Environmental Science & Technology

pubs.acs.org/est

Article

Machine Learning-Assisted Design of Thin-Film Composite Membranes for Solvent Recovery

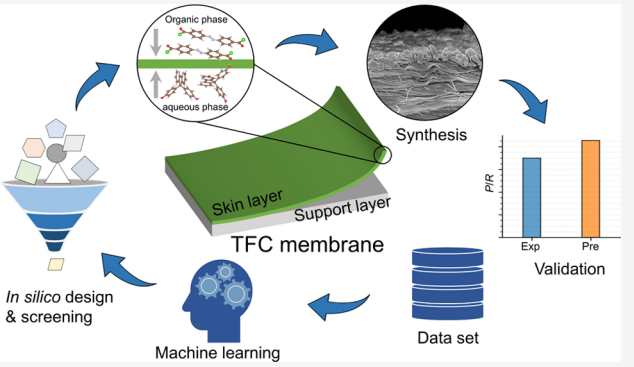
Mao Wang, Gui Min Shi, Daohui Zhao, Xinyi Liu, and Jianwen Jiang*

Cite This: *Environ. Sci. Technol.* 2023, 57, 15914−15924

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Organic solvents are extensively utilized in industries as raw materials, reaction media, and cleaning agents. It is crucial to efficiently recover solvents for environmental protection and sustainable manufacturing. Recently, organic solvent nanofiltration (OSN) has emerged as an energy-efficient membrane technology for solvent recovery; however, current OSN membranes are largely fabricated by trial-and-error methods. In this study, for the first time, we develop a machine learning (ML) approach to design new thin-film composite membranes for solvent recovery. The monomers used in interfacial polymerization, along with membrane, solvent and solute properties, are featurized to train ML models via gradient boosting regression. The ML models demonstrate high accuracy in predicting OSN performance including solvent permeance and solute rejection. Subsequently, 167 new membranes are designed from 40 monomers and their OSN performance is predicted by the ML models for common solvents (methanol, acetone, dimethylformamide, and *n*-hexane). New top-performing membranes are identified with methanol permeance superior to that of existing membranes. Particularly, nitrogen-containing heterocyclic monomers are found to enhance microporosity and contribute to higher permeance. Finally, one new membrane is experimentally synthesized and tested to validate the ML predictions. Based on the chemical structures of monomers, the ML approach developed here provides a bottom-up strategy toward the rational design of new membranes for high-performance solvent recovery and many other technologically important applications.



KEYWORDS: *solvent recovery, organic solvent nanofiltration, membrane, machine learning, chemical structure*

1. INTRODUCTION

In the chemical and pharmaceutical industries, organic solvents are commonly used for synthesis, cleaning, and purification,[1] with a global demand of about 20 million metric tons per year.[2] However, solvents in many processes are not effectively used and are usually incinerated, causing resource waste and environmental pollution. Therefore, solvent recovery has become one of the key green research topics for sustainable manufacturing.[3] At present, conventional thermal separation techniques like distillation are prevalently used for solvent recovery, but they involve energy-intensive phase transition. Recently, organic solvent nanofiltration (OSN), also known as solvent-resistant nanofiltration, has emerged as a robust technology for solvent recovery.[4] It can be operated in mild conditions (e.g., room temperature and a low pressure of 5−10 bar) without phase transition, thus energy-efficient, environmentally friendly, and well suited for the separation of thermal sensitive mixtures.[5]

OSN is pressure-driven membrane-based separation, in which membrane materials play a key role in performance. Polymeric integrally selective asymmetric (ISA) membranes are the most common for OSN and they are fabricated through phase inversion and covalent cross-linking of polymer chains. Nevertheless, cross-linking leads to high transport resistance and low permeance. By contrast, thin-film composite (TFC) membranes prepared by coating a selective layer on a substrate demonstrate great advantages over ISA membranes.[6] Because of the layered structure, the chemistry and performance of TFC membranes can be readily tuned. The selective layer is usually produced via interfacial polymerization, which involves two reactive monomers (one in an aqueous phase and the other in an organic phase) reacting at the aqueous/organic interface. This technique has been widely used in the construction of reverse osmosis membranes[7] and has also shown tremendous promise in the fabrication of OSN membranes.[8]

Received: June 20, 2023
Revised: September 26, 2023
Accepted: September 26, 2023
Published: October 10, 2023

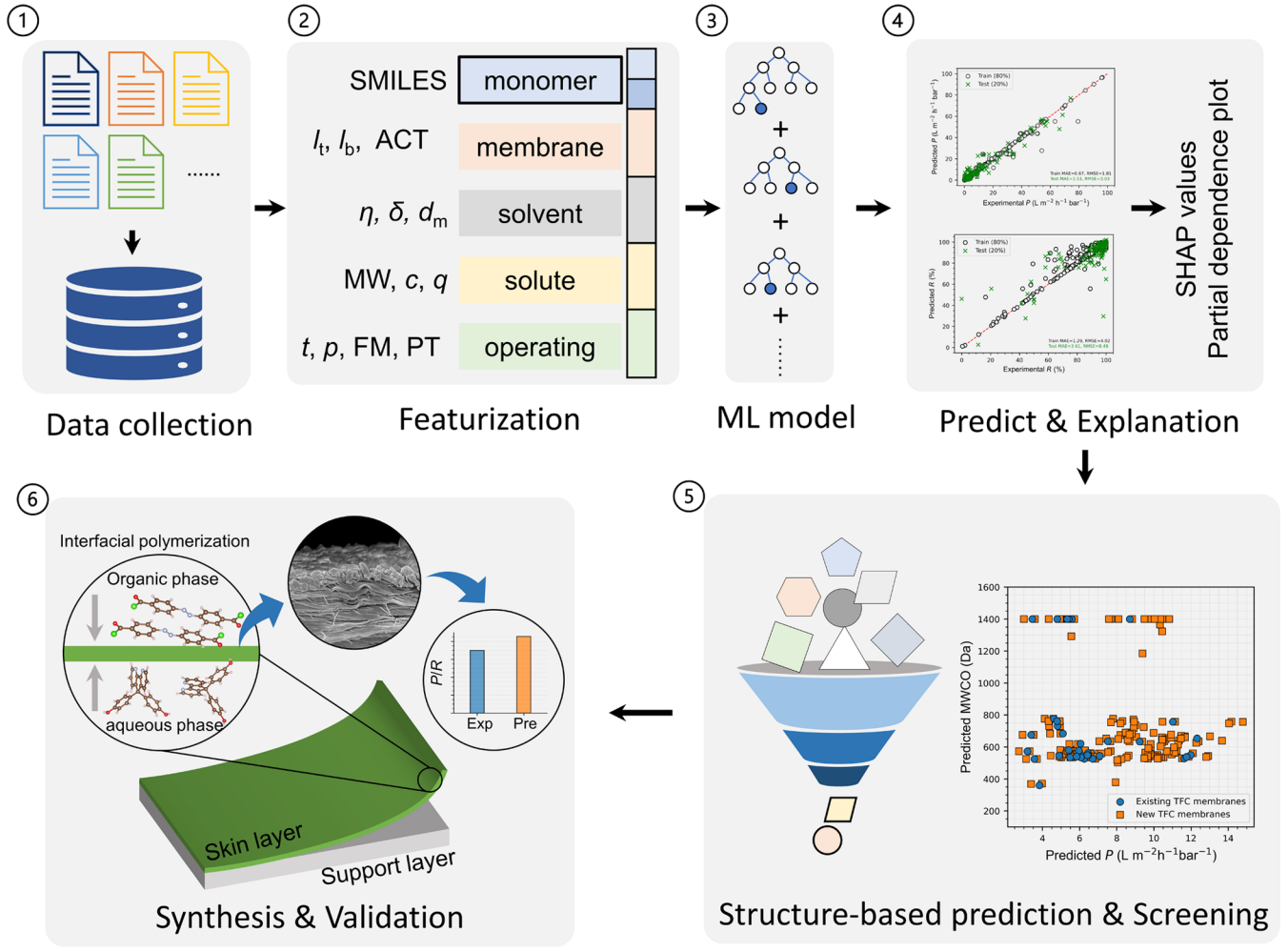


Figure 1. Workflow for the ML-assisted design of TFC membranes.

