Environmental Science & Technology

pubs.acs.org/est

Article

Machine Learning for Polymer Design to Enhance Pervaporation-Based Organic Recovery

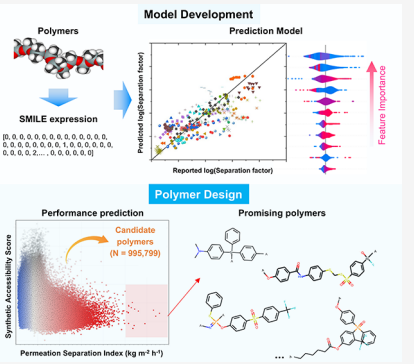
Meiqi Yang, Jun-Jie Zhu, Allyson L. McGaughey, Rodney D. Priestley, Eric M. V. Hoek, David Jassby, and Zhiyong Jason Ren*

Cite This: *Environ. Sci. Technol.* 2024, 58, 10128−10139

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Pervaporation (PV) is an effective membrane separation process for organic dehydration, recovery, and upgrading. However, it is crucial to improve membrane materials beyond the current permeability-selectivity trade-off. In this research, we introduce machine learning (ML) models to identify high-potential polymers, greatly improving the efficiency and reducing cost compared to conventional trial-and-error approach. We utilized the largest PV data set to date and incorporated polymer fingerprints and features, including membrane structure, operating conditions, and solute properties. Dimensionality reduction, missing data treatment, seed randomness, and data leakage management were employed to ensure model robustness. The optimized LightGBM models achieved RMSE of 0.447 and 0.360 for separation factor and total flux, respectively (logarithmic scale). Screening approximately 1 million hypothetical polymers with ML models resulted in identifying polymers with a predicted permeation separation index >30 and synthetic accessibility score <3.7 for acetic acid extraction. This study demonstrates the promise of ML to accelerate tailored membrane designs.



KEYWORDS: *pervaporation, machine learning, LightGBM, SHAP, data leakage management, membrane, wastewater*

1. INTRODUCTION

Pervaporation (PV) is widely used for liquid mixture separation with high product purity, scalability, and high energy efficiency compared to traditional methods like distillation, adsorption, and precipitation.[1,2] PV has gained significant traction in recent years and has been used in solvent recovery, food processing, pharmaceutics, and desalination.[3] More recently, there has been a growing interest in utilizing PV for the separation of low-concentration organic compounds, such as acetic acid, ethanol, and isopropanol, from aqueous solutions.[4] PV relies on a semipermeable membrane to separate the feed and permeate streams. The separation mechanism involves the adsorption of permeating components onto the membrane surface on the feed side, followed by diffusion through the membrane and condensation into the permeate stream.[5] The separation performance is closely tied to the properties of the membrane material, with polymeric membranes being the most commonly investigated due to their low cost and ease of fabrication.

Hydrophilic PV membranes (e.g., poly(vinyl alcohol) (PVA) and sodium alginate (NaAlg)) are used for organic dehydration;[6−8] less commonly, hydrophobic membranes (e.g., poly(dimethylsiloxane) (PDMS) and poly(1-trimethylsilyl-1-propyne) (PTMSP)) are employed for extracting organics from aqueous streams.[9−11] Although membrane modification methods like chemical modification, cross-linking, and incorporating fillers have been implemented to enhance PV performance, the development process still heavily relies on slow and costly empirical methods.[5,12] Alternatively, prediction models have attracted significant interest, as they have the potential to minimize exhaustive experimental investigations and offer insights into the influence of features on performance.

PV performance prediction has traditionally relied on theoretical models and simulation methods, such as the pore-flow model, solution-diffusion model, density functional theory, and molecular dynamics, which have limited applicability and are computationally intensive.[13−15] Machine learning (ML) has emerged as a data-driven approach to solve complex, multivariate problems with high computational efficiency.[16,17] ML models have proved to be promising to facilitate the discovery of new materials and achieve the targeted design of materials.[18−22] ML models have been successfully employed to predict polymer properties (e.g., glass transition temperature, thermal conductivity) based on polymer structure.[23−27] In addition, ML models showed great potential to reveal the relationship between materials'

Received: January 29, 2024
Revised: April 28, 2024
Accepted: April 30, 2024
Published: May 14, 2024

