ARTICLE OPEN

Check for updates

Laying the experimental foundation for corrosion inhibitor discovery through machine learning

Can Özkan[1]✉, Lisa Sahlmann[2], Christian Feiler[2], Mikhail Zheludkevich[2], Sviatlana Lamaka[2], Parth Sewlikar[3], Agnieszka Kooijman[1], Peyman Taheri[1] and Arjan Mol[1]

Creating durable, eco-friendly coatings for long-term corrosion protection requires innovative strategies to streamline design and development processes, conserve resources, and decrease maintenance costs. In this pursuit, machine learning emerges as a promising catalyst, despite the challenges presented by the scarcity of high-quality datasets in the field of corrosion inhibition research. To address this obstacle, we have created an extensive electrochemical library of around 80 inhibitor candidates. The electrochemical behaviour of inhibitor-exposed AA2024-T3 substrates was captured using linear polarisation resistance, electrochemical impedance spectroscopy, and potentiodynamic polarisation techniques at different exposure times to obtain the most comprehensive electrochemical picture of the corrosion inhibition over a 24-h period. The experimental results yield target parameters and additional input features that can be combined with computational descriptors to develop quantitative structure–property relationship (QSPR) models augmented by mechanistic input features.

npj Materials Degradation (2024)8:21 ; https://doi.org/10.1038/s41529-024-00435-z

INTRODUCTION

Corrosion inhibition research has come far since Chyżewski and Evans first categorised sparingly soluble corrosion-decreasing substances as anodic and cathodic inhibitors[1]. Thanks to the advances in computational power and methods, we are observing a paradigm shift in how science is done, and this is also affecting corrosion inhibition research.

There are four contemporary paradigms of science[2,3]. The first is empirical evidence, leading to general laws through 'trial and error'. The second involves theoretical models based on those laws. The third is defined by computational power offered by Moore's law, the application of theoretical models to more complex and specific problems. This results in a data explosion, leading to the fourth paradigm: data-driven scientific discovery—such as using machine learning for categorisation and prediction.

We see examples of this paradigm shift in corrosion inhibitor research in two broad categories: mechanistic and statistical research. Lately, advances in surface analysis, electrochemical characterisation and computational methods have been complementing each other to facilitate the inhibitor discovery process for both of these categories.

On the mechanistic end, a deeper scientific understanding is obtained by controlled experiments and computational models. The critical need for the protection of aerospace aluminium alloys has driven the research that would uncover AA2024-T3 corrosion inhibition of many compounds. Throughout the years, AA2024-T3 corrosion inhibition mechanisms were experimentally uncovered for inorganic compounds such as chromates[4–7], rare-earths[8–11], molybdate[12] and cobalt ions[13], magnesium-based pigments[14–16], lithium salts[17–19], and a vast variety of organic compounds such as imidazole[20,21], triazole/thiazole[22,23], quinoline[24,25], carbamate[26], thiosemicarbazone[27] derivatives, among others[24,28–32]. In addition to uncovering the mechanisms for specific inhibitor species, the physical features of inhibition mechanisms such as the importance of time[33,34] and irreversibility[35] have been investigated.

The pressing demand for novel chromate-free corrosion inhibitors has created the need for high-throughput inhibitor screening methodologies. The approaches inspired by pharmaceutical drug discovery research spanned optical image analysis[36,37], fluorometric detection[38], multi-electrode electrochemical evaluation[36,39–41], surface copper enrichment analysis[42], hydrogen evolution detection[43,44], weight-loss measurements[28,45], and spectroscopic element analysis through multi-channels[46]. These methods rapidly created large datasets but with the trade-off of losing mechanistic information.

The third paradigm supported the mechanistic understanding gained from experiments with computational models that span a continuum to atomistic scales. Finite element method (FEM) models produced previously unattainable information—such as mechanical strains observed for inhibitor dissolution and leaching from coatings[47], local critical pH criteria for pit repassivation[48], and the effect of surface geometry on electrochemical behaviour[49]. Density functional theory (DFT) and molecular dynamics (MD) simulations have introduced a vast amount of quantum mechanical/chemical information that is not directly available from empirical methods, such as density of states, band gap, and other physicochemical electronic properties[50]. The ease of investigation of atomistic properties offered by software/hardware advances has allowed corrosion scientists to replace costly and time-consuming experiments. Molecular modelling was used as a computational microscope to expose the underlying mechanisms of inhibitor structure–substrate sorption phenomena[50–55]. Recent papers[56–58] have reported on how experimental and computational methods are catalyzing one another to combine the strength of empirical and theoretical methods, in which researchers have analysed the influence of type and length of backbone

[1]Department of Materials Science and Engineering, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands. [2]Institute of Surface Science, Helmholtz-Zentrum Hereon, Max-Planck-Straße 1, 21502 Geesthacht, Germany. [3]Department of Materials and Chemistry, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. ✉email: C.Ozkan@tudelft.nl

chains and anchor groups on inhibitor performance by combining carefully controlled experiments with DFT modelling.

The accumulated mechanistic understanding of inhibitors, high-throughput methodologies and FEM/DFT-MD computational approaches generated previously unavailable large datasets about the mechanical and physicochemical behaviour of inhibitors, which paved the road for data-driven statistical investigations. This involved classification and predictive analytics of inhibitors. Properties of inhibitor molecules obtained from DFT calculations, and experimental inhibitor efficiencies gathered from high-throughput methods have been combined to build correlations using machine learning-based quantitative structure–property relationships (QSPR). Winkler et al.[59] used QSPR to reveal empirical molecular descriptors most relevant for AA2024 and AA7075 inhibition and identified that chemical descriptors solely using input features obtained from *in vacuo* DFT did not contain sufficient information to generate predictive models. Würger et al.[60,61] have demonstrated a data-driven inhibitor prediction workflow for magnesium alloys, which combined the results of atomistic simulations and high-throughput experiments with unsupervised machine learning clustering algorithms and supervised learning approaches to predict the behaviour of untested inhibitors. Feiler et al.[44] have demonstrated that the combination of structural information with input features derived from DFT leads to robust predictive models for corrosion inhibition responses of small organic molecules based on an artificial neural network for pure magnesium, as well as Mg-based alloys[62]. The optimisation of machine learning approaches is an ongoing process, whether it is coming up with better methods of identifying the most relevant molecular descriptors[62], or analysis of different inhibitor classification algorithms and creation of new descriptors with intrinsic mechanistic meanings[63].

All in all, in silico inhibitor screening combined with smart high-throughput testing has enabled overcoming the physical limitations of previous paradigms. However, a complete jump to the fourth paradigm will require a strong empirical foundation. A recent review by Coelho et al.[64] has identified the main challenge of utilising machine learning for corrosion research as the lack of high-quality datasets. Corrosion datasets are found to be typically noisy, rarely shared in a systematic machine-readable way, and lacking in time-dependent multidimensional input, which was shown to increase the accuracy of studied models. On the one hand, recent inhibitor data management initiatives such as CORDATA database[65] introduced open-source philosophies to inhibitor discovery and selection - however although database contains hundreds of entries, inhomogeneous data is still a problem. The database contains data acquired on different raw batches of alloys, different or poorly controlled ambient temperatures, and different experimental methods and conditions. On the other hand, dedicated state-of-the-art high-throughput datasets for aluminium alloys have created data for hundreds of organic compounds[28,36,59,63]. However, the lack of multidimensional input is a distinctive shortcoming of high-throughput methods, where only one parameter is collected to represent the inhibition performance. For an alloy prone to localised degradation, such as pitting corrosion of AA2024-T3, a data creation procedure that obtains information on both the open circuit state as well as behaviour under applied potentials is crucial to get the full mechanistic picture. All in all, in silico inhibitor screening combined with smart high-throughput testing has enabled overcoming physical limitations of previous paradigms. However, a complete jump to the fourth paradigm will require a strong empirical foundation. A recent review by Coelho et al.[64] has identified the main challenge of utilising machine learning for corrosion research as the lack of high-quality datasets. Corrosion datasets are found to be typically noisy, rarely shared in a systematic machine-readable way, and lacking in time-dependent multidimensional input, which was shown to increase the accuracy of studied models. On the one hand, recent inhibitor data management initiatives such as CORDATA database[65] introduced open-source philosophies to inhibitor discovery and selection—however although database contains hundreds of entries, inhomogeneous data is still a problem. The database contains data acquired on different raw batches of alloys, different or poorly controlled ambient temperatures, and different experimental methods and conditions. On the other hand, dedicated state-of-the-art high-throughput datasets for aluminium alloys have created data for hundreds of organic compounds[28,36,59,63]. However, the lack of multidimensional input is a distinctive shortcoming of high-throughput methods, where only one parameter is collected to represent the inhibition performance. For an alloy prone to localised degradation, such as pitting corrosion of AA2024-T3, a data creation procedure that obtains information on both the open circuit state as well as behaviour under applied potentials is crucial to get the full mechanistic picture.

We aim to address the need for a robust multidimensional time-dependent electrochemical database with this study. We also show the best practices for applying this multidimensional data to train a predictive machine-learning model. AA2024-T3 samples exposed to around 80 small organic molecules containing electrolytes are electrochemically characterised through linear polarisation resistance, electrochemical impedance spectroscopy and potentiodynamic polarisation. The goal of this brute force 'high-throughput' approach that combines proven electrochemical methods is to demonstrate a methodology to create robust data that contains mechanistic time-dependent information. Gained mechanistic information spans double layer capacitance, charge transfer resistance, diffusion of corrosive ions through a protective inhibitor layer from electrochemical impedance spectroscopy, time-resolved corrosion resistance response from linear polarisation resistance, and corrosion rate, potential, breakdown potential, the kinetics of the electrochemical reactions and nature of anodic and cathodic reactions at biased electrical potentials from potentiodynamic polarisation. The obtained experimental parameters can be employed directly as target parameters for training a machine learning model that is predictive of the performance of untested compounds to create a shortlist of promising candidates. Moreover, the experimental investigation yields additional input features that can be combined with molecular descriptors derived from the molecular structure and atomistic simulations. These input features exhibit the great potential to develop augmented quantitative structure–property relationships as they allow the direct inclusion of information on the underlying mechanisms in the model training. The results of this study are expected to support the development of faster inhibitor screening techniques in the future, which can leverage the link between the molecular structure of the inhibitor and its corrosion inhibition activity.

RESULTS AND DISCUSSION

Experimental results

Figure 1 plots the potentiodynamic polarisation (PDP), electrochemical impedance spectroscopy (EIS), and linear polarisation resistance (LPR) measurements of AA2024-T3 samples exposed to 0.1 M NaCl solution with and without the presence of 1 mM inhibitor candidates of benzotriazole, 2,5-dimercapto-1,3,4-thiadiazole, 2-mercaptobenzimidazole, 2-mercaptobenzoate, sodium acetate, sodium mercaptoacetate, or ammonium pyrollidinedithiocarbamate. The summary of values obtained from the experiments is presented in Table 1. In order to showcase the broad spectrum of behaviours observed in the electrochemical experiments, inhibitor candidates with contrasting characteristics were selected.