To produce a highly selective layer of a TFC membrane via interfacial polymerization, the proper selection of monomers is of central importance. For instance, water-insoluble cucurbit[6]uril (CB6) was ameliorated by piperazine (PIP) to react with trimethyl chloride (TMC), generating two types of selective tunnels (enlarged polyamide network tunnels and rotaxane tunnels) in a TFC membrane with exceptionally high solvent permeance.[9] From a bridged-bicyclic triptycene tetra-acyl chloride, a defect-free TFC membrane was fabricated with a large fraction of finely tuned submicroporosity (pore size < 4 Å) and superb performance in removing small organic microcontaminants.[10] A TFC membrane was produced using per-6-amino-cyclodextrin as a monomer and found to exhibit high permeance and intriguing shape selectivity of solvents due to its ultrathin thickness, regulated microporosity, and layered macrocycle.[11] These experiments demonstrate that the performance of TFC membranes can be tailored at the monomer level. There are a wide variety of monomers for interfacial polymerization, thus implying plenty of room is available to develop new TFC membranes. Due to the large chemical space involved, however, conducting experiments to achieve this is time-consuming and labor-intensive.

As a transformative technology, ML has been increasingly applied to the discovery of new materials. In the field of membranes, it has been increasingly used for gas separation,[12] water treatment,[13] biofuel purification,[14,15] and other liquid separation.[16] Nevertheless, only a few ML studies were reported on OSN, specifically, to predict solvent permeances in a polydimethylsiloxane membrane[17] and solute rejections in PuraMem S600 and PuraMem 280 membranes,[18] to evaluate the OSN performance of commercial and polyimide membranes,[19] to predict the rejections of more than 400 solutes in three polyamide membranes,[20] to optimize TFC membranes for OSN,[21] to examine solvent impact on solute rejection,[22] and to predict the separation of complex organic mixtures in linear polymer membranes.[23] Most of these ML studies primarily investigated the effects of operating conditions as well as solvent and solute properties on OSN performance in a handful of membranes, and they could not be applied to design new membranes or predict OSN performance of membranes with new structures because the chemical structures of membranes were not considered.

In this study, we develop an ML approach to design new TFC membranes for solvent recovery. Based on the chemical structures of monomers involved in interfacial polymerization, new TFC membranes are rationally designed, and their performance is evaluated by ML models. In our very recent study, although ML models were trained to predict methanol permeabilities in three polymer of intrinsic microporosity (PIM) membranes (PIM-A1, CX-PIM-A1, and PIM-8), no new membranes were designed.[24] To the best of our knowledge, this study is the first ML-assisted design of new

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

membranes for solvent recovery. While the standard membrane development is top-down, our approach is based on monomer structures at a molecular level. Such a bottom-up strategy would accelerate the development of new membranes for important applications.

2. METHODS

Figure 1 illustrates the workflow with five steps for the ML-assisted design of TFC membranes. ① Experimental data of TFC membranes for OSN were collected. ② Monomer structures, membrane, solvent and solute properties, and operating conditions were featurized as input features; performance metrics (solvent permeance and solute rejection) were used as targets. ③ ML models were trained and interpreted to establish quantitative relationships between features and targets. ④ New TFC membranes were designed and their performance was predicted to identify top-performing membranes. ⑤ One top-performing membrane was experimentally synthesized and tested.

2.1. Experimental Data Set. From 41 journal articles, we collected 1347 experimental data on TFC membranes for solvent recovery. If the data were tabulated or listed, the values were directly acquired from articles. If the data were graphical, the values were digitized using WebPlotDigitizer v4.6.[25] If missing data were found, the median values of collected data were used for estimation. Different from other OSN applications (e.g., dewaxing of lube oil and biodiesel production), solvent recovery in this work is to examine the recovery of small organic solvents (e.g., methanol and acetone) by removing solute molecules (e.g., dyes). Such solvent recovery is common in the chemical and pharmaceutical industries. The data set included six different categories of parameters: monomers, membrane properties, solvent and solute properties, operating conditions, and performance metrics (Table 1 and Figure S1).

For monomers, the focus was on the top selective layer produced by interfacial polymerization between the A monomer in an aqueous phase and the B monomer in an organic phase. The data set contained 34 A monomers and 6 B monomers, as shown in Figure S1a. The most common monomers were *m*-phenylenediamine and TMC in the aqueous and organic phases, respectively. In principle, these monomers (34 A and 6 B) could mutually polymerize into 204 different polymers. However, only 37 distinct polymers existed in the data set as the selective layers of TFC membranes. Thus, there are a larger number of unexplored polymers and our aim here is to design and identify top-performing new candidates. For membrane properties, the top selective layer thickness l_t (nm), bottom support layer thickness l_b (μm), and membrane activation state (whether treated by a solvent or not) ACT were included. For solvent properties, viscosity η (mPa·s), solubility δ ($MPa^{1/2}$), and molecular diameter d_m (nm) were considered. Among 19 distinct solvents in the data set (Figure S1b), methanol and acetone were commonly examined. We should note that the solvent used for TFC membrane preparation may affect membrane formation, structure, and performance. For example, the miscibility and viscosity of solvent were found to affect the diffusivity and solubility of monomers in the opposite phase; moreover, the pH of solvent might affect the growth rate of polymer.[8] However, in this work, we primarily focus on the selection of different monomers, and thus the effect of solvent on membrane formation was not considered. For solute properties, molecular weight MW ($g·mol^{-1}$), concentration c ($g·L^{-1}$), and charge q were utilized. There were 59 diverse solutes with common oligomers (e.g., styrene oligomers and polyethylene glycol) and dyes (e.g., rose bengal and methyl orange) (Figure S1c). For operating conditions, temperature t (°C), pressure p (bar), flow mode FM (cross-flow or dead-end), and permeate types PT (pure solvent or solvent−solute mixture) were employed. Finally, performance metrics for OSN were quantified by solvent permeance P ($L·m^{-2}·h^{-1}·bar^{-1}$) and solute rejection R (%).

Table 1. Parameters in Experimental Data Set

category	symbol	physical meaning	unit
monomers	A	monomer in the aqueous phase of the top selective layer	
	B	monomer in the organic phase of the top selective layer	
membrane properties	l_t	thickness of the top selective layer	nm
	l_b	thickness of the bottom support layer	μm
	ACT	membrane activation (0: no, 1: yes)	
solvent properties	η	viscosity	mPa·s
	δ	Hildebrand solubility parameter	$MPa^{1/2}$
	d_m	molecular diameter	nm
solute properties	MW	molecular weight	$g·mol^{-1}$
	c	concentration	$g·L^{-1}$
	q	charge (0: neutral, 1: positive, −1: negative)	
operating conditions	t	temperature	°C
	p	pressure	bar
	FM	flow mode (0: cross-flow, 1: dead end)	
	PT	permeate type (0: pure, 1: mixture)	
performance metrics	P	permeance	$L·m^{-2}·h^{-1}·bar^{-1}$
	R	rejection	

