

# Predicting the anion conductivities and alkaline stabilities of anion conducting membrane polymeric materials: development of explainable machine learning models

Yin Kan Phua <sup>a</sup>, Tsuyohiko Fujigaya <sup>a,b,c</sup> and Koichiro Kato <sup>a,c,d</sup>

<sup>a</sup>Department of Applied Chemistry, Graduate School of Engineering, Kyushu University, Fukuoka, Japan;

<sup>b</sup>International Institute for Carbon Neutral Energy Research, Kyushu University, Fukuoka, Japan;

<sup>c</sup>Center for Molecular Systems, Kyushu University, Fukuoka, Japan;

<sup>d</sup>Research Institute for Information Technology, Kyushu University, Fukuoka, Japan

## ABSTRACT

Anion exchange membranes (AEMs) are core components in fuel cells and water electrolyzers, which are crucial to realize a sustainable hydrogen society. The low anion conductivity and durability of AEMs have hindered the commercialization of AEM-based devices, and research and development (R&D) to improve AEM materials is often resource-intensive. Although machine learning (ML) is commonly used in many fields to accelerate R&D while reducing resource consumption, it is rarely used in the AEM field. Three problems hinder the adoption of ML models, namely, the low explainability of ML models; complication with expressing both homopolymers and copolymers in unity to train a single ML model; and difficulty in building a single ML model that comprehends various polymer types. This study presents the first ML models that solve all three problems. Our models predicted the anion conductivity for a diverse set of unseen AEM materials with high accuracy (root mean squared error = 0.014 S cm<sup>-1</sup>), regardless of their state (freshly synthesized or degraded). This enables virtual pre-synthesis screening of novel AEM materials, reducing resource consumption. Moreover, human-comprehensible prediction logic revealed new factors affecting the anion conductivity of AEM materials. Such capability to reveal new important variables for AEM materials design could shift the paradigm of AEM R&D. This proposed method is not limited to AEM materials, instead it presents a technology that is applicable to the diverse set of polymers currently available.

## ARTICLE HISTORY

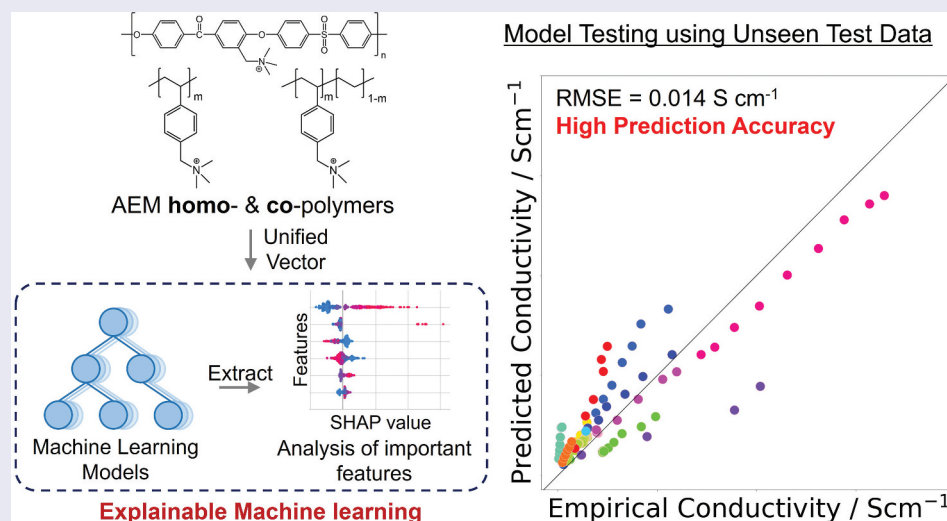
Received 22 June 2023

Revised 5 September 2023

Accepted 18 September 2023




## KEYWORDS


Machine learning models; explainable AI; anion exchange membrane; functional polymers; fuel cell; data-driven



## IMPACT STATEMENT

This study reports a transparent and trustable machine learning model for use in researching polymeric materials, fueling the momentum towards data-driven research, which will significantly reduce research-originating environmental impact.

**CONTACT** Tsuyohiko Fujigaya  [fujigaya.tsuyohiko.948@m.kyushu-u.ac.jp](mailto:fujigaya.tsuyohiko.948@m.kyushu-u.ac.jp); Koichiro Kato  [kato.koichiro.957@m.kyushu-u.ac.jp](mailto:kato.koichiro.957@m.kyushu-u.ac.jp)  Department of Applied Chemistry, Graduate School of Engineering, Kyushu University, 744 Motoooka, Nishi-ku, Fukuoka 819-0395, Japan

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/14686996.2023.2261833>

© 2023 The Author(s). Published by National Institute for Materials Science in partnership with Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

## 1. Introduction

The ever-increasing demand for clean-energy sources that emit minimal to no greenhouse gases during operation has resulted in a shift in attention from fossil fuels to fuel cells, especially toward polymer-electrolyte membrane fuel cells [1]. Polymer-electrolyte membrane fuel cells are receiving increasing industrial and academic attention as they can be used to achieve a zero-carbon emission society [2]. Polymer-electrolyte membrane fuel cells are energy-conversion devices that can efficiently generate electrical energy from chemical energy by breaking the bonds in gaseous hydrogen and oxygen molecules [3]. Water is the only side product of this process [4]. Polymer electrolyte membrane fuel cells can be primarily classified into two categories, namely, proton-exchange membrane fuel cells (PEMFCs) and anion-exchange membrane fuel cells (AEMFCs). PEMFCs are well-developed systems that are fabricated using perfluorinated sulfonic acid membranes such as Nafion, which is considered a benchmark system for developing proton-exchange membranes (PEMs)). PEMs exhibit high proton conductivities, excellent mechanical properties, and good chemical stabilities [5]. In contrast, anion-exchange membranes (AEMs), which are used to fabricate AEMFCs, do not employ such benchmarks [6]. Several commercial AEMs have been reported, and these are characterized by low anion conductivities ( $<0.1 \text{ S cm}^{-1}$ ) [7–10] and poor chemical stabilities ( $<1,000$  hours) [8,11–16]. Such poor AEM properties limit their application prospects. Several AEMs with anion conductivities exceeding  $0.1 \text{ S cm}^{-1}$  have been reported in recent years [17–20]. Although these AEMs exhibit high anion conductivities, their chemical stabilities remain below several thousand hours [12]. Moreover, the anion conductivities and chemical stabilities of most reported AEMs were poorer than those of PEMs [21].

Improving the ion conductivity of AEMs is intrinsically harder than that of PEMs. This can be attributed to the fact that  $\text{OH}^-$ , which is a major contributor of ion conduction in AEMs, exhibits low conductivity. Even in aqueous solution,  $\text{OH}^-$ -based conductivity is equivalent to just 0.568, or slightly more than half of  $\text{H}^+$ -based conductivity ( $\text{H}^+$  often contributes to ion conduction in PEMs) [10]. Addressing the problems associated with low chemical stability is difficult, as  $\text{OH}^-$  readily attacks the anion-conducting functional groups or polymer backbones present in the systems [11,12,22–24]. The ion-exchange capacities of AEMs can be improved using various methods, for example, by incorporating a large number of ion-conducting functional groups into the membrane. This method helps improve ion conductivity but simultaneously increases the water uptake and deteriorates the mechanical properties of these materials [10,25].

Thus, designing AEMs that simultaneously exhibit high anion conductivities and chemical stabilities is difficult. Efforts have been made to simultaneously improve the anion conductivities and chemical stabilities of AEMs by altering the molecular structure of the polymers [8]. Such enhancement efforts are aimed at improving the extent of hydrophilic/hydrophobic phase separation, hydrophilic domain connectivity, and controlling the water uptake of the materials [8]. Flexible side chains [26], multi-cationic groups [27], multi-block copolymer backbones [28], comb-shaped polymers [29], and layered backbones [30] are choices for incorporation into polymers during design as a structural modification to improve anion conductivity. The many combinations available for the incorporation of these building blocks makes it necessary to identify the optimal structures of the materials. The best combination of the design units must be identified to achieve higher anion conductivities and chemical stabilities.

