Molecular fingerprint-aided prediction of organic solute membrane rejection in reverse osmosis and nanofiltration

Journal of Membrane Science

Sangsuk Lee[1], Michael R. Shirts[2,3], and Anthony P. Straub[1,3*]

[1]*Department of Civil, Environmental and Architectural Engineering, University of Colorado Boulder, Boulder, Colorado 80309*

[2]*Department of Chemical and Biological Engineering, University of Colorado Boulder, Boulder, Colorado 80309*

[3]*Materials Science and Engineering Program, University of Colorado Boulder, Boulder, Colorado 80309*

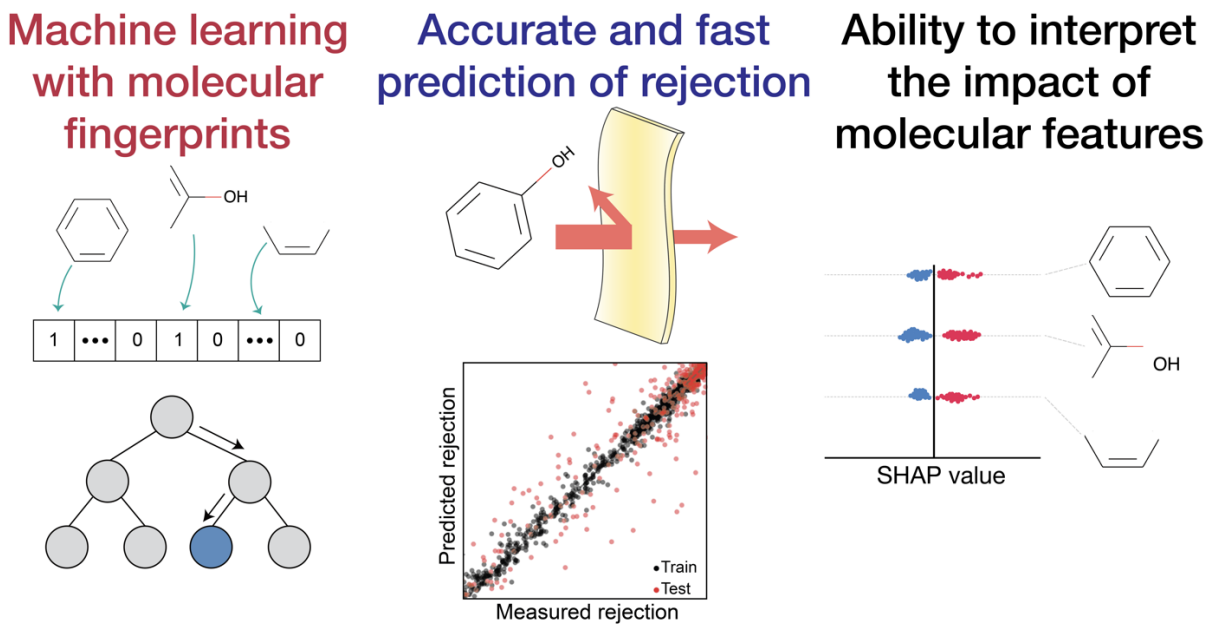
*Corresponding author: Anthony Straub, Email: anthony.straub@colorado.edu

ABSTRACT

Reverse osmosis and nanofiltration are used to purify feedwaters that contain a range of harmful organic solutes. The rejection of many of these solutes is poorly understood due to our limited ability to experimentally measure removal of any given compound. In this work, we present a machine learning approach that predicts organic solute rejection using molecular fingerprints that encode chemical structure features, such as functional groups and rings, into simple binary vectors. We trained machine learning models on a database of 1906 membrane rejection measurements including 228 organic compounds and 39 types of reverse osmosis and nanofiltration membranes. Three types of molecular fingerprint models (structural key, circular, and path based) were compared, and we observed that the Molecular Access System (MACCS) structural key had high performance (coefficient of determination of 0.87 with the testing set), fast calculation time due to its short bit-length, and easy interpretability. In addition to evaluating prediction performance, Shapley Additive Explanations (SHAP) analysis was implemented to gain a better molecular-scale understanding of membrane rejection, identifying molecular substructures that are important in determining their rejection. Overall, this work presents a method to predict the rejection of compounds that uses readily available molecular structure information and improves our ability to understand rejection mechanisms.

Keywords: Machine learning, molecular fingerprints, reverse osmosis, nanofiltration, solute rejection

Graphical Abstract



Highlights

- Machine learning models for solute rejection used molecular fingerprints as input.
- The models showed comparable performance to traditional models.
- MACCS was found to be the best fingerprint for membrane rejection applications.
- SHAP and clustering help interpret how chemical features affect rejection.

1. Introduction

3.6 billion people experience water scarcity at least one month per year, and water shortages are becoming more severe because of climate change and increasing water demands [1]. Processes that utilize unconventional water resources such as desalination, brine treatment, and water reuse have the potential to sustainably alleviate water stress by augmenting clean water supplies beyond those available from existing water resources. Reverse osmosis (RO) and nanofiltration (NF) are widely considered in advanced water treatment trains because of their ability to remove a broad spectrum of compounds, including salts and harmful organic contaminants, more reliably than other processes [2]. In addition to water treatment, RO and NF are increasingly considered in a broad suite of separations including drug purification, protein concentration, and food processing [3–6].

Emerging water treatment applications, such as municipal and industrial wastewater reuse, include myriad organic compounds in the feed streams. Some of these compounds, such as 1,4-dioxane and *N*-nitrosodimethylamine (NDMA), are of particular concern because they are known to be harmful to human health and are well-documented as being poorly removed by membrane processes [7]. However, many other potentially harmful compounds exist in feedwaters with poorly understood rejection in NF and RO [8]. It is thus of critical importance to develop the ability to determine the removal of a broad range of compounds in membrane systems. To date, much of our understanding of membrane rejection has relied on experimental measurements of rejection for individual compounds in studies that generally examined between 5 and 30 compounds [9–11]. These individual experimental measurements provide valuable insights, but alternative methods are necessary to approximate rejection of the multitude of compounds in feedwaters and develop a greater understanding of rejection mechanisms [12,13].

Machine Learning (ML) models are powerful tools to address complex real-world problems in many science and engineering fields related to material development, chemical processing, biomedical studies, and environmental science [14]. ML models are constructed from data, enabling complicated non-linear relationships to be captured in predictive algorithms. In the field of membrane separations, ML models have been used to predict rejection of emerging compounds based on collected rejection data, membrane properties, and compound properties [15–19]. These ML methods have been valuable for the selection of a suitable membrane for a given

application and been used to relate operating conditions and membrane properties to process performance [20,21].

Despite their growing utility in membrane science, current ML algorithms used to predict rejection are unable to predict the removal of new compounds without known molecular properties and face challenges in describing the mechanisms by which compound molar structure influences rejection. Previous ML models for membrane rejection have used input parameters based on compound descriptions such as molecular weight, dipole moment, and octanol-water partition coefficient [14,22,23]. While these descriptors can effectively quantify the physicochemical properties of a compound, they lack the ability to intuitively express structural information and atom connectivity [24,25]. Recent rejection studies have found that the interactions between functional groups and aromatic rings on compounds and polyamide membranes are strongly related to rejection [26–28], and considering the molecular structure is therefore critical to understand how functional groups and their positions within the molecule influence membrane rejection [29]. Moreover, phenomena such as proton dissociation, which enhances rejection through strong electrostatic repulsion, are closely related to molecular structures [30,31]. To gain molecular-level mechanistic insights into membrane rejection, models that include details on molecular structure are therefore needed [14,32]. Furthermore, some experimental molecular descriptors, such as acid dissociation constants and octanol-water partition coefficients, are not readily available for certain compounds, meaning it is not possible to predict the rejection of new and poorly studied compounds using existing ML methods.

Converting organic compound information into detailed features that can be understood by computers is essential to use machine learning to predict membrane rejection and for other chemical applications [33] . Molecular fingerprints (MFs) are a promising tool to present detailed molecular structure information. MFs transform chemical structural features into binary vectors (0's and 1's) that account for the absence or presence of molecular substructures [34,35]. Recently, multiple studies successfully built data-driven models using MFs as input data to predict molecular properties including the refractive index, viscosity, acid dissociation constant, and reaction rate of organic compounds with hydroxyl radical [35–38]. A key advantage of MFs is that they can encode information using Simplified Molecular Input Line Entry System (SMILES) codes, which are available for any compound with a known chemical structure. In addition, MFs can be combined with other analysis techniques such as Clustering and Shapley Additive Explanations (SHAP),

which are evaluation tools to find similarity and disparity of data and explain the input and output relationships [39]. For example, the binary vector information from MF can enhance the pattern recognition power of clustering by coupling molecular sub-structure information and rejection behavior. In addition, ML models do not provide clear relations between input and output, and thus, explainable machine learning techniques such as SHAP can help to unveil the decision-making process of ML models. In studying RO and NF membrane rejection, the combination of MF data and advanced analysis techniques has potential to elucidate the interplay of different molecular fragments and membranes to reveal complex RO/NF membrane rejection mechanisms [15–19]. However, no studies to date have used MFs generated from organic solutes to predict these solutes' rejection in RO and NF water treatment processes.

In this study, we compare three different molecular fingerprint categories (path-based, circular, and structural key) to develop ML models for predicting contaminant rejection and examining the potential of MFs in explaining underlying rejection mechanisms. The models are trained with the different fingerprints and 1906 rejection measurements for 228 unique organic compounds. The prediction performances of the trained models are quantified through evaluation metrics such as the coefficient of determination and the Spearman coefficient. We change the parameters for fingerprints (e.g. maximum path length, maximum radius, structural key number) to explore the optimal fingerprints in terms of prediction power, calculation expense, and interpretability. Molecular fragments created by each fingerprint and the analyses of clustering and SHAP are used to gain an understanding of molecular-scale compound-membrane interactions. Our analysis provides a valuable tool for rejection prediction and yields insights into compound properties that dictate rejection.

