Journal of Materials Chemistry A

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. A, 2019, 7, 11847

Rational design of hydrocarbon-based sulfonated copolymers for proton exchange membranes†

Lunyang Liu,[ab] Wenduo Chen,[ac] Tingli Liu,[ac] Xiangxin Kong,[ac] Jifu Zheng[a] and Yunqi Li*[ac]

Received 19th January 2019
Accepted 11th April 2019

DOI: 10.1039/c9ta00688e

rsc.li/materials-a

Developing novel hydrocarbon-based proton exchange membranes is at the Frontier of research on fuel cells, batteries and electrolysis, aiming to reach the demand for advanced performance in proton conductivity, fuel retardation, swelling, mechanical and thermal stability *etc.* Sulfonated copolymers are a class of highly focused materials used to fabricate proton exchange membranes. Sampling the space of chemical structure and architecture, as well as the composition of sulfonated copolymers, generated more than 2800 original reports (till June 2018), which highlight the feasibility and necessity to screen novel materials with computer-aided design strategies. Through an investigation of the 166 hydrophilic and 175 hydrophobic monomers that were used to build hydrocarbon-based sulfonated copolymers, correlation relationships among performance indexes and intrinsic properties of copolymers were explored. Reliable predictive models, in both regression and classification manners, for proton conductivity, methanol permeability, tensile modulus and degradation temperature were constructed and validated. Based on the top ranked monomers with superior performance and their optimal fractions and combinations, novel copolymers that have better predicted performance at a 0.05 significance level were presented. These predicted formulas provide directions for the synthesis of novel hydrocarbon-based sulfonated copolymers for advanced proton exchange membranes.

Introduction

Fuel cells are one of the most potent clean energy techniques owing to their attractive high energy conversion efficiency and zero emission. Polymer electrolyte fuel cells (PEFCs) are at the Frontier of research, and the key components, proton exchange membranes (PEMs) and the associated catalyst layers, are the bottleneck for their large-scale applications at the current stage. The membranes should meet strict requirements, mainly scaled by five classes of performance indexes: (i) high proton conductivity, (ii) low fuel permeability, (iii) good thermal and hydrolytic stability, (iv) outstanding mechanical properties in both dry and hydrated states, and (v) sufficient water uptake and moderate swelling.[1] Various polymer materials have been explored, abundant data have been accumulated and the necessity to provide data supported knowledge and wisdom for further advancement of PEMs has become apparent.

For commercial usage, the state-of-the-art PEM materials are perfluorinated sulfonic acid (PFSA) polymers, such as Nafion®,[2] Flemion®[3] *etc.* They readily satisfy the five performance indexes, but also suffer from a limited operation temperature window (0–80 °C), high cost and high methanol permeability. Alternative candidates, mainly hydrocarbon-based sulfonated copolymers, including acid-functionalized aromatic hydrocarbon-based polymers,[4] poly(arylene ether nitrile),[5] poly(phenylene ether ketone),[6] poly(ether sulfone),[7,8] poly(arylene ether),[9] poly(ether ketone),[10] poly(arylene ether ketone),[11–14] polyimides,[15–17] polyphenylene oxide[18] *etc.*, have been reported with high expectations. These novel polymers were extensively explored in a trial-and-error manner to some degree. According to a recent comprehensive review,[4] hydrocarbon-based sulfonated copolymers designed to be competitive with PFSA polymers should maintain fine microphase separated structures to achieve a sufficiently high proton conductivity. The assembled structures of copolymers have a continuous matrix made from hydrophobic monomers, and percolated ionic channels stabilized by hydrophilic monomers in a relative humidity window to allow proton conduction. The candidates for both hydrophobic and hydrophilic monomers span a broad space of chemical structure and architecture, and the compositions for a given combination of monomers also create a huge number of formulas. Inspired by the widely distributed block copolymers and their phase-separated structures,[19–22] it is feasible and

[a]Key Laboratory of High-Performance Synthetic Rubber and Its Composite Materials, Key Laboratory of Polymer Ecomaterials, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, PR China. E-mail: yunqi@ciac.ac.cn; Tel: +86 431 85262535

[b]University of Chinese Academy of Sciences, Beijing 100049, PR China

[c]University of Science and Technology of China, Hefei 230026, China

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ta00688e

highly demanded to design novel polymers that can be used to fabricate advanced PEMs using computer-aided rational design strategies and material genome approaches.

There have been more than 2800 academic reports (till June 2018) published on the exploration of various hydrocarbon-based sulfonated copolymers. For a given copolymer, considering the complexity to compute a given performance index such as proton conduction, mechanical modulus, thermal stability, fuel permeability *etc.*, either in a spatial or temporal scale, conventional theoretical and simulation approaches become inadequate. The highly tuneable nature of copolymers for PEMs requires a multi-scale consideration of composition, the assembly of atoms, groups, molecular segments, and molecules, and their changes under various processing and operation conditions. Current discovery of novel copolymers is largely guided by chemical intuition and serendipity, which makes providing solid knowledge on the boundary and optimal solutions difficult. QSPR (Quantitative Structure–Property Relationship)-based approaches have gradually established success in tackling such problems in the development of purpose-driven materials through rational design.[23–27] Takahashi *et al.*[28] used Random Forest (RF) classification to pick 11 out of 15 000 perovskite materials with a predicted ideal band gap that would facilitate the preparation of superior solar cells. Giorgos *et al.*[29] employed a machine learning method for large-scale screening of metal–organic frameworks (MOFs) for gas storage. We recently constructed[30] implicit and explicit predictive models for the proton conductivities of Nafion®-based PEMs using RF and multiple linear regression (MLR), and the most potent fillers and processing treatments were found. However, using data-driven rational design approaches to explore hydrocarbon-based sulfonated copolymers for PEMs, to the best of our knowledge, has not been reported yet.

We performed data mining based on reports closely related to the exploration of hydrocarbon-based sulfonated copolymers to fabricate PEMs. Through the introduction of a unified L_{sig} value, data and feature engineering, predictive model evaluation and optimal solution identification with a unified setting parameter can be readily regulated. Then, predictive models for performance indexes including proton conductivity (σ), the degradation temperature at 5% weight loss (T_{d5}), methanol permeability (MePerm) and tensile modulus (TM) were constructed and their robustness was validated. Further, the hydrophobic and hydrophilic monomers used to build copolymers were clustered and the optimal compositions were screened out. Potent formulas for the copolymers with better overall performance at a significant level were presented. This work provides realizable strategies to guide the synthesis of hydrocarbon-based sulfonated copolymers in the fabrication of advanced PEMs.

