Elucidating Structure-Property relationships in Aluminum Alloy Corrosion Inhibitors by Machine Learning

Tiago L. P. Galvão[a,*], Gerard Novell-Leruth[a,b], Alena Kuznetsova[a], João Tedim[a], José R. B. Gomes[b]

[a] *CICECO-Aveiro Institute of Materials, Department of Materials and Ceramic Engineering, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal*

[b] *CICECO-Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal*

*Corresponding author.

E-mail addresses: tlpgalvao@ua.pt (T. L. P. Galvão); Phone: (+351) 938403485.

ABSTRACT. Organic corrosion inhibitors are playing a crucial role to substitute traditional protective technologies, which have acute toxicity problems associated. However, why some organic compounds inhibit corrosion and others do not, is still not well understood. Therefore, we tested different machine learning (ML) methods to distinguish efficient corrosion inhibitors for aluminum alloys commonly used in aeronautical applications. In this work, we have obtained information that can greatly contribute to automate the search for new and more efficient protective solutions in the future: i) a ML algorithm was selected that is able to classify correctly efficient inhibitors (i.e., with more than 50 % efficiency) and non-inhibitors (i.e. with lower-equal than 50 % efficiency), even when information about different alloys at different pHs is included in the same dataset, which can significantly increase the information available to train the model; ii) new descriptors related to the self-association of the molecules were evaluated, but improvements to the predictive power of the models are limited; iii) average differences concerning the descriptors in this work were identified for inhibitors and non-inhibitors, having the potential to serve as guidelines to select potentially inhibitive molecular systems. This work demonstrates that ML can significantly accelerate research in the field by serving as a tool to perform an initial virtual screen of the molecules.

1. INTRODUCTION

Corrosion Science is a scientific field aiming at preventing or, at least, controlling the deterioration of materials with the ultimate goal of maintaining the safety of infrastructures, thus adding great value to the global economy. Computer guided research has been giving valuable insights into corrosion mechanisms and the role of potentially corrosion inhibitive molecules[1,2]. This work makes a significant contribution to what can be a process of accelerating the search for novel compounds able to inhibit corrosion, by identifying the best machine learning (ML) methods for this purpose.

Metallic corrosion is typically the result of a redox electrochemical reaction characterized by an anodic part, responsible for the oxidation of the metal, and a cathodic part, resulting in the reduction of oxygen and water molecules from the environment[3]. The chemical reactivity of the process is described in Figure 1 for aluminum.

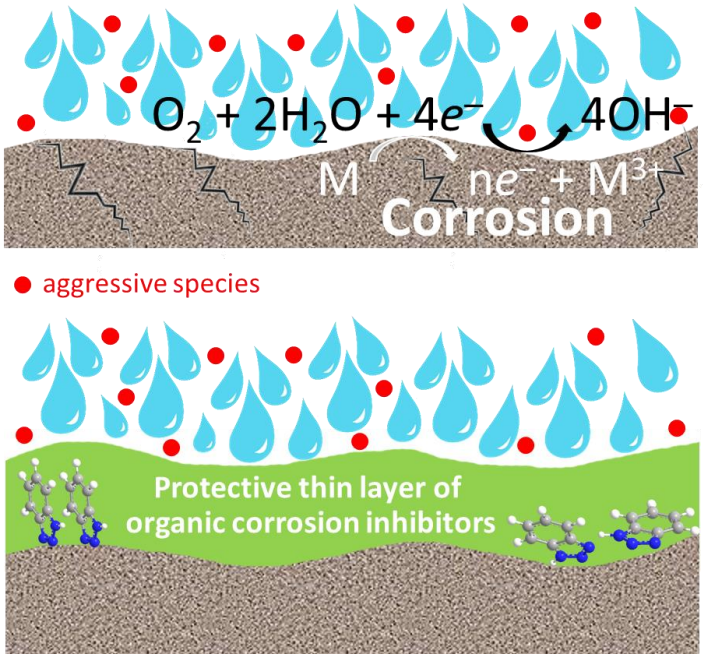

Figure 1. Corrosion electrochemical processes exemplified for an aluminum surface, involving cathodic and anodic reactions (above), and adsorption of a typical corrosion

inhibitor, such as 1,2,3-benzotriazole, to displace water molecules and protect the surface from aggressive species (below).

Aggressive species, such as, for example, Cl^-, SO_4^{2-} and OH^- can accelerate these chemical processes and, hence, the degradation of the surface. Therefore, organic corrosion inhibitors, such as, for example, 1,2,3-benzotriazole, are able to displace water molecules and protect the surface against aggressive species by adsorbing directly onto the metallic surface as represented in Figure 1.[3] Since the seminal work of Kokalj and his coworkers that used molecular modeling to understand what drives the inhibition efficiency of organic corrosion inhibitors[4], there have been important efforts in the atomistic simulation of corrosion processes[5,6] and protective technologies, such as nanostructured conversion films[7] and smart coating additives[8] (Figure 2).

The need to provide chromate-free corrosion protection technologies and the lack of understanding of how structural features of chemical compounds influence their corrosion efficiency, has led to the development of high-throughput testing methodologies of corrosion inhibitors[9–12]. This resulted in inhibition databases of different sizes[11,13] and for different types of metals[14], which in turn catalyzed the application of quantitative structure–property relationship (QSPR) approaches led by Winkler *et al.*[2], whose efforts even received the financial support of a major end-user, such as the Boeing company[11,15].

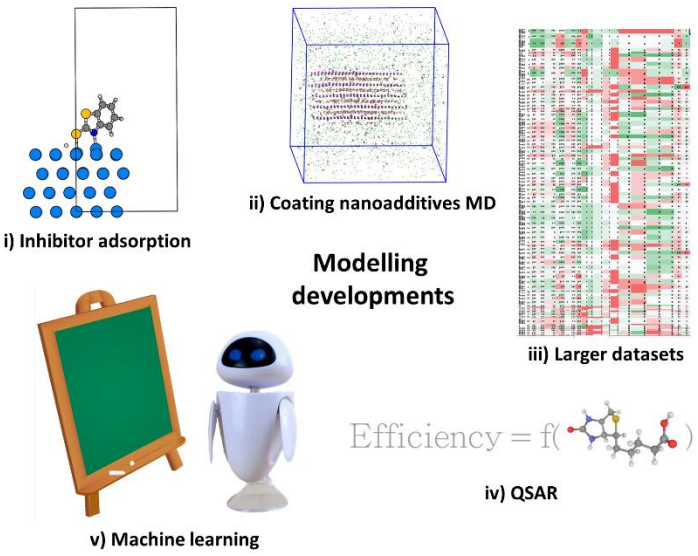

Figure 2. Some recent modelling developments in corrosion protection related to the understanding of phenomena at the molecular scale.

The availability of larger datasets led to the application of sophisticated ML approaches. Neural networks have been applied to model the inhibition efficiency of data sets containing information regarding 28 [13] and 100 [11] small organic compounds for aluminum alloys. Winkler *et al.*[11,15] concluded that molecular properties based on the structure of the inhibitors allowed to obtain a more reliable model than quantum chemical properties of the individual compounds (such as, for example, electronic affinity, ionization potential or electronegativity). To improve this situation, Fernandez *et al.*[16] developed a new methodology to obtain 3D distributions of electronegativity, polarizability and van der Waals volume for the test set containing 28 small organic molecules. Once the 3D properties were obtained, advanced algorithms were applied to verify what the inhibitor and non-inhibitor molecules had in common, and also predict the performance of new structures. This methodology yielded impressive predictive results for a small test set. Würger *et al.*[17] used a multidisciplinary approach combining high-throughput experimental screening of inhibitors efficiencies, unsupervised ML based on a clustering technique, plus data mining and density functional theory calculations, to estimate the corrosion inhibition efficiency of untested molecules for magnesium alloys. Despite the relatively low

amount of available experimental input data for benchmarking their methodology, the proposed workflow showed the existence of a clear relationship between corrosion inhibition efficiency and the molecular structure of molecules acting as magnesium corrosion inhibitors. Feiler *et al.*[18] have built upon this last work producing very promising results of model molecules for the corrosion protection of magnesium alloys.

In the present work, it will be performed a comparative analysis between different ML algorithms for the classification of chemical compounds as inhibitors or non-inhibitors for the corrosion protection of aluminum alloys commonly used in aeronautical applications. This work will enable a better understanding of the pros and cons of each algorithm to predict the corrosion inhibition potential of organic compounds.

We will also provide new descriptors related to the self-association between corrosion inhibitors in their simplest form, i.e., dimerization enthalpies and Gibbs energies. These descriptors can be a rough approximation of the strength of the interactions between the molecules themselves to form a cohesive thin film on the metallic surface. This film can also evolve to the formation of a multilayer[19], and prevents or, at least, diminishes the interaction of aggressive species with the metallic surface[20]. Although the calculation of dimerization energies is generally more computationally demanding than obtaining energetic quantum properties, it provides a more close parallelism with mechanistic processes (self-interaction of molecules during protective film formation) that cannot be achieved by correlating inhibition efficiencies with electronic properties alone.