2. Materials and Methods

2.1. Dataset shape

Membrane properties, compound properties, SMILES notation for each compound, and rejection values were obtained from the previous literature [14,40]. The dataset included 1906 experimental rejection data points for 228 organic compounds and 39 types of RO and NF membranes. The distribution of membrane materials is as follows: there are 1,317 samples of polyamide membranes, 501 samples of poly(piperazinamide) membranes, 69 samples of polyethersulfone membranes, and

159 19 samples of cellulose acetate membranes. The retrieved SMILES notations for the organic
160 compounds were converted into three classes of MFs: path-based, circular, and structural key.
161 RDKit and Morgan fingerprints were chosen as the path-based and circular fingerprints,
162 respectively. Molecular Access System (MACCS) and PubChem fingerprints were chosen as
163 structural key fingerprints.

164 Table 1 summarizes the information available in the input datasets excluding the MFs. The
165 datasets used for model training contained three membrane properties (membrane molecular
166 weight cut-off (MWCO), membrane water contact angle, and solution pH), three operating
167 conditions (pressure, measurement time, and initial concentration), and MFs that represent the
168 molecular structural and chemical properties. We note that insufficient data were available for
169 membrane surface charge in the original dataset, so solution pH was used as an indirect feature of
170 membrane surface charge because solution pH governs the extent of protonation on the membrane
171 surface [41]. The distribution of output (rejection, R) demonstrated a severely left-skewed shape
172 since rejection values tended to be high, with most rejection values greater than 90%. In this study,
173 the original R distribution, -log(1-R) transformation, and -sqrt(1-R) transformation were examined
174 to address skewing of the data and allow for improved fitting resolution; the -sqrt(1-R)
175 transformation was finally applied on the output because the sqrt(1-R) transformation best shifted
176 the left-skewed distribution closer to a normal distribution. All input and output values were scaled
177 from 0 to 1 to render the range of values equal for training, and then the outputs were transformed
178 back to the original scale when evaluating model performance. The raw dataset and processed
179 dataset are available in a GitHub repository (https://github.com/leesang-boulder/machine-
180 learning_molecular-fingerprint_RO-NF-rejection), and the detailed descriptions about MFs can be
181 found in the next section.

182 **Table 1.** Overall distribution of the input data from previous literature including 1906 data
183 points, 228 organic compounds, and 39 types of RO and NF membranes.

	pH	Membrane MWCO (Da)	Membrane water contact angle (°)	Pressure (kPa)	Measurement time (min)	Initial concentration of compound (mg/L)	Rejection (%)
25th Percentile	7.0	100	41.4	500	10	0.053	72.06
50th Percentile	7.0	152	53.8	690	10	0.5	90.82
75th Percentile	7.0	300	63.2	1000	300	10.0	96.90

Min value	2.2	65	14.4	240	10	0.00072	0
Max value	11	460	79.4	3500	5760	2000	100

2.2. Molecular fingerprints generation

Molecular fingerprints (MFs) encode the structural information of chemical compounds into binary vectors. Each binary value (0 or 1) indicates whether a certain molecular fragment exists in a compound and the position of the binary vector displays which molecular fragment it is. This molecular representation enables the computationally efficient management and comparison of chemical structures (e.g. Tanimoto and Dice similarity coefficients) [42]. Fingerprinting has played a key role in virtual screening, quantitative structure-property relationships (QSPR) analysis, similarity-based compound search, target molecule ranking, and other chemical compound discovery processes [43]. Moreover, previous studies proved that fingerprinting is useful for ML models to predict variables that are highly dependent on molecular structures [44–48].

There are five main categories of 2D fingerprints, namely structural key fingerprints, topological or path-based fingerprints, circular fingerprints, pharmacophore, and neural fingerprints. Three categories of fingerprints were applied in this study to convert SMILES into binary vectors: path-based (RDKit), circular (Morgan), and structural key (PubChem and Molecular Access System, MACCS). The generating algorithm of each fingerprint is briefly displayed in Figure 1A. Path-based fingerprints encode all possible connectivity for the fragments of a compound following a linear path along the molecular graph from a central atom up to a given maximum length. Under a path-based approach, any compound can produce a meaningful fingerprint; however, bit collision can be an issue because a bit can be set by multiple different fragments [49]. RDKit was used as the path-based fingerprint, and maximum path lengths of 1 and 3 and bit lengths of 1024 and 32768 were implemented. Although it increases computational complexity, an expanded bit length of 32768 was used for the maximum path length of 3 to avoid bit collision; a detailed explanation of bit collision counts can be found in Section 2 of the Supporting Information (Figure S4). The circular fingerprint utilizes a similar approach to the path-based fingerprint but constructs fragments within a radius of the starting atom instead of on linear paths. The circular Morgan fingerprint was used with a maximum radius of 1 and 3 (which denotes the radial distance from the center atom), and bit lengths of 4096 and 16384 were chosen.

Structural key fingerprints define the bits depending on the presence and absence in the compound of certain fragments from a given list of structural keys. Therefore, these fingerprints are meaningful when the molecules' fragments are well-represented by the pre-defined keys [49]. PubChem and MACCS were used as structural key fingerprints. PubChem fingerprints are based on 881 structural keys defined by the PubChem database system. MACCS uses 166 keys developed by the Molecular Design Limited (MDL) Information Systems. During data generation, one dummy bit was padded to the head of the MACCS keys, which resulted in 167 keys in total.

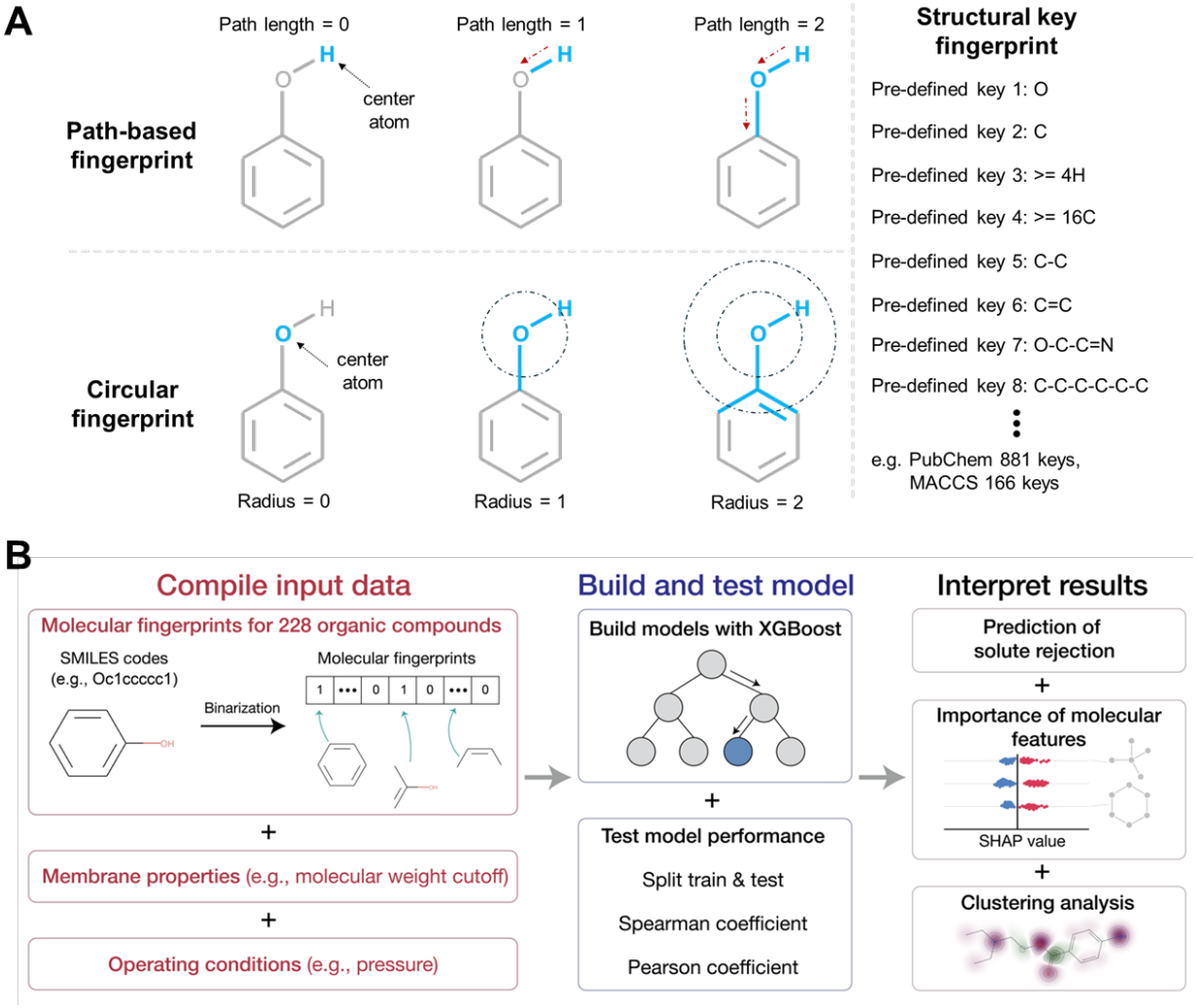


Figure 1. (A) Schematic diagrams of fingerprints used in the study. The pre-defined keys for structural key fingerprints are arbitrary keys, and not related to PubChem and MACCS

fingerprints. (B) Overall workflow of this study demonstrating the input data processing, model development procedures, and approaches for model interpretation.

