npj | materials degradation

Article

Published in partnership with CSCP and USTB

https://doi.org/10.1038/s41529-025-00713-4

Gaining scientific understanding with small data machine learning: explainable molecule representations and their consensus

Can Özkan[1] ✉, Lisa Sahlmann[2], Tim Würger[2], Christian Feiler[2], Sviatlana Lamaka[2], Mikhail Zheludkevich[2], Peyman Taheri[1] & Arjan Mol[1]

Despite the remarkable success of machine learning in materials science, challenges persist in gaining mechanistic insights, especially in low-data regimes where dataset sizes limit the precise applicability of machine learning. The prevailing reliance on high-confidence predictions from the models often leaves the underlying decision-making mechanisms opaque, limiting scientific understanding. This study presents an alternative approach that emphasizes understanding the model decision-making process over individual predictions, enabling the extraction of scientifically meaningful insights from small datasets. Focusing on 107 small organic molecules and their corrosion inhibition properties as a case study, we systematically evaluate 29 molecular featurization methods and 9 target representations, generating over 12 thousand model configurations to identify robust feature-target pairings. We reveal common trends by reverse engineering the best-performing models based on featurization methods of physicochemical descriptors, hashed fingerprints, and structural keys, which we integrate with domain knowledge to create a molecular substructure template for candidate molecules. Using this template, we filter a toxicity database to identify non-toxic corrosion inhibitors, aiming to replace the de facto but hazardous corrosion inhibitor hexavalent chromium. The resulting candidate's efficacy is validated through electrochemical testing, illustrating the feasibility of achieving mechanistic insights from statistical models in data-scarce environments.

In Douglas Adams' Hitchhikers Guide to the Galaxy, an alien race seeks the "Answer to the Ultimate Question of Life, the Universe, and Everything". A planet-sized computer, Deep Thought, is tasked with calculating the answer. After seven and a half million years, it finally reveals the answer: 42. This result baffles the programmers, as they realize they don't actually know what the "Ultimate Question" is. Deep Thought explains that without understanding the true nature of the question, the answer has no real meaning.

We face a similar challenge today in machine learning applications to materials science. On the one hand, machine learning models have been widely successful in materials sciences for autonomous experimentation, materials property interpolations, structure optimization, and chemical space exploration[1–6]. Recent advances, such as creating continuous material property representations through generative autoencoder approaches[7,8], and utilizing the inherent graph structure of molecules through graph neural network architectures[9,10], allowed researchers to investigate previously unexplored chemical spaces. On the other hand, training such methods is resource-intensive, requiring substantial computational power and large datasets—more than ten thousand, in some cases tens of millions of data points for predictive accuracy. Data scarcity and expensive target generation prevent cutting-edge architectures to be used with experimental datasets, and limit the use of such models to simplified DFT simulation predictions. Yet, even with high benchmark scores, a recent study showed that model predictions may not generalize well to new materials spaces[11]. Most importantly, generalized or not, the scientific insights within these models remain largely opaque, meaning that even when models make accurate predictions, the mechanisms behind these "answers" are not transferred to scientists[12,13]. Like Deep Thought's

[1]Department of Materials Science and Engineering, Delft University of Technology, Delft, the Netherlands. [2]Institute of Surface Science, Helmholtz-Zentrum Hereon, Geesthacht, Germany. ✉e-mail: c.ozkan@tudelft.nl

answer, these predictions are limited in their impact without a deeper understanding of the "questions" they are addressing.

The materials science community has already made significant progress in building the foundation for explainable and interpretable models[14–18]. However, most of the published literature on explainability focuses on bigger models, despite most materials discovery problems happening in low- to no-data regime. For this reason, with this paper, we intend to go in the opposite direction of the present mainstream deep-learning focus and ask: Is it possible to gain scientific insights from predictive models with lower prediction metrics based on small datasets? The answer, we believe, is yes. We see that representation is key; with the right representation and reverse engineering machine learning models based on different data representations, we can identify common trends—which combined with domain expertise, can allow the scientist to see new trends that were previously unattainable.

Here, we present an unorthodox framework to transform statistical models into scientific insight. Instead of relying on low-confidence individual predictions of models trained with scarce data, we train and then reverse engineer multiple models at once to understand their decision-making mechanisms. We combine the resulting insights with the scientific intuition of the domain expert. Our work is performed on small organic molecules and their corrosion inhibition properties, but it can principally be applied to any materials discovery task.

Although approaches to interpret machine learning models, such as permutation feature importance[19], local interpretable model agnostic explanations[20], and Shapley additive explanations[21] have been widely used in other scientific fields, specifically in the domain of corrosion and inhibition prediction, there have been limited works[22,23]. Compared to neighboring scientific fields more often than not two points plagued the corrosion inhibition researchers (i) the features most important in predictive performance were identified, but no scientific reasoning were built on top of the models to reason why the model predictions were improving when they were present in the prediction process due to the complex multi-scale nature of corrosion phenomena, and (ii) the limited size of the datasets (often around 20 samples or so) caused eager but early conclusions to be drawn, such as the debated correlations between quantum chemical descriptors and corrosion inhibition performance[23,24]. We aimed to instill more trustworthy conclusions to be drawn from predictive models by coming up with a method of reasoning of how different model representations can be aligned to solve the same problem, hypothesize for potential influences of the features, and validate our hypothesis experimentally.

For gaining scientific insights from statistical investigations, achieving the best representation is important. Here we converge to the best representation for our dataset by systematically analyzing 29 widespread open-access methods that convert molecules into a set of features (hereinafter referred to as "featurization"), and 9 different target representations of experimentally acquired electrochemical data on around 100 molecules (for details see our previous work[25]), to be used as model targets. After settling on the optimum description by looking at different feature-target combinations (which results in more than 12,000 model configurations), we can use complementary descriptions with the lowest root mean squared error (RMSE) together to capture common trends existing in all. We use such trends to come up with a searchable template molecule that can be used to filter existing larger databases. In our case, we use a toxicity database, as our goal is to find a non-toxic molecule with the potential to replace the currently in-use domineering hexavalent chromium: a highly inhibitive but deadly corrosion inhibitor prohibited by the EU REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) regulation[26]. We validate the insight gained from this approach by performing electrochemical measurements of the recommended molecule, showing that gaining mechanistic insight with statistical models is possible for small datasets. The result of this work is expected not only to assist in developing green and sustainable alternative inhibition approaches to corrosion, which eats away 3.1% of the global GDP[27], but also expected to serve as a guiding framework for other data-scarce materials discovery problems.

Results and discussion

Describing corrosion inhibition

The description of the problem is key for the machine learning models to extract all possible information from the signal present in the data. This involves not only selecting the appropriate set of features (the set of numbers used for prediction) but also choosing the correct form for both the features and targets (the set of numbers to predict). This is not only a factor to consider when improving the prediction performance of models. It becomes even more important when models are used for gaining mechanistic insights similar to various spectroscopies, as demonstrated in this paper.

To reveal the best description of corrosion inhibition, we have set up a large span of target and feature representations: 9 targets, 29 featurization methods, 3 feature selection approaches (include all, recursive feature elimination (RFE), SHAP-based), combined with 4 different feature scaling methods (no-scaling, minmax, standard, power) and 4 different regression model architectures (random forest (RF), XGBoost (XGB), support vector machine (SVR), k-nearest neighbors (KNN)). This search space of the optimal description consisted of more than 12,000 configurations. The models based on these configurations were trained with 95 small organic molecules with 10-fold cross validation, and validated with a left-out validation set of 12 other molecules. Target values were obtained from time-resolved electrochemical experiments of electrochemical impedance spectroscopy (EIS) and linear polarization resistance (LPR), discussed in more detail in our previous publication[25]. After training, the best models are identified by comparing cross-validation root mean squared (CV-RMSE) values (all converted into the scale of inhibition power IP). Afterwards, the models' hyperparameters are optimized with Bayesian optimization, and the left-out validation set was used to check the prediction performance of the models. The ranking of the predictive performances of the models can be found in the Supplementary Figs. 1–2.

Figure 1 presents the ranking of different featurization methods for the best 4 targets, and their mean. In this case, "best" means the ranking of models with the lowest CV-RMSE error for a given representation. The inset shows the CV-RMSE performance distribution of models with different featurizations, feature selection methods, feature scaling, and model architectures; pooled for different targets. For the best target (IE_EIS24h), a similar pooling for all other configurations but the featurization methods resulted in the performance distribution of different featurization schemes unfolded on the right, with featurization schemes corresponding to the same labels as featurization ranking for different targets on the left. This results in featurization methods ordered from best to worst for target IE_EIS24h. Lowercase labels correspond to hashed fingerprints or structural keys (also highlighted with _fp suffix), capitalized labels correspond to physicochemical descriptors.

Targets. The 9 targets are the combination of data coming from 3 different electrochemical experimental methods, denoted with respective suffixes (Bode modulus at 10^{-2} Hz measured at 24th hour through EIS, _EIS24h; linear polarization resistance measured at 24th hour, _24 h; linear polarization resistances averaged through the first 24 h, _avg), represented in 3 different forms (raw electrochemical data, R_p; inhibition efficiency, IE; inhibition power, IP).

Compared to the rest of the factors, target representation by far had the most impact on the predictive performance of the models. Looking at the CV-RMSE distributions for the different targets, the best model for the best target IE_EIS24h resulted in a CV-RMSE of 2.73, whereas the best model for the worst target Rp_avg resulted in a CV-RMSE of 6.87, an increase of 152%.

Models based on time-averaged experiments (_avg) performed worse than others, indicating that the prediction of time-dependent phenomena might be more difficult than the prediction of a stabilized reaction after a given time, in this case, after 24 h. R_p models also showed poorer prediction performance compared to IE and IP models.

It was observed that although the best models were in the form of IE, the prediction performance from IP models was more consistent. IE_avg and IE_24h had very large distributions of predictive performance. It is

