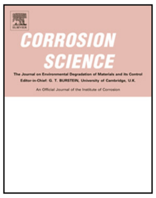
Corrosion Science 106 (2016) 229–235

Contents lists available at ScienceDirect

Corrosion Science

journal homepage: www.elsevier.com/locate/corsci



# Using high throughput experimental data and *in silico* models to discover alternatives to toxic chromate corrosion inhibitors

D.A. Winkler[a,b,c,d,*], M. Breedon[a], P. White[a], A.E. Hughes[a,c], E.D. Sapper[e], I. Cole[a]

[a] *CSIRO Manufacturing, Clayton 3168, Australia*
[b] *Monash Institute of Pharmaceutical Sciences, 399 Royal Parade, Parkville 3052, Australia*
[c] *Latrobe Institute for Molecular Science, Bundoora 3086, Australia*
[d] *School of Chemical and Physical Sciences Flinders University, Bedford Park, 5042, Australia*
[e] *Boeing Research & Technology, Seattle, WA 98124-2207, USA*

ARTICLE INFO

*Article history:*
Received 11 August 2015
Received in revised form 12 January 2016
Accepted 5 February 2016
Available online 11 February 2016

*Keywords:*
A. alloy
A. aluminium
B. modelling studies
C. neutral inhibition

ABSTRACT

Restrictions on the use of toxic chromate-based corrosion inhibitors have created important issues for the aerospace and other industries. Benign alternatives that offer similar or superior performance are needed. We used high throughput experiments to assess 100 small organic molecules as potential inhibitors of corrosion in aerospace aluminium alloys AA2024 and AA7075. We generated robust, predictive, quantitative computational models of inhibitor efficiency at two pH values using these data. The models identified molecular features of inhibitor molecules that had the greatest impact on corrosion inhibition. Models can be used to discover better corrosion inhibitors by screening libraries of organic compounds for candidates with high corrosion inhibition.

## 1. Introduction

Corrosion is clearly a very important and expensive environmental issue for large number of metallic materials used for a myriad of engineering purposes. In the aerospace industry, problems are particularly acute due to the high cost of the infrastructure and the catastrophic consequences of materials failure due to corrosion. Chromates are very effective corrosion inhibitors and have been the treatment of choice for the aerospace industry to date [1]. However, they pose an unacceptable risk to workers and the environment, particularly during production, because of their toxicity and carcinogenicity. Studies have shown that they are 'hot spot' pollutants. Recent epidemiological data from a study of chrome chemical production workers found the excess lifetime risk of death from lung cancer due to occupational exposures to be 255,000 per million workers [2], massively larger than the 'acceptable' risk of 1 death per million. Consequently, chromates are progressively being restricted or removed from service by legislation [3].

There has been an intense search for more benign replacements for chromate corrosion inhibitors that have equivalent or superior performance [4]. Small organic compounds are showing considerable promise as replacement inhibitors, and the recent literature has identified a number of them that show useful corrosion inhibition properties and potentially lower toxicity than exiting inhibitors. Finsgar has reviewed the efficacy and mechanism of action of benzotriazoles in preventing corrosion of copper [5]. Gece has likewise reviewed the potential of accessible small molecule drugs as corrosion inhibitors [6]. Very recently Kuznetsov reviewed the physicochemical aspects of protection of metals by organic corrosion inhibitors [7]. It has been estimated [8] that there are an immense number ($\sim 10^{80}$) of small organic molecules that could be synthesized, providing an essentially infinite source of potentially high performance, relatively benign corrosion inhibitors. Preliminary computational modelling studies by Winkler et al. have shown that the potential of a range of small organic compounds to inhibit corrosion in aerospace alloys can be predicted reliably [9]. However, this work demonstrated that small changes to the molecular structure of such organic molecules could result in dramatic changes in the performance of inhibitors. Considering the number of families of possible inhibitors and their isomers, there are tens of thousands of candidate inhibitor molecules.

Materials science is increasingly using automated methods of synthesis and characterization to accelerate the discovery and optimization of new materials. Modern machine learning methods using these large and rich data sets are showing have proven useful for the quantitative prediction of a wide range of materials prop-

[*] Corresponding author at: CSIRO Manufacturing, Bag 10, Clayton South MDC 3169, Australia. Fax: +61 395452446.
*E-mail address:* dave.winkler@csiro.au (D.A. Winkler).

http://dx.doi.org/10.1016/j.corsci.2016.02.008
0010-938X/Crown Copyright © 2016 Published by Elsevier Ltd. All rights reserved.

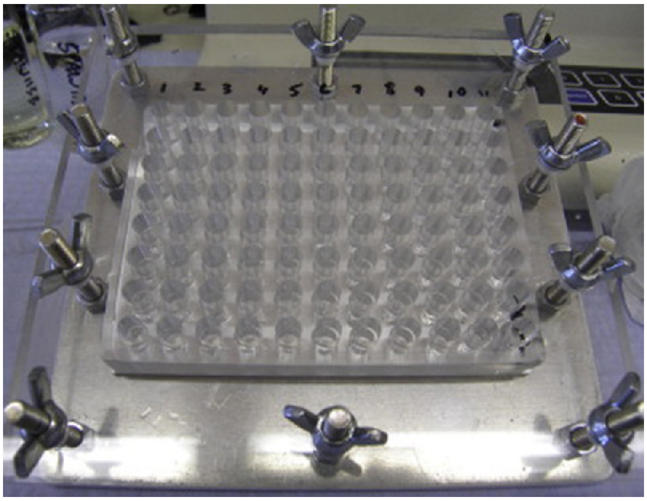


**Fig. 1.** High throughput corrosion inhibition rig consisting of a 10 mm thick polycarbonate clamped to a 10 mm thick block of polydimethylsiloxane rubber, an abraded plate of alloy and a 5 mm thick metal baseplate. The polycarbonate and polydimethylsiloxane sheets have an 8 × 11 grid of 6 mm diameter holes for test solutions.

erties [10]. Small organic molecules are particularly amenable to a high throughput synthesis approach as the technology to generate large libraries of these materials was developed for the pharmaceutical industry more than two decades ago. Here we combine novel high throughput corrosion inhibition testing experiments with machine learning methods to model and predict the corrosion inhibition performance of a large library of small organic candidate corrosion inhibitors at two initial pH values.