2.3. Model development

Three different algorithms (linear Stochastic Gradient Descent, Support Vector Machine, and Extreme Gradient Boosting) were used and compared for training preliminary models. In addition, randomized hyperparameter search was implemented for fine-tuning the three models. This algorithm selection stage is important for evaluating the performance of each algorithm model and the suitability of an algorithm for our data characteristics. Detailed comparisons between algorithms can be found in Section 1 of the Supporting Information (Figure S1-3). XGBoost (Extreme Gradient Boosting) algorithm was chosen to build main models for predicting contaminant rejection with multiple fingerprint datasets (Figure 1B) based on its high prediction performance, low tendency for overfitting, resistance to outliers, and ability to handle a large number of input features as shown in Section 1 of the Supporting Information. XGBoost falls under the category of ensemble methods. The algorithm creates a series of additive estimators, where each estimator in the series is fit to the residual errors of the previous estimator. Training is generally fast and capable of handling overfitting due to improved regularization [50]. The algorithm has a small probability to predict a value outside of the given output range (rejection 0–100%), so we defined boundary conditions to ensure predicted values always fall in this range. Incremental learning was used to handle the 70 samples with the same output value (100% rejection) with different inputs. Conceptual explanation about incremental learning can be found in Figure S5. In each cycle, 50% of the training set was randomly selected to train the model, and positive normally distributed random noises (1% mean and 0.5% standard deviation of the 100% rejection) were deducted from the 100% rejection values to jitter the points. This process was repeated over 100 cycles and the coefficients of a model were incrementally updated at each cycle. The testing set that was not seen during the training was used for evaluation. To verify the models' validity and robustness, datasets were randomly split into a training set and a testing set with a ratio of 4 to 1 using the stratified splitting with 10 bins. Even with the transformation used to make the dataset more uniform in output distribution, the datasets were significantly imbalanced with few points in the low rejection range, so a stratified split was applied with 5 bins to preserve the percentage of samples for each bin in the training and testing sets. In addition, it was wrapped with

a Bayesian optimization wrapper to fine-tune XGBRegressor's parameters while implementing incremental learning. 5-fold cross validation was implemented for this hyperparameter tuning and training stages. 'max_depth', 'learning_rate', 'n_estimators', 'min_child_weight', 'subsample', 'colsample_bytree', 'reg_alpha', 'reg_lambda', and 'gamma' were the parameters to be tuned. To evaluate model performance, the mean square error (MSE), Spearman coefficient, and coefficient of determination were calculated with each training and testing set, respectively. Coefficient of determination is commonly used to evaluate prediction models and allow for comparison with previous studies [14,51]. Spearman coefficients are used to assess non-linear relationships and are therefore useful for membrane rejection data, which relies on complex interactions between membrane properties, operating conditions, and compounds. All the procedures above were executed with the XGBoost and scikit-learn libraries in Python. The datasets and codes used to generate the results were deposited in a GitHub repository (https://github.com/leesang-boulder/machine-learning_molecular-fingerprint_RO-NF-rejection).

2.4. Clustering and Shapley values for model analysis

Clustering is a powerful unsupervised machine learning technique that can help to group similar data points. Clustering can help to identify patterns and relationships between different groups of data points. By examining the characteristics of each cluster, we can understand what makes them similar or different. T-distributed Stochastic Neighborhood Embedding (t-SNE) was used for nonlinear dimension reduction and clustering, which can be found in the Scikit Learn Library in Python. The number of clusters is a hyperparameter, so the inertia (sum of distance squares between a central point and data points in clusters) curve was drawn for finding a suitable number of clusters with the MACCS fingerprint as high-dimensional input data. Based on the inertia curve (see Figure S6), the number of clusters was decided as 4. The MACCS fingerprint was used to implement the clustering of the 228 organic compounds. Sub-dataset was generated with conditions of MWCO < 225 Da and pH = 7 to eliminate the effects of MWCO and pH, and focus on the effect of the molecular structures. A point on the plot represents an organic compound, and the values of the x and y axes are equivalent to the reduction of 167 structural keys of MACCS. The averaged rejections of the compounds were displayed with colors.

Shapley additive explanations (SHAP) were used for evaluating the influence of each single input feature on rejection. The process is based on cooperative game theory that analyzes

the significance of each input variable in the outcome [52]. With SHAP, models are trained with all possible combinations of the input variables, and the differences of the predicted outcome including and excluding an input variable of interest are computed. The obtained SHAP scores can be either positive or negative to present the trends of output, and the scores indicate which compound fragments (input) most affect rejection (output) and thus help explain the rejection mechanisms. Molecular fragments of compounds were drawn using the RDKit library to further investigate molecular structures that are important in determining rejection. The importance score of a feature was obtained by averaging absolute SHAP values of each data point for the feature.

3. Results and Discussion

3.1. Validation of learning process and prediction performance

The learning processes and prediction performances of the three types of MF models (path-based, circular, and structural key) were evaluated with data divided into training and testing sets with stratified split, and the results indicated that models were well-trained. The datasets included the same input features and output rejections, except the type of MF used as input was varied. Figure 2A–C shows the learning curves with different fingerprints where the maximum path length for the path-based fingerprint was 3, the maximum radius for the circular fingerprint was 3, and the bit length of the PubChem fingerprint was 881. The default maximum path length and radius are 7 in RDKit, and here we chose 3 because organic compounds in the data are small compared to other molecules studied in RDKit (e.g., pharmacoactive organic molecules) and larger lengths increase computational time and likelihood of overfitting. The learning curves are broadly used to see how models learn and minimize errors over time. Figure 2A-C demonstrate that error decreases with an increasing number of iterations, eventually reaching a lower limit of error at a high number of iterations. For the circular and PubChem fingerprints, the testing errors did not start increasing after hitting the minimum; suggesting that the models were not overfit, which would be reflected by testing error increasing while training error decreases. For example, the testing error during the learning process with the circular fingerprint was about 0.0013, and the training error was 23% lower at around 0.0010. On the other hand, the testing error of the path-based model increased by 16% from the 90^{th} iteration to the end of the 100^{th} iteration, which indicates that the model may start being overfitted at approximately the 90^{th} iteration and should likely be stopped early.

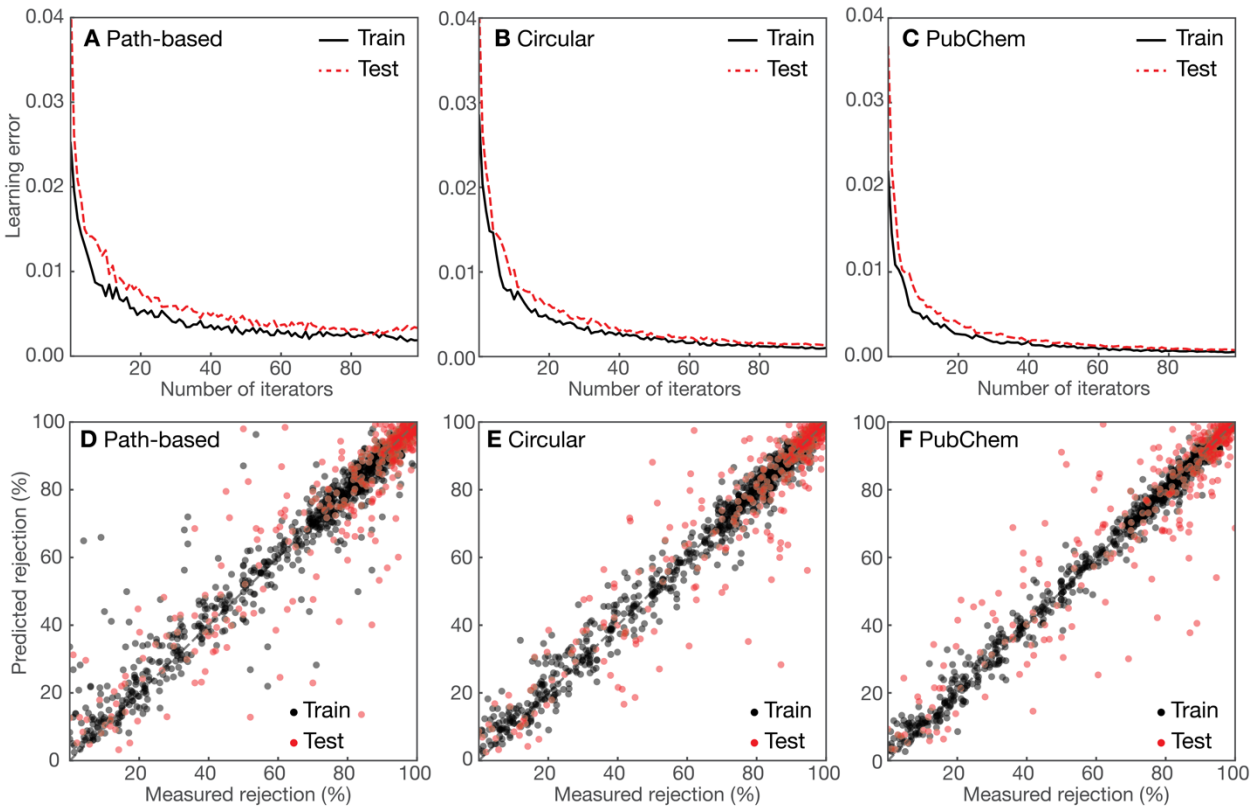


Figure 2. (A–C) Learning curves and (D–F) fit curves of models trained with different molecular fingerprints to compare the distributions of the measured and predicted data. The maximum path length and bit length for the path-based fingerprint were 3 and 32768, respectively. The maximum radius and bit length for the circular fingerprint were 3 and 16384, respectively. The bit length of the PubChem fingerprint was 881. Light black and red colors correspond to single data points, whereas darker shades correspond to a high density of overlapping data points.

