ARTICLE   OPEN

Check for updates

Machine learning assisted discovery of high-efficiency self-healing epoxy coating for corrosion protection

Tong Liu[1,2,3], Zhuoyao Chen[1,2], Jingzhi Yang[1,2], Lingwei Ma[1,2,4], Arjan Mol [5] and Dawei Zhang [1,2,4 ✉]

Machine learning is a powerful means for the rapid development of high-performance functional materials. In this study, we presented a machine learning workflow for predicting the corrosion resistance of a self-healing epoxy coating containing ZIF-8@Ca microfillers. The orthogonal Latin square method was used to investigate the effects of the molecular weight of the polyetheramine curing agent, molar ratio of polyetheramine to epoxy, molar content of the hydrogen bond unit (UPy-D400), and mass content of the solid microfillers (ZIF-8@Ca microfillers) on the low impedance modulus ($\lg|Z|_{0.01Hz}$) values of the scratched coatings, generating 32 initial datasets. The machine learning workflow was divided into two stages: In stage I, five models were compared and the random forest (RF) model was selected for the active learning. After 5 cycles of active learning, the RF model achieved good prediction accuracy: coefficient of determination (R^2) = 0.709, mean absolute percentage error (MAPE) = 0.081, root mean square error (RMSE) = 0.685 ($\lg(\Omega \cdot cm^2)$). In stage II, the best coating formulation was identified by Bayesian optimization. Finally, the electrochemical impedance spectroscopy (EIS) results showed that compared with the intact coating (($4.63 \pm 2.08) \times 10^{11}$ $\Omega \cdot cm^2$), the $|Z|_{0.01Hz}$ value of the repaired coating was as high as ($4.40 \pm 2.04) \times 10^{11}$ $\Omega \cdot cm^2$. Besides, the repaired coating showed minimal corrosion and 3.3% of adhesion loss after 60 days of neutral salt spray testing.

INTRODUCTION

Epoxy (EP) resin is widely used in the field of corrosion protection because of strong adhesion properties, high corrosion resistance, excellent mechanical properties and low cost. However, cracks may arise inside or at the surface of the EP matrix during long-term service and reduce its corrosion protection performance with time, thus increasing potential safety hazards during its service life[1]. The application of self-healing coatings will be the most common and cost-effective method of improving the corrosion protection and thus the durability of metallic structures. A wide range of engineering structures from vehicles to aircrafts, from factories to house-hold equipment can be effectively protected via the self-healing coating systems. Recent efforts have focused on improving the durability of EP coatings in the presence of damage by granting them self-healing functions, which can be realized through intrinsic repair of the material matrix by reversible covalent bonds[2] and noncovalent bonds[3], or via extrinsic strategies depending on the release of healing agents[4] and corrosion inhibitors[5] into coating defects. In contrast to these extrinsic self-healing mechanisms, the intrinsic one endows the coating with the ability to simulate natural systems and repeated repairability. Such mechanisms are typically based on reversible covalent bonds via disulfide bonds[6], Diels–Alder reactions[7], and hydrazone bonds[8], or non-covalent interactions via metal-ligand[9] and hydrogen bonding[10–12]. Among these mechanisms, the most promising one is based on dynamic hydrogen bonds because of their high reversibility and mild repair conditions, in combination with their directional and tunable self-association properties[13]. As an indication of the self-healing ability of the coating, the low-frequency impedance modulus, such as according to the electrochemical impedance spectroscopy (EIS) data measured at

0.01 Hz ($|Z|_{0.01Hz}$), were extensively used to estimate the overall corrosion resistance of the test area[14,15]. A higher $|Z|_{0.01Hz}$ value represents a higher barrier ability of the coating. Based on the previous studies[16], in our view the design of an ideal self-healing corrosion protective coating should have the following main index: (1) The $|Z|_{0.01Hz}$ value of the self-healed coating is nearly close to that of the intact coating; (2) excellent barrier ability, $|Z|_{0.01Hz}$ value more than 10^{10} $\Omega \cdot cm^2$; (3) long-term stability in corrosive environments both before and after repair. For example, in a previous work by our group[11], an intrinsic self-healing EP coating was developed by grafting 2-ureido-4[1H]-pyrimidinone (UPy) as a quadruple hydrogen bonding unit onto the backbones of an EP-matrix. The UPy/EP coating demonstrated high-efficient self-healing functionality within 5 min in 3.5 wt.% NaCl solution. The self-healed coating still had high $|Z|_{0.01Hz}$ value of 4.8×10^{10} $\Omega \cdot cm^2$ even after 60 days of immersion in NaCl solution.

Often, the achievement of the target performance of self-healing implies synergy between multiple components of the EP coating formulation, including different resins, curing agents, liquid/solid additives, etc. The conventional trial-and-error design strategy for coating formulation is time-consuming and labor-intensive. Recently, machine learning methods have show to represent a promising option for materials design and optimization, especially for systems with complex properties or compositions[17–21]. For example, Haik et al.[22] developed a machine learning model to predict the stress relaxation properties of EP matrix composites, based on a three-layer neural network model using initial stress, test temperature and operating time as input variables and stress relaxation behavior as output. The final model was obtained by training 9000 experimental data samples. This model can predict efficiently the time-dependent mechanical

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China. [2]National Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, Beijing 100083, China. [3]College of Materials Science and Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China. [4]Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, China. [5]Department of Materials Science and Engineering, Delft University of Technology, Mekelweg 2, Delft 2628CD, The Netherlands. ✉email: dzhang@ustb.edu.cn
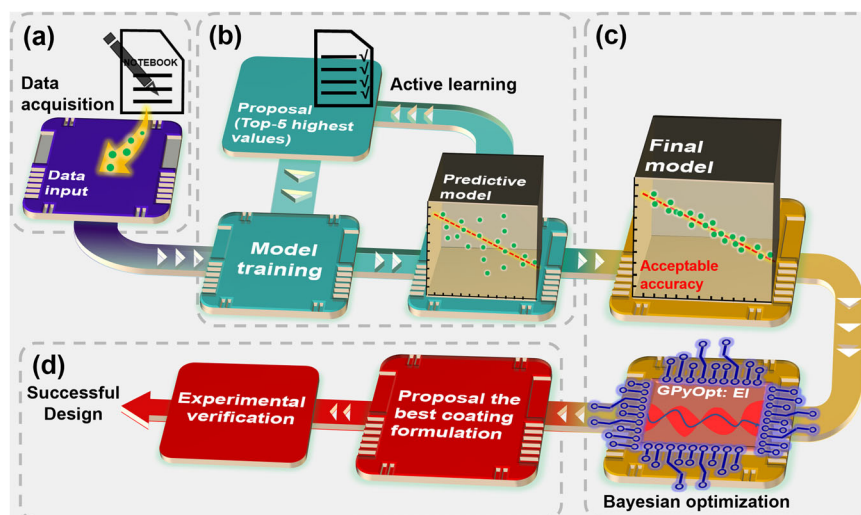
Fig. 1 A machine learning workflow for performance optimization in self-healing EP composite coating. Four steps are involved in machine learning workflow, from **a** data acquisition, **b** active learning, **c** Bayesian optimization, and **d** experimental verification.

behavior of a viscoelastic or a viscoplastic material. Kan et al.[23] constructed a molecular recognition model for predicting 2000 molecular descriptors from chemical structures using a gated graph neural network, and extracted 32-dimensional vectors representing 2000 molecular descriptors through the molecular recognition model to complete the dimension reduction. This 32-dimensional vector was used as the input value for the next Gaussian regression, and the machine learning model for predicting electrical conductivity was finally built by training a large amount of data. Typically, the establishment of an accurate machine learning requires vast training data, which is difficult to be obtained for polymer resin formation considering the heavy experimental workload in the synthesis and characterization[24,25]. Therefore, the construction of small sample datasets in the machine learning aspect of the research method has major implications for polymer design.

The problem of machine learning under small sample data conditions (<1000 samples) has received much attention in recent years[26,27]. For the processing of small sample data, the most common methods are the neural-network-based methods[28], hierarchical machine learning[29], active-learning-based method[30] and so on. For instance, Li et al.[31] proposed a model combined with nearest neighbor interpolation (NNI), synthetic minority oversampling technique (SMOTE) and extreme gradient boosting (XGBoost) models to predict the abrasion of rubber composites with small samples. NNI and SMOTE are two classical models in image processing that aim at increasing the sample size and solving the problem of sample unevenness. Combining these two models, the original dataset was expanded from 23 to 710 samples. Finally, the abrasion was predicted by the XGBoost model to yield a better prediction accuracy (MSE = 0.001). Similarly, active learning has been applied to discover EP adhesive strength[30], polymer molecular dynamics[32], high-T_g polymers[33,34] and among others from the small initial datasets.