## 2. Materials and methods

### 2.1. Corrosion data

Two aluminium alloys, AA2024 and AA7075 were used for the inhibition study. The compositions (weight%) were determined by ICP as AA2024 (Cu 5.3, Mg 1.6, Mn 0.6, Fe 0.2, Zn < 0.1) and AA7075 (Cu 1.4, Mg 2.4, Mn < 0.1, Fe 0.2, Zn 5.4). While iron and magnesium levels were similar, AA2024 contained significantly more copper than AA7075, and AA7075 had substantially higher zinc content than the AA2024 alloy.

Corrosion inhibition data were generated by a high-throughput testing rig previously described [11] (see also Fig. 1). The two alloys were exposed to 100 different organic candidate corrosion inhibitors at a concentration of$10^{-3}$ M in 0.1 M NaCl solutions with initial pH values of 3, 4, 5, 6, 8, 10. As the components of most buffers are also potential corrosion inhibitors they were not used to control the pH of the experiments after initial adjustment of the pH. The inhibition of corrosion by the test compounds was assessed by image processing after a 24-h period. The degree of corrosion could be accurately estimated from the brightness of the images of the circular regions on the alloy exposed to the solution in each well. The image based corrosion scores were highly correlated with results of mass loss corrosion experiments. Potassium dichromate was used as a positive control, and 0.1 M saline (added to all wells) was the negative control. The data at initial pH 4 of and 10 were the most complete and were used in the modelling study. Note that the scoring scheme used in the previous publication (0 = no corrosion, and 100 = most corrosion)[11] were rescaled and reversed for the current inhibition study. Consequently, the scores used in this study ranged from zero (no inhibition) to 10 (maximum inhibition). The estimated measurement errors were ±1 in the 10 point scale. The scores are unitless.

### 2.2. Speciation

The data set consists of 100 small organic molecules with substantial chemical diversity (Supplementary Table 1). In some cases the identity of the organic species in solution (the various ionized forms of the molecules containing acidic or basic heteroatoms) is unambiguous at the pH of the experiment. However, for the heteroaromatic compounds, and indeed, for some of the inhibitors that contain, for example, carboxylate and thiol acidic moieties, the number of ionic species in solution can be quite high. As these ionic forms may have different binding affinities for metals and they will be chemically quite distinct, it might be relevant to understand which ionic species coexist at the experimental pH. However, the pH in the corrosion pits on the alloys is likely to be quite different to that in the bulk. It was also clear from inspection of the inhibition data that the initial pH had a relatively small effect on inhibition in the majority of cases over a 24-h exposure (Supplementary Table 2). Consequently, the compounds were modelled in the neutral form.

Supplementry material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.corsci.2016.02.008.

### 2.3. Quantum mechanical and modelling studies

The 100 molecules in the data set were built and their geometries optimized using the Sybyl x2.0 molecular modelling package (Certera Limited). These structures were used to generate a range of molecular properties (descriptors) such as the molecular surface area, molecular volume, molar refractivity (molecular size and polarizability), polar surface area, numbers of hydrogen bond donors and acceptors, logP (log of the octanol-water partition coefficient), HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital) energies and band gap (calculated using the semiempirical molecular orbital package MOPAC/AM1 and density functional theory (DFT) (see below)). A large pool of computed molecular descriptors was calculated using the DRAGON program [12] and our in-house modelling package, BioModeller [13–15]. Where possible, to aid interpretation of models, we used functional group descriptors that describe to the existence of certain chemical moieties or fragments in molecules (e.g. number of ionized carboxylic acid groups, number of heterocyclic nitrogen atoms, number of thiol groups etc.). Such descriptors make interpretation of models easier for organic chemists as they serve as design rules for new compounds to have good corrosion inhibitory properties. Descriptors such as total molecular charge, and several of the DRAGON descriptors that describe either the existence or frequency of molecular fragments containing atoms a specified number of bonds apart (topological distance) can also be interpreted relatively easily.

We also used quantum chemical properties for inhibitors calculated by DFT using the Spanish Initiative for Electronic Simulations with Thousands of Atoms (SIESTA) [16]. The exchange correlation functional of Perdew–Burke–Ernzerhof (PBE) [17] within the generalized gradient approximation (GGA), and a double zeta plus polarization (DZP) basis set was employed throughout, as per our earlier study [18]. The following molecular properties were calculated *in vacuo*: electron affinity, ionization potential, Mulliken electronegativity, chemical potential, chemical hardness, and the fundamental gap/HOMO-LUMO gap.

We generated machine-learning models that related molecular properties of inhibitors to corrosion inhibition using the BioModeller software. This software uses novel sparse Bayesian feature selection methods to identify relevant molecular properties from large pools of molecular descriptors, and sparse modelling methods to generate optimal structure-inhibition models, both linear and nonlinear. Linear models were simply multiple linear regressions (MLR, weighted combinations of molecular descriptors) to which sparse feature selection has been applied to eliminate molecular descriptors of lower importance. Nonlinear models were generated using a feed forward Bayesian regularized neural networks that used a Gaussian prior (BRANNGP) or sparsity-inducing Laplacian prior (BRANNLP) [13,19–22]. Unlike standard neural networks and related machine learning methods, Bayesian regularization automatically optimizes the complexity of nonlinear models. As sparser models are generally more predictive, this ensures that the correct balance between bias (model too simple to capture the structure-property relationship completely) and variance (model is over fitted) is attained. This also means that a single hidden layer containing a small number of nodes (usually 2–3) is adequate for most nonlinear models. The stopping criterion for neural network training was a maximum in the Bayesian evidence for the models. The hidden layer nodes used nonlinear sigmoidal transfer functions and the input and output layer nodes used linear transfer functions. As is standard practice in structure-property modelling, we divided the data set of inhibitors into a training set (80%), used to generate the models, and a test set (20%), used to assess the ability of the models to predict new data. However, Bayesian regularized neural networks do not strictly need a test set. The predictive power of the corrosion inhibition models was assessed using the standard error of the prediction for the training set (SEE) and test set (SEP), and the less statistically robust squared correlation coefficient ($r^2$) values for the models [23]. The SEE and SEP values described the degree of uncertainty in the predictions of the corrosion inhibition scores by the molecules in the training and test sets respectively.