After training, measured rejections and predicted rejections were compared (Figure 2D–F), and mean square error (MSE), Spearman coefficient, and coefficient of determination of the three fingerprint models were calculated (Table 2). MSE values in the learning curves were obtained using scaled rejections, and the MSE values in Table 2 are reported with rejections in the original scale (0–100%). All three fingerprint models (path-based, circular, and PubChem) had high Spearman coefficients (greater than 0.98) with the training sets. For the testing sets, the Spearman coefficients of the path-based, circular, and PubChem models decreased to 0.91, 0.93, and 0.92, respectively. The coefficient of determination results were similar to the Spearman results in that

the coefficient of determination with the training sets were high (greater than 0.96), and the coefficients with the testing sets fell in the range of 0.84–0.88 with the path-based model again showing the lowest coefficient.

Table 2. Training and testing performances of the models with different fingerprint algorithms and their parameters. Mean squared error (MSE), Spearman coefficient, and coefficient of determination (R^2) of percent rejection were used as evaluation metrics. Spearman coefficient and coefficient of determination are unitless, and MSE has units squared rejection percentage.

Name	Fingerprint parameter	Bit length	Train MSE(% rejection^2)	Test MSE(% rejection^2)	Train Spearman	Test Spearman	Train R^2	Test R^2
Path-based 3	Maximum path length 3	32768	2.60	6.03	0.98	0.91	0.96	0.84
Circular 3	Maximum radius 3	16384	1.92	5.25	0.99	0.93	0.99	0.88
PubChem	-	881	1.37	5.61	0.99	0.92	0.99	0.86

To compare the performance of ML algorithms using MFs to prior ML work, results from 11 previous studies that used ML to evaluate membrane rejection are summarized in Table 3 [22–24,40,53–59]. The datasets used in prior work included information on different combinations of molecular properties, membrane properties, membrane fabrication conditions, operating conditions, and RO and NF membrane rejections. Although the most common ML algorithm used for prior studies has been neural networks, gradient boosting has recently become popular. The ranges of the data size varied widely from 19 to 1906 rejection points for aqueous systems and up to 6910 rejection points for datasets using organic solvents. Although some studies did not report performance metrics that can be compared to those in this work, the available coefficients of determination with the testing set (R^2 test) fell in the range between 0.84 and 0.99. Our work had a coefficient of determination of 0.88 with the testing set. Direct comparison of these values is difficult since data sets and study methods are widely varied. However, these results indicate that predictive performances of the fingerprint-based models introduced in this work are comparable to those of prior work that used molecular descriptors or other metrics.

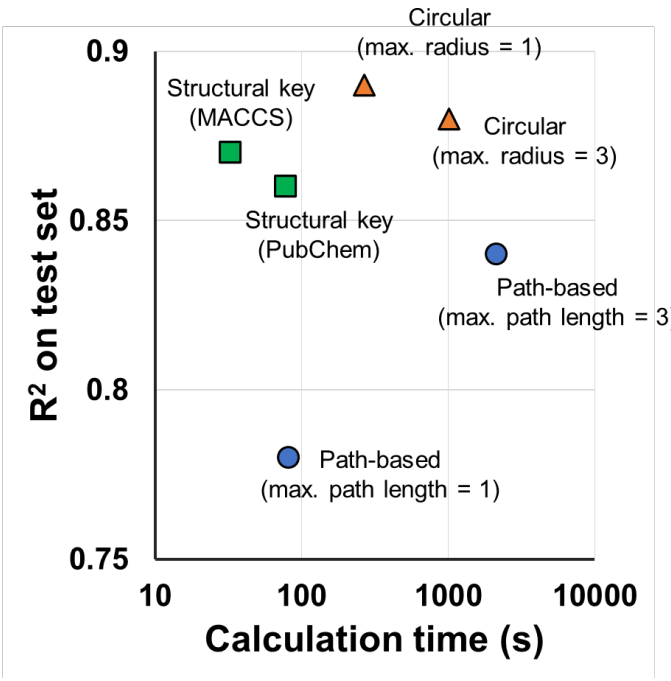
Table 3. Summary of previous studies using machine learning to predict contaminant rejection in nanofiltration (NF), reverse osmosis (RO), and pervaporation (PV). Note that the root mean square error (RMSE) and mean absolute error (MAE) values have been adjusted to the same scale.

Ref.	Year	Algorithm	Process	Data size	Input feature	Output feature	RMSE train	RMSE test	MAE train	MAE test	R^2 train	R^2 test	RMSE CV	R^2 CV
[60]	2000	Neural Net	NF	342	operating conditions, salt type, membrane type	solute rejection								
[61]	2008	Neural Net	RO	50	solute properties	passage/sorbed/ rejected fractions								
[62]	2009	Neural Net	RO/NF	124	solute properties	solute rejection			6.113	4.56	0.91	0.97		
[63]	2015	Neural Net	RO/NF	965	solute properties, membrane properties, operating conditions	solute rejection	10.78	11.53			0.921	0.904		
[64]	2016	Neural Net	RO/NF	436	-	solute rejection								
[65]	2020	Random Forest	RO/NF	701	solute properties, membrane properties, operating conditions	solute rejection	2.5	9.2			0.947	0.907		
[66]	2020	Genetic Algorithm	NF	19	solute properties	solute rejection								
[67]	2021	Neural Net, Support Vector Machine, Random Forest	NF	6910	solute properties, membrane properties, operating conditions	solvent permeance, solute rejection	4.42	12.14			0.989	0.914		
[58]	2022	Extreme Gradient Boosting, Categorical Gradient Boosting	NF	1524	monomer, fabrication conditions, operating conditions	water permeability, solute rejection	4.17	11.74			0.980	0.840		
[68]	2022	Partial Least Square, Convolutional Neural Net	NF	6910	solute properties, membrane properties, operating conditions	solute rejection							7.95	0.89
[59]	2022	Gradient Boosting, Kernel Ridge	PV	681	molecular fragments, membrane properties, operating conditions	solvent flux, separation factor			0.077	0.126	0.995	0.987		
This study	2024	Extreme Gradient Boosting	RO/NF	1906	molecular fingerprints, membrane properties, operating conditions	solute rejection	1.92	5.25			0.99	0.88		

3.2. Effect of varying fingerprint hyperparameters on model performance and bit collision

In the previous section, we found that all three fingerprint types (path-based, circular, and structural key) were able to produce comparable predictive performances when applied to the rejection dataset. We also observed that the path-based model became more susceptible to overfitting compared to the other models. In this section, we vary the hyperparameters of MFs (maximum path length for the path-based fingerprint, maximum radius for the circular fingerprint, and length of the pre-defined structural keys for the PubChem fingerprint) to further investigate how the simulation conditions affect the performance and interpretability of the models. The hyperparameters control the maximum boundary of fragmentation, and hence, the molecular fragments of a compound can change based on the hyperparameter size. The default hyperparameters used in the previous section were 3 for the maximum path length in the RDKit path-based fingerprint, 3 for the maximum radius in the circular fingerprint, and 881 pre-defined structural keys for the PubChem fingerprint, respectively. We selected length or radius of 3 because the organic compounds in the data are relatively small compared to other compounds studied using fingerprints, such as proteins and macromolecules. The maximum path length of 1, maximum radius of 1, and the MACCS fingerprint as a shorter structural key fingerprint (167 keys) were additionally used to compare performances and changes in calculation time.

The predictive performances with different hyperparameters showed that the test coefficient of determination of the path-based model noticeably dropped when the path length was changed from 3 to 1, with values of 0.84 and 0.78, respectively (Figure 3). The coefficient of determination of the circular and PubChem models did not considerably change when lower hyperparameters were used (the coefficient varied by 0.01). In the case of the circular fingerprints, this indicates that the organic compounds in the data are not large and the circular radius of 1 was enough to represent the compounds. The Spearman coefficients, which are good at estimating monotonic association between two variables, were similar (0.90–0.93) for different fingerprints and hyperparameters because non-linearity of the Spearman coefficient makes its evaluation less restrictive than the coefficient of determination.



335

336 **Figure 3.** Comparing prediction performance and calculation time of different types of molecular
337 fingerprints (path-based, circular, and structural key) with different hyperparameters. Structural
338 key fingerprints showed the fastest calculation times while maintaining high prediction
339 performance as indicated by a high coefficient of determination of the testing set. Further
340 information on prediction performance and calculation times is shown in Table S2.

341

342 Calculation time of the different fingerprints and hyperparameters was also evaluated
343 (Figure 3). For a given fingerprint and hyperparameter, calculation times were estimated using the
344 shortest fingerprint bit length that avoided bit collision, which occurs when different substructures
345 are stored in the same bit [69]. Bit collision typically only happens in the hashed style fingerprints
346 (in this study, the path-based and circular MFs) as a result of short bit lengths, and can lead to
347 inaccurate interpretation of rejection behavior. Increasing the bit length to prevent bit collision
348 results in larger calculation times, and small amounts of bit collision may be considered acceptable
349 in some ML work. However, we chose to avoid bit collision in this study since substructures were
350 later used for rejection analysis, and bit collision would interfere with this analysis. In our tests,
351 the path-based fingerprint with a maximum path length of 3 and a bit length of 32768 had the
352 longest calculation time (2154 seconds). The long calculation time indicates that the path-based
353 fingerprint with the maximum length of 3 holds the highest number of combinations of

substructures among six different cases; this high number of combinations possibly explains the overfitting observed in Figure 2A. The circular fingerprint with the maximum radius of 3 required 16384 bits for no bit collision, and its calculation time was 1011 seconds. Structural key fingerprints including PubChem and MACCS had low bit lengths and calculation times compared to other fingerprints. Their lengths were below 1000 bits, and the calculation times were only 78 and 33 seconds for PubChem and MACSS, respectively.