Methods

Datasets

More than 2800 original reports were retrieved from the Web of Science with the keywords of "sulfonate proton exchange membrane & copolymer" and "sulfonate polymer electrolyte membrane & copolymer" till June 2018. In total, 156 of them were kept after excluding those reports either with insufficient and incomplete information as PEM materials, or with other various purposes where the correlation between the intrinsic features of the copolymers and the performance indexes of PEMs was not in focus. 3518 records were accumulated, with all or some of the nine performance indexes, *i.e.* σ, MePerm, water uptake (WU), swelling ratio (SR), glass transition temperature (T_g), T_{d5}, tensile yielding strength (TS), TM and the elongation at break (EAB). The probability density functions (PDFs) for four of these indexes are presented in Fig. S1,† and the values at the preferred side with L_{sig} are labelled. There are, in total, 198 copolymers (166 hydrophilic monomers with at least one sulfonate acid group, and 175 hydrophobic monomers, the full list of these monomers can be found in Tables S1 and S2 in the ESI†) in a broad window of temperatures (in °C) and relative humidity (in %). Then, the chemical structure and architecture of each monomer were encoded as SMILES. The number of records and the corresponding simple statistical values are presented in Table S3.†

Features

Descriptive features for the composition, structure, processing and operational conditions from the reports were accumulated and organized using the protocol introduced in our recent works.[30,31] Four features were tabulated from the original reports; they are the fraction of hydrophilic monomer in the copolymer (FraLic), ion exchange capacity (IEC, meq. g^{-1}), temperature (T, °C), and relative humidity (RH, %). Features for the chemical structure and architecture of the hydrophobic and hydrophilic monomers were computed using RDKit.[32] They include 1D constitutional molecular properties (such as the count of atoms, groups and bonds), 2D topological descriptors (such as the Balaban' J index), the 2D connectivity index (*e.g.* Chi indices) and 2D MOE-type descriptors[33] (such as VSA, EState descriptors). In total, 196 features for each copolymer from the weighted (by FraLic) addition of hydrophilic and hydrophobic monomers were calculated.

Engineering of the large number of features to remove redundancy and information-insufficient noise is necessary to construct robust predictive models. The variance of a feature across the whole dataset is defined as

$$\mathrm{VAR}(x)^2 = \frac{\sum_{i=1}^{N}(x_i - \langle x \rangle)^2}{N} \quad (1)$$

Here, N is the number of records for a feature x, and $\langle x \rangle$ is the mean value. A feature was excluded if VAR was no more than L_{sig}. Then, the pairwise correlation coefficients of two features and performance indexes were calculated according to

$$\mathrm{CORR}(x,y)_p = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}\sqrt{N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2}} \quad (2)$$

$$\mathrm{CORR}(x,y)_{\mathrm{sp}} = \frac{\mathrm{cov}(\mathrm{rg}_x - \mathrm{rg}_y)}{\mathrm{VAR}_{\mathrm{rg}_x}\mathrm{VAR}_{\mathrm{rg}_y}} \quad (3)$$

If the CORR between any two features was larger than $1 - L_{\mathrm{sig}}$, then the feature with less VAR was excluded to remove redundancy. The subscripts p and sp represent Pearson and Spearman correlations, respectively. $\mathrm{VAR}_{\mathrm{rg}_x}$ and $\mathrm{VAR}_{\mathrm{rg}_y}$ are the standard deviations of the rank variables, and $\mathrm{cov}(\mathrm{rg}_x - \mathrm{rg}_y)$ is the covariance of the rank variables. In this work, L_{sig} is fixed at 0.05.

For the dataset used in this work, 88 and 57 features were excluded due to noise ($\mathrm{VAR} \leq L_{\mathrm{sig}}$) and redundancy ($\mathrm{CORR} > 1 - L_{\mathrm{sig}}$), and finally, 51 features remained. Further engineering to construct robust predictive models with as few features as possible was also performed using the random forest (RF) algorithm[34] following our recent reported strategies,[30] and strengthened using the genetic algorithm (GA).[35] The information of the names, abbreviations, groups of molecules and nomenclature for the 55 features is presented in Table S4 in the ESI.†

It is worth noting that using the same procedure, the correlation matrix for performance indexes expressed by Pearson ($\mathrm{CORR_p}$) and Spearman ($\mathrm{CORR_{sp}}$) was also calculated. Alternative to $\mathrm{CORR_{sp}}$, the maximal information coefficient (MIC)[36] was also computed to present the correlation in pairs of performance indexes.

Predictive models and evaluation

Before the construction of predictive models, the combination of the remaining 55 features was optimized by GA enhanced regression RF analysis. For a given dataset and a combination of features, the root mean square error (RMSE) between the predicted and experimental performance indexes was computed. Then, the importance of each feature in predicting a given performance index can be determined by the $\mathrm{CORR_p}$, %IncMSE and $\mathrm{CORR_{sp}}$ in linear, nonlinear and monotonous correlation relationships. Here, we used 2000 decision trees in each RF computation. The optimal combination of features with the least RMSE and the highest stability against data permutation was selected to construct the final predictive models.

To evaluate the performance of the regression predictive models, the root mean squared error (RMSE) and coefficient of determination, *e.g.*, the square of R_{p}, were calculated through

$$R_{\mathrm{p}}{}^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i'\right)^2}{\sum_{i=1}^{N}\left(y_i - \langle y \rangle\right)^2} \quad (4)$$

Here, y_i' is the predicted value for the ith record. $R_{\mathrm{p}}{}^2$ is in the range of 0 and 1, where a larger value means a better prediction.

We also built classification models with the logistic regression (LR) algorithm[37] to classify copolymers. To evaluate the performance of the classification predictive models, threshold values to split binary classes were set as $(1 \pm L_{\mathrm{sig}}) \times \langle y \rangle$ (as listed in Table S3†). Either plus or minus was used depending on the preference of whether a performance index is at the larger or smaller side. The area under the receiver operating characteristic (ROC) curve (AUC) was used to evaluate the goodness-of-fit for the predictive models. The ROC curve represents a graphical representation of the discriminatory power of a binary classification system and it is created by plotting the true positive rate (TPR) against the false positive rate (FPR). AUC values range from 0 to 1, where 1 represents a perfect model and less than 0.5 indicates an ideally random model.

In the evaluation of the robustness of the predictive models, 5-fold cross-validation was carried out using home-made scripts implanted in R-project.

Monomer clustering and copolymer screening

A sequence of elementary units in SMILES was used to record the chemical structure and architecture of each monomer. Then, the matrix for monomer clustering was encoded by Morgan fingerprints,[38] which composed a series of binary digits (bits) for the presence or absence of particular substructures in a monomer. An aligned region in a pair of sequences was used to determine the similarity between two monomers.

The distribution of the chemical structures and architectures of the 166 hydrophilic and 175 hydrophobic monomers was presented using a hierarchical clustering dendrogram. The optimal number of clusters at each hierarchical level was determined using the Elbow method, which minimizes the total intra-cluster variation or the total within-cluster sum of the square[39] through

$$\min\left(\sum_{i=1}^{k} W(C_{\mathrm{k}})\right) \quad (5)$$

where C_{k} is the kth cluster and $W(C_{\mathrm{k}})$ is the within-cluster variation.

Based on the clustering of monomers and their contribution to each performance index, the screening of copolymer compositions was carried out. The fractions of hydrophilic monomers were set from 0.1 to 0.9 with a 0.1 step to get 9 grades. This generates a screening library consisting of 262 450 copolymers, and only 198 of them have been reported. Combinations of hydrophilic and hydrophobic monomers that have predicted performance values with L_{sig} at least at the preferred side of the distribution of performance for the 198 reported copolymers were selected for further analysis.

