TU Delft

Delft University of Technology

Random forest incorporating ab-initio calculations for corrosion rate prediction with small sample Al alloys data

Ji, Yucheng; Li, Ni; Cheng, Zhanming; Fu, Xiaoqian; Sun, Xiaoguang; Chowwanonthapunya, Thee; Zhang, Dawei; Ren, Jingli; Dey, Poulumi; Dong, Chaofang

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

ARTICLE   OPEN

Check for updates

Random forest incorporating ab-initio calculations for corrosion rate prediction with small sample Al alloys data

Yucheng Ji[1,2,3], Ni Li[1,3], Zhanming Cheng[1,3], Xiaoqian Fu[1,3], Min Ao[1,3], Menglin Li[1,3], Xiaoguang Sun [ID][4], Thee Chowwanonthapunya[5], Dawei Zhang [ID][1,3], Kui Xiao [ID][1], Jingli Ren [ID][6], Poulumi Dey[2], Xiaogang Li[1,3] and Chaofang Dong[1,3 ✉]

Corrosion jeopardizes the materials longevity and engineering safety, hence the corrosion rate needs to be forecasted so as to better guide materials selection. Although field exposure experiments are dependable, the prohibitive cost and their time-consuming nature make it difficult to obtain large dataset for machine learning. Here, we propose a strategy Integrating Ab-initio Calculations with Random Forest (IACRF) to optimize the model, thereby estimating the corrosion rate of Al alloys in diverse environments. Based on the thermodynamic assessment of the secondary phases, the ab-initio calculation quantities, especially the work function, significantly improved the prediction accuracy with respect to small-sample Al alloys corrosion dataset. To build a better generic prediction model, the most accessible and effective features are identified to train IACRF. Finally, the independent field exposure experiments in Southeast Asia have proven the generalization ability of IACRF in which the average prediction accuracy is improved up to 91%.

INTRODUCTION

Corrosion is the most severe ordeal that materials are confronted with and negatively affects engineering life and safety[1]. Due to the environmental diversity and service complexity, various corrosion, such as pitting[2], galvanic corrosion[3], and stress corrosion cracking[4] (SCC), have occurred in engineering structural materials. Aluminum (Al) alloys, the most widely used lightweight metal in aviation and transportation, suffer from corrosion problems[5]. Although their application reduces fossil energy consumption, their degradation is not to be neglected[6]. Currently, there are two solutions for monitoring metal corrosion. One is deploying corrosion sensors online to transmit the corrosion current in the real environment[7]. However, its obvious problem is that the transmission is unreliable. Once the protective film or corrosion products are formed on the surface, the corrosion current may be extremely noisy and even lose its functionality, especially for Al alloys. In addition, another trustworthy solution is field exposure experiments offline in the actual environment for a period of time (years) and then collect the samples to calculate the corrosion rate[8]. This solution, although highly dependable, is also costly; any prolonged field exposure time increases the risk of losing the sample.

Several research projects have been conducted to study the intrinsic essence of material corrosion under long-term field exposure in different areas. For instance, Sun reported that the corrosion rates of AA2024 and AA7075 in industrial environments were larger than those in coastal and urban environments[9]. However, corrosion is a complex process that is influenced by many factors, such as composition and environmental parameters[10]. In view of the limited capabilities of traditional modeling, only minority factors are considered[11]. With the development of big data, machine learning (ML) has shown significant advantages in data mining and forecasting[12]. Numerous studies have used miscellaneous ML algorithms to fit corrosion data and compare their pros and cons[13–17]. For the noisy data generated by the corrosion sensor in the first solution, Pei confirmed that the random forest (RF) is more suitable for corrosion data than the artificial neural network and supporting vector regression[18]. Despite the considerable noisy data, corrosion sensors which are mainly used for steel generate massive data for ML training. Evidently, reliable field exposure experiments lack data due to prohibitive acquisition costs. Predicting the corrosion rate of Al alloys using a small-sample dataset combined with environmental parameters is still a major obstacle. Establishing ML models with outstanding generalization ability will be of great significance to prolong the service life of materials and ensure engineering safety[19].

We therefore propose a strategy that utilizes the ab-initio calculations to describe Al alloys intermetallic compounds (IMCs) and the precipitated phases, so as to improve the prediction accuracy of the Al alloys corrosion rate on the Earth with latitude ranging from South 6° to North 52° (See Fig. 1). Compared with the traditional dataset containing composition and heat treatment, the effect of secondary phases in Al alloys on corrosion rate is more comprehensive depicted by the calculated work functions (WFs), surface energies, and Cl adsorption energies (Cl AEs). By introducing the calculated phases descriptor, the generalization ability of the corrosion RF model was greatly improved. Besides, the model had been verified by the independent field exposure experiments in Southeast Asia (Singapore, Jakarta, and Bangkok), where the annual rainfall and particle composition were beyond the range of the training set. Based on this computation and experiment combined model, we deduced what statistical strategy should be employed when the ab-initio calculation is

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, National Materials Corrosion and Protection Data Center, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing, China. [2]Department of Materials Science and Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, the Netherlands. [3]Shunde Innovation School, University of Science and Technology Beijing, Foshan, China. [4]Technical Engineering Department, CRRC Qingdao Sifang Co. Ltd, Qingdao, China. [5]Faculty of International Maritime Studies, Kasetsart University, Chonburi, Thailand. [6]Henan Academy of Big Data, Zhengzhou University, Zhengzhou, China. ✉email: cfdong@ustb.edu.cn
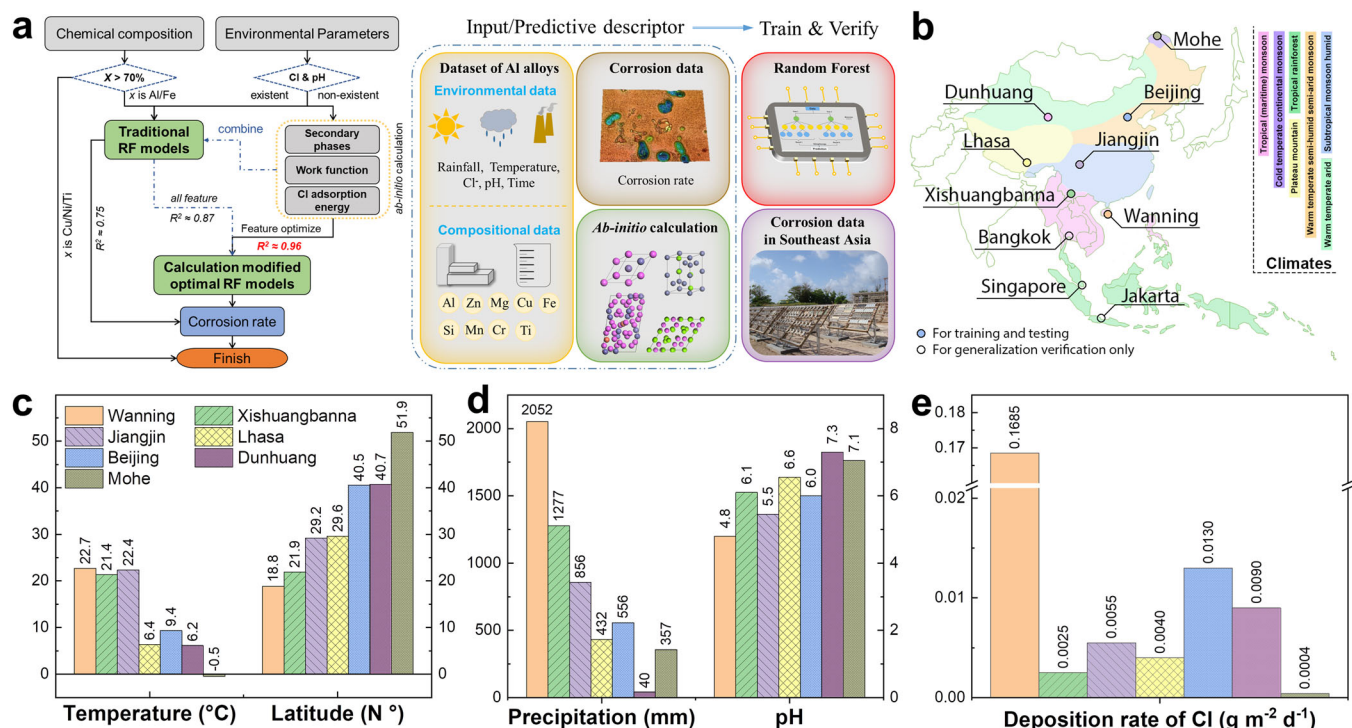
Fig. 1 Calculations modified RF model flow and the field exposure sites. a ML flow diagram and dataset features. b Placement map. Corrosion data obtained in Singapore, Jakarta, and Bangkok are only used for generalization verification. c–e The climate characteristics of regions. The field exposure time of Al alloys is range from 0.5 to 4 years.
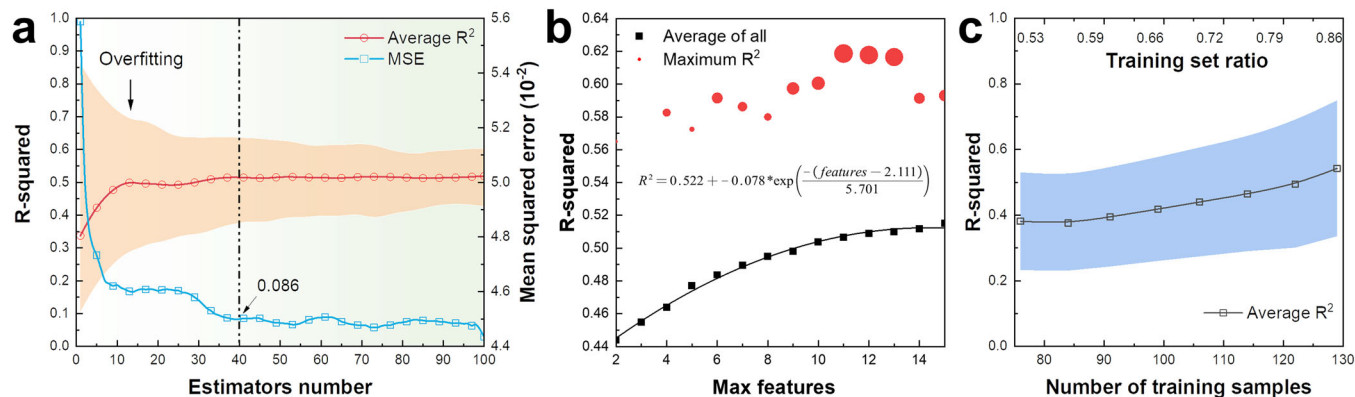