The analysis of various MF types and hyperparameters showed the coefficient of determination and Spearman coefficient of PubChem, MACCS, and circular fingerprint models were all similarly high. However, the path-based fingerprint models with the maximum path length of 1 and 3 gave lower prediction performance than other cases. We hypothesize that this is because the path-based fingerprint model with the maximum path length of 3 suffered from overfitting, even with the training algorithm minimizing overfitting due to sparse data in the larger substructure information, and the path-based fingerprint model with the maximum path length of 1 generated less meaningful substructures to predict rejection due to the low substructure path length. Circular and structural key fingerprints were able to reach higher prediction performances than the path-based fingerprint cases. Structural key fingerprints had the lowest calculation times, and the shorter pre-defined structural keys of MACCS were especially fast to compute. MACCS may therefore be preferred over other fingerprints such as PubChem, circular, and path-based fingerprints to analyze membrane rejection of organic compounds because of the combination of performance and speed.

3.3. Model interpretation with SHAP and clustering analyses

The use of MFs in rejection models allows us to interpret the role of specific molecular substructures in determining overall rejection performance. We used SHAP analysis to investigate the interpretability of each algorithm by evaluating the importance of individual molecular substructures and input features. In addition, we demonstrated how the relation between molecular fragments and rejection mechanisms can be derived using a clustering technique.

SHAP importance scores were compared for the different datasets to identify which input parameters most substantially impacted the rejection (Figure 4). We first examined the importance of membrane properties and operating conditions. All cases show very clear trends correlating rejection with membrane molecular weight cut-off (MWCO) and pH with importance scores of

0.49 and 0.16, respectively. High MWCO decreased steric rejection of compounds. High pH increased rejection since a high pH value results in more negatively charged membranes and, in some instances, more negatively charged compounds.

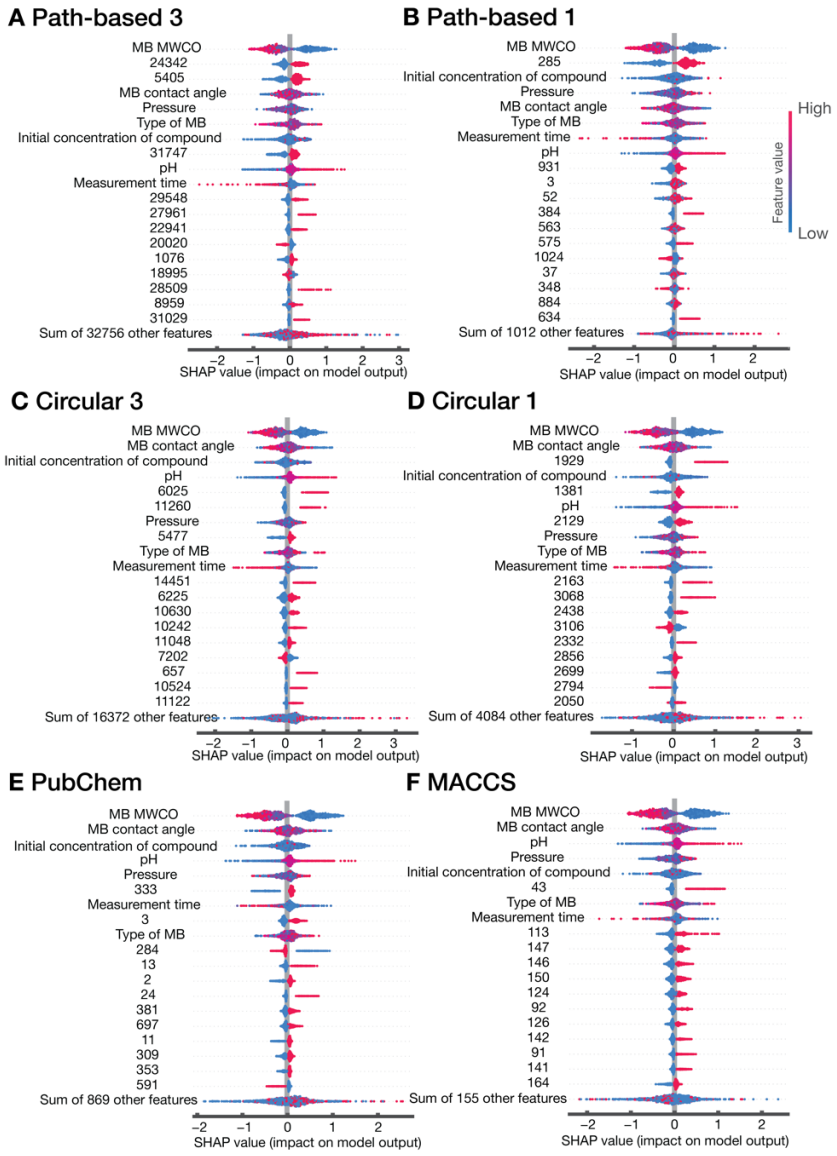


Figure 4. SHAP values describing the importance of input features in determining membrane rejection for different fingerprints and hyperparameters. Input features include operating conditions, membrane properties, and molecular fingerprint bit number. Higher and lower SHAP values for a given feature indicate that the feature is related to increases or decreases in the output rejection value, respectively. Red color indicates an input feature's value is high, and blue color indicates an input feature's value is low. For example, high MWCO values (red points) have

negative SHAP values. This means high MWCO membranes are related to a decrease in the output membrane rejection. Similarly, low MWCO values (blue points) have positive SHAP values, meaning that low membrane MWCO is related to an increase in rejection. The magnitude of the SHAP value corresponds to the change in rejection related to a given input feature.

Thus, the higher pH enhances electric repulsion between membranes and the compounds [70]. The results of measurement time indicate that shorter measurement times are associated with higher rejection values, while longer measurement times tend to result in lower rejection values. This probably arises because prolonged measurements offer a more accurate representation of steady-state system performance, capturing the impact of factors such as adsorption that short-term tests may overlook. Other membrane properties and operating conditions including contact angle, operating pressure, and type of membrane did not show consistent trends with all algorithms.

Analysis of the individual MF substructures with SHAP also provided some insights on the physical implications of molecular structures on membrane rejection. Figure 4 lists the feature codes for MFs that showed high importance scores (structure schematics are shown in Table S3-S6). Some of the substructures with high impacts on membrane rejection were associated with molecular size or charge. The presence of a tetrahedral group of non-H atoms bonded to a central carbon atom (285 Path-based 1, 24343 Path-based 3, 5405 Path-based 3, 113 MACCS, 2163 Circular 1, and 14451 Circular 3) led to high rejections presumably because this structure is correlated with a branched and bulky molecule, and thus, it indicates a larger size and shape [71]. There were functional groups frequently appearing over different fingerprints, such as carboxyl (-COOH), hydroxyl (-OH), and carbonyl (C=O) groups, which are related to electric charge and polarity. For example, organic acids (e.g., salicylic acid, clofibric acid, and benzoic acid) in the dataset contain the -COOH substructure. The hydrogen in -COOH can be deprotonated and result in a negative charge at neutral pH, which increases rejection due to strong electrostatic repulsion [72]. It is noteworthy that the presence of fluorine was found to be of high importance in the SHAP results (384 Path-based 1, 27961 Path 3, 24 PubChem, and 43 MACCS). However, this is likely linked to the large molecular size correlated with fluorination, rather than the effect of the fluorine itself. The dataset has 9 PFAS (perfluoroalkyl substances), and all the compounds have long carbon chains containing 7–19 fluorine atoms with the average molecular weight of PFAS being

67% higher than the average molecular weight of the entire dataset. The PFAS therefore showed high rejections of 86–98% due to their long chains, and fluorine indirectly represented large molecular size.

To further investigate the association between fingerprint substructures and rejection, we analyzed the molecular fragments of MACCS structural key fingerprints. In the previous subsection, we found that MACCS fingerprints are possibly the most efficient candidate for studying membrane rejection based on fast computation time and higher accuracy. MACCS fingerprints are also more interpretable because they are based on the presence or absence of specific structural features in a molecule. Each bit in these fingerprints corresponds to a particular structural feature, so it is relatively straightforward to understand what each bit represents. The top-ranked fragments can explain the interactions between compounds and membranes in a straightforward manner. In Figure 5, the 12 most important substructures of MACCS and their importance scores are displayed. The importance score was obtained by averaging absolute SHAP values in Figure 4. MACCS features 113, 126, 146, and 164 were important because the presence of rings or a tetrahedral geometry with non-H atoms is related to increased molecular size, e.g. the average McGowan molecular volume of compounds having tetrahedrons is 32% higher than the average McGown molecular volume of the entire data. Feature 124 represented carboxylic groups, and features 141 and 147 represented the counts of oxygen in compounds. Larger amounts of oxygen are possibly linked to negative surface charge due to oxygenated groups often having high electronegativity. Features 91, 92, 142, and 150 were associated with alkyl groups, which can affect size and polarity. Overall, the easily identifiable molecular features in the MACCS fingerprint, combined with SHAP analysis, allowed us to identify solute molecular structures most strongly linked to rejection.