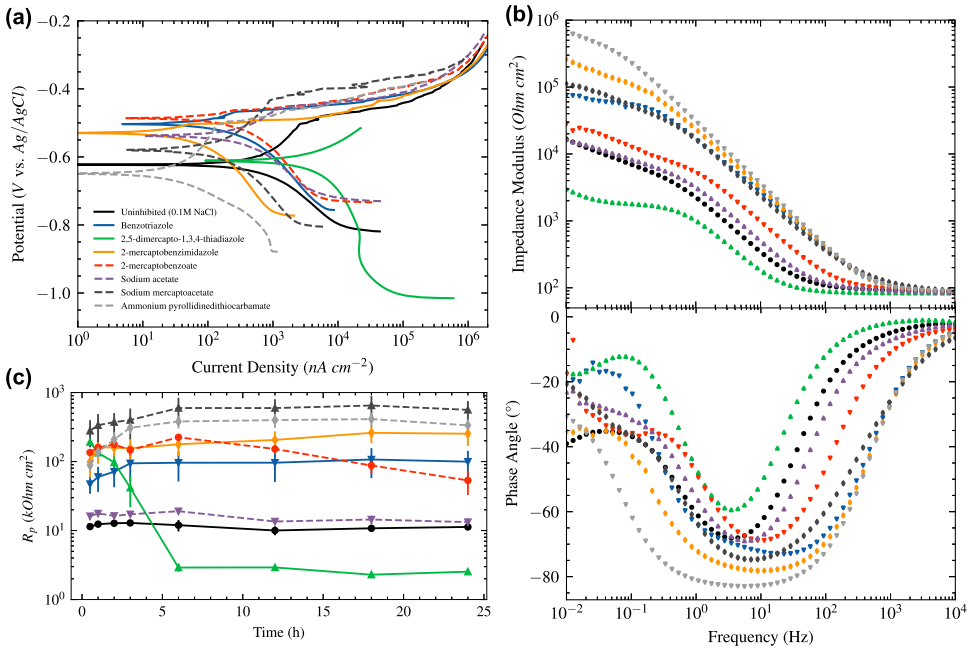

Fig. 1 AA2024-T3 samples exposed to 0.1 M NaCl solution in presence and absence of 1 mM inhibitors. a Potentiodynamic polarisation curves and **b** electrochemical impedance spectroscopy Bode modulus and phase angle plots recorded after 24 h of immersion, **c** linear polarisation resistance R_p values as functions of exposure time.

Table 1. Electrochemical information obtained from potentiodynamic polarisation, electrochemical impedance spectroscopy, and linear polarisation resistance measurements of AA2024-T3 samples exposed to inhibitor-containing solutions

Inhibitor	j_{corr} (nA cm^{-2})	E_{corr} (mV)	E_{br} (mV)	$\|Z\|_{2h}$ (kΩ cm^2)	$\|Z\|_{24h}$ (kΩ cm^2)	$R_p\|_{24h}$ (kΩ cm^2)	$\langle R_p \rangle$ (kΩ cm^2)
Uninhibited (0.1 M NaCl)	604 (±108)	−620 (±12)	−486 (±2)	14 (±0)	14 (±3)	11 (±1)	11 (±1)
Benzotriazole	216 (±38)	−500 (±3)	−475 (±4)	79 (±31)	107 (±48)	100 (±44)	94 (±43)
2,5-dimercapto-1,3,4 thiadiazole	3822 (±399)	−604 (±6)	−479 (±6)	51 (±18)	3 (±0)	3 (±0)	16 (±4)
2-mercaptobenzimidazole	79 (±18)	−523 (±6)	−496 (±8)	80 (±30)	265 (±80)	253 (±85)	207 (±66)
2-mercaptobenzoate	261 (±65)	−527 (±16)	−472 (±19)	130 (±50)	38 (±16)	53 (±21)	135 (±7)
Sodium acetate	396 (±54)	−563 (±14)	−473 (±13)	16 (±2)	16 (±1)	13 (±2)	15 (±1)
Sodium mercaptoacetate	57 (±13)	−572 (±25)	−435 (±34)	203 (±47)	203 (±64)	561 (±191)	555 (±205)
Ammonium pyrollidinedithiocarbamate	38 (±4)	−636 (±14)	−488 (±14)	346 (±34)	480 (±106)	335 (±73)	356 (±173)

Corrosion current density j_{corr}, corrosion E_{corr} and breakdown E_{br} potentials vs. Ag|AgCl, impedance modulus values $|Z|$ observed at 10^{-2} Hz evaluated for 2 and 24 h, linear polarisation resistance R_p evaluated at 24 h and the time-weighted average of the measurements $\langle R_p \rangle$ are presented.

Figure 1a presents polarisation curves of AA2024-T3 samples recorded after 24 h of immersion in inhibitor-containing solutions. Polarisation curves show that the addition of small organic molecules results in corrosion current densities varying up to 2 orders of magnitude. It is noteworthy that the best inhibitor candidates reduced the corrosion current densities more than 10-fold compared to the uninhibited samples. Analysis of corrosion potentials shows that inhibitors act as mixed or anodic inhibitors. Anodic inhibitors reduce the current densities of partial oxidation reactions without affecting the partial reduction reactions, causing the shift of the corrosion potential in the positive direction (and vice versa for cathodic inhibitors)[66]. Albeit small, the addition of organic molecules shifts the corrosion potentials to more positive values, with the exception of ammonium pyrollidinedithiocarbamate. However, when breakdown potentials (potentials where a sudden increase in current for the anodic curves) are observed it is seen that the introduction of molecules resulted in negligible shifts with the exception of 2,5-dimercapto-1,3,4-thiadiazole. The distribution of electrochemical potentials among all inhibitor candidates is analysed more deeply in section "Understanding and Prediction Inhibition: Experimental Input Features for the Machine Learning Model".

Figure 1b shows the EIS impedance Bode modulus plots after 24 h of immersion in inhibitor-containing solutions. The impedance modulus $|Z|$ values observed at 10^{-2} Hz frequency are treated as the R_p values calculated from EIS, as it was shown that it reflects the corrosion resistance of the inhibitor–substrate interface[67]. This approach is based on a simplification since the low-frequency impedance modulus includes contributions from the oxide film resistance, the charge transfer resistance, and often from the diffusion-controlled processes. Moreover, in addition to the real component, it includes the imaginary part. $|Z|$ values show more than a 2-order of magnitude range as was seen for corrosion current density measurements. Corrosion resistance with respect to the uninhibited samples showed more than a 30-fold increase. A comparison of low-frequency impedance modulus values

observed at the 2nd and 24th hour presented in Table 1 shows significant variation in inhibitor behaviour. This change from the 2nd to the 24th hour is more clearly observed in LPR plots, which correspond well with EIS results.

Figure 1c shows estimated R_p results calculated from the LPR measurements conducted throughout 24 h. The instantaneous corrosion resistance of a system can be indirectly assessed by measuring the polarisation resistance R_p. A higher R_p indicates a more resistive interface between the electrode and the electrolyte. The resistive interface hinders the flow of electrons and ions, increasing the corrosion resistance[68]. From the LPR measurements, it is clear that the action of inhibitor species is highly time- and species-dependent. In some cases such as sodium acetate, there is negligible change in behaviour compared to the uninhibited solution. However, in most cases, it was observed that instead of having a constant behaviour, R_p values evolve with time. In cases such as benzotriazole, 2-mercaptobenzimidazole, sodium mercaptoacetate and ammonium pyrollidinedithiocarbamate, there is an initial increase in R_p, and further development of corrosion protection until the 6th hour and stable corrosion protection after that. For 2-mercaptobenzoate it was seen that after an initial increase and a gradual development of corrosion resistance, the protection started to decrease to lower than initial values. For 2,5-dimercapto-1,3,4-thiadiazole it was seen that after the initial, more than an order of magnitude increase in R_p, the protection starts to decrease. This decline continues until the 6th hour and signifies stable active corrosion behaviour afterward.

In the specific case of 2,5-dimercapto-1,3,4-thiadiazole, we conclude that this accelerated corrosion was caused by the pH change of the electrolyte after the introduction of the inhibitor. Analysis of pH measurements of the electrolytes prior to the electrochemical experiments shows that compared to the pH value of 6 of the uninhibited 0.1 M NaCl solution, 2,5-dimercapto-1,3,4-thiadiazole containing solution had an acidic pH value of 3. This is at the boundary of the thermodynamically stable region of Al at 1M Al^{3+} but in the region of preferential stability of Al^{3+} at lower than 1 M contraception of Al^{3+} [69], which is expected for OCP corrosion of AA2024-T3. Therefore the considerable decrease in pH must have disrupted the stable aluminium (hydr)oxide layer and led to active corrosion of the samples.

Due to this dynamic corrosion and inhibition behaviour, it is vital to capture the performance during the whole time-span. One method to achieve this is to estimate the mean value of R_p through a trapezoidal integration over time:

$$\langle R_p \rangle = \frac{1}{t_f - t_0} \int_{t_0}^{t_f} R_p(t) dt \tag{1}$$

$$\approx \frac{1}{t_f - t_0} \sum_{k=1}^{N} \frac{R_p(t_{k-1}) + R_p(t_k)}{2} (t_k - t_{k-1}) \tag{2}$$

where t_f is the final measurement time, t_0 is the initial measurement time, and k is the indices for the performed discrete measurements. The mean estimated this way can be used as a screening metric that contains all time-dependent information in one number. The power of this approach as an inhibitor screening tool was recently shown for pure copper substrates exposed to small organic molecules[33].

Quantifying inhibitor performance

The electrochemical information obtained from the techniques PDP, EIS and LPR can be used to compare the performance of inhibitors. However, it is not possible to directly compare the electrochemical information obtained from different measurement techniques. To enable a more direct comparison between techniques, the results can be converted into relative protection values by comparing the results obtained from the inhibited solutions to the uninhibited ones.

The most widely used metric for comparing the inhibitor performance in the literature is the *inhibition efficiency (IE)*. The inhibition efficiencies are calculated from polarisation resistances R_p obtained from LPR or EIS, the cases when the inhibitor value is higher than blank:

$$\eta = \frac{{R_p}^{inh} - {R_p}^{blank}}{{R_p}^{inh}} = \left(1 - \frac{{R_p}^{blank}}{{R_p}^{inh}}\right) \times 100\% \tag{3}$$

and corrosion current densities j_{corr} obtained from PDP, the cases when the inhibitor value is lower than blank:

$$\eta = \frac{{j_{corr}}^{blank} - {j_{corr}}^{inh}}{{j_{corr}}^{blank}} = \left(1 - \frac{{j_{corr}}^{inh}}{{j_{corr}}^{blank}}\right) \times 100\% \tag{4}$$

where superscripts inh and blank stand for inhibited and uninhibited samples, respectively.

Inhibition efficiency is used widely because it is an easy-to-understand comparison tool. For inhibition, it has values between 0 (no protection at all) to 100% (complete prevention of corrosion). Negative values indicate an acceleration of corrosion compared to the uninhibited case. It is also favoured as under simplifying assumptions it can directly be correlated to the surface coverage by the inhibitor molecules. However, this ease of use obscures the fact that as a mathematical function, this mapping introduces a mathematical bias and as a result is highly non-linear. Due to its form $\left(1 - \frac{a}{b}\right)$, inhibition efficiency introduces an arbitrary 1 next to the relative values $\left(\frac{a}{b}\right)$ that is of actual interest. As a result, minor differences in performance are seen as large jumps for the lower efficiencies (<90%), and major differences are hidden from view at higher efficiencies (>90%). This also causes researchers to wrongly conclude that good-performing inhibitors would also have lower standard deviations since even major variations in electrochemical values are suppressed at the higher end of the inhibition efficiency metric. Therefore, it is not an optimal metric to compare the protection performance of strong inhibitors.

An alternative metric, *inhibition power (IP)*, has recently been proposed to address the limitations of inhibition efficiency[52]. It is the ratio of inhibited and uninhibited inhibition information presented in a logarithmic fashion. For polarisation resistance R_p it is defined as

$$P_{inh} = 10\log_{10}\left(\frac{{R_p}^{inh}}{{R_p}^{blank}}\right) \tag{5}$$

and for corrosion current densities j_{corr} it is defined as

$$P_{inh} = 10\log_{10}\left(\frac{{j_{corr}}^{blank}}{{j_{corr}}^{inh}}\right) \tag{6}$$

By taking only the ratio of electrochemical values into account, inhibitor power eliminates the influence of bias introduced by the arbitrary $(1 - \frac{a}{b})$ form of inhibitor efficiency determination. In this form, an inhibition power increase of 10 from an uninhibited condition corresponds to a corrosion resistance increase by 10-fold, while an increase of 20 corresponds to a 100-fold corrosion resistance increase.