Fig. 2 Analysis of long-term Al alloys field exposure corrosion data by RF model. a The effect of the estimator number on R^2 and MSE. The R^2 improvement with b features and c training set ratio variation. The red symbol size clearly represents the precision of the model.

utilized to express the corrosion resistance of various phases that coexist in Al alloys.

RESULTS

Parsing small-sample Al alloys field exposure corrosion data

To evaluate the characteristics of the Al alloys corrosion data, the effect of the basic RF parameters was first investigated. It can be seen from Fig. 2a that the average goodness-of-fit (R^2) increased with the estimator rise. The growth of the model average accuracy was limited to 0.81% when the estimator exceeded 40. Nevertheless, the model R^2 was anomalous when the estimator was lower than 40. Traditional optimal RF model accuracy was greatly improved with the decrease of the estimator, where the maximum R^2 reached 0.86. Unlike the average accuracy increment, the maximum R^2 of the optimal model slightly dropped by 1.22%. A low estimator was generally considered to be overfitting the data.

Combined with the mean squared error (MSE), it can be confirmed that the error of the model whose estimator was less than 30 was significantly higher than the average error. After the estimator amount exceeded 40, the MSE of the model dropped to 0.086. In addition, the minimum MSE occurred at the model with 100 estimators which was only 0.0443. Taken as a whole, the estimator of RF models for small-sample corrosion dataset must be greater than 40, the optimal value should be 100.

Realistically, our model accuracy still lags far behind the material mechanical prediction models[20,21]. To improve the model accuracy, we therefore deeply analyzed the features and data distribution of corrosion dataset. When the maximum feature number exceeded 10, the models average R^2 remained around 0.51 (Fig. 2b). Interestingly, superfluous features did not positively affect the establishment of the optimal corrosion model (red symbol in Fig. 2b). Only when the maximum feature number was 11, the traditional RF model achieved the optimum ($R^2 \approx 0.62$).
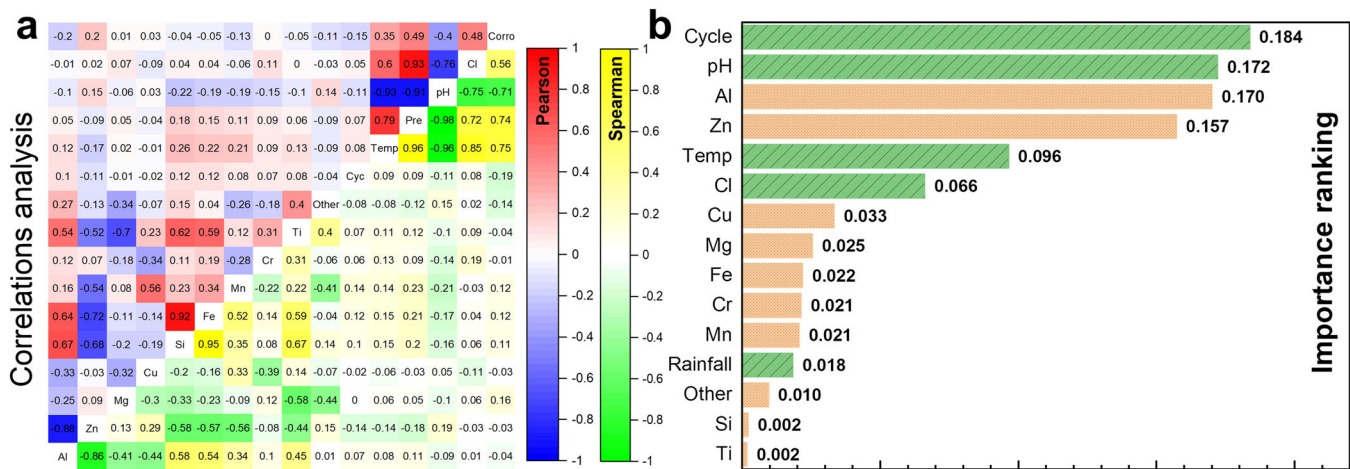
Fig. 3 Analysis of long-term exposure experiment data. a Pearson and Spearman correlation coefficient. **b** Importance ranking of corrosion data features based on the RF model.
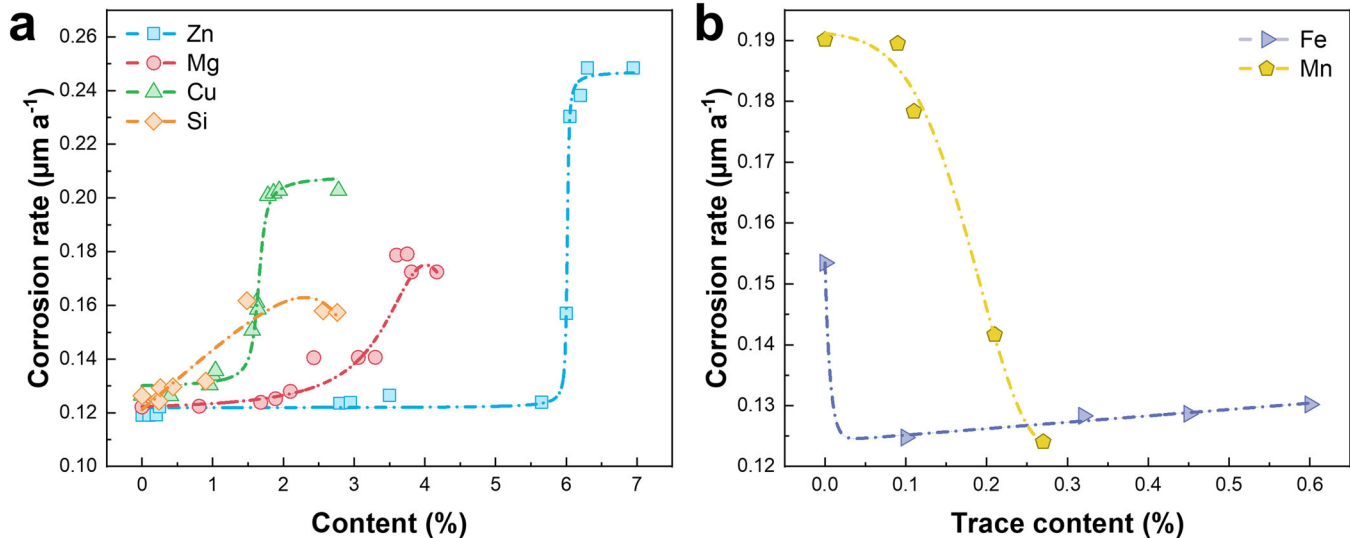

Fig. 4 Elaboration of element effect on corrosion rate. a Common elements. **b** Trace elements.

