *Efficient learning machines*, 67–80
561 (Springer, 2015).

562 [48] Okujeni, A. *et al.* A comparison of advanced regression algorithms for quantifying urban land
563 cover. *Remote Sensing* **6**, 6324–6346 (2014).

564 [49] Wehbe, B., Hildebrandt, M. & Kirchner, F. Experimental evaluation of various machine
565 learning regression methods for model identification of autonomous underwater vehicles. In
566 *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 4885–4890 (IEEE,
567 2017).

568 [50] Schölkopf, B., Luo, Z. & Vovk, V. *Empirical inference: Festschrift in honor of Vladimir N.*
569 *Vapnik* (Springer Science & Business Media, 2013).

570 [51] Mauri, A. alvades: A tool to calculate and analyze molecular descriptors and fingerprints. In
571 *Ecotoxicological QSARs*, 801–820 (Springer, 2020).

572 [52] Frisch, M. *et al.* Gaussian 16 revision c. 01, 2016. *Gaussian Inc. Wallingford CT* **1** (2016).

573 [53] Devinyak, O., Havrylyuk, D. & Lesyk, R. 3d-morse descriptors explained. *Journal of Molecular*
574 *Graphics and Modelling* **54**, 194–203 (2014).

575 [54] Caballero, J. Computational modeling to explain why 5, 5-diarylpentadienamides are trpv1
576 antagonists. *Molecules* **26**, 1765 (2021).

577 [55] Reid, B., Agri-Minerals, P. E. & Headquarters, C. Nop petition for inclusion of magnesium
578 oxide to the national list of substances allowed. *Cell* **850**, 261–0807 (2013).

579 [56] Case, D. R., Zubieta, J., Gonzalez, R. & Doyle, R. P. Synthesis and chemical and biological
580 evaluation of a glycine tripeptide chelate of magnesium. *Molecules* **26**, 2419 (2021).

581 [57] Murakami, Y. Complexing behavior of kojic acid with metal ions. i. mg (ii) and mn (ii)
582 chelates. *Bulletin of the Chemical Society of Japan* **35**, 52–56 (1962).

- [58] Kufelnicki, A. Complexes of uracil (2, 4-dihydroxypyrimidine) derivatives. part i. cu (ii), ca (ii) and mg (ii) coordination with uracil and related compounds in aqueous solution. *Polish Journal of Chemistry* **76**, 1559–1570 (2002).
- [59] Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia* **126**, 1763–1768 (2018).
- [60] Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine learning* **46**, 389–422 (2002).
- [61] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- [62] Deyab, M. Corrosion inhibition of heat exchanger tubing material (titanium) in msf desalination plants in acid cleaning solution using aromatic nitro compounds. *Desalination* **439**, 73–79 (2018).
- [63] Aslam, J. *et al.* Inhibitory effect of 2-nitroacridone on corrosion of low carbon steel in 1 m hcl solution: An experimental and theoretical approach. *Journal of Materials Research and Technology* **9**, 4061–4075 (2020).
- [64] Eddy, N. O., Ameh, P. O. & Essien, N. B. Experimental and computational chemistry studies on the inhibition of aluminium and mild steel in 0.1 m hcl by 3-nitrobenzoic acid. *Journal of Taibah University for Science* **12**, 545–556 (2018).
- [65] Würger, T. *et al.* Data-driven selection of electrolyte additives for aqueous magnesium batteries. *Journal of Materials Chemistry A* (2022).
- [66] Ju, H., Kai, Z.-P. & Li, Y. Aminic nitrogen-bearing polydentate schiff base compounds as corrosion inhibitors for iron in acidic media: a quantum chemical calculation. *Corrosion Science* **50**, 865–871 (2008).
- [67] Barouni, K. *et al.* Amino acids as corrosion inhibitors for copper in nitric acid medium: Experimental and theoretical study. *J. Mater. Environ. Sci* **5**, 456–463 (2014).
- [68] Kao, M.-Y. *Encyclopedia of algorithms* (Springer Science & Business Media, 2008).
- [69] Cristianini, N., Shawe-Taylor, J. *et al.* *An introduction to support vector machines and other kernel-based learning methods* (Cambridge university press, 2000).
- [70] Hofmann, T., Schölkopf, B. & Smola, A. J. Kernel methods in machine learning. *The annals of statistics* **36**, 1171–1220 (2008).

- 613 [71] Daqi, G. & Tao, Z. Support vector machine classifiers using rbf kernels with clustering-based
614 centers and widths. In *2007 international joint conference on neural networks*, 2971–2976
615 (IEEE, 2007).
- 616 [72] Das, S. R., Panigrahi, P. K., Das, K. & Mishra, D. Improving rbf kernel function of sup-
617 port vector machine using particle swarm optimization. *International Journal of Advanced*
618 *Computer Research* **2**, 130–5 (2012).
- 619 [73] Javed, F. *et al.* Rbf kernel based support vector regression to estimate the blood volume and
620 heart rate responses during hemodialysis. In *2009 annual international conference of the IEEE*
621 *engineering in medicine and biology society*, 4352–4355 (IEEE, 2009).
- 622 [74] Azimi, H., Bonakdari, H. & Ebtehaj, I. Design of radial basis function-based support vector
623 regression in predicting the discharge coefficient of a side weir in a trapezoidal channel. *Applied*
624 *Water Science* **9**, 1–12 (2019).
- 625 [75] Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).
- 626 [76] Chang, C.-C. & Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions*
627 *on intelligent systems and technology (TIST)* **2**, 1–27 (2011).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SIPredictingcorrosioninhibitionefficienciesofsmallorganicmoleculesusingdatadriventechiniques.pdf](#)