Determining ion activity coefficients in ion-exchange membranes with machine learning and molecular dynamics

Hishara Keshani Gallage Dona[1, ‡], Teslim Olayiwola[2, ‡], Luis A. Briceno-Mena[2], Christopher G. Arges[3, *], Revati Kumar[1, *], Jose A. Romagnoli[2, *]

[1]*Department of Chemistry, Louisiana State University, Baton Rouge, Louisiana 70803, United States.*
[2]*Cain Department of Chemical Engineering, Louisiana State University, Baton Rouge, Louisiana 70803, United States.*
[3]*Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 160802, United States.*

[*]Corresponding author: jose@lsu.edu

[‡]Authors contributed equally

Abstract

The activity coefficients of ions in polymeric ion-exchange membranes (IEMs) dictates the equilibrium partitioning coefficient of the ions between the membrane and the liquid. It also affects ion transport processes, such as conductivity, in ion-exchange membranes. Accurately predicting the ion activity coefficient without experimental data has been elusive as most models are empirical or semi-empirical. This work employs an embedding process that maps microscopic and macroscopic properties for modeling of ion activity coefficients in IEMs with molecular dynamics and machine learning (ML). This strategy is effective for accurately predicting activity coefficients in various IEMs materials – including random copolymer and block copolymer systems. ML algorithms are increasingly being used for the analysis of complex systems when limited knowledge is available. The framework uses small experimental activity coefficient datasets in conjunction with polymer structure information and molecular attributes describing the solvation of ions and polymers to predict the ion activity coefficient in IEMs. Two different ML models were developed to estimate the molecular attributes and the ion activity coefficient. The best ML model accurately predicts the solvation descriptors and ion activity coefficient with an average mean absolute error of <7% and 10%, respectively. Adopting the said approach allow for the estimation of ion activity coefficients in IEMs without the need for new time-consuming MD simulation runs and experiments.

1. Introduction

Electrically-driven ionic separations has been historically important for water desalination and the production of ultra-pure water for semiconductor, food, and pharmaceutical manufacturing. They are also emerging as platforms for the recovery of organic acids from processed biomass streams[1,2] and valuable metals from waste and hydrometallurgy streams[3] as well as remediating water streams with bad actors (e.g., PFAS[4],[5] and heavy metals[6]). Electrically-driven ionic separations include electrodialysis (ED), membrane capacitive deionization (MCDI) and electrodeionization (EDI) [2–6], and each of these processes utilize ion-exchange membranes (IEMs). The two main types of IEMs are anion exchange membranes (AEMs) and cation exchange membranes (CEMs)

The chemistry and molecular architecture of IEMs influence their ionic conductivity and permselectivity and other transport properties such as osmotic drag.[10–13] Thus, chemistry and microstructure of IEMs have a profound impact on the said properties. The ionic conductivity of IEMs is inversely commensurate to the ohmic overpotential of electrochemical separations. In other words, it can have a large effect on the energy efficiency of electrochemical separations when trying to operate at high current density values. Operating at high current density is important to reducing the size of the separation system and reducing capital costs that arise from ion-exchange membranes and electrodes.[1,2,14] There are many parameters within polymeric IEMs that affect conductivity and permselectivity and they include repeat unit sequence[14] repeat unit chemistry, microstructure[14], concentration of tethered charge groups, the external salt concentration,[14–17] the types of salts,[15,18] and water content[16]. Establishing the said descriptors to ion transport rates within IEMs is time consuming and expensive because many experiments must be performed for a given structure and chemistry. Hence, it is desired to create accurate computational models to estimate these properties with small data sets for training and validation.

Computational models that can bridge molecular descriptors to material properties and performance in electrochemical separation platforms are vital for the design and optimization of these systems.[19] Moreover, finding descriptors that accurately capture the ionic transport between the solution and polymer could reveal insights to the various mechanisms of transport processes in materials. One such descriptor, which is the focus of this work, is the activity coefficient of co-ions and counterions in polymeric IEMs. The activity coefficient is related to the Gibbs excess energy (when the Gibbs energy is selected as the partial molar property)[20] – which is the difference between the actual Gibbs free energy change and the ideal Gibbs free energy change. Many of the counterion and co-ion properties in IEMs is a function of concentration. Under non-ideal behavior, the activity coefficient deviates from one and captures the activity these ions exert within membranes and liquids – which is often lower than the actual concentration of these species. Under constant temperature and pressure, the equilibrium is defined by the electrochemical potential for a given ion being equal in the liquid phase to that in the membrane phase.[21] Exploiting the chemical equilibrium criteria allows one to determine the partitioning coefficient for one ion between the

IEMs and liquid phase. The permselectivity of an IEMs, which is the propensity of the membrane to transport one ion or another, is a product of the partitioning coefficient multiplied by the ratio of the ionic mobility coefficients. Equation 1 depicts permselectivity (P_j^i) and the partitioning coefficient (K_j^i) of one ion ('i') over another ion ('j') as a function of ion activity (a_i) in the solution (i.e., liquid) and membrane phases (superscripts 's' and 'm', respectively).[22] Equation 2 is the definition of ion activity. γ_i is the activity coefficient and C_i is the concentration.

$$P_j^i = \frac{a_i^m / a_i^s}{a_j^m / a_j^s} \cdot \frac{u_i^m}{u_j^m} = K_j^i \cdot \frac{u_i^m}{u_j^m} \tag{<1>}$$

$$a_i = \gamma_i C_i \tag{<2>}$$

Despite its relevance, accurate ion activity coefficient models based upon the physics and chemistry of the ion-exchange membranes are lacking. Activity coefficient models for ions in liquids are semi-empirical or empirical and tend to be complex (e.g., Debye–Hückel and Pitzer).[20] To model the activity coefficient in IEMs, models such as Manning-Donnan correlation[21] (and derivatives of this model) were proposed to determine the ion activity coefficient in IEMs when interfaced with an aqueous electrolyte solution (e.g., NaCl, or KI). The Manning-Donnan showed reasonable predictive ability for the studied scenario provided the dielectric constant (ε) within the hydrated IEMs can be reasonably determined and the average distance between fixed charges (b) can be estimated. Notably, the Manning-Donnan, and its modified versions, is chemistry agnostic and has been shown to fail in certain cases when IEMs are interfaced with dilute salt solutions (e.g., $\leq$ 0.1 M).[21] In the former scenario of chemical specificity, Arges and Kumar and co-workers showed that AEMs with tethered imidazolium groups rather than quaternary ammonium favor organic acid anion uptake.[1,2] The current embodiments of the Manning-Donnan model are unable to account for the said observations with organic acid anions with AEMs. In the latter case of the Manning-Donnan model failing under dilute solutions, an alternative approach is to fit the Manning parameter (ξ) to datasets rather than estimating ε and b to calculate the Manning parameters[23]. It is also worth noting that the role of hydration has a profound impact on ion transport and activity coefficients, and it is only captured in the Manning-Donnan model with ε.

The utility of the empirical and semi-empirical activity coefficient models often needs large data sets for estimating the empirical constants in these models. Since there is a lack of data for ion activity coefficients in various IEMs materials, attaining effective and accurate activity coefficient models for new IEMs systems is elusive. To address this problem, we have developed a strategy of ML tools with molecular dynamics simulations to model and predict activity coefficients of ions in IEMs. ML has been successfully applied in other research endeavors ranging from discovery and synthesis of new materials to prediction of material properties and device level performance.[8,19] ML is a subfield of artificial intelligence and computer science which aims to simulate human learning processes using data and algorithms, gradually increasing the accuracy of the results. ML represents a unique tool to streamline model development for attaining accurate activity coefficients of IEMs. Specifically, in polymer informatics, these methods have found successful applications in predicting key properties such as glass transition temperature,[23,24] gas permeability,[25–27] and phase diagrams.[28]

Increase in computational power has boosted the growth of physics-based tools like molecular dynamics (MD). For example, MD simulation has been used to study polymer phase behavior[29,30] and estimate properties such as thermal conductivity,[31,32] water diffusion constant,[14] and diffusion coefficient.[33] MD simulations provide valuable information about the solvation and spatial distribution of counterions and ionic groups along the polymer backbone, which helps to explain the measured conductivity and ionic activity in experiments. In previous studies, Arges et al.[17,14,2] demonstrated the importance of capturing water activity in IEMs materials using the results from experiments and simulations to understand ion activity and ion dissociation. MD simulations provide information about ion-ion and ion-IEMs pair distribution functions, coordination and hydration numbers around charged moieties using the corresponding radial distribution function.

Here, a computational framework is proposed to predict the activity coefficient of ions in charged IEMs. This approach uses a ML model that leverages molecular scale attributes from MD simulations such as on solvation properties with the IEMs. Both existing and novel data were used in the development of this ML-MD modeling strategy. More specifically, ML methods such as Support Vector Regression (SVR), Artificial Neural Network (ANN), and Random Forest (RFR) were used to connect the polymer fingerprint (based on chemical structure) and molecular level attributes to the macroscopic attribute of IEMs (i.e., activity coefficient). To train the ML methods and to validate predictions, experimental data on ion activity coefficient ($ɣ^{M+}ɣ^{M-}$) in polymeric IEMs (and thin films variants) were obtained from published literature, which contains detailed information regarding the polymer structure, the ion exchange capacity (IEC), water uptake, and ion activity coefficient. The ML-MD modeling strategy accurately predicted activity coefficients of ions in IEMs with less than 10% error and only required small data sets for ML training. The activity coefficient obtained from this ML-MD framework study will be used in a future study to quantify the resistance of the IEM, specifically the Donnan potential responsible for the actual separation of ions, within physics-aware models for electrochemical separations systems, providing a more comprehensive understanding of the underlying mechanisms of the electrically-driven process.

2. Methods

2.1. Datasets

2.1.1. Experimental data

Data consists of experimentally measured ion activity coefficients in 13 different copolymer IEMs and thin films interfaced with varying salt concentrations. The compiled data includes ionomers with different arrangements of monomeric units (i.e., random – termed random copolymer electrolyte (RCE) and block – termed block copolymer electrolyte (BCE)), different number of side chains and different counterions for the charged monomeric unit. Examples of these copolymers are AR103,[21] AR204,[21] CR61,[21,34] Nafion,[15] polyvinyl alcohol sulfate PVAS,[21,35] poly(styrene-*block*-2-vinylpyridine/n-methylpyridinium iodide) PSbNMP,[14,17] poly(2-acrylamido-2-methylpropanesulfonic acid-block-diethylene glycol dimethacrylate) CEM,[16] poly(ethylene glycol) diacrylate and 2-acrylamido-2-methyl-1-propanesulfonic acid (XL-AMPS-PEGDA).[36] Note: AR103, AR204, and CR61 are commercial IEMs. **Figure 1** illustrates the chemical representations of the sampled ionomers. The data curation process considered information about the chemical structure of the IEMs, IEC (expressed as mequiv/g) and water uptake (expressed as g of water per g of dry IEMs). Lastly, each entry in our data set was compared with other sources to ensure the accuracy of the data.