2. MODELING DETAILS

Molecules and descriptors. In this work, a total of 102 organic compounds were examined, which include mostly aromatic moieties and/or amino, carboxylic, hydroxyl and

thiol groups, with the names of the compounds and their chemical structures presented in the SI (Table S1). Such compounds were previously tested experimentally as potential corrosion inhibitors by other authors for aerospace aluminum alloys AA2024 and AA7075. The compounds were examined under mild acidic (pH = 4) and basic (pH = 10) conditions using a high throughput experimental methodology[11], thus totaling 408 data entries. The efficiency of corrosion inhibition was assessed using an image processing methodology validated in a previous study through mass loss corrosion tests[21]. The efficiency scale ranged from zero (no inhibition) to 10 (maximum inhibition)[11], which can be readily rescaled to the percentage scale[21].

Besides obtaining experimental inhibition efficiencies, Winkler *et al.*[11] calculated a large number of descriptors (~2000) and made some available, such as molecular weight, molecular refractivity, octanol/water partition coefficient, polar surface area, molecular volume, molecular area, polar volume, number of donor atoms, number of rings (either aromatic or non-aromatic), number of hydrophobes, number of acceptor atoms and number of rotational bonds, which will also be used herein. The values of these descriptors were taken from Winkler *et al.*[11] without further change. Those authors employed three different codes to obtain the descriptors. The Sybyl x2.0 molecular modelling package (Certera Limited) was used to optimize the molecular structures, while the Dragon[22] and the Biomodeller[23,24] codes were used to calculate the descriptors.

Winkler *et al.* selected the most appropriate descriptors, while relying on linear regression and neural networks to obtain predictive models. They used between 11 and 31 descriptors depending on the alloy and pH[11]. Herein, we will use the same descriptors and different ML algorithms to model the four datasets, with the main aim of evaluating the performance of each algorithm. The dataset used in this work is available online [25]: http://dx.doi.org/10.17632/v5p322m2t8.2.

Dimerization energies. Prior to the optimization of the dimer structures to obtain the dimerization energetic parameters, potential conformations of dimers were generated with a conformational sampling method using a modified version of the Monte Carlo Multiple Minimum (MCMM) algorithm of Chang, Guida and Still[26], as described by Paton *et al.*[27,28]. The initial conformational sampling was performed using the MOPAC 2016 code[29] with the semi-empirical PM6 method incorporating dispersion and hydrogen bonding correction terms (PM6-DH2)[30,31], together with implicit solvation using the COSMO (Conductor-like Screening Model) method[32].

The three most favorable conformations of each dimer system from the preceding step were further optimized using the Gaussian 09 code[33] and dimerization enthalpies and Gibbs energies were calculated. The optimizations of the dimer and corresponding monomer molecular structures were performed with the hybrid density functional theory functional of Truhlar and Zhao (M06-2X)[34], together with the 6-31++G(d,p) basis set. The structures were characterized as true minima on the potential energy surface by the absence of imaginary modes in vibrational frequencies analyses performed also at the M06-2X/6-31++G(d,p) level of theory. The vibrational frequencies calculations were also used to obtain the thermal corrections to enthalpies and Gibbs free energies of dimerization, at T = 298.15 K. The influence of water was considered implicitly in these calculations by using the Polarizable Continuum Model (PCM)[35] as a self-consistent reaction field (SCRF) relying on the default values in the Gaussian code. The optimized dimer structures were made available online through iochem-bd[36].

Machine learning details. ML algorithms with a broad range of artificial intelligence applications and implemented in R were used herein to classify the compounds as inhibitors or non-inhibitors, in order to obtain a predictive analytical model. The algorithms tested were: *k*-nearest neighbors, decision trees, decision trees with Boosting, decision trees with defined error

costs, bagging, random forests, classification rules, artificial neural networks and support vector machine.

In order to evaluate our model, a 5-fold cross-validation method was employed (Figure 3). Although 10-fold cross-validation is more common, 5-fold cross-validation already achieves a similar low bias towards the test sample and equal mean square error for different methods and test sets[37], while being faster to employ for more computationally intensive ML techniques such as artificial neural networks. In 5-fold cross-validation, each algorithm is tested five times against five independent test samples corresponding to 20 % of the dataset after the model is also trained five different times against the remaining 80 %. Each algorithm is also tested for two different alloys at two different pHs. Furthermore, a test was also performed for the four datasets merged into one, thus totaling 408 data entries, including the type of alloy and pH as categorical variables. For the best method, it was employed both 5- and 10-fold cross-validation.

Moreover, for the composite dataset, where the 2 alloys and 2 pHs were modeled together, the three algorithms with the highest sensitivities were further evaluated employing 10-fold cross-validation and an independent test set with 20 % of the data.

In order to evaluate the role of the experimental error in the results obtained, further tests were performed with different thresholds to label the compounds as inhibitors or non-inhibitors: above 40 %, 50 %, 60 %, 70 % or 80 % (for equal or below these values, the compounds were labeled as non-inhibitors).

Prior to the training phase, the values of the descriptors were normalized using min-max normalization, according to equation 1.

$$x_{\mathrm{norm}} = \frac{x_{\mathrm{i}} - \min(x)}{\max(x) - \min(x)} \qquad (1)$$

In order to perform the classification task, the compounds were labeled in such a way that the group of compounds with the characteristic that we are searching for, corrosion inhibition, has in fact a high probably of not only inhibiting corrosion, but also doing so with above average efficiency. Therefore, the compounds with corrosion efficiencies higher than 50 % were labeled as corrosion inhibitors, whereas those with efficiencies of 50 % or lower were labeled as non-inhibitors. Note that the structural differences between inhibitors with efficiencies lower or equal to fifty percent (labeled as non-inhibitors) and higher but close to fifty percent (labeled as inhibitors) might be small, which introduces additional complexity that has to be dealt with by the algorithms. However, we had to split the full set of compounds in inhibitors and non-inhibitors using a specific cutoff line. In the present work, the threshold was chosen as the 50% inhibition efficiency, derived after rescaling the efficiency scale from Winkler *et al.*[11] as explained above.

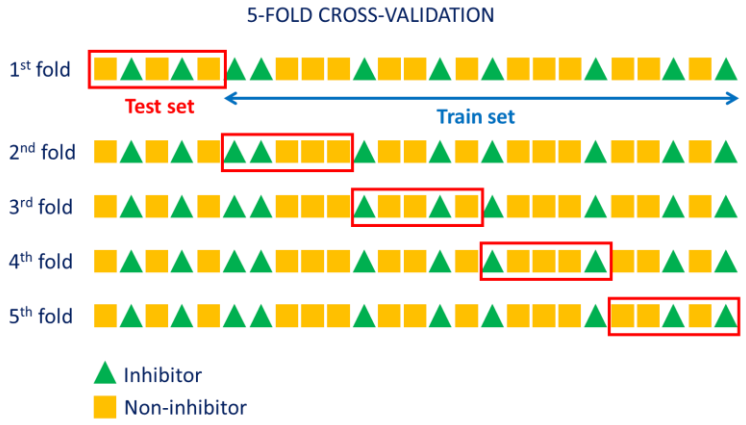

Figure 3. Exemplification of the 5-fold cross-validation statistical method employed in this work.

k-Nearest Neighbors. The *k*-nearest neighbors algorithm[38] measures the similarity between certain features to predict if a new entry in the dataset should be classified as belonging to a certain class or another, according to the labels of existing entries with similar features.

The knn() function from the R package *class* was used[39], since it provides a standard, classic implementation of the algorithm, and several *k* values from 1 to 10 were tested.

Decision trees. Decision trees is a ML method that divides the data into smaller and smaller portions until patterns are identified[40,41], thus forming a tree, similar to a flowchart, where the nodes indicate the decision made based on a certain descriptor. These nodes split into branches that indicate several decision possibilities. After several branches are subdivided into other branches, the tree ends with leaf nodes that denote the result of a combination of decisions at each branch node. The division of the data is stopped and a terminal node is reached when nearly all data entries at the final node share the same class, or there are no features to divide the data even more, or the three has grown until a predefined size.

This approach of dividing data into smaller subsets can be valuable to understand corrosion inhibitors, since some quantum chemical molecular properties have been successful to understand very small sets of corrosion inhibitors[4], but when applying a regression analysis to large datasets the same has not been true[11,15]. This approach can be worthwhile because, for example, the reason why smaller compounds inhibit might be different from why larger compounds inhibit, according to the representation in Figure 4.