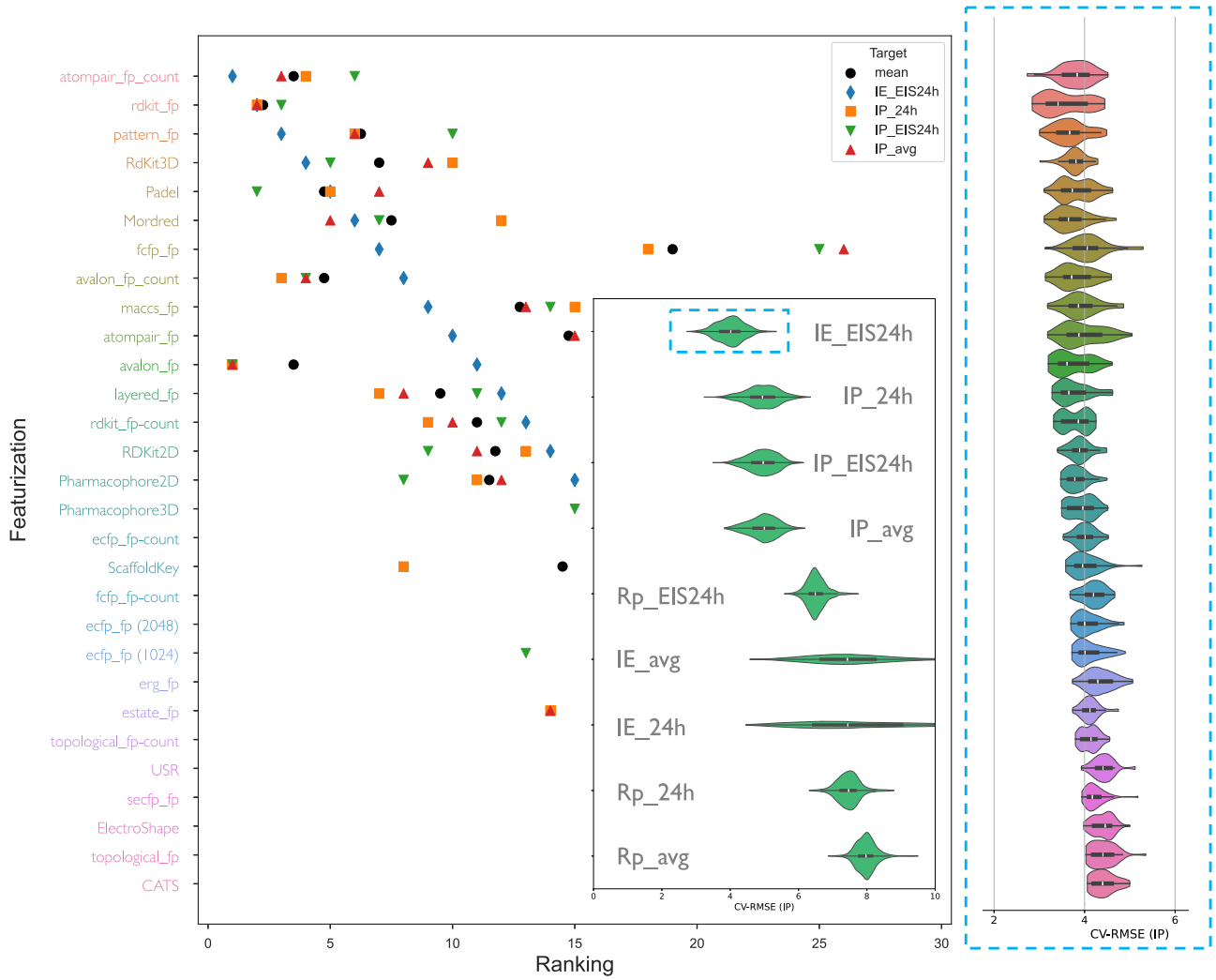


Fig. 1 | Ranking of featurization methods based on prediction performance for the top four targets, and their mean. The *y*-axis lists featurization methods, ordered from best to worst for the target IE_EIS24h, which resulted in models with the lowest cross-validation root mean square error (CV-RMSE). Physicochemical descriptors are capitalized, whereas structural keys/fingerprints are denoted with suffix _fp in case of absence/presence based encoding, and _fp_count for count based encoding. The inset displays CV-RMSE distributions for models across all targets, incorporating variations in featurization methods, feature selection approaches, feature scaling, and model architectures. For the target IE_EIS24h, the distribution of prediction performance for each featurization method is shown on the right, aligned with the corresponding rankings on the left.

hard to say with certainty whether consistent behavior from IP models is dataset-independent—if that is the case, this would mean that the logarithmic form of IP ($c\log_{10}(R_p)$), in comparison to the hyperbolic form of IE ($c\,R_p^{-1}$), stabilizes the model prediction performance, which would make sense for finding edge cases such as good corrosion inhibitors. R_p models had a lesser distribution, but their performances were consistently worse. Worst IP and IE_EIS24h models were almost always better than the best R_p models. This underlines the importance of target representation—it seems that normalization offered by IE and IP transformations allows models to capture corrosion inhibition in a more accurate manner.

Features. The 29 different featurization methods create a wide span of features in a different manner, consisting of 0D (bulk properties and physicochemical descriptors that contain no information about molecule geometry or atom connectivity, e.g., molecular weight, logP octanol-water partition coefficient, contained atom presence and counts), 1D (representations that include information on bonding or bonding fragments, e.g., presence/absence of molecular fragments, hydrogen bond donor or acceptor counts, number of rings, number of functional groups), 2D (molecule graph invariant properties, e.g., topological polar surface area, autocorrelation descriptors), and 3D (topographical molecule shape information, e.g., geometrical, three-dimensional distances and connectivities) descriptors. The featurization methods can broadly be split into three different categories: physicochemical descriptors (e.g., `PaDEL`[28]), structural keys (e.g., `MACCS`[29]), and hashed fingerprints (e.g., `atompair-count`[30]).

Physicochemical calculators generate information about the physical and chemical properties of the whole molecule, such as the surface area occupied by polar atoms and their attached hydrogens, the number of electronegative atoms that can act as hydrogen bond donor/acceptor, the molecule Van der Waals radii surface area of its atoms, or even more obscure and derivative properties such as the topological Balaban index that measures the branching and connectivity of a molecular graph, among many others. The combination of these types of features holds promise in highlighting the properties most relevant for the target molecule behavior we are interested in.

Structural keys encode the molecule structure into a binary bit value (0 or 1) where each bit corresponds to a *pre-defined* structural feature, such as the presence/absence of a benzene ring. If the molecule has the pre-defined

feature, the bit position corresponding to this feature is set to 1; otherwise, it is set to 0. It is important to realize that structural keys cannot encode structural features not pre-defined in their fragment library.

Hashed fingerprints solve this problem by not requiring a pre-defined fragment library, where all possible molecular fragments smaller than the specified size are converted into numeric values using various algorithms. A data of arbitrary size can then be converted into a fixed-size vector using a hash function. The size of this vector is often chosen to be a power of two, default option used being 1024 or 2048. The values of such a vector correspond to the absence/presence of particular molecular fragments, which are denoted as "bits".

One such molecular fragment generation approach would be path-based fingerprints (e.g., Daylight fingerprint based `rdkit_fp`[31]), where branching paths in the molecular graph are analyzed for a given length and hashed in a fixed vector. A different path-based approach would be atom pairs[30], where pairs of atoms with the shortest path connecting them would form substructures to be hashed. Circular fingerprints offer another option, where the circular environments of each atom up to a given radius are used to construct molecular fragments (e.g., extended-connectivity `ECFP`, functional-class `FCFP` fingerprints[32]). Such fingerprints can be encoded in binary to indicate the presence/absence of a given molecular fragment, or can specify the number of occurrences of that molecular fragment by taking integer values for the count-based fingerprinting approach. The flexibility offered by hashing might also cause problems in interpretability, however, since a molecular database may contain a very large amount of molecular fragments, and hashing them into a fixed range can result in "bit collisions", where different molecule fragments would be converted into the same hashed bit value.

For the best target IE_EIS24h, the top three featurization methods were all based on hashed fingerprints: `atompair_fp-count`, `rdkit_fp` and `pattern_fp`. Looking at the mean ranking of the best four targets, `atompair_fp-count`, `rdkit_fp`, `pattern_fp`, `avalon_fp`, `avalon_fp-count`, `layered_fp` hashed fingerprints, and `PaDEL`, `RdKit3D`, `Mordred` physicochemical descriptors resulted in average rankings of less than 10, on average producing more predictive models than other featurization approaches. Compared to alternatives, `ECFP` and `pharmacophore` featurization methods commonly used in many drug discovery problems were inferior in describing corrosion inhibition.

The distribution of the ranking for different targets shows that the choice of the featurization method is heavily dependent on the target. For all IP target representations, however, it was remarkable that avalon_fp consistently resulted in models with the lowest CV-RMSE values. The trends of using fingerprints based on presence/absence compared to counts were also dependent on the featurization method: `atompair` performed better for counts, but `rdkit`, `fcfp` and `avalon` fingerprints performed better in presence/absence binary form. Addition of 3D descriptors to the `RDKit2D` improved ranking consistently, highlighting the importance of 3D molecule effects for corrosion inhibition. Whereas for `pharmacophore` descriptors, ranking of 2D seems better, but both are quite similar in CV-RMSE values, indicating models do not take advantage of additional 3D descriptors offered by this featurization method. Given that `pharmacophore` descriptors were created to work with molecule interactions with a specific biological target, such as a protein or an enzyme, it is normal that the important 3D features do not directly transfer to other problem domains.

Looking at the best target (IE_EIS24h) CV-RMSE distributions for the featurization methods, the best model for the best featurization method, atompair_fp-count, resulted in a CV-RMSE of 2.73, whereas the best model for the worst featurization method resulted in a CV-RMSE of 4.06, an increase of 49%. The distribution for every featurization method was a result of scaling, feature selection, and model architecture.

There was no one best method for choosing any of these details for model configurations, as they all result in similar distributions for CV-RMSE (see Supplementary Figs. 3–5). However, by examining the ranking of all models based on CV-RMSE (see accompanying file), we can qualitatively identify trends among the top-performing models.

Out of the ten models with the lowest CV-RMSE, all had IE_EIS24h as a target. For featurization methods, two were based on `atompair_fp-count`, six on `rdkit_fp`, one was `pattern_fp` and one on `RdKit3D` descriptors. It is interesting to note that only one model was based on physicochemical descriptors, and the rest were based on hashed fingerprints. Feature scaling showed a mix of methods, with the key takeaway being that any scaling is beneficial compared to none: only one model did not use scaling (which was based on RF architecture, which is a scale-independent model), while the others were evenly split, with three models each using minmax, power, and standard scaling.

For feature selection, nine out of ten models used RFE, suggesting it may be a better choice for standardized use despite the drawback of requiring manual selection of the number of features beforehand. The rest of the analysis in this paper used RFE-based feature selection to refine the feature set, ultimately selecting down to the top 10 features of every configuration.

For model architectures, seven models used SVR, two RF, and one XGB. This indicates that the SVR architecture's robustness to outliers and noisy data may be particularly valuable when working with real-world experimental data.

Gaining mechanistic insight through algorithmic feature selection

Having established an optimal model description, we now shift focus to our primary objective: leveraging this refined description to uncover novel mechanistic insights. The premise is that the features that make a model more predictive are also likely to be those most relevant to the underlying physicochemical mechanisms of corrosion inhibition. This makes feature selection methods in machine learning not only a routine for improving model prediction performance, but also a tool for extracting scientific insight hidden in the data statistics.

After identifying key features through algorithmic selection, it is essential to develop an intuition about their relevance to the system. Our goal was to identify whether and how the algorithmically chosen features can be used as a tool to gain mechanistic insight about corrosion inhibition. With that goal in mind, we selected one featurization method from each category for further experimentation: `PaDEL` for physicochemical descriptors, `maccs_fp` for predefined structural keys, and `atompair_fp-count` for hashed fingerprints. The selection was based on the highest predictive performance of each category.

The choice of `MACCS` (Molecular ACCess System) over a potentially more predictive featurization method is based on two key reasons. First, given the narrow range of IE_EIS24h CV-RMSE distributions, explainability takes precedence over pure predictive performance. While accurately predicting the behavior of individual molecules may be challenging for such small-scale models, the primary objective is to predict general mechanistic trends based on the selected features, a task that is comparatively more feasible. `MACCS` is ideal for this, as due to its predefined nature, it is extremely clear what each feature corresponds to. Second, not every featurization method allows explainability in the first place. In our preliminary experiments with visualizing `rdkit_fp` features, we have observed a considerable amount of bit collision—where different molecule substructures are mapped to the same feature column bit. This not only most likely reduces predictive performance but also complicates explainability, as it becomes unclear which substructure is driving the prediction. Other fingerprints, such as Avalon, does not allow visualization in the first place, preventing explainability.