2.1.2. Molecular dynamics data

Classical canonical MD simulations were carried out to gain molecular insight into the activity coefficients of ion exchange polymeric IEMs in ionic separations. Approximately, 100 simulations were conducted using the LAMMPS[37] software package for a ~40 A^0 cubic system containing IEMs, ions, and water. Herein, the IEMs were modeled with either the OPLS-AA[38,39] or GAFF2[40] forcefield and the water molecules with the TIP3P[41] forcefield. Depending on the forcefield, the partial charges of the constituent elements in the IEMs were computed with GAUSSIAN[42] and ANTECHAMBER.[43] Since IEMs are long-chain molecules composed of several repeating units of one (or two) monomers that are difficult to represent completely, we adopted a polymer chain with 6–9 repeating units for all our studied IEMs and thin film variants. Despite their shorter length, these 6-mer (or 9-mer) chains have many of the same characteristics as the finitely long chain. To conduct MD simulations that are representative of experimental conditions, the water molecules per tethered ions (for a given volume) were calculated as described by Kohl and co-workers.[44] The chemistry of the tethered charges, counterions, and co-ions were selected based upon the IEMs materials' chemistry and salt used in the external solution. In each MD run, a total of 17 polymer chains were solvated in a cubic water box with a box length of around 40 A^0 to replicate the experimental setup. (see **Table S1** for details on IEMs in various simulations). For the different IEMs the solvation properties, such as radial distribution function in the first hydration shell around ions/charge moiety, were investigated using canonical MD simulations at 300K temperature. Detailed information on the simulation setup is provided in the supporting information.

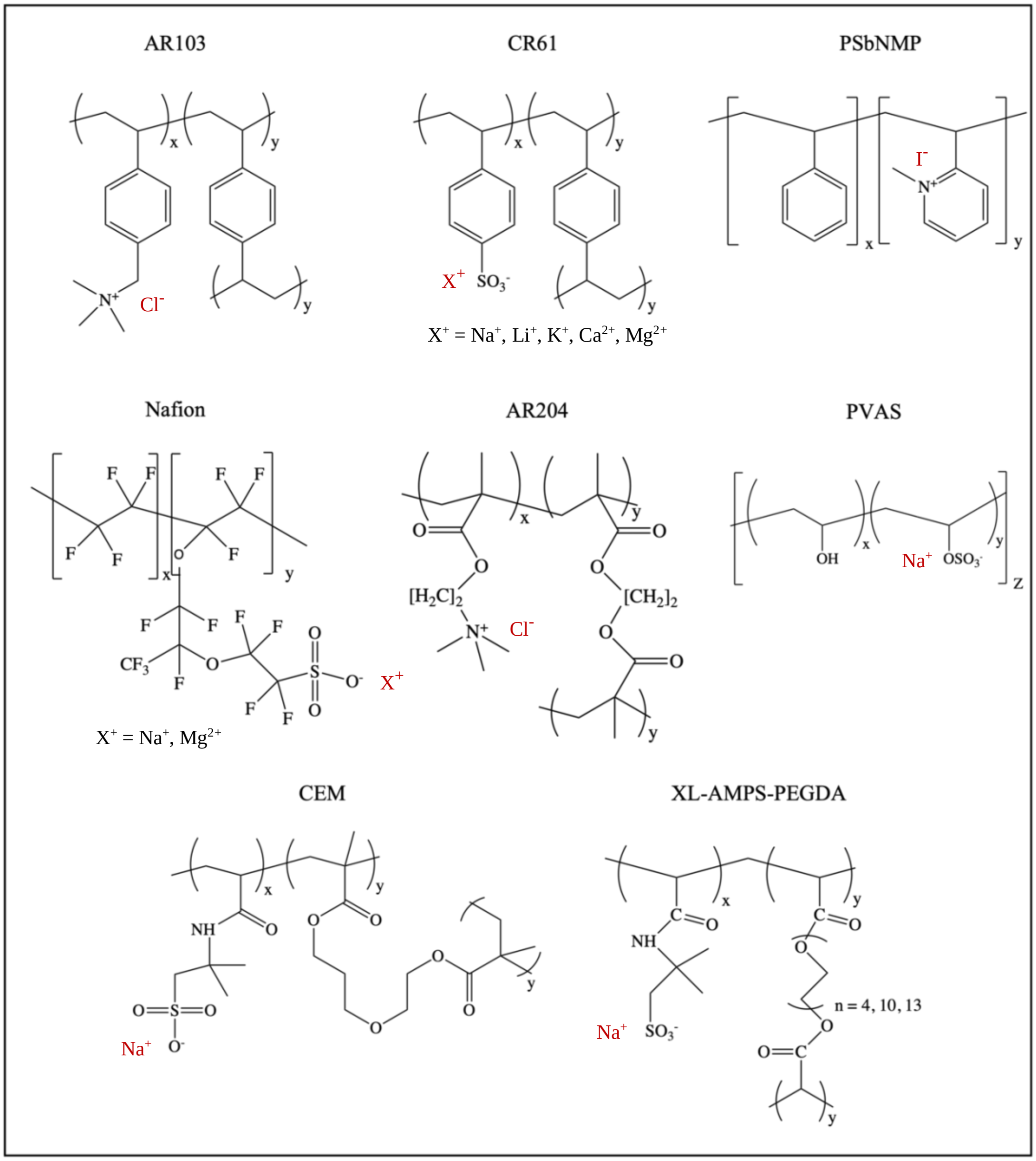

Figure 1. Molecular structure of the ionomers, along with the accompanying counter-ions (X^+).

2.2. Feature Engineering

Information about molecular properties and system conditions must be properly extracted and organized for a ML model to use them as predictors. In this work, three strategies are combined to produce input features that integrate molecular structure information, system conditions, and molecular dynamics calculations. First, Morgan Fingerprints (MF),[45] which have found tremendous applications in ML,[23,27,28] was selected to represent the molecular structure of the different IEMs. The MF algorithm can identify substructures in a molecule, and then represent them in a sparse bit vector. In this work, the MF of each polymer was generated using the RDKit package.[46] The radius and size of the fingerprint were set to 3 and 128-bit, respectively. Furthermore, each IEMs contains different ion types (i.e., tethered ion, co-ion, and counterion species) and there is a need to encode their effects in the modeling framework. Thus, these ion species were represented as binary vectors obtained via one-hot encoding. This one-hot encoding involves converting categorical features into numeric arrays. One-hot encoding was implemented via the Scikit-Learn package.[47] Finally, twelve (12) properties obtainable from equilibrium MD as described in ***Section 2.1.2*** were considered as additional descriptors to quantify the microscopic behavior of the IEM and thin film variant systems. These MD parameters defined as solvation descriptors including coordination numbers from the first solvation shell and various radial distribution functions (RDF). Specifically, the chosen RDF features include position of first minima and peak height, peak position and coordination numbers of counterion and oxygen of water, tethered charge and oxygen of water, and tethered charge and counterion.

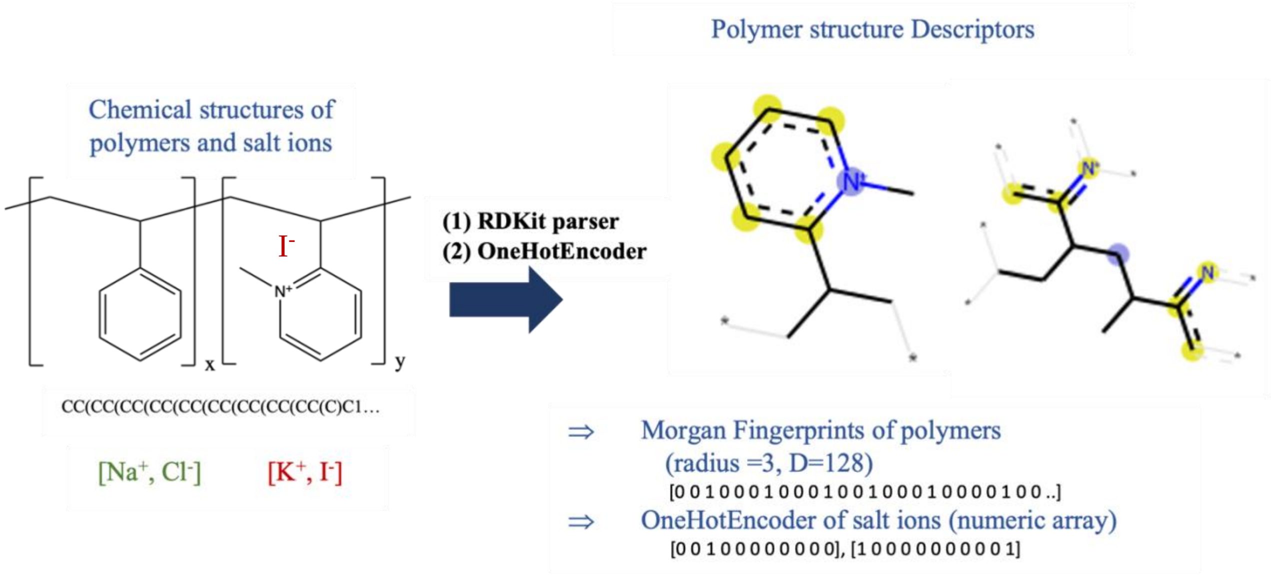
Polymer structure Descriptors
Chemical structures of polymers and salt ions
I-
N+
x
y
CC(CC(CC(CC(CC(CC(CC(CC(CC(C)C1…
[Na+, Cl-]
[K+, I-]
(1) RDKit parser
(2) OneHotEncoder
N+
N
N
⇒ Morgan Fingerprints of polymers
(radius =3, D=128)
[0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0 0 ..]
⇒ OneHotEncoder of salt ions (numeric array)
[0 0 1 0 0 0 0 0 0 0 0], [1 0 0 0 0 0 0 0 0 0 1]

Figure 2. Workflow for handling the preprocessing of the input features for ML model development. The chemical structure of the polymers were generated in Avogadro and then converted to SMILES and finally converted to Morgan fingerprints with the RDKit package. The salt ions i.e., categorical variables are converted into numeric arrays with OneHotEncoder in the Scikit-Learn package. The highlighted areas of polymer structure descriptors illustrate the selected fingerprint bits. Herein, the blue, yellow, and the gray colors depicts the central atom in the environment, aromatic atoms, and the aliphatic ring atoms, respectively.