The C5.0 algorithm developed by J. Ross Quinlan and implemented in the R package *C50* was used in this work, since it is the industry standard for producing decision trees[42].

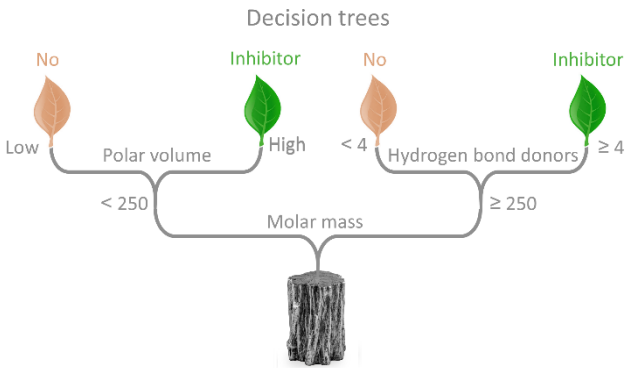

Figure 4. Example of how a decision trees ML method could work in the case of identifying corrosion inhibitors.

The application of the C5.0 algorithm was also accompanied by boosting[43], which is an ensemble technique that combines several weak classifiers to obtain a much more accurate one. When the algorithm attempts to model a class of one or several data entries in the dataset, if these data entries are modeled correctly, they are less likely to appear in further iterations of the decision tree when boosting is used. Therefore, additional iterations will focus solely on the more difficult data entry to classify.

The C5.0 algorithm also allows to define the cost of errors. Therefore, the cost of identifying a real corrosion inhibitor as a non-inhibitor was increased in comparison with identifying a non-inhibitor as an inhibitor. This can result in more false positives (predicted inhibitors which in reality are not) and less true negatives (less non-inhibitors are classified correctly), but might contribute to find more inhibitors overall increasing the true positives even if, in certain cases, the overall accuracy is lower.

Classification rules. The classification rules algorithm uses a set of *if-else* statements, to create conditions that allow to logically understand the data[44]. An example of a possible rule would be: "if organic compounds have high polar surface area and favorable dimerization energy, then they are corrosion inhibitors". When a rule can be applied to a subset of examples in the training data, then this subset is separated from the remaining of the training data. As more rules are added, additional subsets of data are separated until no more data entries remain in the training data. Ideally, classification rules are best employed to non-numerical data, but it was applied in this work because the rules are often readable in plain language.

In this work, it was used the repeated incremental pruning to produce error reduction (RIPPER) algorithm[44] implemented in R through the *RWeka* package, which is an advanced version of more basic classification rules.

Ensemble methods: bagging and random forests. One way to improve standalone ML methods is by using ensemble methods. Ensemble methods create and combine different models, by artificially varying the input data, to achieve better results than the original model alone. Two examples are bootstrapped aggregation (bagging)[45] and random forests[46] that work as represented in Figure 5.

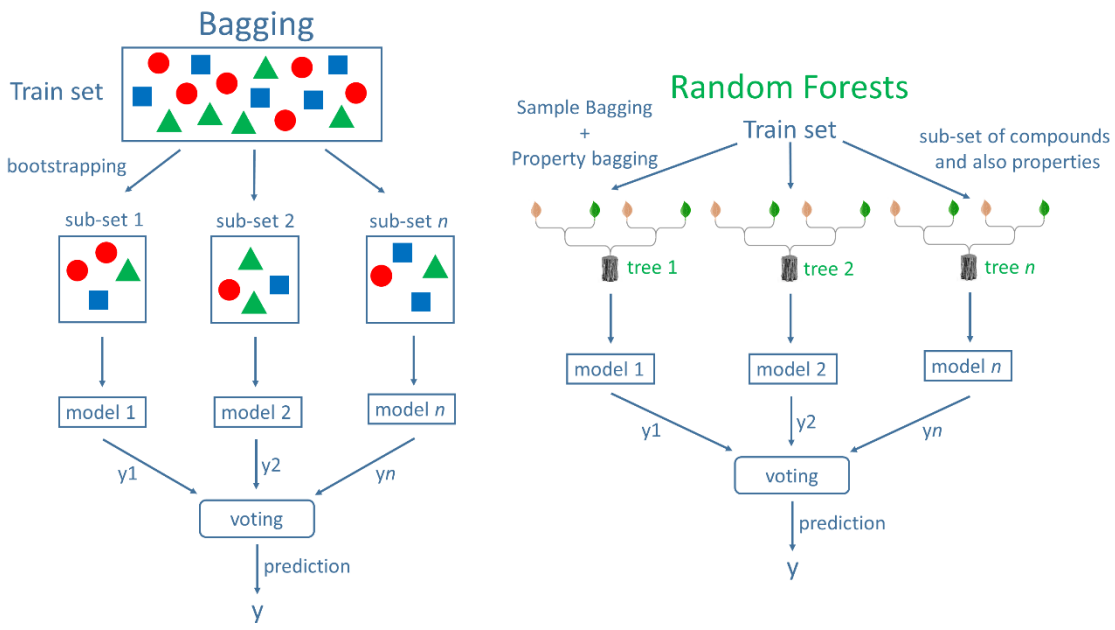

Figure 5. How the bagging (left) and random forests (right) ensemble methods work.

Bagging creates multiple models of the same type (usually, decision trees, but can be applied to other algorithms) from different sub-samples of the overall dataset, whereas random forests (or decision tree forests) combines samples bagging with property bagging. Random forests also act on the properties available to describe the outcome by reducing the number of properties used in each model and providing different trees at each decision node to generate

random sub-sets of the samples (with repetition) and properties (without repetition to decrease the correlation between different trees). Therefore, an ensemble of several decision trees is obtained, hence a forest. The *ipred*[45] and *randomForest*[46] packages were used for bagging and random forests, respectively.

Artificial Neural networks. Artificial neural networks use concepts inspired on how the biological brain works[47], employing a collection of nodes called artificial neurons, which can transmit information from one to another like the synapses in a biological brain. Artificial neural networks have an input layer (the properties of the inhibitors) and an output layer (the inhibition efficiencies). Depending on the algorithm formulation, it can include several intermediary layers (called hidden layers), each layer having a certain number of nodes (hidden nodes), as shown in Figure 6. The number of hidden layers and the number of nodes can be changed when formulating a model. If too few nodes are defined, the model might not describe well enough the train set, whereas, if too many nodes are defined, the model might over fit the train set, resulting usually in a bad performance for the test set.

Connecting the nodes, there are different weights that are iterated over the learning process, and an activation function. An activation function is a function that tells the neuron to pass the information onto the next neuron if a certain criterion is meet. In a first phase, the neurons are activated in sequence from the input layer to the output layer, applying each neuron's a random weight and an activation function along the way. When the signal reaches the output layer, a final signal with the outcome is produced. After this phase, a backward phase is initiated in which the network's final output signal resulting from the previous phase is compared with the true outcome of the training data. The difference between the model's output and the true output results in an error that is propagated backwards in the network to modify the weights that connect the different neurons and reduce future errors. To model our data, the *neuralnet* package was used[48], and it was tested up to ten nodes for each of a combination of

one, two or three hidden layers. The best results for AA2024 pH = 4, AA2024 pH = 10, AA7075 pH = 4, AA7075 pH = 10, and the composite mode were obtained with (8 × 7), (7 × 6 × 5), (2 × 10), (7 × 3 × 5), and (2 × 2) hidden nodes, respectively.

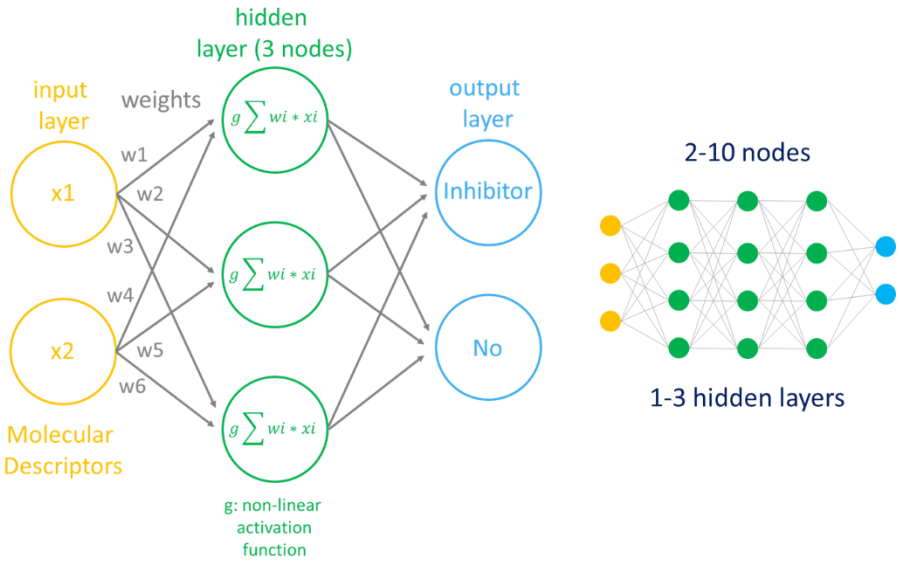

Figure 6. Basic demonstration of how a neural network works.