Results and discussion

Predictive models

Prior to the construction of the predictive models, a correlation matrix containing the performance indexes using $\mathrm{CORR_{sp}}$ and MIC was computed and is presented in Fig. 1. The key performance indexes for PEMs, σ and MePerm, are positively correlated; as selective membranes, they are expected to interact in a trade-off manner.[40] This is reasonable for sulfonated PEMs,

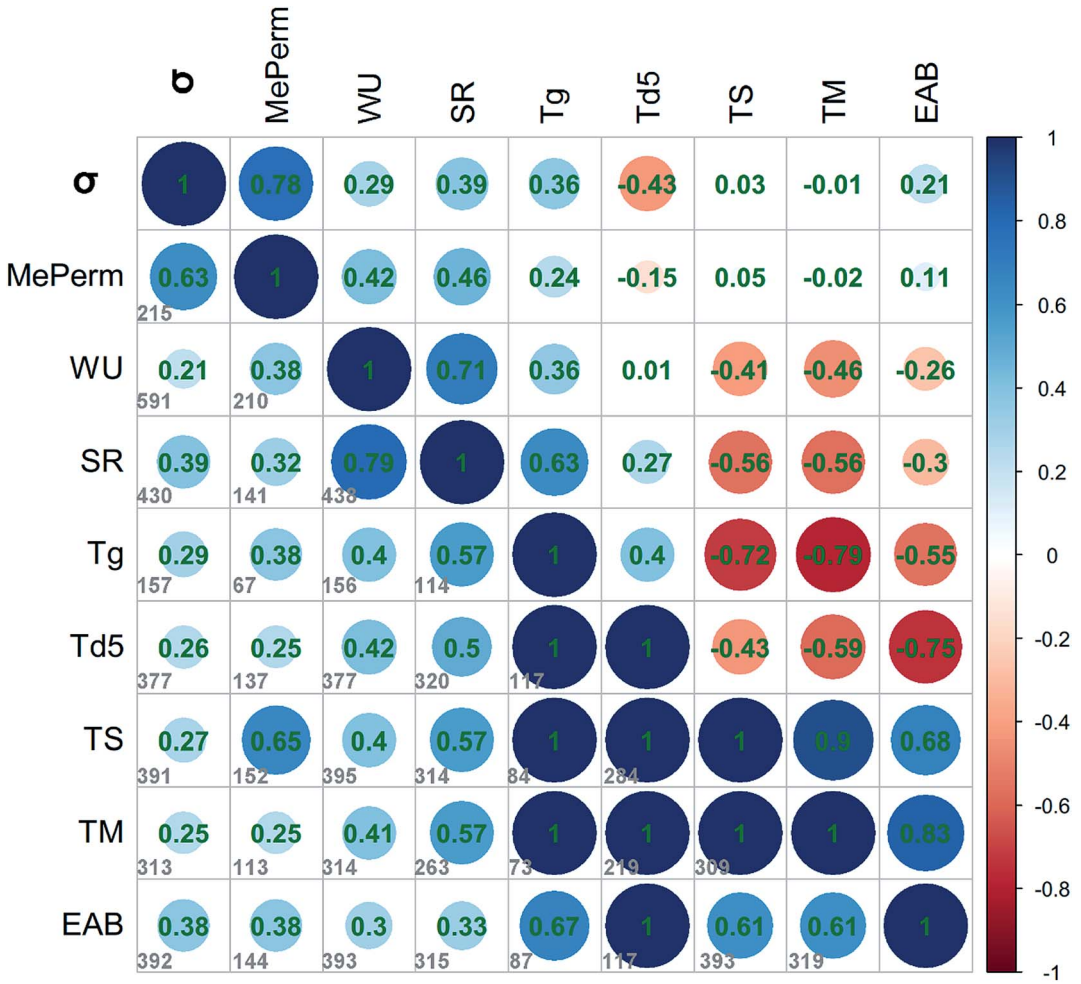


Fig. 1 Correlation matrix of PEM performance indexes. The lower and upper triangles are MIC and $CORR_p$, respectively. The grey numbers are sample capacities for the pairs of performance indexes. The color bar and the size of the circles indicate the strength and direction of correlation from −1 to 1.

where protonated water clusters (such as in Zundel and Eigen forms) and methanol have comparable steric volume and polarity, and both of them transport across the membrane through the interconnected ionic domains.[41] WU, SR and σ are all positively correlated in linear and nonlinear forms;[31] they are crucial indexes for proton conduction and dimensional stability. Regarding the thermal properties, the MIC between T_g and T_{d5} is 1, while the $CORR_p$ is relatively low. This indicates that they have strong nonlinear correlation. Meanwhile, all mechanical properties are strongly correlated. To construct representative predictive models, analogous to the procedure for feature engineering, we selected 4 out of 9 performance indexes, σ, MePerm, T_{d5} and TM. Their probability distributions can be found in Fig. S1.†

The optimization of the combination of features to construct predictive models for the four performance indexes is shown in Fig. S2.† The RMSE for both training and test sets in the 5-fold cross validation levels off after ~300 generations of permutation in the GA enhanced RF algorithm. Finally, the optimized combination of 20, 29, 18 and 27 features was selected to predict σ, T_{d5}, MePerm and TM, separately. These features are bolded and labeled in Table S4,† and the correlation matrix for these selected features is shown in Fig. S3.† It clearly shows that through the feature engineering using L_{sig} criteria, finally selected features are representative and informative. Further evaluation of the importance of features through the mean decrease in node impurity computed in RF, *i.e.*, %IncMSE, together with $CORR_p$ and $CORR_{sp}$ for the correlation of individual features with performance indexes, is illustrated in Fig. S4.† These three coefficients only show partial consistence, which indicates that the predictive models for the four performance indexes cannot be constructed straightforwardly through simple combination of these features. According to the order of %IncMSE, the top ranked features reasonably correlate with the performance indexes. For example, proton concentration in the membranes, which is tightly coupled with Ion Exchange Capacity (IEC), is the most important feature for proton conduction.[6] It is well known that proton conductivity is commonly positively associated with methanol permeability,[42] thus IEC is also the most importance feature for MePerm. For

T_{d5}, the polarity of exposed Van der Waals surface area (PEOE_VSA) is a well-known determinative factor for the thermal stability of polymers.

Predictive models for these four performance indexes in a regression and classification manner were constructed and are shown in Fig. 2. For all four performance indexes, from the test set in the 5-fold cross validation, the correlation between predicted and experimental values using R_p^2 was found to be in the range of 0.79 to 0.91, and the AUC is in the range of 0.78 to 0.91. Such a high confidence score indicates that these predictive models are quite reliable and can be used to screen hydrocarbon-based copolymers from the huge number of copolymer candidates in combination and the fraction of hydrophilic and hydrophobic monomers.