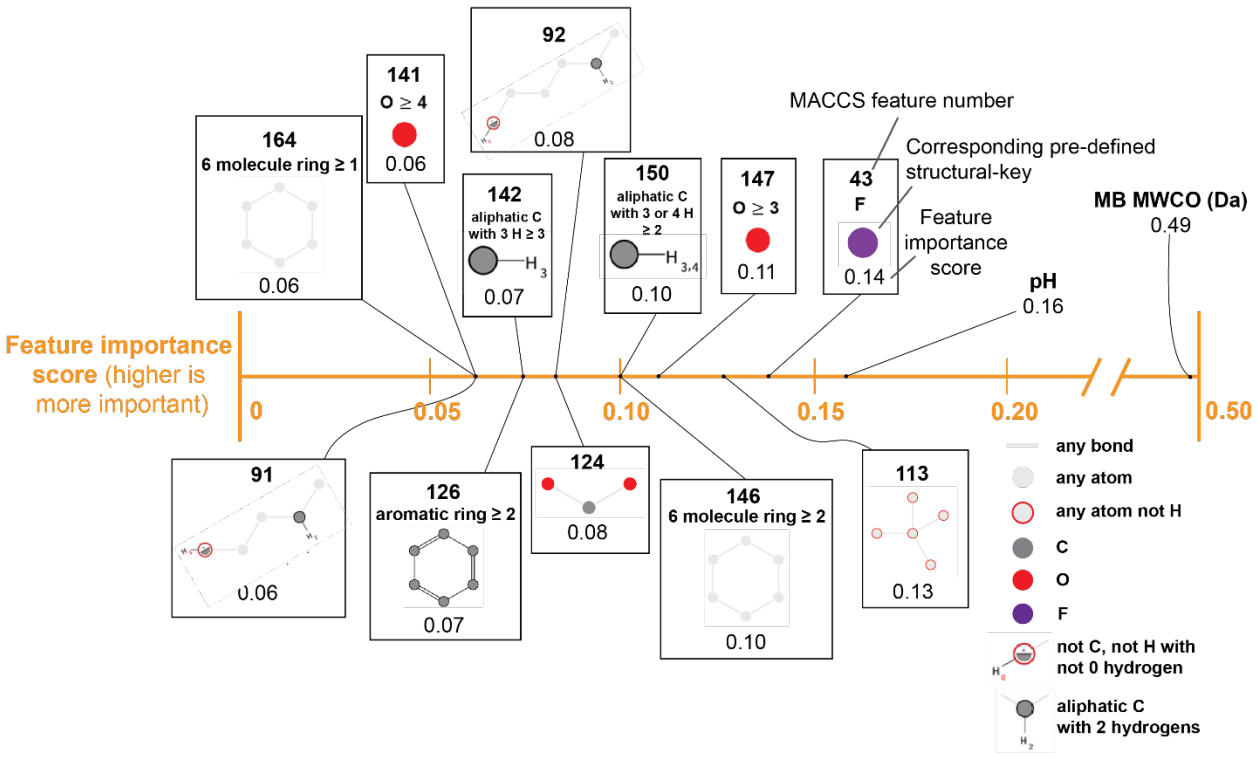


Figure 5. The 12 most important MACCS fingerprint features (importance score greater than 0.05). Feature importance scores were obtained from the SHAP analysis. Note that MWCO and pH have the first and second highest importance scores, and they are presented on the graph for comparison.

Unsupervised clustering algorithms can group molecules based on similarities in their fingerprints without prior labeling and reveal natural groupings or patterns in the data, often uncovering relationships that are not immediately apparent. Clustering analysis was used to categorize compounds based on their molecular structures and connect the clusters' rejection behaviors to their structural similarities and disparities. The analysis, which used MACCS fingerprints, revealed distinct clusters of compounds with different characteristics. Figure 6A shows four clusters with their mean and individual compound rejection values. Figure 6B demonstrates the most shared sub-structures in a cluster by counting frequently appeared fingerprint sub-structures. The top-10 features for each cluster can be found in Table S8. Cluster 2 is the cluster with the highest average cluster rejection (86%) and the least molecular structure variation; it contained compounds such as diltiazem, lidocaine, and propyphenazone which share common features, such as one or more benzene rings with various side chains and functional groups attached to the aromatic core (see Figure S7). Cluster 3 contained molecules with a long

linear path and multiple oxygen atoms including glucose, xylose, and triethyleneglycol (see Figure S7). On the other hand, Cluster 4, which showed the lowest average rejection, had very different sub-structures and had zero fingerprint features over 80% appearance frequency. The cluster also included small aldehydes and alcohols, which resulted in the low rejection in the range of 0 and 40%. It is noteworthy that Cluster 3 had a lower average molecular weight than Cluster 4, but Cluster 3 still showed a higher average cluster rejection (77% vs. 61%), possibly due to its larger molecular volume (e.g. increasing van der Waals radius), which made it difficult for a compound to pass through the pore [73]. The clustering analysis showed that categorizing compounds based on their molecular fingerprints can help pinpoint sub-structures affecting rejection and aids in better defining classes of compounds with similar rejection behavior.

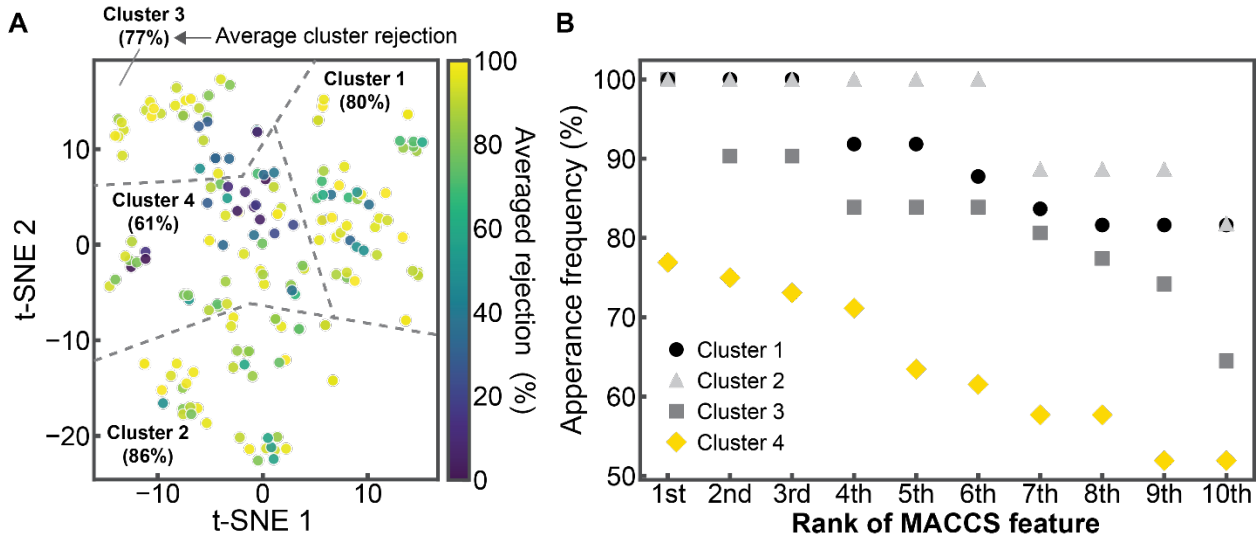


Figure 6. (A) Clustering of compounds and their associated rejection. Each point represents a unique chemical compound, and its color displays an average rejection of the compound. T-distributed Stochastic Neighborhood Embedding (t-SNE) was used for nonlinear dimension reduction from 167 to 2 keys (t-SNE 1 and t-SNE 2) and clustering based on MACCS fingerprints. A sub-dataset with MWCO < 225 Da & pH = 7 was used in clustering. (B) Frequency appearance of difference molecular features in each cluster. The appearance of all the MACCS features in each cluster was counted and converted to a percentage by dividing the count by the number of compounds in the cluster. Only the 10 most frequent features were shown, and the list of the features can be found in Table S8.

Conclusions

In this study, we analyzed membrane rejection of organic compounds using machine learning with molecular fingerprint features, membrane properties, and operating conditions. The molecular fingerprint-based approach estimates rejection accurately using structural information, bypassing the need for experimental or computationally intensive molecular properties. Our model's prediction performance, using molecular fingerprints, was comparable to previous studies on RO and NF solute rejection with data-driven models. We compared three fingerprint generation algorithms—path-based, circular, and structural key—with varying hyperparameters. The MACCS structural key fingerprint demonstrated the best combination of predictive power, calculation efficiency, and interpretability.

Our findings highlight that molecular fingerprints enhance our understanding and visualization of molecular substructures influencing rejection in NF and RO. SHAP analysis quantified the importance of individual molecular substructures (e.g., non-H atom bonded tetrahedrons, ring structures, functional groups) in determining rejection, revealing sensible rejection mechanisms. Clustering analysis grouped compounds by the MACCS fingerprint, identifying common or unique substructures in each cluster, which were related to membrane rejection. Future work may be able to further elucidate specific molecular structures that are strongly linked to solute-membrane interactions impacting membrane rejection. Eventually, the molecular-level insights into the rejection behavior can allow for the development of membranes that have tailored rejection of target compounds.

Acknowledgements

The authors are grateful to financial support from the National Science Foundation under Award Number CBET 2136835. S.L. acknowledges support from American Water Works Association – HDR One Water Institute Scholarship.

Note

The authors declare no competing financial interest.

References

[1] UN Water, UN World Water Development Report, Nature-based Solutions for Water, (2018). http://repo.floodalliance.net/jspui/handle/44111/2726 (accessed July 25, 2022).