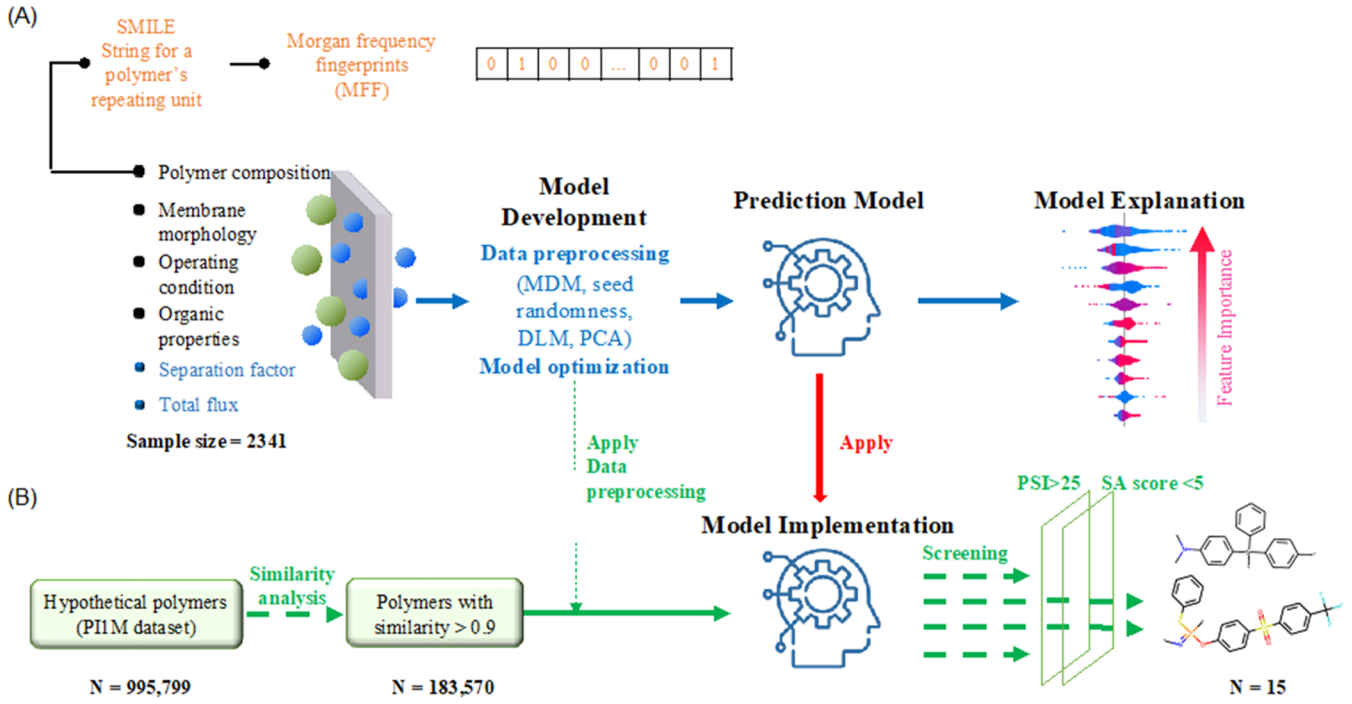


Figure 1. Workflow of ML-model-assisted polymer screening. (A) Prediction model development using the MFF: Generation of MFF from the simplified molecular input line system (SMILE) expression of polymers' repeating units. For model development, different model processing methods are involved, including missing data management (MDM), DLM, and PCA. (B) The developed ML models are then implemented for high-throughput screening of hypothetical polymers in the PI1M data set (N = 995,799) with promising acetic acid extraction performance; the evaluation metrics include similarity score, PSI, and SA score.

structure and their properties and could assist the design of membranes for different processes, but their use in PV membrane design is limited.[28−30] Previous studies linked 15 representative chemical functional groups with PV membrane performance, but the impacts were limited due to the small sample size (681 samples) compared to the total number of samples in the literature.[31] Using only 15 chemical functional groups to describe a wide range of polymers runs into high risks of missing topological features, and lack of seed randomness and data leakage management (DLM) also compromise the model's robustness and accuracy. Our previous ML study addressed some of these limitations for acetic acid/water separation, and we revealed that the mass ratio was the most significant parameter to predict separation selectivity.[32] Still, current models lack critical functions such as separation performance prediction and material design.

In this study, we collected and curated the largest database to date for PV through literature mining, and we then developed a robust ML model for the first time that can facilitate the screening of polymers for PV membranes. The model can greatly advance PV membrane's manufacturing and application. The workflow of ML-assisted polymer discovery is outlined in Figure 1. Polymers' chemistry was described using Hashed Morgan frequency fingerprints (MFFs) and reconstructed via principal component analysis (PCA). The performance of various ML algorithms was evaluated, and the light gradient-boosting machine regression (LGBMR) models demonstrated the highest prediction accuracy on the testing data set (logarithmic scale). For the separation factor, the LGBMR models achieved R^2 and RMSE values of 0.61 and 0.45, respectively, and for total flux, the LGBMR models achieved R^2 and RMSE values of 0.40 and 0.36, respectively. Shapley additive explanations (SHAP) were employed to elucidate feature importance at the atomic level, revealing the most relevant structures for enhancing membrane selectivity. Furthermore, hypothetical polymers from the PI1M data set (where PI1M refers to 1 million polymers from the Polymer Informatics database) were first screened based on their similarity to the studied polymers in the ML model database. Out of 995,799 polymers, 183,570 were identified as potential candidates, based on their high similarity with the studied polymers from the data set. Applying the developed ML models to the PI1M data set, we identified 88,363 polymers with promising potential for acetic acid extraction from water, which are expected to surpass the current upper limit (i.e., permeation separation index, PSI > 4.36). Finally, the viability of fabricating the screened polymers was assessed using a synthetic accessibility (SA) score.

2. METHODS

2.1. Data Collection. The sample sizes of the two data sets are summarized in Table S1. The PV data set was established based on literature data using data mining-assisted data collection, details in Text S1.[33] The PV data set consisted of 2341 data points (Supporting Excel), with 52 unique polymers and 32 types of organic solutes. The PI1M data set included 995,799 hypothetical polymers learned via an RNN trained on SMILES strings of existing polymers in PoLyInfo, as constructed by Ma and Luo.[34] These 995,799 hypothetical polymers have been widely employed for ML-assist polymer discovery.[35] The SMILE of polymers from both the collected data set and PI1M data set includes polymerization points and the bonding information between monomers using "*".