This proved that there were features that are unfavorable for corrosion rate prediction in Al alloys. Whereas the slow increase of average precision (after max features = 10) did not account for the unfavorable features number being 5. Another solution to improve the accuracy is to enlarge the proportion of the training set or multiply the data size. Figure 2c shows that with the increase in training set proportion, the model accuracy had been greatly improved (0.75 maximum, 24% higher than the original model).

Both the Pearson and Spearman's correlation coefficients were carried out to analyze the linear and monotonic relationships between the corrosion data, so as to eliminate the useless features (Fig. 3a). There are two isolated regions in composition that show a strong correlation, one is the correlation between Al and Zn (exceed 0.8). Specifically, Zn is demonstrated to be a strengthening element in AA7xxx to form the η/η' phase[22]. However, owing to the high SCC susceptibility[23] of AA7xxx, other elements such as Cu, Fe, and Si are supplemented to enhance its corrosion resistance. Therefore, Zn addition resulting the synchronous increase of other alloying elements so as to decrease the Al content. Another strong correlation is seen between Si and Fe which are often co-added to enhance the alloys strength and manufacturing fluidity[24]. Moreover, the correlation of annual

rainfall with precipitation pH and Cl deposition were both above 0.9. This indicated that the annual rainfall could replace other environmental parameters during training. Besides, Fig. 3a also demonstrated the environmental parameters were more directly related to corrosion rate than alloy composition, as their coefficient with corrosion rate is 0.3–0.5. We investigated the importance of the features in the corrosion dataset through the traditional RF model (the training set ratio was 0.7 and the R^2 was 0.62). A total of 6 features had an importance index exceeding 0.05 in Fig. 3b, which were exposure time, precipitation pH value, Al content, Zn content, annual average temperature, and Cl deposition, respectively. The trace elements did not have a significant impact (<0.03) on the corrosion prediction.

By means of the RF model, the effect of alloying elements on corrosion rate could be investigated based on the commonly used AA7B05 for underframe corbels. Each element's effect on corrosion rate was reflected by the monotonical variation which was summarized in Fig. 4. It could be found that the alloy corrosion rate fluctuated slightly when Zn content was less than 5.7%. Nevertheless, once the addition of Zn exceeded 6%, the alloys corrosion rate increased by 100.67%. What's more, both superfluous Mg and Si are observed to deteriorate the corrosion resistance. Likewise, Cu displayed a surge in corrosion rate, but its
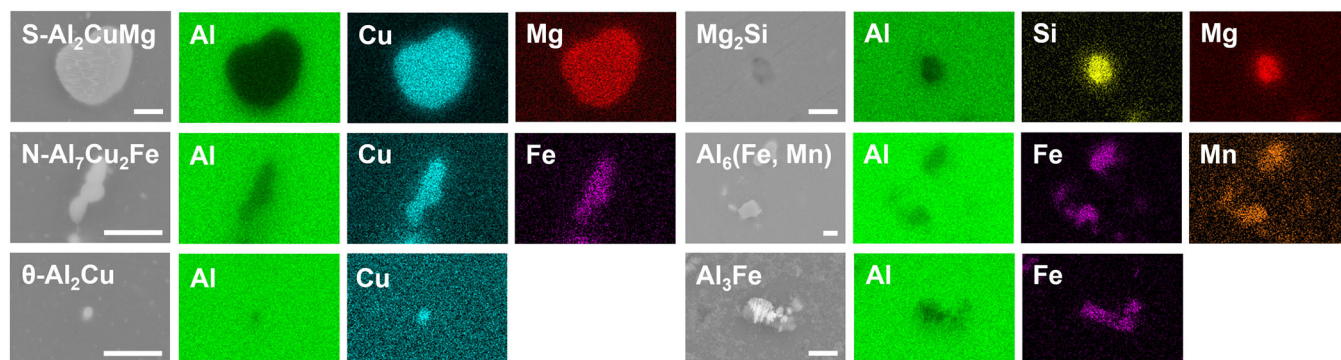
Fig. 5 Micron-scale IMCs or precipitated phases in Al alloys. 2xxx (θ-Al$_2$Cu, S-Al$_2$CuMg, N-Al$_7$Cu$_2$Fe, Mg$_2$Si), 5xxx (Mg$_2$Si, Al$_6$(Fe, Mn)), 6xxx (Mg$_2$Si, S-Al$_2$CuMg, Al$_6$(Fe, Mn)), and 7xxx (Al$_3$Fe, Al$_{23}$Fe$_4$Cu, Mg$_2$Si) Al alloys. Excluding the micro-scale secondary phases, the 5xxx Al alloys also contain β-Al$_3$Mg$_2$ and 7xxx Al alloys have η-MgZn$_2$ phase. The length of bars is 2 μm.
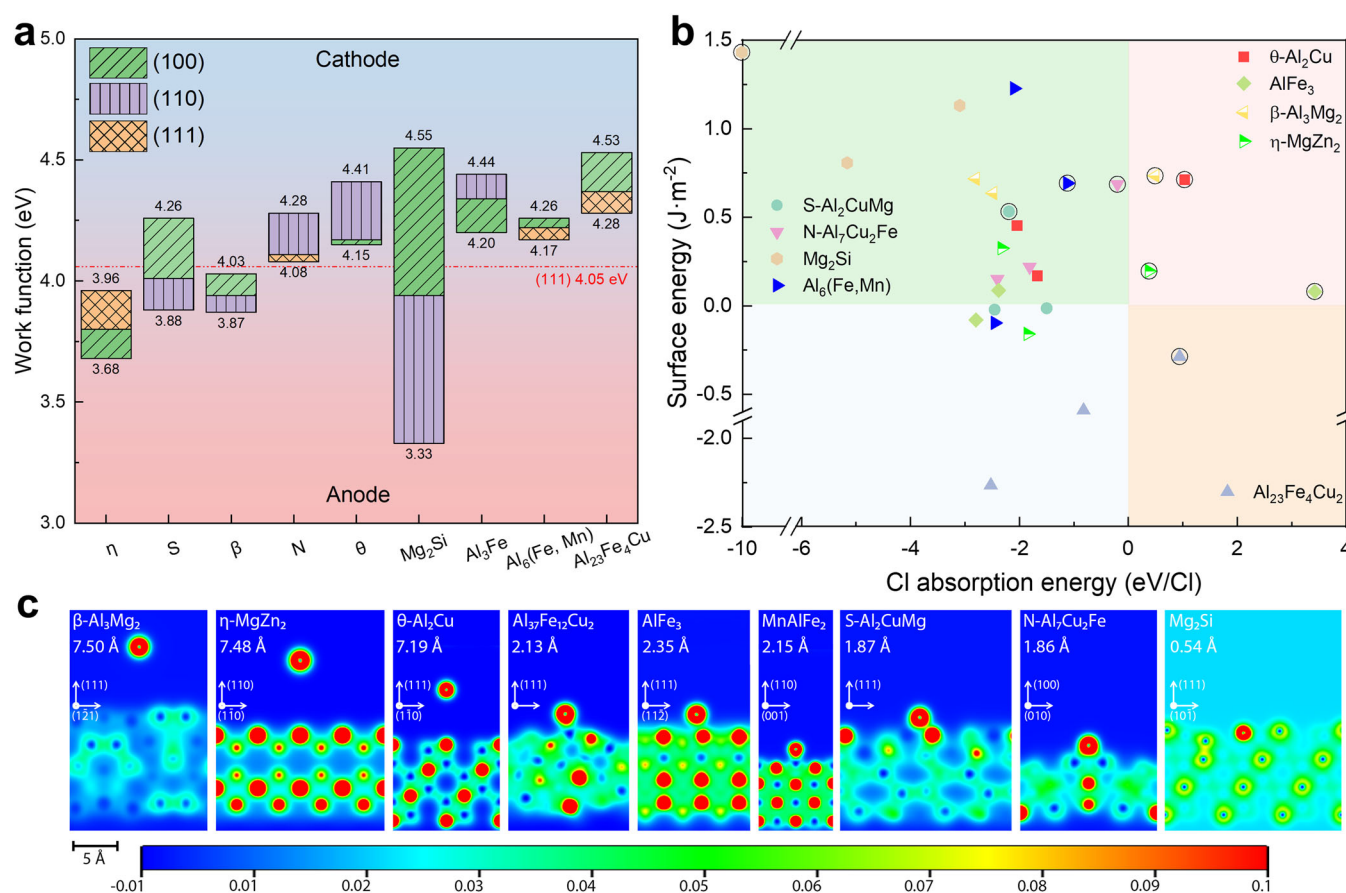