Herein, we employed a machine learning framework to develop self-healing composite coatings for corrosion protection applications. A flowchart of the machine learning workflow is shown in Fig. 1. In the machine learning framework, active learning and Bayesian optimization to model and maximize the common logarithm of the low-frequency impedance modulus ($lg|Z|_{0.01Hz}$) obtained from EIS measurements for various scratched self-healing EP composite coatings to improve its self-healing property. This coating formulation consists of an EP resin, polyetheramines, amino-terminated urea-pyrimidinone monomers (UPy-D400) and ZIF-8@Ca microfillers. The EP resin mixed

Table 1. Summary of variable parameters for coating formulation used at the active learning stage.

Serial number	Variable parameter			
	MW_c (g·mol^{-1})	r	UPy-D400 content (mol%)	ZIF-8/Ca content (wt.%)
1	230	0.55	5	5.5
2	400	0.70	10	7.0
3	2000	0.85	15	8.5
4	4000	1.00	20	10.0

Variable parameters include the molecular weight of polyetheramine curing agent, molar ratio of polyetheramine to EP (r), molar content of UPy-D400 and mass content of ZIF-8/Ca microfillers.

with polyetheramine can react to form an EP-based polymer, and the UPy-D400 acts as a quadruple hydrogen bonding unit that can be grafted into the EP network to provide a self-healing function for the EP polymer via the self-association process; The ZIF-8@Ca microfiller, which is an empty CaCO$_3$ carbonate microcontainer with ZIF-8 nanoparticles assembled on the surface, is incorporated as a model filler that can not only enhance the barrier property of EP coating, but also present a pH-sensitive response to release loaded substance (e.g., inhibitors) to achieve useful functions. For the machine learning process, four-parameter variables, molecular weights of polyetheramine, the molar ratio of polyetheramine to EP, UPy-D400 content, and ZIF-8@Ca content, were used as input, and the $lg|Z|_{0.01Hz}$ value of the scratched coatings was used as output; 32 initial dataset were obtained from the preliminary experiment. Among the five common models, the model with the best accuracy was selected, and trained to achieve the best accuracy by active learning. Subsequently, the Bayesian optimization method was used to search for the scratched self-healing EP composite coating with an extremely high $lg|Z|_{0.01Hz}$ value. Finally, the self-healing and corrosion protective properties of the optimal coating were verified by EIS and salt spray testing.

RESULTS AND DISCUSSION

Experimental results from the initial dataset
As seen in Table 1, four parameters with four initial condition levels were set (total experimental conditions = $4^4 = 256$ sets).

Table 2. Experimental results of lg|Z|$_{0.01Hz}$ values of scratched coatings prepared under various conditions (32 initial dataset), the lg|Z|$_{0.01Hz}$ values represent the average ± standard deviations.

Serial number	Variable parameter				
	MW$_c$ (g·mol^{-1})	r	UPy-D400 content (mol%)	ZIF-8/Ca content (wt.%)	Measured lg(|Z|/Ω·cm^2)
1	230	0.55	5	5.5	4.89 ± 0.72
2	230	0.70	10	8.5	5.12 ± 0.69
3	230	0.85	15	10.0	6.06 ± 0.22
4	230	1.00	20	7.0	8.91 ± 0.76
5	400	0.55	10	7.0	4.75 ± 0.83
6	400	0.70	5	10.0	5.39 ± 0.45
7	400	0.85	20	8.5	10.08 ± 0.72
8	400	1.00	15	5.5	10.55 ± 0.52
9	2000	0.55	15	8.5	8.35 ± 0.41
10	2000	0.70	20	5.5	10.05 ± 0.76
11	2000	0.85	5	7.0	9.12 ± 0.69
12	2000	1.00	10	10.0	7.23 ± 0.82
13	4000	0.55	20	10.0	8.94 ± 0.70
14	4000	0.70	15	7.0	8.04 ± 0.59
15	4000	0.85	10	5.5	8.43 ± 0.28
16	4000	1.00	5	8.5	6.44 ± 0.65
17	230	0.55	20	7.0	4.88 ± 0.70
18	230	0.70	5	5.5	4.93 ± 0.63
19	230	0.85	10	8.5	5.59 ± 0.69
20	230	1.00	15	10.0	7.97 ± 0.70
21	400	0.55	15	5.5	5.01 ± 0.81
22	400	0.70	10	7.0	7.31 ± 0.42
23	400	0.85	5	10.0	8.12 ± 0.62
24	400	1.00	20	8.5	10.87 ± 0.80
25	2000	0.55	10	10.0	6.14 ± 0.75
26	2000	0.70	15	8.5	9.29 ± 0.62
27	2000	0.85	20	5.5	8.98 ± 0.74
28	2000	1.00	5	7.0	6.92 ± 0.70
29	4000	0.55	5	8.5	6.93 ± 0.62
30	4000	0.70	20	10.0	8.35 ± 0.52
31	4000	0.85	15	7.0	9.15 ± 0.66
32	4000	1.00	10	5.5	6.95 ± 0.79

Fig. 2 Distribution of lg|Z|$_{0.01Hz}$ experimental values from the 32 initial dataset. This task aims to confirm the distribution of target property values under initial experimental conditions.

Measurements of lg|Z|$_{0.01Hz}$ experimental values of scratched coatings that comprise our initial dataset are reported in Table 2. Figure 2 shows the distribution of lg|Z|$_{0.01Hz}$ experimental values. As shown in Fig. 2, the average lg|Z|$_{0.01Hz}$ experimental values were widely distributed in the range of 4.75–10.87 (lg(Ω·cm^2)). According to a previous experimental study[11], the scratched coatings with different self-healing abilities are involved in this distribution, indicating that the selection of the initial preparation conditions using the orthogonal Latin square method is reasonable.

Assessment and selection of an lg|Z|$_{0.01Hz}$ values prediction model

Next step, different experimental conditions and corresponding lg|Z|$_{0.01Hz}$ value of scratched coating were used as the input and output of the machine learning process, respectively, and five common machine learning models were trained using 32 initial datasets. A comparison of the predicted and measured lg|Z|$_{0.01Hz}$ values for each model is shown in Fig. 3a, e. A black dashed straight line indicates equal measured and predicted values. A comparison of the accuracy of each model is shown in Fig. 3f. Compared with the other models, the RF model yielded the best accuracy in terms of a higher coefficient of determination (R^2) value, and lower mean absolute percentage error (MAPE) and root mean square error (RMSE) values. This may be due to its deeper layers of model structure than general machine learning models; RF models possessed a good processing ability for data with high variability[38,39]. Hence, the RF model was chosen to predict the lg|Z|$_{0.01Hz}$ values in subsequent steps.

Active learning and machine learning model performance

For the active learning process, the RF model first predicted the lg|Z|$_{0.01Hz}$ values of all (256 − 32 = 224 sets) possible experimental conditions from the 32 initial dataset. The predicted lg|Z|$_{0.01Hz}$ values were ranked in descending order. The five top-ranked experimental conditions from 224 sets of conditions were selected as proposals for subsequent measurements to be performed in the laboratory. These five measurements were added to the initial 32 datasets. Then, the machine learning model for the prediction of the lg|Z|$_{0.01Hz}$ values was trained again on this improved (32 + 5) dataset. The new measurements were re-used in the RF model to improve the accuracy, as this can enhance the prediction accuracy for high-target performance samples in a targeted manner and improve the active learning efficiency. This process, from the prediction phase to the reuse phase, represents one