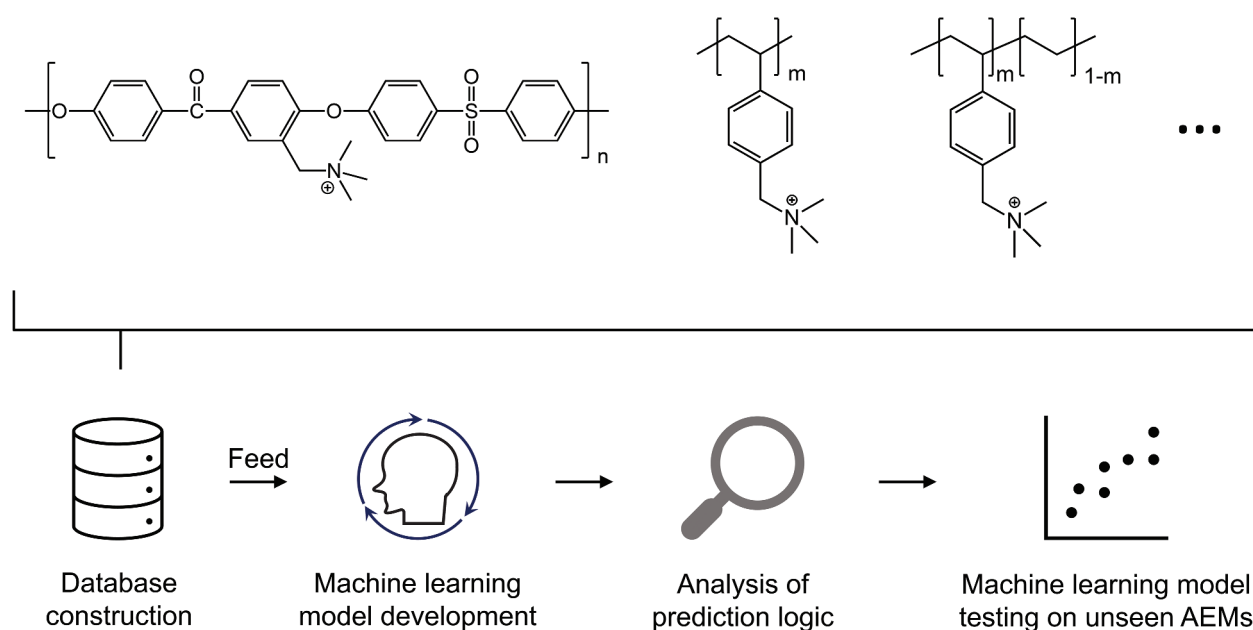
Many studies have been conducted to improve the conductivities of AEMs [26–31]. The trial-and-error-centric (experiment-centric) approach has been primarily used to conduct these studies. This approach involves the consumption of a significant amount of resources, such as funds, labor, and time [32]. Thus, synthesizing large numbers of polymer candidates for AEMs within a realistic timeframe becomes difficult. Various types and combinations of main-chain and side-chain structures have been explored for the fabrication of efficient AEMs [8,23], but numerous polymers and side chains that were not previously used to fabricate AEMs await exploration. New methods should be developed to identify structural unit combinations that can be used to fabricate ideal AEMs with excellent anion conductivities and chemical stabilities in a short time span. Recent advances in information technology have facilitated the emergence of the materials informatics (MI) technique, which uses a data-driven (data-centric) approach and employs the concepts of machine learning (ML), big data analytics, and data mining (associated with the field of materials science) to assist the exploration, discovery, optimization, and development of new materials [32–34]. Such advancements in the research and development (R&D) of materials science is dubbed as the ‘fourth paradigm of materials science’ [35]. These methods have been used in The Harvard Clean Energy Project [36] and The Materials Project [37].

The principles of MI have also been used in the field of polymer science to develop the field of ‘polymer informatics’ (PI) [38–42]. PI is primarily used to predict glass transition temperatures, dielectric constants, and other properties [42,43] directly associated with the main-chain structure of polymers. The relevant data used in PI are often available in open-source polymer databases such as PolyInfo [44]. Zou et al.

[45] and Zhai et al. [46] are the very few groups that studied AEMFCs using MI and reported the results obtained from the analysis of AEMs. Both studies built their own AEM database and used neural networks to develop their ML models. Zou et al. [45] predicted the chemical stabilities (particularly the alkaline stabilities) of AEMs, and Zhai et al. [46] used MI to predict their anion conductivities. Notably, both studies simplified the complexity of AEM polymers by using decomposed structural units as inputs in the ML model that was used to study the AEM polymers. Zou et al. [45] used quaternary ammonium groups as the anion-conducting functional groups and exclusively used the structural units from the main-chain structure as inputs. Zhai et al. [46] fixed the main-chain polymer structure to poly(2,6-dimethyl phenylene oxide) and varied the anion-conducting functional groups. As such, these studies did not focus on the properties and differences of homopolymers and copolymers. Studies wherein the complete chemical structures of AEM polymers were used as input data have not been reported. Cases wherein homopolymers and copolymers were incorporated into a single ML model have also not been reported. Hence, it is important to develop a general model that can be used to explore structurally diverse AEM polymers. Moreover, the prediction logic and results obtained using the ML models should be analyzable and explainable in a scientifically accurate manner to develop a relation of trust between researchers studying AEMs and the ML model. ML models often yield hard-to-interpret prediction logic, which can be attributed to the use of

either exceedingly complex or proprietary functions [47]. Such explainability issues were not dealt with in the reports by Zou et al. [45] and Zhai et al. [46]. This raises questions on the accuracy and reliability of the results. To date, methods for interpreting the prediction logic of ML models have been developed to increase trust between researchers and ML models. Shapley additive explanations (SHAP) [48] and local interpretable model-agnostic explanations (LIME) have been used to achieve this [49]. LIME is not commonly used in the field of MI, whereas SHAP values have been increasingly used for the elucidation of prediction logic [50–54]. Neither SHAP nor LIME has been used in the AEMFC field to date.

Herein, the authors present a comprehensive method that can be used to develop accurate and explainable anion-conductivity- and alkaline-stability-prediction models. This method is trained using the complete chemical structures of a diverse set of AEM polymers (Schemes 1 and S1). Unlike previously reported methods, this method can be used to analyze an extensive range of AEM polymers because it does not limit the polymer structures to be included in the database, with homopolymers and copolymers both being recorded as well. Together with the inclusion of anion conductivity for both freshly synthesized and alkaline-stability-tested AEMs, the applicability of this method is not limited to any specific main-chain structural units, anion-conducting functional groups, or anion conductivity measuring conditions. An AEM polymer database containing a mixture of homopolymers and copolymers was first created. Then, ML



**Scheme 1.** Schematic illustration showing the steps followed in this study. Various polymer structures were collected and included in the database to train the ML model, and the prediction logic corresponding to the model was analyzed to increase the level of model transparency. To evaluate the applicability of the model in real-world settings, it was tested by determining its ability to make predictions for AEM polymer structures that were not used to train the model.

models were built, and each model was combined with SHAP to construct easy-to-understand ML models. These models could accurately predict the anion conductivities of various freshly synthesized AEM polymers and the anion conductivities of AEM polymers subjected to alkaline-stability tests. Furthermore, the ability to interpret the prediction logic corresponding to the three models allowed for the determination of the dominant and controlling factor that dictates the anion-conducting properties of the materials. This could help design new AEM polymers. Thus, this method can be used as a platform to design new AEM polymers.

## 2. Methods

### 2.1. Database construction

Data corresponding to the structural and experimental properties of AEM polymers were collected and coalesced from previously published papers and incorporated into a database to train the ML models. Five review papers [2,8,23,55,56] were selected as the sources for AEM papers with high quality AEM data. This analysis utilizing review papers identified the high-significance AEM papers for prioritization, simultaneously avoiding subjective bias during selection. The extracted data included information on the chemical structure, molar ratio of repeating units, anion conductivity, temperature at which the anion conductivity was measured, anion conductivity achieved during the alkaline stability test, duration of the alkaline stability test, temperature at which the alkaline stability test was conducted, and concentration of the alkaline solution used in the alkaline stability tests. Written data was extracted directly, whereas graphical data was extracted using an online open-source tool ('WebPlotDigitizer' [57]). Data on the conductivity attributable to  $\text{OH}^-$ ,  $\text{HCO}_3^-$ , and  $\text{Cl}^-$  was extracted during data collection. The anion conductivity data was extracted and recorded, and the information was segregated based on the types of ions used during the experiment. Subsequently, the data was incorporated into the database. Although experimental conditions, such as relative humidity, atmospheric gas composition (presence or absence of  $\text{CO}_2$ ), and pre-treatment conditions for membranes significantly affect anion conductivity, detailed information on these parameters was not presented in most papers. Therefore, only the measurement temperature was included as the experimental condition for the anion-conductivity measurements. For the same reasons, the experimental conditions of the alkaline-stability test were limited to the previously listed conditions.

Homopolymer and copolymer data was extracted and incorporated into a single database. The chemical

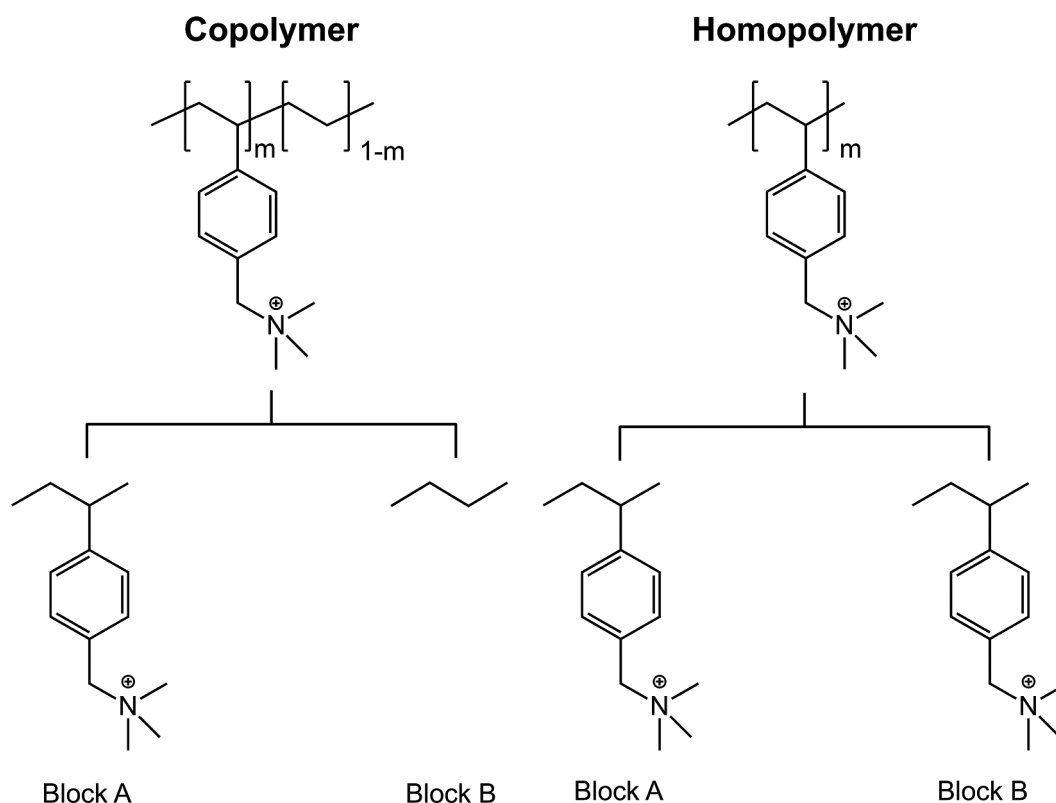
structures of the AEM polymers were segregated into blocks A and B (Scheme 2). For copolymers, block A represents structures containing anion-conducting functional groups, whereas block B does not. For homopolymers, chemical structures containing anion-conducting functional groups were presented by both blocks. The  $\text{CH}_3$  unit was used as the capping unit in each block.