Fig. 6 Ab-initio calculations for the Al alloys secondary phases. a WF. Three orientations, (100), (110), and (111), were considered. The lowest WF is depicted as a line, and the remaining two surfaces are expressed in the form of summation on this line. **b** The phase surface energies and their adsorption energies for the Cl atom. The surface with the lowest WF of each phase is surrounded by a black circle. **c** Adsorption distance and charge distribution between the Cl atom and the phase surface with the lowest WF.

content required was small (>1.5%). On the contrary, Fe and Mn could improve the Al alloys corrosion resistance (Fig. 4b). Usually, Fe is considered as impurity element for mechanical properties in Al alloys. However, it can be seen from ML that a small amount of Fe can significantly improve the corrosion resistance (although excessive addition would increase the corrosion rate). This was mainly attributed to the low formation energy of the Fe-phases, redundant Fe was easy to form coarse phases with high WFs which lead to galvanic corrosion. Mn also improved the corrosion resistance, but unlike Fe, the increase of Mn did not substantially

reduce their corrosion resistance. In general, the gigantic effect of Zn on the corrosion rate of Al alloys can also be confirmed by materializing to specific alloys. Therefore, combining the correlation of Zn with Si, Fe, Mn, and Ti (Fig. 3a), it is feasible to predict the corrosion rate by optimizing the features (using Zn). However, it is unreasonable to only consider Zn, especially for Al alloys without Zn, so the Al content should also be considered. Although optimizing the features could improve the RF model accuracy, it is not optimal (up to 0.75).

Ab-initio calculations for the secondary phases

Complex secondary phases (including IMCs and precipitated phases) were another important characteristic in Al alloys, which were experimentally observed as shown in Fig. 5. Whereas how to obtain the phases information efficiently and quickly is a conundrum. The experimental means is troubled by the consequences of tiny element changes in the phases. However, the corrosion information of the secondary phases varied by the exposed element (type, positions, and content) can be considered as much as possible by calculations[25].

WF, surface energy, and Cl AE of the secondary phases with different orientation were calculated to enrich the phase information in the corrosion dataset. The most easily formed Al surface was (111) and its WF was 4.05 eV[26]. The theoretical WF difference (Eq. (13)) between the various phases and (111) surface can qualitatively determine the role of phases in the corrosion process: anode or cathode[27,28]. To sum up, η-MgZn$_2$ and β-Al$_3$Mg$_2$ belonged to the anode phase in Al alloys, the N-Al$_7$Cu$_2$Fe, θ-Al$_2$Cu, Al$_3$Fe, Al$_6$ (Fe, Mn), and Al$_{23}$Fe$_4$Cu were the cathode phase (see Fig. 6a). In particular, the corrosion tendency of Mg-containing phases, such as S-Al$_2$CuMg and Mg$_2$Si, was strongly correlated with the crystallographic orientation and the atomic type of the outermost termination. Moreover, the difference between the minimum/maximum WF of the Mg$_2$Si and that of Al was too large to cause corrosion. When Mg$_2$Si acts as an anode, the Al alloys tend to have pitting corrosion[29]. Conversely, the Al alloys will undergo galvanic corrosion when Mg$_2$Si enacts as a cathode. Fortunately, the surface energy of the most easily formed Mg$_2$Si surface (100) was 0.81 J m^{-2} which indicated that the Mg$_2$Si size was not enormous (experimentally measured close to 1 μm). Meanwhile, the (100) WF was 4.55 eV so that the Al alloys containing Mg$_2$Si were relatively less prone to galvanic corrosion.

Hazardously, all orientation surface energies of Al$_{23}$Fe$_4$Cu were negative (the lowest value among them being −2.87 J m^{-2}), which meant that large Al$_{23}$Fe$_4$Cu can be formed in Al alloys. Combined with the influence of its high WF, it was easy to cause galvanic corrosion. Whereas the positive surface energies did not mean that the phase cannot be precipitated/formed in the Al matrix, it just indicated that the surface formation process was a non-spontaneous reaction. In addition, except for the Al$_{23}$Fe$_4$Cu, the surface energies of the other eight phase orientations with lowest WF were all positive. Most of the phase orientation (>81.48%) showed a clear affinity with respect to Cl. The lowest WFs of β-Al$_3$Mg$_2$, θ-Al$_2$Cu, η-MgZn$_2$, AlFe$_3$ and Al$_{23}$Fe$_4$Cu$_2$ had poor affinity for the Cl atom, so there was no obvious Cl additive effect[30]. Furthermore, the adsorption distance and charge distribution shown in Fig. 6c clearly divided the secondary phases into three categories. The first category was the precipitated phase that interacts weakly with the Cl atom, for instance the β-Al$_3$Mg$_2$, θ-Al$_2$Cu, and η-MgZn$_2$. The distance between the

outermost surface of phase and the Cl atom was larger than 7 Å, and there was no obvious charge transfer at the intermediate points. The second category was the IMCs that generate strongly interacting ionic bonds with the Cl atom. Such phases were IMCs containing Fe and S-Al$_2$CuMg. The distance between the Cl atom and the outermost atoms was 2 Å, and there was an apparent charge sharing. Finally, the worst phase was the Mg$_2$Si, where the Cl atom had visibly eroded into the phase interior (distance <1 Å). Based on the above calculations, we numerically describe the role of the secondary phase in the corrosion process of Al alloys.

Facilitation and generalization for syncretic RF model

To establish a more accurate corrosion prediction model, various statistical methods were applied to handle the Al secondary phase information (WFs and Cl AEs) to refine the corrosion dataset, namely the Integrating Ab-initio Calculations with Random Forest models (IACRF). Top six important features (in Fig. 3) were utilized to further compare the statistical method differences. Figure 7 specifically shows the effect of the WF and Cl AE descriptor on the model accuracy. Evidently, using WFs to amend the corrosion dataset improved the mean prediction accuracy by 43.56 ± 1.03% (Fig. 7a). Besides, the strategies using the arithmetic mean method (AMM) had extremely similar improvements in accuracy as their mathematical expectations difference was less than 0.06%. The optimal model under the AMM appeared when using the average WFs with the lowest surface energy; its accuracy reached 0.861. Then, ignoring the phases type, only the extrema WF was utilized to supplement the corrosion dataset, the model accuracy fluctuated significantly. Especially, the mean accuracy of the model only considering the lowest WF improved to 0.737. On the contrary, the accuracy of the model that only considers the highest WF had dropped drastically, its Gaussian distribution had shifted to the left (see Supplementary Fig. 1). The model using the lowest WF achieved an optimal accuracy of 0.869.