2.2. Fingerprint Generation and Process Description. Features related to polymer composition, membrane morphology, operating conditions, and solute properties were used to

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

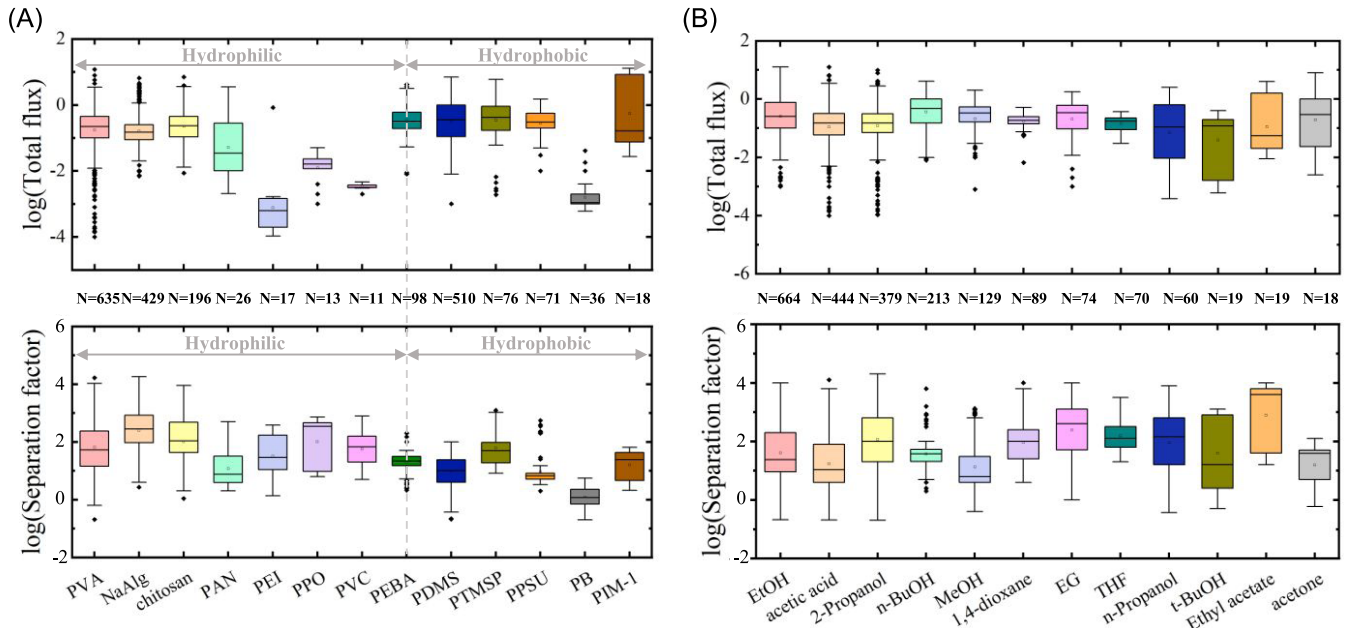


Figure 2. Summary of PV performance based on different (A) polymers and (B) organic solutes. N in the plot represents the sample size for each category. EtOH stands for ethanol, *n*-BuOH stands for *n*-butanol, MeOH stands for methanol, EG stands for ethylene glycol, and THF stands for Tetrahydrofuran. The box represents the interquartile range (IQR) of the data, which is the range between the first quartile (Q1) and the third quartile (Q3). The black line inside the box represents the mean value. The whiskers extend from the edges of the box to the minimum and maximum data points.

develop the prediction models for separation factor and total flux, separately. Molecular fingerprints as numerical representations of molecules are widely used in cheminformatics and drug discovery, with high flexibility for new or hypothetical chemical structures.[19] The MFF is a type of circular fingerprint that not only captures the substructures present in a molecule but also includes overall topological features (details in Text S2). MFFs for polymers were generated (highlighted in orange). In a fingerprint vector, each bit represents certain substructures, and the number of occurrences of each substructure is calculated from the substructure frequency.[36] Taking PDMS as an example (Figure S1), the feature at bit 22 corresponds to substructure (0,2), where the center (Si) atom was labeled as 0 (blue circle) and 2 represents the substructure cropped with a radius of 2 (solid line). In this work, all polymer MFFs were generated with a maximum radius of 2, and bit length = 1024.

Twelve key input features were selected based on domain knowledge, including selective layer thickness (l), filler size, filler concentration, cross-linker concentration, experimental temperature (T), downstream pressure (P), mass ratio between components A and B ($\alpha_{A/B}$), effluent type, effective area (A), organics' Hildebrand solubility parameter (δ), organics' dielectric constant (ε), and organics' molar volume (V). Detailed descriptions of these input features are included in Text S3 and Tables S2 and S3. We selected two metrics as output features to evaluate the PV performance, namely, total flux J (kg m^{-2} h^{-1}) and separation factor $\beta_{B/A}$. For a given feed stream composition, a membrane with a higher $\beta_{B/A}$ indicates better selectivity. We also included PSI to represent the PV performance affected by the trade-off between J and $\beta_{B/A}$, while PSI = $(\beta_{B/A} - 1) \times J$.

2.3. Data Preprocessing. To obtain robust results, we followed rigorous data preprocessing and model development recommendations outlined in Zhu et al.[37] Collected data were reprocessed via removal, replacement, filling, dimension reduction, and feature scaling before and during the modeling; details can be found in Text S4.[32]

2.4. Model Development and Optimization. We followed the procedures introduced in our previous work (Yang et al.) to develop a more robust DLM-based model. For all modeling in this work, data points from the same study were bundled into the same subset (e.g., training, testing, pretraining, and validation) to avoid data leakage. We first assessed the variance of a reference model based on 101 seeds (from 0 to 1000 with an interval of 10) because the choice of seed could lead to different data splitting, resulting in a huge discrepancy in the final prediction performance. We selected a representative seed to generate relatively average results based on the overall accuracy index (OAI) (eq 1), which represents overall model performance from R^2, RMSE, and MAE; the details of calculating OAI are shown in Text S5.