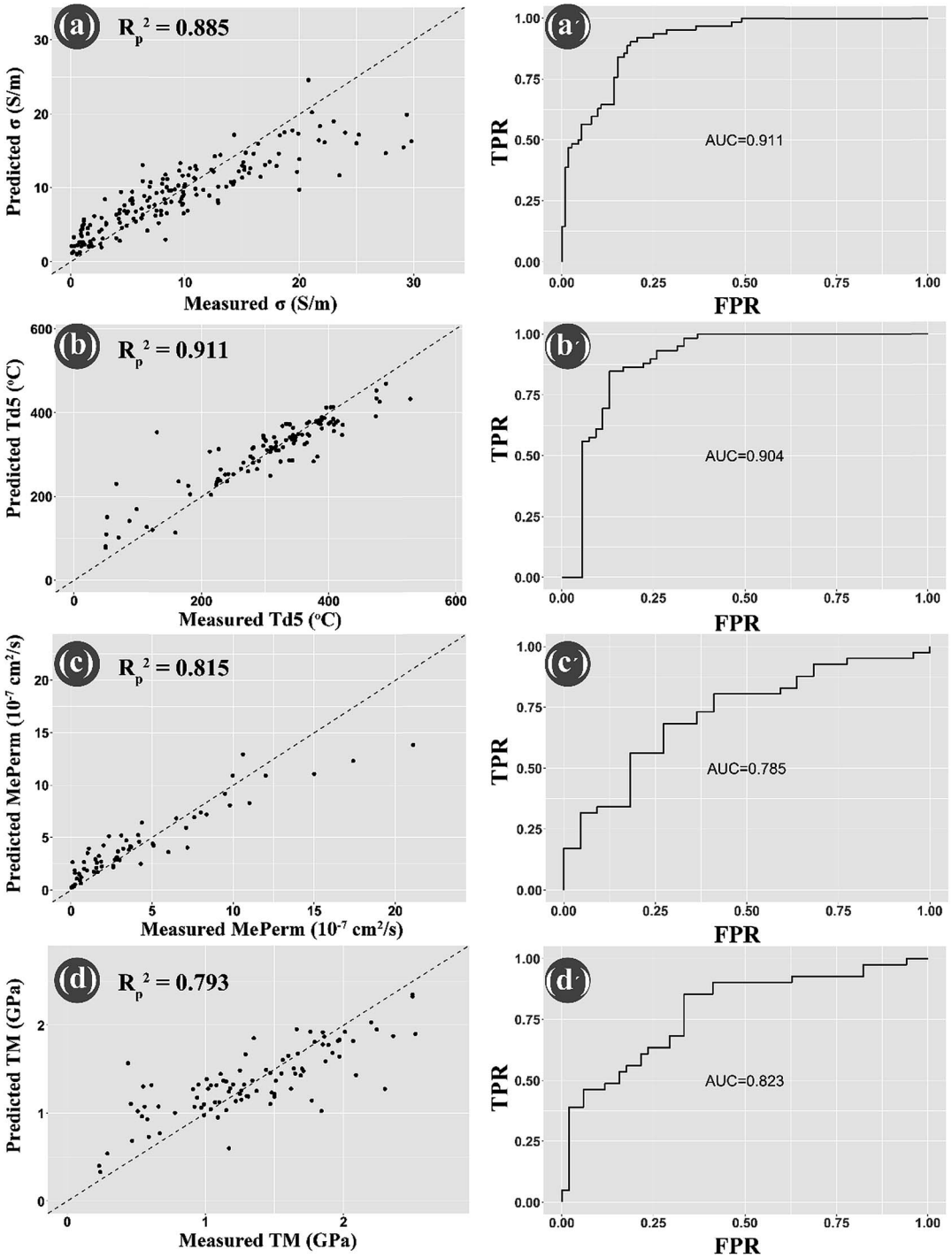


Fig. 2 Performance of predictive models using the test dataset for (a, a′) proton conductivity (σ), (b, b′) the degradation temperature at 5% weight loss (T_{d5}), (c, c′) methanol permeability (MePerm) and (d, d′) tensile modulus (TM) in regression using random forest and classification using logistic regression.

Clustering of monomers

Through the minimization of the total intra-cluster variation, the determination of the optimal number of clusters at the elbow point, all 166 hydrophilic monomers and 175 hydrophobic monomers can be clustered into 4 and 3 clusters, respectively (shown in Fig. 3). In the 4 clusters of hydrophilic monomers, there are 95, 29, 33 and 9 monomers in each cluster. The significance level among the distributions of chemical space spanned by the monomers in each cluster was also computed using the *t*-test. All except the pair of cli2 and cli4 show significant difference in chemical space at L_{sig}. Similarly, there are 94, 12 and 69 monomers in the 3 hydrophobic clusters. Only the pair of clo1 and clo2 slightly overlapped in chemical space. These results indicate that the clustering based on SMILES sequences provides reasonable discrimination for both hydrophilic and hydrophobic monomers.

We further investigated the correlation between chemical space and performance indexes. The distribution of performance indexes associated with the chemical space represented by the clusters of monomers using a violin plot is shown in