2.2. Featurization. The monomer structures, membrane, solvent and solute properties, as well as operating conditions, were featurized as input features. Specifically, the monomer structures were digitalized by molecular fingerprints using the simplified molecular input-line entry system (SMILES) strings.[26] Through such digitalization, we could establish quantitative structure−performance relationships and provide design guidelines. Five different molecular fingerprinting methods were adopted and compared in this work, including the fragment fingerprints (Frag),[15] Molecular ACCess System (MACCS) fingerprints,[27] Morgan fingerprints with bit vectors (MorganBit), and Morgan fingerprints with count vectors (MorganCount),[28] as well as RDKit descriptors (RDKitDP).[29] The MACCS and Frag fingerprints are based on 166 and 255 fixed substructure keys, respectively. The digit of each bit in a MACCS fingerprint indicates whether a substructure is present (1) or absent (0), whereas the digit of each bit in a fragment fingerprint is used to count the occurrence of a substructure. As for the Morgan fingerprints, substructural information is hashed into bit vectors, with the advantage of not requiring a predefined substructure library and allowing the length to be

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

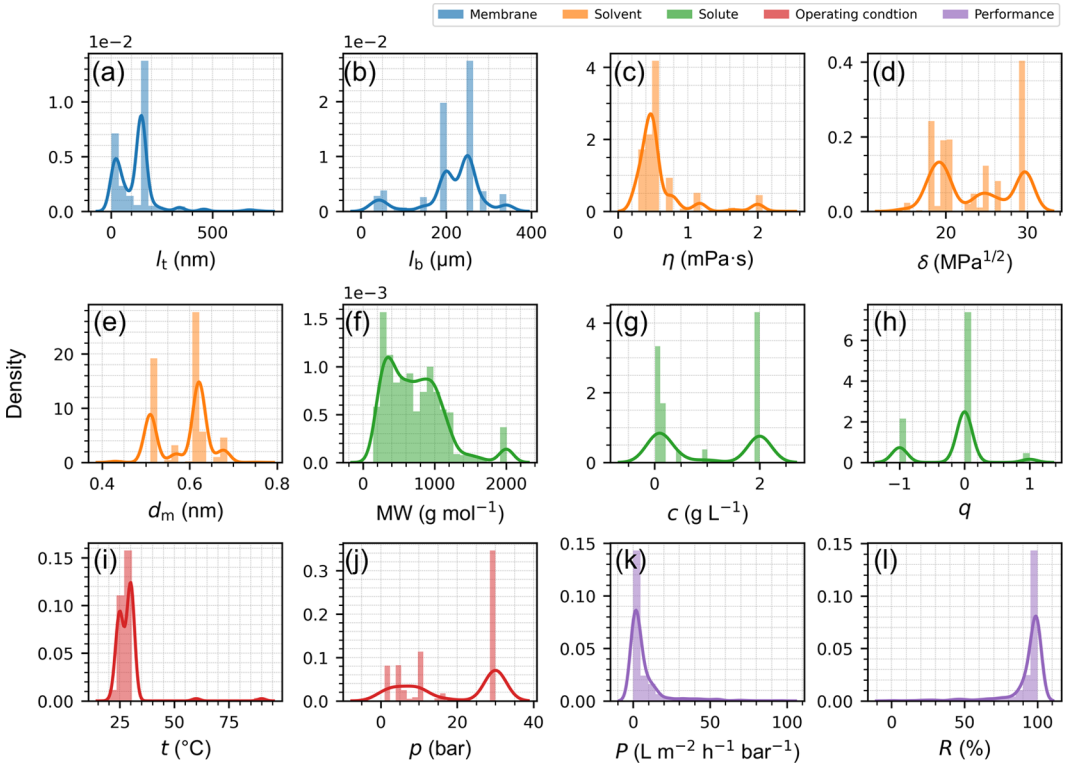


Figure 2. Distributions of different parameters in the experimental data set. Membrane properties: (a) top selective layer thickness l_t and (b) bottom layer thickness l_b. Solvent properties: (c) viscosity η, (d) Hildebrand solubility parameter δ, and (e) molecular diameter d_m. Solute properties: (f) molecular weight MW, (g) concentration c, and (h) charge q. Operating conditions: (i) temperature t and (j) pressure p. Performance metrics: (k) permeance P and (l) rejection R.

changed. Both the 0/1 digit and the occurring digit were employed here. The RDKitDP consists of 208 different physicochemical parameters for a chemical structure, which are commonly used in ML, providing electrons, topologies, functional groups, and other physicochemical information. All five different molecular fingerprints in this study were created by the RDKit cheminformatics package.[29]

2.3. Machine Learning. To train ML models, the performance metrics P and R were used as targets, while the monomer structures, membrane, solvent and solute properties, as well as operating conditions were input features. Gradient boosting regression (GBR)[30] was used for training. The GBR algorithm combines numerous weak regressors to create a powerful regressor, which has been shown to have a high predictive capability for various applications. The training was carried out by using the *scikit-learn* toolkit.[31] Unless otherwise noted, the experimental data were randomly divided into 80% and 20% in the training and test sets, respectively. It is worthwhile to note that data leakage may occur if training and test sets contain data from the same source, which can be mitigated by grouping the data in terms of source prior to the splitting of training and test sets.[32] Due to the limited data points in this work, out-of-sample validation was conducted by experimental synthesis and test of a new membrane. To tune hyperparameters, the GridSearchCV object with 5-fold cross-validation was adopted. The hyperparameters are listed in Table S1. The performance of ML models was evaluated by coefficient of determination (R^2), mean absolute error (MAE), and root-mean-square error (RMSE)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{exp}^{(i)} - y_{pre}^{(i)})^2}{\sum_{i=1}^{n} (y_{exp}^{(i)} - \overline{y}_{exp})^2} \quad (1)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_{exp}^{(i)} - y_{pre}^{(i)}|}{n} \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_{exp}^{(i)} - y_{pre}^{(i)})^2}{n}} \quad (3)$$

where y_{exp} and y_{pre} are the experimental and predicted results, respectively, and n is the number of samples. The units of MAE and RMSE are identical to those of P and R, where R is often represented as a percentage (dimensionless).

To provide in-depth insights, the partial dependence plot (PDP)[30] and the SHapley Additive exPlanations (SHAP)[33] were adopted to assess how different features play their roles in the ML models. The PDP was used to depict the interplay between input features and ML predictions, while the SHAP value was used to quantify the importance of each feature. Thereafter, the ML models were utilized to predict the performance of 167 new TFC membranes designed by combining different monomers, and the top-performing membranes were identified.

2.4. Experiment. To validate the ML predictions, we experimentally synthesized and tested one top-performing TFC membrane. A commercially available polyethylene (PE) membrane with a symmetric structure and a nominal pore size of 20 nm was purchased from the SEMCORP group. Methanol, acetone, *n*-hexane, and dimethylformamide