$$\text{OAI} = \frac{R^2}{\text{MAE} \times \text{RMSE}} \tag{1}$$

The reference model we used was CatBoost regression (CBR) (number of iterations = 300), the optimal model identified by Yang et al.[32] In addition to CBR, extra trees regression (ETR), random forest regression (RFR), and LGBMR were applied to compare the performance. LGBMR was selected due to its outperforming computational speed and memory consumption (details in Text S6).[38] Second, by fixing the seed, we split the data into training and testing (18−22%) subsets, and the training subset was further divided into 3 predefined folds based on data distribution and DLM to minimize a serious data imbalance. Third, we applied cross-validation and grid search using the 3-fold predefined training data to search the optimal method and model structure. The detailed information for the hyperparameters in the grid search is explained and summarized in Text S6 and Table S6. Finally, once the optimal

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

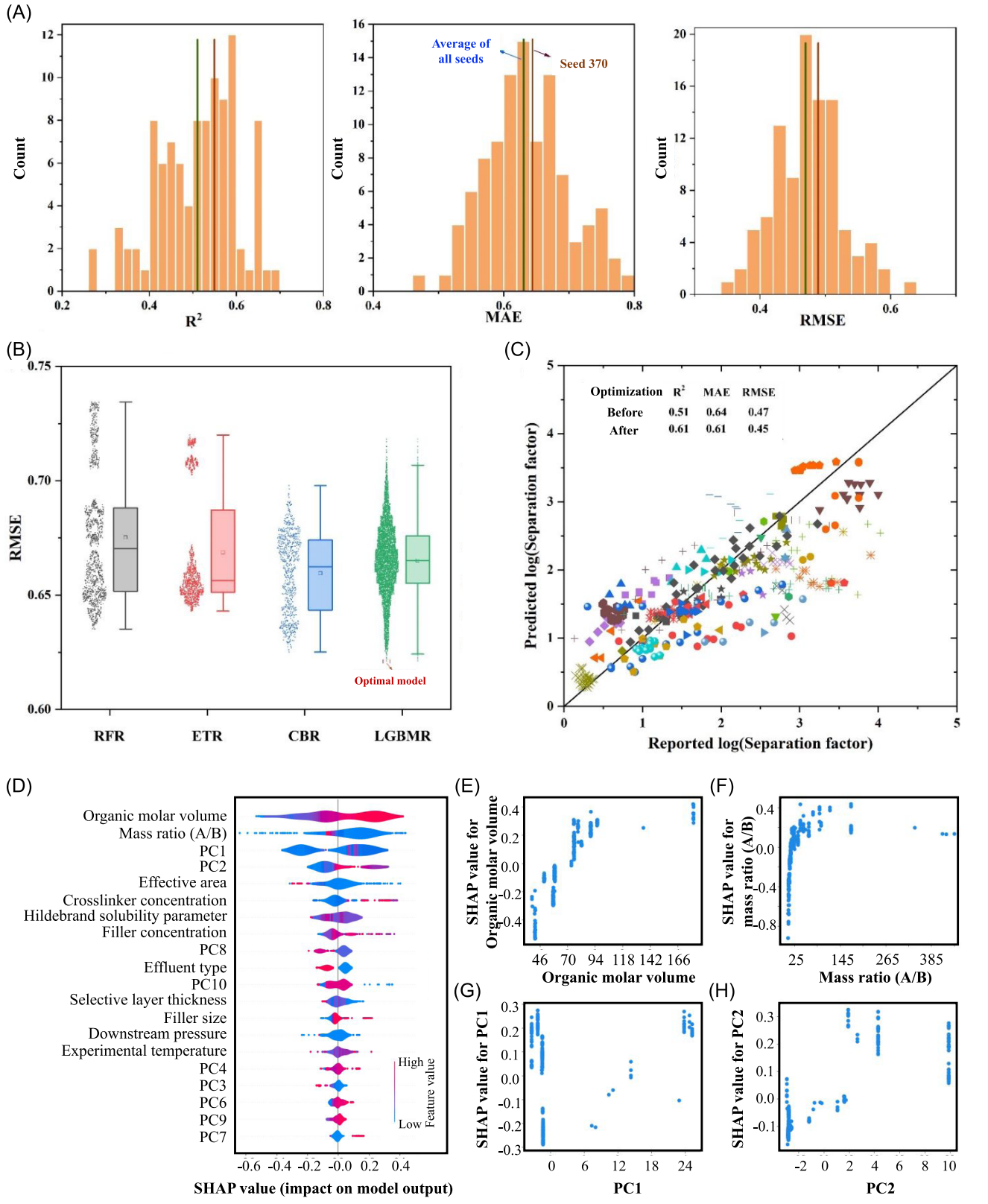


Figure 3. Model development and prediction results for the separation factor. (A) Primary seeds assessment using R^2, MAE, and RMSE based on the testing data set using CBR. (B) Box plots showing performance (RMSE) comparison between RFR, ETR, CBR, and LGBMR using DLM-based data segmentation and a predefined CV approach based on the training data set (seed 370). (C) Comparison between reported and predicted separation factor (logarithmic) data based on the testing data set using the optimal LGBMR model; note that data from the same (anonymous) study are shown in the same mark. Prediction performance using the testing data set before and after optimization is also displayed. (D–H) Model interpretation by analyzing the contribution of features using SHAP for separation factor prediction. SHAP summary plot of (D) features' contribution, (E) organics' molar volume, (F) mass ratio, (G) PC1, and (H) PC2.

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

model was identified, we deployed the model structure to the entire training data set to update the model parameters and then assessed the model generalization using the testing data set. The above procedures were used to determine optimal models.

2.5. Model Explanation. In this work, SHAP was employed to explain the established model by analyzing the contribution between different features, where the importance of each feature is calculated into the output value (Text S7).[39] Specifically, the Shapley value was retrained with all possible features' subsets, and the model prediction variance was calculated. To further unveil the contribution from specific chemical structures for the important PC identified from SHAP analysis, we also uncovered the original fingerprints and traced the fingerprints back to the chemical substructures.