Comparison of electrochemical techniques: inhibition efficiency vs. inhibition power

The comparison of electrochemical results converted into inhibition efficiency and inhibition power metrics is presented in Fig. 2. Example correlations between EIS measured at 24th hour with time-weighted LPR average $\langle R_p \rangle$ for individual experimental runs are visible in Fig. 2a. Figure 2b quantifies the correlation between different electrochemical measurement techniques in the form of

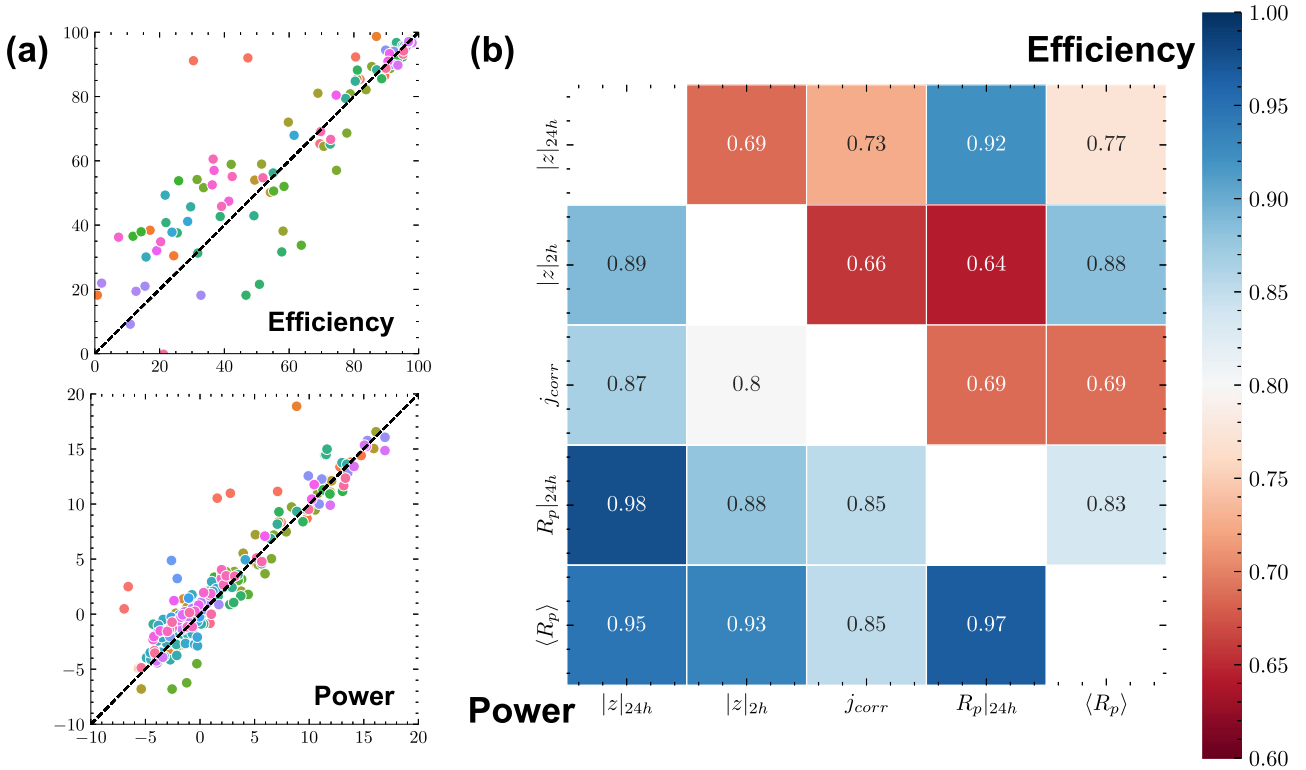

Fig. 2 The correlation between different electrochemical measurement techniques: EIS performed at the 2nd and 24th hour $|Z|_{2h}$ and $|Z|_{24h}$, potentiodynamic polarisation performed at 24th-hour j_{corr}, LPR performed at the 24th hour $R_p|_{24h}$ and the time-weighted average of LPR measurements $\langle R_p \rangle$. a Example correlation between $|Z|_{24h}$ and $\langle R_p \rangle$, values from electrochemical measurements converted into top: inhibition efficiency, bottom: inhibition power. Each dot represents an individual measurement, categorised in colours with respect to their inhibitor species. **b** Pearson correlation coefficients between different electrochemical measurements, converted in top-right triangle: inhibition efficiency, bottom-left triangle: inhibition power metrics.

Pearson correlations. *P*-values (value describing how likely it is that your data would have occurred under the null hypothesis of your statistical test) of the Pearson statistical test correlations were between 10^{-134} and 10^{-51}, much lower than the commonly used criteria 10^{-6}, indicating statistical significance.

The differences between inhibitor efficiency and power correlations are vividly seen in Fig. 2a. For inhibition efficiency, the correlations are weak except for the top right part, the best-performing inhibitors. This might falsely lead to the impression that an increase in inhibitor performance results in a higher correlation between experiments. This impression is misleading and is an artefact of the mathematical function used for converting raw electrochemical information into inhibition efficiency. When the correlations are visualised in the form of inhibitor power, higher correlations between the good-performing inhibitors are lost. All compounds behave in a similar way and cluster around the perfect correlation diagonal.

The only exceptions to the strong correlation seen for inhibitor power are the compounds that change their corrosion protection behaviour throughout time. Given that $|Z|_{24h}$ measures the protective properties at the 24th hour, and $\langle R_p \rangle$ captures additional time-dependent information, this behaviour is completely expected.

Apart from being more consistent, inhibitor power facilitates discerning between the better and best inhibitors. As more conceptually argued in the previous section, the inhibition efficiency metric squeezes the high-performing compounds together. This is clearly visible from the clustering of experiments for the efficiency metric, versus individually identifiable best-performing compounds for the power metric in Fig. 2a.

The clustering seen for the inhibition efficiency metric also creates an issue for training a predictive model. Imbalanced data usually results in models that have poor predictive performance, especially for the minority class[70]. The homogeneous distribution of results is crucial in training an unbiased machine learning model, which is better provided with the inhibition power metric.

Figure 2b presents the correlations between the electrochemical measurement techniques more quantitatively in the form of Pearson correlations. The top-right triangle shows the correlations between different electrochemical measurement technique results converted into inhibition efficiency, and the bottom-left triangle shows the same results converted into inhibition power.

Pearson's bivariate sample correlations quantify linear correlations between two sets of data with the following formula:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad (7)$$

where n is the total number of experiments (in this work ~ 300), i is the index representing different experiments, x_i, y_i individual sample points from two different electrochemical measurement methods, $\overline{x}$, $\overline{y}$ sample means obtained from the different electrochemical methods. The correlation coefficient $r_{x,y}$ can take values from −1 to 1. 1 indicates a perfect linear relationship between x and y, where all data points lie on a line where x increases as y increases, and vice versa for −1. A value of 0 indicates that there is no linear relationship between the two variables. For time-invariant electrochemical behaviour, a correlation coefficient of 1 is expected between the different electrochemical measurement results[71].

A quick comparison of the inhibition efficiency and power correlation triangles shows that the correlations between

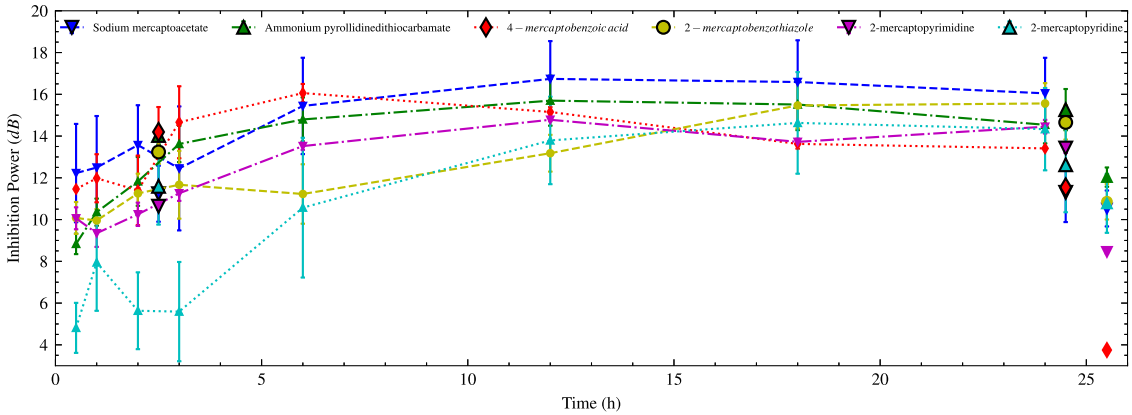

Fig. 3 The representation of PDP, EIS, and LPR measurements converted into inhibition power for best performing among the tested inhibitors. Small line-scatter plots represent LPR, larger plots with black edges represent EIS at 2nd and 24th hours, and the final larger scatter plots represent the PDP measurement results converted into inhibition power. The solubilities of inhibitors denoted with italics were <1 mM.

measurements are consistently lower for the inhibition efficiency metric. For the inhibition efficiency, the correlations between different techniques are all below 0.9, with the exception of LPR and EIS measurements performed at the 24th hour. For the inhibition power, LPR and EIS measurements carried out at the 24th hour and time-weighted LPR average $\langle R_p \rangle$ show very high correlations. EIS performed at 2nd hour shows the lowest correlations with the rest of the measurements. Trustworthy EIS measurements require the electrochemical system to be linear, causal, and time-invariant within the time-frame of the measurement[33,34]. However, for dynamic systems similar to the ones shown in Fig. 1c, time-invariance would not be often observed at measurements done at 2nd hour, which would explain the low correlations. The highest correlation of EIS performed at 2nd hour was observed with a time-weighted parameter, $\langle R_p \rangle$. This again emphasises the time-variable inhibitor behaviour. For inhibition power, higher correlation was observed between $\langle R_p \rangle$ and EIS performed at 24th hour, compared to $\langle R_p \rangle$ and EIS performed at 2nd hour indicates that measurements at the 24th hour were more representative of the time-dependent corrosion inhibition behaviour. Surprisingly for inhibition efficiency, the opposite is the case. This might be due to the volatility inherent to the inhibition efficiency transformation.

PDP measurements show lower correlations with the rest. This is most likely due to altered electrochemical behaviour caused by the high overpotentials (±250 mV) necessary for the PDP experiments. Due to the high overpotentials encountered during the potentiodynamic scans, the physicochemical properties of the surface are modified, potentially leading to an altered substrate surface chemistry[33,72]. Another reason could be the increased user input during the Tafel slope analysis required for corrosion current density calculations, which is much higher than required for EIS or LPR. Specifically for the case of AA2024-T3, the use of the Tafel approach is not straightforward. On one side, the cathodic behaviour is significantly influenced by oxygen diffusion limitations. On the other, the anodic processes are not solely governed by charge transfer but rather occur at localised regions such as intermetallic particles and grain boundaries. Therefore, the conventional Tafel approach cannot be employed since it is applicable only under activation-controlled processes. Tafel analysis in such conditions is a very simplified approach and can lead to deviations.

For the reasons presented above, we argue that inhibition power is a more 'efficient' way of discerning between better and best inhibitors, and a better approach to training an unbiased predictive machine learning model. Therefore, it is used to compare and rank the inhibitor performance in the next section.

Ranking of inhibitors

Figure 3 demonstrates the electrochemical measurement results converted into inhibition power for the best-performing inhibitors. LPR measurements are shown with line-scatter, EIS with black framed scatter, and PDP with the final individual scatter plots. The width of the EIS and PDP symbols was chosen so that it would convey the time it takes to perform the measurements.