Four parameter variables included the molecular weight of polyetheramine, molar ratio of polyetheramine to EP, the molar content of UPy-D400, and mass content of the ZIF-8@Ca microfillers. An initial 32 sets of experimental conditions were extracted from the 256 sets by orthogonal Latin square design method[35]. This is a method based on mathematical statistics and the orthogonality principle, which can achieve the equivalent results of a large number of comprehensive tests with the minimum number of tests. It selects a part of points which can represent the whole experiment according to the orthogonality of the experiments. And these selected points are uniformly distributed in the whole space[36,37]. Then, the coatings were prepared for EIS measurements according to these 32 conditions, the corresponding the low impedance modulus (lg|Z|$_{0.01Hz}$ value) of different scratched coatings was obtained. The reason for selecting lg|Z|$_{0.01Hz}$ value as the output instead of using |Z|$_{0.01Hz}$ value is to eliminate the undesirable effects caused by sample dataset with high variability.
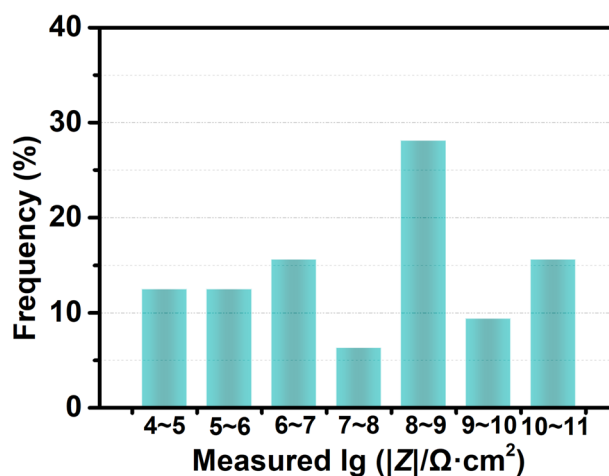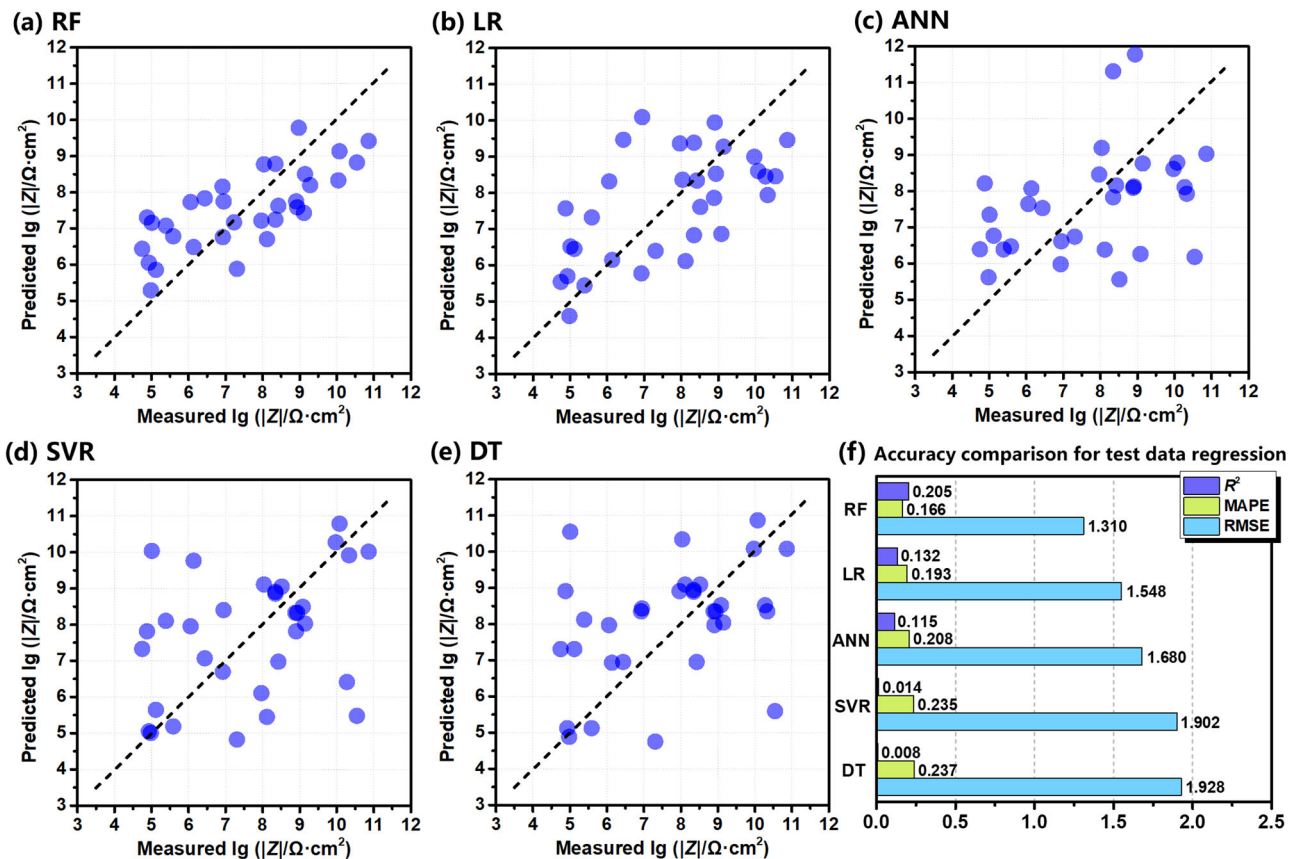
Fig. 3 The selection of the best machine learning model. Distribution of predicted versus measured $lg|Z|_{0.01Hz}$ values from successive test sets used in the 10-fold cross-validation using different machine learning models, **a–e** correspond to artificial neural network (ANN), linear regression (LR), support vector regression (SVR), decision tree (DT) and random forest (RF) model, respectively. **f** A comparison of the accuracy for each model, including R^2, MAPE, and RMSE values.

cycle of active learning (see Table 3). This active learning process is repeated until the preliminary goal of the best accuracy of the machine learning model is achieved. In this study, the active learning cycle was stopped if all the evaluation indices (MAPE, RMSE and R^2) stopped increasing.

Figures 4a–g present scatter plots of the predicted versus measured $lg|Z|_{0.01Hz}$ values from the initial dataset to the last cycle. The blue and red dots indicate existing and new measurements, respectively. The evolution of the corresponding R^2, MAPE and RMSE values for each cycle is summarized in Fig. 4h, i. As shown in Figs. 4a–g, the predicted and measured values gradually approached the black dashed straight line from the initial dataset to the last cycle, indicating that an increase in the dataset size resulted in predicted $lg|Z|_{0.01Hz}$ values that are closer to measured $lg|Z|_{0.01Hz}$ values. As the dataset size increased, R^2 clearly increased, and the MAPE and RMSE decreased gradually. After five active learning cycles, the R^2, MAPE and RMSE values reached equilibrium, at this time, the active learning process was terminated. For the dataset of 62 samples, the RF model achieved R^2, MAPE and RMSE values of 0.709, 0.081 and 0.685 ($lg(\Omega \cdot cm^2)$), respectively. Compared to the accuracy of the initial dataset, improvements of 246%, 51% and 47% were achieved for R^2, MAPE, and RMSE, respectively. In this case, R^2 was greater than 0.7 and both MAPE and RMSE were stabilized at a low level, indicating that the RF model reached acceptable accuracy. Therefore, the active learning procedure was stopped at this stage and the RF model was fixed based on the existing dataset.

In addition, Table 3 lists the top-five proposed experiments for the five cycles of active learning with the corresponding predicted and measured $lg|Z|_{0.01Hz}$ values. Several measured $lg|Z|_{0.01Hz}$

values in Table 3 that were greater than 11.00 ($lg(\Omega \cdot cm^2)$), which is greater than the highest value in the initial dataset, showed that the RF model allowed us to predict the experimental conditions of the coating with a potentially high self-healing ability. These additional data on high-performance self-healing coatings are beneficial for further maximization using Bayesian optimization. In addition, the proposed experiments required polyetheramine of molecular weights 400 and 2000 g·mol^{-1}, with an r value greater than 0.85, 10-20 mol% of UPy-D400, and ZIF-8@Ca microfiller content in the full range. This provided the main guidance for refining the test conditions in the subsequent step.

Bayesian optimization for screening optimal candidate

In this step, three experimental conditions were refined: r values, molar ratio of UPy-D400, and microfiller content were varied from 0.85 to 1.00, 10 to 20 mol%, and 5.5 to 10.0 wt.%, by increments of 0.1, 1 mol%, and 0.1 wt.%, respectively. The molecular weights of the polyetheramine curing agents were fixed at 400 and 2000 g·mol^{-1}. Obviously, this search space for the coating formulation is vast, and the machine learning model has limited utility if it do not incorporate uncertainty and the expected improvement process. Since a machine learning model is built using a limited amount of training data, the selection of candidates using that model may be limited to a local search. Therefore, we speculate that Bayesian optimization may give better results because this optimization technique considers the uncertainty of the prediction and the balance between local and global search[40].

Bayesian optimization works on a surrogate model and evaluates a utility function[41]. The utility function uses the mean and standard deviation of the candidates estimated by the

Table 3. Experimental results of predicted and measured lg|Z||$_{0.01Hz}$ values of scratched coatings prepared under various proposed conditions.