Support vector machine. The support vector machine algorithm combines concepts of *k*-nearest neighbors and linear regression modeling to produce a surface, called a hyperplane, which separates data according to different features. The hyperplane is calculated in such a way that the separation between classes of points is as wide as possible. Moreover, by adding new dimensions to the data, it is possible to separate non-linear data, using what is the so-called kernel trick, as represented in Figure 7. The *kernlab* package that implements a support vector machine algorithm[49] was used in this work.

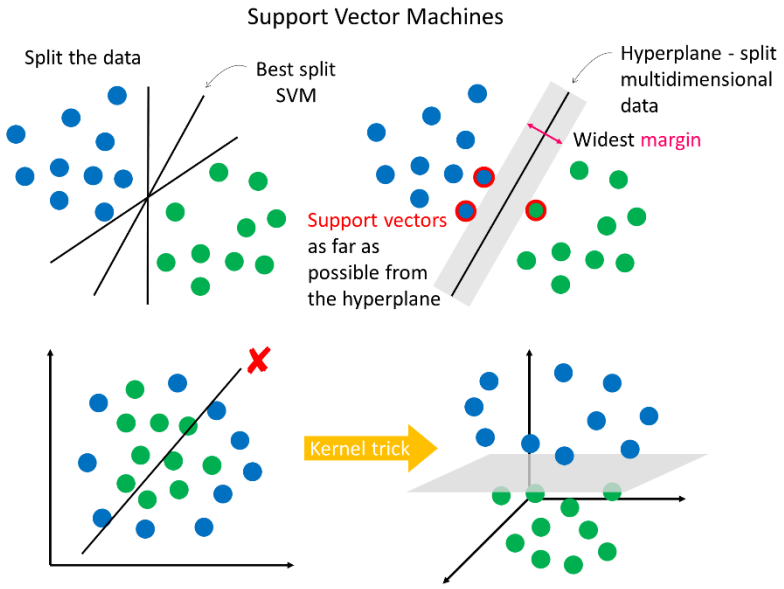

Figure 7. Hyperplane and kernel trick in the support vector machine algorithm.

Statistical parameters to evaluate the performance of the models. For classification tasks, performance measures consider the classes of the test set. The class of interest, herein corrosion inhibitors, is known as the positive class, while the other, the non-inhibitors, is known as negative. The relationship between both classes can be depicted in Figure 8 comprising four possible outcomes: i) true positives (TP): correctly classified corrosion inhibitors; ii) true negatives (TN): correctly classified non-corrosion inhibitors; iii) false positives (FP): compounds classified as corrosion inhibitors that are actually non-inhibitors; iv) false negatives (FN): compounds that were classified as non-corrosion inhibitors, but that are in fact corrosion inhibitors. Knowing this, the balanced accuracy, sensitivity and specificity defined by equations (2), (3) and (4) were used as measures of performance for classification.

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \quad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

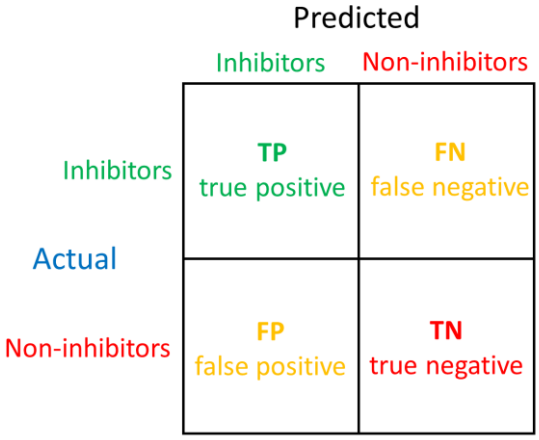

Figure 8. Four possible outcomes from the classification of inhibitors and non-inhibitors.

4. RESULTS AND DISCUSSION

Exploratory data analysis. Figure 10 schematically presents the main features of the experimental inhibition efficiencies evaluated herein by ML. It is possible to deduce that the

inhibitors are less efficient under basic conditions than under acidic conditions, according to the box plot.

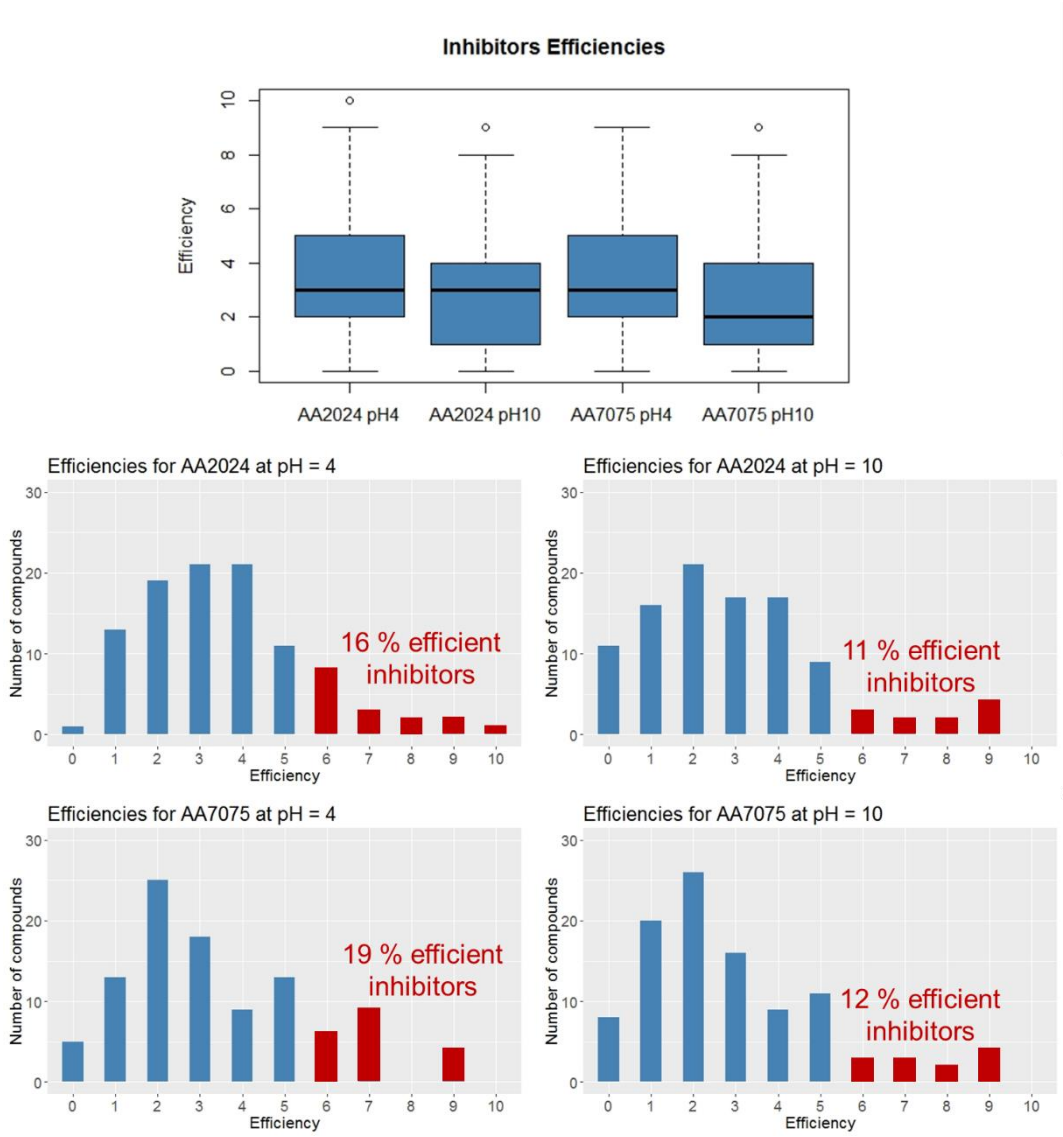

Figure 9. Box plot, showing the minimum, first quartile, median, third quartile, and maximum, (above) and bar plots (4 charts bellow) of the experimental inhibition efficiencies under different conditions evaluated in this work.

As can be noticed by the skewness of the bar plots (Figure 9) to the higher inhibition efficiencies, there is a considerably lower number of efficient corrosion inhibitors than weak inhibitors and non-inhibitors, which make it more challenging for ML algorithms to learn about the data and correctly identify compounds as inhibitors.