2.6. Potential Polymer Screening. The developed ML models were then implemented in the PI1M data set to screen promising polymers for acetic acid extraction due to the industrial need for efficient acetic acid/water separations and interest in resource recovery.[40] MFFs of polymers in the PI1M data set were calculated with the same fingerprint vector. To improve the feasibility of polymer synthesis and to facilitate the understanding of the effect of chemical composition on separation performance, we first screened 995,799 polymers based on their similarity (i.e., similarity score ≥0.9) to the 52 studied polymers (Text S8). In addition, PCA obtained from model training was applied to the candidate polymers' MFF. Other input features were replaced with their mean values in the PV data set. The two best models established in Section 2.4 were then applied to the previously selected hypothetical polymers, and their predicted values (i.e., $\beta_{B/A}$ and J) were obtained. Their PSIs were calculated and acted as evaluation metrics for the polymer screening.

In addition to performance, we used the SA score as an indicator of the economic and practical feasibility of using hypothetical polymers to further screen out targeted polymers.[41] SA score is calculated based on a combination of fragment contributions and complexity penalties, which can quantitatively indicate the difficulty of polymer synthesis (details in Text S9).[42] As a result, the final polymer candidates are predicted to be more advantageous in terms of both separation performance and synthetic feasibility.

3. RESULT AND DISCUSSION

3.1. Data Sets and Data Analysis. The PV performance ($\beta_{B/A}$ and J) for the PV data set is summarized in Figure S2. The PSI for organic dehydration has a maximum value of 12,723.80 and an average of 105.96, while the PSI for organic extraction has a maximum value of 1285 and an average of 87.20. In addition, Figure S3 provides the distributions of selected key features, where all followed the normal distribution. Figure S4 summarizes the extensively studied polymers and organic solutes from the PV data set. Figure 2A displays histograms of the 13 most investigated polymers, with their corresponding $\log(\beta_{B/A})$ and $\log(J)$ distributions. Table S7 summarizes the sample sizes for a total of 52 polymer types. Among these 52 polymers, PVA, NaAlg, and chitosan are the most extensively studied hydrophilic polymers. Notably, PVA emerges as the most frequently used polymer overall, accounting for almost 25% of the data points in the database. Among hydrophilic polymers, PPO exhibits the highest median value of $\log(\beta_{B/A})$, followed by NaAlg. PEBA membrane is widely investigated for organic extraction (e.g., phenol and isopropyl alcohol) and exhibits good selectivity for high boiling bioproducts, and it exhibits a favorable median value of $\log(J)$. Among hydrophobic polymers, PDMS received the most attention, followed by PTMSP and PPSU. Interestingly, PTMSP demonstrates better selectivity of organics from H_2O (higher $\log(\beta_{B/A})$) compared to other hydrophobic polymers and the highest average $\log(J)$. The density of the 1024 fingerprint bits for the 52 studied polymers is shown in Figure S5.

Figure 2B illustrates the 12 most extensively studied organic solutes in the collected data set. Detailed descriptions of the sample sizes for all 32 types of organics are summarized in Table S8. Among the organic solutes, ethanol (EtOH), acetic acid, and 2-propanol (isopropanol) emerged as the most studied, followed by *n*-butanol (*n*-BuOH) and methanol (MeOH). It is worth noting that these organic solutes possess boiling points similar to those of H_2O and can form azeotropic mixtures, making traditional distillation methods ineffective. Interestingly, the separation of ethylene glycol and 1,4-dioxane from H_2O appears to achieve both a satisfactory total flux and relatively good selectivity. This outcome can likely be attributed to the significant differences in liquid properties compared with H_2O, thus enabling effective separation.

The literature data analysis did not reveal significant correlations between the selected input features (experimental temperature, mass ratio, and selective layer thickness) and the separation performance indicators (Figure S6). Furthermore, according to the Pearson correlation coefficient (r) and Spearman correlation coefficient (r_s) analysis (Figures S7–S8), the organics' dielectric constant exhibited a high correlation with the organics' Hildebrand solubility parameter. To prevent issues of multicollinearity in the model, the organics' dielectric constant was subsequently excluded from the input features.

3.2. Performance of ML Models for Separation Factor Prediction. The preliminary seed test for prediction of separation factor shows a large discrepancy (e.g., R^2 from 0.269 to 0.686) of prediction performance (R^2, MAE, and RMSE) for different seeds based on the testing data set, indicating the necessity of seed randomness assessment (Figure 3A). Seed 370 (OAI ≈ 1.82) was selected, as it provided similar results compared to the average performance (OAI ≈ 1.80). Before optimization, the preliminary model exhibited a severe overfitting issue, which could be caused by the mixed quality of the data and overtraining of the model. To minimize overfitting due to overtraining, we used a series of strategies, including CV, model regularization, feature subsampling, and bootstrap/bagging. Model hyperparameter optimization (HPO) experiments suggested that LGBMR was able to achieve the lowest RMSE (average ≈ 0.62) during CV, with a feature fraction of 0.4, a learning rate of 0.05, a maximum depth of 12, a minimum child samples of 5, a number of estimators of 400, a number of leaves of 30, and a regularization lambda of 3 (Figure 3B). In other words, only 40% of features and the L_2 regularization penalty based on a magnitude of 3 were used to train the model to minimize overfitting. The 1024 fingerprint features were transformed into 11 important PCs (Figure S9A). The final prediction results show that the model obtained better performance after optimization in terms of R^2 (0.61), MAE (0.61), and RMSE (0.45), and the OAI was increased from 1.69 to 2.23 (Figure 3C). Nonetheless, the substantial performance gap between the training and testing subsets indicates that the overfitting