2.3. Machine learning models

ML algorithms tend to perform differently depending on the nature of the training data. Although there are some guidelines that help in determining which algorithm to use, in most application several approaches have to be explored to find the best one for the given problem. In this work, three representative algorithms, namely, support vector regression, decision trees, and artificial neural networks were investigated.

A support vector machine (SVM) is an algorithm that tries to find the hyperplane that gives the best separation between two linearly separable regions. The idea of a SVM was first proposed by Boser et al.[48] in 1996 and remains a powerful ML modeling method, especially for low-dimensional and small datasets. A useful extension of this method is the introduction of a kernel that transforms non-linearly separable data into linearly separable data by adding an extra dimension. The classification problem, that is finding the boundary between two or more classes, can be reframed as a regression problem by setting the algorithm to find the two hyperplanes that contain both the training data and the predicted values, while minimizing the distance between said hyperplanes.[49] This latter implementation is called "Support Vector Regression".

Decision-tree based regression (DT) and its extensions have been shown to be versatile modeling frameworks with good performance in many different applications.[50] The most common implementations use the idea of random forest (RF) in which multiple models (decision tree models) are fit to the training data, and the final prediction is made via a voting mechanism. This type of approach, typically referred to as ensemble methods, helps prevent overfitting by training multiple models, each with a different bias, and then averaging their predictions. Another advantage of DT is that they tend to perform well with high-dimensional data.

Artificial neural networks (ANN) are a type of machine learning method that constructs a model as a network of nodes connected by edges, arranged in layers (also known as fully connected layers). Each node multiplies the outputs from the previous layer's nodes by their corresponding weights (represented by the edges), adds them, and passes the result through an activation function to feed the next layer of nodes. To train ANN, the predicted output is calculated for a given set of inputs, the error between the predicted and true values is computed, and the weights are adjusted accordingly. The high tunability of ANNs, especially their architecture (number of nodes and layers), makes them flexible and thus attractive for a wide variety of applications. A considerable drawback, however, is the fact that the amount of data required to effectively train the model increases greatly with model complexity.

2.4. Training of models

We explored two different model development frameworks to estimate the ion activity coefficient. The first framework involves predicting the solvation descriptors of IEM type materials. In developing this model, the polymer fingerprints, salt concentration, number of water molecules per tethered ions and the one-hot encoded categorical features of the salt ions are considered as the input while the solvation descriptors as the output. In the second modeling task, a new ML model was constructed to predict the ion activity coefficient in the IEM and thin film variants. In developing this ML model, the polymer fingerprints, encoded salt ions, number of water molecules per tethered ions, solvation descriptors (obtainable from the first ML model) and salt concentration were considered as the model input while the activity coefficient as the target variable. In this work, all the model development phases were implemented with the aid of the python packages Scikit-Learn[47] and Pymoo.[51]

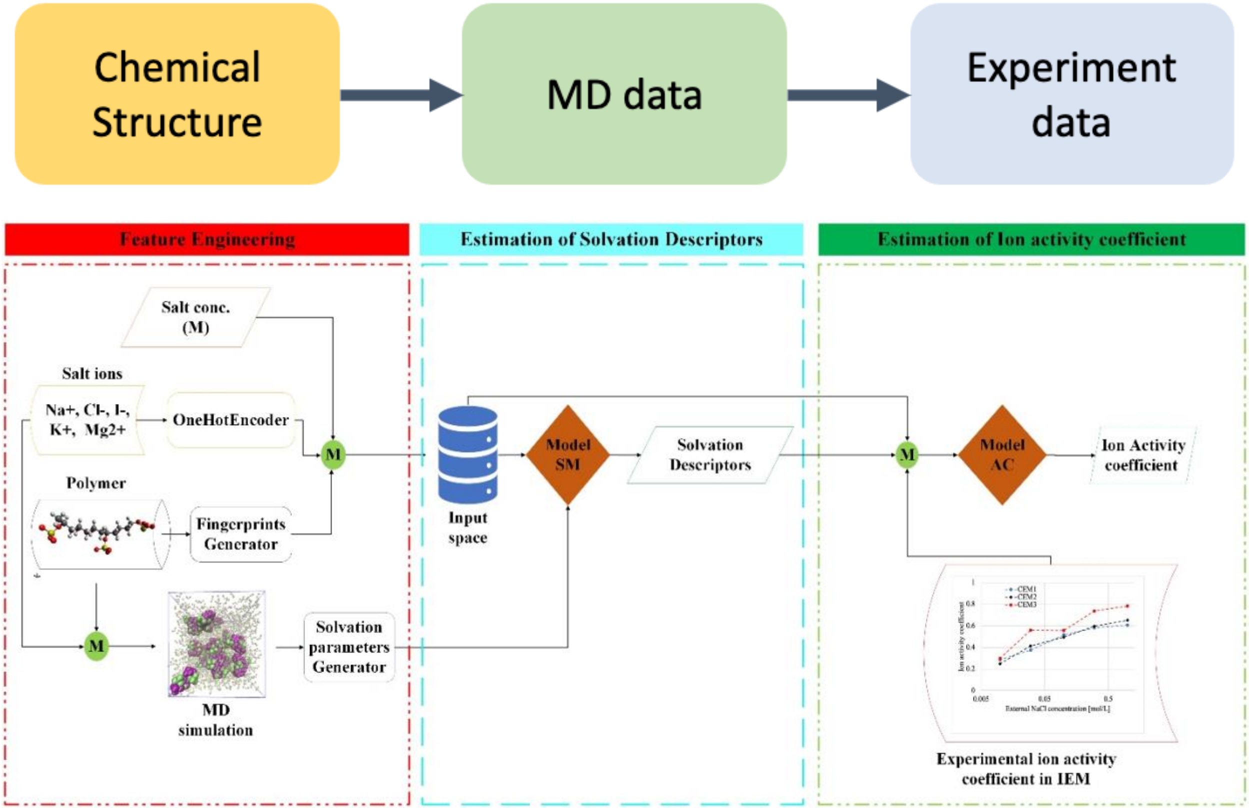

Figure 3. The workflow of the modeling strategies involved in this study. Two major steps in setting up a proper ML model to achieve the two model frameworks. Going through the steps, it shows how the chemical structures (molecular fingerprints) are fed to predict the solvation parameters (RDF, coordination numbers, ect.) and then finally, the experiment data (i.e., activity coefficient).

One major step in ML development is to ensure that the input features are in the same scale without losing its original distribution, else the disparity in data ranges can significantly alter the generalizability of the resulting model. Data normalization was needed only for the solvation descriptors which were converted into a standard range of -1 and 1. Also, normalization was implemented because variables such as the coordination numbers show small changes in values even when their controlling variables (i.e., salt concentration) changes significantly. By normalizing the input variables, any small changes in any variables becomes more prominent. The resulting model from this framework should serve as a useful tool for future users when there exists no MD data for their chosen IEMs.

In developing the two models, we split the different corresponding dataset based on a chosen ratio. Data splitting is frequently used in ML to prevent overfitting. In this study, 80% of the dataset served as training data, and the remaining as test data. To identify issues like overfitting or selection bias and to provide insight into how the model will generalize to an independent dataset during the training process, the idea of cross validation was implemented during the training phase. The training data in this study was further divided into 5 folds such that 1 out of every 5 data points in the training data was used to test the model's ability to predict data that was not used in its estimation. It is noteworthy to mention that the original test data was never used during the training process and thus, any prediction with these data gives the best representation of the developed model. Furthermore, statistical measures namely coefficient of determination (R^2) and mean absolute error (MAE) were computed to assess the performance of the developed models.

To obtain the best model, there is a need to obtain the best hyperparameters that accurately define the best-performing model. Manually finding the right set of hyperparameters for a ML model can be challenging and time-consuming. Therefore, a systematic way of choosing the right combination of hyperparameters is needed to efficiently deploy the models and understand their performance. Hyperparameters can be explored using various methods, such as random search, grid search, and Bayesian search, to find the optimal combination. However, evolutionary algorithms have proven useful in addressing multi-objective optimization problems,[8,52,53] and shown to perform better for multi-objective hyperparameter optimization[54]. For this study, hyperparameter selection for the regression methods was defined as a Mixed Integer Nonlinear Programming problem and solved using the Non-dominated Sorting Genetic Algorithm (NSGA-II), [55] a multi-objective evolutionary algorithm implemented in Pymoo.[51]

Principal component analysis (PCA) is a dimensionality reduction technique used for data exploration and analysis with the presence of big data, including numerous variables, multiple observations for each variable, and closely corelated data. By employing this technique, a low-dimension space with reduced set of features that represents the original data set can be identified. The original data are projected to the first principal axis to create the first principal component, which captures the greatest variation in the data while the second principal component, which accounts for most of the variance in the data not covered by the first principal component. Up until the entire data matrix is deconstructed, each succeeding principal component explains the maximum amount of variance it can give the restriction that it is orthogonal to the preceding principal components. The principal element with the lowest variance in the data may be eliminated in accordance with user requirements. The advantages of this technique include the elimination of collinearity, noise reduction, need for less storage, and the ability to visualize in 2D and 3D spaces.[56]

3. Results and Discussion

3.1. Polymer space under exploration

The dataset under study consists of 13 IEMs (and thin film variants) linked to experimental ion activity coefficients. To assess the difference between the available IEMs, the Tanimoto similarity[57,58] based on the Morgan Fingerprints of the SMILES representation of each polymer was computed. Similarity metrics including Tanimoto or Dice or Cosine similarity has been found to be a better metric compared to other metrics that work based on Manhattan or Euclidean distance.[59] Stemming from this discovery, Tanimoto similarity is the most popular and widely accepted similarity metrics for polymer chemistry. In this study, the Tanimoto (Tc) and Dice (Dc) similarity index were computed to understand the similarities in sampled polymer space. The results of the analysis are depicted in **Figure 4** and **Figure S1**. Two polymers are considered identical when the Tc (or Dc) equals 1 and completely different when the Tc (or Dc) equals 0. The similarity matrix showed that our selected polymers are very different from one another. Only the RCE and BCE forms of a polymer have high similarity values (>0.7). This is expected because these polymer forms have the same chemistry except with different arrangement of monomers. Overall, a model generated from these polymers will have a good generalization to other polymers with similar constituents because the sampled polymers exhibit great variation.