The presented inhibitors show stable behaviour after 6 h, with the exception of 2-mercaptobenzothiazole which develops inhibition until after 18 h, and of 4-mercaptobenzoic acid which seems to show a minor decrease in inhibition performance with time. LPR and EIS results correlate strongly with each other (except for 2-mercaptopyridine EIS around 2 h), as expected from the analysis in the previous section "Comparison of electrochemical techniques: inhibition efficiency vs. inhibition power". PDP results demonstrated a lower inhibition power for all cases. This systematic difference was attributed to the destructive nature of PDP measurements.

Although a qualitative analysis is possible through such plots, the quantitative ranking of a high number of inhibitors is not feasible through such visualisations. To this end, the time-weighted LPR average $\langle R_p \rangle$ is advantageous as it captures the complete time-dependent behaviour in a single number. Additionally, it shows high correlation with other electrochemical techniques as seen in Fig. 2b.

Figure 4 presents the ranking of inhibitor candidates in the form of a box-plot, created from the time-weighted LPR average $\langle R_p \rangle$ values converted into inhibition power through Eq. (5). Inhibitor candidates are ranked with respect to their mean inhibition power values, and their medians are represented by horizontal bars. The box part shows the main portion of the data, the interquartile range. The edges of the box show the 25th and 75th percentile. Whiskers show the minimum and maximum measurement results.

The importance of heteroatom presence, aromatic ring and π-bond containing molecular structures on inhibitor performance has been consistently mentioned in the literature[30,73–75]. It has been argued that the availability of non-bonded lone pair electrons of heteroatoms and π-electrons of double/triple bonds facilitate electron transfer from the inhibitor to the *d*-orbitals of the metal, acting as adsorption centres during metal–inhibitor interactions. To identify such trends in this experimental data set, the inhibitors have been categorised according to their molecular structures: the presence of N, O, and S heteroatoms and their aromatic vs. aliphatic bond structures.

Almost half of the inhibitor candidates behaved as corrosion accelerators. This was in contrast to the findings of previous studies[28,59], which was the basis of the inhibitor selection procedure

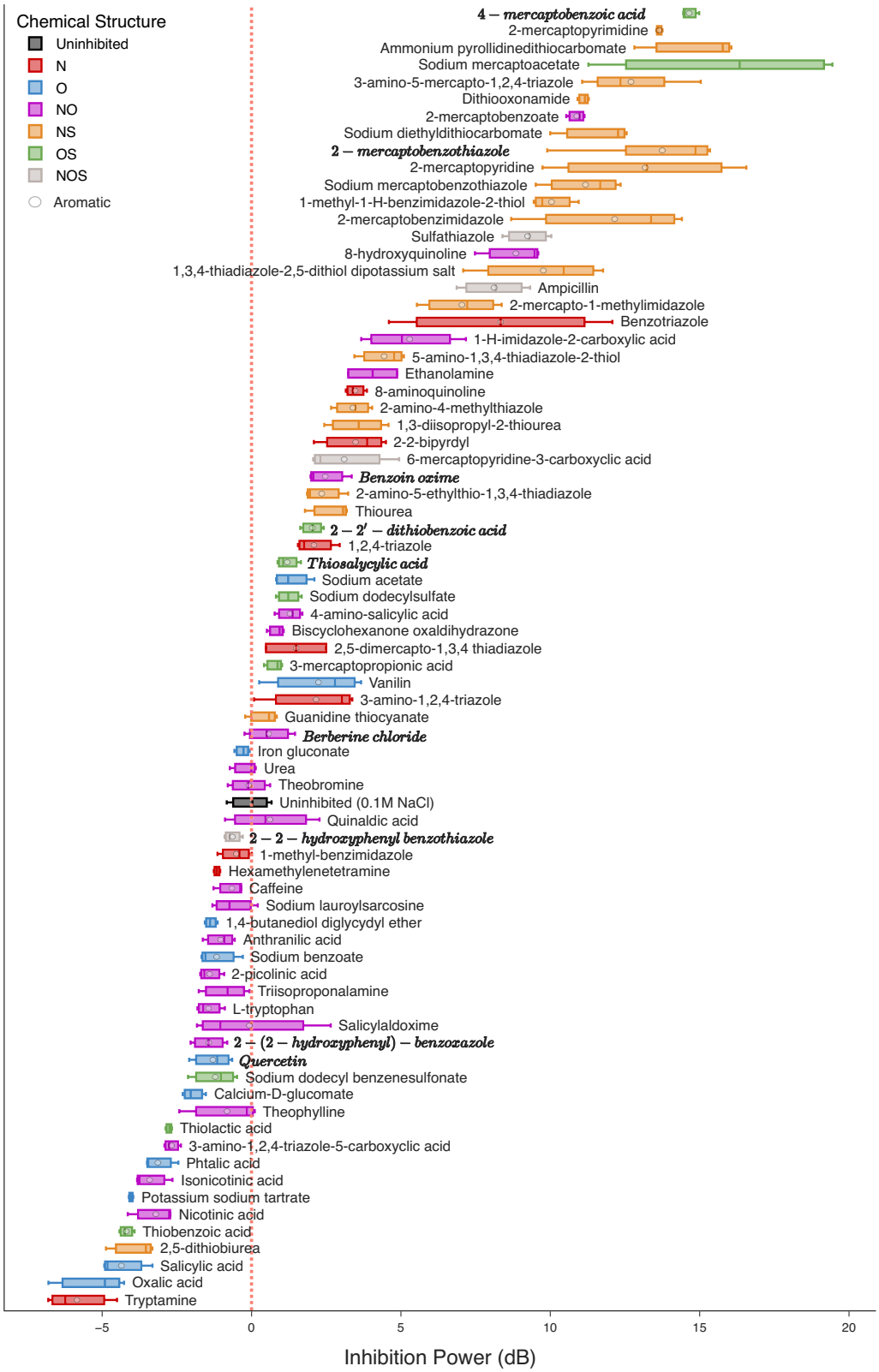

Fig. 4 The inhibitor candidate ranking visualised as boxplots. Colours indicate the nitrogen (N), oxygen (O), sulfur (S) heteroatom content and presence/absence of aromatic ring structures. The solubility of molecules denoted with italics was <1 mM.

of our paper. Non-adjusted pH could be one reason for this behaviour. 80% of sole O heteroatom-containing compounds behaved as accelerators, with the exception of sodium acetate and vanillin. On the other hand, N and S heteroatom-containing organic molecules performed consistently well. They had the highest inhibition power values with none of them performing as corrosion accelerators. Compounds that contained N, S and O together had in-between inhibition properties. This leads us to suggest that N, S

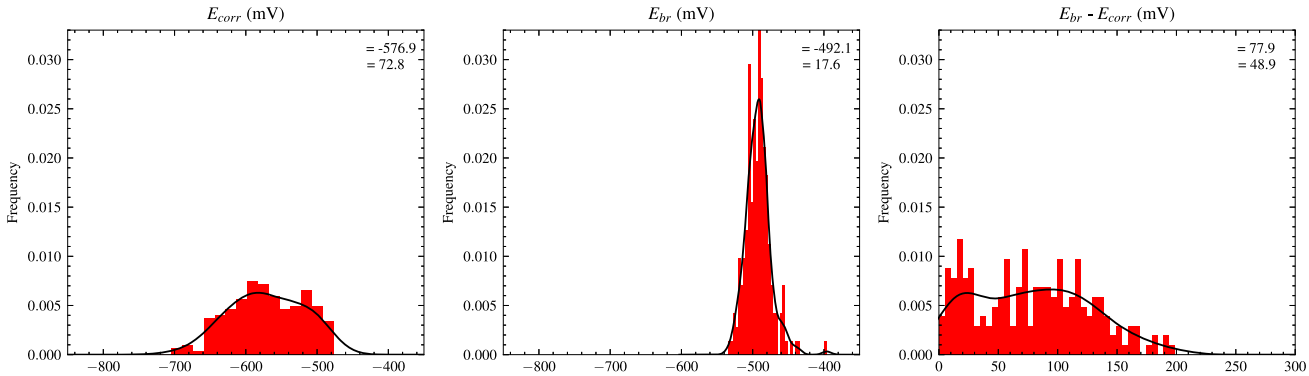

Fig. 5 The distribution of corrosion potentials E_{corr}, pitting breakdown potentials E_{br} where a sudden jump in anodic current is observed, and the differences between the two. Histograms are shown as red bars, and kernel density estimates of the probability functions are shown as black curves. μ and σ represent the mean and standard deviation, respectively.

heteroatoms grant inhibitive properties to the organic molecule, whereas O could potentially hinder inhibition. Specifically for AA2024-T3 corrosion inhibition, it was observed that functional groups with N and S heteroatoms form coordination complexes with Cu-containing intermetallics, reducing the corrosion rate[76–78]. The heteroatom trend is generally in line with the previously suggested heteroatom electronegativity-inhibition effect, where heteroatoms provided inhibition with inverse order of their calculated individual electronegativity: P > S > N > O[73]. It is proposed that lesser electronegativity results in increased charge transfer and provide inhibition. However, the real situation in a small organic molecule is much more complex as the electronegativities of the heteroatoms will change depending on the molecular structure.

The discussion above addresses only one of the important molecular descriptors. Trends are not clear for the rest. The comparison of aromatic and aliphatic behaviour shows no significant difference. This is most likely because the tested aliphatic molecules already contain excess π-bonds in their linear chain. The behaviour of N and O, and S and O containing molecules are most complex. The molecules are spread throughout the inhibitor/accelerator spectrum, seemingly without an underlying order and act as best and worst performing compounds, as seen in the behaviour of 4-mercaptobenzoic acid and thiobenzoic acid.

Understanding and predicting inhibition: experimental input features for the machine learning model

It is clear that any predictive corrosion inhibition model requires a more comprehensive description of the system than the presence or absence of heteroatoms. Compared to analysing individual properties like the presence of certain heteroatoms, π-bonds and functional moieties, quantitative structure-property relationship (QSPR) models have potential in exploring more complex physical phenomena[62,79–88]. QSPR inhibition models relate predictor variables, which can be physicochemical properties and/or theoretical molecular descriptors of inhibitor compounds, to the experimentally measured inhibition performance. Quantified physicochemical properties or descriptors (obtained through theoretical calculations and molecular modelling techniques such as density functional theory and molecular dynamics) expressed in a mathematical relationship, a quantitative structure-property relationship, can be established to predict the performance of untested organic molecules.

The inclusion of experimental physicochemical descriptors is the next logical step to supplement the input feature pool and to concomitantly improve the robustness of the predicted values as well as the generalisability of QSPR models for small organic corrosion inhibitors. Some important physical and chemical experimental input features that are capable of increasing prediction quality are presented below.

Molecular weight. Molecular weight–inhibitor power relationship can be found in Supplementary Fig. 2. It seems that most organic molecules cluster in the range of 100–200 g mol^{-1}, and after around 250 g mol^{-1} there seems to be a decrease in the inhibitor performance. This is most likely due to steric hindrance effects, where an increase in the size of the molecule would hamper the adsorption reactions with the substrate[89]. Based on this observation we suggest that as a rule of thumb, small organic molecules with molecular weights lower than 250 g mol^{-1} can hold more promise to be inhibitor candidates. This would limit the chemical space to be explored and facilitate the efficiency of novel inhibitor discovery.

Inhibitor concentration. The influence of concentration is certainly important for inhibition behaviour. An exploratory comparison of 6 molecules at 0.1 and 1 mM concentrations shows that with increasing concentration, inhibitor systems become more protective and accelerator systems become more corrosive (provided in Supplementary Fig. 5). Typically as concentration increases, a corresponding increase in inhibition is observed until a critical concentration is reached. After this critical concentration the inhibition either reaches a plateau or in certain cases starts to decline[90]. It was previously argued that the decline in inhibition was related to the formation of oligomers: either the molecule concentration higher than the critical value causes adsorbed inhibitor molecules to desorb due to interaction with free molecules present in the solution, forming oligomers, or oligomers that form in the solution beforehand reduce the concentration of inhibitor available for adsorption[91]. Any analysis of an inhibition system has to be aware of such behaviour when comparing inhibition performance at different conditions.