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

issue primarily stems from the mixed quality of data obtained from different studies. Among the 37 different studies in the testing data set, it is evident that data from some studies (e.g., dark gray diamonds) align closely with the perfect diagonal line (i.e., results were accurately predicted by the model), whereas data from some others (e.g., green pluses) deviate more strongly from the line (Figure 3C). Because of the mixed-quality nature of the training data set, the model was trained with both high-quality and poorly fitted data sets. A potential solution lies in utilizing additional relevant, high-quality experimental studies in the future, which may alleviate the issue. It is also worth noting that ML models without DLM (as have been reported in many previous ML studies) can easily achieve overoptimistic results (e.g., $R^2 > 0.9$).[32,37]

The influence of input features on the predicted separation factors was analyzed by using the Shapley values. Shapley values quantify the feature's contribution to the predicted separation factor, as shown in the hierarchy in the SHAP summary plot (Figure 3D). Specifically, the color in the plot represents the value of the input feature, ranging from low (blue) to high (red). For example, the blue color represents a low temperature, while the red color indicates a high temperature. The absolute SHAP value denotes the contribution of the input feature with a higher absolute value representing a greater contribution. A positive Shapley value indicates that an increase in the input feature is associated with a higher predicted parameter. Conversely, a negative Shapley value implies that an increased input feature is associated with a lower predicted parameter. A higher feature value with a higher predicted parameter indicates a positive correlation and vice versa.

Figure 3D presents the SHAP summary plot for the predicted separation factor. Among all of the input features, organic molar volume ranks first and shows a positive correlation with SHAP values. This indicates that a larger organic molar volume is associated with a higher predicted separation factor (Figure 3E). In the PV data set, it is observed that H_2O has a relatively small molar volume (18.07 cm^3 mol^{-1}), while most organic solutes have larger molar volumes (e.g., MeOH with 40.49 cm^3 mol^{-1}, acetic acid with 57.25 cm^3 mol^{-1}). Therefore, the permeation rate of organic solutes from water is expected to be higher for solutes with smaller molar volumes compared to those with larger molar volumes.[43] Consequently, the selectivity would be higher for organic solutes with larger molar volumes. The mass ratio (A/B) ranks second (Figure 3F), and in agreement with our previous study, a positive correlation is observed.[32] This means that a higher mass ratio is associated with a higher separation factor. In the context of organic dehydration, a lower ratio of H_2O in the feed corresponds to a higher selectivity for H_2O, which aligns with experimental observations.[44] Typically, as the water saturation in the feed solution decreases, the water solubility and diffusivity tend to increase. This can lead to increased intermolecular friction and reduced water permeation during transport through the membrane. Consequently, when the feed solution has a higher concentration of organic solute (a larger mass ratio), water permeation through the membrane becomes more pronounced. In the case of organic extraction, as the concentration of organic solutes increases, they exhibit stronger sorption interactions with the hydrophobic membrane. This increased interaction causes the membrane to swell further, resulting in the creation of free volume and mobility within the membrane. Consequently, the diffusion rate of both organic solutes and H_2O are enhanced. However, the diffusion rate of H_2O increases more significantly due to its smaller molecular diameter (e.g., 0.26 nm for H_2O vs 0.52 nm for ethanol). As a result, the selectivity decreases.[45,46]

PC1 ranked third for predicting the separation factor. To elucidate the correlation between PC1 and the separation factor (Figure 3G), we uncovered the contribution of 1024 fingerprint bits to PC1, which investigated which molecular substructures contribute the most to the prediction (Figure S10 and Table S9). The top 10 fingerprint bits of PC1 include 849, 726, 356, 718,1019, 715, 896, 322, 650, and 891. Both Bit_849 and Bit_356 represent atomic sites on the benzene ring, but the unbranched carbon on the benzene ring (Bit_849) is more important than the linked branched carbon (Bit_356). Bit_726 represented the carbon−carbon bond on the benzene ring. Additionally, Bits 718, 715, and 896 represented substructures with two chemical bonds away from the central carbon atom on the benzene ring (Radius = 2), and these substructures include atoms on the branch chain, such as O, S, and C. The importance between these atoms follows O > S > C as the rest of the substructure remained similar. This difference possibly originates from the differences in electronegativity and polarity. Because O has greater electronegativity than S and C, the C−O bond exhibits stronger polarity compared to C−S and C−C bonds. The presence of a benzene ring is important, as benzene rings consist of relatively nonpolar C−C and C−H bonds, which are not effectively solvated by water, making benzene rings highly stable and hydrophobic. It is also noteworthy that two of the top 10 most important bits are unrelated to the benzene ring. Bit_1019 is related to the carbon atom on a six-carbon ring, which forms a nonpolar bond with adjacent carbon atoms. Second, Bit_650 corresponds to the oxygen atom on an oxygen−carbon bond, which is a polar covalent bond.

PC2 ranked fourth and demonstrates a positive correlation with the separation factor (Figure 3H). The most important bit in PC2 is Bit_1019 (Figure S11 and Table S10), which corresponds to the carbon atom on a six-carbon ring. Additionally, Bits 807, 463, and 233 are associated with the hydroxyl group, which contributes to a higher selectivity for H_2O due to its polarity. Furthermore, Bits 695, 897, 299, and 983 correspond to substructures related to the ether group, where ethers can form hydrogen bonds with water since the oxygen atom is attracted to the partially positive hydrogens in water molecules. This interaction may also contribute to higher selectivity for H_2O.