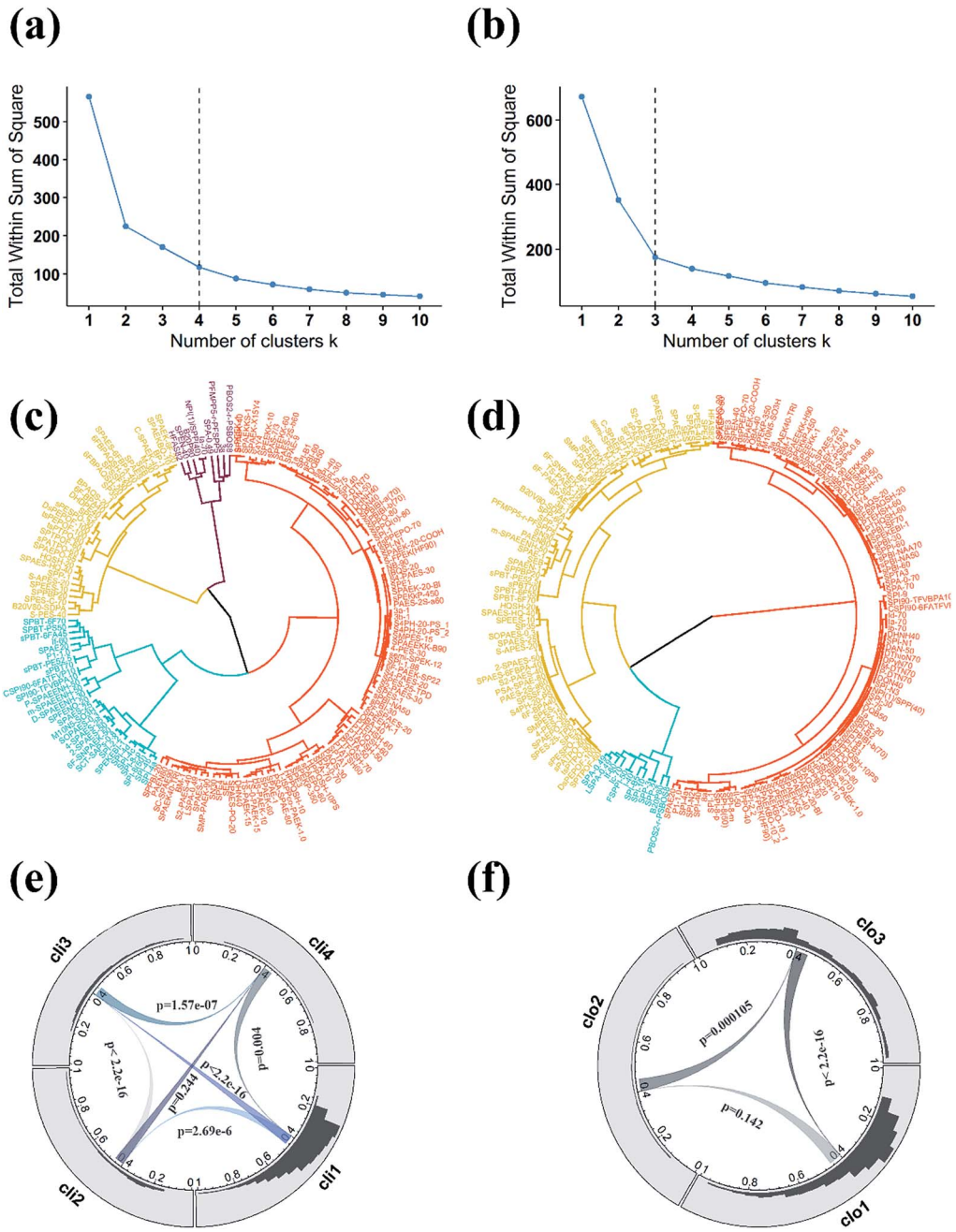


Fig. 3 Determination of the optimal number of clusters at the elbow point (a, b), the distribution of monomers in the dendrogram of clusters (c, d), and the histogram of monomer similarity in each cluster (e, f) for 166 hydrophilic (a, c, e) and 175 hydrophobic (b, d, f) monomers. The *p*-values in (e and f) indicate the significant difference for the ribbon connection between two clusters using the *t*-test.

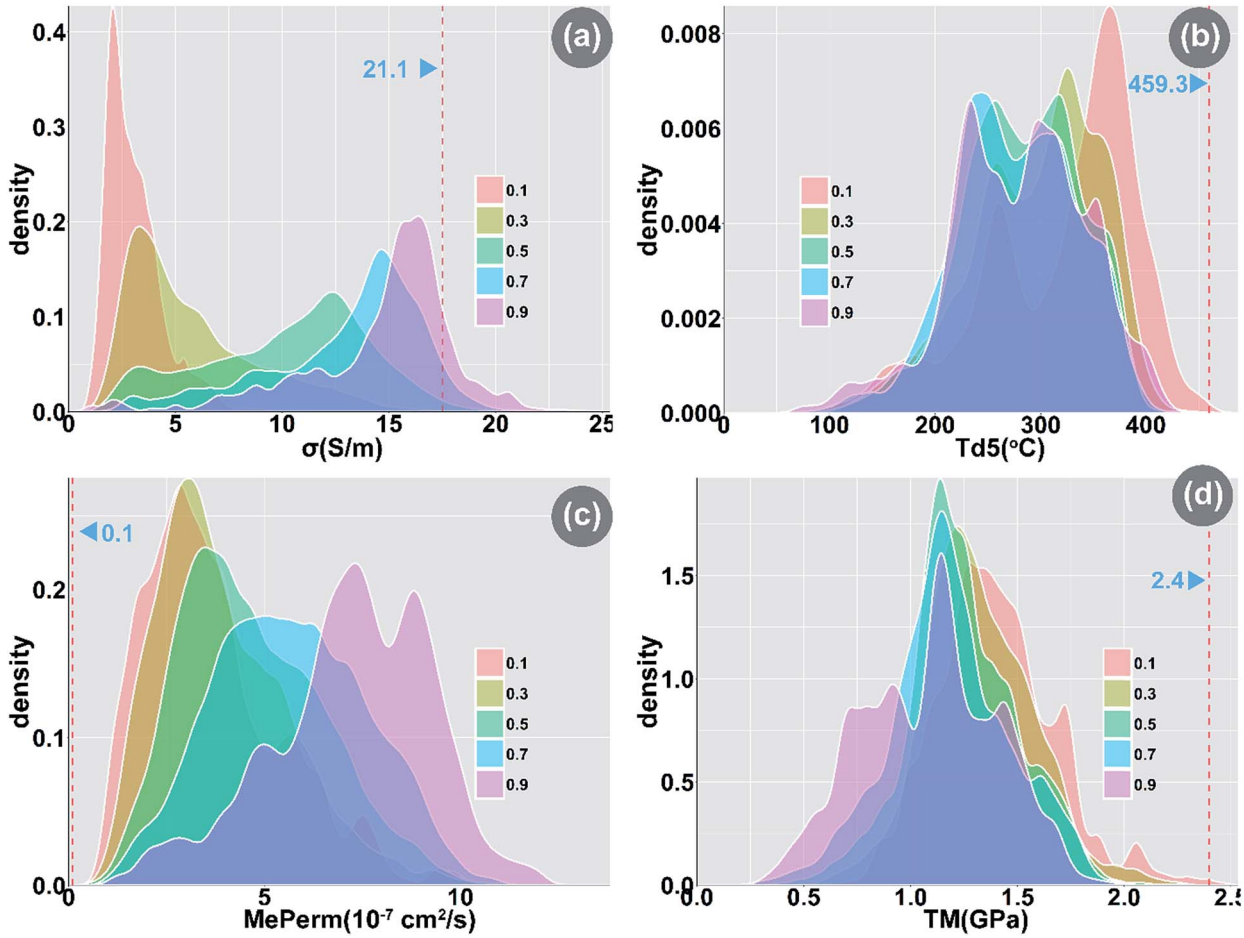


Fig. 4 PDF plots of RF model predictions at different fractions of hydrophilic monomer ratio: (a) proton conductivity (σ), (b) the degradation temperature at 5% weight loss (T_{d5}), (c) methanol permeability (MePerm) and (d) tensile modulus (TM). Vertical dashed lines show the best value from all reports with L_{sig} that refers to the distribution of reported copolymers.

Fig. S5.† ANOVA analysis shows that all except one have significantly different distributions in performance indexes. This suggests that the location of copolymers in the chemical space can strongly affect these four performance indexes. The exception is for hydrophobic clusters and MePerm that is due to the majority of hydrophobic monomers having been designed to retard methanol permeability. For the hydrophilic monomers, cli1 shows the largest median TM value, cli2 has the highest proton conductivity, and cli3 shows the highest median T_{d5}. The hydrophilic monomers in cli4 have a higher portion with a high σ, TM, and low MePerm at the preferred side, but with a very low T_{d5} that may inhibit the simple utilization of these monomers. For the hydrophobic clusters, the monomers in clo2 show a bimodal distribution in σ, but T_{d5} is not high enough, which also prohibits their usage. Hydrophobic monomers play the major role as a supporting matrix, and the large number of monomers in clo1 that exhibited well balanced performances should encourage the development of hydrocarbon-based copolymers for PEMs. This cluster analysis provides clear evidence for how structural modification influences performance indexes. It again validates the feasibility of the rational design of copolymers for targeted performance indexes.