## 3. Results and discussion

### 3.1. *Relationship between the two aluminium alloys*

The inhibition results for all inhibitors over 24 h exposure correlate only weakly for the two alloys ($r^2 = 0.35$, initial pH4; $r^2 = 0.27$ initial pH10). The corrosion inhibition exhibited by the compounds was significantly lower for the AA7075 alloy than for the AA2024 alloy. The inhibition results identified sulfur-containing ligands as being overrepresented in the effective corrosion inhibitors for the AA2024 alloy than for the AA7075 alloy (e.g. 2-mercaptopyrimidine, 3-mercaptobenzoate, 2-mercaptonicotinate, 2,3-dimercaptosuccinate, mercapto-acetate, mercaptopropionate, 2,5-dimercapto-1,3,4-thiadiazole). This is likely to be due to the 4x higher copper levels in the AA2024 alloy. Additional evidence for this hypothesis is provided by experiments in which sheets of pure copper and pure aluminium were exposed to the small molecule inhibitors. There appeared to be little interaction with the aluminium but substantial reaction with pure copper [24].

### 3.2. *Relationship between corrosion inhibition and DFT properties*

Breedon at al. [18]. reported recently that, for a small set of organic corrosion inhibitors, molecular properties calculated by quantum chemical methods such as density functional theory (DFT) did not correlate strongly with corrosion inhibition. We investigated this finding for the larger and more chemically diverse data set, and again found there were essentially no correlation between the investigated DFT calculated molecular properties and the experimentally measured corrosion inhibition efficiency. This is in contrast to a significant number of literature reports [25–35] that claim frontier orbital energies are relevant to the corrosion inhibition. However, many of these studies employed very small numbers of inhibitors, as low as four in some cases, making the probability of chance correlations high if a significant number of potential model descriptors were tried in a model. We used six molecular descriptors for the inhibitor molecules calculated *in vacuo* by high-level DFT calculations to confirm that quantum chemical descriptors were not useful in modelling corrosion inhibition for this set of inhibitors. Only two of the parameters showed small correlations with inhibition and these proved inadequate to generate useful structure-inhibition models (see Supplementary Table 3).

**Table 1**
Summary of corrosion inhibition models for AA7075 at initial pH 4. Explanation of the relevant descriptors is provided in Supplementary Table 4.

| Model | $N_{descr}$ | $r^2_{train}$ | SEE | $r^2_{test}$ | SEP | $N_{eff}$ |
|---|---|---|---|---|---|---|
| MLREM | 17 | 0.78 | 1.2 | | | 18 |
| BRANNGP 2 nodes | 17 | 0.73 | 1.0 | | | 17 |
| BRANNGP 3 nodes | 17 | 0.72 | 1.0 | | | 18 |
| BRANNLP | 17 | 0.72 | 1.0 | | | 19 |
| MLREM 20% test | 17 | 0.77 | 1.2 | 0.79 | 1.1 | 18 |
| BRANNGP 2 nodes 20% test | 17 | 0.71 | 1.0 | 0.82 | 0.9 | 18 |
| BRANNGP 3 nodes 20% test | 17 | 0.71 | 1.0 | 0.82 | 0.9 | 17 |
| BRANNLP 20% test | 17 | 0.71 | 1.1 | 0.79 | 1.0 | 19 |

Descriptors: B05[N-S], HATS4p, C-030, F05[O-S], F02[S-S], B03[C-S], RDF050m, F03[N-S], GATS8m, GATS3v, F02[N-S], MATS5v, Infective-80, RDF100m, C-001, S-106, B02[N-N].

Supplementry material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.corsci.2016.02.008.

### 3.3. *Machine learning-based quantitative structure-inhibition modelling*

We have recently shown that a combination of calculated molecular descriptors, sparse feature selection methods, and nonlinear machine learning algorithms can robustly and quantitatively model the corrosion inhibition of small organic molecules [9]. Such models were able to make quantitative predictions of the inhibitory activity of new candidate organic corrosion inhibitors, and provided insight into the molecular features that contributed most to the inhibition of aerospace alloy corrosion.

Given this prior validation, we employed similar methods for the current study. DRAGON and in-house descriptor algorithms generated almost 2000 molecular descriptors for each inhibitor. After removing those that did not vary significantly with inhibition, 1515 descriptors remained. Clearly this large number needed to be processed to exclude all but 10–20 of the most relevant descriptors, or the models could be over fitted and be incapable of predicting the inhibition of new molecules. We used the sparse feature selection capabilities of BioModeller to select the most relevant subset of descriptors from this large pool in a context dependent manner. The description of the most relevant molecular descriptors used in this study is provided in Supplementary Table 4.

See Excel sheet 1 as supplementary file. Supplementry material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.corsci.2016.02.008.

The modelling methods generated models of good statistically significance that could make quantitative predictions of the corrosion inhibition properties of compounds in an external test set. In most cases between 8 and 22 descriptors were sufficient to generate linear and nonlinear models that predicted the degree of inhibition for molecules in the training and test data sets. The models used descriptors derived from the neutral form of the inhibitors as explained above. Tables 1–4 summarize the performance of the corrosion inhibition models.