The next sections first analyze the interpretation of hashed `atompair_fp-count` fingerprints through bit visualization. We show that by making artificial changes in the prediction queries, we can gauge the response of a given feature. Since visualization into substructures is not possible for physicochemical descriptors and other featurization methods, in the following section, we show how to use Bayesian optimization as a tool for understanding the model decision-making process with `PaDEL` featurization. Finally, we use SHAP analysis for deciphering feature influence

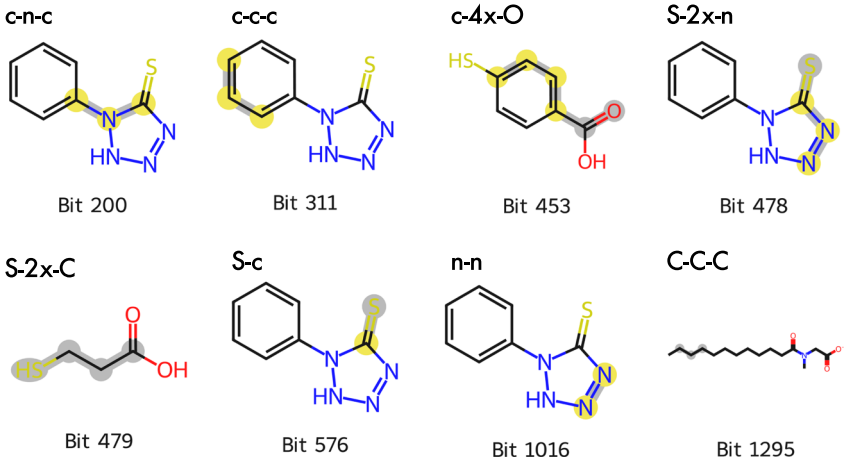


Fig. 2 | Example molecules that contain key features identified by the feature selection algorithm, visualized as "bits". The SMILES-like strings corresponding to the bit are presented in the top left, and the corresponding structures are highlighted in yellow and gray on the molecules. Aliphatic atoms are highlighted in gray and in uppercase letters, aromatic atoms in yellow and in lowercase letters.

for the best models based on `atompair-count`, `PaDEL`, and `MACCS`, and we demonstrate how we can use models based on different featurization methods to gain a united mechanistic insight.

Visualizing algorithmically selected fingerprints for finding the corrosion inhibition structural building blocks

Figure 2 demonstrates the algorithmically selected features from the best `atompair-count` featurization visualized as corresponding molecular substructures, here denoted as 'bits'. The substructures are recorded at the time the fingerprints are generated. Later, these substructures are used to map the features back to the corresponding parts of each molecule. As mentioned before, `atompair` featurization describes the molecular substructures as two atoms and the number of atoms between them. For example, the bit corresponding to feature 478 is a sulfur and nitrogen with two atoms in between, and can be written in a SMILES-like format as S-(2x)-n, where x corresponds to any atom, and uppercase/lowercase denotes aliphaticity/aromaticity. In the same manner, bit 479 would be S-(2X)-C, bit 576 S=C, and so on.

Based on the selected features we can see that the model "thinks" that substructures involving triplets of aromatic (bit 200, bit 311) and aliphatic carbons (bit 1295), nitrogen-nitrogen couples (bit 1016), sulfur directly attached to carbon (bit 576), sulfur attached to nitrogen with two atoms in between (bit 478), sulfur attached to carbon with two atoms in between (bit 479), and aliphatic carbon attached to aromatic carbon with three atoms in between (bit 453) are important in predicting corrosion inhibition. Feature selection identifies these substructures as corrosion inhibition-critical substructures.

These substructures align with the mainstream literature conclusions, which highlight that sulfur and nitrogen atoms often serve as anchoring points to the surface. Additionally, aromatic groups may contribute not only through steric effects that help repel detrimental chloride ions but also through their electron-donating or -withdrawing properties, which can influence direct interactions with the metallic surface or modulate electron density redistribution within attached functional groups. Meanwhile, long aliphatic chains further enhance steric hindrance, both factors being critical for effective corrosion inhibition[23]. Especially the fact that sulfur, critical for the inhibition of copper intermetallics of AA2024-T3[33,34], was identified as important solely by the model, with no prior domain expertise or previous scientific insight, shows that models can capture physicochemical insights. This shows promise in reverse engineering statistics into mechanisms, and for that, understanding the model decision-making process is key.

The visualized molecule substructures already give mechanistic tips, but to understand how every feature contributes to the model in a detailed manner, and to explore what would've happened if only that particular feature had a different value, we have produced counterfactual predictions. To form counterfactuals, we have kept every other feature constant while changing only the analyzed feature to its maximum or minimum value found in the dataset, and then examined how the predictions of the model changed.

Figure 3 presents examples from the counterfactual predictions for top-performing molecules for the selected features and their corresponding visualized bits. For a counterfactual prediction, the feature to be analyzed is modified to its maximum and minimum value found in the dataset, while keeping all the other features at their original values. In this way, the effect of every feature on a set of given molecules can be analyzed independently from other features.

Here, we present the influence of feature 478 vs. 479, and feature 576 vs. 1016, as they capture interesting interpretable trends. Feature 576 is directly related to the surface bonding opportunity offered by the sulfur atom, where the predicted efficiency increases with an increase in the number of thione bonds. The only cases where the maximum does not correspond to an increase are the two derivatives of 1,2,4 triazoles. This makes sense structurally, as the maximum sulfur amount found in the dataset is four, which in the case of the smaller five-ring structures, might hinder bonding instead of supporting. Therefore, depending on the ring size, excess sulfur not contributing to bonding might not be beneficial for inhibition.

Feature 478 and 479 are also connected to the sulfur behavior. Feature 478 and 479 correspond to very similar substructure bits with the only difference being the end atom: bit 478 is a sulfur atom connected to a nitrogen atom with two atoms in between (S-2x-n), bit 479 is a sulfur atom connected to a carbon atom with two atoms in between (S-2x-c). Despite their similar structures, an increase of feat 478 resulted in a decrease of predicted inhibition efficiency for all molecules, whereas an increase of feat 479 on the contrary, increased predictions.

Clearly, there is something important about this bond distance to be present in 20% of the features. In cyclic structures made up of five or six atoms, in either case where sulfur is in the ring or attached as a branching functional group, this sulfur-nitrogen distance would put sulfur and nitrogen on opposite sites of the ring. If the molecule benefits from sulfur and nitrogen being close to one another for corrosion inhibition—such as the formation of bidentate chelates—this position would prevent nitrogen from working together with sulfur in bonding, and highly electronegative nitrogen would draw excess electrons necessary for bonding away from the sulfur donation centers. In combination with trends of feature 1016, where maximum nitrogen-nitrogen pairs cause a decrease in predicted corrosion efficiency, it is clear that the position of nitrogen is very important for maximizing corrosion inhibition performance.

This can also be used as a design principle: (i) as the atompair distance at a topological distance of 4 between C-S increases relative to N-S, and (ii) as the presence of neighboring nitrogen atoms that do not contribute to surface binding decreases, the corrosion inhibition performance increases.

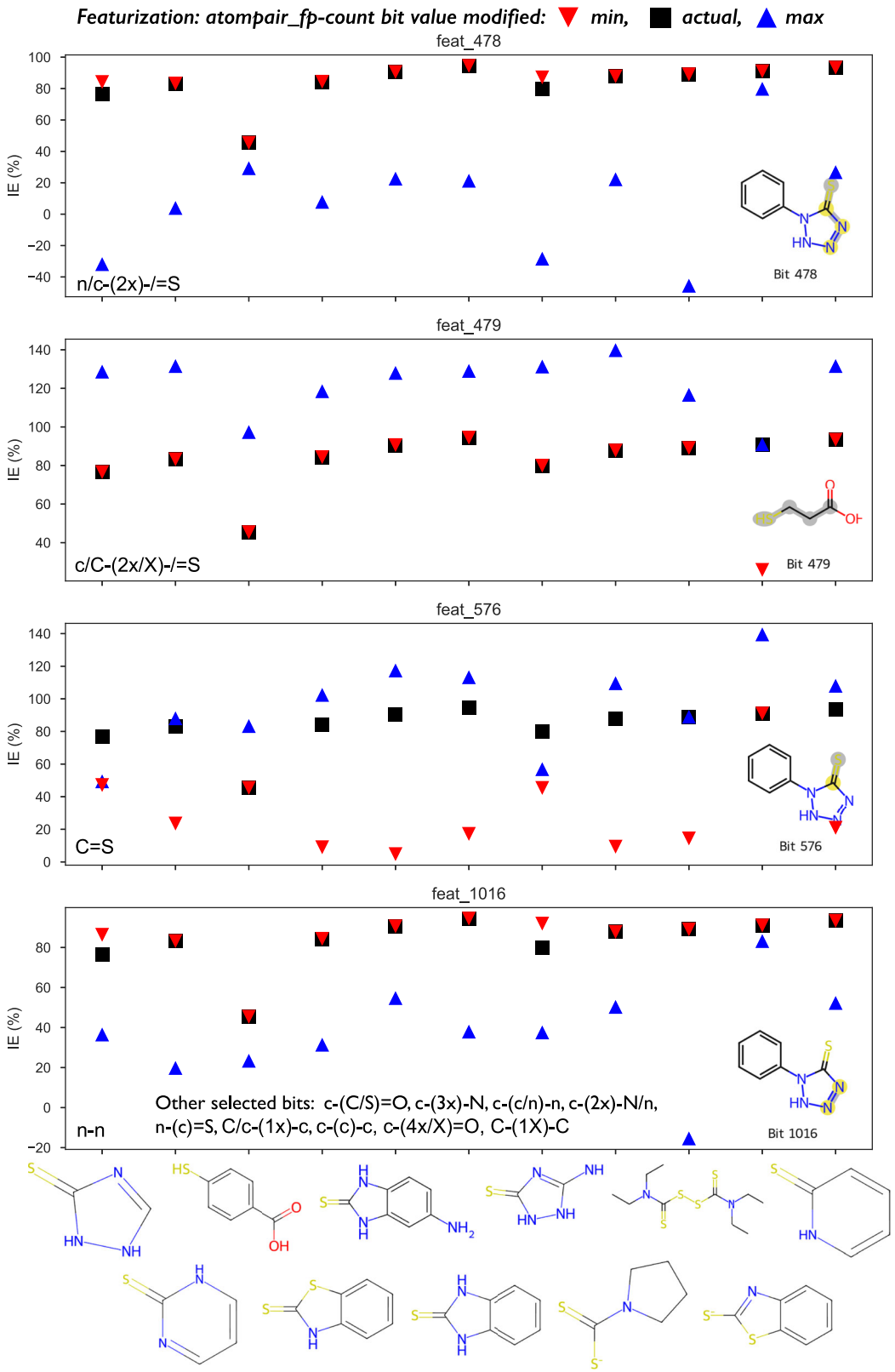


Fig. 3 | Creating counterfactual predictions for molecules with highest target values in the experimental dataset. The *x*-axis shows the molecules, and the *y*-axis shows the predictions. For actual predictions, the `atompair-count` model with the lowest CV-RMSE is used for predictions. For min/max, only the value for the corresponding feature is changed to the min/max found in the original featurization dataset, and then the same model is used for predictions. Visualization of the features as bits on example molecules is shown at the bottom right of the plots, and corresponding SMILES-like strings are shown on the bottom left.

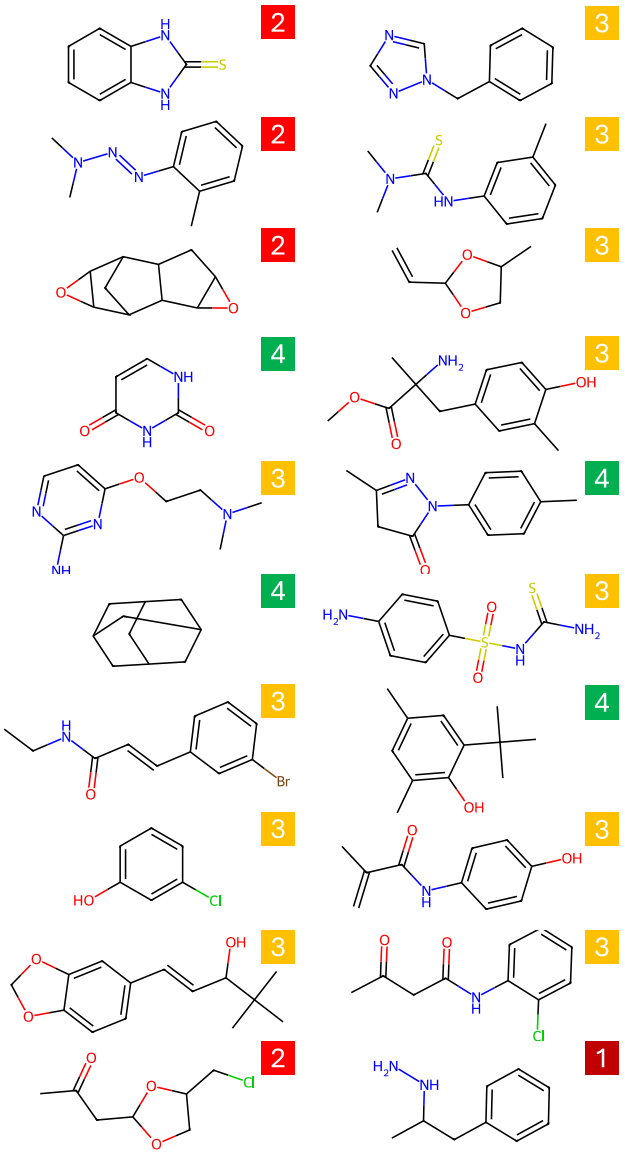


Fig. 4 | Molecules most similar to the pseudomolecule for the similarity metric cosine similarity. EPA toxicity classification shown next to the molecules: 1 highly toxic, 2 moderately toxic, 3 slightly toxic, and 4 practically non-toxic. Molecular featurization was done with PaDEL, and the dataset used was a toxicity database curated from previously published work[35].