Screening copolymers

Based on the predictive models and clustered structure–performance relationships, we carried out a combinatorial and high-throughput screening for potential copolymers with improved performances to fabricate PEMs. All 166 hydrophilic and 175 hydrophobic monomers, in 9 grades for the fraction of hydrophilic monomers, were used to generate the screening material library. This library consists of 261 252 novel copolymers after excluding the 198 reported ones. The prediction by the regression RF model and classification LR models for the four performance indexes was employed to screen the copolymer library at a fixed temperature and relative humidity of 80 °C and 100%, where most PEMs were measured and showed the best performance for σ. The color–map plot for the predicted values by RF and LR is shown in Fig. S6 and S7,† and the distributions of these predicted values are presented using PDF in Fig. 4. We can see that an increase of the fraction of hydrophilic monomers leads to higher σ and MePerm values, but it is

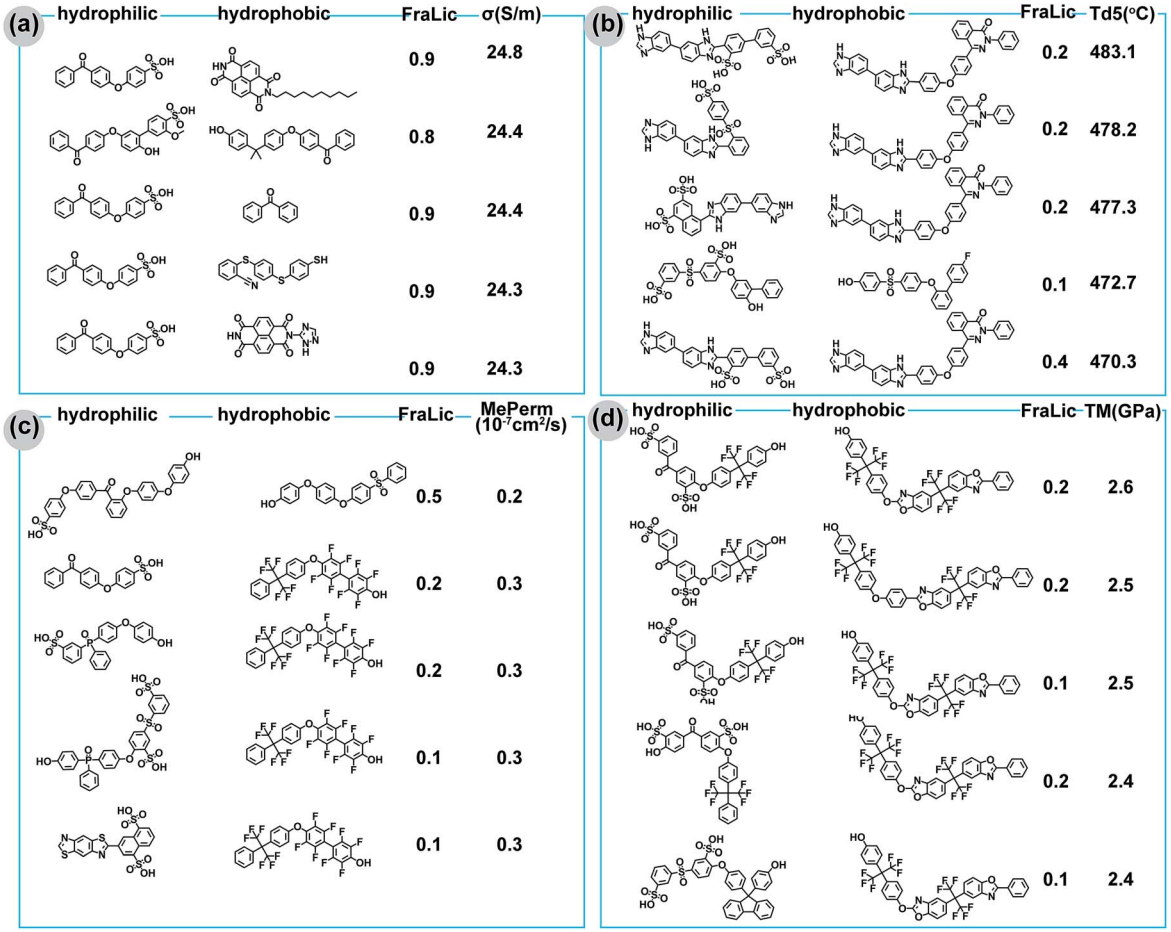


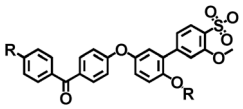
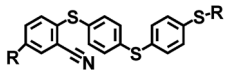
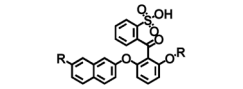
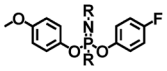
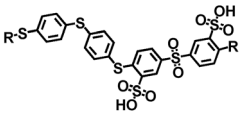
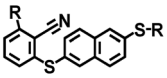
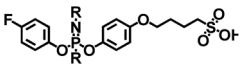
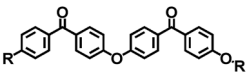
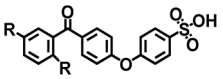
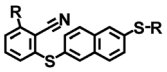
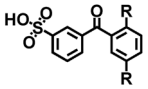
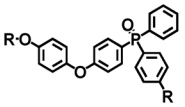
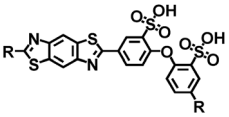
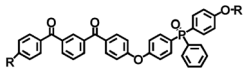
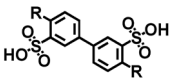
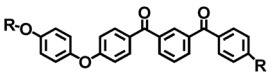
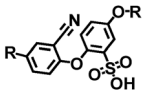
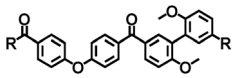
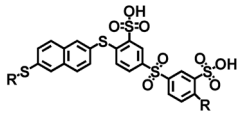
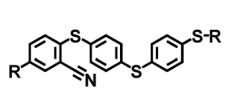
Fig. 5 Top 5 ranked copolymers with the combination of hydrophilic and hydrophobic monomers, and the fraction of hydrophilic monomers for each performance index of (a) proton conductivity (σ), (b) the degradation temperature at 5% weight loss (T_{d5}), (c) methanol permeability (MePerm) and (d) tensile modulus (TM).

deleterious to T_{d5} and TM. The evolution of PDF against FraLic shows different overall distributions, which suggests that the explored chemical space is quite limited. Otherwise, the overall PDF profile should be similar to a normal distribution.

In reference to the distribution of 198 reported copolymers at L_{sig} (values shown in Table S3†), 319, 32, 156 and 48 novel copolymers show superior performance in σ, T_{d5}, MePerm and TM, respectively. We then sorted these copolymers by using a single performance index, and the top 5 ranked copolymers are presented in Fig. 5. The σ of the top 5 ranked copolymers is in the range of 24.82–24.32 S m^{-1}, which is better than the value of 17.5 S m^{-1} for Nafion 117.[43] The σ values of these copolymers are also comparable to that of sulfonated poly(arylene-*co*-naphthalimide) based membranes, which were reported by Zhang *et al.*,[44] where a maximum of 30.2 S m^{-1} was achieved at 100 °C. But, as expected, a higher content of sulfonate facilitates both proton conduction and methanol permeability; these top ranked copolymers have their fraction of hydrophilic monomers ranging from 0.8 to 0.9, which may not be practicable. Accordingly, the top ranked copolymers with low MePerm occur in copolymers with a low FraLic. Meanwhile, a low FraLic is also favored for copolymers with superior performance in T_{d5} and TM. This is reasonable in the design of copolymers for PEMs, where the hydrophobic part is mainly in charge of the thermal and mechanic stability.