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

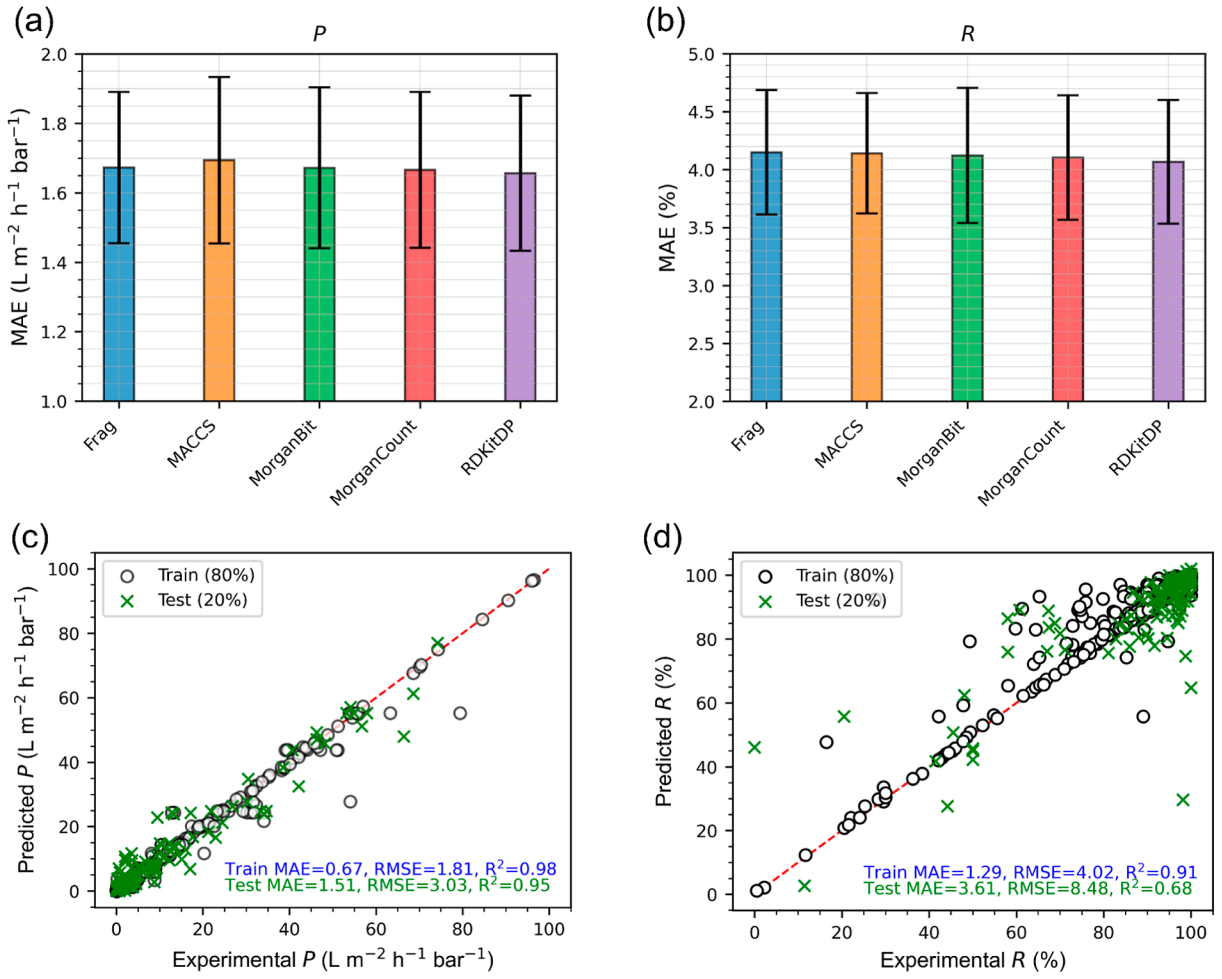


Figure 3. (a,b) Average MAE values for P and R predictions in the test set using five different molecular fingerprints with 50 different random splits of training (80%)/test (20%) sets. (c,d) ML predicted versus experimental P and R.

(DMF) with an analytical reagent grade were used as solvents. Three dyes including remazol brilliant blue, rose bengal, and alcian blue with molecular weights of 627, 974, and 1299 g·mol^{-1}, respectively, were purchased from Sigma-Aldrich and used as model solutes to test the OSN performance. PIP and azobenzene-4,4′-dicarbonyl dichloride (AZ) were purchased from Tokyo Chemical Industry Pte. Ltd. and Sigma-Aldrich Pte. Ltd., respectively. All chemicals were used as received.

A PE substrate was sandwiched between two aluminum frames for interfacial polymerization between PIP and AZ. The PE substrate was used in its original state or plasma treated. The plasma treatment was conducted in a plasma system supplied by Femto Science Inc. (Model CUTE-MPR) at a frequency of 50 kHz with pure Ar gas flowing at a flow rate of 20 sccm. The interfacial polymerization was carried out by first pouring a 1.5 wt % aqueous PIP solution for 2 min on the PE substrate. Subsequently, a 0.15 wt % AZ in n-hexane/dichloromethane (ratio 4:1) solution was poured on the dried substrate to react with PIP for 3 min. The PIP-AZ membrane after interfacial polymerization was rinsed with demineralized (DI) water and stored in DI water before test. The membrane morphology and thickness were evaluated by using a JSM-6700F field emission scanning electron microscope (FESEM).

Solvent permeation and solute rejection of the PIP-AZ membrane were measured by a dead-end permeation cell with a volume of 300 cm^3 under 5−6 bar at room temperature. The permeate sample was weighed with a precision balance (Denver Instrument, TB214). The permeance P was calculated by $P = Q/(A{\cdot}\Delta p)$, where Q denotes the volumetric flow rate of solvent, A refers to the effective filtration area, and Δp represents the transmembrane pressure. The solute rejection R was calculated by $R = (1 - C_p/C_f){\cdot}100\%$, where C_p and C_f are the solute concentrations in permeate and feed solutions, respectively, measured by using a UV−Vis spectrometer (Pharo 300, Merck).

3. RESULTS AND DISCUSSION

3.1. Analysis of Experimental Data Set. Figure 2 shows the distributions of different parameters in the experimental data set. The thickness of the top selective layer l_t ranges from 10 nm to several hundred nm, with many <100 nm in ultrathin membranes (Figure 2a). The thickness of the bottom support layer l_b is in the range of 100 to 400 μm, while a thin support layer around 30 μm also exists (Figure 2b). There are 19 distinct solvents in the data set, with viscosity η ranging from 0.29 to 2.27 mPa·s (Figure 2c), Hildebrand solubility parameter δ from 14.9 to 29.7 MPa$^{1/2}$ (Figure 2d), molecular diameter d_m from 0.43 to 0.75 nm (Figure 2e). For 59 solutes in the data set, molecular weight MW extends from 142 to 2009 g·mol^{-1}, as in wide variety of OSN applications (Figure 2f). The solute concentration c is mostly below 1 g·L^{-1},