From Fig. 7b, it could be concluded that the precision of the Cl AEs supplementation was significantly behind that of the WF, the average accuracy of the Cl AEs model was 0.630 only. The reason was that it was completely different from the WF which directly pointed out the ability of the secondary phases to gain or lose electrons. Furthermore, another reason is that oxide films are easily formed on the Al surface, the probability of Cl directly contacting with the Al matrix is low. Although the oxide film on the secondary phases may be weak, some of them, such as β-Al$_3$Mg$_2$, θ-Al$_2$Cu, and η-MgZn$_2$, were resistant to Cl adsorption. Similar to WF, the optimal model of Cl AE modification appeared in the extreme value method. Whereas the maximum model accuracy (0.837) was the Cl AE with the lowest surface energy. Besides, it is unadvisable to average the Cl AEs of all orientations for all phases, which significantly masked the pitting sensitive
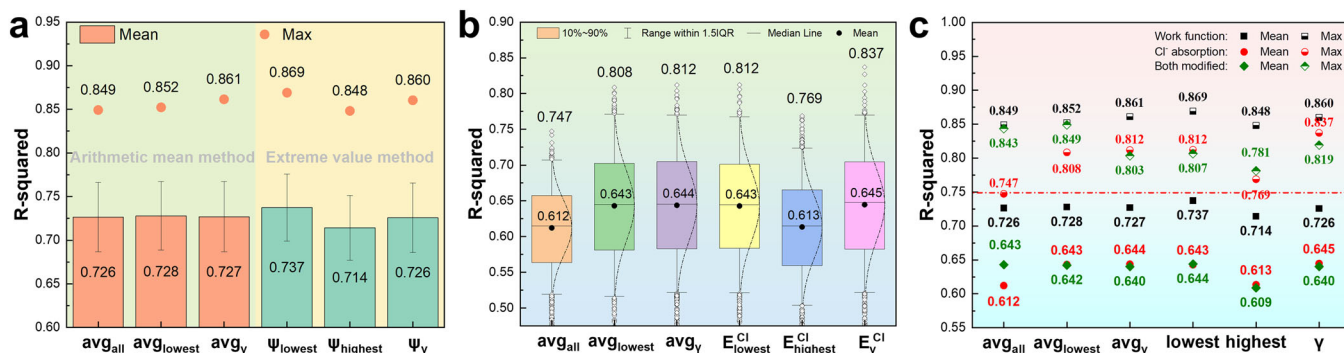

Fig. 7 Contrast of different statistical strategies for calculation modified corrosion dataset. a Mean accuracy and optimal strategies using WF. **b** Statistics of model accuracy after supplementation with the Cl AE. **c** Average and maximum accuracy comparison of models with a single factor (WF or Cl AE) or both the factors.
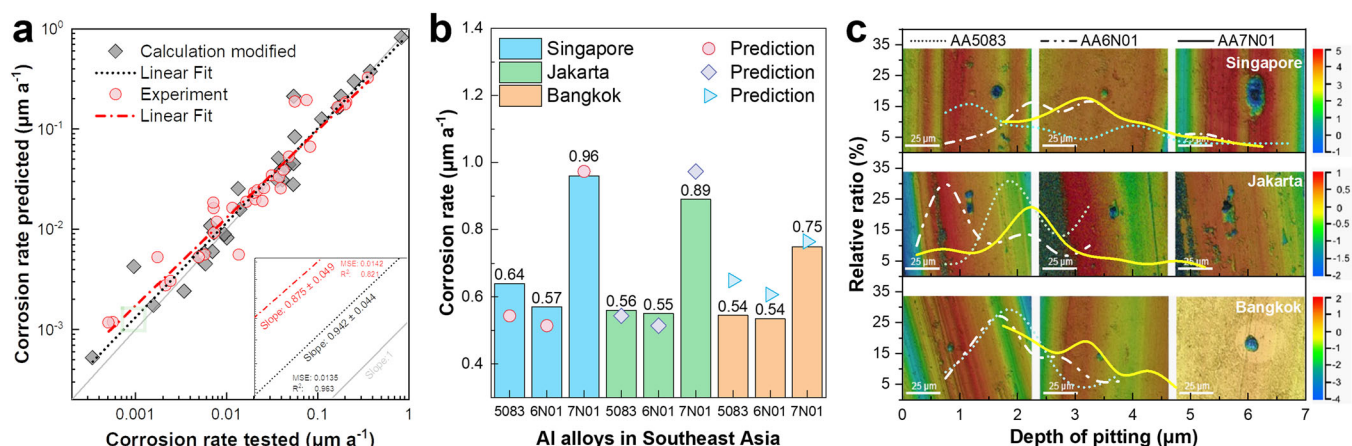
Fig. 8 Independent verification of two feasible models for Al alloys field exposure data. a Comparison of RF models trained by the pure experimental corrosion dataset or ab-initio calculation modified dataset. **b** Three different Al alloys, AA5083, AA6N01, and AA7N01, were placed in Southeast Asia (Singapore, Jakarta, and Bangkok) for 2 years to examine the generalization ability of the RF model with phase calculation. **c** Corrosion morphology and statistics on the number and depth of pits in Al alloys after field exposure.

phases (optimal accuracy of avg_{all} is 0.747, Fig. 7b). From Supplementary Fig. 2, the prediction accuracy of IACRF modified with WF or Cl AE varies slightly with the estimators increased.

A comprehensive comparison of the two ab-initio calculated descriptors and three amendatory methods (WF, Cl AE and both modified) are summarized in Fig. 7c. When the corrosion dataset was supplemented with both WF and Cl AE, the model average accuracy was similar to the Cl adsorption modification. Only the model amended by the average all phases strategy improved to 0.64. In terms of mean accuracy, the model using only WFs was significantly higher than the other two strategies. For the optimal model with the highest accuracy, the model combined with WF and Cl AE partially (in AMM) achieved the accuracy of model using only the WF. Specifically, the differences of the avg and avg_{lowest} strategies with that of WF were only 0.71% and 0.35%, respectively. Further, the accuracy of the other four statistical method models indicated that the Cl adsorption descriptors cancelled the beneficial effect of WF to some extent. Therefore, the single WF descriptor is the most beneficial calculated feature for the field exposure corrosion dataset of Al alloys.

To enhance the generalization ability of the corrosion prediction model, we further trained the model with different feature combinations. Finally, the selected features (exposure time, Al content, annual rainfall, and calculated WF) were retained to train the model. Figure 8a shows two corrosion RF models after repeated tuning: the experimental model and the calculation modified model (IACRF). Since some experimental features were removed from the model, the R^2 of the experimental optimal model dropped to 0.821, where the MSE was 0.0142. Despite the low R^2, in terms of corrosion data prediction, the fitting slope between the experimental data and its prediction value was 0.875 ± 0.049. Encouragingly, the IACRF model had the R^2 as high as 0.963. The MSE of this model dropped to 0.0136, and the slope of linear fitting reached 0.942 ± 0.044. This illustrated that the secondary phases were a non-negligible factor in the process of studying the corrosion rate prediction. Adding the phases calculated descriptors to the corrosion dataset had greatly reduced the prediction difficulty and increased prediction accuracy by 17.30%.