In the majority of reports, alternative polymers were developed for usage in PEMs and compared with the "golden standard" material, namely Nafion. We collected peak values of Nafion 117 for the four performance indexes, σ: 17.5 S m^{-1},[43] T_{d5}: 266.2 °C,[45] MePerm: 11.5 × 10^{-7} cm^2 s^{-1} (ref. 46) and TM: 0.3 GPa.[47] Then, all copolymers in the library that have overall better predicted performance could be found. There are, in total, 2838 novel copolymers. These polymers are composed by 59 (out of 166) hydrophilic and 154 (out of 175) hydrophobic monomers. The FraLic locates at 0.79 ± 0.14 in the range from 0.2 to 0.9. Maximization of the selectivity (defined as σ/MePerm at 100% RH and 80 °C) may provide the best candidates for PEMs. Table 1 shows that the top ranked copolymers with unique hydrophilic monomers fall into the ideal range for PEM applications. The full list of the 2838 candidate copolymers with the predicted σ, T_{d5}, TM and MePerm can be found in the ESI.† It is worth noting that these copolymers shown in Table 1 are not included in the 198 reported copolymers, and only the copolymer with the best selectivity associated with a unique hydrophilic monomer was presented. The σ of the top copolymers ranged from 17.6 to 21.2

Table 1 Top ranked copolymers with best predicted values and overall proton conductivity (σ), the degradation temperature at 5% weight loss (T_{d5}), methanol permeability (MePerm) and tensile modulus (TM) performance better than Nafion 117. The table is sorted by selectivity (σ/MePerm) in decreasing order. The *R* group in each monomer denotes the position of polymerization. Refi and Refo represent the original reports for the hydrophilic and hydrophobic monomers used in the fabrication of PEMs

	Hydrophilic	Hydrophobic	FraLic	σ (S m^{-1})	T_{d5} (°C)	MePerm (10^{-7} cm^2 s^{-1})	TM (GPa)	Selectivity (10^5 S s cm^{-3})	Refi	Refo
	Nafion 117	—		17.5	266.2	11.5	0.3	—		
1			0.7	18.2	331.8	2.7	1.3	6.5	48	49
2			0.6	19.1	304.2	3.2	1.2	5.9	50	51
3			0.4	17.9	322.7	3.1	1.2	5.7	52	49
4			0.6	17.6	300.7	3.3	1.2	5.2	51	10
5			0.8	20.9	269.9	4.0	1.1	5.1	53	49
6			0.6	20.1	297.8	3.9	1.1	5.1	43	54
7			0.6	19.3	328.4	3.8	1.3	5.1	55	56
8			0.4	21.2	296.9	4.1	1.1	5.0	57	58
9			0.4	18.2	278.8	3.6	1.5	4.9	5	59
10			0.4	19.1	326.5	3.9	1.1	4.8	49	49

S m^{-1}, the thermal and mechanical performances are all superior to those of Nafion 117, and the MePerm of these copolymers is only around one quarter of that of Nafion 117. The top hydrophilic monomer was reported by Dong *et al.*[48] and a series of sulfonated methoxyphenyl-containing poly(arylene ether ketone) based membranes was fabricated; the σ in the original report was in the range of 7.1–29.4 S m^{-1}. The corresponding hydrophobic monomer was reported by Shin *et al.*,[49] and the σ of the membranes spanned the range of 14.4–25.3 S m^{-1}. For the MePerm and TM, all have preferable values as compared with those of Nafion 117. This indicates that novel hydrocarbon-based sulfonated copolymers with overall performance better than the "golden standard" do exist. Better candidate materials found through introducing densely clustered ion-conducting groups (*e.g.*, multiple sulfonic groups on aromatic rings *etc.*) or ketone functional groups are worth exploring to solve the bottleneck problem in proton exchange membranes.

Conclusions

In this work, through the introduction of the L_{sig}, we can regulate feature engineering, predictive model evaluation and optimal formula identification with a unified setting parameter. Reliable and robust predictive models for the four critical

performance indexes of proton exchange membranes made from hydrocarbon-based sulfonated copolymers, in both regression and classification manners, were constructed. These predictive models were applied to exhaustively screen the combination of hydrophobic and hydrophilic monomers, and this provided 319, 32, 156, and 48 novel copolymers that potentially have superior single performance at the 0.05 significance level in reference to the distribution of the 198 reported copolymers in proton conductivity, methanol permeability, tensile modulus and degradation temperature, respectively. We also found 2838 novel copolymers that have better overall performance than Nafion 117 for fully hydrated membranes at 80 °C. This work provides a way to learn from past knowledge, and rationally design novel materials with the aid of machine learning. Further experimental validations for these predicted formulas should greatly promote our understanding of the materials used to fabricate advanced proton exchange membranes.

There are also some defects of the predicted copolymer candidates used for proton exchange membranes found in this work. This is also a general problem for many machine learning, rational design and material genome approaches, where system-dependant factors may dominate the increase or decrease of a performance index. For example, the feasibility of synthesis, desulfonation in hydrophilic monomers, acid–base interactions *etc.* may exclude some top-ranked copolymer candidates screened out in this work. On the other hand, for materials used in proton exchange membranes, they may be operated under harsh conditions, such as dehydrated and low temperature conditions. In this situation, Nafion still outperforms most reported hydrocarbon-based sulfonated copolymers nowadays. Meanwhile, one of the other important performance indexes, the long-term reliability considering the operation stability of materials during hydration–dehydration cycles and heating–cooling cycles measured in membranes or in a membrane electrode assembly, was not predicted, due to the limited available data. Further enclosing of the long-term reliability may further reduce the candidates that have overall performance better than Nafion 117. These problems, as shown from the narrowing-down of candidates by picking out 59 out of 166 hydrophilic and 154 out of 175 hydrophobic monomers whose overall best performance was better than Nafion, can be solved based on this study. Overall, we are expecting experts to practically validate these candidate copolymers selected through machine learning based rational design.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (21774128), Major State Basic Research Development Program (2015CB655302), Key Research Program of Frontier Sciences (QYZDY-SSW-SLH027) and One Hundred Talent Project of the Chinese Academy of Sciences, and State Key Laboratory of Separation Membranes and Membrane Processes (Tianjin Polytechnic University) No. M2-201707. We are grateful to the Computing Center of Jilin Province for essential support.