[2] C.Y. Tang, Z. Yang, H. Guo, J.J. Wen, L.D. Nghiem, E. Cornelissen, Potable Water Reuse through Advanced Membrane Technology, Environ Sci Technol 52 (2018) 10215–10223. https://doi.org/10.1021/ACS.EST.8B00562

[3] Q. Yang, K.Y. Wang, T.S. Chung, A novel dual-layer forward osmosis membrane for protein enrichment and concentration, Sep Purif Technol 69 (2009) 269–274. https://doi.org/10.1016/j.seppur.2009.08.002.

[4] Y. Cui, T.S. Chung, Pharmaceutical concentration using organic solvent forward osmosis for solvent recovery, Nature Communications 2018 9:1 9 (2018) 1–9. https://doi.org/10.1038/s41467-018-03612-2.

[5] K. Lutchmiah, A.R.D. Verliefde, K. Roest, L.C. Rietveld, E.R. Cornelissen, Forward osmosis for application in wastewater treatment: A review, Water Res 58 (2014) 179–197. https://doi.org/10.1016/j.watres.2014.03.045.

[6] V. Sant'Anna, L.D.F. Marczak, I.C. Tessaro, Membrane concentration of liquid foods by forward osmosis: Process and quality view, J Food Eng 111 (2012) 483–489. https://doi.org/10.1016/j.jfoodeng.2012.01.032.

[7] A. Egea-Corbacho Lopera, S. Gutiérrez Ruiz, J.M. Quiroga Alonso, Removal of emerging contaminants from wastewater using reverse osmosis for its subsequent reuse: Pilot plant, Journal of Water Process Engineering 29 (2019) 100800. https://doi.org/10.1016/J.JWPE.2019.100800.

[8] B. Petrie, R. Barden, B. Kasprzyk-Hordern, A review on emerging contaminants in wastewaters and the environment: Current knowledge, understudied areas and recommendations for future monitoring, Water Res 72 (2015) 3–27. https://doi.org/10.1016/J.WATRES.2014.08.053.

[9] J.L. Acero, F.J. Benitez, F.J. Real, E. Rodriguez, Elimination of selected emerging contaminants by the combination of membrane filtration and chemical oxidation processes, Water Air Soil Pollut 226 (2015) 1–14. https://doi.org/10.1007/S11270-015-2404-8.

[10] V.S. Babu, M. Padaki, L.P. D'Souza, S. Déon, R. Geetha Balakrishna, A.F. Ismail, Effect of hydraulic coefficient on membrane performance for rejection of emerging contaminants, Chemical Engineering Journal 334 (2018) 2392–2400. https://doi.org/10.1016/J.CEJ.2017.12.027.

[11] V. Yangali-Quintanilla, A. Sadmani, M. McConville, M. Kennedy, G. Amy, A QSAR model for predicting rejection of emerging contaminants (pharmaceuticals, endocrine disruptors) by nanofiltration membranes, Water Res 44 (2010) 373–384. https://doi.org/10.1016/J.WATRES.2009.06.054.

[12] B.D. Coday, B.G.M. Yaffe, P. Xu, T.Y. Cath, Rejection of trace organic compounds by forward osmosis membranes: A literature review, Environ Sci Technol 48 (2014) 3612–3624. https://doi.org/10.1021/ES4038676

[13] J. Radjenović, M. Petrović, F. Ventura, D. Barceló, Rejection of pharmaceuticals in nanofiltration and reverse osmosis membrane drinking water treatment, Water Res 42 (2008) 3601–3610. https://doi.org/10.1016/J.WATRES.2008.05.020.

[14] N. Jeong, T. Chung, T. Tong, Predicting Micropollutant Removal by Reverse Osmosis and Nanofiltration Membranes: Is Machine Learning Viable?, Environ Sci Technol 55 (2021) 11348–11359. https://doi.org/10.1021/ACS.EST.1C04041.

[15] B. Van Der Bruggen, J. Schaep, D. Wilms, C. Vandecasteele, Influence of molecular size, polarity and charge on the retention of organic molecules by nanofiltration, J Memb Sci 156 (1999) 29–41. https://doi.org/10.1016/S0376-7388(98)00326-3.

[16] J.G. Wijmans, R.W. Baker, The solution-diffusion model: a review, J Memb Sci 107 (1995) 1–21. https://doi.org/10.1016/0376-7388(95)00102-I.

[17] X.L. Wang, T. Tsuru, S.I. Nakao, S. Kimura, The electrostatic and steric-hindrance model for the transport of charged solutes through nanofiltration membranes, J Memb Sci 135 (1997) 19–32. https://doi.org/10.1016/S0376-7388(97)00125-7.

[18] W.M. Deen, Hindered transport of large molecules in liquid-filled pores, AIChE Journal 33 (1987) 1409–1425. https://doi.org/10.1002/AIC.690330902.

[19] T. Chaabane, S. Taha, M. Taleb Ahmed, R. Maachi, G. Dorange, Coupled model of film theory and the Nernst–Planck equation in nanofiltration, Desalination 206 (2007) 424–432. https://doi.org/10.1016/J.DESAL.2006.03.577.

[20] C. Su, H. Yeo, Q. Xie, X. Wang, S. Zhang, Understanding and optimization of thin film nanocomposite membranes for reverse osmosis with machine learning, J Memb Sci 606 (2020) 118135. https://doi.org/10.1016/j.memsci.2020.118135.

[21] T. Bonny, M. Kashkash, F. Ahmed, An efficient deep reinforcement machine learning-based control reverse osmosis system for water desalination, Desalination 522 (2022). https://doi.org/10.1016/J.DESAL.2021.115443.

[22] R. Goebel, T. Glaser, M. Skiborowski, Machine-based learning of predictive models in organic solvent nanofiltration: Solute rejection in pure and mixed solvents, Sep Purif Technol 248 (2020) 117046. https://doi.org/10.1016/J.SEPPUR.2020.117046.

[23] Y. Ammi, L. Khaouane, S. Hanini, Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks, Korean Journal of Chemical Engineering 2015 32:11 32 (2015) 2300–2310. https://doi.org/10.1007/S11814-015-0086-Y.

[24] G. Ignacz, G. Szekely, Deep learning meets quantitative structure–activity relationship (QSAR) for leveraging structure-based prediction of solute rejection in organic solvent nanofiltration, J Memb Sci 646 (2022) 120268. https://doi.org/10.1016/J.MEMSCI.2022.120268.

[25] S. Zhong, J. Hu, X. Yu, H. Zhang, Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation, Chemical Engineering Journal 408 (2021) 127998. https://doi.org/10.1016/J.CEJ.2020.127998.

[26] Y. ling Liu, K. Xiao, A. qian Zhang, X. mao Wang, H. wei Yang, X. Huang, Y.F. Xie, Exploring the interactions of organic micropollutants with polyamide nanofiltration membranes: A molecular docking study, J Memb Sci 577 (2019) 285–293. https://doi.org/10.1016/J.MEMSCI.2019.02.017.

[27] T. Fujioka, H. Kodamatani, W. Yujue, K.D. Yu, E.R. Wanjaya, H. Yuan, M. Fang, S.A. Snyder, Assessing the passage of small pesticides through reverse osmosis membranes, J Memb Sci 595 (2020) 117577. https://doi.org/10.1016/J.MEMSCI.2019.117577.

[28] M.G. Shin, W. Choi, S.J. Park, S. Jeon, S. Hong, J.H. Lee, Critical review and comprehensive analysis of trace organic compound (TOrC) removal with polyamide RO/NF membranes: Mechanisms and materials, Chemical Engineering Journal 427 (2022). https://doi.org/10.1016/J.CEJ.2021.130957.

[29] T.R. Nickerson, E.N. Antonio, D.P. McNally, M.F. Toney, C. Ban, A.P. Straub, Unlocking the potential of polymeric desalination membranes by understanding molecular-level interactions and transport mechanisms, Chem Sci 14 (2023) 751–770. https://doi.org/10.1039/D2SC04920A.

[30] L. Xing, R.C. Glen, R.D. Clark, Predicting pKa by Molecular Tree Structured Fingerprints and PLS, J Chem Inf Comput Sci 43 (2003) 870–879. https://doi.org/10.1021/CI020386S.

[31] L. Xing, R.C. Glen, Novel Methods for the Prediction of logP, pKa, and logD, J Chem Inf Comput Sci 42 (2002) 796–805. https://doi.org/10.1021/CI010315D.

[32] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, MoleculeNet: A Benchmark for Molecular Machine Learning, Chem Sci 9 (2017) 513–530. https://doi.org/10.48550/arxiv.1703.00564.

[33] G. Hinselmann, L. Rosenbaum, A. Jahn, N. Fechner, A. Zell, JCompoundMapper: An open source Java library and command-line tool for chemical fingerprints, J Cheminform 3 (2011). https://doi.org/10.1186/1758-2946-3-3.

[34] F.O. Sanches-Neto, J.R. Dias-Silva, L.H. Keng Queiroz Junior, V.H. Carvalho-Silva, “pySiRC”: Machine Learning Combined with Molecular Fingerprints to Predict the Reaction Rate Constant of the Radical-Based Oxidation Processes of Aqueous Organic Contaminants, Environ Sci Technol 55 (2021) 12437–12448. https://doi.org/10.1021/acs.est.1c04326.

[35] Y. Ding, M. Chen, C. Guo, P. Zhang, J. Wang, Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties, J Mol Liq 326 (2021) 115212. https://doi.org/10.1016/J.MOLLIQ.2020.115212.

[36] S. Zhong, J. Hu, X. Fan, X. Yu, H. Zhang, A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants, J Hazard Mater 383 (2020) 121141. https://doi.org/10.1016/J.JHAZMAT.2019.121141.