Coincidentally, it can be found from Fig. 8a that the predictions of both the models at a low corrosion rate ($<0.01 \ \mu m \ a^{-1}$) had a larger error than that at a high corrosion rate ($>0.1 \ \mu m \ a^{-1}$). The field exposure corrosion rate decreased with time, which was attributed to the fact that the corrosion rate mentioned here refers to the average annual corrosion rate. The Al alloys had a higher corrosion rate in the first exposure year. Compared with the

increase in time, the amplification of corrosion weight loss was lower, which lead to a decrease in the corrosion rate. When the corrosion rate was converted to weight loss, the corrosion weight loss increased with exposure time. Besides, it was worth noting that the Al content of the AA6N01 was 98.24%, which clearly showed a 1.41% increase from the maximum training set (originally 96.85%). Additionally, the annual rainfall in Singapore (2345 mm, 1.146 after normalization) was also out of range[31]. Although both the compositional and environmental parameters were outside the range of the input values, the RF model trained using the WF retained its accuracy to some extent. After synthesizing the corrosion rates of three Al alloys in Southeast Asia, the overall R^2 predicted reached 0.822 and the MSE was 0.0031. Moreover, the whole average error between the predicted rate and the experimental rate was 8.89%. Comparing the region difference, it was found that the IACRF model had the best prediction ability for the Al alloys placed in Jakarta; its error was only $6.33 \pm 3.04\%$. However, the worst prediction region of the IACRF model was Bangkok as its error was as high as 11.53%. In Singapore, where the rainfall exceeded the dataset range, the accuracy reached $91.20 \pm 7.43\%$. From the experimental results, the corrosion in Singapore was the most serious no matter which Al alloys series was under consideration. The morphology, size, and number of pits are summarized in Fig. 8c. The more and deeper the pits, the greater the risk of an engineering disaster. In addition, the experiment found that although the pits of the three Al alloys had decreased in Singapore, the pits were far deeper than those in the other two regions. Once the materials perforation occurred, the stress concentration caused the rapid failure of the Al alloy structure. While the other two regions had the same magnitude of pitting number, the average pits depth in Bangkok was deeper.

DISCUSSION

By limiting the maximum features of subtree, the effect of the features number on the corrosion model was reported. Unfortunately, we only know that increasing the features number to a certain extent (≈ 10) was conducive to improving the model accuracy. It was not known exactly what features each subtree used. Based on the analysis of features and learning curve, we determined that the corrosion prediction models could be further optimized. Significantly different from the ML models with sufficient data, their learning accuracy curve consistently demonstrated tiny fluctuation with the increase in the data size[32,33]. It is adequately indicated that the field exposure corrosion dataset

belongs to a small-data sample. Due to the time span of the field exposure experiments, performed from 0.5 years to 4 years where at least five samples of the same alloys were required to ensure the experimental accuracy, such high experimental costs obviously cannot expand the data size. However, learning from small-sample data is still a difficult issue[34,35]. In addition to the solution of enlarging the dataset from academic journals[36], another feasible method is to utilize the partial sample generalization error to fine-tune the learner parameters. For instance, Xue provided feedback input for data with large prediction errors and used only 22 samples of data to construct a support vector machine algorithm that successfully fabricated a Fe-Ni based shape memory alloys[37]. Therefore, to further improve the corrosion model accuracy, the small-sample data needs to be deconstructed to reduce harmful features and add new features.

Creatively, we introduced the Al alloys phases descriptor by observing the relationship between corrosion rates and the secondary phases. According to the element distribution and atomic ratio of the second phase, we approximated the corresponding crystallographic models (Supplementary Fig. 3). In addition, it is worth mentioning that it is experimentally difficult to incorporate various phases information within corrosion dataset. Currently, the scanning Kelvin probe force microscopy is a potential means to determining the corrosion property of the secondary phases[38]. However, the cost of this means is exorbitant, the sequence of phase composition detection (carbon injection changes the surface state) and phase surface potential monitoring (priority implementation may result in repeated testing of the same phase) lowers the experimental efficiency. So, the utilization of ab-initio calculation is the optimal way to refining the corrosion dataset. By comparing the model accuracy with different calculations statistics, these results inspired us to consider more the secondary phases with lowest WF where the ab-initio calculations were carried out to determine the Al alloys corrosion resistance. Although the improvement of model accuracy by Cl AE descriptor is weaker than that of WF, it maybe attributed to the natural dense oxide film[39,40] on the Al surfaces (thickness ~10 nm) and the low Cl concentration in atmosphere environment. The effect of Cl on the oxide/passivation film is controversial, mainly including Cl permeation and Cl-induced film thinning theory[41]. Therefore, the Cl AE feature may be more accurate in the training of metal/alloys without oxide/passivation films, i.e., carbon steel and Mg alloys. Furthermore, the prediction accuracy of stainless steel, whose passivation film[42] (~1–2 nm) is extremely sensitive to Cl, may be significantly improved after Cl AE addition. For the thin iron passivation films accompanied by element doping (Cr, Ni, Mo) and structural transformation (Fe_3O_4, Fe_2O_3, $Fe(OH)_2$, and $Fe(OH)_3$), the best Cl calculated descriptor could be the Cl diffusion in the passivation films[43] (especially for Cl permeation systems). The Cl diffusion in oxide-layer feature may has significantly impact on the accuracy when incorporating ab-initio calculations to predict the corrosion rate of stainless steel. In addition, the same secondary phase has many structures with different stoichiometry and crystallographic space groups during heat treatment[44]. Considering more orientations and different terminal atoms on the same orientation is beneficial to describe the physical properties of secondary phases more accurately. These calculations may improve the accuracy of models but is not certain. However, it should be noted that for the calculated descriptor, especially the WF, the range (min, max) of them should include that of metastable phases as much as possible, e.g., Mg_2Si in Supplementary Table 1.

Actually, some environmental parameters, such as Cl deposition and rainfall pH, are needed to be collected using specialized instruments. If some environmental parameters used in the aforementioned models were missing, the accuracy would have been affected. Therefore, our further work was to use readily accessible environmental parameters to train the model. A model

with $R^2 \approx 0.96$ was successfully trained by the combined feature optimization and ab-initio calculations modified corrosion dataset. Although the verification precision of the field exposure experiments in Southeast Asia dropped to 0.822, the overall error was less than 9%. We reckoned that this accuracy drop was attributed to the fact that the parameters in Southeast Asia exceeded the training set thresholds. Among them, some composition of AA5083 and AA6N01 (never utilized in training process) are below/beyond the minimum/maximum-value of the training set, while the rainfall in Singapore exceeds the max-value of the dataset.

In conclusion, we have proposed an optimal strategy for the Al alloys corrosion rate prediction based on small-sample data from long-term field exposure experiments. The IACRF model is conducive to the selection and utilization of Al alloys in various environments. The use of the ab-initio calculations for the WF and the Cl AE of the secondary phases in Al alloys further characterizes the corrosion dataset, improving the corrosion rate prediction accuracy by 17.30%. In particular, we explicitly interpreted the effect of the phase information on the Al alloys corrosion prediction. Moreover, through comparing various statistical methods of WF and Cl AE, it was evident that the minimum WF of all phases should be specified as the evaluation criterion when utilizing the ab-initio calculation to measure the corrosion resistance of Al alloys. In addition, the WF was more suitable for corrosion prediction of pitting (galvanic) corrosion-prone alloys, while Cl AE tends to play a more significant role in Cl-induced passivation film destruction or general corrosion systems. To generalize the model broadly, we utilized feasible experimental features, Al content, rainfall, and exposure time combined with phases WF, to train the model. Finally, independent field exposure experiments in Southeast Asia demonstrated and verified the generalization ability of IACRF in corrosion prediction.

METHODS

Data pre-processing and random forest model

The corrosion dataset involves 15 experimental features and 2 ab-initio calculated features, with a total of 162 experimental data. The experimental features include ten element content (Al, Zn, Mg, Si, Cu, Fe, Cr, Mn, Ti, and the sum of other elements), four environmental parameters (annual average temperature, annual rainfall, precipitation pH, and the deposition of Cl), and exposure time. Each feature was linearly normalized. Experimental data were open access on Github, and the calculation results of secondary phases are summarized in Supplementary Table 2. In addition, six statistical strategies were utilized to find the optimal description for the secondary phase ab-initio calculation in Al alloys (Supplementary Tables 3 and 4). The detailed statistical strategies (the preprocessing of Cl AE is similar to that of WF) were as follows:

The descriptor ψ_{all}^{avg} refers to averaging the WFs of all crystallographic orientation including all phases in Al alloys, given by Eq. (1).

$$\psi_{all}^{avg} = \left(\sum_i^n \sum \psi_{surf}^i \right)/n \tag{1}$$

where i is the serial number of secondary phases, n is total number of secondary phases in Al alloys, and ψ_{surf}^i indicates the WF of i-th phase on different orientations (surf).