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

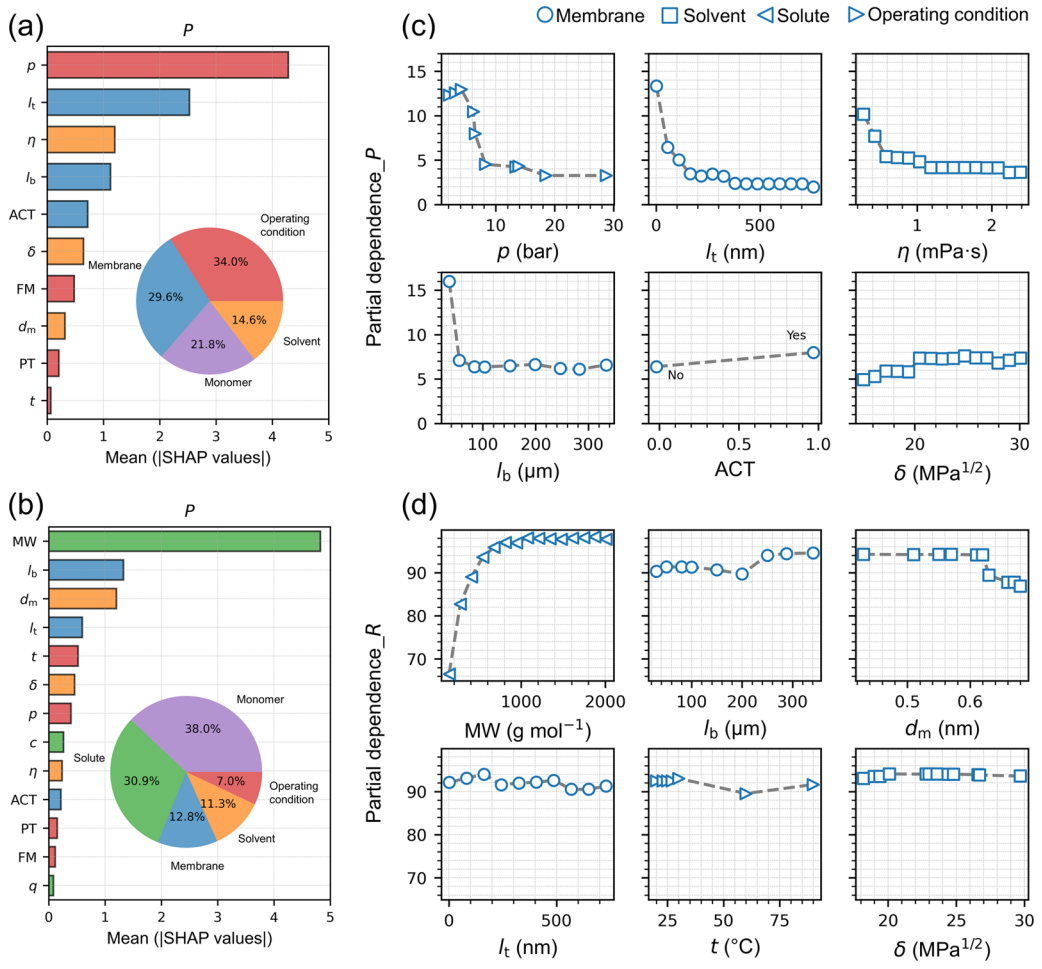


Figure 4. (a,b) Feature importance for *P* and *R*. (c,d) Partial dependence of *P* and *R* on the six most important features (monomer-related features are not included).

whereas 2 g·L^{-1} is often used for polystyrene (PS) as a standard solute in OSN to determine the molecular weight cutoff (MWCO) (Figure 2g). Most solutes are neutral ($q = 0$), though both positively and negatively charged solutes are also examined (Figure 2h). Regarding operating conditions, ambient temperature *t* of 25 °C is most common (Figure 2i) and transmembrane pressure *p* is usually ≤ 10 bar or around 30 bar (Figure 2j). In most cases, permeance *P* is ≤ 10 L·m^{-2}·h^{-1}·bar^{-1} (Figure 2k), while rejection *R* is ≥ 90% (Figure 2l).

Figure S2 shows the permeances *P* of different solvents and the rejections *R* of different solutes. For *P*, acetone, methanol, ethanol, DMF, and ethyl acetate (EA) exhibit a wide distribution, attributed to various membrane materials and operating conditions in different OSN experiments. This suggests that the performance of OSN is influenced by many factors. For *R*, PS and PEG have a broad distribution due to their varying MW. Meanwhile, rose bengal, crystal violet, methyl orange, and rhodamine B also exhibit a wide distribution of *R* as measured in various OSN membranes. Figure S3 shows the pie plots of ACT, flow mode (FM), and permeate type (PT). In the data set, 36.45% of membranes are activated. Dead end (64.14%) is a more common FM than cross-flow (35.86%). The majority of PT is a mixture of solvent and solute (80.4%), while 19.6% is pure solvent.

3.2. Performance of ML Models. In our ML models, five different molecular fingerprints were used to represent monomer structures including Frag, MACCS, MorganBit, MorganCount, and RDKitDP. As seen in Figure 3a,b, the five fingerprints generally have similar average MAE values for *P* and *R* predictions in the test set. Nevertheless, RDKitDP exhibits the lowest MAE in both *P* and *R* predictions. A detailed comparison with a minor tangent is found in Figure S4. Thus, the following results are based on the RDKitDP as the default molecular fingerprint unless otherwise stated. Figure 3c,d shows the ML predicted versus experimental *P* and *R*. In the test set, the MAE, RMSE, and R^2 between predicted and experimental *P* are 1.51, 3.03 L·m^{-2}·h^{-1}·bar^{-1}, and 0.95, respectively. Between predicted and experimental *R*, the MAE, RMSE, and R^2 are 3.61%, 8.48%, and 0.68. Compared to *P*, *R* prediction is less accurate because of the complex interactions of solute with solvent and membrane. To comprehensively evaluate the ML models, the MAE, RMSE, and R^2 using 5-fold cross-validation were also calculated and are shown in Table S2. Figure S5a,b depicts the influence of a random split of 80% (training)/20% (test) on *P* and *R* predictions. The median values of MAE and RMSE are 1.7 and 3.8 L·m^{-2}·h^{-1}·bar^{-1} in the test set for *P* prediction, while they are about 4% and 10% for *R* prediction. Figure S5c,d further shows the learning curve for the *P* and *R* predictions. With increasing size of the training set, the MAE for *P* prediction in the test set drops monotonically; for *R* prediction, the MAE initially drops, reaches a minimum at 80% training size, and then rises at 90% training size. Consequently, an 80%/20% split of training/test sets is appropriate. It is also observed that at a small size of

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

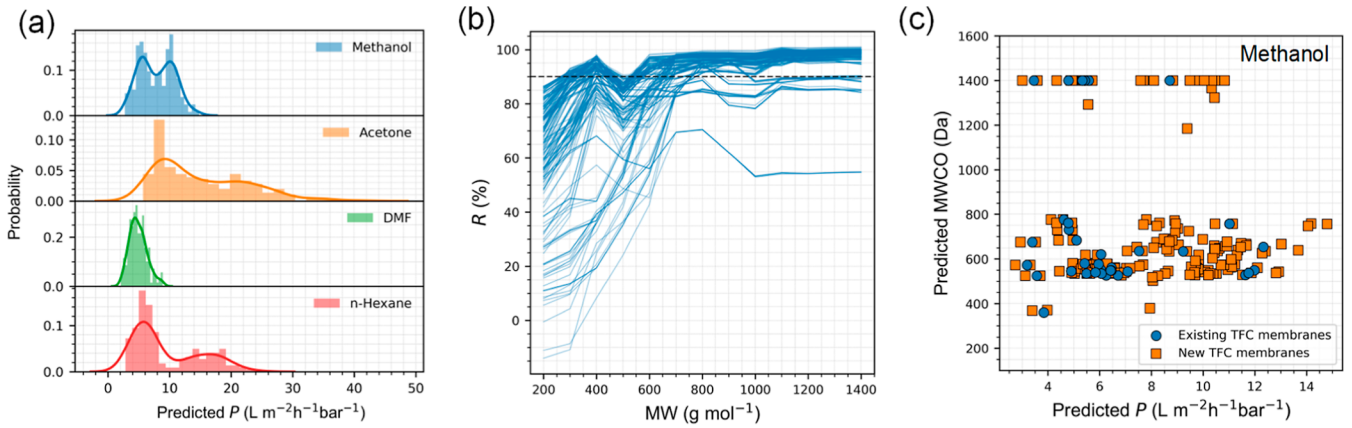


Figure 5. (a) Predicted P for methanol, acetone, DMF, and n-hexane in 167 new TFC membranes. (b) Predicted R in 167 new membranes, with PS as the solute and methanol as the solvent. Each line represents a membrane. (c) MWCO versus P of methanol in 167 new membranes and 37 existing membranes with PS as the solute and methanol as the solvent.

training set, the MAE remains almost constant for P prediction but rises for R prediction. Furthermore, we have incorporated a parity plot, as a reference, with a 60% (training)/40% (test) split (Figure S6). Notably, the accuracy of P prediction remains consistent, suggesting the absence of overfitting in the 80%/20% split. Meanwhile, a decline in R prediction accuracy is observed. This observation can be attributed to the intricate interactions among solute, solvent, and membrane, necessitating a broader diversity of data for accurate R prediction.

3.3. Insights from ML Models. To provide in-depth insights, we analyze how different features in the ML models contribute to P and R based on SHAP and partial dependence analysis. Figure 4a,b shows the feature importance as well as the normalized importance of each category of the features. The normalized feature importance was estimated by the following procedure. First, the SHAP value of each feature was calculated. Next, all the features were grouped into four categories: monomer, membrane, solvent, and operating condition. Finally, the sum of the SHAP values in each category was calculated and normalized. A more detailed normalized feature importance can be found in Figure S7. As indicated by the pie plots, the most important feature for P is operating condition with 34.0% contribution, followed by membrane (29.6%), monomer (21.8%), and solvent (14.6%). For R, monomer contributes the most, which accounts for 38.0% of the total importance; meanwhile, solute also exhibits a significant impact (30.9%) because R is directly related to solute. Other features are less important, with 12.8, 11.3, and 7% contributions by membrane, solvent, and operating condition, respectively.