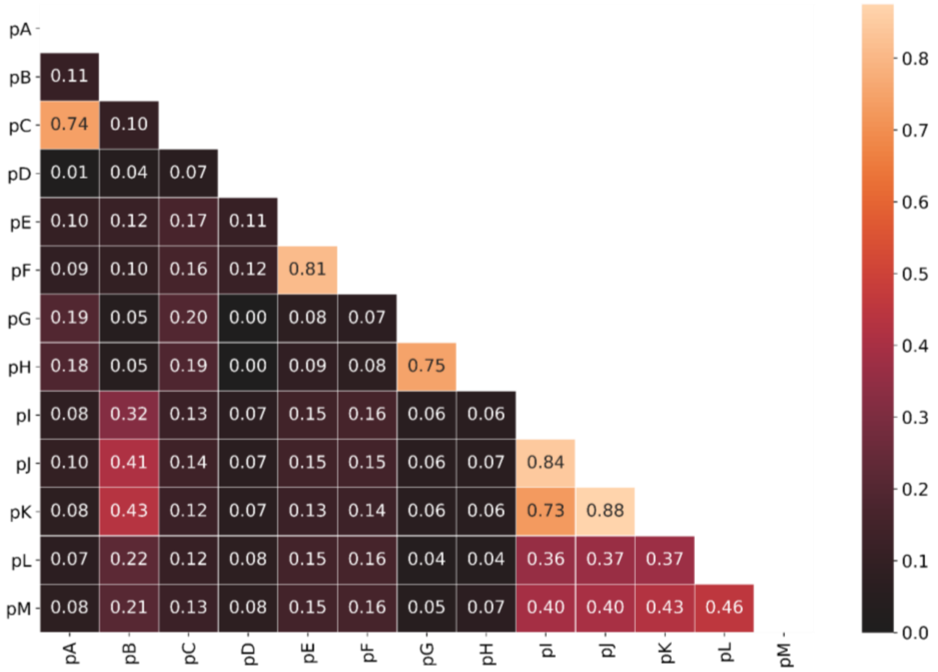

Figure 4. Visualization of the dissimilarity among the studied ionomers. The labels represents AR103 (pA), AR204 (pB), CR61 (pC), Nafion (pD), PVAS (pE, pF), PSbNMP (pG, pH), CEM1 (pI), CEM2 (pJ), CEM3 (pK), XL-(AMPS-PEGDA, pL: n = 4, x/(x+y) = 9%, pM: n = 4, x/(x+y) = 45%). Herein, pA, pB, pC, pI, pJ, pK, pL, pM are crosslinked IEM (model in BCE form for MD simulations), while the PVAS and PSbNMP model in both BCE and RCE forms and Nafion in RCE form. The heat map shows the computed Tanimoto similarity between each polymer. The Tanimoto similarity between ionomers A and B is mathematically equal to a/(a+b+c) where a, b and c are counts of bits in MF of ionomer A but not in B, counts of bits in MF of ionomer B but not in A and counts of bits in MF of ionomer A and B, respectively.

3.2. MD simulations

3.2.1. Solvation descriptors

MD simulations can determine the solvation environment of IEMs at the atomic level. Information on spatial distribution, solvation as well as ion pairing between counterions and co-ions can be studied using the radial distribution function g(r). The first solvation shell of the charged species can be defined using the first minimum in g(r). The water molecules and the counterions are considered in the initial hydration shell of the IEMs if the distance is smaller than this cut-off. Additionally, the peak height and peak position of the g(r) provide more information about the strength of association between tethered charged groups in these IEMs. These g(r) characteristics are all dependent on the polymer structure, type of salt as well as salt concentration of various IEMs. The radial distribution function of the CR61 IEMs at various KCl concentrations are shown in **Figure 5**. The SI provides a detailed explanation of the calculations used to determine these dynamical quantities.

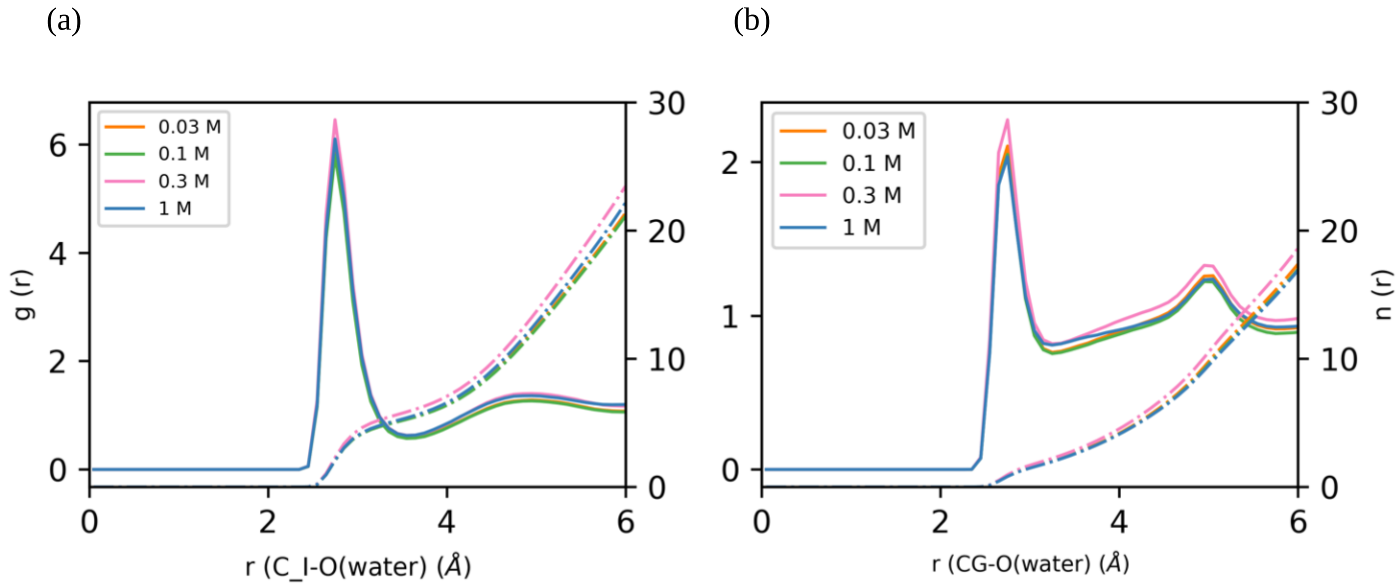

Figure 5. The radial distribution function, g(r), (solid line) and coordination number, n(r), (dashed line) (a) between the counter ion (K+) and (O) water, (b) Charged group (SO_3^-) and O (water) in a CR61 - KCl - water system

3.2.2. Identification of influential descriptors

The important features among the collected polymer properties and solvation descriptors were related to the ion activity coefficient values. Computing the pairwise correlations between the features and target data can offer preliminary insights into the descriptive performance of the selected features. In other words, a correlated feature-target data assists the model to predict the desired properties. The pairwise correlation was computed based on spearman rank correlation. The correlation heat map indicated that the selected features are correlated with the ion activity coefficient, as seen in **Figure 6**.

The spearman's correlations for the number of water molecules per tethered ions and salt concentrations were 0.59 and 0.37, respectively, whereas the correlations for the peak height of RDF between tethered charge (in IEMs) and oxygen of water and coordination number (at first minima) from RDF between tethered charge (in IEMs) and oxygen (in water) were −0.39 and -0.35, respectively. Finally, the peak position of RDF between the tethered charge (in IEMs) with the oxygen of water and coordination number (at first minima) of RDF between counterion with the oxygen of water were found to have correlations of -0.11 and -0.13, respectively. Overall, the computed solvation descriptors correlated significantly with the desired activity coefficients.

Utilizing SHAP (SHapley Additive exPlanations) feature importance analysis, it was possible to examine the precise impact of these features including the polymer fingerprints on the prediction of the ion activity coefficient in ionomers. Some of the features may exhibit some level of relationship between one another, which in turn could possibly lead to multicollinearity if a linear regression algorithm was selected as the desired ML algorithm. Also, it is possible to eliminate multicollinearity by combining two or more collinear variables into a single variable. In this study, data dimensionality reduction and three ML methods were studied.

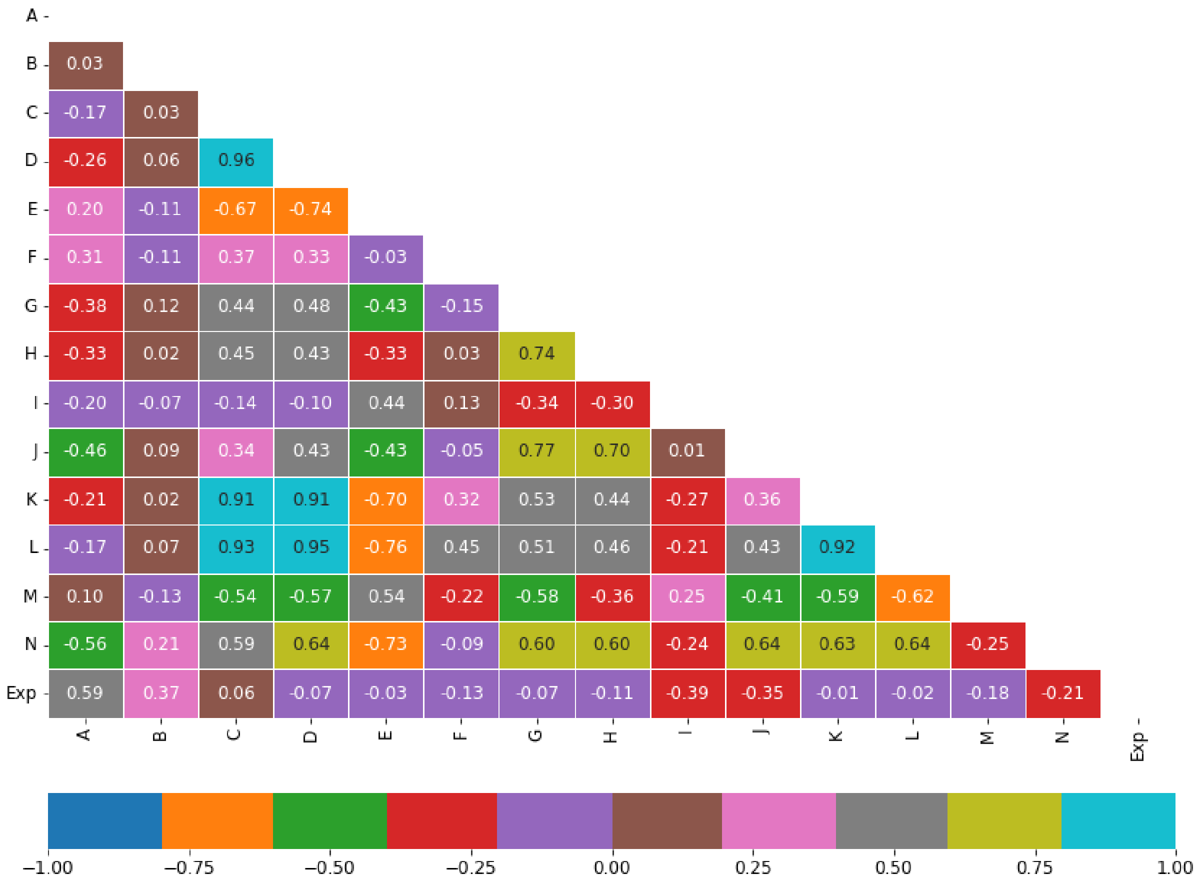

Figure 6: Heatmap showing the pairwise relationship between the solvation parameters, number of water molecules per tethered ions, salt concentration and experimental ion activity coefficient. Table 1 shows the description of the legends (A-N, Exp).