The descriptor ψ_{lowest}^{avg} refers to specifying the lowest WF among all orientation in one phase, and then averaging all phases in Al alloys.

$$\psi_{lowest}^{avg} = \left(\sum_i^n \psi_{lowest}^j \right)/n \tag{2}$$

where the ψ_{lowest}^j expresses the lowest WF corresponding to all orientations in the i-th phase.

The descriptor $\psi_{\gamma_{lowest}}^{avg}$ uses the WF of the orientation with the lowest surface energy to represent the i-th phase, and then

averages all phases in the Al alloys.

$$\psi_{\gamma_{lowest}}^{avg} = \left(\sum_i^n \psi_{\gamma_{lowest}}^i\right)/n \tag{3}$$

where $\psi_{\gamma_{lowest}}^i$ denotes the WF with the lowest surface energy (γ) in the i-th phase.

The descriptors ψ_{lowest} ($\psi_{highest}$) are the minimal (maximal) WF ignoring the number and type of phases.

$$\psi_{lowest} = \min\left(\psi_{lowest}^n\right) \tag{4}$$

$$\psi_{highest} = \max\left(\psi_{highest}^n\right) \tag{5}$$

where the ψ_{lowest}^n ($\psi_{highest}^n$) indicates the minimal (maximal) WF of all phases (n) in Al alloys.

The descriptor $\psi_{\gamma_{lowest}}$ is the WF with the lowest surface energy neglecting the phase type.

$$\psi_{\gamma_{lowest}} = \min\left(\psi_{\gamma_{lowest}}^n\right) \tag{6}$$

The RF algorithm utilized was the classification and regression tree assembled by scikit-learn[45]. Besides, the regression criterion was the squared error loss function, and out-of-bag data was also used to evaluate the generalization ability of the model. Since the dataset size was small (~160), the maximum estimator was limited to 100. Therefore, the estimators increased from 1 to 100, and 28 repeats were performed for each estimator. Specifically, all features were utilized for model training, and the dataset is shown in Supplementary Table 5. To eliminate the effects of data variance, the split number of the dataset was fixed (0.8 training data and 0.2 testing data). In addition, a new leaf can generate when the sample amount reached 2, and the minimum samples number on each leaf was 1. The estimator's number of IACRF model was 100, and the maximum depth of the tree is not limited. Moreover, the specific parameters of IACRF models can be obtained on Github. Furthermore, the evaluation criteria were calculated by MSE and R^2, given by Eqs. (7) and (8).

$$MSE = \frac{1}{N}\sum_i^N \left(y_i - y_{pred_i}\right)^2 \tag{7}$$

$$R^2 = 1 - \frac{\sum_i^N \left(y_i - y_{pred_i}\right)^2}{\sum_i^N \left(y_i - y_{mean_i}\right)^2} \tag{8}$$

where N expresses the total number of data in corrosion dataset, and i-th is the serial number. y_i denotes the true value of the i-th sample. y_{pred_i} is the predicted value and y_{mean_i} represents the average true values of all samples.

Long-term field exposure experiments

Al alloys including 2xxx, 5xxx, 6xxx, and 7xxx were subjected to field exposure experiment. Chemical compositions of the 18 kinds of Al alloys are listed in Supplementary Table 6, and the size of exposure sample was 150 mm × 75 mm × 4 mm. Prior to the field exposure experiments, the samples were all degreased in acetone, cleared with distilled water, and dehydrated by ethyl alcohol. Next, the dry samples were weighted by a balance with an accuracy of 0.01 mg and kept in desiccators. During field exposures, the samples were mounted on racks at an angle of 45° to the horizon, and the exposure time ranged from 0.5 to 4 years. Three Southeast Asian regions, namely Singapore, Jakarta, and Bangkok, were selected for validation. In addition, their available environmental parameters are shown in Supplementary Table 7. Subsequently, the samples were recovered for corrosion assessment according to the Chinese standard (GB/T 16545-2015). To be specific, corrosion products covered on the specimens were removed by immersion in the solution (50 mL H_3PO_4 + 20 g CrO_3 + 1 L H_2O) for 5–10 min at 80–100 °C. Then, the samples were rinsed with distilled water and air dried. The microstructures

(scanning electron microscopy and energy dispersive spectroscopy mapping) of the secondary phases in Al alloys were detected by GeminiSEM 500 (ZEISS, UK). The corrosion rate (ω, $\mu m a^{-1}$) could be calculated as follow:

$$\omega = \frac{1000(G_0 - G_1)}{2t\rho(ab + ac + bc)} \tag{9}$$

where G_0 and G_1 are the original and the final weights (g), a, b, and c express the length, width, and thickness (mm) of the samples respectively, t denotes the exposure time (year, a), and ρ is the density (g cm^{-3}) of the Al alloys.

Ab-initio calculations

All ab-initio calculations were carried out using the Vienna Ab-initio Simulation Package (VASP) with Medea[46]. A total of nine secondary phases in Al alloys were considered to calculate WF, surface energy and Cl AE, which initial structure is presented in Supplementary Fig. 3. In view of the complexity of numerous secondary phases in Al alloys, two criteria were utilized to determine the corresponding calculated secondary phases structures. (I) The most stable structure with the lowest formation energy should be taken into account; (II) The structure observed from actual exposure samples. To guarantee the calculation accuracy, the surface models consisted of at least 6-layer slabs. Besides, using the outermost atoms (containing the adsorbed Cl atom) as base points, a 15-Å-thick vacuum gap was applied to the model to create the surface state (See Supplementary Fig. 4). General gradient approximation[47] and Perdew–Burke–Ernzerhof[48] methods were utilized in the calculations. In addition, the structure optimization energy convergence accuracy was set to 10^{-5} eV Å^{-1}, and the spacing of k-points and cutoff energy for adsorption calculation was set to 0.25 Å^{-1} and 450 eV, respectively. The surface energy and Cl AE can be calculated based on the Eqs. (10) and (11), respectively.

$$\gamma_{surf} = \left(E_{surf} - \sum nE_{atom}\right)/2A \tag{10}$$

where γ_{surf} and E_{surf} represent the surface energy and total energy, respectively. E_{atom} indicates the energy of a single element in bulk and n is the atom number. A expresses the area.

$$E_{ads}^{Cl} = E_{surf}^{Cl} - E_{surf} - E_{Cl} \tag{11}$$

where E_{ads}^{Cl} is the adsorption energy of a single Cl atom on a phase surface, E_{surf}^{Cl} indicates the total energy of the adsorption system with the Cl adsorbed on the surface, E_{surf} represents the pure system without the Cl adsorbed, and E_{Cl} is the energy of a single Cl in bulk.

The WF (ψ) refers to the minimum energy required for the electrons of the metals to escape from the interior to the vacuum. In order to express the corrosion tendency of secondary phases in Al alloys more efficiently and accelerate the training process simultaneously, the WF is calculated by Eq. (12) and then preprocessed to obtain the theoretical WF differences (ψ_{matix}^{phase}) according to Eq. (13)[49].

$$\psi = \phi - E_f \tag{12}$$

$$\psi_{matix}^{phase} = \frac{\psi_{phase} - \psi_{matrix}}{e} \tag{13}$$

where ϕ indicates the vacuum level determined from the average potential in the vacuum gap. E_f is the Fermi level of the system. ψ_{phase} and ψ_{matrix} are the WF of the phase and Al matrix. e is the charge of an electron.

DATA AVAILABILITY
The relevant data is available from the corresponding author upon reasonable request.

CODE AVAILABILITY
Our code and dataset are publicly available on https://github.com/yucheng-ji/IACRF.