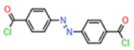
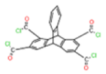
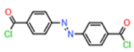
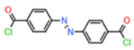
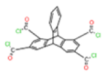
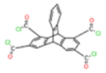
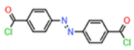
Furthermore, the partial dependence of P and R on the six most important features (excluding monomer) is examined. As shown in Figure 4c, p, l_t, η, l_b, ACT, and δ are the six important features of P. (i) When operating pressure p increases, P drops; because a polymeric membrane trends to be more compact under a higher pressure, thus leading to a lower permeance.[4] (ii) With increasing thickness of either selective layer l_t or support layer l_b, P drops, particularly in an ultrathin membrane. One benefit of TFC membranes is that the thin selective layer contributes to high permeance; moreover, the layer thickness can be controlled by many factors such as reaction time and monomer concentration.[34] (iii) P rises when solvent viscosity η decreases or solvent solubility parameter δ increases. Such a trend is well-acknowledged by a phenomenological model $P \propto \frac{\delta}{\eta d_m^2}$ derived from experiments.[35] (iv) ACT (from 0 to 1) has a positive effect on P. This is because activation removes residual solvent in a membrane, thus producing more transport pathways and enhancing P.[36]

Figure 4d shows the six most important features for R, including MW, l_b, d_m, l_t, t, and δ. (i) With increasing MW, R first rises considerably and then converges after MW ≈ 1000 g·mol^{-1}. Obviously, a solute with a larger MW is more easily rejected during OSN and could be completely rejected if the MW is sufficiently large. (ii) The thickness of the support layer l_b appears to have a positive effect on R, opposite to that on P, whereas the thickness of the selective layer l_t has a negligible effect. (iii) When the molecular diameter of solvent d_m becomes larger, R initially remains unchanged and drops at $d_m \approx 0.66$ nm; the drop in R is plausibly due to increasingly stronger interaction between solvent and membrane. It was reported that polymer chains in a membrane would rearrange in dimethylacetamide (DMAc, a solvent with a molecular diameter of 0.66 nm) and facilitate solute permeation (i.e., reduce R).[37] (iv) Operating temperature t appears to have a minor impact on R, and a drop in R is observed at ∼60 °C. This can be attributed to insufficient data in terms of t. (v) Solvent solubility parameter δ appears to have a negligible effect on R.

3.4. Membrane Design and Experimental Validation. As pointed out earlier, there were 34 A monomers and 6 B monomers in the data set, which could polymerize into a total of 204 polymers. Nevertheless, only a small number of 37 polymers existed in the data set for TFC membranes. To explore possible membranes for OSN, we in silico designed 167 new TFC membranes by combining 34 A monomers and 6 B monomers. In principle, the design methodology can be used to virtually design a wide variety of new polymers, given the chemical structures of A and B monomers in aqueous and organic phases. This is the key salient feature of incorporating chemical structures into ML in the current study. However, we should note that probably not all the 167 new membranes can be realized in practice. The feasibility depends on the reactivity and stoichiometry ratio of monomers at the aqueous/organic interface, as well as the solubility of monomers in aqueous and organic phases. For the 167 new membranes, we predicted their OSN performance (P and R) using the ML models developed for four common solvents (methanol, acetone,

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

Table 2. Top 10 Membranes with the Highest Predicted *P* for Methanol

No.	Aqueous monomer	Organic monomer	Existing	Predicted *P* ($L \cdot m^{-2} \cdot h^{-1} \cdot bar^{-1}$)	Predicted MWCO (Da)
1			No	14.8	756
2			No	14.2	758
3			No	14.1	747
4			No	13.7	639
5			No	13.0	667
6			No	12.9	543
7			No	12.8	537
8			No	12.4	634
9			Yes	12.3	654
10			No	12.3	626

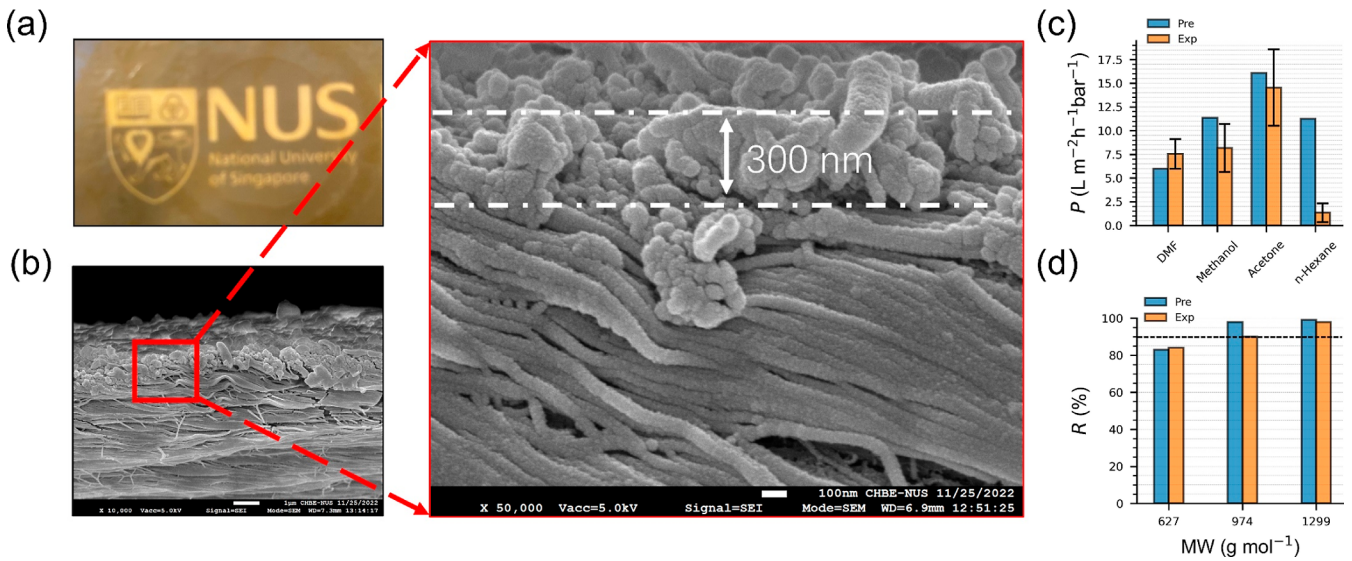


Figure 6. (a,b) PIP-AZ membrane with its cross-section morphology. The NUS logo is a copyright of the National University of Singapore. (c) Permeances of four solvents (DMF, methanol, acetone, and *n*-hexane). (d) Rejections of three dyes (remazol brilliant blue, rose bengal, and alcian blue with molecular weights of 627, 974, and 1299 $g \cdot mol^{-1}$) in methanol. The dashed line indicates 90% rejection.