References

1 H. W. Zhang and P. K. Shen, *Chem. Rev.*, 2012, **112**, 2780–2832.
2 K. A. Mauritz and R. B. Moore, *Chem. Rev.*, 2004, **104**, 4535–4585.
3 R. Devanathan, *Energy Environ. Sci.*, 2008, **1**, 101.
4 D. W. Shin, M. D. Guiver and Y. M. Lee, *Chem. Rev.*, 2017, **117**, 4759–4805.
5 P. Zheng, J. Liu, X. Liu and K. Jia, *Solid State Ionics*, 2017, **303**, 126–131.
6 T. D. Dong, J. H. Hu, M. Ueda, Y. M. Wu, X. Zhang and L. J. Wang, *J. Mater. Chem. A*, 2016, **4**, 2321–2331.
7 S. W. Lee, J. C. Chen, J. A. Wu and K. H. Chen, *ACS Appl. Mater. Interfaces*, 2017, **9**, 9805–9814.
8 R. Chen, G. Li, S. Yang, M. Xiong and J. Jin, *Solid State Ionics*, 2017, **300**, 157–164.
9 M. K. Ahn, S. B. Lee, C. M. Min, Y. G. Yu, J. Jang, M. Y. Gim and J. S. Lee, *J. Membr. Sci.*, 2017, **523**, 480–486.
10 S. Feng, J. Pang, X. Yu, G. Wang and A. Manthiram, *ACS Appl. Mater. Interfaces*, 2017, **9**, 24527–24537.
11 K. Oh, K. Ketpang, H. Kim and S. Shanmugam, *J. Membr. Sci.*, 2016, **507**, 135–142.
12 M. D. T. Nguyen, S. Yang and D. Kim, *J. Power Sources*, 2016, **328**, 355–363.
13 L. Lin, Z. Chen, Z. Zhang, S. Feng, B. Liu, H. Zhang, J. Pang and Z. Jiang, *Polymer*, 2016, **96**, 188–197.
14 L. Chen, S. Zhang, Y. Jiang and X. Jian, *RSC Adv.*, 2016, **6**, 75328–75335.
15 C. Wang, B. Shen, Y. Zhou, C. Xu, W. Chen, X. Zhao and J. Li, *Int. J. Hydrogen Energy*, 2015, **40**, 6422–6429.
16 H. Yao, N. Song, K. Shi, S. Feng, S. Zhu, Y. Zhang and S. Guan, *Polym. Chem.*, 2016, **7**, 4728–4735.
17 H. Yao, K. Shi, N. Song, N. Zhang, P. Huo, S. Zhu, Y. Zhang and S. Guan, *Polymer*, 2016, **103**, 171–179.
18 Y. Z. Liu, L. Ding, J. Liu, Z. J. Yang and T. W. Xu, *Acta Polym. Sin.*, 2018, 797–813.
19 S. W. Lee, J. C. Chen, J. A. Wu and K.-H. Chen, *ACS Appl. Mater. Interfaces*, 2017, **9**, 9805–9814.
20 R. M. Chen, G. Li, S. L. Yang, M. Y. Xiong and J. H. Jin, *Solid State Ionics*, 2017, **300**, 157–164.
21 G. Wang and M. D. Guiver, *J. Membr. Sci.*, 2017, **542**, 159–167.
22 J. F. Zheng, W. H. Bi, X. Dong, J. H. Zhu, H. C. Mao, S. H. Li and S. B. Zhang, *J. Membr. Sci.*, 2016, **517**, 47–56.
23 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
24 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.
25 D. J. Audus and J. J. d. Pablo, *ACS Macro Lett.*, 2017, **6**, 1078–1082.
26 Z. Li, S. Wang, W. S. Chin, L. E. Achenie and H. Xin, *J. Mater. Chem. A*, 2017, **5**, 24131–24138.

27 Y.-C. Lin, Y.-J. Lu, C.-S. Tsao, A. Saeki, J.-X. Li, C.-H. Chen, H.-C. Wang, H.-C. Chen, D. Meng, K.-H. Wu, Y. Yang and K.-H. Wei, *J. Mater. Chem. A*, 2019, **7**, 3072–3082.
28 K. Takahashi, L. Takahashi, I. Miyazato and Y. Tanaka, *ACS Photonics*, 2018, **5**, 771–775.
29 G. Borboudakis, T. Stergiannakos, M. Frysali, E. Klontzas, I. Tsamardinos and G. E. Froudakis, *npj Comput. Mater.*, 2017, **3**, 40.
30 L. Y. Liu, W. D. Chen and Y. Q. Li, *J. Membr. Sci.*, 2018, **549**, 393–402.
31 L. Y. Liu, W. D. Chen and Y. Q. Li, *J. Membr. Sci.*, 2016, **504**, 1–9.
32 G. A. Landrum, *Rdkit: Open-source cheminformatics software*, http://www.rdkit.org/.
33 https://www.chemcomp.com/journal/descr.htm.
34 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
35 M. Melanie, *An introduction to genetic algorithms*, MIT Press, Cambridge, MA, 1996.
36 D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, *Science*, 2011, **334**, 1518–1524.
37 C. R. Boyd, M. A. Tolson and W. S. Copes, *J. Trauma*, 1987, **27**, 370–378.
38 A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé and G. Pujadas, *Methods*, 2015, **71**, 58–63.
39 R. L. Thorndike, *Psychometrika*, 1953, **18**, 267–276.
40 H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech and B. D. Freeman, *Science*, 2017, **356**, eaab0530.
41 C. Y. Wang, D. W. Shin, S. Y. Lee, N. R. Kang, Y. M. Lee and M. D. Guiver, *J. Membr. Sci.*, 2012, **405–406**, 68–78.
42 V. S. Silva, A. Mendesa, L. M. Madeira and S. P. Nunes, *J. Membr. Sci.*, 2006, **276**, 126–134.
43 W. Li, Z. Cui, X. Zhou, S. Zhang, L. Dai and W. Xing, *J. Membr. Sci.*, 2008, **315**, 172–179.
44 F. Zhang, N. Li and S. Zhang, *J. Appl. Polym. Sci.*, 2010, **118**, 3187–3196.
45 M. Feng, R. Qu, Z. Wei, L. Wang, P. Sun and Z. Wang, *Sci. Rep.*, 2015, **5**, 9859.
46 L.-K. Chen, C.-S. Wu, M.-C. Chen, K.-L. Hsu, H.-C. Li, C.-H. Hsieh, M.-H. Hsiao, C.-L. Chang and P. P.-J. Chu, *J. Membr. Sci.*, 2010, **361**, 143–153.
47 L. M. Lin, Z. Chen, Z. P. Zhang, S. N. Feng, B. Liu, H. B. Zhang, J. H. Pang and Z. H. Jiang, *Polymer*, 2016, **96**, 188–197.
48 B. Dong, Y. Wang, J. Pang, S. Guan and Z. Jiang, *RSC Adv.*, 2015, **5**, 107982–107991.
49 D. W. Shin, S. Y. Lee, N. R. Kang, K. H. Lee, M. D. Guiver and Y. M. Lee, *Macromolecules*, 2013, **46**, 3452–3460.
50 E. P. Jutemar and P. Jannasch, *J. Polym. Sci., Part A: Polym. Chem.*, 2011, **49**, 734–745.
51 F. Fu, H. Xu, Y. Dong, M. He, T. Luo, Y. Zhang, X. Hao, T. Ma and C. Zhu, *Solid State Ionics*, 2015, **278**, 58–64.
52 H. Dai, H. Zhang, Q. Luo, Y. Zhang and C. Bi, *J. Power Sources*, 2008, **185**, 19–25.
53 S. Seesukphronrarak, K. Ohira, K. Kidena, N. Takimoto, C. S. Kuroda and A. Ohira, *Polymer*, 2010, **51**, 623–631.
54 L. Fu, G. Xiao and D. Yan, *J. Mater. Chem.*, 2012, **22**, 13714.
55 G. Wang, K. H. Lee, W. H. Lee, D. W. Shin, N. R. Kang, D. H. Cho, D. S. Hwang, Y. Zhuang, Y. M. Lee and M. D. Guiver, *Macromolecules*, 2014, **47**, 6355–6364.
56 L. Fu, G. Xiao and D. Yan, *J. Membr. Sci.*, 2010, **362**, 509–516.
57 K. Si, D. Dong, R. Wycisk and M. Litt, *J. Mater. Chem.*, 2012, **22**, 20907–20917.
58 Y. Gao, G. P. Robertson, M. D. Guiver, S. D. Mikhailenko, X. Li and S. Kaliaguine, *Macromolecules*, 2004, **37**, 6748–6754.
59 Z. Zhang, L. Wu and T. Xu, *J. Mater. Chem.*, 2012, **22**, 13996.