Electrochemical potentials. Electrochemical information obtained from the experiments can serve as target parameters to be predicted (such as previously calculated inhibition power) and also can be utilised as descriptors. This can augment the molecular descriptors of the model by adding mechanistic insights related to the electrolyte–electrode system, which were otherwise lacking from the statistical nature of machine learning models.

The dominant degradation mechanism of AA2024-T3 is pitting corrosion[92,93]. Furthermore, the alloy is used in combination with composite structures in modern aeroplanes which triggers galvanic corrosion. For this reason, parameters that represent pitting and galvanic corrosion hold promise as either target parameters to be predicted, or as additional descriptors that provide mechanistic information to the models.

Figure 5 presents the distribution of corrosion potentials E_{corr}, pitting breakdown potentials E_{br} where an instantaneous large increase in anodic current is observed[94], and the differences between the two. It is seen that inhibitors can modify E_{corr}

significantly, as seen from the 200 mV range and high standard deviation of 72.8 mV. On the other hand, E_{br} values change negligibly, with a standard deviation of 17.6 mV, leading us to believe that this is an intrinsic property of the substrate. This is in line with previous dealloying studies, where an alloy-dependent intrinsic critical potential was observed for activating porosity formation in an otherwise passive surface[95]. E_{br} acts as the threshold potential for preferential dealloying of the active phases, which in the case of AA2024-T3 is the potential for initiating stable pits resulting from active S (Al_2CuMg) and θ (Al_2Cu) phase intermetallics[96]. The difference between potentials $E_{br}-E_{corr}$ describes the overpotentials required to reach this threshold, which was shown to be highly influenced by the introduction of inhibitors.

The influence of difference in potentials $E_{br}-E_{corr}$ is denoted here as passive range, and plotted with respect to inhibitor power in Fig. 6 to see whether there is a correlation between the two parameters. Different chemical groups are denoted with different colours. No significant correlation was observed between the passive range and inhibition performance. It was observed that apart from NS aliphatic and OS aliphatic/aromatic compounds, a weak negative correlation between the two parameters was observed. However, this behaviour was not statistically significant because of the high spread observed for the experiments. In any case, the seemingly unsystematic behaviour with low correlation highlights the need for further study. As the key parameter for localised electrochemical activity, passive range holds promise either as a target to be predicted on its own or as a descriptor to be used in combination with the molecular descriptors.

Bulk pH. Apart from 4-mercaptobenzoic acid, sodium diethyldithiocarbamate, and 1,3,4-thiadiazole-2,5-dithiol-dipotassium salt, the pH did not change in the presence of compounds with good inhibition performance (IP > 10, IE > 90%) and had neutral pH values around 6. On the other hand, IP was lower in the presence of compounds which caused the initial pH of the electrolyte to be out of the 4.5–8.5 Al stability window[97]. The clustering of lower pH values at the lower inhibition power segment suggests that this results in active corrosion becoming the dominant degradation mechanism.

It was seen that there is no correlation of inhibition power with either the average or the difference in pH (Supplementary Figs. 3 and 4). It must be noticed that what is measured as bulk electrolyte pH and what the actual pH observed on the substrate surface can be very different, and bulk electrolyte measurements do not fully reflect local behaviour such as concentration gradients at the electrolyte–substrate interface and throughout the diffusion layer[98].

The lack of correlation does not mean that bulk pH information is useless as a machine-learning model feature. It is very relevant for explaining the outlier behaviour, as the pH difference caused by the inhibitor molecule is not captured directly with computational descriptors. The addition of bulk pH as a feature can capture such pH-based behaviour, and can be used as a forensic analysis tool to explain outliers of the model.

Exploring experimental descriptors for machine learning

Experimentally measured pH shows the power of descriptors obtained from experiments. To produce a short list of compounds with possibly useful properties for further experimental testing. The selection of relevant input features is a crucial step in the development of QSPR models as features with low or no relevance to the target property will degrade the model. The recursive feature elimination (RFE) was carried out for the four distinct groups of input features: structural features only, structural features combined with DFT, structural features combined with average pH, and structural features combined with DFT and average pH. Feature elimination was performed for both IE and IP targets. The whole feature selection process was repeated 100 times with different random seeds and the *n*-tuples that were selected in the majority of the runs can be found in

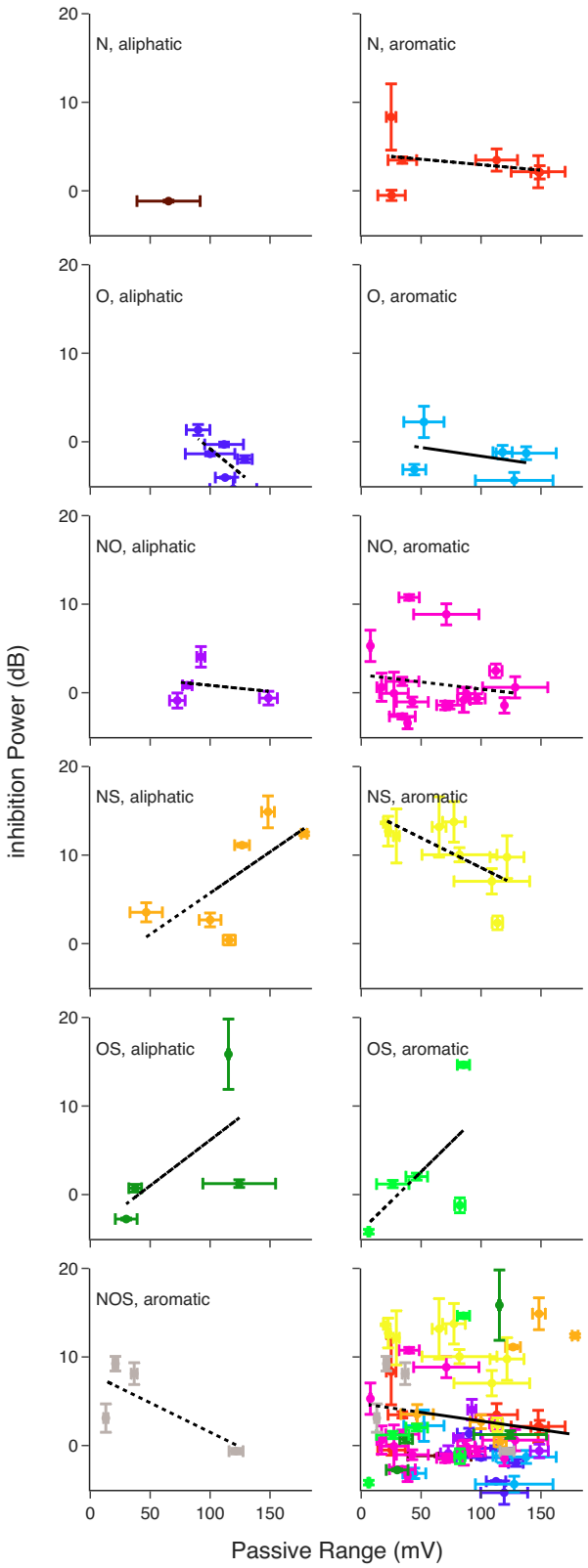

Fig. 6 The correlation between inhibitor power and the passive range ($E_{corr}-E_{br}$).

Table 2. Results of one specific train test split

Target	# Features	RMSE structural	RMSE structural + DFT	RMSE structural + pH	RMSE structural + DFT + pH	R^2 structural	R^2 structural + DFT	R^2 structural + pH	R^2 structural + DFT + pH
IE	10	0.24	0.23	0.18	0.18	0.17	0.19	0.49	0.51
	5	0.22	0.24	0.2	0.19	0.27	0.13	0.42	0.43
IP	10	0.19	0.19	0.18	0.18	0.25	0.31	0.32	0.38
	5	0.2	0.15	0.16	0.15	0.19	0.55	0.49	0.54

Supplementary Tables 2–5. To use the same technique for the QSPR step that was employed for sparse feature selection, random forest (RF) models have been trained using the experimental database. By algorithmically eliminating the weakest features, it allows automatic feature selection without user bias or intervention[62]. Moreover, RF is an ensemble model that builds multiple decision trees and combines their predictions. Naturally, this ensemble approach helps reduce the risk of overfitting, which can be crucial when dealing with small datasets. Another advantage is robustness against outliers: outliers can have a significant impact on smaller datasets, whose influence again can be mitigated by aggregating predictions from multiple trees.

RF regression models predicting the quantitative inhibition performance values were trained to create an active material discovery loop to explore the vast chemical space for promising compounds in an efficient manner. Out of the 78 organic molecules that were tested, only 59 were fully dissolved in solutions. These molecules corresponded to a target concentration of 1 mM and were used to train the ML models. As the input to these models molecular descriptors (MDs) based on the structure of the molecules, descriptors calculated by DFT as well as selected experimental parameters have been used. The accuracy and robustness of the trained models are assessed using a cross-validation (CV) approach.

In aqueous solutions, aluminium alloys have a protective passive (hydr)oxide layer preventing them from corrosion at a pH range roughly between 4 and 10[97]. In this pH range, scratches or mechanical damage to the passive layer are quickly repaired but if the pH drops below or rises above the stable range, aluminium starts to corrode actively. As the oxide layer is no longer stable at such conditions, this influences the inhibitor-substrate interaction. As a result, the pH makes for an effective feature in an ML model because aluminium is typically more likely to corrode at very high or very low pH levels. The pH is selected by the RFE routine every time it is part of the set of input features. This demonstrates emphatically that pH appears to be a key feature in the prediction of the inhibition performance of organic molecules.

In addition to pH, several properties derived from DFT calculations are also selected by the RFE as soon as they are included in the set of input features. The DFT parameter that was selected most frequently, was the highest occupied molecular orbital (HOMO). Additionally, the lowest unoccupied molecular orbital (LUMO) and dipole were selected in at least half of the cases, with $n = 10$ for the RFE step. This contrasts with recent works, which have concluded that the correlation between DFT properties and the corrosion-inhibiting effect of small molecules seems absent.[52,99] However, neither of these works mixed the DFT features with molecular descriptors that encode the molecular structure. It is noteworthy that the correlation between the HOMO energy levels and IE/IP is essentially zero in this work as well, corroborating these prior works.

When examining the results for one specific train test split in Table 2, it is evident that at least for most of the cases where the DFT parameters and/or pH value are added to the set of input features, the R^2 increases and the RMSE decreases. This indicates that including these parameters enhances the prediction and increases the reliability and robustness of the models. The only case where this does not hold is the IE model with five input features combining structural features and DFT. A closer examination reveals that for IE ten input features allow for more accurate predictions than five, whereas for IP the reverse is true. Lowest RMSE and highest R^2 were achieved for the model that uses combined descriptors and IP as the target. In Fig. 7 the measured IP is plotted against the IP predicted by the RF.

In order to perform CV, the dataset was divided into six folds and thus six RF models were trained. The average R^2 and RMSE and the corresponding standard deviation of these models are shown in Table 3. The evaluation of the models using different classes of input features indicates that adding DFT parameters and/or the pH value increases the prediction accuracy in most of the cases according to the determined mean values for R^2 and RMSE. The models with the lowest RMSE and highest R^2 include pH and DFT parameters as input in addition to the structural features, further supporting our claim that molecular descriptors derived from atomistic simulations can be helpful to generate QSPR models that predict the corrosion inhibition responses of small organic molecules to lightweight engineering metals such as aluminium and magnesium alloys. Unlike the specific train test split case, the lowest RMSE was obtained for IE as a target, and the highest R^2 was achieved for IP as the target. Unfortunately, high variation among different folds makes it difficult to state with certainty whether IP or IE performs better targets for such models. However, the comparably low R^2 and RMSE values for all considered models and the high standard deviations of these metrics indicate that more training data is required to achieve better generalisation. Furthermore, they are highly sensitive to outliers in the blind test set.