### 2.2. Pre-processing and feature selection

The data corresponding to the extracted chemical structures were transformed into a machine-processable form before using them to train the ML models. The chemical structures of the AEM polymers were drawn using ChemDraw, following which the data was converted into the simplified molecular input line entry specification (SMILES) form using the native function available in ChemDraw for the purpose. Following the generation of the SMILES code, each chemical structure was converted into molecular descriptors using Mordred, an open-source library for molecular descriptor calculations [58]. Mordred is a library that contains widely used chemical descriptors in the field of chemoinformatics [58], such as the RDKit package [59]. RDKit package is a huge library that contains fingerprints such as MACCS Keys, RDKit fingerprint, Morgan fingerprint, MinHash fingerprint, and Avalon fingerprint. In comparison, Mordred is a library that includes not only complete RDKit package, but also other fingerprints and descriptors used in quantitative structure-activity relationship, thereby placing Mordred as a comprehensive descriptor package. As Mordred generates more than 1,600 molecular descriptors for each chemical structure [58], the descriptors for building ML models were selected using the data-pre-processing function in PyCaret [60]. PyCaret is an open-source automated ML tool that enables the determination of the best ML model from a series of ML models available in its library. Minimal coding is required to construct the desired models [60]. For each experimental condition, the condition with the highest repeating frequency was used as a substitute for missing values of the respective experimental condition. The filled-in values represented the dominant experimental conditions in the field of AEM research.

### 2.3. Model development via automated ML

Categorical Boosting (CatBoost) [61], eXtreme Gradient Boosting (XGBoost) [62], and Random Forest (RF) regression models [63] were used as ML algorithms, and these models were built using PyCaret. The three models are tree-based ensemble models that could be effectively used to conduct the research reported herein, as they do not require



**Scheme 2.** Method used to segregate structural data corresponding to AEM polymers for data curation. For copolymers, blocks a and B represent the chemical structures of molecules containing and devoid of anion-conducting functional groups, respectively. The blocks corresponding to homopolymers represent the same chemical structures.

a large amount of data for execution. The amount of data required for using these models is lower than that required for using deep learning models (such as neural networks). Ten iterations were performed during the process of training the models, and a group 10-fold cross-validation (CV) strategy was employed to ensure that splitting of data into training and validation datasets was done according to the AEM polymer structure, and not randomly. This prevents data leakage by keeping conductivity data points obtained from different experimental conditions but of the same AEM polymer structure together. As such, the data in the validation dataset was not seen by the model during training, and vice versa. The model with the lowest average CV error was used as the final predictive model for anion conductivity.

#### 2.4. Performance metrics

The performance of the models were evaluated by analyzing the root mean squared error (RMSE) and mean absolute error (MAE) values. Although these metrics do not indicate whether the values predicted by the model are over- or under-estimated, they quantify the performance and accuracy of the model. RMSE is highly sensitive to large errors and outliers, whereas the sensitivity of MAE toward these factors is low.

A combination of these metrics was used to evaluate the model performance based on the training, validation, and test datasets.

RMSE was calculated as follows:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (1)$$

where  $n$  represents the number of samples,  $y_i$  represents the empirical anion conductivity with  $i = 1, \dots, n$ , and  $\hat{y}_i$  represents the predicted value of  $y_i$ .

MAE was calculated as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

The variables in Equation (2) represent the same factors as those represented by the variables in Equation (1).

#### 2.5. Evaluation of prediction logic

The model prediction logic was evaluated by extracting feature-importance and SHAP values [48] corresponding to the explanatory variables. Only the data that were deemed important (top 20 variables) were visualized. Feature importance evaluates the importance of a feature/variable by computing the rise in the prediction error of ML model following the

permutation of the particular feature/variable [63,64]. SHAP was developed from game theory and served as a tool for mapping the importance of each feature to a particular prediction [48]. The difference in feature importance can be attributed to the ability of SHAP to present the impact of each feature on the prediction output. SHAP does not reflect the influence on the process of model fitting.

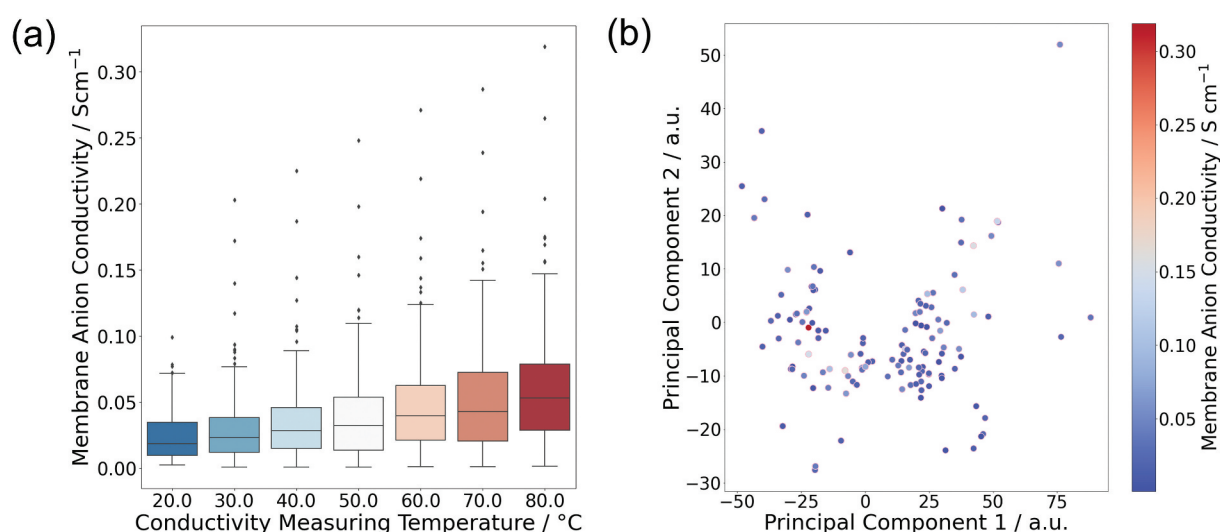
### 3. Results and discussion

#### 3.1. Database construction

A database containing 2,197 anion conductivity-related data from 62 AEM-related papers was built. Data on the temperature-dependent anion conductivity of freshly synthesized AEMs and temperature-independent anion conductivity of AEMs subjected to the conditions of alkaline-stability tests were recorded. The number of papers extracted was comparable to those used to conduct previously reported research [46]. Review papers were chosen as the sources of AEM-related papers. Most of the papers from which the data was extracted were published in or before 2018 (Figure S1a). The temperature-dependent anion-conductivity data for the freshly synthesized AEMs were segregated based on their respective ion forms. Of the 2,197 data points considered to conduct this study, 1,211 corresponded to  $\text{OH}^-$ , 37 corresponded to  $\text{Cl}^-$ , 15 corresponded to  $\text{HCO}_3^-$ , and 2 corresponded to  $(\text{CO}_3)_2^-$ . Thus, a total of 1,265 data points related to the anion conductivity of freshly synthesized AEMs (fresh-anion conductivity) were collected. The data corresponding to  $\text{OH}^-$  was used for ML, considering the differences in the conductivities of  $\text{OH}^-$ ,  $\text{Cl}^-$ , and  $\text{HCO}_3^-$ . The anion conductivity measured during the alkaline-stability tests (degraded-anion conductivity) yielded 932 data

points. These data points corresponded to  $\text{OH}^-$  due to the nature of alkaline-stability tests. The AEMs were submerged in alkaline solutions to simulate the accelerated degradation tests. Hence, a total of 2,197 anion conductivity-related data points were obtained from coalescing fresh- and degraded-anion conductivity data points. The box plot shown in Figure 1(a) presents the relationship between temperature and the distribution of anion conductivity in the case of fresh-anion conductivity. The two parameters were positively correlated with each other. The anion conductivity of the majority of the AEMs was less than  $0.07 \text{ S cm}^{-1}$ . However, the anion conductivity of some AEM polymers was significantly higher than  $0.10 \text{ S cm}^{-1}$  (Figure 1(a)).

Data corresponding to both homopolymers and copolymers were recorded in the constructed database. The method used to incorporate polymer structural data into a single database (which involved segregating AEM polymers into blocks A and B based on their chemical structure) is similar to the method reported by Kuenneth et al. [42], who separated the two monomers of copolymers into two blocks. The previously reported method and the method used in this study differed in the use of capping symbols corresponding to the polymer repeating unit. This study used the  $-\text{CH}_3$  unit (instead of  $*$ ) as the capping unit because it represented a chemical structure, whereas  $*$  is a symbol. The use of the  $-\text{CH}_3$  unit could address the problems faced during the conversion of the data to the Mordred descriptor and the training of the ML models. Capping with  $-\text{CH}_3$  potentially changes the chemical information of the chemical structure since it adds a capping group that carries chemical information on its own. Such side effect can be suppressed by universally applying  $-\text{CH}_3$  capping to all polymers in the database: since all polymers contain  $-\text{CH}_3$  capping, averaging of the



**Figure 1.** (a) boxplot of anion conductivity as a function of the measuring temperature. (b) distribution of the polymer types present in the built database (the data points are colored based on the anion conductivity values).