Figure 10 presents the correlation plot of the molecular descriptors and the inhibition efficiencies to be modeled in this work. Only the number of hydrogen bond donor atoms correlates reasonably well with the inhibition efficiencies, followed by the polar surface areas, polar volumes and number of rings. This demonstrates that the identification of molecular structures capable of being corrosion inhibitors is a highly non-linear problem, for which linear regression is not of much use.

Contrarily to other descriptors which can be self-correlated, the enthalpies and Gibbs energies of dimerization are not correlated at all with the remaining properties, thus making them differentiated properties.

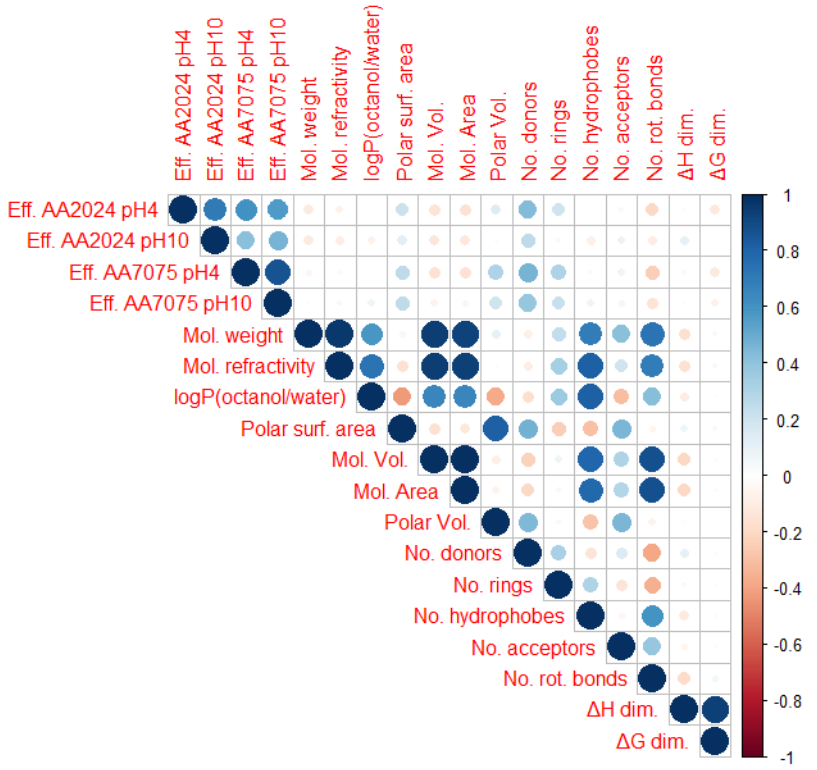

Figure 10. Linear correlation plot from the regression analysis of the inhibition efficiencies and descriptors modelled in this work. The blue color means positive linear correlation, whereas the red color means negative linear correlation. Larger circles mean a higher absolute

linear correlation value, whereas smaller or no circles mean lower absolute correlation value or zero linear correlation.

Classification. In order to identify the structures whose features can lead to corrosion inhibition of aluminum alloys, ML models were obtained for the train set, and then their performances evaluated against the test set, by means of 5-fold cross-validation. Figure 11 presents the average results obtained for each alloy under each pH value. Most methods had specificities of nearly 90 % or even higher. However, the high specificities (probability of identifying correctly the non-inhibitors) can be due to the higher number of non-inhibitors in the dataset. Therefore, their correct identification by chance is statistically more probable. On the other hand, the sensitivity refers to the probability of correctly classifying the true inhibitors, which are in lower number and are also the compounds with the condition of interest in this work (corrosion inhibition). In the dataset used in this work, only 10 to 20 % of the compounds have corrosion inhibition efficiencies above 50 % (labeled as corrosion inhibitors), thus, 80 to 90 % of the compounds have corrosion inhibition efficiencies equal or below 50 % (labeled as non-corrosion inhibitors). As a result, if, for example, a particular algorithm predicts that every compound is a non-corrosion inhibitor, it will have a specificity of 100 %, but its performance still would not be satisfactory, since it would not identify any corrosion inhibitor correctly. Therefore, the most important parameters are the balanced accuracy, used in this work similarly to other literature works because of the unbalanced number of inhibitors and non-inhibitors in the dataset[50], and sensitivity, which is the ratio of true corrosion inhibitors that were identified correctly.

When considering the average balanced accuracy for the four independent datasets corresponding to the 2 alloys at the two pH conditions, the following order of performance of the algorithms is obtained: decision rules < decision trees with boosting < decision trees <

bagging < decision trees with error costs < *k*-nearest neighbors < support vector machine < random forests < artificial neural networks. Regarding the sensitivity (correct identification of true corrosion inhibitors), the following 3 methods had the best performances for the individual datasets: decision trees with error costs < artificial neural networks < random forests. It is evident that the numeric nature of the descriptors does not favor the application of decision rules. Three different ensemble methods were tested with decision trees: boosting, bagging and random forests. Boosting focuses on the most difficult examples as the decision trees final model is optimized, which does not lead to improved results. However, sample bagging (combining the results of several different models obtained for several smaller resamples of the whole training set) leads to better results than the original decision trees. Likewise, random forests, which combines sample bagging with property bagging, improves the results even more, being the most sensitive of all methods. Clustering with *k*-nearest neighbors and support vector machine allow to correctly identify at least half of the total inhibitors, but neural networks and decision trees with error costs can on average identify more than sixty percent and almost seventy percent, respectively, of all the molecules that are able to inhibit corrosion. The error costs in decision trees increase the number of inhibitors that are correctly classified (true positives) at the expense of increasing the number of non-inhibitors that are incorrectly classified as inhibitors (false positives). However, this leads to a lower overall accuracy. This feature is also present in random forests. Neural networks is the most balanced method tested in this work for the individual datasets, with the best compromise between sensitivity and balanced accuracy, while random forests has the highest sensitivity.

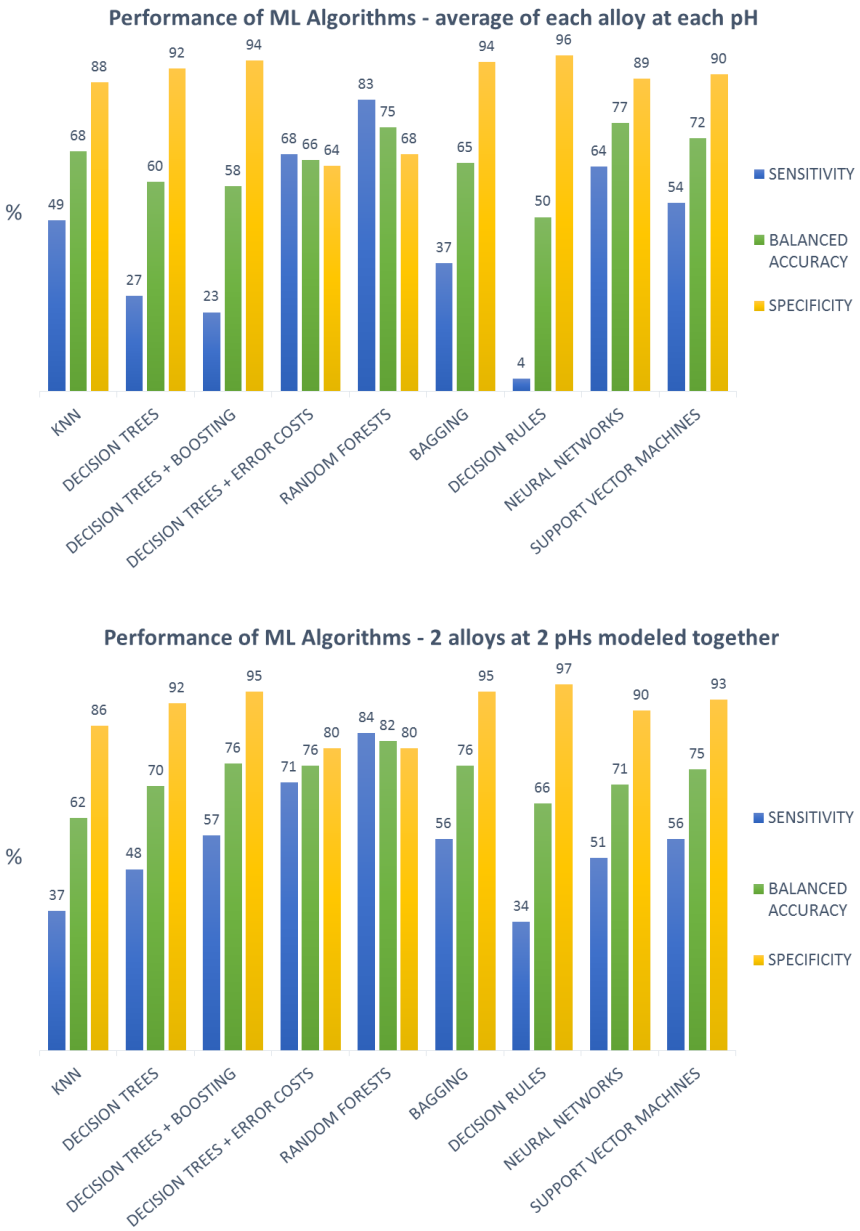

Figure 11. Average performance of ML algorithms for the two alloys under different pHs examined in this work (above) and performance of the algorithms for trained and tested for the composite dataset two alloys at two pHs included in the same dataset (bellow).