Cycle	Rank	Variable parameter					
		MW$_c$ (g·mol^{-1})	r	UPy-D400 content (mol%)	ZIF-8/Ca content (wt.%)	Predicted lg(|Z|/Ω·cm^2)	Measured lg(|Z|/Ω·cm^2)
Initial	1	400	1.00	20	5.5	10.49 ± 0.32	10.15 ± 0.30
	2	400	1.00	20	7.0	9.96 ± 0.24	10.88 ± 0.44
	3	400	0.85	10	5.5	9.71 ± 0.14	10.3 ± 0.41
	4	2000	0.85	15	5.5	8.93 ± 0.29	8.35 ± 0.62
	5	2000	0.85	20	8.5	9.33 ± 0.20	9.05 ± 0.75
Cycle 1	1	400	0.85	20	7.0	10.14 ± 0.19	10.11 ± 0.44
	2	400	1.00	20	10.0	9.88 ± 0.17	10.24 ± 0.13
	3	400	1.00	15	7.0	9.52 ± 0.20	10.52 ± 0.46
	4	2000	1.00	10	7.0	9.57 ± 0.18	8.23 ± 0.29
	5	400	0.85	20	10.0	9.65 ± 0.17	9.52 ± 0.51
Cycle 2	1	400	0.85	15	5.5	10.27 ± 0.21	10.08 ± 0.30
	2	400	1.00	15	8.5	10.23 ± 0.25	11.03 ± 0.38
	3	400	0.85	15	7.0	10.03 ± 0.15	10.76 ± 0.46
	4	400	1.00	15	10.0	9.90 ± 0.20	9.63 ± 0.64
	5	400	0.85	15	8.5	10.31 ± 0.12	10.26 ± 0.71
Cycle 3	1	400	0.85	15	10.0	9.44 ± 0.24	10.25 ± 0.75
	2	2000	0.85	15	7.0	9.12 ± 0.35	9.62 ± 0.54
	3	2000	0.85	15	8.5	9.37 ± 0.36	9.94 ± 0.48
	4	2000	1.00	15	5.5	8.91 ± 0.44	9.41 ± 0.38
	5	2000	1.00	15	8.5	9.43 ± 0.18	9.22 ± 0.15
Cycle 4	1	2000	0.85	15	10.0	9.40 ± 0.30	9.68 ± 0.80
	2	2000	0.85	20	8.5	9.30 ± 0.25	9.03 ± 0.68
	3	2000	1.00	15	10.0	9.30 ± 0.20	9.84 ± 0.51
	4	2000	1.00	15	7.0	9.24 ± 0.22	9.10 ± 0.74
	5	2000	1.00	20	8.5	9.00 ± 0.18	9.62 ± 0.48
Cycle 5	1	2000	0.85	20	7.0	9.04 ± 0.10	9.18 ± 0.84
	2	2000	0.85	20	10.0	9.28 ± 0.08	9.45 ± 0.54
	3	2000	1.00	20	10.0	9.06 ± 0.15	9.21 ± 0.69
	4	2000	1.00	20	5.5	8.99 ± 0.18	9.30 ± 0.50
	5	2000	0.85	20	7.0	9.16 ± 0.20	9.09 ± 0.25

Initial step: the top-five proposed experiments were obtained by a model trained on initial 32 samples in the range of remaining 224 untested experiments; Cycle 1: From the remaining 219 untested experiments, the another top-five proposed experiments were obtained by a model trained on 37 samples. Cycle 2 ~ 5 utilized the same method to obtain new proposed experiment and train the model.

surrogate model. The utility function encodes a trade-off between the exploitation (candidate searching at points with high mean) and exploration (candidate searching at points with high uncertainty). Herein, we have used RF as the surrogate model and expected improvement (EI) as a utility function. The EI is defined as the following Eqs. (1)-(2)[42]:

$$EI(x) = \sigma(x)[z\Phi(z) + \phi(z)] \tag{1}$$

$$z = [\mu(x) - f(x^+) - \varepsilon]/\sigma(x) \tag{2}$$

where $EI(x)$ represents the expected improvement value for each coating formulation candidate. μ and σ are the predicted output and standard deviation of the candidates obtained from the surrogate model, $f(x^+)$ is the maximum value of the target material property observed in the training data set. Φ represents the cumulative distribution function and ϕ is the probability distribution function assuming the target property values follows the normal distribution. The term ε regulates the amount of exploration, higher the value of ε more is the exploration. In this method, the largest EI value represents the most promising coating formulation candidate. Here, we use 1000 iterations for BO

run, as this was sufficiently many to predict the optimal experimental conditions with high accuracy (see Data Availability section for where to access this code), and a series of experiments were conducted starting from rank 1 (Table 4). The new highest lg|Z||$_{0.01Hz}$ values of 11.58 ± 0.28 (lg(Ω·cm^2)) was observed, that is, (4.40 ± 2.04) × 10^{11} Ω·cm^2. This impedance modulus value was considerably high compared with those reported in previous studies on EP-based self-healing coating[11,43–46], which reported a typical lg|Z||$_{0.01Hz}$ value range of 7.48–10.68 (lg(Ω·cm^2)). The suggested experimental conditions from Bayesian optimization showed that a relatively low molecular weight of polyetheramine and a high molar ratio of polyetheramine to EP were promising conditions for achieving a high lg|Z||$_{0.01Hz}$ value, whereas the molar ratio of UPy-D400 and microfillers content should be in the middle of their defined range. According to previous studies[47,48], excessive amine addition improves the shape recovery rate of EP materials. The intrinsic self-repair process mentioned in this study is realized by a self-healing unit (hydrogen bond) self-association process on the premise that the damage can be physically closed. A high shape recovery rate is beneficial for the physical closure of scratched material surfaces[11]. Excess amine
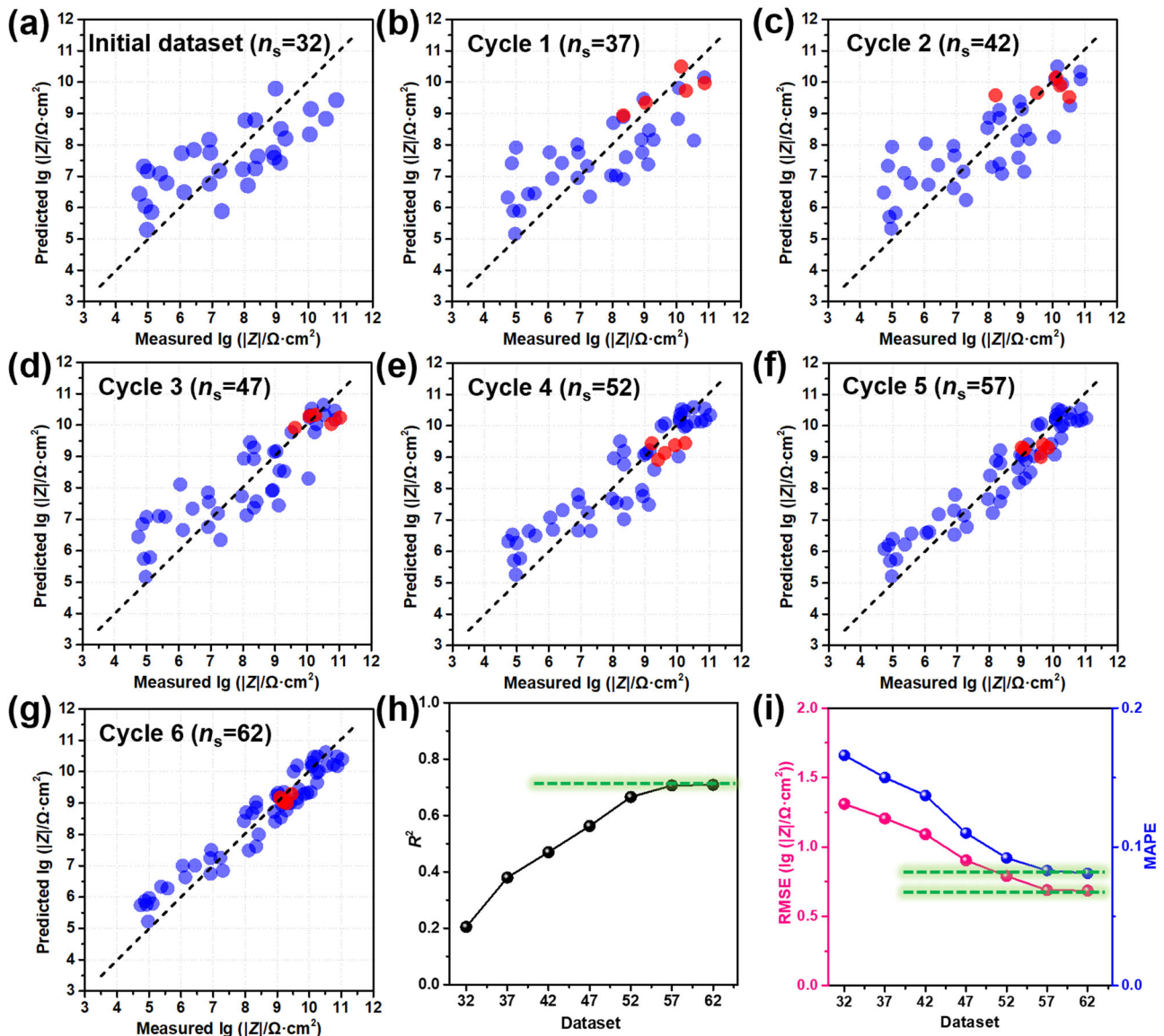
Fig. 4 Active learning process. a–g Correlation scatter plots of predicted and measured $lg|Z|_{0.01Hz}$ values using different datasets, including initial dataset and cycle 1-6 datasets. **h, i** Comparison of the accuracy (R^2, RMSE and MAPE value) of the RF model for different datasets.