In summary, we employed various standard electrochemical techniques at different intervals to investigate the electrochemical behaviour of around 80 small organic molecules. Our aim was to capture the most comprehensive electrochemical picture of AA2024-T3 immersed in inhibitor-containing electrolytes. The performance of inhibitor candidates was quantified through statistical analysis of their electrochemical response. This highlighted the need for complementary information from different techniques to have a mechanistic understanding of an inhibition system. For initial inhibitor screening purposes, time-weighted LPR measurements showed very high correlations with other techniques and are a good substitute for representing the protective behaviour of the inhibitor. Time-dependent measurements showed that for the majority of organic molecules electrochemical measurements performed in less than 6 hours varied in time and were unstable. To understand the true inhibitive properties of inhibitor candidates, electrochemical studies should analyse the inhibition performance at least after 6 hours for more reliable results. Statistical analysis shows that inhibition efficiency is not an 'efficient' way to distinguish between good inhibitors. Inhibition power is a more suitable metric for discerning between "better" and "best" inhibitors. Inhibition power eliminates clustering of data observed in a higher efficiency range (>90%), which is an important condition for training an unbiased machine learning

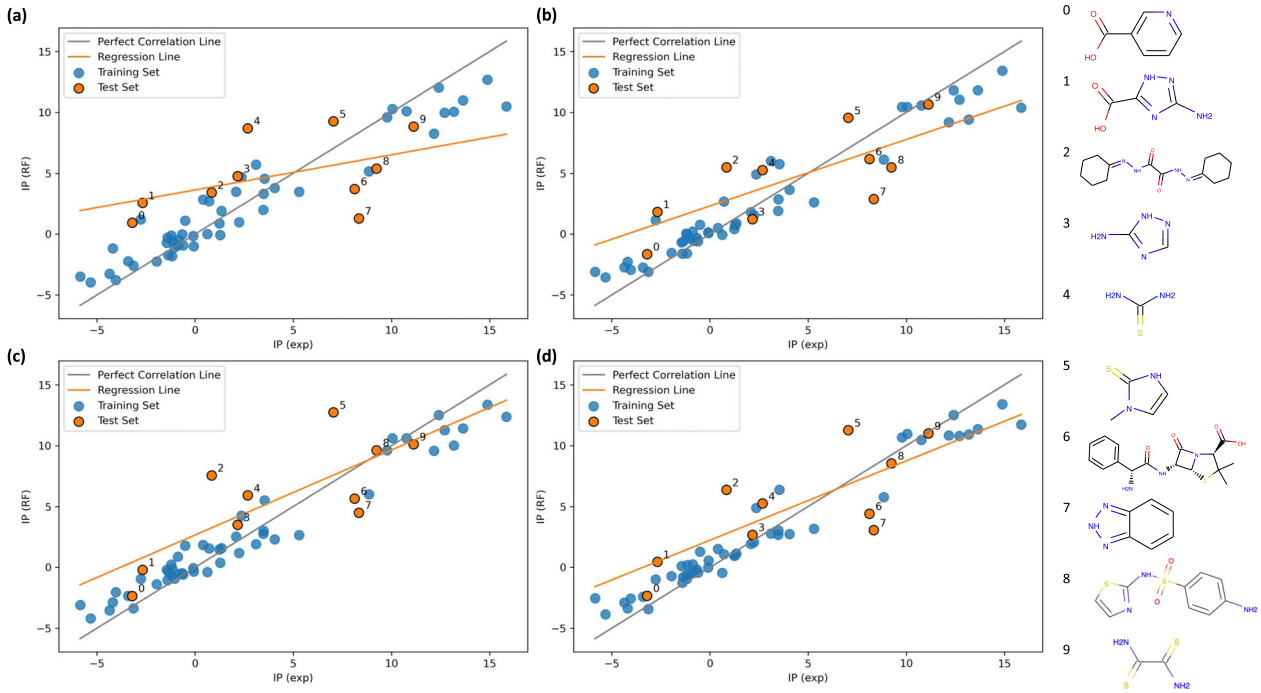

Fig. 7 Prediction results for random forest models with 5 input features that uses the IP as target. Feature pool: **a** only structural features, **b** structural features and DFT parameters, **c** structural features and pH, **d** structural features, pH and DFT parameters.

Table 3. Results of 6-fold cross-validation

Target	# Features	RMSE structural	RMSE structural + DFT	RMSE structura + pH	RMSE structural + DFT + pH	R^2 structural	R^2 structural + DFT	R^2 structural + pH	R^2 structural + DFT + pH
IE	10	0.22 (±0.03)	0.19 (±0.02)	0.14 (±0.02)	0.14 (±0.02)	−0.47 (±0.33)	−0.12 (±0.18)	0.3 (±0.21)	0.35 (±0.21)
	5	0.21 (±0.03)	0.23 (±0.02)	0.14 (±0.02)	0.14 (±0.02)	−0.46 (±0.38)	−1.07 (±0.67)	0.27 (±0.22)	0.35 (±0.22)
IP	10	0.2 (±0.04)	0.19 (±0.04)	0.2 (±0.04)	0.18 (±0.04)	0.3 (±0.14)	0.35 (±0.14)	0.31 (±0.13)	0.41 (±0.13)
	5	0.21 (±0.04)	0.2 (±0.03)	0.2 (±0.03)	0.18 (±0.03)	0.28 (±0.13)	0.33 (±0.14)	0.35 (±0.11)	0.41 (±0.11)

model. The need for more complicated predictive models with advanced descriptors was clear by categorising molecules based on heteroatom content and the presence of aromatic moieties. Compounds that contain both N and S heteroatoms performed consistently well, however, the performance of compounds with other chemical structures was spread over a large range. Electrochemical information coming from corrosion potential and passive range bears no linear correlation to inhibition power and could be either a predictive descriptor in combination with other features for predicting corrosion resistance or can be an important prediction target as it is a key parameter for localised corrosion. The machine learning model augmented with mechanistic information is key in exploring the complexity of corrosion phenomena, which was highlighted by the predictive power of pH. No linear relationship between bulk pH and inhibitor performance was observed, however, information gained from pH assisted in describing the system better by including information about the environment not necessarily found in computational descriptors, which increased the prediction rate and assisted in outlier analysis of the random forest models.

At this stage rather than designing a final prediction system, we have explored the use of machine learning models to create an active learning loop for more efficient experimental discovery. The obtained experimental parameters can be employed directly as target parameters for training a machine learning model that is predictive of the performance of untested compounds to create a shortlist of promising candidates. Moreover, the experimental investigation yielded additional input features like pH that can be combined with molecular descriptors derived from the molecular structure and atomistic simulations. These input features exhibit great potential to develop augmented quantitative structure–activity relationships as they allow the direct inclusion of information about the underlying mechanisms in the training of the models. The results of this study are expected to support the development of (i) faster inhibitor screening techniques that can capture the same high-resolution electrochemical information on a shorter time-scale, (ii) more complex models that can leverage the link between the physicochemical nature of the inhibitor and its protective performance.

METHODS

Sample preparation

Aluminium alloy 2024 with a T3 temper (AA2024-T3) in the form of 2 mm-thick sheets is purchased (from Salomon's Metalen B.V., the Netherlands) to perform the electrochemical experiments. The chemical composition of the alloy measured by the supplier in

accordance with the ASTM-E1251 standard is provided in Supplementary Table 1.

The sheets were cut with an automatic shearing machine to dimensions of 20 mm x 20 mm samples. The samples were mechanically ground on a rotating plate polisher under a stream of water using Struers waterproof SiC sandpapers with progressively finer grits of 320, 800, 1200, 2000 and 4000. Subsequently, the samples were polished using a fine diamond suspension (Struers DiaDuo-2) with 3 and 1 µm particle sizes. After the polishing procedure, samples were cleaned with isopropanol in an ultrasonic bath (EMAG-EMMI 30HC) for 15 minutes and dried with compressed air. Sample preparation resulted in a mirror-like surface finish.

Inhibitors and electrolytes

The salt solutions without the addition of inhibitors (pH 5.9) were prepared with NaCl powder with Milli-Q pure water (15.0 MΩ cm resistance at 25 °C). For inhibitor-containing solutions, inhibitors in quantities corresponding to 1 mM concentrations were also added during the mixture step. No additional compounds were added to modify the pH and/or increase the solubility of inhibitors. 78 small organic molecules were tested as corrosion inhibitors, resulting in 0.1 M NaCl–1 mM inhibitor electrolytes.

Initial organic molecule choice was based on previous inhibitor screening studies[28,59]. Tested organic molecules had both aromatic/aliphatic moieties of thiol, amino, carboxyl and hydroxyl groups. CAS numbers and common names of the compounds are presented in the Supplementary Dataset. All chemicals were purchased from Sigma-Aldrich, with the exception of sodium chloride (J.T. Baker), 3-amino-5-mercapto-1,2,4-triazole, lithium nitrate, cerium carbonate hydrate (Alfa Aesar), cerium chloride heptahydrate, sodium acetate (Fluka), 2-mercaptobenzoate (Thermo Fisher Scientific), 5-mercapto-1-phenyl-1H-tetrazole (TCI Chemicals) and sodium mercaptobenzothiazole (Apollo Scientific). Almost all inhibitors dissolved fully in 1 mM concentrations, with the exception of thiosalycylic acid, 2-mercaptobenzothiazole, *α*-benzoin oxime, 2,2′-dithiodibenzoic acid, 4-mercaptobenzoic acid, 2-(2-hydroxyphenyl)benzothiazole, quercetin hydrate, berberine chloride hydrate and 2-(2-hydroxyphenyl)benzoxazole. The solutions of these compounds were either murky, resulted in muddy suspensions/emulsions or had visible undissolved particles in the solution. The pH of the resulting solutions was measured with Metrohm 913 pH meter, before and after the electrochemical experiments.

Electrochemical experiments

Electrochemical measurements were conducted at room temperature in open-to-air 0.1 M NaCl solutions, with (or without) the added 1 mM inhibitor candidates. A conventional three-electrode electrochemical cell (flat corrosion cell, Corrtest Instruments, China) with the sample as the working electrode, platinum mesh as the counter electrode, and Ag|AgCl (saturated KCl) as the reference electrode were used to perform the experiments. The designated electrolyte volume was 300 ml and the exposed surface area was 0.785 cm^2 (1 cm diameter circle). Electrochemical measurements were controlled with Biologic VSP-300 multichannel potentiostats through EC-Lab software (version 11.33, Biologic, France).

The electrochemical measurements consisted of three different techniques commonly used in the field of corrosion science: linear polarisation resistance (LPR), electrochemical impedance spectroscopy (EIS) and potentiodynamic polarisation (PDP). The electrochemical investigations were initialised after observing the open circuit potential (OCP) for 10 minutes. LPR was measured over a potential range of ±10 mV with a scan rate of 0.5 mV s^{-1} every 10 minutes for 24 hours. The polarisation resistance (R_p) values were calculated by applying a linear fit to the observed linear region of potential vs. current density plots. EIS measurements were conducted at the 2nd and 24th hour. EIS measurements were conducted by applying a sinusoidal AC perturbation with a peak-to-peak amplitude of 10 mV in the 10 kHz–10 mHz frequency range with 10 frequency point per logarithmic decade with 3 repetitions per frequency point. OCP was observed in between LPR and EIS measurements. After the EIS at the 24th hour, potentiodynamic polarisation curves are recorded in a single sweep with a scan rate of 0.5 mV s^{-1} from −250 mV cathodic to +250 mV anodic potentials with respect to open circuit potential. Corrosion potentials and current densities were calculated with Tafel extrapolation, by obtaining the intersection of tangents from linear parts of anodic and cathodic curves of the log|current density|-potential polarisation curves. Visual summary of electrochemical experiments is presented in Supplementary Fig. 1.

All electrochemical experiments were repeated at least three times per inhibitor to ensure the reproducibility of the experiments.