#### 3.3.1. *Corrosion inhibition models for AA7075*

The results of modelling of corrosion inhibition in the AA7075 aerospace alloy are summarized in Tables 1 and 2. The $r^2$ values provide an indication of the fraction of the variance in the training and test set data that is explained by the model. The parameter $N_{descr}$ is the number of molecular descriptors (excluding the MLR intercept) remaining in the models, and $N_{eff}$ is the number of effec-

**Table 2**
Summary of corrosion inhibition models for AA7075 at initial pH 10. Explanation of the relevant descriptors is provided in Supplementary Table 4.

| Model | $N_{descr}$ | $r^2_{train}$ | SEE | $r^2_{test}$ | SEP | $N_{eff}$ |
|---|---|---|---|---|---|---|
| MLREM | 11 | 0.66 | 1.4 | | | 12 |
| BRANNGP 2 nodes | 11 | 0.67 | 1.2 | | | 15 |
| BRANNGP 3 nodes | 11 | 0.72 | 1.1 | | | 17 |
| BRANNLP | 11 | 0.63 | 1.2 | | | 15 |
| MLREM 20% test | 11 | 0.67 | 1.3 | 0.56 | 1.7 | 12 |
| BRANNGP 2 nodes 20% test | 11 | 0.63 | 1.2 | 0.60 | 1.6 | 12 |
| BRANNGP 3 nodes 20% test | 11 | 0.65 | 1.2 | 0.60 | 1.6 | 12 |
| BRANNLP 20% test | 11 | 0.64 | 1.2 | 0.58 | 1.6 | 12 |

Descriptors: B02[N-S], nCp, F06[C-N], B03[N-S], MATS7m, S-107, T(N.N), Mor28m, E1s, C-025, F02[C-S].

**Table 3**
Summary of corrosion inhibition models for AA2024 at initial pH 4. Explanation of the relevant descriptors is provided in Supplementary Table 4.

| Model | $N_{descr}$ | $r^2_{train}$ | SEE | $r^2_{test}$ | SEP | $N_{eff}$ |
|---|---|---|---|---|---|---|
| MLREM | 8 | 0.52 | 1.5 | | | 9 |
| BRANNGP 2 nodes | 11[a] | 0.52 | 1.4 | | | 9 |
| BRANNGP 3 nodes | 11 | 0.51 | 1.4 | | | 9 |
| BRANNLP | 11 | 0.49 | 1.4 | | | 9 |
| MLREM 20% test | 11 | 0.48 | 1.5 | 0.62 | 1.5 | 9 |
| BRANNGP 2 nodes 20% test | 11 | 0.4 | 1.4 | 0.63 | 1.4 | 9 |
| BRANNGP 3 nodes 20% test | 11 | 0.4 | 1.4 | 0.63 | 1.4 | 9 |
| BRANNLP 20% test | 11 | 0.32 | 1.44 | 0.65 | 1.5 | 9 |
| MLREM | 29["b] | 0.84 | 1.0 | | | 30 |
| MLREM 20% test | 29 | 0.85 | 1.0 | 0.75 | 1.2 | 30 |
| BRANNGP 3 nodes | 29 | 0.83 | 0.8 | | | |
| BRANNGP 3 nodes 20% test | 29 | 0.75 | 0.8 | 0.82 | 1.0 | 28 |

**Table 4**
Summary of corrosion inhibition models for AA2024 at pH initial 10. Explanation of the relevant descriptors is provided in Supplementary Table 4.

| Model | $N_{descr}$ | $r^2_{train}$ | SEE | $r^2_{test}$ | SEP | $N_{eff}$ |
|---|---|---|---|---|---|---|
| MLREM | 14[a] | 0.60 | 1.5 | | | 15 |
| BRANNGP 2 nodes | 14 | 0.63 | 1.3 | | | 17 |
| BRANNGP 3 nodes | 14 | 0.67 | 1.3 | | | 18 |
| BRANNLP | 14 | 0.57 | 1.4 | | | 18 |
| MLREM 25% test | 14 | 0.62 | 1.5 | 0.46 | 1.6 | 15 |
| BRANNGP 2 nodes 25% test | 14 | 0.53 | 1.3 | 0.49 | 1.5 | 15 |
| BRANNGP 3 nodes 25% test | 14 | 0.52 | 1.3 | 0.48 | 1.5 | 14 |
| MLREM | 31[b] | 0.87 | 1.0 | | | 32 |
| MLREM 20% test | 31 | 0.88 | 1.0 | 0.76 | 1.2 | 32 |
| BRANNGP 3 nodes | 31 | 0.87 | 0.8 | | | 34 |
| BRANNGP 3 nodes 20% test | 31 | 0.81 | 0.7 | 0.74 | 1.1 | 32 |

[a] Descriptors: C-027, G1u, C-029, F02[N-S], MATS5p, F05[O-S], IDDE, BELv2, Infective-80, H-052, nTriazoles, R2u+, G2v, B02[C-S].

[b] Descriptors: C-029, C-030, F03[N-N], H-049, H-052, Mor08e, O-058, F04[N-N], Ds, S-107, MATS5p, F05[O-S], Infective-80, B02[C-N], B02[C-S], F02[N-S], G1u, B02[O-S], B03[C-N], IDDE, BELv2, R5e+, B04[N-S], G2v, B05[N-S], B05[N-Cl], EEig10d, E1p, nTriazoles, C-005, JGI4.

tive weights used in the neural network models after the sparse Bayesian regularization algorithm prunes out the less relevant network weights. We found that nonlinear models provided a modest but significant improvement in the quality of corrosion inhibition structure-property models compared to linear models (lower standard errors of prediction). We found that the models for inhibition of the 7075 alloy required a smaller number of descriptors than models for the corrosion inhibition of 2024 alloy.