Table 4. Proposed preparations of a composite coating at Bayesian optimization stage with the related experimental $lg|Z|_{0.01Hz}$ values of scratched coatings.

Rank	Variable parameter					
	MW_c (g·mol^{-1})	r	UPy content (mol%)	ZIF-8/Ca content (wt.%)	Predicted $lg(|Z|/$ Ω·cm^2)	Measured $lg(|Z|/\Omega$·cm^2)
1	400	0.94	14	7.8	11.01	11.58 ± 0.28
2	400	0.97	17	8.0	10.92	11.15 ± 0.65
3	400	1.00	16	8.0	10.92	10.98 ± 0.40
4	400	0.95	20	8.8	10.88	10.85 ± 0.74
5	400	1.00	16	7.4	10.88	10.90 ± 0.68

(excessive r value) leads to higher flexibility but lower mechanical strength of EP materials[47], an optimum combination of high strength and good flexibility can be achieved by adjusting the r value precisely through Bayesian optimization. The introduction of

self-healing units and microfillers may also affect the various performance indicators of the coatings, which can balance each addition amount simultaneously to achieve a reasonable design for target property.

Figure 5 shows the distribution of $lg|Z|_{0.01Hz}$ values of scratched coatings from the initial dataset, after the five active learning cycles, and after a Bayesian optimization process. The $lg|Z|_{0.01Hz}$ values from the initial dataset were spread randomly from 4.75 to 10.87 (lg(Ω·cm^2)). By comparison, all samples that followed an active learning cycle exhibited a high $lg|Z|_{0.01Hz}$ value (>8.23 (lg(Ω·cm^2))), and one sample from the Bayesian optimization dataset showed an exceptionally high $lg|Z|_{0.01Hz}$ value. These results demonstrate the potential of our machine learning framework for the design and optimization of high-performance functional materials based on small sample conditions.

Interpretation of machine learning model for coating design

EIS measurements were conducted on the scratched pure commercial EP and ZIF-8@Ca/EP coatings and their corresponding intact coatings to study the self-healing and corrosion resistance
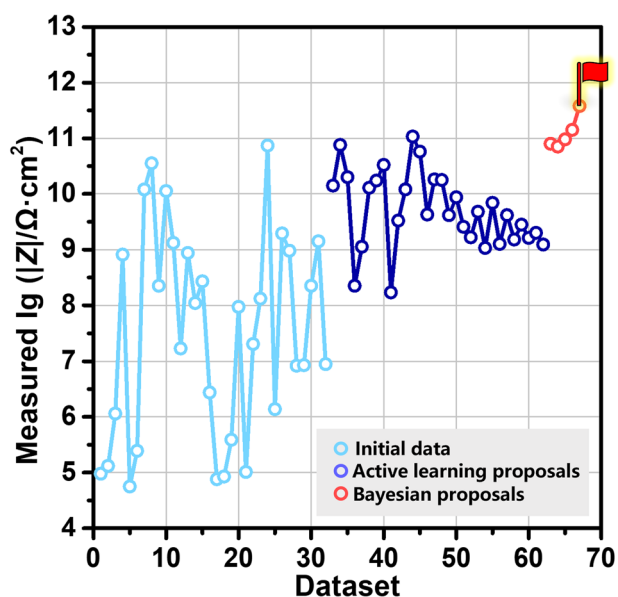
Fig. 5 Comparison of the measured target performance for each machine learning stage. Distribution of measured $\lg|Z|_{0.01Hz}$ values from the initial dataset (blue), after active learning process (dark blue) and after Bayesian optimization (red).

properties. The ZIF-8@Ca/EP coating was prepared based on the best formulation selected by Bayesian optimization. Nyquist and Bode plots of the intact coatings were obtained by EIS after 30 min of immersion in 3.5 wt.% NaCl solution (Fig. 6a–c). Figure 6d–i show the Nyquist and Bode plots of the steels with scratched coatings after immersion for 1, 15, 30 and 60 d. The as-used pure EP coating was prepared by mixing E51 with D400 polyether-amine curing agents at a molar ratio of 5:3. For the pure EP sample, the intact coating initially showed a high barrier property with large capacitive arc in the Nyquist plot (Fig. 6a) and the high $|Z|_{0.01Hz}$ value (3.98×10^{10} $\Omega \cdot cm^2$) in the Bode plot (Fig. 6b). The phase angles in the high frequencies (10^5 Hz) were close to $-90°$ which indicates the capacitive character of the coatings. In contrast to the intact pure EP coating, intact ZIF-8@Ca/EP coating exhibited a slightly larger capacitive arc in terms of Nyquist plot, and $|Z|_{0.01Hz}$ value rose to 3.82×10^{11} $\Omega \cdot cm^2$, indicating substantial improvement in the barrier property of the coating after the machine learning adjustment. The average and standard deviation of the $|Z|_{0.01Hz}$ value for intact coating were calculated using six parallel samples, expressed as $(4.63 \pm 2.08) \times 10^{11}$ $\Omega \cdot cm^2$.

In terms of the scratched coatings, the capacitive arcs of the pure EP coating shrank and the $|Z|_{0.01Hz}$ values declined gradually over the entire immersion time, demonstrating the continuous deterioration of the barrier property (Figs. 6d–e). Subsequently, for the phase diagrams in Fig. 6f, scratched pure EP showed two-time constants: one related to the charge transfer process at the coating/substrate interface ($10^{-2}-10^0$ Hz), and the other related to the resistance increase by means of corrosion product formation in the artificial defect (10^1-10^5 Hz)[49]. Compared with the Bode plots for pure EP coating, the Bode plots of the scratched coating showed approximately $-45°$ straight lines with $|Z|_{0.01Hz}$ values in excess of 3.80×10^{11} $\Omega \cdot cm^2$ at the beginning of immersion. The corresponding phase angles were $-90°$ over the frequency range of $10^{-1}-10^5$ Hz. This implies that during the immersion, a conductive pathway is not formed through the coating, which largely exhibits a capacitive behavior similar to that of an intact coating[50]. During the 60 d of immersion, the $|Z|_{0.01Hz}$ values of the ZIF-8@Ca/EP coating only slightly decreased from 3.80×10^{11} $\Omega \cdot cm^2$ to 1.23×10^{11} $\Omega \cdot cm^2$, confirming that the

scratched ZIF-8@Ca/EP coating had been well repaired and possessed a satisfactory corrosion resistance.

After scratching, the pure EP and ZIF-8@Ca/EP coatings were subjected to salt spray tests following the ASTM B117/D1654 standard. Figures 6b and 7a show the optical images of the coatings after exposure to the salt spray chamber for different periods. According to the visual assessment in Fig. 7a, green corrosion products were observed at the scratches of the pure EP coating within the 1 d of the salt spray test. After 60 d, large-scale coating delamination and corrosion products appeared in the scratched region, indicating that the scratched location of the pure EP coating was highly vulnerable to attack by corrosive species. Compared with pure EP, only slight scratch traces were observed at the scratched positions, and the ZIF-8@Ca/EP coating did not show any signs of degradation (delamination, corrosion, or blistering) after 30 d (Fig. 7b). Furthermore, as the salt spray exposure time increased to 60 d, only one slight corrosion spot was observed at the scratched site, indicating the corrosion of the scratched ZIF-8@Ca/EP coating could be controlled in a salt spray environment for a long time.

The adhesion strength, an important indicator of coating properties, can be measured using a pull-off test. Figure 7d shows the adhesion strength/loss values of intact pure EP and ZIF-8@Ca/EP coating before and after the 60 d salt spray test. The optical images of the remaining coatings following the pull-off test are presented in Fig. 7c. As shown in Fig. 7c, none of the samples exhibits cohesive failure. As shown in Fig. 7c, the dry adhesion strength of the ZIF-8@Ca/EP coatings (9.82 MPa) is higher than that of pure EP (4.70 MPa). This is because the introduction of branched-chain amines and UPy units enhanced the hydrogen bonding between the coating and the metal surface[51]. After salt spraying, the pure EP coating exhibited a considerable adhesion loss of 79.4% (0.97 MPa). In contrast, the ZIF-8@Ca/EP coating demonstrated not only the highest wet adhesion strength (9.50 MPa) but also minimal adhesion loss (3.3%) after a 60 d of salt spray test.