From these results, we argue that the potential of counterfactuals for gaining mechanistic insight is promising. Before diving deeper into mechanistic insights, we would also like to demonstrate a way of analyzing the physicochemical features, and afterwards combine both with a more complete analysis in the section on SHAP analysis.

Bayesian optimization as a tool for understanding model decision-making process

Bayesian optimization is a statistical method for optimizing any black-box objective function that lacks an analytical form and is expensive to evaluate. Instead of the true objective function, Bayesian optimization uses a surrogate model that is an approximation of the objective function. This cheaper-to-analyze alternative is used to extrapolate the function with a measure of uncertainty. An acquisition function is used to select the next point to sample. This selection can be a combination of exploration (searching areas of the *n*-dimensional search space where the surrogate model is uncertain, active learning) and exploitation (searching areas where the surrogate model predicts high objective function values, Bayesian optimization). In a sequential manner, the objective function is evaluated at the selected point, the surrogate is updated based on the gained information, and the acquisition function decides on the next point to be evaluated. This process is repeated iteratively, while with every step, the global optimum of the surrogate model converges towards the global optimum of the objective function.

Bayesian optimization can also assist in illuminating the black-box function of corrosion inhibition. The advantage of using physicochemical descriptors as model features is the clarity of the features—every feature is defined clearly, whether it is heteroatom content, ring number, electronegativity, or any other interesting physicochemical quality. However, the disadvantage is often that the calculated descriptors are quite arcane; therefore, it is difficult to have an intuitive understanding of what kind of molecule structure would result in the quantitative value of a descriptor. This can potentially make interpretation difficult. The reversed problem is even more difficult: given multiple such features, a molecular chemist or a materials scientist would have a hard time converting these quantitative parameters into an actual molecule. Bayesian optimization offers a way out of this thorny reverse-design problem.

In our case, the black-box function to be optimized was the best model based on PaDEL featurization. We were looking for the selected feature values that would result in the highest inhibition efficiency. After initializing the optimization with samples in our dataset combined with 2000 samples with randomized features to be analyzed, Bayesian optimization was run for 1000 iterations. The features that resulted in the optimized maximum were the optimal molecule parameters that the model predicts will lead to the best inhibition efficiency. We call this artificial creation an optimized "pseudomolecule".

This ideal "pseudomolecule" can be used as a template for finding real molecules that are similar. We can compare the pseudomolecule with molecules from any given database and find the molecules most similar to it. For this, we used a previously published toxicity database[35] that contains over 10,000 molecules. The choice of selecting a toxicity database was deliberate—aside from predictions from our model, the database would also provide information on the toxicity of compounds. We calculated similarity between our candidate pseudomolecule and the molecules in the toxicity database with the cosine similarity metric. The most similar 20 molecules are presented in Fig. 4 in descending order in similarity. The EPA toxicity classifications of the molecules are also presented with the molecules, which correspond to: 1 highly toxic, 2 moderately toxic, 3 slightly toxic, and 4 practically non-toxic[36].

What we see from this figure is a complementary picture to the results from the previous model with atompair-count featurization. The most similar real molecule in the dataset to our pseudomolecule is 2-mercaptobenzimidazole, which is a known, good corrosion inhibitor and already present in our training dataset (measured IE 94.6%). Other molecules are not present in our dataset. One common theme across all molecules was the presence of bulky benzene groups. Most benzene groups were without any heteroatoms, but some molecules had pyridine, pyrimidine, or dioxolane rings. Connected to the benzene groups, most of the molecules had hydrocarbon chains with secondary/tertiary amines, ketone, or carboxylic acid functional groups. Some molecules had sulfur, always in close presence to nitrogen. Similar to the `atompair-count` case, S-N at a four-atom distance was not present in the molecules.

This type of analysis is particularly useful for gaining a mechanistic understanding with greater confidence when working with limited data. One key advantage is its ability to serve as a tool for interpolation rather than extrapolation. Since optimization focuses on feature value boundaries already present in the dataset, the model operates within familiar territory, enabling more reliable interpolation within those boundaries. The addition of toxicity values also gives the choice of selecting non-toxic molecules for further experimentation.

In the next section, we show how the combination of this sort of reverse engineering and SHAP analysis helps to understand what the "perfect" molecule for corrosion inhibition would be according to statistical models.

SHAP analysis for deciphering feature influence

SHAP (SHapley Additive exPlanations) analysis is a method originating from game theory[21]. The SHAP value in the context of a machine learning model is the expected individual contribution of a feature to the model prediction. The SHAP value for any given feature i is calculated as:

$$\phi_i(v) = \sum_{C \subseteq N-i} \frac{|C|!\,(n-|C|-1)!}{n!}\{v(C \cup \{i\}) - v(C)\}, \quad (1)$$

where v is a characteristic function that maps every coalition of n features to a prediction. Here, v is the machine learning model, and C is such a coalition —a group of features working together. $|C|$ is the number of features in coalition C. $|C|!$ is the number of ways coalition C can form. $(n-|C|-1)!$ is the number of ways the rest of the features can join to the coalition after feature i joins. $n!$ gives the number of ways to form a coalition from n features. The resulting term $\frac{|C|!\,(n-|C|-1)!}{n!}$ is the weight for marginal contribution, or the probability of feature i making a contribution to coalition C. The term $v(C \cup \{i\}) - v(C)$ is the marginal contribution of feature i to the coalition C. All the marginal contributions of a feature with their probability of making those contributions are weighed with respect to the weights for marginal contribution, then summed over all coalitions that which the feature can make a marginal contribution. This gives the expected marginal contribution, in other terms, the SHAP value. In this way, all the possible coalitions that a feature can contribute to are considered, and a feature's individual contribution as well as the interactions between features are evaluated.

Figure 5 presents SHAP beeswarm plots for (a) atompair-count, (b) PaDEL, and (c) MACCS featurization methods. The beeswarm plot represents the distribution of the feature impact of SHAP values across a dataset. Each point in the plot represents the SHAP value of a feature for a specific instance, with color indicating feature value. Beeswarm plots can identify which features influence the model's predictions the most, through the direction (positive or negative) and magnitude of these influences across the dataset. Positive SHAP value contributions mean that the value of that feature is expected to increase the corrosion inhibition efficiency of a given instance, and vice versa. This helps in interpreting the model, uncovering feature importance, and detecting patterns in the predictions.

The importance of pH has been identified in our previous work[25], and it is also present as a feature in all of the presented model featurizations. A detailed analysis will not be repeated here; however, we would like to highlight that pH was always chosen as one of the most important features of the models, despite having no linear correlation with targets. Based on SHAP dependency plots in our previous works and the beeswarm plots here, we observe that models have very negative SHAP values for very large and very small feature values. The reason becomes clear when one observes the Pourbaix diagram of Al, where aluminum oxide is stable and protective in the pH range of 4–9, but starts to disintegrate below and above this range. The models capture that behavior quite well, proving that, given the right features, mechanistic behavior resulting from the environment-substrate interactions can be captured.

Atompair-count fingerprints. Figure 5a shows SHAP beeswarm plots for `atompair-count` fingerprints. The visualization of the molecule substructures as bits, as seen in the previous section, allows us to explain the feature SHAP behavior. The notation used in this section is used as before: uppercase for aliphatic, lowercase for aromatic atoms, / for denoting structures corresponding to multiple atoms or bonds, and *nx* for the number *n* of any atoms in between.

Feature 576 corresponds to substructure bit c/C=S, a carbon-sulfur double bond. The presence of sulfur is expected to increase the inhibition efficiency predictions, and it has the biggest impact on model predictions. This is in line with trends seen from literature, which mentions the high tendency of sulfur to bond with copper[23,37], which would also allow the organic molecules to bond with Cu-based intermetallics of the AA2024-T3 substrate, which are the root cause for localized corrosion[38–40]. If the intermetallics are protected, it would greatly decrease the microgalvanic driving forces that cause localized corrosion of the alloy. For this reason, it is not surprising that sulfur presence is the most important feature, but it is nonetheless remarkable that the model has learned the importance of bonding with such a clear tendency. Analysis of other model SHAP values shows that this is not a coincidence.

Feature 453 corresponds to substructure bit c/C-4x=O. High values have a positive impact on the model, whereas low values expect to have a minor negative one. Analysis of molecules that contain this bit reveal that the majority had a ccccC=O substructure. That corresponds to an aromatic ring with alcohol, ketone, or carboxylic acid functional groups. Assuming that they are not near the substrate anchoring sulfur/nitrogen groups, such structures would indeed push away the corrosive Cl^- ions through steric hindrance.

Feature 1295 corresponds to the substructure bit C-X-C. Higher feature values result in a sharp decrease of SHAP values, and lower values of it result in minor positive values. The majority of molecules containing this feature include CCC and CNC substructures, which are characteristic of aliphatic hydrocarbon chains. These chains exhibit low reactivity, limiting their interaction with both the surface and the surrounding environment. The prevalence of these groups is a characteristic quality of surfactants, where a long aliphatic tail is typically attached to a carboxylic or amino group. Such surfactants are recognized as effective inhibitors under specific conditions for substrates like carbon steels[41,42]. In these cases, the presence of long tails likely contributes to corrosion inhibition by providing steric hindrance. However, this was not the case for this alloy system. An excessive presence of such chains can lead to a bulky structure with decreased molecule solubility without contributing to surface bonding, which must have been the dominant negative effect.

Feature 479 corresponds to substructure bit c/C-2x-S, sulfur bonded to any two atoms bonded to a carbon atom. Whereas feature 576 contained information on double-bonded sulfur, this one contains information on single-bonded sulfur. Analogous to feature 576, higher values correspond to significantly increased SHAP values. It seems that sulfur with this topological distance to carbon is predicted to contribute significantly to corrosion inhibition. As previously discussed through counterfactual analysis, this sulfur-carbon distance is notably peculiar. For cyclic structures composed of five- or six-membered rings, where sulfur is incorporated within the ring or attached as a functional group, this distance positions sulfur and carbon farther apart, on the opposite sides of the ring. This means that this feature value can be maximized through dithiocarbamate-like structures, or S attached to ring structures. This observation suggests a structural configuration in which sulfur is either bonded as a functional group to one of the ring's vertices or directly integrated into the ring structure.

Trends for the rest of the features are less straightforward to analyze. *Feature* 1016 corresponds to n-n, aromatic nitrogen connected together. It seems that a high n-n presence results in more activity, expected to push the predictions more to higher and lower values. *Feature* 200 corresponds to c-x-c/C, where higher values decrease the SHAP values, which might be related to aromaticity degree. *Feature 478* corresponds to n-2x-/=S, where the presence of it is making a molecule take more extreme SHAP values. *Feature 311* corresponds to ccc, which again is related to the aromaticity degree, and the influence on the model is low.