Table 1. The descriptions of the MD solvation descriptions. (RDF refers to radial distribution function).

A	*number of water molecules per tethered ions*	**I**	*peak height of RDF between tethered charge (in IEMs) and oxygen of water*
B	*concentration of salt (M)*	**J**	*coordination number (at first minima) from RDF between tethered charge (in IEMs) and oxygen (in water)*
C	*first minima of RDF between the counterion and oxygen of water.*	**K**	*first minima of RDF between tethered charge (in IEMs) and counterion (in salt)*
D	*peak position of RDF between counterion and oxygen of water*	**L**	*peak position of RDF between tethered charge (in IEMs) and counterion (in salt)*
E	*peak height of RDF between counterion and oxygen of water*	**M**	*peak height of RDF between tethered charge (in ionomer) and counterion (in salt)*
F	*coordination number (at first minima) from RDF between counterion and oxygen of water*	**N**	*coordination number (at first minima) from RDF between tethered charge (in ionomer) and counterion (in salt)*
G	*first minima of RDF between tethered charge (in IEMs) and oxygen of water*	**Exp**	*experimental ion activity coefficient*
H	*peak position of RDF between tethered charge (in IEMs) and oxygen of water*		

3.3. Performance of Machine learning models

3.3.1. Solvation descriptors

In this section, the performance of the regression algorithms was evaluated to predict the solvation descriptors gathered from MD simulations. **Figure S4** presents a scheme of the modeling and ML development approach utilized in predicting each solvation descriptor. This approach employs 137 input variable types based on the polymer fingerprints, one-hot encoded salt ions, number of water molecules per tethered ions, concentration of salt, and twelve (12) solvation descriptors obtainable from MD simulations as target variables. Specifically, the target variables include position of RDF first minima, position of 1st RDF peak, height of 1st RDF peak and coordination number from RDF. **Table 2** describes the 12 solvation descriptors considered in this study. Prior to model development, the target data was normalized to range between -1 and 1 because data normalization enhances the coherence of entry types resulting in data segmentation and greater data quality. Furthermore, PCA was applied to varying combinations of input features to help eliminate the issue of dimensionality that may arise from having several input features. Specifically, these combinations are as follows: input vector **A0** - no transformation, input vector **A1** - PCA applied only to the polymer fingerprints, and input vector **A2** - PCA applied to all input features. By applying the PCA to the features, the optimal number of components was computed based on the explained variance and the results are shown in **Figure 7**. The computed number of components that ensures a 99% of variance using the two transformed input vectors explained equals 10. Finally, the input vectors were transformed to obtain **A1** and **A2**.

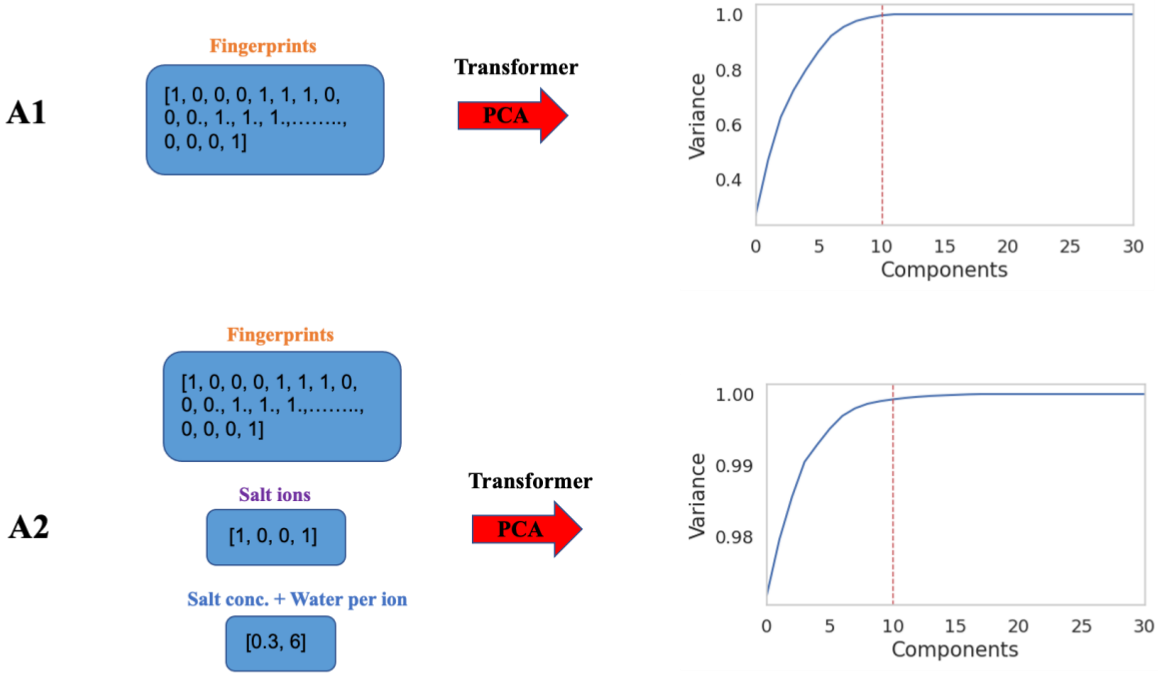

Figure 7. Explained Variance using PCA. For input vector **A1**, all entries remain intact except for Morgan Fingerprints that were transformed using PCA. In input vector **A2**, all the features namely Morgan Fingerprints, one hot encoded salt ions, salt concentrations and number of water molecules per tethered cation were transformed using PCA.

Afterwards, three ML algorithms including Support Vector Regression (SVR), Artificial Neural Network (ANN) and Random Forest Regression (RFR) were created, trained, and validated (as shown in **Figure S3**) using the three input data types (**A0**, **A1**, **A2**). The performance of the trained models was analyzed using two performance metrics (R^2 and MAE) listed in **Table S2**. The NSGA-II algorithm was utilized to optimize the model's hyperparameters on a training dataset, which consisted of 80% of the total 92 data points. The hyperparameters selected by NSGA-II were then used to train the final ML models, as presented in **Table S4**. The final model's predictive performance is illustrated in **Figure 8**, which is based on the testing dataset comprising the remaining 20% of the data. Each solvation descriptor was trained as a Multiple Input Single Output

(MISO) approach and their corresponding optimal hyper-parameters are listed in **Table S1** of the supplementary information.

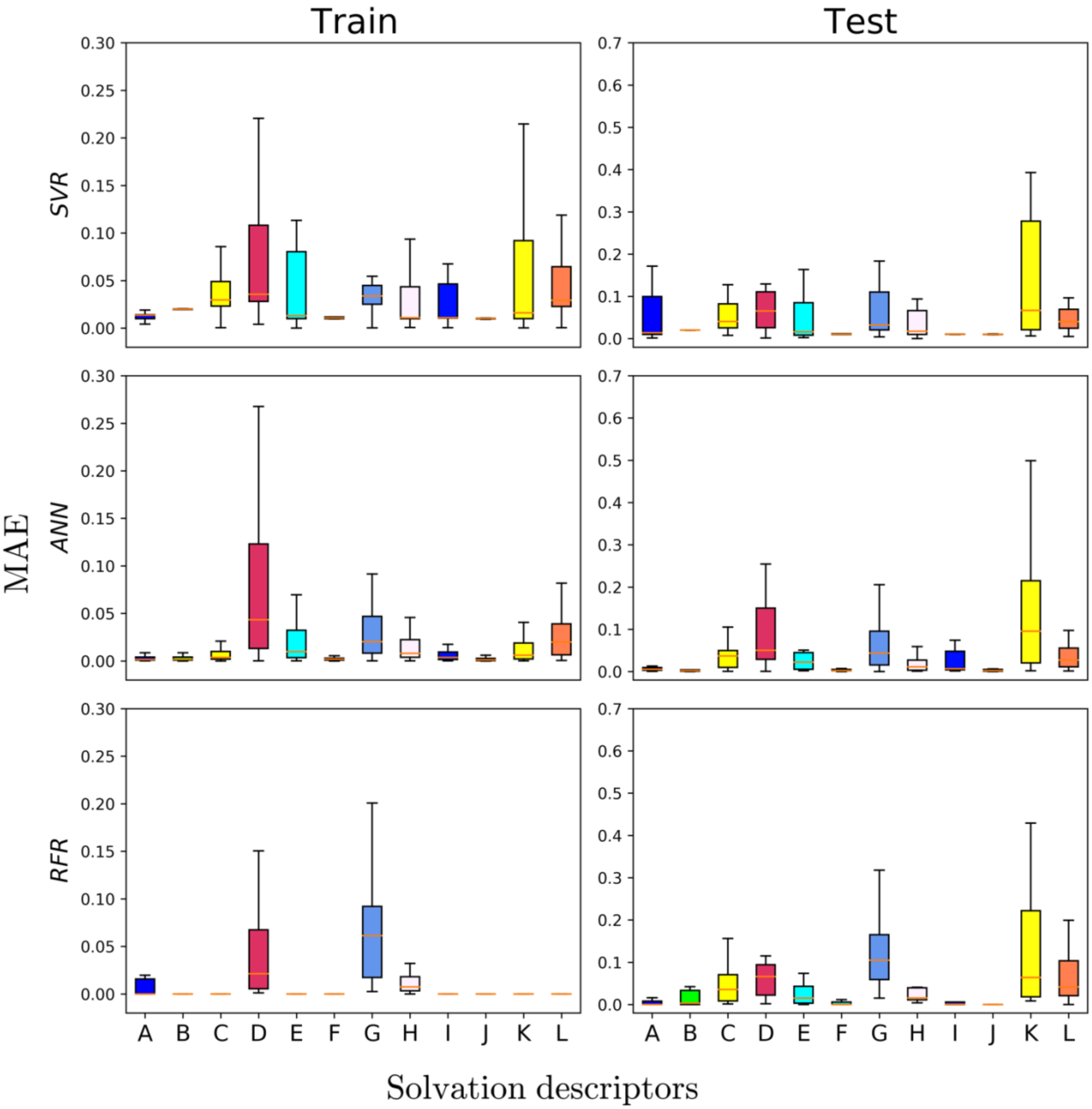

Figure 8. Performance of the ML models (SVR, ANN and RFR) in terms of mean absolute error (MAE) for predicting the solvation descriptors based on the best input matrix. The boxplot indicates the mean absolute error, MAE (mathematically shown in **Table S2** of the Supplementary file) of the predictions and the legends are coded based on **Table 2**.