*3.3.1.1. Inhibition models at initial pH 4.* The performance of the corrosion inhibition models is summarized in Table 1. Both linear and nonlinear modelling methods generated statistically significant models relating the inhibitor structures to the corrosion inhibition. The linear MLREM models could predict the training and test sets with standard errors of prediction of 1.2 and 1.1 corrosion scores respectively. The neural network models had standard errors of prediction of 1.0 and 0.9 corrosion scores, a 15% improvement in prediction error. These models were relatively sparse, employing 17 molecular descriptors and a similar number of effective weights (fitted parameters) in most models. The quality of the models is illustrated in Fig. 2.

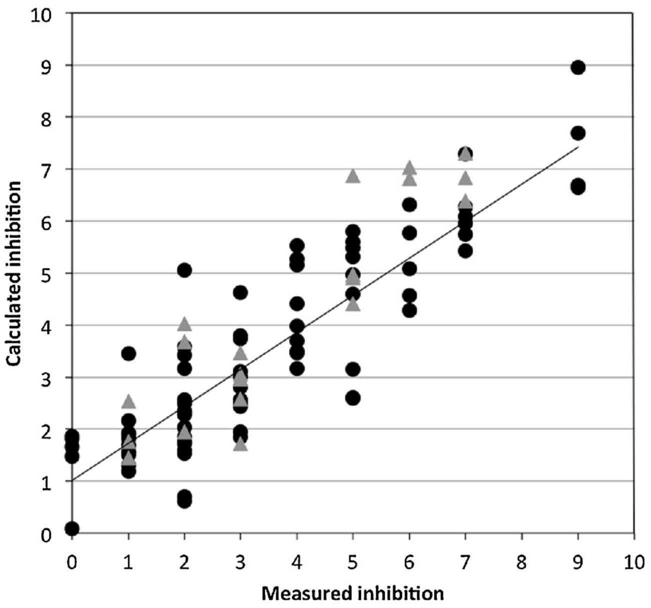


**Fig. 2.** The experimentally measured versus predicted corrosion inhibition scores (0–10) for a Bayesian neural network inhibitor model (2 hidden layer nodes) for the 7075 alloy at initial pH 4. Training set data are shown by circles and test set data by triangles.

*3.3.1.2. Inhibition models at initial pH 10.* Models for the inhibition at pH10 (Table 2) were sparser, employing only 11 descriptors, but had higher standard errors of prediction. The nonlinear models could account for 60–65% of the variance in the data. As with the corrosion inhibition models at pH4, the nonlinear neural network models could predict the performance of inhibitors in the test set with a lower standard error, 1.7 versus 1.6 inhibition scores. The descriptors of these sparser models largely encoded similar properties relating to functional groups containing N and S, C and S, two nitrogen atoms, and thiol functionality.

The ability of the models to predict the degree of inhibition of the external test set compounds is satisfactory, as Fig. 3 illustrates.

*3.3.2. Corrosion inhibition models for AA2024*

The structure-inhibition models for the AA2024 alloy required a larger number of descriptors than the inhibition models for the AA7075 alloy. A larger number of molecules were good inhibitors of corrosion in the AA2024 allow than in the AA7075 alloy. As with the AA7075 alloy, nonlinear corrosion inhibition structure-property models had greater predictivity than linear models. The performance of structure–corrosion inhibition models for the AA2024 alloy is summarized in Tables 3 and 4.

When the models were sparse, using only 9 effective weights in the models, they had only modest predictive power. Less sparse models (more descriptors), summarized in the second half of Table 3 and 4, were of substantially higher statistical power and could predict the properties of the independent test set well. This suggested that the models were not over fitted in spite of the relatively large number of descriptors used. The quality of the prediction of the training and test set corrosion inhibition is illustrated in Figs. 4 and 5.

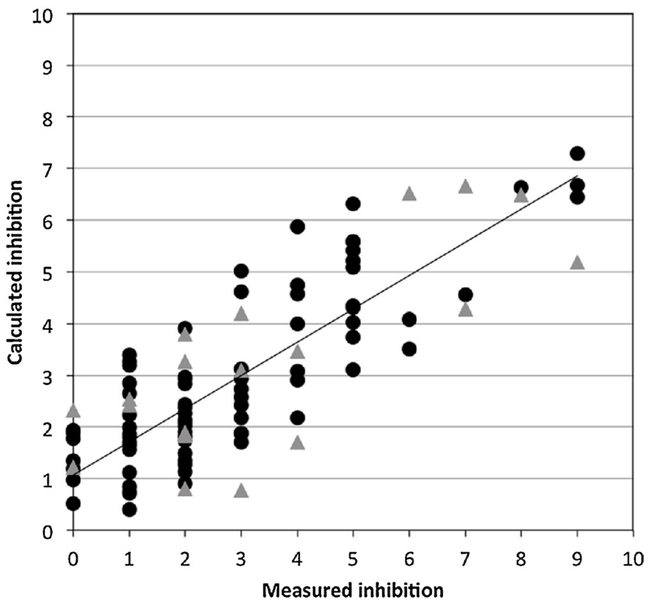


**Fig. 3.** The experimentally measured versus predicted corrosion inhibition scores (0–10) for Bayesian neural network inhibitor model (2 hidden layer nodes) for the 7075 alloy at initial pH 10. Training set data are shown by circles and test set data by triangles.