In summary, the design of experimental techniques combined with an active learning and Bayesian optimization was proposed to predict and optimize the $\lg|Z|_{0.01Hz}$ values of scratched EP self-healing coatings composed of different molecular weights of polyetheramine curing agent, molar ratios of polyetheramine to E51 EP resin, molar content of UPy-D400 and mass contents of ZIF-8@Ca microfillers. The active learning process yielded the preferred experimental conditions to build a predictive RF model of $\lg|Z|_{0.01Hz}$ values with satisfactory accuracy ($R^2 = 0.709$, MAPE = 0.081, RMSE = 0.685 ($\lg(\Omega \cdot cm^2)$)) after five cycles of active learning. Then, an extremely high $\lg|Z|_{0.01Hz}$ values of 11.58 ($|Z|_{0.01Hz} = 3.80 \times 10^{11}$ $\Omega \cdot cm^2$) was achieved using the experimental conditions that were refined by Bayesian optimization. As confirmed by EIS, the ZIF-8@Ca/EP coating exhibited a great healing effect in barrier property (intact sample: 3.82×10^{11} $\Omega \cdot cm^2$, repaired sample: 3.80×10^{11} $\Omega \cdot cm^2$). In addition, in terms of the corrosion resistance after repair, the ZIF-8@Ca/EP coating exhibited slight corrosion after 60 d of the salt spray test, and the adhesion loss of the composite coating after the salt spray test was 3.3%, which was considerably lower than that of the pure EP coating (79.4%).

METHODS

Materials

Polyetheramine curing agents with four different molecular weights (230, 400, 2000 and 4000 g·mol^{-1}) were sourced from the Aladdin Industrial Corporation. The E51 EP resin was sourced from Jiangsu Heli Resin Co., ltd. The ZIF-8@Ca microfillers and the UPy-D400 monomers were obtained using previously published methods[11,51]. The Q235 mild steel was used as the substrate.
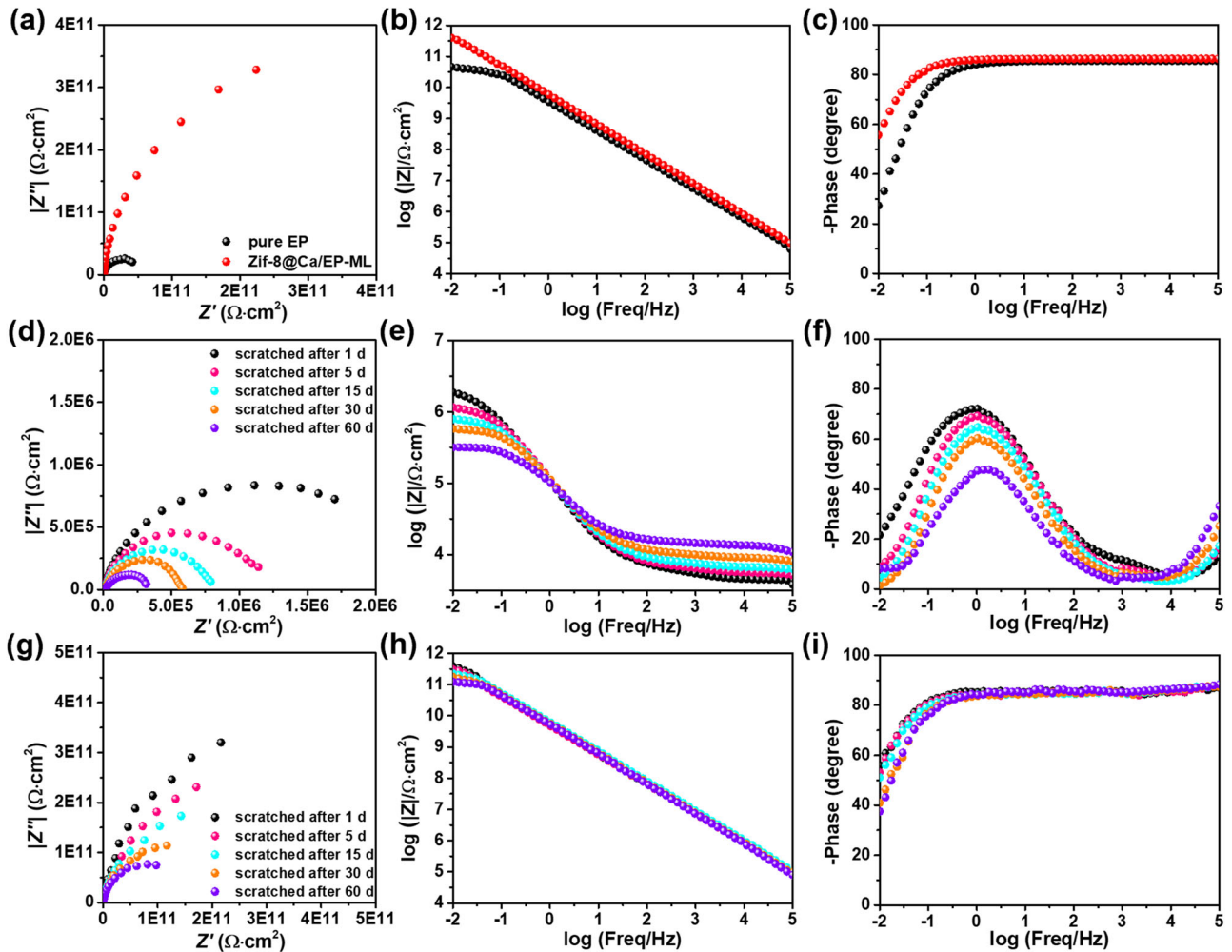
Fig. 6 EIS characterizations of the different intact/scratched coatings. a Nyquist plots and **b, c** Bode plots of the intact pure EP and intact ZIF-8@Ca/EP coatings after 30 min of immersion in 3.5 wt.% NaCl solution. Nyquist plots and Bode plots of different **d–f** scratched pure EP and **g–i** scratched ZIF-8@Ca/EP coating during immersion in 3.5 wt.% NaCl solution for 60 d.

Preparation of coatings and EIS test

Based on the selected 32 experimental conditions, the preparation process of the self-healing EP coating containing ZIF-8@Ca microfillers (ZIF-8@Ca/EP) is shown in Fig. 8. In each case, the ZIF-8@Ca microfillers were first mixed with the E51 EP resin under magnetic stirring. The polyetheramine curing agent and UPy-D400 were then added to the mixture using a mechanical agitator at 500 rpm for 10 min. Prior to the coating preparation, the steel specimens were wet-polished sequentially with 150-, 240- and 400-grit sandpapers, washed with ethanol and blow-dried in an N_2 atmosphere. The resulting mixture was applied to a steel piece using a bar coater. The coated samples were obtained by drying at room temperature for 48 h. The final thickness of each of the dry films was approximately 85 μm.

EIS tests were performed to measure the low-frequency impedance ($|Z|_{0.01Hz}$) values of the coated steel with/without an artificial scratch. Herein, all scratches of the EIS tests are made by a scalpel, and they are reproducible. The EIS results were obtained using a 3.5 wt.% NaCl solution and a CHI-660E electrochemical workstation with a three-electrode cell system comprising a coated steel substrate as a working electrode, a platinum plate electrode as a counter electrode and a saturated calomel electrode (SCE) as a reference electrode. The test parameters were set in the $10^{-2}-10^5$ Hz range with a 0.02 V root mean square amplitude. Prior to EIS measurements, artificial through-coating

scratches (approximately 3 mm in length and approximately 60 μm in width) were made on the different coated steels using a scalpel. The measurements were conducted on the coated steels at least five times to ensure the reproducibility of the EIS results. In EIS results, the $|Z|_{0.01Hz}$ value in the Bode plot usually represents the main performance index for the corrosion resistance of a coating, that is, a higher $|Z|_{0.01Hz}$ value reflects a higher barrier property[52]. Therefore, this index was used to characterize the repair effect of the barrier properties of the coating after scratching.

To further verify the self-healing and long-term anti-anticorrosion ability of the scratched composite coating after machine learning process, salt spray test was performed on the coatings via exposing the samples to salt spray for 60 d in accordance with ASTM D1654.