Based on these observations, it seems that the correct combination of c/C=S, c/C-4x=O, c/C-2x-S, and n-2x-/=S might result in ideal model predictions.

PaDEL descriptors. Figure 5b shows SHAP beeswarm plots for PaDEL descriptors. The descriptions of the computationally generated descriptors are quite often not adequately documented, which requires double-checking multiple sources. Analysis of the descriptors below is primarily based on the book *Molecular descriptors for chemoinformatics*[43], and the documentation pages of numerous descriptor calculator packages.

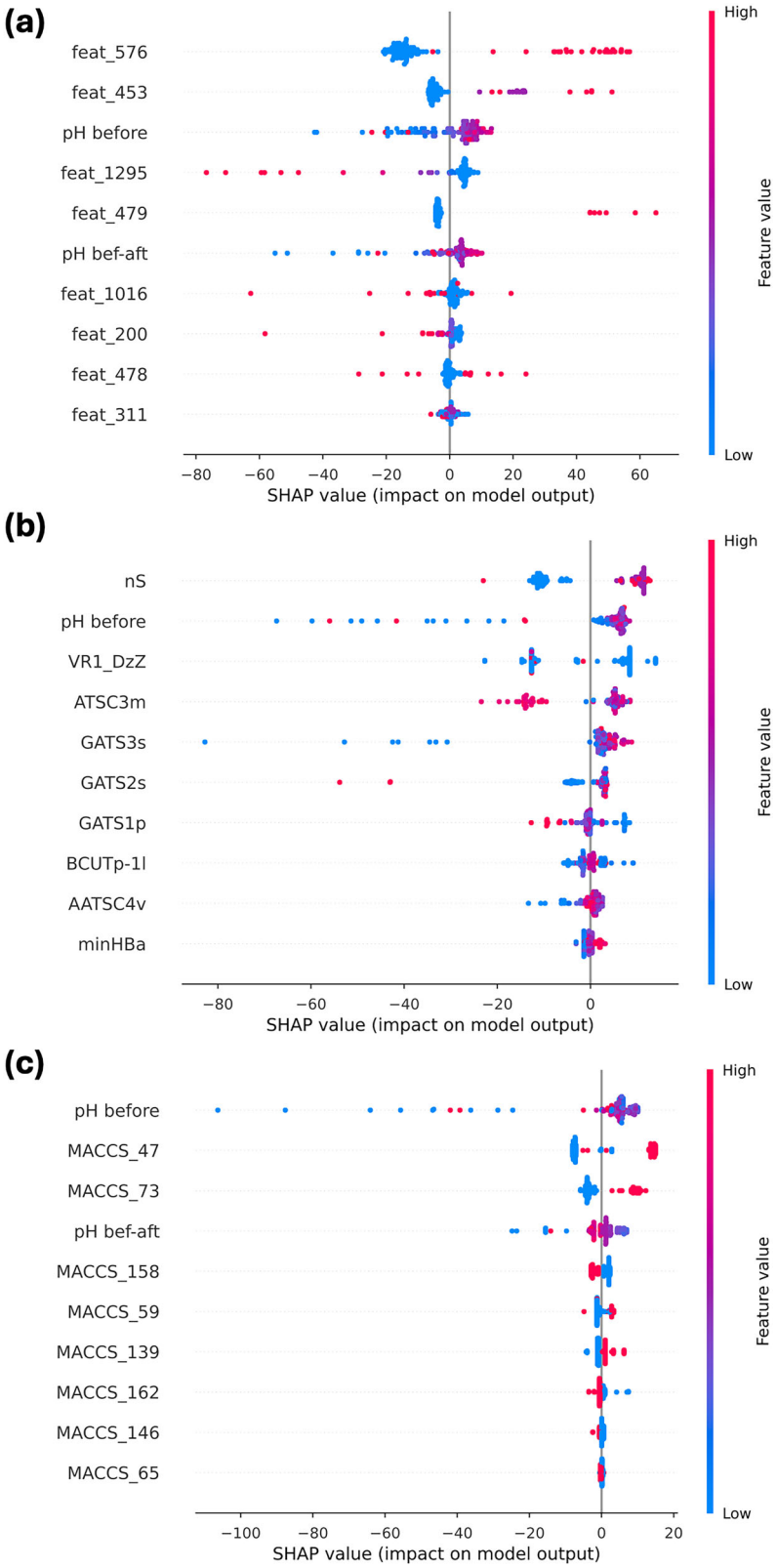


Fig. 5 | SHAP beeswarm plots displaying how features in a dataset impact model output for featurization. a atom pair-count, **b** PaDEL, **c** MACCS. Each dot represents an individual model instance (molecule), which pile up along each feature row to show density. Each row corresponds to one feature, which is sorted by the mean of absolute SHAP values. Color is used to display the original value of a feature, whereas the SHAP value is the impact of a given feature value on the model output. Large values correspond to larger expected model impact.

nS represents the number of sulfur atoms in the molecule. An increase in the number of sulfur is expected to increase model predictions. It is the descriptor with the highest impact, and the presence or absence of sulfur is predicted to be critical in the inhibition property of the molecule. It is directly related to features 576 and 479 of the atompair-count fingerprints.

ATSC stands for Centered Autocorrelation of a Topological Structure (also known as Moreau-Broto autocorrelation). Autocorrelation descriptors calculate the correlation between a specific atomic property, such as atomic mass, at a defined topological distance within the molecule. They capture how a property is distributed across the molecular structure. The property

values are "centered" by subtracting the mean property value across the molecule.

ATSC3m reflects how the atomic mass is distributed and correlated across atoms that are three bonds apart in a molecule. *3* refers to the "lag", which indicates the topological distance between atoms being considered in the molecule, in this case, three bonds. *m* denotes that the descriptor is weighted by atomic mass. A higher ATSC3m value suggests significant variation in atomic masses at this specific distance, indicating that heavier and lighter atoms are more differently positioned in relation to each other. A lower value indicates that there is little variation in the atomic masses of atoms that are three bonds apart. Since high ATSC3m values are expected to result in a significant drop in the majority of prediction values, neighbors that are two atoms apart and similar to one another in atomic mass could be more suitable for inhibitor molecule structures.

AATSC4v quantifies the autocorrelation of atomic van der Waals volumes within a molecule at a topological distance of four bonds, further normalized with respect to molecule size before calculating the autocorrelation. Higher values did not markedly improve prediction performance, but lower values certainly hindered it. This can be observed for straight-chain structures larger than butane. However, for higher values, individual or fused ring systems exhibit greater topological distances and hence greater potential. Examples include carbon atoms in aromatic rings such as benzene, benzimidazole, benzotriazole, or cyclopentanes. Additionally, the presence of two neighboring heteroatoms, such as nitrogen in five-membered rings like imidazoles, can also contribute to these increased distances.

The behavior of *VR1_DzZ* was difficult to analyze. Dz are a modification of ATSC descriptors that use topological distance in conjunction with properties of atoms (see Dz^K pg.33[43]). Official PaDEL documentation describes VR1_DzZ as a "Randic-like indices eigenvector-based index from Barysz matrix / weighted by atomic number" (see Randic-like pg.164, VRA1 pg.717[43], calculation of Barysz distance matrix[44]), which is defined by coefficients of the eigenvector associated with the largest negative eigenvalue. It is related to local vertex invariants able to provide discrimination among graph vertices. However, its non-linear, complicated effect is difficult to analyze in isolation, where high values seem to hinder the inhibition efficiency; therefore, it is not further discussed, as tying it to the molecular structure is not accessible.

GATS stands for Geary Autocorrelation of Topological Structure. Like ATSC, GATS descriptors are used to quantify the autocorrelation of a specific atomic property over a defined topological distance in a molecule. GATS differs from other types of autocorrelation by including a normalization factor, which adjusts for the number of atoms and bonds considered, providing a scale-independent measure. A strong positive correlation produces low GATS values between 0 and 1, negative autocorrelation produces values larger than 1, whereas no correlation corresponds to a value of 1 (pg.32[43]).

GATS3s provides a measure of how the Sanderson electronegativity varies across the molecule at a topological distance of three bonds. *3* again refers to the lag. *s* denotes that the descriptor is weighted by atomic Sanderson electronegativity, which is a specific measure of electronegativity that describes the ability of an atom to attract electrons in a chemical bond. High GATS3s values indicate a significant variation in electronegativity values among atoms that are three bonds apart. This might occur in molecules with a mix of atoms that have widely differing electronegativities, as for heteroatoms (e.g., sulfur, nitrogen, oxygen) in an organic molecule. Low GATS3s values suggest uniformity in electronegativity at this distance. Similar electronegativities would indicate less variation in the ability to attract electrons across the molecule. The SHAP values seem to increase with increasing GATS3s values, suggesting that for ideal inhibitors atoms at 3-bond distance should have a higher electronegativity difference.

GATS2s is similar to GATS3s with the only difference being the topological distance, which is 2 in this case. This suggests molecules with atoms at 2-bond distance with differing electronegativities would result in higher target values. Unlike GATS3s, high GATS2s values decrease the model performance for two outliers.

GATS1p is also similar to GATS3, but in this case, the topological distance is 1, so it considers neighboring atoms. *p* indicates that the descriptor is weighted by atomic polarizability. GATS1p, therefore, is a measure of how the property of atomic polarizability varies over the structure of the molecule for neighboring atoms. Polarizability is a measure of how easily the electron cloud around an atom can be distorted by an electric field, which is related to the size of the atom and its electron density distribution. Lower GATS1p values seem to increase the model predictions. A low GATS1p value suggests that the atomic polarizability of adjacent atoms is quite similar. This would occur in molecules where atoms have similar sizes and electronic environments, leading to little variation in how easily their electron clouds can be distorted.

The *BCUTp-1l* descriptor reflects the distribution of polarizable atoms in a molecule. *BCUT* stands for Burden - CAS - University of Texas eigenvalues. It refers to a set of molecular descriptors derived from the Burden matrix, a matrix which captures a desired property correlation between every atom in a molecule. *p* indicates that the descriptor is weighted by atomic polarizability. *1l* signifies the lowest eigenvalue obtained from the Burden matrix. A low eigenvalue typically indicates that the molecule's polarizability is relatively evenly distributed or that there are no extreme variations in polarizability across the molecule. Conversely, a higher eigenvalue suggests more significant variations, possibly indicating regions of the molecule with high and low polarizability. In the case for this model, its effect was not straightforward to analyze, but it was observed that higher values corresponded to a more limited absolute impact, suggesting less active molecules, which may not be desirable for inhibitor molecule design.

minHBa refers to the calculated minimum hydrogen bond acceptor strength in a molecule. A hydrogen bond is the electrostatic attraction between a hydrogen atom covalently bonded to a more electronegative atom or group, the "donor", and another electronegative atom that has a lone pair of electrons, the "acceptor". Main hydrogen bond donors and acceptors are electronegative atoms like N and O, which have lone pairs of electrons that can attract the hydrogen atom. The minHBa descriptor specifically focuses on identifying the weakest hydrogen bond acceptor within the molecule—a high minHBa value would mean that even the least effective hydrogen bond acceptor in the molecule has a relatively high hydrogen bonding potential. A high minHBa seemed to increase the prediction values. This suggests that C atoms with higher hydrogen bond acceptor values would assist in improving predictions. This might be related to the aromaticity: as it was found that aromatic rings act as hydrogen bond acceptors[45], therefore, compared to aliphatic C chains, the presence of aromatic rings might increase the minHBa values. A high hydrogen bonding capacity would help in the self-assembly process by creating more intact monolayers as the organic molecules adsorb to the surface with one part, and attach with one another through hydrogen bonding.[46,47] A tighter bonding between adsorbed molecules would hinder chlorides from penetrating in between. However, the influence of the descriptor on the model is weaker than the rest of the features.