chemical meaning that  $-\text{CH}_3$  might carry can be expected. The use of  $-\text{CH}_3$  as capping unit is important because it allows the construction of a universal database: one that can be used for both ML and simulation such as density functional theory calculation. This is because  $*$  is incompatible with simulation for use as capping unit, necessitating the use of capping unit that has actual chemical meaning. If simulation is to be done, it will require a separated and dedicated database for simulation that has  $-\text{CH}_3$  capping, making the whole process filled with redundant steps and inefficiency. Such inefficiency could steer potential users away, keeping away the proposed method from becoming a platform for AEM research. Although capping with  $-\text{CH}_3$  has its own disadvantage, the advantage it provides – streamlining the whole process of implementing ML into current research, regardless of the involvement of simulation, outweighs its disadvantage, reinforcing the potential for the proposed approach to establish itself as a robust platform for AEM research. The homopolymer representation mode also differed. This study used the same chemical structure to fill two separate blocks, whereas Kuenneth et al. used the unit monomer corresponding to the homopolymer as the input data for only one of the blocks (block A in this study). The other block (block B in this study) was left empty [42]. Kuenneth et al. multiplied the matrices obtained from the blocks with their respective compositions in the original polymer and added these matrices to form a single matrix of descriptors to be used for ML [42]. This method could not be used in this study as it lowered the degree of explainability of ML by eliminating the information corresponding to the originating blocks of the descriptors. This made mapping the blocks (A or B) with the descriptors and the process of backward tracing difficult. Additionally, matrices with the same value but originating from two different polymers could potentially be obtained due to the summation and multiplication operators involved in the conversion process to a single matrix. Data points corresponding to 15 types of homopolymers and 257 types of copolymers were extracted, indicating that copolymers were the go-to choice for polymer design. The types of polymer main-chains used by researchers and reported in the literature were analyzed, and the relevant data was extracted. Analysis of the distribution plot (Figure 1(b)) generated using the principal component analysis (PCA) method (used to compress the multi-dimensional molecular descriptors into a two-dimensional vector) revealed that the database contains a wide variety of polymer main-chains that span the horizontal axis. PCA is a popular exploratory data analysis technique used to plot the chemical space covered by chemical structures in a particular database [65]. This involves compressing every chemical structure in the database into two dimensions and then plotting the

obtained value to generate a scatter plot. Several relatively dense or sparse areas were identified in the plot, but overall, a database without a significantly biased polymer main-chain-type distribution was built. A gradient of colors (from blue to red) was used to represent the points corresponding to the fresh-anion conductivity values (blue and red represent low and high conductivity, respectively). A clear relationship between the position of the AEM chemical structures in the plot and the conductivity was not demonstrated, albeit AEMs characterized by high conductivities (exceeding  $0.10 \text{ S cm}^{-1}$ ) were primarily situated to the left of the plot (Figure 1(b)). Thus, investigating potential candidates for the development of new AEM polymers by solely analyzing the location of the potential candidates in the distribution plot was difficult. This is where advanced techniques, such as ML, can be used for speeding up the R&D of AEM polymers. The top 10 most commonly used types of polymer main-chains for block A are shown in Figure S1b, with poly(2,6-dimethyl-1,4-phenylene oxide) being the most frequently used main-chain structure for block A. A total of 46 AEMs containing this structural unit were identified. Poly(arylene ether sulfone), which is present in 59 AEMs, is the most frequently used main-chain for block B (Figure S1c). The most widely used anion-conducting moiety is trimethylammonium (Figure S1d), which is present in 91 AEMs. However, the use of the most widely used moiety does not necessarily result in the generation of high conductivity. For example, the fresh-anion conductivity of AEMs such as quaternary trimethylammonium functionalized poly(2,6-dimethyl-1,4-phenylene oxide) is  $<0.01 \text{ S cm}^{-1}$  [66].

### 3.2. Construction of regression models for the prediction of anion conductivity and alkaline-stability tests

The chemical structures were converted to 3,226 chemical descriptors using Mordred. Specifically, 1,613 descriptors were generated for blocks A and B. Descriptors with non-numerical values (such as those with strings or Not a Number values), those with the same value for every AEM polymer, and those with perfect collinearity or multicollinearity were eliminated, and a total of 522 chemical descriptors remained. The remaining descriptors (together with the data corresponding to the molar ratio associated with blocks A and B, the temperature at which the conductivity values were measured, the number of days over which the alkaline stabilities of the compounds were studied, and the temperature at which the alkaline-stability tests were conducted) were used as the explanatory variables. The fresh- and degraded-anion conductivity values measured in the  $\text{OH}^-$ -form were considered as the target variables. During ML,

the database was split in the ratio of 95:5 based on the types of AEM polymer chemical structures. The ratio reflects that 95% of the data was used for training and validation, and the remaining 5% was isolated for testing. A group 10-fold CV was performed using the previously obtained train-validation data. The fresh- and degraded-anion conductivities for most of the AEMs in the database were 0.01–0.05 S cm<sup>-1</sup> (Figures 1(a) and S2).

The RMSE and MAE values (obtained using the trained models (Table 1)) calculated using the fresh- and degraded-anion conductivity values output during training were smaller by an order of magnitude than the range of the empirically measured conductivity values. This indicates that the models were successfully trained. For the validation results, the RMSE and MAE values were also smaller than the empirical conductivity values by an order of magnitude (Table 1). The validation step involved using the models to estimate the conductivity of AEM polymers that were not used during training. Thus, high prediction accuracy and generalizability can be expected using the proposed model in real-world settings. Linear regression models, such as Ridge and Lasso regression, were also used for the analysis, and the RMSE values obtained during validation using these models were 0.0357 and 0.0322 S cm<sup>-1</sup>, respectively. Three separate models were individually built using scikit-learn [67] intended for use as a baseline: support vector regression (SVR; kernel used: rbf) [68], gaussian process regression (GPR; kernel used: ConstantKernel(1.0, constant\_value\_bounds='fixed') \* RBF(1.0, length\_scale\_bounds='fixed')) [69], and multi-layer perceptron regression [70] (MLP; 2 hidden layers, 256 nodes). Their RMSE obtained from cross-validation were 0.0712, 0.0543, and 0.0344 S cm<sup>-1</sup>, respectively. The orders of magnitude of all these RMSE values were the same as those of the AEM anion conductivities. The poor validation accuracy obtained using the Ridge and Lasso regression models can be attributed to the fact that these are linear regression models, which model the relationship between the explanatory and the target variable using linear predictor functions. However, the anion conductivity values of the AEM systems did not exhibit a linear dependency on their chemical structures. This indicated that linear regression models could not effectively analyze the non-linear relationship between the parameters to yield predicted anion conductivity values comparable with the actual

values. For MLP, the ineffectiveness of these models to deal with extremely high dimension yet small sample dataset led to such results, due to their tendency to overfit [71]. As for GPR and SVR, cross-validation, which further splits a small dataset smaller, might have largely contributed to overfitting [72].

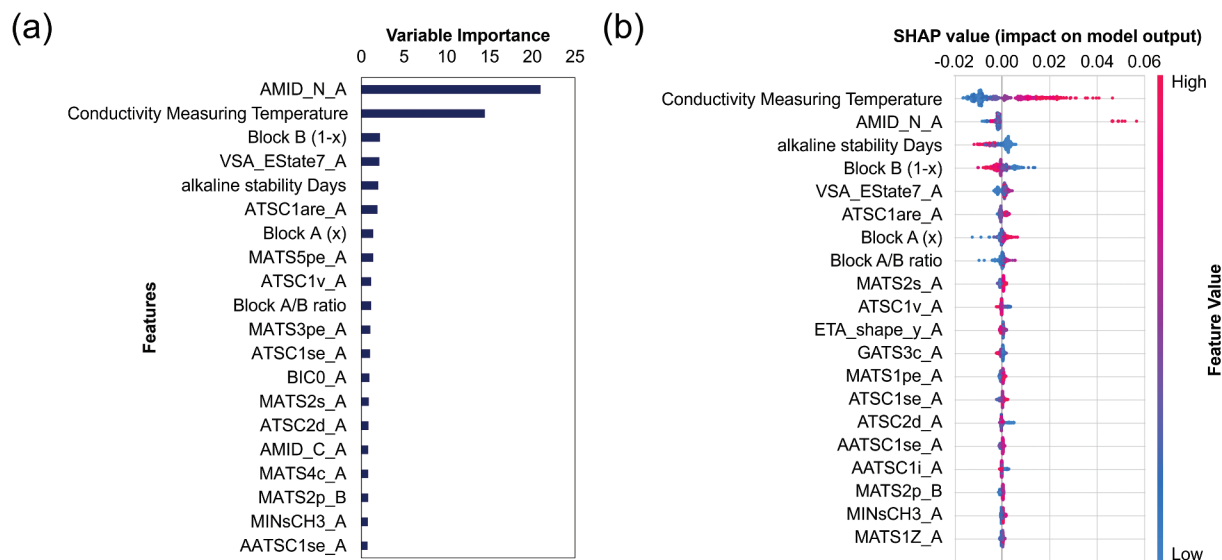
### 3.3. Interpretation of prediction logic

In addition to building ML models with high prediction accuracies, the prediction logic must be effectively analyzed to develop trust between researchers and ML models. The focus on analyzing the prediction logic associated with ML models in the field of MI has recently increased. The methods used for analyzing prediction logic are primarily classified into two categories, namely, cases where feature importance is used for analysis [45] and those where a local interpreter (the SHAP value) is used for analysis [50–54]. The results obtained from both the feature importance and SHAP value were compared and analyzed (Figure 2, CatBoost; Figure S3, XGBoost and RF; the results for the top 20 features are explained in Table S1). The feature importance plots were generated using the three developed models (Figures 2(a) and S3(a–b)). The most common feature of importance was AMID\_N\_A (the average molecular identification number corresponding to the nitrogen atom [73] in block A). This feature is a descriptor of the chemical structure and presents the average of the total number of weighted paths corresponding to nitrogen atoms [73]. The feature importance plot indicates that it is unclear whether a large or small AMID\_N\_A value maximizes anion conductivity (Figures 2(a) and S3(a–b)). Approximately 10–13 features labeled 'ATS' ranked among the top 20 features identified using the three models, with each of them carrying a different meaning that is defined according to their respective complete name. The features that ranked in the top 20 list differed from model to model but could be grouped according to part of the acronym included in their name. All features represented by 'ATS' are related to the autocorrelation coefficient function, which represents the descriptors used to calculate a value representing the chemical structure based on the topology of the molecules [58]. Features containing the term 'Estate' in their names are associated with the electrotopological state of the molecules [74] and reflect the distribution of electrons. Again, analysis of the feature importance plots did not reveal if a large or small value of the feature results in the maximization of anion conductivity (Figures 2(a) and S3(a–b)). For commonly known factors that affect anion conductivity, such as the measuring temperature, it is unnecessary for the plot to reflect whether the feature should take a larger or smaller value since it is known that higher temperatures typically lead to higher

**Table 1.** Summary of the train-validation accuracy corresponding to CatBoost, XGBoost, and RF in terms of RMSE and MAE.