Molecular descriptor generation, feature selection and evaluation of random forest models

The molecular descriptors based on the structure of the molecules for the input to the random forest (RF) model, e.g. the molecular weight or the number of certain functional groups, have been generated using the open-source chemoinformatic software package RDKit[100]. Additionally, DFT computations have been carried out to determine electronic key properties like frontier orbital energy levels using the commercial software package Turbomole[101] resulting in a pool of 216 molecular descriptors (208 structural, 7 derived from DFT simulations and 1 experimental parameter (the average pH, average of before and after electrochemical measurements). The aim of the recursive feature elimination (RFE) was, to reduce this number to five or ten input features. Furthermore, experimental parameters, especially the average pH, which were obtained from the experiments, were used as additional input to the ML model. To determine the influence of DFT and experimental parameters, the RF has been trained on different sets of input features: on the structural features only, on the structural features complemented by DFT or experimental parameters or both.

Prior to training, RFE, a sparse feature selection approach based on RF, has been carried out to select the most pertinent input features. The purpose is to select *n*-tuples of features that perform well together. Features that have low or no relevance to the modelled property would degrade the model and using too many input features will ultimately lead to overfitting on the training data. Therefore, the five and ten most relevant features in each of the four groups have been determined with RFE and subsequently used as the input to the RF model.

RF is a supervised learning method where the output is obtained by averaging the results of a set of decision trees. The RF model can use both the IE and IP as targets. Examining the data distribution for IE and IP (see Supplementary Figure 6), it can be observed that there is no uniform distribution in either case which may lead to an unintentional bias in the training data. Preprocessing step consisted of the removal of minimally varying and highly correlated features, and scaling the rest. Features with variance lower than 0.1 have been removed with the VarianceThreshold function of scikit-learn. Features with correlations higher than 0.8 to rest of the features are dropped. All features have been scaled using MinMaxScaler of scikit-learn. For the implementation of RF models in this work, the default parameters provided by scikit-learn have been utilised.

To evaluate the performance of the models, the coefficient of determination (R^2) and the root mean squared error (RMSE) have been employed. The first step was to divide the data into a training and test set, with the test set containing ten molecules, or

roughly 17% of the total number of molecules in the dataset. To be more confident in the model's performance, in the next step, a CV approach has been used to assess the model's robustness. For this purpose, the dataset was split into six different folds using the KFold function of sci-kit learn and all folds but one are used for training the models; this fold is held back and used as the test set. Each fold also contained roughly 10% of the total number of molecules in the dataset. In total, the models are trained six times and the average of the errors is calculated to assess their robustness. The results of a leave-one-out CV can be found in the Supplementary Tables 6 and 7.

Unless otherwise stated, the error bars and bracketed values (±e.g.) presented throughout the study represent the standard error.

DATA AVAILABILITY

All data generated and analysed during this study are included in this published article (and its supplementary information files).

Received: 16 July 2023; Accepted: 18 January 2024;
Published online: 21 February 2024

REFERENCES

1. Chyżewski, E. & Evans, U. R. The classification of anodic and cathodic inhibitors. *Trans. Electrochem. Soc.* **76**, 215 (1939).
2. Hey, A., Tansley, S. & Tolle, K. The Fourth Paradigm: Data-intensive Scientific Discovery (Microsoft Research, Redmond, WA, 2009).
3. Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *APL Mater.* **4**, 1–10 (2016).
4. Frankel, G. S. & McCreery, R. L. Inhibition of Al alloy corrosion by chromates. *Electrochem. Soc. Interface* **10**, 34–38 (2001).
5. Kendig, M. W. & Buchheit, R. G. Corrosion inhibition of aluminum and aluminum alloys by soluble chromates, chromate coatings, and chromate-free coatings. *Corrosion* **59**, 379–400 (2003).
6. Ilevbare, G. O. Inhibition of pitting corrosion on aluminum alloy 2024-T3: effect of soluble chromate additions vs chromate conversion coating. *Corrosion* **56**, 227–242 (2000).
7. Gharbi, O., Thomas, S., Smith, C. & Birbilis, N. Chromate replacement: what does the future hold? *npj Mater. Degrad.* **2**, 23–25 (2018).
8. Yasakau, K. A., Zheludkevich, M. L., Lamaka, S. V. & Ferreira, M. G. Mechanism of corrosion inhibition of AA2024 by rare-earth compounds. *J. Phys. Chem. B* **110**, 5515–5528 (2006).
9. Matter, E. A., Kozhukharov, S., Machkova, M. & Kozhukharov, V. Comparison between the inhibition efficiencies of Ce(III) and Ce(IV) ammonium nitrates against corrosion of AA2024 aluminum alloy in solutions of low chloride concentration. *Corros. Sci.* **62**, 22–33 (2012).
10. Kosari, A. et al. Editors' choice-dealloying-driven cerium precipitation on intermetallic particles in aerospace aluminium alloys. *J. Electrochem. Soc.* **168**, 041505 (2021).
11. Markley, T. A., Forsyth, M. & Hughes, A. E. Corrosion protection of AA2024-T3 using rare earth diphenyl phosphates. *Electrochim. Acta* **52**, 4024–4031 (2007).
12. Lopez-Garrity, O. & Frankel, G. S. Corrosion inhibition of aluminum alloy 2024-T3 by sodium molybdate. *J. Electrochem. Soc.* **161**, C95–C106 (2014).
13. Jakab, M. A., Presuel-Moreno, F. & Scully, J. R. Effect of molybdate, cerium, and cobalt ions on the oxygen reduction reaction on AA2024-T3 and selected intermetallics. *J. Electrochem. Soc.* **153**, B244 (2006).
14. Kannan, B., Glover, C. F., McMurray, H. N., Williams, G. & Scully, J. R. Performance of a magnesium-rich primer on pretreated AA2024-T351 in full immersion: a galvanic throwing power investigation using a scanning vibrating electrode technique. *J. Electrochem. Soc.* **165**, C27–C41 (2018).
15. Collazo, A., Nóvoa, X. R. & Pérez, C. The role of Mg^{2+} ions in the corrosion behaviour of AA2024-T3 aluminium alloys immersed in chloride-containing environments. *Electrochim. Acta* **124**, 17–26 (2014).
16. Santucci, R. J. & Scully, J. R. Mechanistic framework for understanding ph-induced electrode potential control of AA2024-T351 by protective Mg-based pigmented coatings. *J. Electrochem. Soc.* **167**, 131514 (2020).
17. Kosari, A. et al. Laterally-resolved formation mechanism of a lithium-based conversion layer at the matrix and intermetallic particles in aerospace aluminium alloys. *Corros. Sci.* **190**, 109651 (2021).
18. Visser, P., Gonzalez-Garcia, Y., Mol, J. M. C. & Terryn, H. Mechanism of passive layer formation on AA2024-T3 from alkaline lithium carbonate solutions in the presence of sodium chloride. *J. Electrochem. Soc.* **165**, C60–C70 (2018).
19. Visser, P., Meeusen, M., Gonzalez-Garcia, Y., Terryn, H. & Mol, J. M. C. Electrochemical evaluation of corrosion inhibiting layers formed in a defect from lithium-leaching organic coatings. *J. Electrochem. Soc.* **164**, C396–C406 (2017).
20. Marinescu, M. Recent advances in the use of benzimidazoles as corrosion inhibitors. *BMC Chem.* **13**, 1–21 (2019).
21. Xhanari, K. et al. Green corrosion inhibitors for aluminium and its alloys: a review. *RSC Adv.* **7**, 27299–27330 (2017).
22. Zheludkevich, M. L., Yasakau, K. A., Poznyak, S. K. & Ferreira, M. G. Triazole and thiazole derivatives as corrosion inhibitors for AA2024 aluminium alloy. *Corros. Sci.* **47**, 3368–3383 (2005).
23. Recloux, I. et al. Stability of benzotriazole-based films against AA2024 aluminium alloy corrosion process in neutral chloride electrolyte. *J. Alloys Compd.* **735**, 2512–2522 (2018).
24. Verma, C., Quraishi, M. A. & Ebenso, E. E. Quinoline and its derivatives as corrosion inhibitors: a review. *Surf. Interfaces* **21**, 100634 (2020).
25. Snihirova, D., Lamaka, S. V., Taheri, P., Mol, J. M. & Montemor, M. F. Comparison of the synergistic effects of inhibitor mixtures tailored for enhanced corrosion protection of bare and coated AA2024-T3. *Surf. Coat. Technol.* **303**, 342–351 (2016).
26. Mohammadi, I., Shahrabi, T., Mahdavian, M. & Izadi, M. Sodium diethyldithiocarbamate as a novel corrosion inhibitor to mitigate corrosion of 2024-T3 aluminum alloy in 3.5 wt% NaCl solution. *J. Mol. Liq.* **307**, 112965 (2020).
27. Prakashaiah, B. G., Vinaya Kumara, D., Anup Pandith, A., Nityananda Shetty, A. & Amitha Rani, B. E. Corrosion inhibition of 2024-T3 aluminum alloy in 3.5% NaCl by thiosemicarbazone derivatives. *Corros. Sci.* **136**, 326–338 (2018).
28. Harvey, T. G. et al. The effect of inhibitor structure on the corrosion of AA2024 and AA7075. *Corros. Sci.* **53**, 2184–2190 (2011).
29. Lamaka, S. V., Zheludkevich, M. L., Yasakau, K. A., Montemor, M. F. & Ferreira, M. G. High effective organic corrosion inhibitors for 2024 aluminium alloy. *Electrochim. Acta* **52**, 7231–7247 (2007).
30. Xhanari, K. & Finšgar, M. Organic corrosion inhibitors for aluminum and its alloys in chloride and alkaline solutions: a review. *Arab. J. Chem.* **12**, 4646–4663 (2019).
31. Popoola, L. T. Organic green corrosion inhibitors (OGCIs): a critical review. *Corros. Rev.* **37**, 71–102 (2019).
32. Zhou, B., Wang, Y. & Zuo, Y. Evolution of the corrosion process of AA 2024-T3 in an alkaline NaCl solution with sodium dodecylbenzenesulfonate and lanthanum chloride inhibitors. *Appl. Surf. Sci.* **357**, 735–744 (2015).
33. Taheri, P. et al. On the importance of time-resolved electrochemical evaluation in corrosion inhibitor-screening studies. *npj Mater. Degrad.* **4**, 1–4 (2020).
34. Meeusen, M. et al. A complementary electrochemical approach for time-resolved evaluation of corrosion inhibitor performance. *J. Electrochem. Soc.* **166**, C3220–C3232 (2019).
35. Visser, P., Terryn, H. & Mol, J. M. C. On the importance of irreversibility of corrosion inhibitors for active coating protection of AA2024-T3. *Corros. Sci.* **140**, 272–285 (2018).
36. White, P. A. et al. Towards materials discovery: assays for screening and study of chemical interactions of novel corrosion inhibitors in solution and coatings. *N. J. Chem.* **44**, 7647–7658 (2020).
37. White, P. A. et al. A new high-throughput method for corrosion testing. *Corros. Sci.* **58**, 327–331 (2012).
38. Taylor, S. & Chambers, B. The discovery of non-chromate corrosion inhibitors for aerospace alloys using high-throughput screening methods. *Corros. Rev.* **25**, 571–590 (2007).
39. Muster, T. H. et al. A rapid screening multi-electrode method for the evaluation of corrosion inhibitors. *Electrochim. Acta* **54**, 3402–3411 (2009).
40. Muster, T. H. et al. A review of high throughput and combinatorial electrochemistry. *Electrochim. Acta* **56**, 9679–9699 (2011).
41. García, S. J. et al. The influence of pH on corrosion inhibitor selection for 2024-T3 aluminium alloy assessed by high-throughput multielectrode and potentiodynamic testing. *Electrochim. Acta* **55**, 2457–2465 (2010).
42. Chambers, B. D. & Taylor, S. R. High-throughput assessment of inhibitor synergies on aluminum alloy 2024-T3 through measurement of surface copper enrichment. *Corrosion* **63**, 268–276 (2007).
43. Lamaka, S. V. et al. Comprehensive screening of Mg corrosion inhibitors. *Corros. Sci.* **128**, 224–240 (2017).
44. Feiler, C. et al. In silico screening of modulators of magnesium dissolution. *Corros. Sci.* **163**, 108245 (2020).
45. Zabula, A. V. et al. Screening of molecular lanthanide corrosion inhibitors by a high-throughput method. *Corros. Sci.* **165**, 108377 (2020).
46. White, P. A. et al. High-throughput channel arrays for inhibitor testing: Proof of concept for AA2024-T3. *Corros. Sci.* **51**, 2279–2290 (2009).