DMF, and *n*-hexane). PS was used as a model solute with a concentration of 2 $g \cdot L^{-1}$ and MW ranging from 200 to 1400 $g \cdot mol^{-1}$. Based on earlier analysis (Figure 2), l_t and l_b were set as 150 nm and 200 μm, respectively, which are approximately their median values in the experimental data set. The membranes were supposed to be activated, and the flow mode was a dead-end configuration. The operating conditions, *t* and *p*, were taken as 25 °C and 10 bar, according to a recently proposed standard OSN protocol.[38]

Figure 5a shows the predicted *P* for methanol, acetone, DMF, and *n*-hexane in 167 new TFC membranes. The range of *P* varies with solvent: 3−15 $L \cdot m^{-2} \cdot h^{-1} \cdot bar^{-1}$ for methanol,

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

6−41 $L{\cdot}m^{-2}{\cdot}h^{-1}{\cdot}bar^{-1}$ for acetone, 2−9 $L{\cdot}m^{-2}{\cdot}h^{-1}{\cdot}bar^{-1}$ for DMF, and 3−25 $L{\cdot}m^{-2}{\cdot}h^{-1}{\cdot}bar^{-1}$ for *n*-hexane. The predicted *R* of PS with methanol as the solvent are illustrated in Figure 5b. With increasing MW, *R* apparently rises, as seen earlier from the partial dependence analysis. When the MW of PS is >700 $g{\cdot}mol^{-1}$, *R* approaches 90% in most of the 167 new membranes. From Figure 5b, the MWCO is estimated to be *R* = 90%. For membranes with predicted *R* always <90%, the MWCO is considered >1400 $g{\cdot}mol^{-1}$. Figure 5c shows the MWCO versus *P* in 167 new membranes and 37 existing membranes with PS as the solute and methanol as the solvent. The maximum *P* in the existing membranes is ∼12 $L{\cdot}m^{-2}{\cdot}h^{-1}{\cdot}bar^{-1}$, while several of the 167 new membranes are predicted to possess higher *P*. In both new and existing membranes, the MWCO is populated mostly in the range of 500−800 Da, with a few <400 Da and between 1300 and 1400 Da.

Table 2 lists the top 10 membranes out of the entire 204 membranes (including 167 new and 37 existing) with the highest predicted *P* for methanol. All the predictions in 204 membranes can be found on GitHub. Only one of the 10 top membranes (no. 9) existed in the experimental data set, while the other nine are newly designed in this study. In terms of *P*, eight new membranes are predicted to outperform all the existing membranes in the data set. Furthermore, it is observed that most of the monomers in the aqueous phase of the top membranes are nitrogen-containing heterocyclic. This is consistent with the fact that introducing nitrogen-containing heterocyclic into membranes can enhance microporosity and permeance.[39]

To validate the ML predictions, considering the accessibility of monomers, we synthesized a TFC membrane (PIP-AZ) comprising PIP and AZ, listed as no. 4 in Table 2. The PIP-AZ membrane from interfacial polymerization was confirmed with a color change from white of a PE support to orange, indicating the successful formation of a PIP-AZ selective layer (Figure 6a). The layer thickness was estimated to be 300 nm, as visualized by the FESEM image (Figure 6b). After activation, the performance of the PIP-AZ membrane was tested by measuring the permeances of four solvents (methanol, acetone, DMF, and *n*-hexane) and the rejections of three dyes (remazol brilliant blue, rose bengal, and alcian blue) in methanol. Meanwhile, the developed ML models were used to predict *P* and *R* in the PIP-AZ membrane with a layer thickness of 300 nm. Fairly good agreement is observed in Figure 6c between the predicted and experimental permeances of three solvents (DMF, methanol, and acetone). However, the permeance of *n*-hexane is overpredicted, which is because of the sparse data points in the experimental data set for *n*-hexane, thus causing inaccurate prediction. With more experimental data available in the future, the prediction can be improved. As shown in Figure 6d, the predicted rejections of all three dyes match rather well with the experimental data. With increasing MW of the dye molecule, the rejection rises. Based on 90% rejection, the MWCO is estimated to be around 1000 $g{\cdot}mol^{-1}$ in the PIP-AZ membrane.

4. ENVIRONMENTAL IMPLICATIONS

In this study, we aim to design TFC membranes for solvent recovery by using the ML approach. The chemical structures of monomers, membrane, solvent and solute properties, as well as operating conditions, are used as input features, while solvent permeance *P* and solute rejection *R* are employed as targets. High accuracy is achieved for the ML predicted *P* and *R*. Feature importance analysis reveals that operating conditions contribute mostly to *P*, while the molecular weight of solute dominates *R*. From 40 monomers in the experimental data set, we design 167 new TFC membranes and apply the ML models to predict their OSN performance for four solvents including methanol, acetone, DMF, and *n*-hexane. Top-performing new membranes are identified and found to have methanol *P* values higher than those of existing membranes.

With an increasingly greater focus on sustainability and stricter environmental regulations, solvent recovery has become more important, particularly in the pharmaceutical industry. There has been considerable interest in developing OSN membranes for solvent recovery. At present, OSN membranes are produced largely based on empirical trial-and-error methods. Existing ML studies in the field of OSN focus on the effects of solute and solvent properties, as well as operation conditions; hence, a design strategy is lacking for OSN membranes. By embedding chemical structures into our ML models, we achieve the rational design of new OSN membranes. The newly designed top-performing membranes are potential candidates for solvent recovery. Given the availability of such membranes, it might be feasible to develop membrane modules and configurations to efficiently recover solvents. Furthermore, it should be underlined that the ML-assisted design methodology, based on the chemical structures of monomers, provides a bottom-up strategy and can be readily extended to the design of new membranes for a wide variety of other important applications, e.g., gas separation, water desalination, and biofuel purification.

■ ASSOCIATED CONTENT

Data Availability Statement

Data and codes related to this study can be found on GitHub (https://github.com/maowang-code/machine-learning-for-solvent-recovery).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.3c04773.

> Different types of monomers, solvents, and solutes in the experimental data set; permeances of different solvents and rejections of different solutes; hyperparameters used in ML models; average MAE values for *P* and *R* predictions; ML predicted versus experimental *P* and *R* with 60% (training)/40% (test) split; and feature importance for *P* and *R* (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jianwen Jiang − *Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576, Singapore;* orcid.org/0000-0003-1310-9024; Email: chejj@nus.edu.sg

Authors

Mao Wang − *Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576, Singapore;* orcid.org/0000-0001-7583-1683

Gui Min Shi − *Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576, Singapore*

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924

Daohui Zhao − *Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576, Singapore*

Xinyi Liu − *Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore 117576, Singapore*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.3c04773

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the A*STAR under AME IRG grant (A20E5c0092), the Ministry of Education of Singapore, and the National University of Singapore (R-279-000-598-114 and R-279-000-574-114) for financial support.

■ REFERENCES

(1) Clarke, C. J.; Tu, W. C.; Levers, O.; Brohl, A.; Hallett, J. P. Green and Sustainable Solvents in Chemical Processes. *Chem. Rev.* **2018**, *118*, 747−800.

(2) Clark, J. H.; Farmer, T. J.; Hunt, A. J.; Sherwood, J. Opportunities for Bio-Based Solvents Created as Petrochemical and Fuel Products Transition Towards Renewable Resources. *Int. J. Mol. Sci.* **2015**, *16*, 17101−17159.

(3) Jiménez-González, C.; Poechlauer, P.; Broxterman, Q. B.; Yang, B.-S.; am Ende, D.; Baird, J.; Bertsch, C.; Hannah, R. E.; Dell'Orco, P.; Noorman, H.; Yee, S.; Reintjens, R.; Wells, A.; Massonneau, V.; Manley, J. Key Green Engineering Research Areas for Sustainable Manufacturing: A Perspective from Pharmaceutical and Fine Chemicals Manufacturers. *Org. Process Res. Dev.* **2011**, *15*, 900−911.

(4) Marchetti, P.; Jimenez Solomon, M. F.; Szekely, G.; Livingston, A. G. Molecular Separation with Organic Solvent Nanofiltration: A Critical Review. *Chem. Rev.* **2014**, *114*, 10735−10806.

(5) Shi, G. M.; Feng, Y.; Li, B.; Tham, H. M.; Lai, J.-Y.; Chung, T.-S. Recent Progress of Organic Solvent Nanofiltration Membranes. *Prog. Polym. Sci.* **2021**, *123*, 101470.