Model	RMSE (S cm <sup>-1</sup> )		MAE (S cm <sup>-1</sup> )	
	Train	Validation	Train	Validation
CatBoost	0.00290	0.00600	0.00170	0.00350
XGBoost	0.00220	0.00700	0.00110	0.00380
RF	0.00360	0.00840	0.00190	0.00470





**Figure 2.** (a) feature importance and (b) SHAP plots generated using CatBoost.

conductivity. However, there is a lack of consensus on uncommon or newly found important features. The lack of information on the relationship between the value of the feature and anion conductivity makes designing AEM polymers that reflect the effects of the features difficult, potentially discouraging the use of ML techniques.

The SHAP plots were analyzed (Figure 2(b), CatBoost; Figure S3c and d, XGBoost and RF; details of the top 20 features are explained in Table S1), and the top 20 features identified by the ML models, which can be used to predict anion conductivity, have been presented. A consensus was reached regarding the results obtained using the three models. It is worth noting that main chain capping-related descriptors were not present in the top 20 important variables (SHAP nor feature importance), confirming that using  $-\text{CH}_3$  as capping unit did not result in any adverse effect of concern. The most common features that were deemed to be important are the temperature at which conductivity is measured and AMID\_N\_A. These two features contribute the most to the determination of the anion conductivity of the AEM systems. As mentioned previously, temperature significantly affects the anion conductivities of AEMs, which increase with temperature. Considerable information can be obtained from the SHAP plots: the pink and blue data points represent high and low feature values, respectively. The effects (toward anion conductivity) of the data points located toward the positive direction of the horizontal axis are greater than the effects of the values in the negative direction. Analysis of the SHAP plots revealed a positive relationship between AMID\_N\_A and anion conductivity. This is reflected by the pink data points present on the far right of the plots. The models managed to effectively reflect the important features of

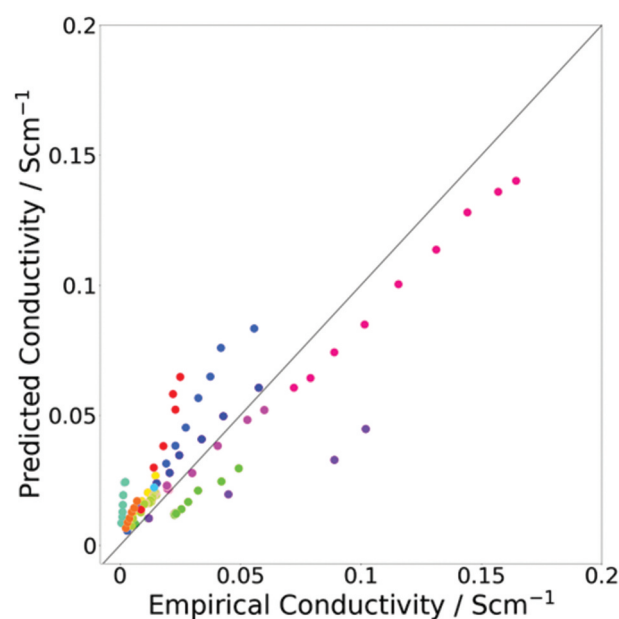
the AEM polymer structure, because AMID\_N\_A is related to nitrogen, which is often present in anion-conducting functional groups. However, an increase in AMID\_N\_A is not directly associated with an increase in the number of anion-conducting functional groups in the AEM polymers, even though the latter often results in high anion conductivity. Instead, it indicates a high average of the total number of weighted paths associated with the nitrogen atom. This can be attributed to the fact that an AEM polymer containing three anion-conducting functional groups (fresh-anion conductivity:  $0.175 \text{ S cm}^{-1}$  at  $80^\circ\text{C}$ ; AMID\_N\_A: 0.20) [27] is characterized by an AMID\_N\_A value (0.20) that is lower than the value characterizing the polymer containing two anion-conducting functional groups (fresh-anion conductivity:  $0.319 \text{ S cm}^{-1}$  at  $80^\circ\text{C}$ ; AMID\_N\_A: 0.43) [75]. In particular, AMID\_N\_A presents the position of anion-conducting functional groups in the polymer, which is represented in numerical form. Increasing AMID\_N\_A also showed three separate trends toward the effect on anion conductivity: i. rapid local maximization towards AMID\_N\_A = 0.079, followed by slow decline; ii. rather gentle local maximization towards AMID\_N\_A = 0.204, followed by steeper decline than the first maxima; iii. steep increase towards global maximum at AMID\_N\_A = 0.433 (Figure S4a). In general, increase in AMID\_N\_A could lead to increase in anion conductivity, but care must be taken to not fall into the valleys in the plot that leads to subpar anion conductivity. Other important features positioned among the top 20 features, identified by the models, are the number of days corresponding to the alkaline-stability test (duration of soaking AEM in alkaline solutions) and the molar ratio between blocks A and B. A common consensus is available for these features: an increase in the soaking

time of an AEM in an alkaline solution results in a decrease in its anion conductivity; an increase in block A of an AEM, which contains a cationic group, typically increases its anion conductivity. However, it is interesting to note that simply increasing block A ratio in an AEM polymer does not always result in high anion conductivity, as too much of block A will lead to sharp decline of anion conductivity (Figure S4b), most likely due to weak mechanical strength of the membrane formed. Most of the features in the list contained the term 'ATS' in their names, and as mentioned previously, these features are chemical descriptors related to the autocorrelation coefficient function. Such descriptors are generally suggested to not be related to the target variable [76,77], but the analysis of the SHAP plots indicates the opposite for the case of using anion conductivity as the target variable. The correlation between all variables deemed to be important by the three models and anion conductivity were confirmed using correlation matrix (Figure S4c). Block A ratio and AMID\_N\_A indeed has positive correlation, and several ATS-es has negative correlation, coherent to the relationship shown by SHAP plots. Developing or designing AEM polymers characterized by a high average total number of weighted paths for the nitrogen atom or topological shape is difficult, as the guidelines presented for them based on such explanations are difficult to understand. Instead, using new approaches to design molecules, such as the Monte Carlo tree search-based molecular generative model [78], can potentially use such guidelines to rapidly design new AEM polymers. For example, novel AEM polymers characterized by high anion conductivities may be developed by setting a high target for AMID\_N\_A during the molecule generation process, preferably those exceeding 0.433 due to the lack of exploration beyond 0.433 (Figure S4a). The advantage of the SHAP plot lies in the ability to analyze and understand the function and effect of each feature on the ease of AEM to conduct anions, whereas such detailed information cannot be obtained by analyzing feature-importance plots. The use of a SHAP plot to gain insights into the results may aid the development of a highly transparent ML model. Therefore, the analysis of SHAP plots is recommended during the implementation of ML models in R&D associated with AEMs.

### 3.4. Unseen data prediction test

The models were used to predict the results for a set of AEMs that were not used during training to ensure the applicability of the models under real-world settings, with the polymers used named as 'unseen AEM polymers'. The models were used to predict both the fresh- and degraded-anion conductivities of 14 unseen AEM polymers (chemical structures shown in Figure S5a – n),

all of which were excluded from the train-validation dataset. The accuracy obtained using the models (Figure 3 for overall results of CatBoost; Figure S6 for XGBoost and RF; Figure S7a – n for the fresh-anion conductivity results of individual AEM polymers, with individual AEM polymers having a call sign of unseen a – n, respectively; Figure S8a – g for the results of degraded-anion conductivities, covering unseen b, c, f, g, i, j, and m polymers) was in the range of 0.01393–0.016037 S cm<sup>-1</sup> for the RMSE and 0.009879–0.011407 S cm<sup>-1</sup> for the MAE of all 14 polymers (Table 2 and Figure S9; distribution plot). Analysis of the overall prediction results revealed that the best prediction performance for unseen AEM polymers can be obtained using CatBoost, with RMSE and MAE values as low as 0.013930 and 0.009879 S cm<sup>-1</sup>, respectively. The worst results were obtained using XGBoost. Overall, the models over- or under-estimated several of the unseen polymers, with prediction results matching the exact value of empirical anion conductivity remain rare. This is attributed to the fact that ML models tend to over- or underestimate the anion conductivity of an AEM polymer depending on the amount of data available for similar polymers in the high or low anion conductivity region, respectively. In this study, unseen a, c, d, k, and l was overestimated, while unseen f was underestimated. The reason behind such trend and behavior of ML models were clarified in the distribution plot for empirical anion conductivity (of all AEM polymers included in the database) against temperature (Figure S10). Large density of high conductivity data for AEM polymers similar to the structure of unseen a, c, d, k, and l (Figure S10b) might



**Figure 3.** Predicted conductivity versus empirical conductivity plot generated using CatBoost. Different colors indicate different unseen AEM polymers. Multiple dots of the same color represent temperature-dependent and alkaline-stability-test-based prediction values for the same AEM polymers, where the empirical data was extracted from their respective paper.

**Table 2.** General prediction accuracies obtained using CatBoost, XGBoost, and RF in terms of RMSE and MAE for 14 unseen AEM polymers.