Two alloys under two pHs in a single dataset. In order to gather more data for the training phase of the algorithms, the four sub-datasets (102 examples each), corresponding to the two alloys, each tested for two pHs, were mixed in a single dataset containing a total of 408 data points. The objective of performing this test is that, if successful, it opens the door to build

larger datasets containing experimental tests corresponding to different alloys, conditions, and, desirably, also different metals.

Each data point was identified in the dataset with a categorical variable indicating the alloy and the pH that it corresponds to. Decision rules, decision trees and related algorithms with different ensemble methods can handle directly categorical variables, whereas *k*-nearest neighbors, neural networks and support vector machine, these variables (alloys and pH) were transformed into binary variables (0 and 1), indicating for each data point in question the alloy or pH it corresponded to.

In fact decision rules, and decision trees and its ensemble variations, all obtain better results, with higher balanced accuracy, sensitivity and specificity, when the two alloys at different pHs are trained together (Figure 11). This should be due to the larger amount of information available to train the models.

On the other hand, support vector machine produces only a marginal improvement for the larger dataset, whereas neural networks and *k*-nearest neighbors actually produce slightly worse results, especially in terms of sensitivity. This can be indicative of the difficulty in handling non-continuous variables by the latter algorithms.

Therefore, the different variations of decision trees, especially random forests, with a balanced accuracy and a specificity of 82 %, and a sensitivity of 84 %, look very encouraging to identify corrosion inhibitors from more complex datasets, including other factors and conditions, such as, for example, different types of metals, concentration of inhibitors and types of corrosion environments.

The Random Forests results obtained with 5-fold cross-validation were also validated using 10-fold cross-validation. The balanced accuracy, sensitivity and specificity are presented in Table 1. For the 10-fold cross-validation test, the results were slightly better than the results obtained with 5-fold cross-validation (Figure 11). It demonstrates the lack of bias towards a portion of the dataset by the less computational expensive 5-fold cross-validation method. This

also demonstrates the positive effect of having more data to train the algorithms, since for 10-fold cross-validation, 90 % of the dataset (~367 examples) is used to train the algorithm, instead of 80 % (~326 examples) as in the case of 5-fold cross-validation.

Moreover, the three algorithms with the highest sensitivities for the composite dataset, random forests, decision trees with error costs and decision trees with boosting, were further evaluated employing a training set, a validation set and a test set, in a two-phase approach, where in the first phase the model was trained and validated using 10-fold cross-validation for 80 % of the data, and, in the second phase, it was further tested against 20 % of the data. The results are presented in Table 1, where it is possible to observe that their performance is still very satisfactory even after using an independent test set of data not considered for cross validation.

Table 1. Sensitivity, accuracy and specificity for the performance tests made on the composite dataset, with the two alloys and pHs modelled together, of the three methods with the highest sensitivities.

Algorithm	Test	Sensitivity / %	Balanced accuracy / %	Specificity / %
Cross-validation on the complete dataset				
Random forests	10-fold cross-validation	87	85	83
Cross-validation on 80 % of the dataset, plus independent test on the remaining 20 %				
Random forests	10-fold cross-validation	83	82	80
Random forests	independent test	80	82	84
Decision trees + error costs	10-fold cross-validation	75	74	73
Decision trees + error costs	independent test	67	75	82

Decision trees + boosting	10-fold cross-validation	50	73	95
Decision trees + boosting	independent test	47	74	100

In order to evaluate the role of the experimental error in the results obtained, the prediction tests were repeated using the method with the highest sensitivity, random forests, for the composite dataset, considering different thresholds to label the compounds as inhibitors: above 40 %, 50 %, 60 %, 70 % or 80 % (for efficiencies equal or below these values, the compounds were labeled as non-inhibitors). The results were presented in SI.

Influence of dimerization. In this work, two new parameters were introduced to rank corrosion inhibitors: dimerization enthalpies and Gibbs energies. These parameters were calculated with state-of-the-art computational chemistry approaches to try to capture in an

indirect way the influence of the intermolecular interactions involved in the formation of protective films of corrosion inhibitors.

The most stable structures of the dimers are stabilized mostly by hydrogen bonding and non-covalent interactions (Table S2). Figure 12 presents the most frequent types of interaction, namely, O–H···O, O–H···N, N–H···N, S–H···N, N–H···S, and π···π.

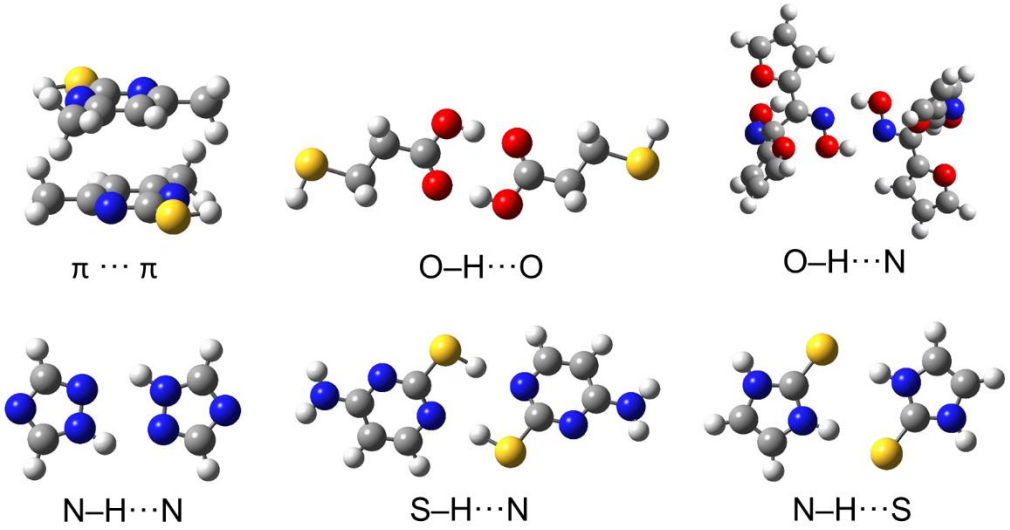

Figure 12. Molecular structures exemplifying the main types of intermolecular interactions resulting from the dimerization of the corrosion inhibitors in the datasets (spheres color code: carbon - grey; nitrogen - blue; oxygen - red; sulphur - yellow; hydrogen - white).

Therefore, the best performing methods for the individual datasets modeled separately were tested with and without these parameters, and it was found that the performance of the algorithms improves when information about dimerization energies is taken into account.

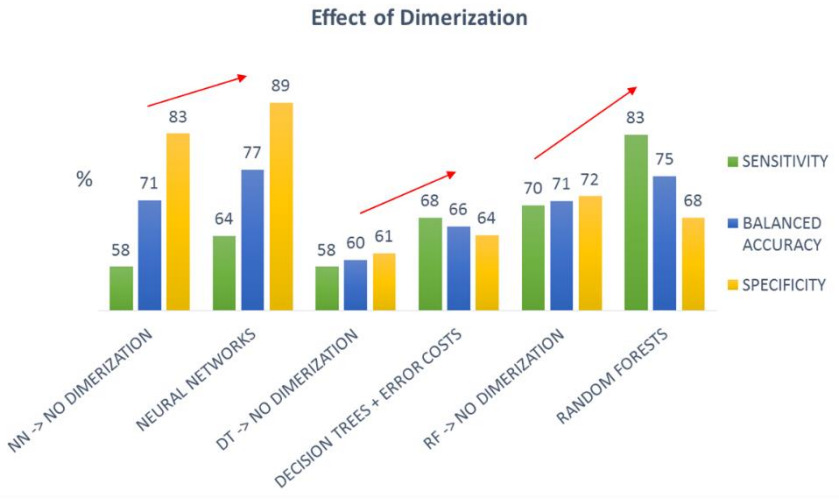

Figure 13. Effect of dimerization enthalpies and Gibbs energies on the average performance of three methods with high sensitivities for the individual alloys under different pHs. Neural networks (NN), decision trees (DT) and random forests (RF) were tested with and without dimerization enthalpies and Gibbs energies.