Based on Figure 3D, the effective membrane area ranked fifth. It is observed that a smaller effective area is associated with a higher separation factor (Figure S12). However, this observation has rarely been discussed, prompting us to delve deeper into representative studies. On the one hand, we examined 7 representative studies that utilized membranes with an area larger than 200 cm^2, of which 3 employed commercial membranes obtained from suppliers. On the other hand, we investigated 15 studies with a membrane area smaller than 8 cm^2, all of which utilized lab-synthesized membranes. This analysis suggests a possible hypothesis: the new membranes developed with high selectivity were fabricated by using new/advanced materials or synthetic processes that are not yet available for scale-up. Specifically, 7 out of the 15 studies developed mixed matrix membranes with inorganic fillers and 2 employed hybrid membranes.

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

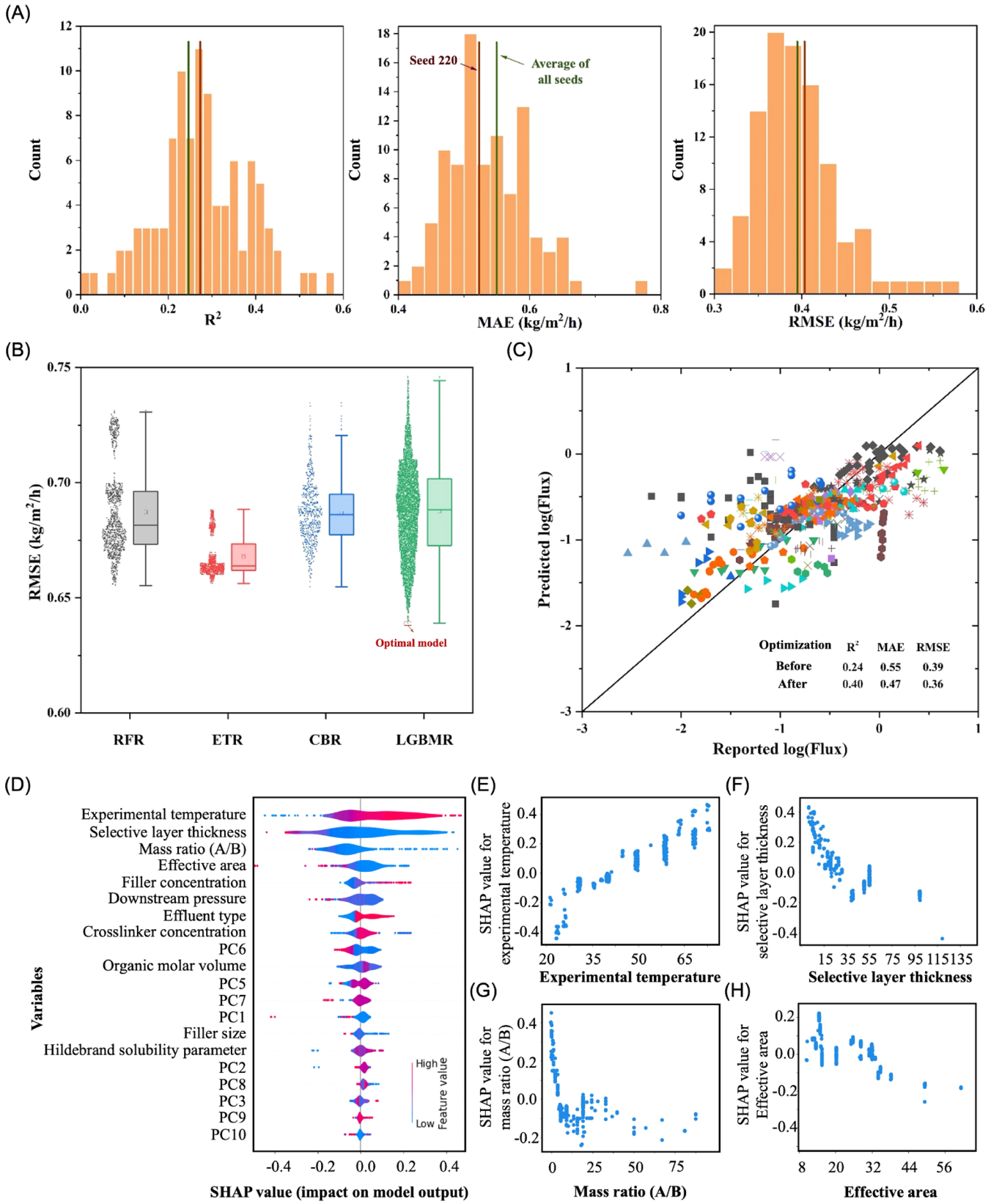


Figure 4. Model development and prediction results for flux. (A) Primary seeds assessment using R^2, MAE, and RMSE based on the testing data set using CBR. (B) Box plots showing performance (RMSE) comparison between RFR, ETR, CBR, and LGBMR using DLM-based data segmentation and a predefined CV approach based on the training data set (seed 220). (C) Comparison between reported and predicted flux (logarithmic) data based on the testing data set using the optimal LGBMR model notes that data from the same (anonymous) study are shown in the same mark. Prediction performance using the testing data set before and after optimization is also displayed. (D−H) Model interpretation by analyzing the contribution of features using SHAP for total flux prediction. SHAP summary plot of (D) features' contribution, (E) experimental temperature, (F) selective layer thickness, (G) mass ratio, and (H) effective area.