Summarizing the strongest interpretable influences that would result in a higher predicted inhibition efficiency for AA2024-T3 alloy:

- High number of sulfur atoms.
  The molecule likely contains multiple sulfur atoms. Sulfur is relatively electronegative (though less so than oxygen and nitrogen) and can participate in various chemical environments, such as thiols (-SH), thioethers (R-S-R'), or disulfides (R-S-S-R').
- High GATS3s and GATS2s: high variation in electronegativity at a three- and two-bond distance.
  This suggests that at a distance of three- and two-bonds, there is a significant difference in the Sanderson electronegativity values. This could mean that there are alternating patterns of atoms with high and low electronegativities. The presence of highly electronegative

heteroatoms such as sulfur, nitrogen, and oxygen, combined with less electronegative atoms such as carbon at three- and two-bonds distance, would result in higher descriptor values.

- Low ATSC3m: low variation in atomic mass at a three-bond distance. The low ATSC3m value indicates minimal variation in atomic mass at a distance of three bonds. This implies that the atoms in the molecule, despite the different types, have similar masses. Since sulfur has a relatively high atomic mass compared to carbon, oxygen, nitrogen, or hydrogen, this would result from structures with only a small amount of sulfur at the periphery of the structure, leading to a more uniform mass distribution.
- Low GATS1p: low variation in polarizability at a one-bond distance. The low GATS1p value indicates uniformity in atomic polarizability for neighboring atoms. This suggests that the atoms connected directly to each other do not vary much in their polarizability, which could be the case if they are similar types of atoms or atoms with similar electronic environments.

Molecular structures corresponding to such trends would include several sulfur atoms, either in a linear arrangement or part of cyclic structures. Alternating electronegativities of N/O with carbon at two/three-bonds distance (structures of -S/N/O-X-C- / -S/N/O-X-X-C-) would result in high GATS2s/GATS3s. Sanderson electronegativities increase in order of C<S<N<O (2.75<2.96<3.19<3.65, in Pauling units); therefore, N/O coupled with C would contribute more to the increase in GATS2s and GATS3s. This also coincides with feature 478 and 479 bits c/C-2x-S and n-2x-/=S from atompair-count fingerprints. Despite the presence of heavy sulfur, the molecule's structure would need to consist mainly of atoms of similar mass, meaning structures with long carbon chains or multiple cyclic structures are necessary to give rise to low ATSC3m. In addition, adjacent atoms should have similar polarizability values, indicating a lack of highly polarizable atoms directly bonded to less polarizable ones, again pointing towards long carbon chains or cyclic structures. Derivatives of larger thiols (R-SH) and thioethers (R_1-S-R_2), as well as cyclic thiophene and benzothiophene-like molecules, would satisfy such criteria.

MACCS keys. Figure 5c shows SHAP beeswarm plots for MACCS keys. Despite lower prediction performance, MACCS was added to the previously studied featurization methods because MACCS features are completely predetermined and very interpretable. The MACCS keys were interpreted based on the Mayachemtools MACCS keys documentation[48].

The *MACCS 47* key corresponds to the S-/=x/X-n/N substructure. The presence of such a substructure is expected to increase the inhibition efficiency predictions for almost all molecules. This matches with the GATS2s requirements from PaDEL featurization, as the S-X-N structure would have a higher electronegativity difference at two-bond distance.

The *MACCS 73* key corresponds to the S=X substructure, which matches the bit 576 from atompair-count featurization. The presence of a double bond with sulfur, along with the S-X-N substructure, has the largest impact on higher inhibition value predictions. The rest of the features influence the model significantly less. As discussed before, sulfur presence is critical for inhibition, and multiple models consistently using related features underline this.

The *MACCS 158* key corresponds to the C-N substructure. Nitrogen presence is expected to decrease the model predictions. This was counterintuitive, as S-X-N presence was expected to increase the predictions. One explanation might be that in the presence of sulfur, the model overshoots the predictions, and this feature decreases it to the expected values. This was actually observed for counterfactual predictions (Fig. 3), where when feature 576 (corresponding to C=S) is artificially replaced with the maximum values found in the dataset, the predictions went above the theoretical maximum of 100% for the majority of well-performing molecules. Meanwhile, if there's no sulfur present, the molecules are often just not expected to work as corrosion inhibitors for the selected AA2024-T3 substrate.

The *MACCS 59* key corresponds to the S-x-x substructure (sulfur bonded to any atom with a non-aromatic bond, whereas that atom is bonded to another with an aromatic bond). For structures where sulfur is bonded to an aromatic ring structure, its presence often can be correlated with an increase in the model predictions, although the underlying relationship seems to be complex.

The *MACCS 139* key corresponds to the OH substructure. This would be present in carboxylic acids and alcohols, and its presence is expected to increase the model predictions. MACCS 139 shows similarity to feature 453 from atompair-count featurization.

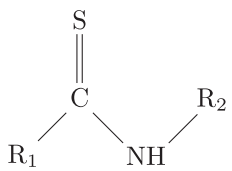
The *MACCS 162* key corresponds to the presence of an aromatic substructure. Its presence is expected to slightly decrease the model predictions. This could be working with MACCS 59, where the aromaticity effect in combination with sulfur presence determines the complete effect of the ring structures.

The *MACCS 146* key corresponds to the condition where O < 2. This would act as a carboxylic acid detector, as one acid group would need at least two oxygen atoms. When this condition is true, and there are no acid groups on the molecule, the predictions of the model are expected to decrease. In combination with MACCS 139, this would determine the influence of single carboxylic acid functional groups.

The *MACCS 65* key corresponds to the c-n substructure. Its influence on the model is very weak and mixed.

Taken together, the combined information of all keys suggests that molecules containing S=C/C-N substructures, coupled with carboxylic acid groups and limited C-N bonding, may exhibit strong corrosion inhibition potential.

The common trends. Combining the insights from all three featurizations, we deduce that the presence of S=C would improve the model predictions. S=C can form in sulfur analogs of carbonyl and carboxyl group thiocarbonyl and dithiocarboxyl groups, and would act as the anchor binding the molecule to the substrate. This substructure would ideally have N as its neighbor to C, which seems to have a positive influence on the inhibition. This might potentially be a result of N assisting with the bonding through the S, or through its electronegativity, stabilize the hydrogen bonding formed between the molecules during the self-assembly process. This gives us a molecular structure template for an ideal corrosion inhibitor:



where R_1 can be S for dithiocarboxylic acids, or a longer chain that starts with S for dithiocarbamate structures. R_1 and R_2 can contain and/or be merged together into single or fused ring structures. This, in combination with carboxyl presence, would fulfill the criteria from different featurizations.

These patterns are found in the structure of the commonly used corrosion inhibitors such as 2-mercaptopyrimidine, ammonium pyrrolidinedithiocarbamate, and 3-amino-1,2-4-triazole-5-thiol[25]. Literature suggests that S and N heteroatom containing organic molecules can stabilize AA2024-T3 aluminum oxide by covering the surface through sulfatization, or can adsorb on the copper-rich intermetallics, suppressing the cathodic reactions, which often is the driving force of corrosion in the surrounding area[49]. Notably, even without any expert-guided feature selection, through observing the trends hidden in the dataset statistics alone, the results of this methodology corroborate the previous spectroscopy results that aimed to uncover mechanisms responsible for structures responsible for corrosion inhibition of various organic molecules[23].

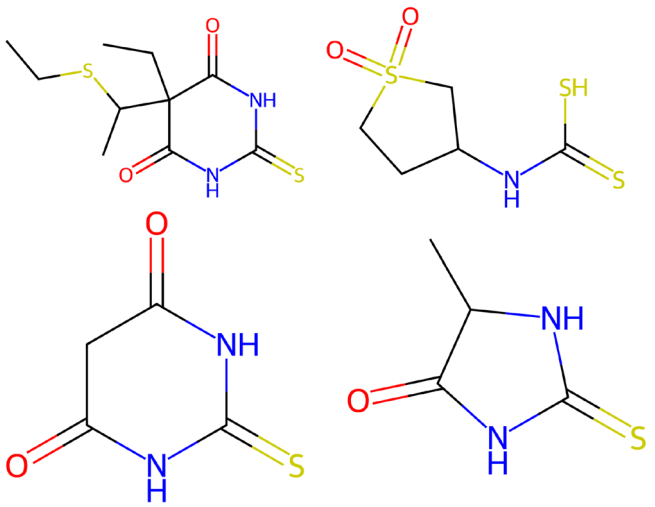


Fig. 6 | Non-toxic molecules from the toxicity database that fit the trends observed from different featurization methods. Top-left: 5-ethyl-5-(1-(ethylthio) ethyl)-2-thiobarbituric acid, top-right: sulfocarbathione, bottom-left: 2-thiobarbituric acid, bottom-right: 5-methyl-2-thiohydantoin.

The C(=S)N (and also c(=S)N, C(=S)n, c(=S)n for aromaticity variants) SMILES string can be converted into a SMARTS pattern, with which molecule databases can searched for this substructure. For the toxicity database we have used in this study (which contains more than 10,000 molecules), this search ends up in 123 hits of this database, which can be used as lead candidates for exploring potential, yet untested, corrosion inhibitors. These resulting lead molecules can be further constrained to include the trend coming from atompair-count featurization of feature 453 c/C-4x=O, where its presence is expected to increase model predictions. The presence of c/C-4x=O molecular fragment further decreases the lead molecules to 10. Among these, the molecules with EPA classification 3 and 4 are displayed in Fig. 6: 5-ethyl-5-(1-(ethylthio)ethyl)-2-thiobarbituric acid, sulfocarbathione,2-thiobarbituric acid, and 5-methyl-2-thiohydantoin.

Out of the displayed four molecules, only 2-thiobarbituric acid (Fig. 6, lower-left) was available to purchase off the shelf. To show the validity of our gained insight, we have conducted electrochemical experiments using the same methodology used to acquire previous targets to curate the original training dataset[25]. We have tested 1 mM 2-thiobarbituric acid and adjusted its pH to 7.0, as the original solution had a pH of 2.3, much lower than the thermodynamic stability window of Al_2O_3, which is between 4 and 8.5[50].

Electrochemical impedance measurements performed after 24 h of electrolyte exposure show that 2-thiobarbituric acid is indeed a promising molecule for corrosion inhibition. A comparison of the diameters of the suppressed semicircles shown in the Nyquist plot of Fig. 7a shows that the addition of thiobarbituric acid enlarges the diameter significantly, which is related to an increase in the polarization resistance and overall corrosion inhibition of the surface. Bode plots of Fig. 7b demonstrate that the addition of thiobarbituric acid increased the impedance modulus values measured at 10^{-2} Hz, which represents the corrosion resistance of the inhibitor-surface interface[51]. Impedance modulus values were raised to 64.2 ± 14.5 kOhm cm^2 in the presence of thiobarbituric acid, which corresponded to an inhibition efficiency of 84.1 ± 3.5%. An interesting observation consistent throughout samples was that while the open circuit potential (OCP) values were constant around −520 mV vs. Ag||AgCl throughout the first 24 h (similar to uninhibited values), the linear polarization resistance values kept increasing throughout time, without showing any signs of slowing down. These observations indicate that thiobarbituric acid can work as a strong corrosion inhibitor, and other corrosion inhibitor candidate molecules presented at Fig. 6 should also be tested for their potential.