To compare the influence of dimerization with the influence of the other features for the composite model, tests were performed using random forests with and without each individual descriptor. From the results presented in Table 2, it is possible to verify that the Gibbs energy and enthalpy of dimerization have a similar effect to the other features, albeit very negligible for this composite model. Nevertheless, for the composite model, the best results are obtained when all the features are combined together and that no single feature is decisively more important than the others. The latter observation is justified by the consistently lower values of balanced accuracy and sensitivity, which are all within a similar range.

Table 2. Random forests performance for the composite model, considering the influence of each feature, by removing one for each test. The results correspond to 5-fold cross-validation.

Features considered	Sensitivity / %	Balanced accuracy / %	Specificity / %
All	84	82	80
No alloy	73	79	84
No pH	71	77	83
No mol. weight	78	78	78

No mol. refractivity	76	78	79
No LogP(octanol/water)	77	78	79
No polar surf. area	78	79	80
No mol. Vol.	75	78	80
No mol. Area	82	81	79
No polar Vol.	76	79	81
No no. donors	76	78	80
No no. rings	78	79	79
No hydrophobes	82	80	78
No no. acceptors	82	81	80
No no. rot. bonds	78	78	78
No ΔH(dim)	76	78	80
No ΔG(dim)	82	79	76

What characterizes a corrosion inhibitor? In order to gain further insights into the properties that distinguish inhibitors from non-inhibitors, the average values were obtained for the whole available data, with the results presented in Figure 14.

The results indicate that inhibitors have one or two aromatic rings with few or no rotational bonds (a large number of rotational bonds does not favor corrosion inhibition). They are also more polar, have between three and four hydrogen bond donor atoms, and have more favorable Gibbs energies of dimerization, despite the enthalpies of dimerization being also favorable but of the same magnitude for both inhibitors and non-inhibitors. The more favorable Gibbs energies, but similar enthalpies of dimerization, also point to the importance of the entropic term ($\Delta G_{dim} = \Delta H_{dim} - 298.15\ K \times \Delta S_{dim}$) to the formation of the dimers, which is also associated with higher inhibition efficiency.

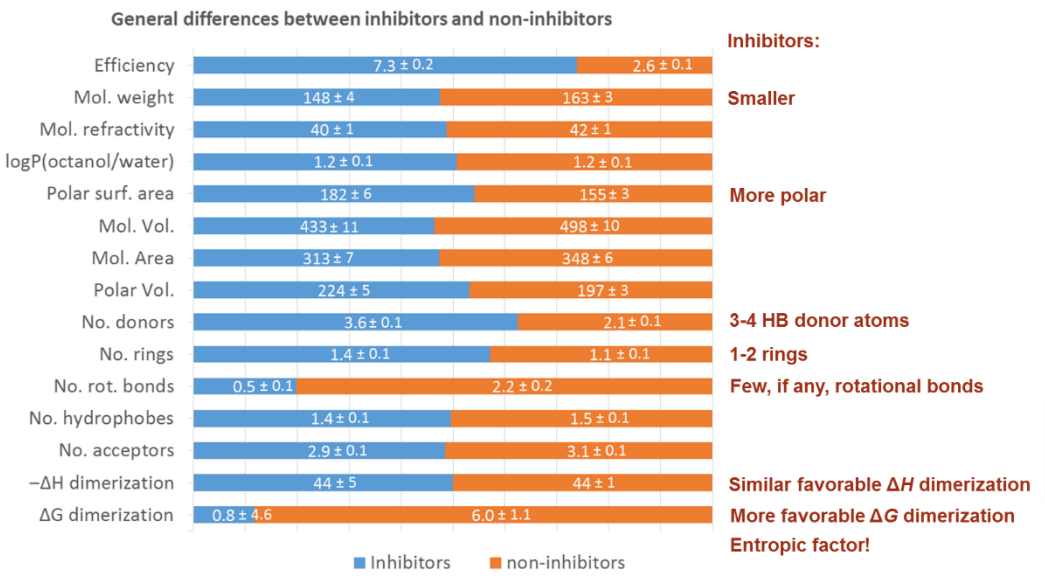

Figure 14. Percentage stack plot of the average properties of inhibitors and non-inhibitors with the respective standard error (left) and main characteristics of corrosion inhibitors in comparison with non-inhibitors (right).

5. CONCLUSIONS

This is foundational work to provide a support methodology that allows to restrict the number of unknown compounds that have to be tested experimentally by excluding the compounds classified as non-inhibitors and selecting those identified as inhibitors. For this purpose, random forests is the most successful method.

For the four independent datasets (AA2024 and AA7075, under pH = 4 and pH = 10), random forests had an average accuracy of 68 %, while being able to identify correctly 83 % of the inhibitors. If the four datasets are used together with the type of alloy and pH are included as categorical variables, the overall accuracy increases to 80 %, while identifying correctly 84 % of the inhibitors. This should be due to the increased amount of data that is provided to train the algorithm, and shows that in the future it might be possible to analyze data that include, not only different types of alloys and pHs, but, possibly, also other electrolyte conditions such as aggressive anions concentration, inhibitor concentration or even type of metal.

The dimerization enthalpy and Gibbs energy were introduced as indirect measures of the intermolecular interactions involved in the formation of protective films. By comparing the accuracy, sensitivity and specificity of algorithms with and without considering these parameters, it was verified that the dimerization thermodynamic properties have some importance to model the smaller individual datasets, but no significant importance to model the composite dataset.

A statistical analysis of the average value of the properties of inhibitors and non-inhibitors was performed, indicating that inhibitors have one or two aromatic rings with few or no rotational bonds. Moreover, a large number of rotational bonds seems to hinder corrosion inhibition. Inhibitors are also more polar, have between three and four hydrogen bond donor atoms in their structure, and have more favorable Gibbs energies of dimerization, despite the enthalpies of dimerization being also favorable, but of the same magnitude for both inhibitors and non-inhibitors, suggesting that the dimerization entropy is an important factor in the protection mechanism.

ASSOCIATED CONTENT

*Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: . The supporting information includes the names and chemical structures comprising the dataset studied in this work, as well as dimerization structures of the molecules.

ACKNOWLEDGEMENTS

This work was developed within the scope of the project CICECO-Aveiro Institute of Materials, UIDB/50011/2020 & UIDP/50011/2020, financed by national funds through the FCT/MEC and when appropriate co-financed by FEDER under the PT2020 Partnership

Agreement. It was also financed in the framework of the project DataCor (refs. POCI-01-0145-FEDER-030256 and PTDC/QUI-QFI/30256/2017), and SELMA (PTDC/QEQ-QFI/4719/2014), Project 3599 - Promover a Produção Científica e Desenvolvimento Tecnológico e a Constituição de Redes Temáticas (3599-PPCDT) and FEDER funds through COMPETE 2020, Programa Operacional Competitividade e Internacionalização (POCI).

References

(1) Galvão, T. L. P.; Wilhelm, M.; Gomes, J. R. B.; Tedim, J. Emerging trends in smart nanocontainers for corrosion applications. In *SMART NANOCONTAINERS*; ELSEVIER, 2019.

(2) Winkler, D. A. Predicting the Performance of Organic Corrosion Inhibitors. *Metals (Basel).* **2017**, *7* (12), 553 DOI: 10.3390/met7120553.

(3) Brycki, B. E.; Kowalczyk, I. H.; Szulc, A.; Kaczerewska, O.; Pakiet, M. Organic Corrosion Inhibitors. In *Corrosion Inhibitors, Principles and Recent Applications*; Aliofkhazraei, M., Ed.; IntechOpen, 2018.

(4) Kokalj, A.; Peljhan, S.; Finšgar, M.; Milošev, I. What Determines the Inhibition Effectiveness of ATA, BTAH, and BTAOH Corrosion Inhibitors on Copper? *J. Am. Chem. Soc.* **2010**, *132* (46), 16657–16668 DOI: 10.1021/ja107704y.

(5) Taylor, C. D. Modeling Corrosion, Atom by Atom. *Interface Mag.* **2014**, *23* (4), 59–64 DOI: 10.1149/2.F04144IF.

(6) Duda, Y.; Govea-Rueda, R.; Galicia, M.; Beltrán, H. I.; Zamudio-Rivera, L. S. Corrosion Inhibitors: Design, Performance, and Computer Simulations. *J. Phys. Chem. B* **2005**, *109* (47), 22674–22684 DOI: 10.1021/JP0522765.

(7) Galvão, T. L. P.; Neves, C. S.; Zheludkevich, M. L.; Gomes, J. R. B.; Tedim, J.; Ferreira, M. G. S. How Density Functional Theory Surface Energies May Explain the Morphology of Particles, Nanosheets, and Conversion Films Based on Layered Double Hydroxides. *J. Phys. Chem. C* **2017**, *121* (4), 2211–2220 DOI: 10.1021/acs.jpcc.6b10860.