AEM polymer	RMSE ( $S\text{ cm}^{-1}$ )			MAE ( $S\text{ cm}^{-1}$ )		
	CatBoost	XGBoost	RF	CatBoost	XGBoost	RF
All	0.013930	0.016037	0.014827	0.009879	0.011407	0.010817

have led to overestimation, while large density of low conductivity data for AEM polymers similar to unseen f (Figure S10f) might have led to underestimation. Nevertheless, the performance of all models was comparable, with differences of only approximately  $\pm 0.001\text{ S cm}^{-1}$ , revealing that the models can appropriately understand the relationship between the chemical structures, anion conductivity, and alkaline-stability test conditions. The fresh- (Table S3) and degraded-anion conductivity (Table S4) values for the unseen b, c, f, g, i, j, and m polymers (referred to as the ‘seven unseen polymers’) were analyzed because degraded-anion conductivity data was present in their respective reports. Comparing the prediction accuracy between the fresh- and degraded-anion conductivities for the seven unseen polymers, the models were more effective at analyzing the degraded-anion conductivity than fresh-anion conductivity of the materials. The RMSE and MAE values for degraded-anion conductivity were lower than  $0.01\text{ S cm}^{-1}$  for most of the seven unseen polymers, regardless of the model used. The best results for fresh-anion conductivity were obtained using CatBoost, with the least number of unseen polymers (four out of seven unseen polymers with alkaline stability test results provided in their respective paper) characterized by RMSE and MAE values  $> 0.01\text{ S cm}^{-1}$ , whereas XGBoost and RF fared worse and yielded equally poor results (five out of seven). Good RMSE and MAE values were obtained when the degraded-anion conductivity of the seven unseen polymers was analyzed. This can be attributed to the fact that detailed information on the systems other than the AEM polymer structure can be obtained by conducting alkaline-stability tests and used as parameters: the period for which AEM was soaked in the alkaline solution, temperature of the alkaline solution during the alkaline-stability test, and concentration of the alkaline solution. In contrast, during the analysis of fresh-anion conductivity, only the temperature at which anion conductivity was measured was considered as a parameter.

The RMSE values obtained using the models were mostly comparable with, and in some cases lower than, the values reported by Zhai et al. (the RMSE values were  $0.007\text{--}0.025\text{ S cm}^{-1}$  depending on the functional group [46]). (Table S2). Only a few unseen AEM polymers were characterized by RMSE values higher than the reported range (Table S2). Generally, it is more challenging and complex to predict the fresh- and degraded-anion conductivities for a set of structurally diverse AEM

polymers than those faced during the prediction of the fresh-anion conductivity of a specific set of AEM polymers, assuming that the same ML model was used. These challenges were overcome, and low RMSE values were obtained when the ML models reported herein were used. This signifies that high prediction accuracy could be achieved even with a limited amount of training and validation data, and the success in unitarily representing homopolymers and copolymers (Scheme 2) might have also contributed to overcoming such challenges. Notably, this method used less data than that required by ML models based on neural networks, such as those reported by Zhai et al. and Zou et al. [45,46]. Thereby, the three models used herein could efficiently and effectively understand the relationship between explanatory variables (chemical structure descriptors, experimental conditions, and polymer structure information) and target variables (anion conductivity). As mentioned, the high prediction accuracy achieved for the test data revealed that a single model can predict both the fresh- and degraded-anion conductivities of various AEM polymers. This demonstrates the high potential for this method to significantly reduce the frequency of conducting a resource-intensive degradation test during materials exploration, which can potentially help to streamline the R&D process associated with the production of novel AEM polymeric materials. Together with the explainable results that were obtained by analyzing the SHAP plots in this study, a versatile, transparent, and easy-to-execute method was developed for incorporation into the AEM polymer R&D cycle. The developed method can be used to predict both the fresh- and degraded-anion conductivities of yet-to-be synthesized novel AEM polymers designed in-house. This helps to accelerate the R&D process by reducing the number of synthetic cycles followed. Although this study focused on developing a method for the analysis of AEM polymers, the proposed method can be used to investigate all types of polymers, including those used to fabricate gas-separation membranes, lithium-ion conducting membranes, and other functional polymers. Naturally, this method can be applied to the R&D of general polymers as well. Similar to the field of AEM polymers, reports on ML models trained using a set of structurally diverse polymers are scarce. Therefore, the results

reported herein can be used to develop a platform for developing methodologies aimed at implementing the ML technique in the field of polymers.

#### 4. Conclusions

The results reported herein revealed that a single model can be used to predict the fresh- and degraded-anion conductivities of materials. This indicates the increased versatility of the ML models used in the R&D processes associated with the materials exploration of AEM polymers. The models (CatBoost, XGBoost, and RF) are easy to build and use, and good prediction accuracies for both fresh- and degraded-anion conductivities for unseen AEM polymers were achieved using them. The SHAP values were analyzed to study the transparency and explainability of these models. The importance of individual features could be understood and an in-depth analysis of the features could be conducted by analyzing how the values of each feature (corresponding to each AEM polymer) affect the anion conductivity and alkaline stability of the materials. The ability to visualize the vector of impact for each feature that is deemed important by ML models is the first step toward achieving transparency in ML prediction logic. Simultaneously, the difficulties in interpreting the important chemical-structure descriptors, which are those obtained from currently available and widely used descriptors in the chemoinformatic field, were also discussed herein. Although it is difficult to manually implement the important features originating from chemical-structure descriptors into the design of AEM polymers, molecular generative models can be used to optimize these features if used as the target variables for such models. This gives hope to rapidly design AEM polymers with high anion conductivities and alkaline stabilities. By overcoming difficulties in manual interpretation by developing highly interpretable chemical descriptors in future studies, the proposed approach can further accelerate the development of explainable ML for use in AEM R&D.

#### Abbreviations

AEM: anion exchange membrane; AEMFC: anion-exchange membrane fuel cell; CatBoost: Categorical Boosting; CV: cross-validation; GPR: gaussian process regression; LIME: local interpretable model-agnostic explanation; MAE: mean absolute error; ML: machine learning; MLP: multi-layer perceptron regression; PCA: principal component analysis; PEMFC: proton-exchange membrane fuel cell; PI: polymer informatics; R&D: research and development; RF: Random Forest; RMSE: root mean squared error; SHAP: Shapley additive explanation; SMILES: simplified molecular input line entry specification; SVR: support vector regression; XGBoost: eXtreme Gradient Boosting.

#### Acknowledgment

Y. K. P. thanks the Sato Yo International Scholarship Foundation of Japan for financial support. The authors thank the Ministry of Education, Culture, Sports, Science and Technology (MEXT) for the financial support. This study was supported by the Japan Science and Technology Agency (JST), ACT-X [Grant Number JPMJAX22AF], Japan, the establishment of university fellowships towards the creation of science technology innovation [grant no. JPMJFS2132], and the “Engineering Research for Pioneering of a New Field” grant provided by the Faculty of Engineering, Kyushu University, Japan. This study was also supported by JSPS KAKENHI [Grant Number JP23H02027]. The authors thank the Robert T. Huang Entrepreneurship Center of Kyushu University (QREC), Japan, for supporting the project via “Academic Challenge 2021” grant.

#### Author contributions

Yin Kan Phua: Investigation and Writing – Original Draft. Tsuyohiko Fujigaya: Supervision, Writing – Review and editing. Koichiro Kato: Conceptualization, Supervision, Writing – Review and Editing.

#### Disclosure statement

The AEM database and source codes that support the findings of this study are not publicly accessible as they are the Intellectual Property of Kyushu University. However, they may be made available promptly to requester upon reasonable request for academic use.

#### Funding

This work was supported by the ACT-X [JPMJAX22AF]; Japan Science and Technology Agency [JPMJFS2132]; Japan Society for the Promotion of Science [JP23H02027]; Sato Yo International Scholarship Foundation [Scholarship program for Foreign Studies]; Robert T. Huang Entrepreneurship Center of Kyushu University (QREC) [Academic Challenge 2021].

#### ORCID

Yin Kan Phua  <http://orcid.org/0000-0002-9668-9775>  
Tsuyohiko Fujigaya  <http://orcid.org/0000-0003-3563-8234>  
Koichiro Kato  <http://orcid.org/0000-0003-4392-8741>

#### References

- [1] Luo Y, Wu Y, Li B, et al. Development and application of fuel cells in the automobile industry. *J Energy Storage*. 2021;42:103124. doi: 10.1016/j.est.2021.103124
- [2] Vijayakumar V, Nam SY. Recent advancements in applications of alkaline anion exchange membranes for polymer electrolyte fuel cells. *J Ind Eng Chem*. 2019;70:70–86. doi: 10.1016/j.jiec.2018.10.026
- [3] Sasaki K. Hydrogen energy engineering: a Japanese perspective. 1st ed. Tokyo (Japan): Springer Japan;