[37] S. Zhong, K. Zhang, D. Wang, H. Zhang, Shedding light on “Black Box” machine learning models for predicting the reactivity of HO radicals toward organic compounds, Chemical Engineering Journal 405 (2021) 126627. https://doi.org/10.1016/J.CEJ.2020.126627.

[38] P.G. Francoeur, D. Peñaherrera, D.R. Koes, Active Learning for Small Molecule pKa Regression; a Long Way To Go, (2022). https://doi.org/10.26434/CHEMRXIV-2022-8W1Q0.
[39] A Unified Approach to Interpreting Model Predictions, (n.d.). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed July 25, 2022).
[40] S. Lee, J. Kim, Prediction of Nanofiltration and Reverse-Osmosis-Membrane Rejection of Organic Compounds Using Random Forest Model, Journal of Environmental Engineering 146 (2020) 04020127. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001806.
[41] A. Martín, F. Martínez, J. Malfeito, L. Palacio, P. Prádanos, A. Hernández, Zeta potential of membranes as a function of pH: Optimization of isoelectric point evaluation, J Memb Sci 213 (2003) 225–230. https://doi.org/10.1016/S0376-7388(02)00530-6.
[42] S. Riniker, G.A. Landrum, Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods, J Cheminform 5 (2013) 1–7. https://doi.org/10.1186/1758-2946-5-43.
[43] K. Gao, D.D. Nguyen, V. Sresht, A.M. Mathiowetz, M. Tu, G.-W. Wei, Are 2D fingerprints still valuable for drug discovery? †, Phys. Chem. Chem. Phys 22 (2020) 8373. https://doi.org/10.1039/d0cp00305k.
[44] D.S. Wigh, J.M. Goodman, A.A. Lapkin, A review of molecular representation in the age of machine learning, Wiley Interdiscip Rev Comput Mol Sci (2022) e1603. https://doi.org/10.1002/WCMS.1603.
[45] Z. Cang, G.W. Wei, Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction, Int J Numer Method Biomed Eng 34 (2018). https://doi.org/10.1002/CNM.2914.
[46] Z. Cang, L. Mu, G.W. Wei, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening, PLoS Comput Biol 14 (2018) e1005929. https://doi.org/10.1371/JOURNAL.PCBI.1005929.
[47] K. Wu, G.W. Wei, Quantitative Toxicity Prediction Using Topology Based Multitask Deep Neural Networks, J Chem Inf Model 58 (2018) 520–531. https://doi.org/10.1021/ACS.JCIM.7B00558.
[48] K. Wu, Z. Zhao, R. Wang, G.W. Wei, TopP–S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility, J Comput Chem 39 (2018) 1444–1454. https://doi.org/10.1002/JCC.25213.
[49] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, Molecular fingerprint similarity search in virtual screening, Methods 71 (2015) 58–63. https://doi.org/10.1016/J.YMETH.2014.08.005.
[50] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (n.d.). https://doi.org/10.1145/2939672.
[51] V. Yangali-Quintanilla, A. Sadmani, M. McConville, M. Kennedy, G. Amy, A QSAR model for predicting rejection of emerging contaminants (pharmaceuticals, endocrine disruptors) by nanofiltration membranes, Water Res 44 (2010) 373–384. https://doi.org/10.1016/j.watres.2009.06.054.

[52] K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artif Intell 298 (2021) 103502. https://doi.org/10.1016/J.ARTINT.2021.103502.
[53] W.R. Bowen, M.G. Jones, J.S. Welfoot, H.N.S. Yousef, Predicting salt rejections at nanofiltration membranes using artificial neural networks, Desalination 129 (2000) 147–162. https://doi.org/10.1016/S0011-9164(00)00057-6.
[54] D. Libotean, J. Giralt, R. Rallo, Y. Cohen, F. Giralt, H.F. Ridgway, G. Rodriguez, D. Phipps, Organic compounds passage through RO membranes, J Memb Sci 313 (2008) 23–43. https://doi.org/10.1016/J.MEMSCI.2007.11.052.
[55] V. Yangali-Quintanilla, A. Verliefde, T.U. Kim, A. Sadmani, M. Kennedy, G. Amy, Artificial neural network models based on QSAR for predicting rejection of neutral organic compounds by polyamide nanofiltration and reverse osmosis membranes, J Memb Sci 342 (2009) 251–262. https://doi.org/10.1016/J.MEMSCI.2009.06.048.
[56] L. Khaouane, Y. Ammi, S. Hanini, Modeling the Retention of Organic Compounds by Nanofiltration and Reverse Osmosis Membranes Using Bootstrap Aggregated Neural Networks, Arabian Journal for Science and Engineering 2016 42:4 42 (2016) 1443–1453. https://doi.org/10.1007/S13369-016-2320-2.
[57] J. Hu, C. Kim, P. Halasz, J.F. Kim, J. Kim, G. Szekely, Artificial intelligence for performance prediction of organic solvent nanofiltration membranes, J Memb Sci 619 (2021) 118513. https://doi.org/10.1016/J.MEMSCI.2020.118513.
[58] H. Gao, S. Zhong, W. Zhang, T. Igou, E. Berger, E. Reid, Y. Zhao, D. Lambeth, L. Gan, M.A. Afolabi, Z. Tong, G. Lan, Y. Chen, Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization, Environ Sci Technol 56 (2022) 2572–2581. https://doi.org/10.1021/ACS.EST.1C04373.
[59] M. Wang, Q. Xu, H. Tang, J. Jiang, Machine Learning-Enabled Prediction and High-Throughput Screening of Polymer Membranes for Pervaporation Separation, ACS Appl Mater Interfaces 14 (2022). https://doi.org/10.1021/ACSAMI.1C22886.
[60] W.R. Bowen, M.G. Jones, J.S. Welfoot, H.N.S. Yousef, Predicting salt rejections at nanofiltration membranes using artificial neural networks, Desalination 129 (2000) 147–162. https://doi.org/10.1016/S0011-9164(00)00057-6.
[61] D. Libotean, J. Giralt, R. Rallo, Y. Cohen, F. Giralt, H.F. Ridgway, G. Rodriguez, D. Phipps, Organic compounds passage through RO membranes, J Memb Sci 313 (2008) 23–43. https://doi.org/10.1016/J.MEMSCI.2007.11.052.
[62] V. Yangali-Quintanilla, A. Verliefde, T.U. Kim, A. Sadmani, M. Kennedy, G. Amy, Artificial neural network models based on QSAR for predicting rejection of neutral organic compounds by polyamide nanofiltration and reverse osmosis membranes, J Memb Sci 342 (2009) 251–262. https://doi.org/10.1016/J.MEMSCI.2009.06.048.
[63] Y. Ammi, L. Khaouane, S. Hanini, Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks, Korean Journal of Chemical Engineering 2015 32:11 32 (2015) 2300–2310. https://doi.org/10.1007/S11814-015-0086-Y.
[64] L. Khaouane, Y. Ammi, S. Hanini, Modeling the Retention of Organic Compounds by Nanofiltration and Reverse Osmosis Membranes Using Bootstrap Aggregated

Neural Networks, Arabian Journal for Science and Engineering 2016 42:4 42 (2016) 1443–1453. https://doi.org/10.1007/S13369-016-2320-2.
[65] S. Lee, J. Kim, Prediction of Nanofiltration and Reverse-Osmosis-Membrane Rejection of Organic Compounds Using Random Forest Model, Journal of Environmental Engineering 146 (2020) 04020127. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001806.
[66] R. Goebel, T. Glaser, M. Skiborowski, Machine-based learning of predictive models in organic solvent nanofiltration: Solute rejection in pure and mixed solvents, Sep Purif Technol 248 (2020) 117046. https://doi.org/10.1016/J.SEPPUR.2020.117046.
[67] J. Hu, C. Kim, P. Halasz, J.F. Kim, J. Kim, G. Szekely, Artificial intelligence for performance prediction of organic solvent nanofiltration membranes, J Memb Sci 619 (2021) 118513. https://doi.org/10.1016/J.MEMSCI.2020.118513.
[68] G. Ignacz, G. Szekely, Deep learning meets quantitative structure–activity relationship (QSAR) for leveraging structure-based prediction of solute rejection in organic solvent nanofiltration, J Memb Sci 646 (2022) 120268. https://doi.org/10.1016/J.MEMSCI.2022.120268.
[69] S. Riniker, G.A. Landrum, Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods, J Cheminform 5 (2013) 43. https://doi.org/10.1186/1758-2946-5-43.
[70] C. Bellona, J.E. Drewes, The role of membrane surface charge and solute physico-chemical properties in the rejection of organic acids by NF membranes, J Memb Sci 249 (2005) 227–234. https://doi.org/10.1016/J.MEMSCI.2004.09.041.
[71] Y. Kiso, Y. Sugiura, T. Kitao, K. Nishimura, Effects of hydrophobicity and molecular size on rejection of aromatic pesticides with nanofiltration membranes, J Memb Sci 192 (2001) 1–10. https://doi.org/10.1016/S0376-7388(01)00411-2.
[72] L.N. Breitner, K.J. Howe, D. Minakata, Effect of Functional Chemistry on the Rejection of Low-Molecular Weight Neutral Organics through Reverse Osmosis Membranes for Potable Reuse, Environ Sci Technol (2019) 11401–11409. https://doi.org/10.1021/acs.est.9b03856.
[73] J.R. Werber, C.J. Porter, M. Elimelech, A Path to Ultraselectivity: Support Layer Properties to Maximize Performance of Biomimetic Desalination Membranes, Environ Sci Technol 52 (2018) 10737–10747. https://doi.org/10.1021/ACS.EST.8B03426.