(8) Pérez-Sánchez, G.; Galvão, T. L. P.; Tedim, J.; Gomes, J. R. B. A molecular dynamics framework to explore the structure and dynamics of layered double hydroxides. *Appl. Clay Sci.* **2018**, *163*, 164–177 DOI: 10.1016/J.CLAY.2018.06.037.

(9) Chambers, B. D.; Taylor, S. R.; Kendig, M. W. Rapid Discovery of Corrosion Inhibitors and Synergistic Combinations Using High-Throughput Screening Methods. *CORROSION* **2005**, *61* (5), 480–489 DOI: 10.5006/1.3280648.

(10) Kallip, S.; Bastos, A. C.; Zheludkevich, M. L.; Ferreira, M. G. S. A multi-electrode cell for high-throughput SVET screening of corrosion inhibitors. *Corros. Sci.* **2010**, *52* (9), 3146–3149 DOI: 10.1016/J.CORSCI.2010.05.018.

(11) Winkler, D. A.; Breedon, M.; White, P.; Hughes, A. E.; Sapper, E. D.; Cole, I. Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors. *Corros. Sci.* **2016**, *106*, 229–235 DOI: 10.1016/j.corsci.2016.02.008.

(12) García, S. J.; Muster, T. H.; Özkanat, Ö.; Sherman, N.; Hughes, A. E.; Terryn, H.; de Wit, J. H. W.; Mol, J. M. C. The influence of pH on corrosion inhibitor selection for

2024-T3 aluminium alloy assessed by high-throughput multielectrode and potentiodynamic testing. *Electrochim. Acta* **2010**, *55* (7), 2457–2465 DOI: 10.1016/J.ELECTACTA.2009.12.013.
(13) Harvey, T. G.; Hardin, S. G.; Hughes, A. E.; Muster, T. H.; White, P. A.; Markley, T. A.; Corrigan, P. A.; Mardel, J.; Garcia, S. J.; Mol, J. M. C.; et al. The effect of inhibitor structure on the corrosion of AA2024 and AA7075. *Corros. Sci.* **2011**, *53* (6), 2184–2190 DOI: 10.1016/j.corsci.2011.02.040.
(14) Lamaka, S. V.; Vaghefinazari, B.; Mei, D.; Petrauskas, R. P.; Höche, D.; Zheludkevich, M. L. Comprehensive screening of Mg corrosion inhibitors. *Corros. Sci.* **2017**, *128*, 224–240 DOI: 10.1016/J.CORSCI.2017.07.011.
(15) Winkler, D. A.; Breedon, M.; Hughes, A. E.; Burden, F. R.; Barnard, A. S.; Harvey, T. G.; Cole, I. Towards chromate-free corrosion inhibitors: structure–property models for organic alternatives. *Green Chem.* **2014**, *16*, 3349 DOI: 10.1039/c3gc42540a.
(16) Fernandez, M.; Breedon, M.; Cole, I. S.; Barnard, A. S. Modeling corrosion inhibition efficacy of small organic molecules as non-toxic chromate alternatives using comparative molecular surface analysis (CoMSA). *Chemosphere* **2016**, *160*, 80–88 DOI: 10.1016/j.chemosphere.2016.06.044.
(17) Würger, T.; Feiler, C.; Musil, F.; Feldbauer, G. B. V.; Höche, D.; Lamaka, S. V.; Zheludkevich, M. L.; Meißner, R. H. Data Science Based Mg Corrosion Engineering. *Front. Mater.* **2019**, *6*, 53 DOI: 10.3389/fmats.2019.00053.
(18) Feiler, C.; Mei, D.; Vaghefinazari, B.; Würger, T.; Meißner, R. H.; Luthringer-Feyerabend, B. J. C.; Winkler, D. A.; Zheludkevich, M. L.; Lamaka, S. V. In silico Screening of Modulators of Magnesium Dissolution. *Corros. Sci.* **2019**, 108245 DOI: 10.1016/J.CORSCI.2019.108245.
(19) Samide, A.; Iacobescu, G.; Tutunaru, B.; Grecu, R.; Tigae, C.; Spînu, C. Inhibitory Properties of Neomycin Thin Film Formed on Carbon Steel in Sulfuric Acid Solution: Electrochemical and AFM Investigation. *Coatings* **2017**, *7* (11), 181 DOI: 10.3390/coatings7110181.
(20) Taghavikish, M.; Dutta, N.; Roy Choudhury, N. Emerging Corrosion Inhibitors for Interfacial Coating. *Coatings* **2017**, *7* (12), 217 DOI: 10.3390/coatings7120217.
(21) White, P. A.; Smith, G. B.; Harvey, T. G.; Corrigan, P. A.; Glenn, M. A.; Lau, D.; Hardin, S. G.; Mardel, J.; Markley, T. A.; Muster, T. H.; et al. A new high-throughput method for corrosion testing. *Corros. Sci.* **2012**, *58*, 327–331 DOI: 10.1016/J.CORSCI.2012.01.016.
(22) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.
(23) Burden, F. R.; Polley, M. J.; Winkler, D. A. Toward Novel Universal Descriptors: Charge Fingerprints. *J. Chem. Inf. Model.* **2009**, *49* (3), 710–715 DOI: 10.1021/ci800290h.
(24) Winkler, D. A.; Burden, F. R. Robust QSAR Models from Novel Descriptors and Bayesian Regularised Neural Networks. *Mol. Simul.* **2000**, *24* (4–6), 243–258 DOI: 10.1080/08927020008022374.
(25) DATACOR Dataset V2. **2019**, *2* DOI: 10.17632/V5P322M2T8.2.
(26) Chang, G.; Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111* (12), 4379–4386 DOI: 10.1021/ja00194a035.
(27) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9* (3), 2313–2323 DOI: 10.1021/acscatal.8b04043.
(28) Paton, R. S. Full Monte https://github.com/bobbypaton/FullMonte (accessed May 7,

2019).
(29) Stewart, J. J. P. MOPAC2016. *Stewart Comput. Chem. Color. Springs, CO, USA* **2016**.
(30) Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. A Transferable H-Bonding Correction for Semiempirical Quantum-Chemical Methods. *J. Chem. Theory Comput.* **2010**, *6* (1), 344–352 DOI: 10.1021/ct900541n.
(31) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5* (7), 1749–1760 DOI: 10.1021/ct9000922.
(32) Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, No. 5, 799–805 DOI: 10.1039/P29930000799.
(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; et al. Gaussian 09, Rev A.1. *Gaussian Inc Wallingford CT*. Gaussian, Inc. 2009.
(34) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other function. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241 DOI: 10.1007/s00214-007-0310-x.
(35) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **2005**, *105* (8), 2999–3093 DOI: 10.1021/cr9904009.
(36) Galvão, T. L. P. Datacor dimer structures https://iochem-bd.bsc.es/browse/handle/100/192939 (accessed May 7, 2019).
(37) Molinaro, A. M.; Simon, R.; Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **2005**, *21* (15), 3301–3307 DOI: 10.1093/bioinformatics/bti499.
(38) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46* (3), 175–185 DOI: 10.1080/00031305.1992.10475879.
(39) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*; Statistics and Computing; Springer New York: New York, NY, 2002.
(40) Quinlan, J. R. Learning Efficient Classification Procedures and Their Application to Chess End Games. In *Machine Learning*; Springer Berlin Heidelberg: Berlin, Heidelberg, 1983; pp 463–482.
(41) Quinlan, J. R. Simplifying decision trees. *Int. J. Man. Mach. Stud.* **1987**, *27* (3), 221–234 DOI: 10.1016/S0020-7373(87)80053-6.
(42) Salzberg, S. L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **1994**, *16* (3), 235–240 DOI: 10.1007/BF00993309.
(43) Schapire, R. E.; Freund, Y. *Boosting : foundations and algorithms*; MIT Press, 2012.
(44) Cohen, W. W. Fast Effective Rule Induction. *Mach. Learn. Proc. 1995* **1995**, 115–123 DOI: 10.1016/B978-1-55860-377-6.50023-2.
(45) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123–140 DOI: 10.1023/A:1018054314350.
(46) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32 DOI: 10.1023/A:1010933404324.
(47) McCulloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5* (4), 115–133 DOI: 10.1007/BF02478259.
(48) Riedmiller, M.; Braun, H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In *IEEE International Conference on Neural Networks*; IEEE; pp 586–591.

(49) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1–27 DOI: 10.1145/1961189.1961199.
(50) Choi, J.-S.; Ha, M. K.; Trinh, T. X.; Yoon, T. H.; Byun, H.-G. Towards a generalized toxicity prediction model for oxide nanomaterials using integrated data from different sources. *Sci. Rep.* **2018**, *8* (1), 6110 DOI: 10.1038/s41598-018-24483-z.

Table of Contents