Data pre-processing, data splitting and machine learning models

Data pre-processing and data splitting were performed and different machine learning models were simulated using the Python package scikit-learn (version 1.1.1). The four variable parameters (Table 4) in this study were standardized following a standard Gaussian distribution of a mean of 0 and a variance of 1[53]. The purpose of normalization is to make the preprocessed data be limited to a certain range (e.g., [0,1] or [−1,1]), thus
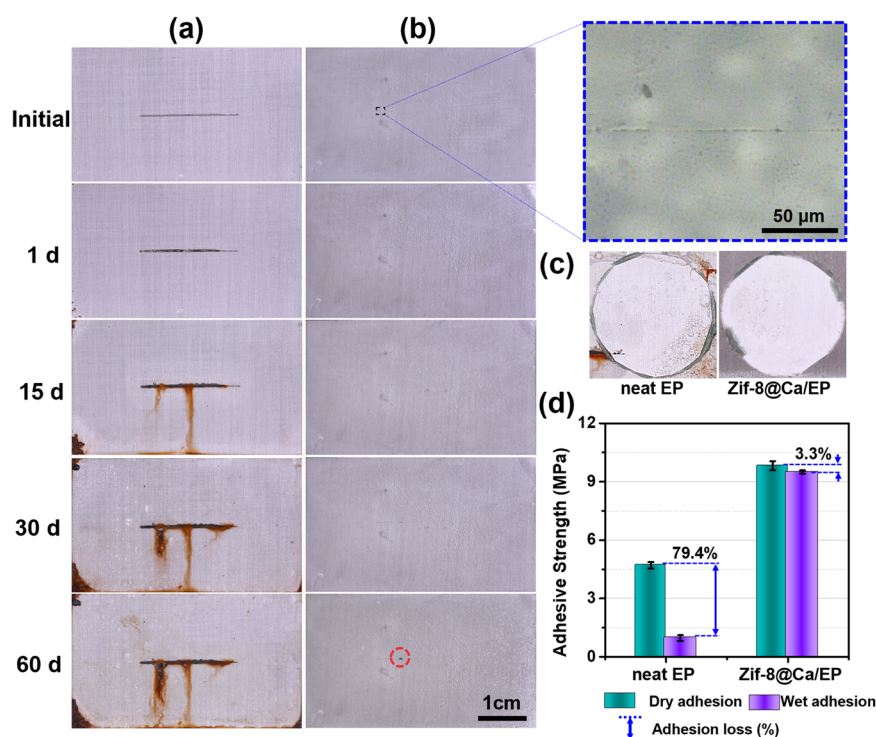
Fig. 7 Salt spray analysis of the different intact/scratched coatings. a, b Optical images of the pure EP and ZIF-8@Ca/EP coating. **c** Optical images of the pure EP and ZIF-8@Ca/EP coating after pull-off test at the end of salt spray test. **d** The adhesion strength values of the pure EP and ZIF-8@Ca/EP coating before and after 60 d of salt spray exposure, the adhesion strength values represent the average ± standard deviations.
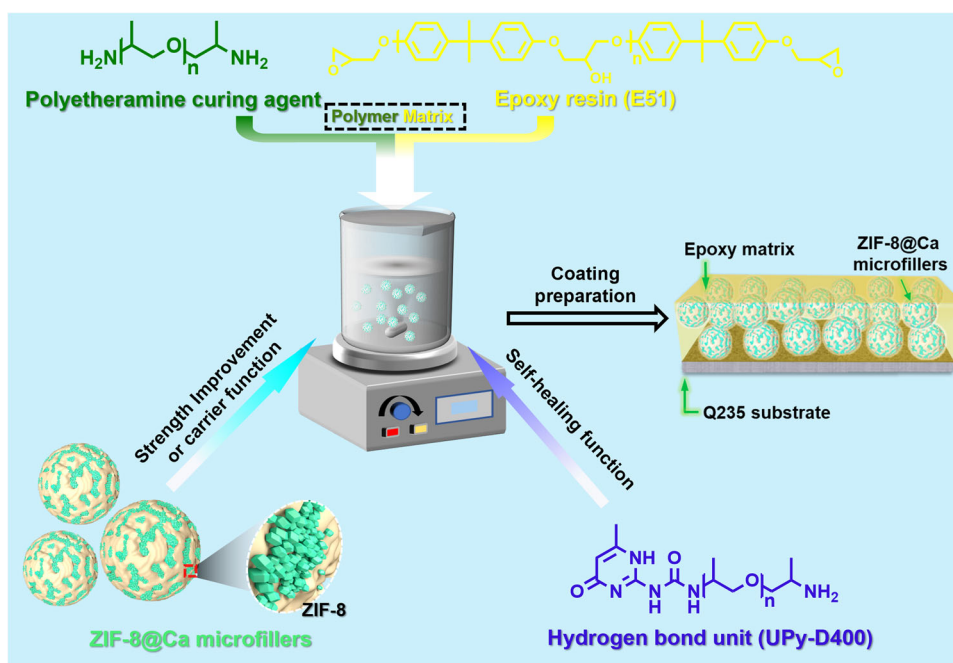

Fig. 8 Schematic illustration of the preparation process for self-healing EP composite coating. The coating formulation consists of the EP resin, polyetheramines, hydrogen bond unit (UPy-D400) and ZIF-8@Ca microfillers.

eliminating the undesirable effects caused by sample dataset with high variability. The validity and accuracy of all employed machine learning models were evaluated using k-fold cross-validation. In this step, the data were randomly arranged and divided into 10 groups. Nine groups were allocated for training purposes, and the remaining group was assigned to validate of the model. The average value was obtained by repeating the same process 10 times. To obtain the performance level of the model, the MAPE,

RMSE and R^2 were introduced to evaluate the k-fold cross-validation, using the following Eqs. (3)-(5):[54–56]

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5}$$

where n is the number of samples, and y_i and $\hat{y}_i$ are the experimental and predicted values of the ith sample, respectively.

The accuracy of the machine learning model was accessed using its MAPE (MAPE value is in between 0 and 1, a value closer to 0 indicates greater accuracy[57]) and RMSE (a lower value of each indicates greater accuracy[30]) and R^2 (a value closer to 1 indicates greater accuracy; when the R^2 coefficient is greater than 0.7, the model represents acceptable accuracy[58].)

Five machine learning models were applied as regression tools to the dataset: LR, ANN, SVR, DT and RF models. The machine learning methods are described in detail in the related reference[59]. The interested reader should refer to the Data Availability section for where to access our code used to run these algorithms.

Bayesian optimization

Bayesian optimization[40] was used to determine the highest $\lg|Z|_{0.01\text{Hz}}$ values by refining the variable conditions from Table 1. Bayesian optimization was performed using the Python package GPyOpt.