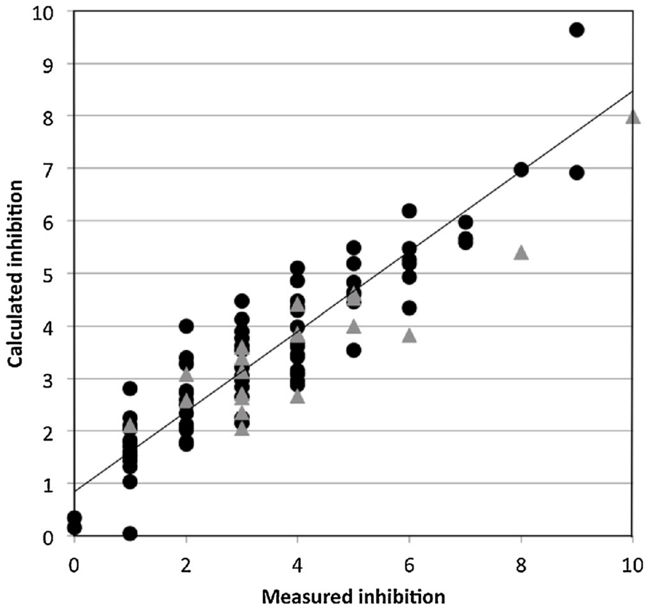


**Fig. 4.** The experimentally measured versus predicted corrosion inhibition scores (0–10) for Bayesian neural network inhibitor model (3 hidden layer nodes) for the 2024 alloy at initial pH 4. Training set data are shown by circles and test set data by triangles.

*3.3.2.1. Inhibition models at initial pH 4.* The sparsest models for corrosion inhibition at pH4 contained only 9 parameters (Table 3), and could predict the training and test sets with a standard error of 1.4 corrosion scores. The less sparse models were substantially better at predicting training and test sets, with standard errors of prediction of 0.8 and 1.08 corrosion scores for training and test sets respectively, consistent with the estimated level of experimental error. The ability of the models to predict the training and test sets with similar SEP values (Fig. 4) shows that the models are not over fitted and are robust.

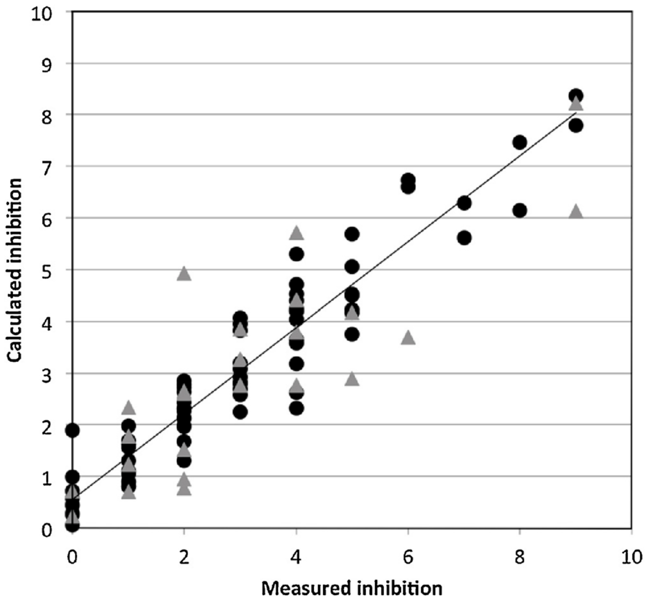


**Fig. 5.** The experimentally measured versus predicted corrosion inhibition scores (0–10) for Bayesian neural network inhibitor model (3 hidden layer nodes) for the 2024 alloy at initial pH 10. Training set data are shown by circles and test set data by triangles.

*3.3.2.2. Inhibition models at initial pH 10.* As with the models for inhibition at initial pH 4, the sparsest initial pH 10 models were not as effective at predicting the properties of the training set as the less sparse models (Table 4). The latter models predicted the training and test sets with similar efficacy, suggesting they were robust. The best models could explain 75–80% of the variance in the data, and could predict the training and test sets with standard errors of 0.7 and 1.1 corrosion score respectively (Fig. 5).

### 3.4. *Interpretation of the models*

The models required the use of a relatively large number of descriptors, suggesting the relationship between the molecular properties of the small organic compounds and corrosion inhibition was quite complex. Although molecular descriptors generated by DRAGON and BioModeller are extremely useful, and allow good predictive models of materials properties to be generated, interpretation of these models is often quite difficult [36]. In the current corrosion inhibitor study, the number of descriptors necessary to generate useful models, and the relatively opaque or arcane nature of many of them makes mechanistic interpretation of the models problematic. Attempts were made to generate models with smaller numbers of descriptors, or with descriptors that were more chemically interpretable, but such models were of substantially lower predictive power. This is the most comprehensive inhibitor modelling study, in terms the number and molecular diversity of inhibitors, range of alloys and pH conditions, and quality of the model predictions that has been reported to date. The lack of mechanistic interpretability must be balanced against the usefulness of the models in selecting new inhibitors from large libraries of possibilities by virtual screening. However, some interpretation of the models is possible.

It is clear by inspection that in many cases the presence of a sulfur atom, particularly when it is an ionizable thiol moiety near a heteroatom in a ring, generates compounds with very good corrosion inhibition properties. A substantial number of descriptors relate to properties of sulfur and nitrogen atoms accordingly. In particular, many of the selected descriptors relate to molecular fragments that are important for inhibition. Most easily interpreted of these are the Bn[X-Y] and Fn[X-Y] descriptors that describe the topological relationships between atoms X and Y (usually heteroatoms, and most frequently S, N, and less often O) and topological distance n. The topological distance is the smallest number of bonds between atoms X and Y. For example, the descriptor B02[C-S] or F02[C-S] is important in most models. It represents the presence/absence or frequency of occurrence of a carbon and sulfur atom that are two bonds apart. Also commonly represented are the F03[N-S] and F04[N-S] topological descriptors. These represent the frequency of nitrogen and sulfur atoms that are 3 and 4 bonds apart. The number of triazoles and topological descriptors related to nitrogen heterocycles such as Bn[N-N] and Fn[N-N] are also common descriptors in the models (where n represents a specific number of bonds separating the descriptors). In some cases the presence of extended conjugation ($NC_{conj}$) is important, as the following discussion describes.