47. Visser, P. et al. Li leaching from Li carbonate-primer: transport pathway development from the scribe edge of a primer/topcoat system. *Prog. Org. Coat.* **158**, 106284 (2021).
48. Moraes, C. V., Santucci, R. J., Scully, J. R. & Kelly, R. G. Finite element modeling of chemical and electrochemical protection mechanisms offered by mg-based organic coatings to AA2024-T351. *J. Electrochem. Soc.* **168**, 051505 (2021).
49. Binggeli, M., Shen, T.-H. & Tileli, V. Simulating current distribution of oxygen evolution reaction in microcells using finite element method. *J. Electrochem. Soc.* **168**, 106508 (2021).
50. Obot, I. B., Macdonald, D. D. & Gasem, Z. M. Density functional theory (DFT) as a powerful tool for designing new organic corrosion inhibitors: Part 1: An overview. *Corros. Sci.* **99**, 1–30 (2015).
51. Kokalj, A. & Costa, D. *Molecular Modeling of Corrosion Inhibitors* (Elsevier, 2018).
52. Kokalj, A. et al. Simplistic correlations between molecular electronic properties and inhibition efficiencies: do they really exist? *Corros. Sci.* **179**, 108856 (2021).
53. Kokalj, A. On the alleged importance of the molecular electron-donating ability and the HOMO–LUMO gap in corrosion inhibition studies. *Corros. Sci.* **180**, 109016 (2021).
54. Luo, X. et al. Computational simulation and efficient evaluation on corrosion inhibitors for electrochemical etching on aluminum foil. *Corros. Sci.* **187**, 109492 (2021).
55. Costa, D., Ribeiro, T., Cornette, P. & Marcus, P. DFT modeling of corrosion inhibition by organic molecules: carboxylates as inhibitors of aluminum corrosion. *J. Phys. Chem. C* **120**, 28607–28616 (2016).
56. Milošev, I. et al. Electrochemical, surface-analytical, and computational DFT study of alkaline etched aluminum modified by carboxylic acids for corrosion protection and hydrophobicity. *J. Electrochem. Soc.* **166**, C3131–C3146 (2019).
57. Milošev, I. et al. Editors' choice—the effect of anchor group and alkyl backbone chain on performance of organic compounds as corrosion inhibitors for aluminum investigated using an integrative experimental-modeling approach. *J. Electrochem. Soc.* **167**, 061509 (2020).
58. Milošev, I. et al. The effects of perfluoroalkyl and alkyl backbone chains, spacers, and anchor groups on the performance of organic compounds as corrosion inhibitors for aluminum investigated using an integrative experimental-modeling approach. *J. Electrochem. Soc.* **168**, 071506 (2021).
59. Winkler, D. A. et al. Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors. *Corros. Sci.* **106**, 229–235 (2016).
60. Würger, T. et al. Data science based mg corrosion engineering. *Front. Mater.* **6**, 1–9 (2019).
61. Würger, T. et al. Exploring structure–property relationships in magnesium dissolution modulators. *npj Mater. Degrad.* **5**, 1–10 (2021).
62. Schiessler, E. J. et al. Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning models. *npj Comput. Mater.* **7**, 39–41 (2021).
63. Galvão, T. L., Novell-Leruth, G., Kuznetsova, A., Tedim, J. & Gomes, J. R. Elucidating structure-property relationships in aluminum alloy corrosion inhibitors by machine learning. *J. Phys. Chem. C* **124**, 5624–5635 (2020).
64. Coelho, L. B. et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. *npj Mater. Degrad.* **6**, 1–16 (2022).
65. Galvão, T. L. P. et al. CORDATA : an open data management web application to select corrosion inhibitors. 4–7 (2022).
66. Andreatta, F. & Fedrizzi, L. Corrosion Inhibitors. In *Active Protective Coatings*, 233 edn (eds Hughes, A. E., Mol, J. M., Zheludkevich, M. L. & Buchheit, R. G.) Ch. 4, 59–84 (Springer, Netherlands, Dordrecht, 2016).
67. Barsoukov, E. & Macdonald, J. R. (eds) *Impedance Spectroscopy: Theory, Experiment, and Applications* (John Wiley & Sons, 2005), 2nd edn.
68. Scully, J. R. Polarization resistance method for determination of instantaneous corrosion rates. *Corrosion* **56**, 199–217 (2000).
69. Pourbaix, M. *Atlas of Electrochemical Equilibria in Aqueous Solutions* (NACE, 1966).
70. Thabtah, F., Hammoud, S., Kamalov, F. & Gonsalves, A. Data imbalance in classification: experimental evaluation. *Inf. Sci.* **513**, 429–441 (2020).
71. Benesty, J., Chen, J., Huang, Y. & Cohen, I. Pearson correlation coefficient. In *Springer Topics in Signal Processing*, Eds J. Benesty & W. Kellermann, Springer, Vol. **2**, 1–4 (2009).
72. Kelly, R. G., Scully, J. R., Shoesmith, D. & Buchheit, R. G. *Electrochemical Techniques in Corrosion Science and Engineering* (CRC Press, New York, 2002).
73. Verma, C., Verma, D. K., Ebenso, E. E. & Quraishi, M. A. Sulfur and phosphorus heteroatom-containing compounds as corrosion inhibitors: an overview. *Heteroatom Chem.* **29**, 1–20 (2018).
74. Rani, B. E. & Basu, B. B. J. Green inhibitors for corrosion protection of metals and alloys: an overview. *Int. J. Corros.* **2012** (2012).
75. Verma, C., Ebenso, E. E. & Quraishi, M. A. Corrosion inhibitors for ferrous and non-ferrous metals and alloys in ionic sodium chloride solutions: a review. *J. Mol. Liq.* **248**, 927–942 (2017).
76. Neupane, S. et al. Study of mercaptobenzimidazoles as inhibitors for copper corrosion: down to the molecular scale. *J. Electrochem. Soc.* **168**, 051504 (2021).
77. Kozlica, D. K., Kokalj, A. & Milosev, I. Synergistic effect of 2-mercaptobenzimidazole and octylphosphonic acid as corrosion inhibitors for copper and aluminium-an electrochemical, XPS, FTIR and DFT study. *Corros. Sci.* **182**, 109082 (2021).
78. Wu, X., Wiame, F., Maurice, V. & Marcus, P. Molecular scale insights into interaction mechanisms between organic inhibitor film and copper. *npj Mater. Degrad.* **5**, 1–8 (2021).
79. Özçelik, R., van Tilborg, D., Jiménez-Luna, J. & Grisoni, F. Structure-based drug discovery with deep learning. *ChemBioChem* 202200776. http://arxiv.org/abs/2212.13295, https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cbic.202200776 (2023).
80. Harren, T., Matter, H., Hessler, G., Rarey, M. & Grebner, C. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *J. Chem. Inf. Model.* **62**, 447–462 (2022).
81. Miyao, T., Kaneko, H. & Funatsu, K. Inverse QSPR/QSAR analysis for chemical structure generation (from *y* to *x*). *J. Chem. Inf. Model.* **56**, 286–299 (2016).
82. Lo, Y. C., Senese, S., Damoiseaux, R. & Torres, J. Z. 3D Chemical similarity networks for structure-based target prediction and scaffold hopping. *ACS Chem. Biol.* **11**, 2244–2253 (2016).
83. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion Drug Discov.* **16**, 949–959 (2021).
84. Amar, Y., Schweidtmann, A. M., Deutsch, P., Cao, L. & Lapkin, A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chem. Sci.* **10**, 6697–6706 (2019).
85. Stanley, M. et al. FS-Mol: a few-shot learning dataset of molecules. In 35th Conference on Neural Information Processing Systems *NeurIPS* (2021).
86. Obrezanova, O., Csányi, G., Gola, J. M. & Segall, M. D. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **47**, 1847–1857 (2007).
87. Moret, M. et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* **14**, 1–12 (2023).
88. Sutojo, T. et al. A machine learning approach for corrosion small datasets. *npj Mater. Degrad.* **7**, 1–10 (2023).
89. Edwards, D. A. Steric hindrance effects on surface reactions: applications to BIAcore. *J. Math. Biol.* **55**, 517–539 (2007).
90. Yun, L. et al. Evaluation and optimization of corrosion inhibitor system. *IOP Conf. Ser. Mater. Sci. Eng.* **729** (2020).
91. Finšgar, M., Lesar, A., Kokalj, A. & Milošev, I. A comparative electrochemical and quantum chemical calculation study of BTAH and BTAOH as copper corrosion inhibitors in near neutral chloride solution. *Electrochim. Acta* **53**, 8287–8297 (2008).
92. Kosari, A. et al. Dealloying-driven local corrosion by intermetallic constituent particles and dispersoids in aerospace aluminium alloys. *Corros. Sci.* **177**, 108947 (2020).
93. Kosari, A. et al. In-situ nanoscopic observations of dealloying-driven local corrosion from surface initiation to in-depth propagation. *Corros. Sci.* **177**, 108912 (2020).
94. Frankel, G. S. Fundamentals of corrosion kinetics. In *Active Protective Coatings: New-Generation Coatings for Metals* 17–32 (2016).
95. Erlebacher, J. An atomistic description of dealloying. *J. Electrochem. Soc.* **151**, C614 (2004).
96. Hughes, A. E., Parvizi, R. & Forsyth, M. Microstructure and corrosion of AA2024. *Corros. Rev.* **33**, 1–30 (2015).
97. Kolics, A., Besing, A. S., Baradlai, P., Haasch, R. & Wieckowski, A. Effect of ph on thickness and ion content of the oxide film on aluminum in NaCl media. *J. Electrochem. Soc.* **148**, B251 (2001).
98. Lamaka, S. V. et al. Local pH and its evolution near Mg alloy surfaces exposed to simulated body fluids. *Adv. Mater. Interfaces* **5**, 1800169 (2018).
99. Winkler, D. A. Predicting the performance of organic corrosion inhibitors. *Metals* **7**, 1–8 (2017).
100. *RDKit: Open-source Cheminformatics,* G. Landrum, Github and SourceForge (accessed 4 May 2023). https://www.rdkit.org.
101. TURBOMOLE, A development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

ACKNOWLEDGEMENTS

This work is a part of the VIPCOAT project (Virtual Open Innovation Platform for Active Protective Coatings Guided by Modelling and Optimisation) funded by the Horizon 2020 research and innovation programme of the European Union by grant agreement no. 952903.

AUTHOR CONTRIBUTIONS

Can Özkan: Conceptualisation, methodology, validation, formal analysis, investigation, data curation, writing—original draft, visualisation. Lisa Sahlmann: Methodology, software, formal analysis, writing—original draft, visualisation. Christian Feiler: Methodology, formal analysis, writing—original draft, supervision. Sviatlana Lamaka: Writing—review & editing, supervision. Mikhail Zheludkevich: Writing—review & editing, supervision. Parth Sewlikar: Investigation, writing—review & editing. Agnieszka Kooijman: Investigation, writing—review & editing. Peyman Taheri: Writing—review & editing, supervision. Arjan Mol: Conceptualisation, resources, writing—review & editing, supervision.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41529-024-00435-z.

Correspondence and requests for materials should be addressed to Can Özkan.