A good ML model is labeled good if R^2 is close to 1 and MAE is very close to zero. Though high R^2 indicates a good model, it may be misleading. In this study, the MAE was selected as the choice parameter to assess the predictive performance of the developed model. ML models typically exhibit higher accuracy levels in predicting outcomes on the dataset used for training as compared to the testing dataset. A ML model does not suffer from underfitting or overfitting if and only if it predicts training and test data with similar accuracy. Based on the three data types (input vectors A0, A1, A2) and available ML algorithms (SVR, ANN, RFR), we formulated 9 different models for each solvation descriptor. The best set of hyperparameters for the SVR, ANN, and RFR model based on the three different input vectors are shown in **Table S4, S6** and **S8**, respectively. For simplicity, only the prediction performance of the three ML models based on the best input vector is presented in **Figure 8**. The results illustrate that the small size of the train data is enough to train the different models without underfitting or overfitting. After tuning the hyperparameters for each solvation descriptors, the MAE values for the SVR model falls averagely to 0.059, 0.071 and 0.122 for A0, A1 and A2 training data respectively, while R^2 value rose to 0.931, 0.937, and 0.763 for A0, A1 and A2 training data, respectively. Meanwhile, the average MAE values for the ANN model decreased to 0.021 for A0, 0.026 for A1, and 0.065 for A2 training data, while R^2 value rose to 0.986, 0.984, and 0.928 in the same order. Similarly, the change in MAE and R^2 values follows the same trend for the testing data. In comparison to the ANN and SVR model, the tuned RFR model based on the training data resulted in averaged R^2 values of 0.989, 0.984 and 0.844, and averaged MAE values of 0.013, 0.022 and 0.084, respectively for A0, A1 and A2.

The results showed that the dimensionality reduction via PCA does not significantly alter the performance of the different ML models when applied only to the Morgan fingerprints - i.e., input vector A1. If applied to the full input vector (i.e., input vector A2), the use of PCA significantly resulted in the loss of predictability of the selected features. Thus, it can be inferred that the number of input variables arising from the combinations of polymer fingerprints, one-hot encoded salt ions, salt concentration, and number of water molecules do not exhibit a negative multicollinearity or curse of dimensionality effects.

Furthermore, the developed ML models exhibit a low MAE and high R^2 values. **Table S3** summarized the performance of different ML models based on best input vector representation. The detailed statistical performance of the different ML models based on the three input vectors are listed in **Table S5, S7** and **S9**. The MAE values indicate that the prediction performance of the various ML models followed this trend: ANN > RFR > SVR. Overall, the ANN model outperformed the other models. The results demonstrate that the proposed method for feature generation and ML architecture optimization, with the assistance of the NSGA-II algorithm, facilitated the selection of the most appropriate hyperparameters for the selected ML models. Based on this study, it is suggested that the integration of ANN models with NSGA-II can be a valuable approach for predicting solvation descriptors without requiring MD simulations. This can significantly reduce computational costs and time, with minimal input parameters.

Table 2. The descriptions of the MD descriptions shown by the boxplot in **Figure 9**. (RDF refers to radial distribution function).

A	*first minima of RDF between the counterion and oxygen of water*	G	*peak height of RDF between tethered charge (in IEMs) and oxygen of water*
B	*peak position of RDF between counterion and oxygen of water*	H	*coordination number (at first minima) from RDF between tethered charge (in IEMs) and oxygen (in water)*
C	*peak height of RDF between counterion and oxygen of water*	I	*first minima of RDF between tethered charge (in IEMs) and counterion (in salt)*
D	*coordination number (at first minima) from RDF between counterion and oxygen of water*	J	*peak position of RDF between tethered charge (in IEMs) and counterion (in salt)*
E	*first minima of RDF between tethered charge (in IEMs) and oxygen of water*	K	*peak height of RDF between tethered charge (in IEMs) and counterion (in salt)*
F	*peak position of RDF between tethered charge (in IEMs) and oxygen of water*	L	*coordination number (at first minima) from RDF between tethered charge (in IEMs) and counterion (in salt)*

3.3.2. Ion activity coefficient in IEMs and ion-exchange thin films

In this section, the performance of the regression algorithms was evaluated to predict the experimental activity coefficients of ion exchange membranes. Initial Cheminformatics included 152 input variable types based on the polymer fingerprints (Morgan fingerprint), one-hot encoded salt ions, number of water molecules per tethered ions and concentration of salt and twelve (12) solvation descriptors obtainable from MD simulations as input variables. These solvation attributes are particularly important for understanding atomistic level properties in IEMs in ionic separations (described in **Table 2)**. Figure **S7** illustrates the ML approach used in predicting activity coefficients.

By employing a dimensionality reduction technique, we created a ML model with the fewest components possible in order to prevent overfitting and conserve computational resources. To keep both models comparable, this approach uses the same three combinations as in the solvation property prediction model. These include input vector **A0** - no transformation, **A1** - PCA applied only to the polymer fingerprints and **A2** - PCA applied to all input features in the input vector. By applying the PCA to the features, the optimal number of components was computed based on the explained variance and the results are shown in **Figure S8**. Model construction was based on the principal component analysis findings when explained variance exceeded 99%. Using the three input vectors, three ML algorithms were developed, trained, and verified (as shown in **Figure S3)**. These algorithms included Support Vector Regression (SVR), Artificial Neural Network (ANN), and Random Forest Regression (RFR). Two performance indicators (R^2, and MAE) were utilized to (shown in **Table S2**) evaluate the models' performance. The same approach as in section 3.3.1 was employed to optimize hyperparameters using 80% of the total 80 data points from the training dataset with the NSGA-II algorithm. The final ML models were then trained using the optimized hyperparameters, as presented in Table S10, and their performance was evaluated in Figure 3 through prediction on the testing dataset (the remaining 20% of the data).

We formulated 9 different models to predict activity coefficients based on the three input vectors (A0, A1, A2) and three ML algorithms mentioned above (SVR, ANN, RFR). The prediction performance of the three ML models based on the different input vectors with A0, A1 and A2 is presented in **Figures 9 and S9** respectively. Even with the small set of data, these models show better performance of predicting the activity coefficients. The MAE values for the SVR model fall on average to 0.0165, 0.0164, and 0.0185 for the A0, A1, and A2 training data, respectively, using the best set of hyperparameters for the SVR model. In contrast, the R^2 value slightly increases from 0.994 to 0.998 and 0.997 for the same training data while the opposite trend can be seen for the R^2 value for test data. The average MAE values for the ANN model are 0.0156 for A0, 0.0315 for A1, and 0.0298 for A2 training data, while R^2 value shoes best fit with 0.99. Moreover, the RFR model based on the training data produced averaged R^2 values with 0.98 and averaged MAE values of 0.0286, 0.0291, and 0.0273 for A0, A1, and A2, respectively. Overall, the performance of the testing set of data for the SVR and ANN algorithms shows better performance compared to the RFR model (see **Table 3** for more details). Furthermore, the effect of PCA on testing data compared to the training set is not that significant in this study. (See **S8** and **S9** for more details) Similar observations can be seen in the solvation descriptors prediction model in section 3.3.1.

Table 3: Performance of the activity models based on the input vector type in predicting the activity coefficients. (Input vector **A0** - no transformation, input vector **A1** - PCA applied only to the polymer fingerprints, and input vector A**2** - PCA applied to all input features.)

Metric	Data	Model	A0	A1	A2
MAE	Train	SVR	0.0165	0.0164	0.0185
		ANN	0.0156	0.0315	0.0298
		RFR	0.0286	0.0291	0.0273
	Test	SVR	0.0535	0.0626	0.0730
		ANN	0.0412	0.0677	0.0714
		RFR	0.0783	0.1067	0.1141
R^2	Train	SVR	0.9941	0.9980	0.9973
		ANN	0.9966	0.9901	0.9916
		RFR	0.9884	0.9898	0.9885
	Test	SVR	0.9776	0.8724	0.8427
		ANN	0.9880	0.8406	0.8147
		RFR	0.9508	0.6711	0.6114

Understanding the reasoning behind a model's prediction can be equally crucial as determining its accuracy. By identifying the class of addictive feature importance methods, the SHAP approach is utilized to describe how ML models produce their output in a test data set. SHAP has a unified approach to assign each feature an importance value for a particular prediction. The bar plot of the SHAP values displays a global feature importance plot, where the global relevance of each feature is determined as the mean absolute value of that feature across all provided samples. **Figure 9** summarizes the results of SHAP global feature importance plot with top 20 features based on their average SHAP values with different ML models importance with no transformation of initial features (A0). A summary of the definitions of these descriptors that are used in **Figure 9** is provided in **Table 1.** In general, from top 20 features, higher contribution can be seen from Morgan fingerprint bits followed by solvation descriptors in all the three ML models. Despite some solvation descriptors not showing much contribution to existing models, salt concentrations and RDF characteristics between charge bearing groups/counter ion and waters are significant contributors to all three models. A correlation between above solvation descriptors and activity coefficient values can be further verified based on spearman's correlation matrix. Using these frameworks, the materials properties can then be correlated to device settings for cost-effective operation and for the design of effective materials. Microscopical properties and macroscopic properties can be mapped easily into the manufacturing process of such materials and devices using these ML frameworks.