DATA AVAILABILITY

REFERENCES

1. He, Y. et al. Micro-crack behavior of carbon fiber reinforced Fe_3O_4/graphene oxide modified epoxy composites for cryogenic application. *Compos. Part A Appl. Sci. Manuf.* **108**, 12–22 (2018).
2. Huang, S. et al. An overview of dynamic covalent bonds in polymer material and their applications. *Eur. Polym. J.* **141**, 110094 (2020).
3. Utrera-Barrios, S., Verdejo, R. & López-Manchado, M. A. & Hernández Santana, M. Evolution of self-healing elastomers, from extrinsic to combined intrinsic mechanisms: a review. *Mater. Horiz.* **7**, 2882–2902 (2020).
4. Samadzadeh, M., Boura, S. H., Peikari, M., Kasiriha, S. M. & Ashrafi, A. A review on self-healing coatings based on micro/nanocapsules. *Prog. Org. Coat.* **68**, 159–164 (2010).
5. Shchukin, D. G. Container-based multifunctional self-healing polymer coatings. *Polym. Chem.* **4**, 4871–4877 (2013).
6. Canadell, J., Goossens, H. & Klumperman, B. Self-healing materials based on disulfide links. *Macromolecules* **44**, 2536–2541 (2011).
7. Kuang, X. et al. Facile fabrication of fast recyclable and multiple self-healing epoxy materials through diels-alder adduct cross-linker. *J. Polym. Sci. Pol. Chem.* **53**, 2094–2103 (2015).
8. Wen, N. et al. Recent advancements in self-healing materials: Mechanicals, performances and features. *React. Funct. Polym.* **168**, 105041 (2021).
9. Han, Y., Wu, X., Zhang, X. & Lu, C. Self-healing, highly sensitive electronic sensors enabled by metal–ligand coordination and hierarchical structure design. *ACS Appl. Mater. Inter.* **9**, 20106–20114 (2017).
10. Nardeli, J. V., Fugivara, C. S., Taryba, M., Montemor, M. F. & Benedetti, A. V. Self-healing ability based on hydrogen bonds in organic coatings for corrosion protection of AA1200. *Corros. Sci.* **177**, 108984 (2020).
11. Liu, T. et al. Ultrafast and high-efficient self-healing epoxy coatings with active multiple hydrogen bonds for corrosion protection. *Corros. Sci.* **187**, 109485 (2021).
12. Kim, G., Caglayan, C. & Yun, G. J. Epoxy-based catalyst-free self-healing elastomers at room temperature employing aromatic disulfide and hydrogen bonds. *ACS omega* **7**, 44750–44761 (2022).
13. Bosnian, A., Brunsveld, L., Folmer, B. & Sijbesma, R. & Meijer, E. *Macromol. Symp.* **201**, 143–154 (2003).
14. Rosero-Navarro, N. C., Pellice, S. A., Durán, A. & Aparicio, M. Effects of Ce-containing sol–gel coatings reinforced with SiO_2 nanoparticles on the protection of AA2024. *Corros. Sci.* **50**, 1283–1291 (2008).
15. Wang, J. et al. Two birds with one stone: Nanocontainers with synergetic inhibition and corrosion sensing abilities towards intelligent self-healing and self-reporting coating. *Chem. Eng. J.* **433**, 134515 (2022).
16. Fan, Z. et al. Self-healing mechanisms in smart protective coatings: a review. *Corros. Sci.* **144**, 74–88 (2018).
17. Tao, Q., Xu, P., Li, M. & Lu, W. Machine learning for perovskite materials design and discovery. *npj Comput. Mater.* **7**, 23 (2021).
18. Li, Z. et al. Machine learning in concrete science: applications, challenges, and best practices. *npj Comput. Mater.* **8**, 127 (2022).
19. Zhong, X. et al. Explainable machine learning in materials science. *npj Comput. Mater.* **8**, 204 (2022).
20. Taylor, C. D. & Tossey, B. M. High temperature oxidation of corrosion resistant alloys from machine learning. *npj Mater. Degrad* **5**, 38 (2021).
21. Li, Q. et al. Long-term corrosion monitoring of carbon steels and environmental correlation analysis via the random forest method. *npj Mater. Degrad* **6**, 1 (2022).
22. Al-Haik, M. S., Hussaini, M. Y. & Garmestani, H. Prediction of nonlinear viscoelastic behavior of polymeric composites using an artificial neural network. *Int. J. Plast.* **22**, 1367–1392 (2006).
23. Hatakeyama-Sato, K., Tezuka, T., Umeki, M. & Oyaizu, K. AI-assisted exploration of superionic glass-type Li(+) conductors with aromatic structures. *J. Am. Chem. Soc.* **142**, 3301–3305 (2020).
24. Askland, K. D. et al. Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy. *Int. J. Methods Psychiatr. Res.* **24**, 156–169 (2015).
25. Shao, M., Zhu, X.-J., Cao, H.-F. & Shen, H.-F. An artificial neural network ensemble method for fault diagnosis of proton exchange membrane fuel cell system. *Energy* **67**, 268–275 (2014).
26. Xu, P., Ji, X., Li, M. & Lu, W. Small data machine learning in materials science. *npj Comput. Mater.* **9**, 42 (2023).
27. Sutojo, T. et al. A machine learning approach for corrosion small datasets. *npj Mater. Degrad* **7**, 18 (2023).
28. Xiang, K.-L., Xiang, P.-Y. & Wu, Y.-P. Prediction of the fatigue life of natural rubber composites by artificial neural network approaches. *Mater. Des.* **57**, 180–185 (2014).
29. Menon, A., Thompson-Colón, J. A. & Washburn, N. R. Hierarchical machine learning model for mechanical property predictions of polyurethane elastomers from small datasets. *Front. Mater.* **6**, 87 (2019).
30. Pruksawan, S., Lambard, G., Samitsu, S., Sodeyama, K. & Naito, M. Prediction and optimization of epoxy adhesive strength from a small dataset through active learning. *Sci. Technol. Adv. Mater.* **20**, 1010–1021 (2019).
31. Li, D., Liu, J. & Liu, J. NNI-SMOTE-XGBoost: A novel small sample analysis method for properties prediction of polymer materials. *Macromol. Theory Simul.* **30**, 2100010 (2021).
32. Novikov, I. S., Shapeev, A. V. & Suleimanov, Y. V. Ring polymer molecular dynamics and active learning of moment tensor potential for gas-phase barrierless reactions: Application to S + H_2. *J. Chem. Phys.* **151**, 224105 (2019).
33. Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Commun.* **9**, 860–866 (2019).
34. Jha, A., Chandrasekaran, A., Kim, C. & Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures. *Model. Simul. Mater. Sci. Eng.* **27**, 024002 (2019).
35. Mandl, R. Orthogonal Latin squares: an application of experiment design to compiler testing. *Commun. ACM* **28**, 1054–1058 (1985).
36. Balak, Z. & Zakeri, M. Application of Taguchi L_{32} orthogonal design to optimize flexural strength of ZrB_2-based composites prepared by spark plasma sintering. *Int. J. Refract. Met. H.* **55**, 58–67 (2016).
37. Wu, C. J. & Hamada, M. S. *Experiments: planning, analysis, and optimization*. (John Wiley & Sons), (2011).
38. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
39. Ji, Y. et al. Random forest incorporating ab-initio calculations for corrosion rate prediction with small sample Al alloys data. *npj Mater. Degrad* **6**, 83 (2022).
40. Packwood, D. *Bayesian Optimization for Materials Science*. 11-28 (Springer), (2017).

41. Wagner, T., Emmerich, M., Deutz, A. & Ponweiser, W. in *Parallel Problem Solving from Nature, PPSN XI: 11th International Conference, Kraków, Poland, September 11-15, 2010, Proceedings, Part I 11*. 718-727 (Springer).

42. Mohanty, T., Chandran, K. & Sparks, T. D. Machine learning guided optimal composition selection of niobium alloys for high temperature applications. *APL Mach. Learn.* **1**, 036102 (2023).

43. Cui, G. et al. Research progress on self-healing polymer/graphene anticorrosion coatings. *Prog. Org. Coat.* **155**, 106231 (2021).

44. Nawaz, M., Habib, S., Khan, A., Shakoor, R. A. & Kahraman, R. Cellulose microfibers (CMFs) as a smart carrier for autonomous self-healing in epoxy coatings. *N. J. Chem.* **44**, 5702–5710 (2020).

45. Zhang, C., Wang, H. & Zhou, Q. Preparation and characterization of microcapsules based self-healing coatings containing epoxy ester as healing agent. *Prog. Org. Coat.* **125**, 403–410 (2018).

46. Wang, T. et al. Photothermal nanofiller-based polydimethylsiloxane anticorrosion coating with multiple cyclic self-healing and long-term self-healing performance. *Chem. Eng. J.* **446**, 137077 (2022).

47. Zheng, N., Fang, G., Cao, Z., Zhao, Q. & Xie, T. High strain epoxy shape memory polymer. *Polym. Chem.* **6**, 3046–3053 (2015).

48. Li, J., Rodgers, W. R. & Xie, T. Semi-crystalline two-way shape memory elastomer. *Polymer* **52**, 5320–5325 (2011).

49. Oliveira, C. & Ferreira, M. Ranking high-quality paint systems using EIS. Part I: intact coatings. *Corros. Sci.* **45**, 123–138 (2003).

50. Hao, Y., Sani, L. A., Ge, T. & Fang, Q. Phytic acid doped polyaniline containing epoxy coatings for corrosion protection of Q235 carbon steel. *Appl. Surf. Sci.* **419**, 826–837 (2017).

51. Liu, T. et al. Self-healing and corrosion-sensing coatings based on pH-sensitive MOF-capped microcontainers for intelligent corrosion control. *Chem. Eng. J.* **454**, 140335 (2023).

52. Tavandashti, N. P. et al. Inhibitor-loaded conducting polymer capsules for active corrosion protection of coating defects. *Corros. Sci.* **112**, 138–149 (2016).

53. Zheng, X., Zheng, P. & Zhang, R.-Z. Machine learning material properties from the periodic table using convolutional neural networks. *Chem. Sci.* **9**, 8426–8432 (2018).

54. De Myttenaere, A., Golden, B., Le Grand, B. & Rossi, F. Mean absolute percentage error for regression models. *Neurocomputing* **192**, 38–48 (2016).

55. Fukutani, T., Miyazawa, K., Iwata, S. & Satoh, H. G-RMSD: Root mean square deviation based method for three-dimensional molecular similarity determination. *Bull. Chem. Soc. Jpn.* **94**, 655–665 (2021).

56. Uyanık, T., Karatuğ, Ç. & Arslanoğlu, Y. Machine learning approach to ship fuel consumption: A case of container vessel. *Transp. Res. D.-T. E.* **84**, 102389 (2020).

57. Chen, S., Cao, H., Ouyang, Q., Wu, X. & Qian, Q. ALDS: An active learning method for multi-source materials data screening and materials design. *Mater. Des.* **223**, 111092 (2022).

58. Faraji Niri, M., Reynolds, C., Román Ramírez, L. A. A., Kendrick, E. & Marco, J. Systematic analysis of the impact of slurry coating on manufacture of Li-ion battery electrodes via explainable machine learning. *Energy Storage Mater.* **51**, 223–238 (2022).

59. Bishop, C. M. & Nasrabadi, N. M. *Pattern Recognition and Machine Learning*. 4 (Springer), (2006).

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Correspondence and requests for materials should be addressed to Dawei Zhang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.