REFERENCES
1. Li, X. et al. Materials science: share corrosion data. *Nature* **527**, 441–442 (2015).
2. Frankel, G. Pitting corrosion of metals: a review of the critical factors. *J. Electrochem. Soc.* **145**, 2186 (1998).
3. Song, G. Potential and current distributions of one-dimensional galvanic corrosion systems. *Corros. Sci.* **52**, 455–480 (2010).
4. King, A., Johnson, G., Engelberg, D., Ludwig, W. & Marrow, J. Observations of intergranular stress corrosion cracking in a grain-mapped polycrystal. *Science* **321**, 382–385 (2008).
5. Raabe, D., Tasan, C. C. & Olivetti, E. A. Strategies for improving the sustainability of structural metals. *Nature* **575**, 64–74 (2019).
6. Reuter, M. A., van Schaik, A., Gutzmer, J., Bartie, N. & Abadías-Llamas, A. Challenges of the circular economy: a material, metallurgical, and product design perspective. *Annu. Rev. Mater. Sci.* **49**, 253–274 (2019).
7. Popova, K. & Prošek, T. Corrosion monitoring in atmospheric conditions: a review. *Metals* **12**, 171 (2022).
8. Man, C., Dong, C., Wang, L., Kong, D. & Li, X. Long-term corrosion kinetics and mechanism of magnesium alloy AZ31 exposed to a dry tropical desert environment. *Corros. Sci.* **163**, 108274 (2020).
9. Sun, S., Zheng, Q., Li, D. & Wen, J. Long-term atmospheric corrosion behaviour of aluminium alloys 2024 and 7075 in urban, coastal and industrial environments. *Corros. Sci.* **51**, 719–727 (2009).
10. Renner, F. et al. Initial corrosion observed on the atomic scale. *Nature* **439**, 707–710 (2006).
11. Sheikh, A. K., Boah, J. K. & Hansen, D. A. Statistical modeling of pitting corrosion and pipeline reliability. *Corrosion* **46**, 190–197 (1990).
12. Fogel, D. B. *Evolutionary computation: toward a new philosophy of machine intelligence*. 3 edn, (John Wiley & Sons, 2006).
13. Coelho, L. B. et al. Reviewing machine learning of corrosion prediction in a data-oriented perspective. *npj Mater. Degrad.* **6**, 1–16 (2022).
14. Roy, A. et al. Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys. *npj Mater. Degrad.* **6**, 1–10 (2022).
15. Dong, C. et al. Integrated computation of corrosion: modelling, simulation and applications. *Corros. Commun.* **2**, 8–23 (2021).
16. Zhao, H., Zhang, X., Ji, L., Hu, H. & Li, Q. Quantitative structure–activity relationship model for amino acids as corrosion inhibitors based on the support vector machine and molecular design. *Corros. Sci.* **83**, 261–271 (2014).
17. Ao, M. et al. Image deep learning assisted prediction of mechanical and corrosion behavior for Al-Zn-Mg alloys. *IEEE Access* **10**, 35620–35631 (2022).
18. Pei, Z. et al. Towards understanding and prediction of atmospheric corrosion of an Fe/Cu corrosion sensor via machine learning. *Corros. Sci.* **170**, 108697 (2020).
19. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
20. Ding, R. et al. Chemical boundary engineering: A new route toward lean, ultra-strong yet ductile steels. *Sci. Adv.* **6**, eaay1430 (2020).
21. Zou, C. et al. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. *Acta Mater.* **202**, 211–221 (2021).
22. Garner, A. et al. Multiscale analysis of grain boundary microstructure in high strength 7xxx Al alloys. *Acta Mater.* **202**, 190–210 (2021).
23. Ji, Y., Dong, C., Chen, L., Xiao, K. & Li, X. High-throughput computing for screening the potential alloying elements of a 7xxx aluminum alloy for increasing the alloy resistance to stress corrosion cracking. *Corros. Sci.* **183**, 109304 (2021).
24. Trink, B., Weißensteiner, I., Uggowitzer, P. J., Strobel, K. & Pogatscher, S. High Fe content in Al-Mg-Si wrought alloys facilitates excellent mechanical properties. *Scr. Mater.* **215**, 114701 (2022).
25. Obot, I., Macdonald, D. & Gasem, Z. Density functional theory (DFT) as a powerful tool for designing new organic corrosion inhibitors. Part 1: an overview. *Corros. Sci.* **99**, 1–30 (2015).
26. Schöchlin, J., Bohnen, K. P. & Ho, K. M. Structure and dynamics at the Al(111)-surface. *Surf. Sci.* **324**, 113–121 (1995).
27. Taylor, C. D., Wasileski, S. A., Filhol, J.-S. & Neurock, M. First principles reaction modeling of the electrochemical interface: consideration and calculation of a tunable surface potential from atomic and electronic structure. *Phys. Rev. B* **73**, 165402 (2006).
28. Ji, Y., Dong, C., Kong, D. & Li, X. Design materials based on simulation results of silicon induced segregation at AlSi10Mg interface fabricated by selective laser melting. *J. Mater. Sci. Technol.* **46**, 145–155 (2020).
29. Li, N. et al. Insight into the localized strain effect on micro-galvanic corrosion behavior in AA7075-T6 aluminum alloy. *Corros. Sci.* **180**, 109174 (2021).
30. Foley, D. L., Leff, A. C., Lang, A. C. & Taheri, M. L. Evolution of β-phase precipitates in an aluminum-magnesium alloy at the nanoscale. *Acta Mater.* **185**, 279–286 (2020).
31. Liu, Y. et al. Short-term rainfall forecast model based on the improved BP–NN algorithm. *Sci. Rep.* **9**, 19751 (2019).
32. Skinnider, M. A., Stacey, R. G., Wishart, D. S. & Foster, L. J. Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.* **3**, 759–770 (2021).
33. Dellaferrera, G., Woźniak, S., Indiveri, G., Pantazi, A. & Eleftheriou, E. Introducing principles of synaptic integration in the optimization of deep neural networks. *Nat. Commun.* **13**, 1885 (2022).
34. Agrawal, A. et al. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integr. Mater. Manuf. I.* **3**, 90–108 (2014).
35. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *npj Comput. Mater.* **4**, 25 (2018).
36. Wang, C., Fu, H., Jiang, L., Xue, D. & Xie, J. A property-oriented design strategy for high performance copper alloys via machine learning. *npj Comput. Mater.* **5**, 1–8 (2019).
37. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 1–9 (2016).
38. Rohwerder, M. & Turcu, F. High-resolution Kelvin probe microscopy in corrosion science: scanning Kelvin probe force microscopy (SKPFM) versus classical scanning Kelvin probe (SKP). *Electrochim. Acta* **53**, 290–299 (2007).
39. Strohmeier, B. R. An ESCA method for determining the oxide thickness on aluminum alloys. *Surf. Interface Anal.* **15**, 51–56 (1990).
40. Evertsson, J. et al. The thickness of native oxides on aluminum alloys and single crystals. *Appl. Surf. Sci.* **349**, 826–832 (2015).
41. Wei, X. et al. Electrochemical measurements and atomistic simulations of Cl-induced passivity breakdown on a Cu2O film. *Corros. Sci.* **136**, 119–128 (2018).
42. Olefjord, I. & Elfstrom, B.-O. The composition of the surface during passivation of stainless steels. *Corrosion* **38**, 46–52 (1982).
43. Chen, D. et al. Revealing the inner rules of PREN from electronic aspect by first-principles calculations. *Corros. Sci.* **189**, 109561 (2021).
44. Van Huis, M., Chen, J., Zandbergen, H. & Sluiter, M. Phase stability and structural relations of nanometer-sized, matrix-embedded precipitate phases in Al–Mg–Si alloys in the late stages of evolution. *Acta Mater.* **54**, 2945–2955 (2006).
45. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
46. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
47. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
48. Hammer, B., Hansen, L. B. & Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Phys. Rev. B* **59**, 7413 (1999).
49. Jin, Y. et al. First-principle calculation of Volta potential of intermetallic particles in aluminum alloys and practical implications. *J. Electrochem. Soc.* **164**, C465 (2017).

AUTHOR CONTRIBUTIONS
Y.J., N.L., and C.D. designed the algorithms and wrote the manuscript. Y.J., N.L., Z.C., M.L., K.X., T.C., M.A., and X.S. collected the field exposure experiment dataset. Y.J., X.F., D.Z., J.R., X.L., and P.D. analyzed the results. X.L. and C.D. supervised the project.

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/s41529-022-00295-5.

Correspondence and requests for materials should be addressed to Chaofang Dong.