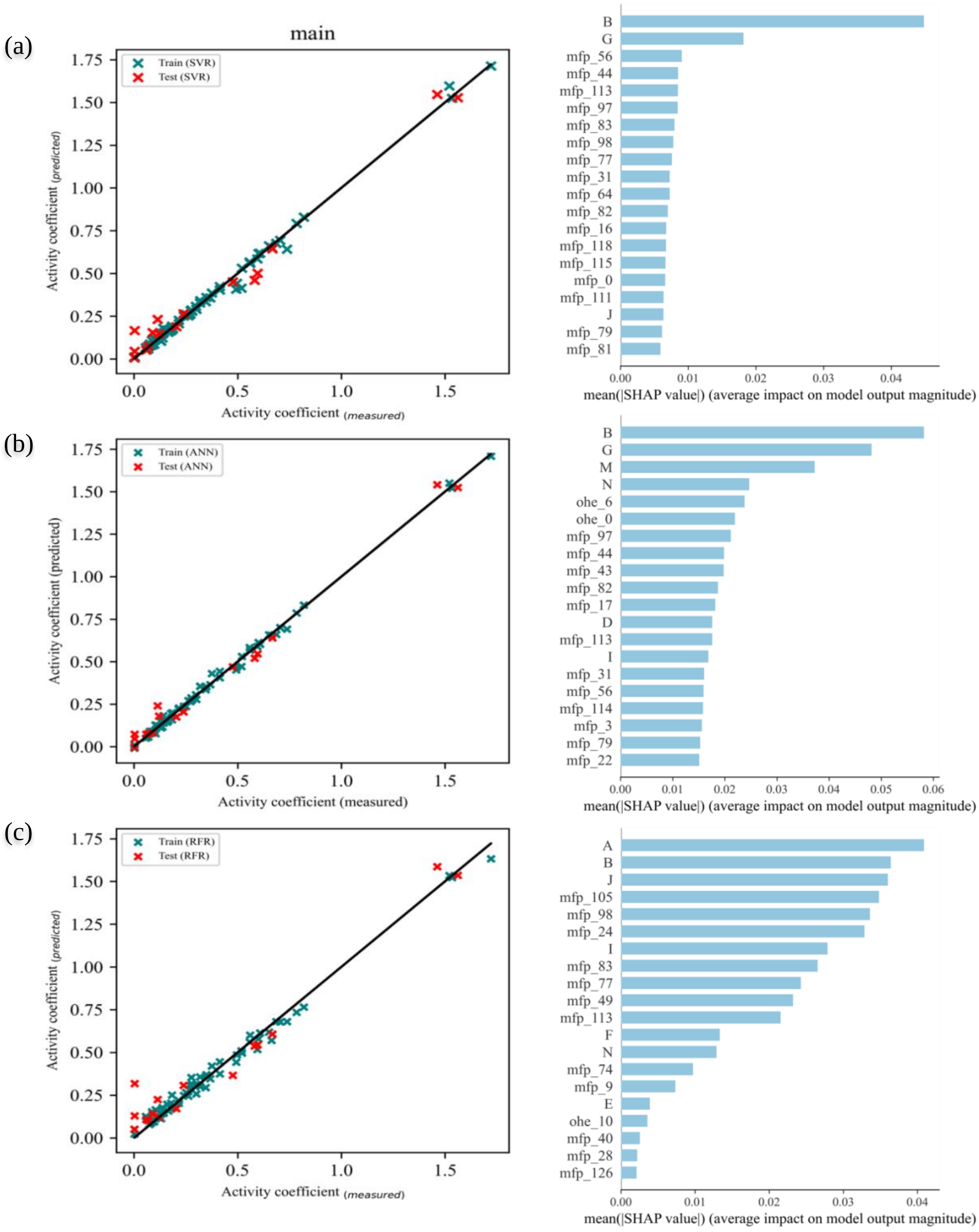

Figure 9. Activity coefficient predictions versus measured values; Activity coefficients as a function of input vector (left) and corresponding feature importance (SHAP) values (right) for **A0** - no transformation, (a) SVR (b) ANN (c) RFR. Here, Mfp_'n' (n = 0 - 127) and Ohe_'n' (n = 0

- 9) represents the Morgan fingerprint bits of the IEM and One hot encoding of the salt ions, respectively. **Table 1** shows the description of the legends (A-N)

4. Conclusions

In this work, an integrated framework was introduced by leveraging the advantages of molecular dynamics and ML algorithms to map the interrelationships between molecular descriptors and the ion activity coefficients in ionomer media (i.e., IEMs and ion-exchange thin films). The study established a two-step link between the molecular fingerprints of the co-polymers, the solvation properties obtained from molecular dynamics (MD) and their corresponding ion activity coefficients parameters. This approach enabled the acquisition of activity coefficients for existing and emerging ion-exchange materials even in the absence of solvation data or experimental data, which effectively reduced the complexity associated with computing the solvation properties from MD and the measuring of activity coefficients from experiments.

To construct the framework, different molecular representation approaches and ML algorithms were assessed to predict the solvation properties and activity coefficient of the ion-exchange materials. The performance of the trained models was verified with the aid of various statistical parameters. Overall, the ANN model showed the best predictive ability compared to other ML algorithms in solvation prediction model. Specifically, the results demonstrated the efficacy of the ANN-predicted parameters in modeling desired properties based on selected features with minimal errors compared to available MD/experimental data, with errors less than 7% for solvation properties. Overall, all the three ML models accurately predict the ion activity coefficient with an average mean absolute error of <10%. SVR and ANN shows slightly better performance compared to RFR. Moreover, the trained solvation parameters model and activity coefficient models proved to be capable of characterizing the ionic activity coefficient with acceptable deviations. The sparse available dataset can give rise to discrepancies; however, these can be improved by using larger datasets. This, in turn, necessitates the creation of a library of polymeric IEMs for a variety of electrochemical technologies that have IEMs interfaced with liquid solutions with dissolved ions. In addition, the use of newly developed force fields (though computationally expensive) for the MD simulations can help improve the accuracy of the proposed framework.

Even in situations where there is insufficient experimental data available, the results of this work demonstrate the potential of the integrated approach for predicting the activity coefficient of IEMs, facilitating the search for new materials with accepted behavior that meet both technical and industrial requirements for successful electrochemical separations.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, under the Office of Basic Energy Sciences Separation Science program under Award No. DE-SC0022304. We gratefully acknowledge the computer time allotted by the high-performance computing center at LSU and the Louisiana Optical Network Initiative. Luis A. Briceno-Mena thanks the support received from Universidad de Costa Rica.

References

(1) Jordan, M. L.; Kulkarni, T.; Senadheera, D. I.; Kumar, R.; Lin, Y. J.; Arges, C. G. Imidazolium-Type Anion Exchange Membranes for Improved Organic Acid Transport and Permselectivity in Electrodialysis. *J. Electrochem. Soc.* **2022**, *169* (4), 043511. https://doi.org/10.1149/1945-7111/ac6448.

(2) Jordan, M. L.; Kokoszka, G.; Gallage Dona, H. K.; Senadheera, D. I.; Kumar, R.; Lin, Y. J.; Arges, C. G. Integrated Ion-Exchange Membrane Resin Wafer Assemblies for Aromatic Organic Acid Separations Using Electrodeionization. *ACS Sustain. Chem. Eng.* **2023**, *11* (3), 945–956. https://doi.org/10.1021/acssuschemeng.2c05255.

(3) Feng, Y.; Yang, L.; Liu, J.; Logan, B. E. Electrochemical Technologies for Wastewater Treatment and Resource Reclamation. *Environ. Sci. Water Res. Technol.* **2016**, *2* (5), 800–831. https://doi.org/10.1039/C5EW00289C.

(4) Kim, K.; Baldaguez Medina, P.; Elbert, J.; Kayiwa, E.; Cusick, R. D.; Men, Y.; Su, X. Molecular Tuning of Redox-Copolymers for Selective Electrochemical Remediation. *Adv. Funct. Mater.* **2020**, *30* (52), 2004635. https://doi.org/10.1002/adfm.202004635.

(5) Román Santiago, A.; Baldaguez Medina, P.; Su, X. Electrochemical Remediation of Perfluoroalkyl Substances from Water. *Electrochimica Acta* **2022**, *403*, 139635. https://doi.org/10.1016/j.electacta.2021.139635.

(6) Alkhadra, M. A.; Jordan, M. L.; Tian, H.; Arges, C. G.; Bazant, M. Z. Selective and Chemical-Free Removal of Toxic Heavy Metal Cations from Water Using Shock Ion Extraction. *Environ. Sci. Technol.* **2022**, *56* (19), 14091–14098. https://doi.org/10.1021/acs.est.2c05042.

(7) Tang, C.; Bruening, M. L. Ion Separations with Membranes. *J. Polym. Sci.* **2020**, *58* (20), 2831–2856. https://doi.org/10.1002/pol.20200500.

(8) Briceno-Mena, L. A.; Romagnoli, J. A.; Arges, C. G. PemNet: A Transfer Learning-Based Modeling Approach of High-Temperature Polymer Electrolyte Membrane Electrochemical Systems. *Ind. Eng. Chem. Res.* **2022**, *61* (9), 3350–3357. https://doi.org/10.1021/acs.iecr.1c04237.

(9) Wilberforce, T.; Rezk, H.; Olabi, A. G.; Epelle, E. I.; Abdelkareem, M. A. Comparative Analysis on Parametric Estimation of a PEM Fuel Cell Using Metaheuristics Algorithms. *Energy* **2023**, *262*, 125530. https://doi.org/10.1016/j.energy.2022.125530.

(10) Palakkal, V. M.; Rubio, J. E.; Lin, Y. J.; Arges, C. G. Low-Resistant Ion-Exchange Membranes for Energy Efficient Membrane Capacitive Deionization. *ACS Sustain. Chem. Eng.* **2018**, *6* (11), 13778–13786. https://doi.org/10.1021/acssuschemeng.8b01797.

(11) Strathmann, H.; Grabowski, A.; Eigenberger, G. Ion-Exchange Membranes in the Chemical Process Industry. *Ind. Eng. Chem. Res.* **2013**, *52* (31), 10364–10379. https://doi.org/10.1021/ie4002102.

(12) Li, N.; Guiver, M. D. Ion Transport by Nanochannels in Ion-Containing Aromatic Copolymers. *Macromolecules* **2014**, *47* (7), 2175–2198. https://doi.org/10.1021/ma402254h.

(13) Sata, T. *Ion Exchange Membranes*; 2004. https://doi.org/10.1039/9781847551177.

(14) V. Ramos-Garcés, M.; Li, K.; Lei, Q.; Bhattacharya, D.; Kole, S.; Zhang, Q.; Strzalka, J.; P. Angelopoulou, P.; Sakellariou, G.; Kumar, R.; G. Arges, C. Understanding the Ionic Activity and Conductivity Value Differences between Random Copolymer Electrolytes and Block Copolymer Electrolytes of the Same Chemistry. *RSC Adv.* **2021**, *11* (25), 15078–15084. https://doi.org/10.1039/D1RA02519H.

(15) Sujanani, R.; Katz, L. E.; Paul, D. R.; Freeman, B. D. Aqueous Ion Partitioning in Nafion: Applicability of Manning's Counter-Ion Condensation Theory. *J. Membr. Sci.* **2021**, *638*, 119687. https://doi.org/10.1016/j.memsci.2021.119687.