This paper demonstrates that mechanistic insights can be derived from machine learning models to design novel functional molecules. Rather than

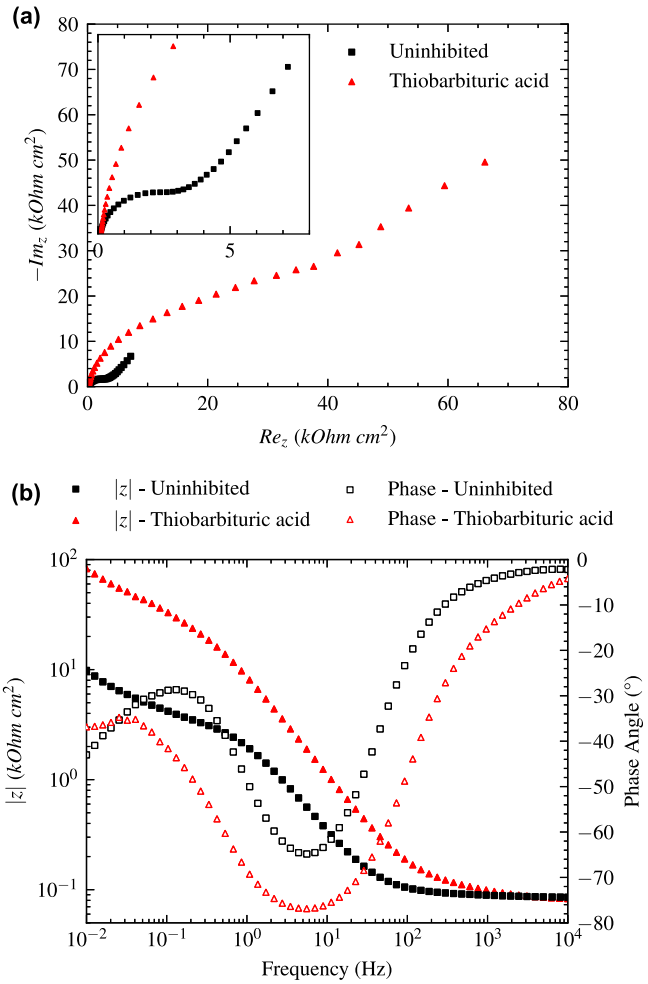


Fig. 7 | Electrochemical impedance spectroscopy of AA2024-T3 alloy exposed to 0.1M NaCl electrolytes with or without 2-thiobarbituric acid for 24 h. a Nyquist, **b** Bode modulus and phase angle plots, Inset shows a zoom in of lower resistance values.

focusing on predicting individual molecule performance from small datasets—a task that is inherently limited by dataset size—we can reverse-engineer statistical models to understand their decision-making processes. By representing molecules using various featurization methods and applying feature elimination techniques to identify the most important features, we gain insight into which feature combinations represent the problem best. These insights can then be integrated with the domain knowledge of scientists to utilize machine learning models beyond their typical "black-box" functionality.

However, it is crucial to remain aware of the limitations of different molecular representations. For example, while hashing-based methods are more generalizable, they may lead to bit collisions, making models volatile and less interpretable. Fingerprint techniques, though useful, may overlook subtle molecular changes if these do not alter the structural fragments being represented, whereas physicochemical descriptors may be more suitable for capturing such nuances. Nevertheless, fingerprints can effectively capture broader trends, as they are closely tied to molecular structure, and their interpretation is often more straightforward since they represent visualizable substructures—provided there are no bit collisions.

The combination of diverse molecular representations holds significant, largely untapped potential for scientific discovery via statistical models. Agreement among models that use different featurization methods can enable feature selection to serve as a powerful tool—similar to how multiple spectroscopic techniques are combined in materials science to achieve a more comprehensive understanding of the studied scientific phenomena. Additionally, SHAP (SHapley Additive exPlanations) analysis

offers promise in isolating the effects of complex trends, and it is highly effective in creating controlled variables within a materials research framework. Insights gained from different representations can complement one another, forming the basis for testable hypotheses, as illustrated here in the discovery of a novel corrosion inhibitor, 2-thiobarbituric acid for AA2024-T3.

Next to what has been studied in this work, further insights can be gained by manipulating molecular structures—such as adding or removing fragments—at no additional cost after the model has been trained, allowing for the testing of trends in the material properties of interest. If these insights can then be integrated into generative chemical foundation models, it would enable the rapid design of new molecules at a fraction of the original cost.

Methods

Generation of electrochemical targets

Target generation experiments are discussed in a detailed manner in our previous publications[25]. This work is based on 107 organic molecules tested as corrosion inhibitors for the same substrate, AA2024-T3. The targets corresponded to three different electrochemical experiments: EIS performed at 24th hour (_EIS24h), linear polarization resistance experiments performed at 24th hour (_24 h), and linear polarization resistance values averaged through time (_avg) with trapezoidal integration[52]:

$$\left\langle R_p \right\rangle = \frac{1}{t_f - t_0} \int_{t_0}^{t_f} R_p(t)dt \tag{2}$$

$$\approx \frac{1}{t_f - t_0} \sum_{k=1}^{N} \frac{R_p(t_{k-1}) + R_p(t_k)}{2} (t_k - t_{k-1}) \tag{3}$$

where t_f is the final measurement time, t_0 is the initial measurement time, and k is the index for the performed discrete measurements. The results from these experiments are represented in three different forms: raw electrochemical polarization resistance values R_p, R_p values converted into inhibition efficiencies (IE) with:

$$\mathrm{IE} = \frac{R_p^{\mathrm{inh}} - R_p^{\mathrm{blank}}}{R_p^{\mathrm{inh}}} \tag{4}$$

and into inhibition power (IP) with

$$\mathrm{IP} = 10\log_{10} \frac{R_p^{\mathrm{inh}}}{R_p^{\mathrm{blank}}} \tag{5}$$

where superscripts "inh" and "blank" stand for samples exposed to organic molecules or only to NaCl, respectively. These three different experimental approaches with three different representations resulted in nine different potential target values.

Generation of fingerprints and physicochemical descriptors

The SMILES strings of 107 small organic molecules are first desalted (removal of ionic metal parts from the strings) for correct descriptor calculation, then converted into structural fingerprints and physicochemical descriptors with the open-source Python cheminformatics packages `RDKit` (v. 2023.03.3)[53] and `molfeat` (v. 0.9.2)[54] to use as features of the machine learning models. Using these packages, 29 different methods were chosen for converting molecules into tabular numeric features. These represented the most popular cheminformatics tools for digitizing molecules. Every feature dataset was supplemented with pH-based experimental features: $\mathrm{pH_{before}}$ (pH measurement before the experiments), $\mathrm{pH_{after}}$ (pH measurement after the experiments), $\mathrm{pH_{average}}$ (average of before and after values), and $\mathrm{pH_{bef\text{-}aft}}$ (difference between before and after values). The resulting datasets contained between 18 and 2052 features.

Machine-learning model training and comparison

Targets 107 samples are split into two sets: a training set with 95 samples, and a set-aside validation set with 12 samples. This was achieved through the `verstack` (v. 3.9.2)[55] package using a continuous stratified split so that the target data distributions in both sets are statistically similar.

Features The feature datasets for training samples are cleaned with the help of `scikit-learn` package (v. 1.5.0)[56]. If the training datasets had missing values for any samples, these are filled with the median of that feature. To eliminate redundancies found in the molecular representation, the model features with variances lower than 0.1, and features correlated to others with a Pearson correlation value of more than 0.8 are removed. Afterwards, features are scaled with three different `scikit-learn` scaler functions to assist the model learning process: `MinMaxScaler`, `StandardScaler`, `PowerTransformer`. Features with and without scaling are algorithmically selected with two different sparse feature selection methods: one based on the RFE method of `scikit-learn` which uses impurity-based feature importance on RF estimators (RFE), and the other based on RFE of the `Probatus` package (v. 3.0.0)[57], which uses SHAP-based feature importance ($\mathrm{RFE_{SHAP}}$). RFE was repeated 1000 times with random seeds, $\mathrm{RFE_{SHAP}}$ used 5-fold randomized cross-validation search. Feature selection was carried out to prevent flooding the model with irrelevant features, as high-dimensionality of the feature space would result in fitting the noise rather than the signal, commonly known as overfitting. Another reason was to capture only the features most relevant to the mechanism of corrosion inhibition, which is expected from highly predictive features. The top ten selected features from RFE, or the optimum number of selected features revealed from $\mathrm{RFE_{SHAP}}$ were used for the actual models.

Models After scaling and feature selection, different featurization schemes are combined with different target representations to be modeled with four different regression architectures, three implemented in `scikit-learn`: `random forest`[58], `support vector machine`[59], `k-nearest neighbors`[60], and `xgboost` implemented in the `xgboost` package (v. 1.7.6)[61]. The optimization scoring function used was negative root mean squared error. Bayesian optimization was employed using the `bayes-opt` package (v. 1.4.3)[62] to find the optimal hyperparameters based on a 10-fold cross-validation score, where the data is divided into 10 random subsets. In each iteration, 9 subsets are used for training, and the remaining 1 subset is used for testing, allowing the model's generalization performance to be evaluated across different train-test splits. Optimized hyperparameters and their ranges were:

- Random forest: number of trees (10, 1000), maximum tree depth (1, 50), minimum number of samples required to split (2, 25), maximum ratio of used features: (0.1, 1)
- Support vector machine: regularization parameter C (0.001, 1000), the margin of tolerance ϵ (0.001, 10), kernel coefficient γ (0.001, 100), radial basis function kernel
- K-nearest neighbors: number of neighbors (1, 10), weighing for the neighbors (uniform or distance-weighed), the distance metric to be used for calculating 'neighborhood' (Euclidean, Manhattan, or Minkowski)
- XGBoost: number of trees (100, 1000), maximum tree depth (2, 10), learning rate (0.01, 0.1), fraction of the training data to be randomly sampled for tree construction (0.1, 1.0), fraction of features randomly sampled for tree construction (0.1, 1.0), minimum loss reduction gamma required for further leaf node partition (0.1, 1.0), L1 regularization (0.001, 100), L2 regularization (0.001, 100)

Detailed explanation of hyperparameters can be found in the `scikit-learn` documentation and textbooks[63]. Learning curves and prediction plots are recorded for further analysis. Regression performance was quantified with R^2, RMSE, and MAE. After quantification, models are retrained with all training set with optimized hyperparameters and saved as pickle files for further experiments.

Visualizing fingerprints

atompair-count, rdkit-fp, and ECFP fingerprints are analyzed to identify which molecular fragments the features correspond to, and further visualized through the code provided.

Generating best pseudomolecules through Bayesian optimization

The retrained models are optimized with Bayesian optimization. Now, the optimized parameters were not the model hyperparameters, but the model input values of algorithmically selected features and the predictions of the selected model. The acquisition function used was upper confidence bound with the default implemented hyperparameters. The bounds for optimization for each feature were set based on the minimum and maximum values observed in the original database, ensuring interpolation rather than extrapolation. On top of all initial real molecule samples, 2000 random samples were used to initialize the optimization, and 1000 iterations were performed for optimizing the pseudomolecule. Pseudomolecule feature scaling is inverted for further use for similarity analysis. The resulting features represent the optimal artificial molecule parameters according to the model, leading to the best target property and creating an ideal pseudomolecule.

Curating the toxicity dataset for pseudomolecule similarity hits

To find the molecules most similar to the pseudomolecule, a query molecule database is necessary. A database with a large collection of SMILES strings and experimental toxicity values was chosen as the candidate database[35]. The choice of selecting a toxicity database was deliberate. Aside from predictions from our model, such a database would provide information on the toxicity of compounds. The training and evaluation sets from Supplementary Table 2 of the original study were combined, and the SMILES strings along with their experimentally determined U.S. Environmental Protection Agency (EPA) toxicity hazard classifications were extracted. The EPA classifications corresponded to: I highly toxic, II moderately toxic, III slightly toxic, IV practically non-toxic[36]. After desalting the molecules, generating descriptors, and cleaning the dataset, this process yielded over 10,000 candidate molecules. The similarity between molecules was calculated with cosine similarity $S_{\cos}$, where the similarity between two vectors is calculated as:

$$S_{\cos}(\mathbf{M},\mathbf{P}) = \frac{\mathbf{M}\cdot\mathbf{P}}{||\mathbf{M}||||\mathbf{P}||} = \frac{\sum_{i=1}^{n} M_i P_i}{\sqrt{\sum_{i=1}^{n} M_i^2}\cdot\sqrt{\sum_{i=1}^{n} P_i^2}}, \quad (6)$$

where M_i and P_i are the i^{th} components of vectors $\mathbf{M}$ and $\mathbf{P}$, respectively, corresponding to the vector of the query molecule and the optimized pseudomolecule, respectively. The resulting cosine similarity values span from 1 to −1, from 1 meaning the vectors are oriented in the same direction (complete similarity), to −1 meaning vectors are oriented in the opposite direction (complete dissimilarity), and 0 indicating orthogonal vectors (decorrelation). In-between values indicate intermediate similarity/dissimilarity.