2016. Chapter 2, Current Status: General; p. 15–35. doi: [10.1007/978-4-431-56042-5\\_2](https://doi.org/10.1007/978-4-431-56042-5_2)
- [4] Wang Y, Chen KS, Mishler J, et al. A review of polymer electrolyte membrane fuel cells: technology, applications, and needs on fundamental research. *Appl Energy*. 2011;88(4):981–1007. doi: [10.1016/j.apenergy.2010.09.030](https://doi.org/10.1016/j.apenergy.2010.09.030)
- [5] Mauritz KA, Moore RB. State of understanding of nafion. *Chem Rev*. 2004;104(10):4535–4586. doi: [10.1021/cr0207123](https://doi.org/10.1021/cr0207123)
- [6] Hickner MA. Strategies for developing New anion exchange membranes and electrode ionomers. *Electrochem Soc Int*. 2017;26(1):69–73. doi: [10.1149/2.F08171if](https://doi.org/10.1149/2.F08171if)
- [7] Zhegur-Khais A, Kubannek F, Krewer U, et al. Measuring the true hydroxide conductivity of anion exchange membranes. *J Membr Sci*. 2020;612:118461. doi: [10.1016/j.memsci.2020.118461](https://doi.org/10.1016/j.memsci.2020.118461)
- [8] Gottesfeld S, Dekel DR, Page M, et al. Anion exchange membrane fuel cells: current status and remaining challenges. *J Power Sources*. 2018;375:170–184. doi: [10.1016/j.jpowsour.2017.08.010](https://doi.org/10.1016/j.jpowsour.2017.08.010)
- [9] Ferriday TB, Middleton PH. Alkaline fuel cell technology - a review. *Int J Hydrogen Energy*. 2021;46(35):18489–18510. doi: [10.1016/j.ijhydene.2021.02.203](https://doi.org/10.1016/j.ijhydene.2021.02.203)
- [10] Pan J, Chen C, Zhuang L, et al. Designing advanced alkaline polymer electrolytes for fuel cell applications. *Acc Chem Res*. 2012;45(3):473–481. doi: [10.1021/ar200201x](https://doi.org/10.1021/ar200201x)
- [11] Cheng J, He G, Zhang F. A mini-review on anion exchange membranes for fuel cell applications: stability issue and addressing strategies. *Int J Hydrogen Energy*. 2015;40(23):7348–7360. doi: [10.1016/j.ijhydene.2015.04.040](https://doi.org/10.1016/j.ijhydene.2015.04.040)
- [12] Arges CG, Zhang L. Anion exchange membranes' evolution toward high hydroxide ion conductivity and alkaline resiliency. *ACS Appl Energy Mater*. 2018;1(7):2991–3012. doi: [10.1021/acsaem.8b00387](https://doi.org/10.1021/acsaem.8b00387)
- [13] You W, Noonan KJT, Coates GW. Alkaline-stable anion exchange membranes: a review of synthetic approaches. *Prog Polym Sci*. 2020;100:101177. doi: [10.1016/j.progpolymsci.2019.101177](https://doi.org/10.1016/j.progpolymsci.2019.101177)
- [14] Adhikari S, Pagels MK, Jeon JY, et al. Ionomers for electrochemical energy conversion & storage technologies. *Polymer*. 2020;211:123080. doi: [10.1016/j.polymer.2020.123080](https://doi.org/10.1016/j.polymer.2020.123080)
- [15] Thompson ST, Peterson D, Ho D, et al. Perspective—the next decade of AEMFCs: near-term targets to accelerate applied R&D. *J Electrochem Soc*. 2020;167(8):084514. doi: [10.1149/1945-7111/ab8c88](https://doi.org/10.1149/1945-7111/ab8c88)
- [16] Lin CX, Wang XQ, Li L, et al. Triblock copolymer anion exchange membranes bearing alkyl-tethered cycloaliphatic quaternary ammonium-head-groups for fuel cells. *J Power Sources*. 2017;365:282–292. doi: [10.1016/j.jpowsour.2017.08.100](https://doi.org/10.1016/j.jpowsour.2017.08.100)
- [17] Mahmoud AMA, Elsaghier AMM, Otsuji K, et al. High hydroxide ion conductivity with enhanced alkaline stability of partially fluorinated and quaternized aromatic copolymers as anion exchange membranes. *Macromolecules*. 2017;50(11):4256–4266. doi: [10.1021/acs.macromol.7b00401](https://doi.org/10.1021/acs.macromol.7b00401)
- [18] Mandal M, Huang G, Hassan NU, et al. Poly(norbornene) anion conductive membranes: homopolymer, block copolymer and random copolymer properties and performance. *J Mater Chem A*. 2020;8(34):17568–17578. doi: [10.1039/d0ta04756b](https://doi.org/10.1039/d0ta04756b)
- [19] Zeng L, He Q, Liao Y, et al. Anion exchange membrane based on interpenetrating polymer network with ultrahigh ion conductivity and excellent stability for alkaline fuel cell. *Research*. 2020;2020:4794706. doi: [10.34133/2020/4794706](https://doi.org/10.34133/2020/4794706)
- [20] Douglin JC, Varcoe JR, Dekel DR. A high-temperature anion-exchange membrane fuel cell. *J Power Sources Adv*. 2020;5:100023. doi: [10.1016/j.powera.2020.100023](https://doi.org/10.1016/j.powera.2020.100023)
- [21] Mustain WE, Chatenet M, Page M, et al. Durability challenges of anion exchange membrane fuel cells. *Energy Environ Sci*. 2020;13(9):2805–2838. doi: [10.1039/d0ee01133a](https://doi.org/10.1039/d0ee01133a)
- [22] Varcoe JR, Atanassov P, Dekel DR, et al. Anion-exchange membranes in electrochemical energy systems. *Energy Environ Sci*. 2014;7(10):3135–3191. doi: [10.1039/c4ee01303d](https://doi.org/10.1039/c4ee01303d)
- [23] Xue JD, Zhang JF, Liu X, et al. Toward alkaline-stable anion exchange membranes in fuel cells: cycloaliphatic quaternary ammonium-based anion conductors. *Electrochem Energy Rev*. 2022;5(2):348–400. doi: [10.1007/s41918-021-00105-7](https://doi.org/10.1007/s41918-021-00105-7)
- [24] Yang Z, Ran J, Wu B, et al. Stability challenge in anion exchange membrane for fuel cells. *Curr Opin Chem Eng*. 2016;12:22–30. doi: [10.1016/j.coche.2016.01.009](https://doi.org/10.1016/j.coche.2016.01.009)
- [25] Li Q, Liu L, Miao Q, et al. A novel poly(2,6-dimethyl-1,4-phenylene oxide) with trifunctional ammonium moieties for alkaline anion exchange membranes. *Chem Commun*. 2014;50(21):2791–2793. doi: [10.1039/c3cc47897a](https://doi.org/10.1039/c3cc47897a)
- [26] Hibbs MR. Alkaline stability of poly(phenylene)-based anion exchange membranes with various cations. *J Polym Sci B Polym Phys*. 2013;51(24):1736–1742. doi: [10.1002/polb.23149](https://doi.org/10.1002/polb.23149)
- [27] Zhu L, Pan J, Wang Y, et al. Multication side chain anion exchange membranes. *Macromolecules*. 2016;49(3):815–824. doi: [10.1021/acs.macromol.5b02671](https://doi.org/10.1021/acs.macromol.5b02671)
- [28] Tanaka M, Fukasawa K, Nishino E, et al. Anion conductive block poly(arylene ether)s: synthesis, properties, and application in alkaline fuel cells. *J Am Chem Soc*. 2011;133(27):10646–10654. doi: [10.1021/ja204166e](https://doi.org/10.1021/ja204166e)
- [29] Li N, Leng Y, Hickner MA, et al. Highly stable, anion conductive, comb-shaped copolymers for alkaline fuel cells. *J Am Chem Soc*. 2013;135(27):10124–10133. doi: [10.1021/ja403671u](https://doi.org/10.1021/ja403671u)
- [30] Lee WH, Park EJ, Han J, Shin DW, Kim YS, Bae C. Poly(terphenylene) anion exchange membranes: the effect of backbone structure on morphology and membrane property. *ACS Macro Lett*. 2017;6(5):566–570. doi: [10.1021/acsmacrolett.7b00148](https://doi.org/10.1021/acsmacrolett.7b00148)
- [31] Rao AHN, Nam S, Kim T-H. Comb-shaped alkyl imidazolium-functionalized poly(arylene ether sulfone)s as high performance anion-exchange membranes. *J Mater Chem A*. 2015;3(16):8571–8580. doi: [10.1039/c5ta01123j](https://doi.org/10.1039/c5ta01123j)
- [32] Ramakrishna S, Zhang T-Y, Lu W-C, et al. Materials informatics. *J Intell Manuf*. 2018;30(6):2307–2326. doi: [10.1007/s10845-018-1392-0](https://doi.org/10.1007/s10845-018-1392-0)
- [33] Takahashi K, Tanaka Y. Materials informatics: a journey towards material design and synthesis. *Dalton Trans*. 2016;45(26):10497–10499. doi: [10.1039/C6DT01501H](https://doi.org/10.1039/C6DT01501H)
- [34] Senderowitz H, Tropsha A. Materials informatics. *J Chem Inf Model*. 2018;58(12):2377–2379. doi: [10.1021/acs.jcim.8b00927](https://doi.org/10.1021/acs.jcim.8b00927)