(6) Li, Y.; Guo, Z.; Li, S.; Van der Bruggen, B. Interfacially Polymerized Thin-Film Composite Membranes for Organic Solvent Nanofiltration. *Adv. Mater. Interfaces* **2021**, *8*, 2001671.

(7) Lee, K. P.; Arnot, T. C.; Mattia, D. A Review of Reverse Osmosis Membrane Materials for Desalination − Development to Date and Future Potential. *J. Membr. Sci.* **2011**, *370*, 1−22.

(8) Raaijmakers, M. J. T.; Benes, N. E. Current Trends in Interfacial Polymerization Chemistry. *Prog. Polym. Sci.* **2016**, *63*, 86−142.

(9) Cao, X.; Guo, J.; Cai, J.; Liu, M.; Japip, S.; Xing, W.; Sun, S. The Encouraging Improvement of Polyamide Nanofiltration Membrane by Cucurbituril-Based Host−Guest Chemistry. *AIChE J.* **2020**, *66*, No. e16879.

(10) Ali, Z.; Ghanem, B. S.; Wang, Y. G.; Pacheco, F.; Ogieglo, W.; Vovusha, H.; Genduso, G.; Schwingenschlogl, U.; Han, Y.; Pinnau, I. Finely Tuned Submicroporous Thin-Film Molecular Sieve Membranes for Highly Efficient Fluid Separations. *Adv. Mater.* **2020**, *32*, 2001132.

(11) Huang, T. F.; Puspasari, T.; Nunes, S. P.; Peinemann, K. V. Ultrathin 2D-Layered Cyclodextrin Membranes for High-Performance Organic Solvent Nanofiltration. *Adv. Funct. Mater.* **2020**, *30*, 1906797.

(12) 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*, No. eaaz4301.

(13) Gao, H.; Zhong, S.; Zhang, W.; Igou, T.; Berger, E.; Reid, E.; Zhao, Y.; Lambeth, D.; Gan, L.; Afolabi, M. A.; et al. Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization. *Environ. Sci. Technol.* **2021**, *56*, 2572−2581.

(14) Wang, M.; Xu, Q. S.; Tang, H. J.; Jiang, J. W. Machine Learning-Enabled Prediction and High-Throughput Screening of Polymer Membranes for Pervaporation Separation. *ACS Appl. Mater. Interfaces* **2022**, *14*, 8427−8436.

(15) Wang, M.; Jiang, J. W. Accelerating Discovery of High Fractional Free Volume Polymers from a Data-Driven Approach. *ACS Appl. Mater. Interfaces* **2022**, *14*, 31203−31215.

(16) Xu, Q.; Jiang, J. W. Recent Development in Machine Learning of Polymer Membranes for Liquid Separation. *Mol. Syst. Des. Eng.* **2022**, *7*, 856−872.

(17) Goebel, R.; Skiborowski, M. Machine-Based Learning of Predictive Models in Organic Solvent Nanofiltration: Pure and Mixed Solvent Flux. *Sep. Purif. Technol.* **2020**, *237*, 116363.

(18) Goebel, R.; Glaser, T.; Skiborowski, M. Machine-Based Learning of Predictive Models in Organic Solvent Nanofiltration: Solute Rejection in Pure and Mixed Solvents. *Sep. Purif. Technol.* **2020**, *248*, 117046.

(19) Hu, J.; Kim, C.; Halasz, P.; Kim, J. F.; Kim, J.; Szekely, G. Artificial Intelligence for Performance Prediction of Organic Solvent Nanofiltration Membranes. *J. Membr. Sci.* **2021**, *619*, 118513.

(20) Ignacz, G.; Szekely, G. Deep Learning Meets Quantitative Structure−Activity Relationship (QSAR) for Leveraging Structure-Based Prediction of Solute Rejection in Organic Solvent Nanofiltration. *J. Membr. Sci.* **2022**, *646*, 120268.

(21) Wang, C.; Wang, L.; Soo, A.; Bansidhar Pathak, N.; Kyong Shon, H. Machine Learning Based Prediction and Optimization of Thin Film Nanocomposite Membranes for Organic Solvent Nanofiltration. *Sep. Purif. Technol.* **2023**, *304*, 122328.

(22) Ignacz, G.; Alqadhi, N.; Szekely, G. Explainable Machine Learning for Unraveling Solvent Effects in Polyimide Organic Solvent Nanofiltration Membranes. *Adv. Membr.* **2023**, *3*, 100061.

(23) Lee, Y. J.; Chen, L.; Nistane, J.; Jang, H. Y.; Weber, D. J.; Scott, J. K.; Rangnekar, N. D.; Marshall, B. D.; Li, W.; Johnson, J. R.; Bruno, N. C.; Finn, M. G.; Ramprasad, R.; Lively, R. P. Data-Driven Predictions of Complex Organic Mixture Permeation in Polymer Membranes. *Nat. Commun.* **2023**, *14*, 4931.

(24) Xu, Q.; Gao, J.; Feng, F.; Chung, T. S.; Jiang, J. W. Synergizing Machine Learning, Molecular Simulation and Experiment to Develop Polymer Membranes for Solvent Recovery. *J. Membr. Sci.* **2023**, *678*, 121678.

(25) Rohatgi, A. *WebPlotDigitizer*. https://automeris.io/WebPlotDigitizer (accessed December 20, 2022).

(26) Weininger, D. SMILES: A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(27) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(28) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(29) Landrum, G. *RDKit: Open-Source Cheminformatics Software*. http://www.rdkit.org/(accessed February 10, 2023).

(30) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189−1232.

(31) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(32) Yang, M.; Zhu, J. J.; McGaughey, A.; Zheng, S.; Priestley, R. D.; Ren, Z. J. Predicting Extraction Selectivity of Acetic Acid in Pervaporation by Machine Learning Models with Data Leakage Management. *Environ. Sci. Technol.* **2023**, *57*, 5934−5946.

(33) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st international conference on neural information processing systems*, 2017; pp 4768−4777.

(34) Gorgojo, P.; Karan, S.; Wong, H. C.; Jimenez-Solomon, M. F.; Cabral, J. T.; Livingston, A. G. Ultrathin Polymer Films with Intrinsic Microporosity: Anomalous Solvent Permeation and High Flux Membranes. *Adv. Funct. Mater.* **2014**, *24*, 4729−4737.

(35) Karan, S.; Jiang, Z.; Livingston, A. G. Sub-10 nm Polyamide Nanofilms with Ultrafast Solvent Transport for Molecular Separation. *Science* **2015**, *348*, 1347−1351.
(36) Jimenez Solomon, M. F.; Bhole, Y.; Livingston, A. G. High flux membranes for organic solvent nanofiltration (OSN)—Interfacial polymerization with solvent activation. *J. Membr. Sci.* **2012**, *423*−*424*, 371−382.
(37) Razali, M.; Didaskalou, C.; Kim, J. F.; Babaei, M.; Drioli, E.; Lee, Y. M.; Szekely, G. Exploring and Exploiting the Effect of Solvent Treatment in Membrane Separations. *ACS Appl. Mater. Interfaces* **2017**, *9*, 11279−11289.
(38) Le Phuong, H. A.; Blanford, C. F.; Szekely, G. Reporting the Unreported: The Reliability and Comparability of the Literature on Organic Solvent Nanofiltration. *Green Chem.* **2020**, *22*, 3397−3409.
(39) Ren, D.; Li, Y.-h.; Ren, S.-p.; Liu, T.-Y.; Wang, X.-L. Microporous Polyarylate Membrane with Nitrogen-Containing Heterocycles to Enhance Separation Performance for Organic Solvent Nanofiltration. *J. Membr. Sci.* **2020**, *610*, 118295.

https://doi.org/10.1021/acs.est.3c04773
Environ. Sci. Technol. 2023, 57, 15914−15924