SHAP (SHapley Additive exPlanations) analysis

SHAP values are a concept from cooperative game theory used to fairly distribute the *payout* among players based on their contributions. In machine learning, each feature value of the instance is a player in a game where the prediction is the payout. SHAP values are applied to interpret complex models by attributing the contribution of each feature to the model's prediction for a specific instance. `SHAP` package (v. 0.42.1)[64] was used to create SHAP beeswarm plots for optimized models based on 3 different featurization methods: atompair-count, PaDEL, and MACCS.

Validation experiments through electrochemical measurements

AA2024-T3 sheets with a thickness of 2 mm (Salomon's Metalen B.V., the Netherlands) were cut into 20 × 20 mm specimens using an automatic shear cutter. The samples were then sequentially ground with 320, 800, 1200, 2000, and 4000 grit papers on a rotating plate sander under running water, followed by cleaning in isopropanol for 15 min and drying with compressed air. The resulting specimens were used for electrochemical measurements. The electrochemical investigations consisted of observing the OCP for 24 h, where a linear polarization resistance (LPR) measurement was performed every hour to observe the time-dependent behavior. For LPR measurements, the potentials were scanned from −10 to +10 mV vs. OCP at a rate of 0.5 mV/s. After concluding the 24-h observation, EIS measurements were performed, where a 10 mV peak-to-peak amplitude sinusoidal AC perturbation was applied from 10 to 10 mHz frequency range with 10 frequency points per logarithmic decade. Flat three-electrode electrochemical cells (Corrtest Instruments, China) were used to perform the experiments at room temperature. The sample was used as the working electrode, platinum mesh was used as the counter electrode, and Ag|AgCl (saturated KCl) was used as the reference electrode. The exposed surface area was 0.785 cm^2, exposed to a 250 ml 0.1 M NaCl 1 mM 2-thiobarbituric acid electrolyte. The pH of the electrolyte was adjusted to 7.0 with an adequate amount of NaOH, by analysing the solution pH with a Metrohm 913 pH meter. All chemicals were purchased from Sigma-Aldrich. The electrochemical measurements were controlled with Biologic VSP-300 multichannel potentiostats with the help of EC-Lab software. The electrochemical experiments were repeated three times to confirm the reproducibility of the experiments.

Data availability

The data supporting this study's findings are available as supplementary information.

Code availability

The codes used to calculate the results of this study are provided in the accompanying files.

Received: 6 August 2025; Accepted: 9 November 2025;
Published online: 12 December 2025

References

1. Mater, A. C. & Coote, M. L. Deep learning in chemistry. *J. Chem. Inf. model.* **59**, 2545–2559 (2019).
2. Yano, J. et al. The case for data science in experimental chemistry: examples and recommendations. *Nat. Rev. Chem.* **6**, 357–370 (2022).
3. Karande, P., Gallagher, B. & Han, T. Y.-J. A strategic approach to machine learning for material science: how to tackle real-world challenges and avoid pitfalls. *Chem. Mater.* **34**, 7650–7665 (2022).
4. Rodrigues, J. F., Florea, L., de Oliveira, M. C., Diamond, D. & Oliveira, O. N. Big data and machine learning for materials science. *Discov. Mater.* **1**, 1–27 (2021).
5. Chong, S. S., Ng, Y. S., Wang, H.-Q. & Zheng, J.-C. Advances of machine learning in materials science: ideas and techniques. *Front. Phys.* **19**, 13501 (2024).
6. Choudhary, K. et al. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* **8**, 59 (2022).
7. Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
8. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
9. Reiser, P. et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **3**, 93 (2022).
10. Zhong, Z., Li, C.-T. & Pang, J. Hierarchical message-passing graph neural networks. *Data Min. Knowl. Discov.* **37**, 381–408 (2023).
11. Li, K., DeCost, B., Choudhary, K., Greenwood, M. & Hattrick-Simpers, J. A critical examination of robustness and generalizability of machine

learning prediction of materials properties. *npj Comput. Mater.* **9**, 55 (2023).
12. Krenn, M. et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* **4**, 761–769 (2022).
13. Friederich, P., Krenn, M., Tamblyn, I. & Aspuru-Guzik, A. Scientific intuition inspired by machine learning-generated hypotheses. *Mach. Learn. Sci. Technol.* **2**, 025027 (2021).
14. Zhong, X. et al. Explainable machine learning in materials science. *npj comput. Mater.* **8**, 204 (2022).
15. Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E. & Hoffmann, H. Explainability methods for graph convolutional neural networks. In *Proc. the IEEE/CVF conference on computer vision and pattern recognition*, 10772–10781 (IEEE, 2019).
16. Jin, W., Barzilay, R. & Jaakkola, T. Multi-objective molecule generation using interpretable substructures. In *Proc. International conference on machine learning*, 4849–4859 (PMLR, 2020).
17. Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Acc. Mater. Res.* **3**, 597–607 (2022).
18. Vu, T.-S. et al. Towards understanding structure–property relations in materials with interpretable deep learning. *npj Comput. Mater.* **9**, 215 (2023).
19. Pham, T. H. et al. A data-driven QSAR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques. *RSC Adv.* **14**, 11157–11168 (2024).
20. Ribeiro, M. T., Singh, S. & Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144 (ACM, 2016).
21. Shapley, L. S. A value for n-person games. *Contrib. Theory Games* **2**, 307–317 (1953).
22. Diao, Y., Yan, L. & Gao, K. Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features. *Mater. Des.* **198**, 109326 (2021).
23. Winkler, D. A. et al. Impact of inhibition mechanisms, automation, and computational models on the discovery of organic corrosion inhibitors. *Prog. Mater. Sci.* 101392 (2024).
24. Kokalj, A. et al. Simplistic correlations between molecular electronic properties and inhibition efficiencies: do they really exist? *Corros. Sci.* **179**, 108856 (2021).
25. Özkan, C. et al. Laying the experimental foundation for corrosion inhibitor discovery through machine learning. *npj Mater. Degrad.* **8**, 21 (2024).
26. The European Parliament and the Council. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, amending Directive 199 (2006).
27. Bender, R. et al. Corrosion challenges towards a sustainable society. *Mater. Corros.* **73**, 1730–1751 (2022).
28. Allouche, A.-r Software news and updates Gabedit—a graphical user interface for computational chemistry software. *J. Comput. Chem.* **32**, 174–182 (2012).
29. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).
30. Smith, D. H., Carhart, R. E. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73 (1985).
31. Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* **C-22**, 1025–1034 (1973).
32. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
33. Prakashaiah, B., Kumara, D. V., Pandith, A. A., Shetty, A. N. & Rani, B. A. Corrosion inhibition of 2024-T3 aluminum alloy in 3.5% NaCl by thiosemicarbazone derivatives. *Corros. Sci.* **136**, 326–338 (2018).
34. 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).
35. Gadaleta, D. et al. SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. *J. Cheminform.* **11**, 1–16 (2019).
36. U.S. National Archives and Records Administration. Code of federal regulations, protection of environment, title 40, sec. 156.62 (2006). https://www.ecfr.gov/current/title-40/part-156/section-156.62.
37. York, J. T., Bar-Nahum, I. & Tolman, W. B. Copper–sulfur complexes supported by n-donor ligands: Towards models of the Cuz site in nitrous oxide reductase. *Inorg. Chim. Acta* **361**, 885–893 (2008).
38. Hughes, A. E., Parvizi, R. & Forsyth, M. Microstructure and corrosion of aa2024. *Corros. Rev.* **33**, 1–30 (2015).
39. 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).
40. Kosari, A. et al. Dealloying-driven local corrosion by intermetallic constituent particles and dispersoids in aerospace aluminium alloys. *Corros. Sci.* **177**, 108947 (2020).
41. Ganjoo, R. & Kumar, A. Current trends in anti-corrosion studies of surfactants on metals and alloys. *J. Bio- Tribo-Corros.* **8**, 1–35 (2022).
42. Abdelmonem, H., Al-Bonayan, A. M. & Fouda, A. E.-A. S. Some surfactants as corrosion inhibitors for carbon steel in acidic solutions. *Surf. Eng. Appl. Electrochem.* **58**, 412–423 (2022).
43. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (John Wiley & Sons, 2009).
44. Barysz, M., Jashari, G., Lall, R. S., Srivastava, V. K. & Trinajstic, N. On the distance matrix of molecules containing heteroatoms. Chemical applications of topology and graph theory. *Studies in physical and theoretical chemistry* (1983).
45. Levitt, M. & Perutz, M. F. Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol.* **201**, 751–754 (1988).
46. Huang, S. et al. Hydrogen bond induces hierarchical self-assembly in liquid-crystalline block copolymers. *Macromol. Rapid Commun.* **39**, 1700783 (2018).
47. Sikder, A. & Ghosh, S. Hydrogen-bonding regulated assembly of molecular and macromolecular amphiphiles. *Mater. Chem. Front.* **3**, 2602–2616 (2019).
48. Manish Sud. MACCS (Molecular ACCess System) documentation (2024). http://www.mayachemtools.org/docs/modules/pdf/MACCSKeys.pdf.
49. Özkan, C. et al. Quasi-stable adsorption as a stepping stone to stable corrosion inhibition. *Appl. Surf. Sci.* **712**, 164060 (2025).
50. Pourbaix, M. Atlas of electrochemical equilibria in aqueous solutions. *NACE* (1966).
51. Barsoukov, E. & Macdonald, J. R. (eds.) *Impedance Spectroscopy: Theory, Experiment, and Applications* (John Wiley & Sons, 2005), second edn. https://doi.org/10.1002/0471716243.
52. Taheri, P. et al. On the importance of time-resolved electrochemical evaluation in corrosion inhibitor-screening studies. *npj Mater. Degrad.* **4**, 1–4 (2020).
53. Landrum, G. et al. RDKit: Open-source cheminformatics (2020). https://www.rdkit.org.
54. Noutahi, E. et al. datamol-io/molfeat (2023). https://github.com/datamol-io/molfeat.
55. Zherebtsov, D. Verstack. https://github.com/DanilZherebtsov/verstack (2020).
56. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

57. Koops, R. Probatus: Validation (like Recursive Feature Elimination for SHAP) of (multiclass) classifiers & regressors and data used to develop them. https://github.com/ing-bank/probatus (2023).
58. Ho, T. K. Random decision forests. In *Proc. 3rd international conference on document analysis and recognition*, vol. 1, 278–282 (IEEE, 1995).
59. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
60. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
61. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proc. the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, New York, NY, USA, 2016). https://doi.org/10.1145/2939672.2939785
62. Nogueira, F. Bayesian Optimization: open source constrained global optimization tool for Python (2014). https://github.com/bayesian-optimization/BayesianOptimization.
63. Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (O'reilly, 2019).
64. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. et al. (eds.) *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017). http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Acknowledgements

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

Author contributions

Can Özkan: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing—original draft. Lisa Sahlmann: Writing—review and editing. Tim Würger: Writing—review and editing. Christian Feiler: writing—review and editing. Sviatlana Lamaka: Writing—review and editing. Mikhail Zheludkevich: Writing—review and editing. Peyman Taheri: Writing—review and editing, supervision. Arjan Mol: Resources, writing—review and 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/s41529-025-00713-4.

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