In our previous study we surmised that aromatic inhibitors might form ordered layers on the surface of the metal. Literature reports also suggest that molecules resembling thiophenolates may form polymeric complexes on the surface [37,38]. Clearly, the interaction of small organic molecules with metal surfaces is quite complex and largely unknown and there is scope for other inhibition mechanisms to also play a role. Consequently, given the relatively small size of the data set and limited chemical diversity, and the obvious complexity of inhibition and corrosion processes happening at the surface, one must be cautious not to over interpret the models.

Given the ability of the models to predict the corrosion inhibition efficiency of organic compounds in the test set, these models are capable of predicting the likely corrosion inhibition of additional new small molecules not yet tested, or in some cases not yet synthesized. The models would allow libraries or databases of real or virtual molecules to be computationally screened to identify putative corrosion inhibitors. However, given the relatively small size of the data set, care must be taken to ensure that compounds in a virtual screen had molecular descriptors within, or close to, the domain of applicability of the models where corrosion inhibition predictions have the highest reliability.

### 3.5. Comparison with previous QSPR models of corrosion inhibition

This study differs from our earlier work [9] on quantitative modelling of corrosion inhibition by small organic molecules. The focus of that study was determining whether robust machine learning methods could generate quantitative, predictive structure-inhibition models, and whether the quantum chemical descriptors reported to be important in earlier papers were indeed useful. Our previous study used much smaller data set of 28 compounds only 18 of which showed corrosion inhibitory effects. The pH in this earlier study was not controlled. The study we report here contains 100 small molecule inhibitors whose efficacy was measured using fast, high throughput experimental methods at two pH values. The current study was able to assess corrosion inhibition after only 24 h using optical methods, allowing substantially more compounds to be assessed and modelled than for the 28 day mass loss corrosion inhibition methods employed in our previous study. The larger number and chemical diversity of the compounds in this study means that the models have a larger range of applicability than those generated in our earlier and more limited study. Although our prior study strongly suggested that molecular parameters calculated by DFT or other quantum chemically derived methods were not particularly useful for modelling structure-inhibition relationships, our much larger study has shown that there is effectively no correlation between these parameters and inhibition. This strongly suggests that other published structure-inhibition models in the literature are incorrect, presumably because their models were based on very small numbers of inhibitors. The promising results of our earlier study, which suggested that quantitative prediction of corrosion inhibition properties of small organic molecules was feasible, have been put on a much firmer footing by the comprehensive study we report here.

## 4. Conclusions

We have shown how a combination of high throughput corrosion screening technology and modern machine learning methods can generate predictive models of the molecular structure-corrosion inhibition relationships in aerospace aluminium alloys AA2024 and AA7075, for a large set of molecules at two pH values. This is the one of the first studies where corrosion inhibition data from high throughput experiments with good chemical diversity has been modelled in a manner that provides useful predictions of the properties of small organic molecules as chromate replacements. As with our earlier studies, we have shown that, for this larger more diverse set of organic inhibitors, the descriptors generated by *in vacuo* DFT calculations for the inhibitors alone do not contain sufficient information to be useful in generating models. Future work examining organic inhibitor-surface interactions may yield more pertinent DFT level descriptors, and warrants further examination. The results from this study illustrate the considerable potential for computational modelling and high throughput experimentation to work together to accelerate development of novel materials for corrosion control.

## Acknowledgements

The support of the Boeing Company and the CSIRO Advanced Materials Transformational Capability is gratefully acknowledged. Aman Jain is acknowledged for running selected DFT calculations.

## References

[1] M.W. Kendig, R.G. Buchheit, Corrosion inhibition of aluminum and aluminum alloys by soluble chromates, chromate coatings, and chromate-free coatings, Corrosion 59 (2003) 379–400.

[2] R.M. Park, J.F. Bena, L.T. Stayner, R.J. Smith, H.J. Gibb, P.S.J. Lees, Hexavalent chromium and lung cancer in the chromate industry: a quantitative risk assessment, Risk Anal. 24 (2004) 1099–1108.

[3] C.B.K. Max Costa, Toxicity and carcinogenicity of chromium compounds in humans, Crit. Rev. Toxicol. 36 (2006) 155–163.

[4] A.E. Hughes, I.S. Cole, T.H. Muster, R.J. Varley, Designing green, self-healing coatings for metal protection, NPG Asia Mater. 2 (2010) 143–151.

[5] M. Finsgar, I. Milosev, Inhibition of copper corrosion by 1,2,3-benzotriazole: a review, Corros. Sci. 52 (2010) 2737–2749.

[6] G. Gece, Drugs: a review of promising novel corrosion inhibitors, Corros. Sci. 53 (2011) 3873–3898.

[7] Y.I. Kuznetsov, Physico-chemical aspects of protection of metals by organic corrosion inhibitors, Protect. Met. Phys. Chem. 51 (2015) 1111–1121.

[8] D. Winkler, Adrien Albert award: how to mine chemistry space for new drugs and biomedical therapies, Aust. J. Chem. 68 (2015) 1174–1182.

[9] D.A. Winkler, M. Breedon, A.E. Hughes, F.R. Burden, A.S. Barnard, T.G. Harvey, I. Cole, Towards chromate-free corrosion inhibitors: structure-property models for organic alternatives, Green Chem. 16 (2014) 3349–3357.

[10] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Quantitative Structure-Property Relationship modeling of diverse materials properties, Chem. Rev. 112 (2012) 2889–2919.

[11] P.A. White, G.B. Smith, T.G. Harvey, P.A. Corrigan, M.A. Glenn, D. Lau, S.G. Hardin, J. Mardel, T.A. Markley, T.H. Muster, N. Sherman, S.J. Garcia, J.M.C. Mol, A.E. Hughes, A new high-throughput method for corrosion testing, Corros. Sci. 58 (2012) 327–331.

[12] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH Weinheim, 2000.