(16) Jang, E.-S.; Kamcev, J.; Kobayashi, K.; Yan, N.; Sujanani, R.; Talley, S. J.; Moore, R. B.; Paul, D. R.; Freeman, B. D. Effect of Water Content on Sodium Chloride Sorption in Cross-Linked Cation Exchange Membranes. *Macromolecules* **2019**, *52* (6), 2569–2579. https://doi.org/10.1021/acs.macromol.8b02550.

(17) Lei, Q.; Li, K.; Bhattacharya, D.; Xiao, J.; Kole, S.; Zhang, Q.; Strzalka, J.; Lawrence, J.; Kumar, R.; Arges, C. G. Counterion Condensation or Lack of Solvation? Understanding the Activity of Ions in Thin Film Block Copolymer Electrolytes. *J. Mater. Chem. A* **2020**, *8* (31), 15962–15975. https://doi.org/10.1039/D0TA04266H.

(18) Galizia, M.; Paul, D. R.; Freeman, B. D. Co-Ion Specific Effect on Sodium Halides Sorption and Transport in a Cross-Linked Poly(p-Styrene Sulfonate-Co-Divinylbenzene) for Membrane Applications. *J. Membr. Sci.* **2020**, *612*, 118410. https://doi.org/10.1016/j.memsci.2020.118410.

(19) Briceno-Mena, L. A.; Venugopalan, G.; Arges, C. C.; Romagnoli, J. A. A Machine Learning Approach for Device Design from Materials and Operation Data. In *Computer Aided Chemical Engineering*; Computer Aided Chemical Engineering; Elsevier B.V., 2021; pp 279–285. https://doi.org/10.1016/B978-0-323-88506-5.50045-0.

(20) Pablo, J. J. de; Schieber, J. D. *Molecular Engineering Thermodynamics*; Cambridge University Press, 2014.

(21) Kamcev, J.; Paul, D. R.; Freeman, B. D. Ion Activity Coefficients in Ion Exchange Polymers: Applicability of Manning's Counterion Condensation Theory. *Macromolecules* **2015**, *48* (21), 8011–8024. https://doi.org/10.1021/acs.macromol.5b01654.

(22) Luo, T.; Roghmans, F.; Wessling, M. Ion Mobility and Partition Determine the Counter-Ion Selectivity of Ion Exchange Membranes. *J. Membr. Sci.* **2020**, *597*, 117645. https://doi.org/10.1016/j.memsci.2019.117645.

(23) Tao, L.; Varshney, V.; Li, Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. *J. Chem. Inf. Model.* **2021**, *61* (11), 5395–5413. https://doi.org/10.1021/acs.jcim.1c01031.

(24) Zhang, Y.; Xu, X. Machine Learning Glass Transition Temperature of Styrenic Random Copolymers. *J. Mol. Graph. Model.* **2021**, *103*, 107796. https://doi.org/10.1016/j.jmgm.2020.107796.

(25) Yuan, Q.; Longo, M.; Thornton, A. W.; McKeown, N. B.; Comesaña-Gándara, B.; Jansen, J. C.; Jelfs, K. E. Imputation of Missing Gas Permeability Data for Polymer Membranes Using Machine Learning. *J. Membr. Sci.* **2021**, *627*, 119207. https://doi.org/10.1016/j.memsci.2021.119207.

(26) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing Exceptional Gas-Separation Polymer Membranes Using Machine Learning. *Sci. Adv.* **2020**, *6* (20), eaaz4301. https://doi.org/10.1126/sciadv.aaz4301.

(27) Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine Learning Enables Interpretable Discovery of Innovative Polymers for Gas Separation Membranes. *Sci. Adv.* **2022**, *8* (29), eabn9545. https://doi.org/10.1126/sciadv.abn9545.

(28) Ethier, J. G.; Casukhela, R. K.; Latimer, J. J.; Jacobsen, M. D.; Rasin, B.; Gupta, M. K.; Baldwin, L. A.; Vaia, R. A. Predicting Phase Behavior of Linear Polymers in Solution Using Machine Learning. *Macromolecules* **2022**, *55* (7), 2691–2702. https://doi.org/10.1021/acs.macromol.2c00245.

(29) Loo, W. S.; Fang, C.; Balsara, N. P.; Wang, R. Uncovering Local Correlations in Polymer Electrolytes by X-Ray Scattering and Molecular Dynamics Simulations. *Macromolecules* **2021**, *54* (14), 6639–6648. https://doi.org/10.1021/acs.macromol.1c00995.

(30) Tsai, E.; Gallage Dona, H. K.; Tong, X.; Du, P.; Novak, B.; David, R.; Rick, S. W.; Zhang, D.; Kumar, R. Unraveling the Role of Charge Patterning in the Micellar Structure of Sequence-Defined Amphiphilic Peptoid Oligomers by Molecular Dynamics Simulations. *Macromolecules* **2022**, *55* (12), 5197–5212. https://doi.org/10.1021/acs.macromol.2c00141.

(31) Wu, C.-W.; Ren, X.; Zhou, W.-X.; Xie, G.; Zhang, G. Thermal Stability and Thermal Conductivity of Solid Electrolytes. *APL Mater.* **2022**, *10* (4), 040902. https://doi.org/10.1063/5.0089891.

(32) Wei, X.; Ma, R.; Luo, T. Thermal Conductivity of Polyelectrolytes with Different Counterions. *J. Phys. Chem. C* **2020**, *124* (8), 4483–4488. https://doi.org/10.1021/acs.jpcc.9b11689.

(33) Fong, K. D.; Self, J.; Diederichsen, K. M.; Wood, B. M.; McCloskey, B. D.; Persson, K. A. Ion Transport and the True Transference Number in Nonaqueous Polyelectrolyte Solutions for Lithium Ion Batteries. *ACS Cent. Sci.* **2019**, *5* (7), 1250–1260. https://doi.org/10.1021/acscentsci.9b00406.

(34) Galizia, M.; Benedetti, F. M.; Paul, D. R.; Freeman, B. D. Monovalent and Divalent Ion Sorption in a Cation Exchange Membrane Based on Cross-Linked Poly (p-Styrene Sulfonate-Co-Divinylbenzene). *J. Membr. Sci.* **2017**, *535*, 132–142. https://doi.org/10.1016/j.memsci.2017.04.007.

(35) Nagasawa, M.; Izumi, M.; Kagawa, I. Colligative Properties of Polyelectrolyte Solutions. V. Activity Coefficients of Counter- and by-Ions. *J. Polym. Sci.* **1959**, *37* (132), 375–383. https://doi.org/10.1002/pol.1959.1203713208.

(36) Yan, N.; Sujanani, R.; Kamcev, J.; Galizia, M.; Jang, E.-S.; Paul, D. R.; Freeman, B. D. Influence of Fixed Charge Concentration and Water Uptake on Ion Sorption in AMPS/PEGDA Membranes. *J. Membr. Sci.* **2022**, *644*, 120171. https://doi.org/10.1016/j.memsci.2021.120171.

(37) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in ’t Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a Flexible Simulation Tool for Particle-Based Materials Modeling at the Atomic, Meso, and Continuum Scales. *Comput. Phys. Commun.* **2022**, *271*, 108171. https://doi.org/10.1016/j.cpc.2021.108171.

(38) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666. https://doi.org/10.1021/ja00214a001.

(39) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236. https://doi.org/10.1021/ja9621760.

(40) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(41) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(42) Frisch, M.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, Ga. Gaussian 09, Revision D. 01, 2009.

(43) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247–260. https://doi.org/10.1016/j.jmgm.2005.12.005.

(44) Liu, L.; Kohl, P. A. Anion Conducting Multiblock Copolymers with Different Tethered Cations. *J. Polym. Sci. Part Polym. Chem.* **2018**, *56* (13), 1395–1403. https://doi.org/10.1002/pola.29020.

(45) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(46) Landrum, G.; Tosco, P.; Kelley, B.; Ric; sriniker; gedeck; Vianello, R.; NadineSchneider; Cosgrove, D.; Kawashima, E.; Dalke, A.; N, D.; Jones, G.; Cole, B.; Swain, M.; Turk, S.; AlexanderSavelyev; Vaucher, A.; Wójcikowski, M.; Take, I.; Probst, D.; Ujihara, K.; Scalfani, V. F.; godin, guillaume; Pahl, A.; Berenger, F.; JLVarjo; strets123; JP; DoliathGavid. RDKit: Open-Source Cheminformatics., 2022. https://doi.org/10.5281/zenodo.7235579.

(47) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.

(48) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*; COLT '92; Association for Computing Machinery: New York, NY, USA, 1992; pp 144–152. https://doi.org/10.1145/130385.130401.

(49) Drucker, H.; Burges, C. J. C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*; NIPS'96; MIT Press: Cambridge, MA, USA, 1996; pp 155–161.

(50) Tan, P.-N.; Steinbach, M.; Karpatne, A.; Kumar, V. Introduction to Data Mining (2nd Edition).

(51) Blank, J.; Deb, K. Pymoo: Multi-Objective Optimization in Python. *IEEE Access*, 2020, *8*, 89497–89509. https://doi.org/10.1109/ACCESS.2020.2990567.

(52) Khan Mashwani, W.; Salhi, A.; Yeniay, O.; Hussian, H.; Jan, M. A. Hybrid Non-Dominated Sorting Genetic Algorithm with Adaptive Operators Selection. *Appl. Soft Comput.* **2017**, *56*, 1–18. https://doi.org/10.1016/j.asoc.2017.01.056.

(53) Wang, Y.; Shen, Y.; Zhang, X.; Cui, G.; Sun, J. An Improved Non-Dominated Sorting Genetic Algorithm-II (INSGA-II) Applied to the Design of DNA Codewords. *Math. Comput. Simul.* **2018**, *151*, 131–139. https://doi.org/10.1016/j.matcom.2018.03.011.

(54) Vishwakarma, G.; Haghighatlari, M.; Hachmann, J. Towards Autonomous Machine Learning in Chemistry via Evolutionary Algorithms. **2019**. https://doi.org/10.26434/chemrxiv.9782387.v1.

(55) Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 2002, *6*, 182–197. https://doi.org/10.1109/4235.996017.

(56) *Chapter 12 - Principal component analysis | Elsevier Enhanced Reader*. https://doi.org/10.1016/B978-0-12-815739-8.00012-2.

(57) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 163–166. https://doi.org/10.1021/ci990316u.

(58) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1151–1157. https://doi.org/10.1021/ci030285+.

(59) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminformatics* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.