- [35] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science. *APL Mater.* 2016;4(5):053208. doi: [10.1063/1.4946894](https://doi.org/10.1063/1.4946894)
- [36] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett.* 2011;2(17):2241–2251. doi: [10.1021/jz200866s](https://doi.org/10.1021/jz200866s)
- [37] Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* 2013;1(1):011002. doi: [10.1063/1.4812323](https://doi.org/10.1063/1.4812323)
- [38] Audus DJ, de Pablo JJ. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* 2017;6(10):1078–1082. doi: [10.1021/acsmacrolett.7b00228](https://doi.org/10.1021/acsmacrolett.7b00228)
- [39] Afzal MAF, Haghghatdari M, Ganesh SP, et al. Accelerated discovery of high-refractive-index polyimides via first-principles molecular modeling, virtual high-throughput screening, and data Mining. *J Phys Chem C.* 2019;123(23):14610–14618. doi: [10.1021/acs.jpcc.9b01147](https://doi.org/10.1021/acs.jpcc.9b01147)
- [40] Wu Y, Guo J, Sun R, et al. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *NPJ Comput Mater.* 2020;6(1):120. doi: [10.1038/s41524-020-00388-2](https://doi.org/10.1038/s41524-020-00388-2)
- [41] Chen L, Pilia G, Batra R, Huan TD, Kim C, Kuenneth C, Ramprasad R. Polymer informatics: current status and critical next steps. *Mater Sci Eng R Rep.* 2021;144:100595. doi: [10.1016/j.mser.2020.100595](https://doi.org/10.1016/j.mser.2020.100595)
- [42] Kuenneth C, Schertzer W, Ramprasad R. Copolymer informatics with multitask deep neural networks. *Macromolecules.* 2021;54(13):5957–5961. doi: [10.1021/acs.macromol.1c00728](https://doi.org/10.1021/acs.macromol.1c00728)
- [43] Doan Tran H, Kim C, Chen L, et al. Machine-learning predictions of polymer properties with polymer genome. *J Appl Phys.* 2020;128(17):171104. doi: [10.1063/5.0023759](https://doi.org/10.1063/5.0023759)
- [44] Otsuka S, Kuwajima I, Hosoya J, et al. PoLyInfo: polymer database for polymeric materials design. In: Khafa F, Barolli L, Bessis N, editors. Proceedings of the 2011 International Conference on Emerging Intelligent Data and Web Technologies; 2011 Sep 7–9; Tirana, Albania. Conference Publishing Services: Institute of Electrical and Electronics Engineers; 2011. p. 22–29. doi: [10.1109/eidwt.2011.13](https://doi.org/10.1109/eidwt.2011.13)
- [45] Zou X, Pan J, Sun Z, et al. Machine learning analysis and prediction models of alkaline anion exchange membranes for fuel cells. *Energy Environ Sci.* 2021;14(7):3965–3975. doi: [10.1039/d1ee01170g](https://doi.org/10.1039/d1ee01170g)
- [46] Zhai F-H, Zhan Q-Q, Yang Y-F, et al. A deep learning protocol for analyzing and predicting ionic conductivity of anion exchange membranes. *J Membr Sci.* 2022;642:119983. doi: [10.1016/j.memsci.2021.119983](https://doi.org/10.1016/j.memsci.2021.119983)
- [47] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206–215. doi: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x)
- [48] Lundberg SM, Lee S-I A unified approach to interpreting model predictions. In: Luxburg U, Guyon I, Bengio S, et al., editors. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA. Red Hook (NY): Curran Associates Inc.; 2017. p. 4768–4777. doi: [10.5555/3295222.3295230](https://doi.org/10.5555/3295222.3295230)
- [49] Ribeiro MT, Singh S, Guestrin C “Why should I trust you?”: explaining the predictions of any classifier. In: Krishnapuram B, Shah M, Smola A, et al., editors. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA. New York (NY): Association for Computing Machinery; 2016. p. 1135–1144. doi: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778)
- [50] Jablonka KM, Jothiappan GM, Wang S, et al. Bias free multiobjective active learning for materials design and discovery. *Nat Commun.* 2021;12(1):2312. doi: [10.1038/s41467-021-22437-0](https://doi.org/10.1038/s41467-021-22437-0)
- [51] Honrao SJ, Yang X, Radhakrishnan B, et al. Discovery of novel Li SSE and anode coatings using interpretable machine learning and high-throughput multi-property screening. *Sci Rep.* 2021;11(1):16484. doi: [10.1038/s41598-021-94275-5](https://doi.org/10.1038/s41598-021-94275-5)
- [52] Hatakeyama-Sato K, Umeki M, Adachi H, et al. Exploration of organic superionic glassy conductors by process and materials informatics with lossless graph database. *NPJ Comput Mater.* 2022;8(1):170. doi: [10.1038/s41524-022-00853-0](https://doi.org/10.1038/s41524-022-00853-0)
- [53] Anker AS, Kjær ETS, Juulsholt M, et al. Extracting structural motifs from pair distribution function data of nanostructures using explainable machine learning. *NPJ Comput Mater.* 2022;8(1):213. doi: [10.1038/s41524-022-00896-3](https://doi.org/10.1038/s41524-022-00896-3)
- [54] Mastelini SM, Cassar DR, Alcobaça E, et al. Machine learning unveils composition-property relationships in chalcogenide glasses. *Acta Materialia.* 2022;240:118302. doi: [10.1016/j.actamat.2022.118302](https://doi.org/10.1016/j.actamat.2022.118302)
- [55] Pan ZF, An L, Zhao TS, et al. Advances and challenges in alkaline anion exchange membrane fuel cells. *Prog Energy Combust Sci.* 2018;66:141–175. doi: [10.1016/j.pecs.2018.01.001](https://doi.org/10.1016/j.pecs.2018.01.001)
- [56] Park EJ, Kim YS. Quaternized aryl ether-free polyaromatics for alkaline membrane fuel cells: synthesis, properties, and performance – a topical review. *J Mater Chem A.* 2018;6(32):15456–15477. doi: [10.1039/c8ta05428b](https://doi.org/10.1039/c8ta05428b)
- [57] Rohatgi A. WebPlotDigitizer [internet]. Pacifica (CA); 2021 [cited 2022 Nov 22]. Available from: <https://automeris.io/WebPlotDigitizer>
- [58] Moriwaki H, Tian YS, Kawashita N, et al. Mordred: a molecular descriptor calculator. *J Cheminform.* 2018;10(1):4. doi: [10.1186/s13321-018-0258-y](https://doi.org/10.1186/s13321-018-0258-y)
- [59] Landrum G, Tosco P, Kelley B, et al. Rdkit: open-source cheminformatics [internet]; 2021 [cited 2022 Nov 22]. Available from: <https://www.rdkit.org/>
- [60] Ali M. PyCaret: an open source, low-code machine learning library in python [Internet]; 2021 [cited 2022 Nov 22]. Available from: <https://www.pycaret.org>
- [61] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, et al., editors. Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, Canada. Red Hook (NY): Curran Associates Inc.; 2018. p. 6639–6649. doi: [10.5555/3327757.3327770](https://doi.org/10.5555/3327757.3327770)
- [62] Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Krishnapuram B, Shah M, Smola A, et al., editors. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA.

- New York (NY): Association for Computing Machinery; 2016. p. 785–794. doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)
- [63] Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. doi: [10.1023/a:1010933404324](https://doi.org/10.1023/a:1010933404324)
- [64] Molnar C. Interpretable machine learning. British Columbia (Canada): Leanpub; 2020. Model-Agnostic Methods, Feature Importance; p. 190–199.
- [65] Dobson CM. Chemical space and biology. *Nature.* 2004;432(7019):824–828. doi: [10.1038/nature03192](https://doi.org/10.1038/nature03192)
- [66] Gopi KH, Peera SG, Bhat SD, et al. Preparation and characterization of quaternary ammonium functionalized poly(2,6-dimethyl-1,4-phenylene oxide) as anion exchange membrane for alkaline polymer electrolyte fuel cells. *Int J Hydrogen Energy.* 2014;39(6):2659–2668. doi: [10.1016/j.ijhydene.2013.12.009](https://doi.org/10.1016/j.ijhydene.2013.12.009)
- [67] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* 2011;12:2825–2830. doi: [10.5555/1953048.2078195](https://doi.org/10.5555/1953048.2078195)
- [68] Drucker H, Burges CJC, Kaufman L, et al. Support vector regression machines. In: Jordan M, Petsche T, editors. Proceedings of the 9th International Conference on Neural Information Processing Systems; 1996 Dec 3–5; Denver, CO. Cambridge (MA): MIT Press; 1996. p. 155–161. doi: [10.5555/2998981.2999003](https://doi.org/10.5555/2998981.2999003)
- [69] Williams CKI, Rasmussen CE Gaussian processes for regression. In: Touretzky D, Mozer M, Hasselmo M, editors. Proceedings of the 8th International Conference on Neural Information Processing Systems; 1995 Nov 27–Dec 2; Denver, CO. Cambridge (MA): MIT Press; 1995. p. 514–520. doi: [10.5555/2998828.2998901](https://doi.org/10.5555/2998828.2998901)
- [70] Rosenblatt F. The perceptron, a perceiving and recognizing automaton project para. NY (USA): Cornell Aeronautical Laboratory; 1957.
- [71] Liu B, Wei Y, Zhang Y, et al. Deep neural networks for high dimension, low sample size data. In: Sierra C, editor. Proceedings of the 26th International Joint Conference on Artificial Intelligence; 2017 Aug 19–25; Melbourne, Australia. Pennsylvania Ave, NW (DC): AAAI Press; 2017. p. 2287–2293. doi: [10.5555/3172077.3172206](https://doi.org/10.5555/3172077.3172206)
- [72] Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res.* 2010;11:2079–2107. doi: [10.5555/1756006.1859921](https://doi.org/10.5555/1756006.1859921)
- [73] Randic M. On molecular identification numbers. *J Chem Inf Comput Sci.* 1984;24(3):164–175. doi: [10.1021/ci00043a009](https://doi.org/10.1021/ci00043a009)
- [74] Hall LH, Kier LB. Electrotopological state indexes for atom types - a novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci.* 1995;35(6):1039–1045. doi: [10.1021/ci00028a014](https://doi.org/10.1021/ci00028a014)
- [75] Lai AN, Zhou K, Zhuo YZ, et al. Anion exchange membranes based on carbazole-containing polyolefin for direct methanol fuel cells. *J Membr Sci.* 2016;497:99–107. doi: [10.1016/j.memsci.2015.08.069](https://doi.org/10.1016/j.memsci.2015.08.069)
- [76] Hollas B. An analysis of the autocorrelation descriptor for molecules. *J Math Chem.* 2003;33(2):91–101. doi: [10.1023/a:1023247831238](https://doi.org/10.1023/a:1023247831238)
- [77] Comesana AE, Huntington TT, Scown CD, et al. A systematic method for selecting molecular descriptors as features when training models for predicting physicochemical properties. *Fuel.* 2022;321:123836. doi: [10.1016/j.fuel.2022.123836](https://doi.org/10.1016/j.fuel.2022.123836)
- [78] Yang X, Zhang J, Yoshizoe K, et al. ChemTS: an efficient python library for de novo molecular generation. *Sci Technol Adv Mater.* 2017;18(1):972–976. doi: [10.1080/14686996.2017.1401424](https://doi.org/10.1080/14686996.2017.1401424)