[13] F.R. Burden, M.J. Polley, D.A. Winkler, Toward novel universal descriptors: charge fingerprints, J. Chem. Inf. Model. 49 (2009) 710–715.

[14] D.A. Winkler, F.R. Burden, Robust QSAR models from novel descriptors and Bayesian regularised neural networks, Mol. Simul. 24 (2000) 243.

[15] D.A. Winkler, F.R. Burden, A.J.R. Watkins, Atomistic topological indices applied to benzodiazepines using various regression methods, Quant. Struct. Act. Rel. 17 (1998) 14–19.
[16] J.M. Soler, E. Artacho, J.D. Gale, A. García, J. Junquera, P. Ordejón, D. Sánchez-Portal, The SIESTA method for ab initio order-N materials simulation, J. Phys. Cond. Matt. 14 (2002) 2745–2779.
[17] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868.
[18] M. Breedon, M.C. Per, I.S. Cole, A.S. Barnard, Molecular ionization and deprotonation energies as indicators of functional coating performance, J. Mater. Chem. A 2 (2014) 16660–16668.
[19] F.R. Burden, D.A. Winkler, Robust QSAR models using Bayesian regularized neural networks, J. Med. Chem. 42 (1999) 3183–3187.
[20] F.R. Burden, D.A. Winkler, New QSAR methods applied to structure-activity mapping and combinatorial chemistry, J. Chem. Inf. Comp. Sci. 39 (1999) 236–242.
[21] F.R. Burden, D.A. Winkler, An optimal self-pruning neural network and nonlinear descriptor selection in QSAR, QSAR Comb. Sci. 28 (2009) 1092–1097.
[22] F.R. Burden, D.A. Winkler, Optimal sparse descriptor selection for QSAR using Bayesian methods, QSAR Comb. Sci. 28 (2009) 645–653.
[23] D.L. Alexander, A. Tropsha, D.A. Winkler, Beware of $R^2$: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models, J. Chem. Inf. Model. 55 (2015) 1529–1533.
[24] T.G. Harvey unpublished data CSIRO (2011).
[25] F. Bentiss, M. Lebrini, M. Lagrenee, M. Traisnel, A. Elfarouk, H. Vezin, The influence of some new 2,5-disubstituted 1,3,4-thiadiazoles on the corrosion behaviour of mild steel in 1 M HCl solution: AC impedance study and theoretical approach, Electrochim. Acta 52 (2007) 6865–6872.
[26] E.E. Ebenso, T. Arslan, F. Kandemirli, I. Love, C. Odretir, M. Saracoglu, S.A. Umoren, Theoretical studies of some sulphonamides as corrosion inhibitors for mild steel in acidic medium, Int. J. Quant. Chem. 110 (2010) 2614–2636.
[27] E.E. Ebenso, D.A. Isabirye, N.O. Eddy, Adsorption and quantum chemical studies on the inhibition potentials of some thiosemicarbazides for the corrosion of mild steel in acidic medium, Int. J. Mol. Sci. 11 (2010) 2473–2498.
[28] E.E. Ebenso, M.M. Kabanda, L.C. Murulana, A.K. Singh, S.K. Shukla, Electrochemical and quantum chemical investigation of some azine and thiazine dyes as potential corrosion inhibitors for mild steel in hydrochloric acid solution, Ind. Eng. Chem. Res. 51 (2012) 12940–12958.
[29] N.O. Eddy, E.E. Ebenso, U.J. Ibok, Adsorption, synergistic inhibitive effect and quantum chemical studies of ampicillin (AMP) and halides for the corrosion of mild steel in $H_2SO_4$, J. Appl. Electrochem. 40 (2010) 445–456.
[30] E.H. El Ashry, A. El Nemr, S.A. Essawy, S. Ragab, Corrosion inhibitors part V: QSAR of benzimidazole and 2-substituted derivatives as corrosion inhibitors by using the quantum chemical parameters, Prog. Org. Coat. 61 (2008) 11–20.
[31] K.F. Khaled, Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: a QSAR model, Corros. Sci. 53 (2011) 3457–3465.
[32] K.F. Khaled, N.S. Abdel-Shafi, Quantitative structure and activity relationship modeling study of corrosion inhibitors: genetic function approximation and molecular dynamics simulation methods, Int. J. Electrochem. Soc. 6 (2011) 4077–4094.
[33] K.F. Khaled, K. Babic-Samardzija, N. Hackerman, Theoretical study of the structural effects of polymethylene amines on corrosion inhibition of iron in acid solutions, Electrochim. Acta 50 (2005) 2515–2520.
[34] A.Y. Musa, A.B. Mohamad, A.A.H. Kadhum, M.S. Takriff, W. Ahmoda, Quantum chemical studies on corrosion inhibition for series of thio compounds on mild steel in hydrochloric acid, J. Ind. Eng. Chem. 18 (2012) 551–555.
[35] M. Outirite, M. Lagrenee, M. Lebrini, M. Traisnel, C. Jama, H. Vezin, F. Bentiss, AC impedance, X-ray photoelectron spectroscopy and density functional theory studies of 3,5-bis(n-pyridyl)-1,2,4-oxadiazoles as efficient corrosion inhibitors for carbon steel surface in hydrochloric acid solution, Electrochim. Acta 55 (2010) 1670–1681.
[36] T. Fujita, D.A. Winkler, Reconciling the "two QSARs", J. Chem. Inf. Mod. ASAP 56 (2016), http://dx.doi.org/10.1021/acs.jcim.5b00229.
[37] X.R. Ye, X.Q. Xin, Coordination chemical-reaction of MoS42- on the surface of copper, Acta Chim. Sin. 53 (1995) 462–467.
[38] X.R. Ye, X.Q. Xin, J.J. Zhu, Z.L. Xue, Coordination compound films of 1-phenyl-5-mercaptotetrazole on copper surface, Appl. Surf. Sci. 135 (1998) 307–317.