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

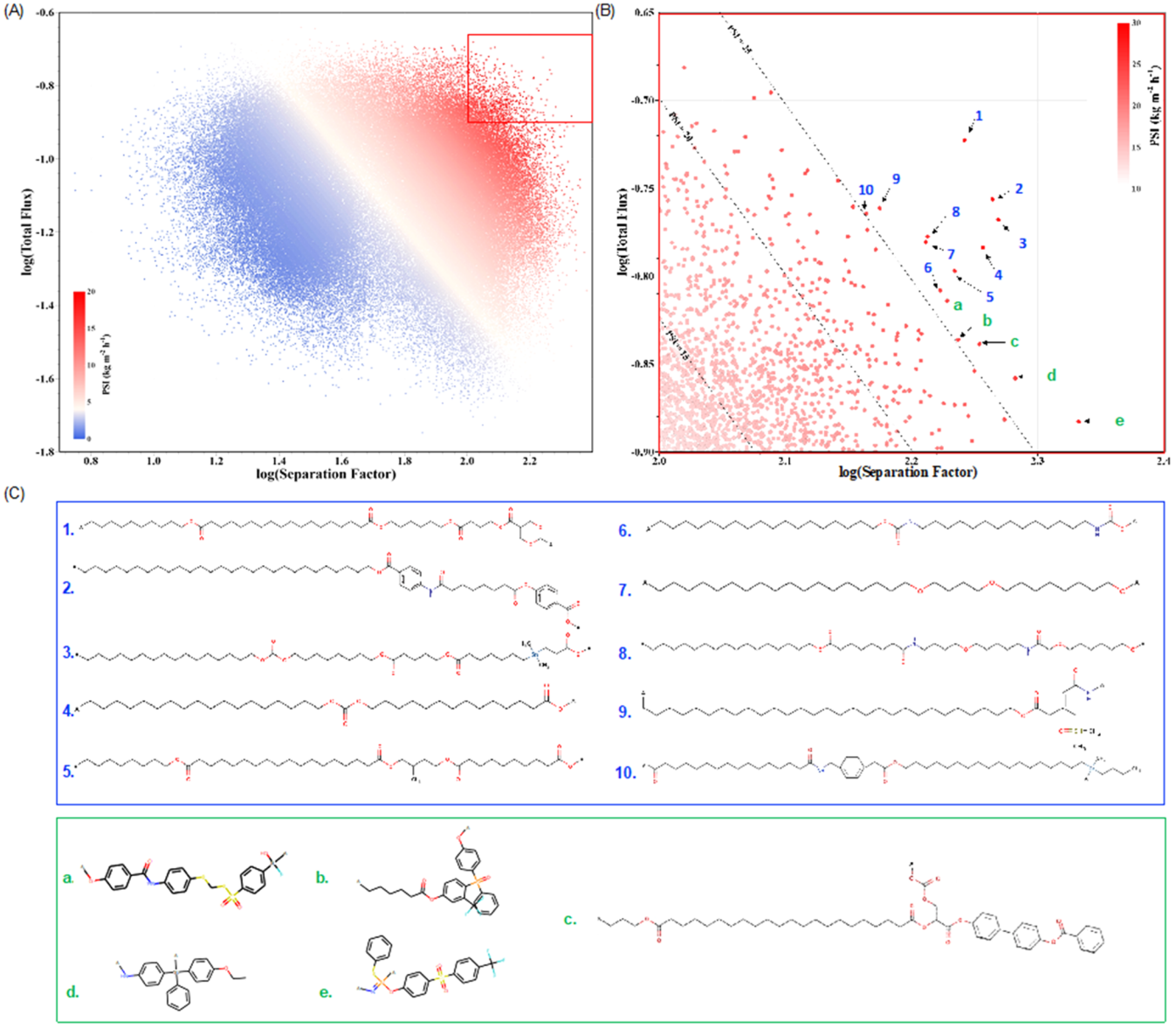


Figure 5. Screening of polymers from PI1M data set. (A) Summary of predicted PV performance of polymers in PI1M data set. The light-yellow slice means the critical value 5, as the current state-of-the-art PSI for acetic acid extraction is 4.6. (B) Expanded graph indicating screened polymers with the highest predicted PSI. The dashed lines represent PSI = 15, 20, and 25. (C) Highlight polymers in (B), where "A" indicate the polymerization points.

3.3. Performance of ML Models for Total Flux Prediction. In general, similar results, albeit with relatively lower accuracy, were obtained for flux prediction compared to the separation factor prediction. Based on the preliminary seed test (Figure 4A), we selected seed 220 (OAI ≈ 1.32) as it was very similar to the average OAI (≈1.31). Similar to separation factor prediction, LGBMR outperformed the other tree-based methods and obtained the lowest RMSE (average ≈ 0.64). Interestingly, the optimal model structure was also similar, with only the exception being the number of leaves of 50 (Figure 4B). This difference reflects the need for greater model complexity to gain more useful information for total flux prediction compared with separation factor prediction. Another minor difference was found when information on the fingerprint features was extracted based on 10 PCs (Figure S9B). The optimal model helped to increase R^2 (0.24 → 0.40) and reduce MAE (0.55 → 0.47 kg/m^2/h) and RMSE (0.39 → 0.36 kg/m^2/h) (Figure 4C). Among the 40 different studies in the testing data set, the discrepancy in data positions for different studies seems to be more significant for total flux than for the separation factor. For example, data from one study (hexagons in wine color on the right side) exhibited almost a vertical line, suggesting minor changes in the total flux with substantial differences in experimental conditions (causing different predicted values). Data from another study (triangles in light blue on the left side) showed an opposite pattern as a horizontal line, indicating that the model is not sensitive to the differences in the experimental conditions explored in that study. The model was also found to be overoptimized (e.g., R^2 > 0.85) and less robust when data splitting with strict DLM was not integrated.

In Figure 4D, the experimental temperature ranked first among all input features and demonstrated a positive linear correlation with J (Figure 4E), indicating that higher

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

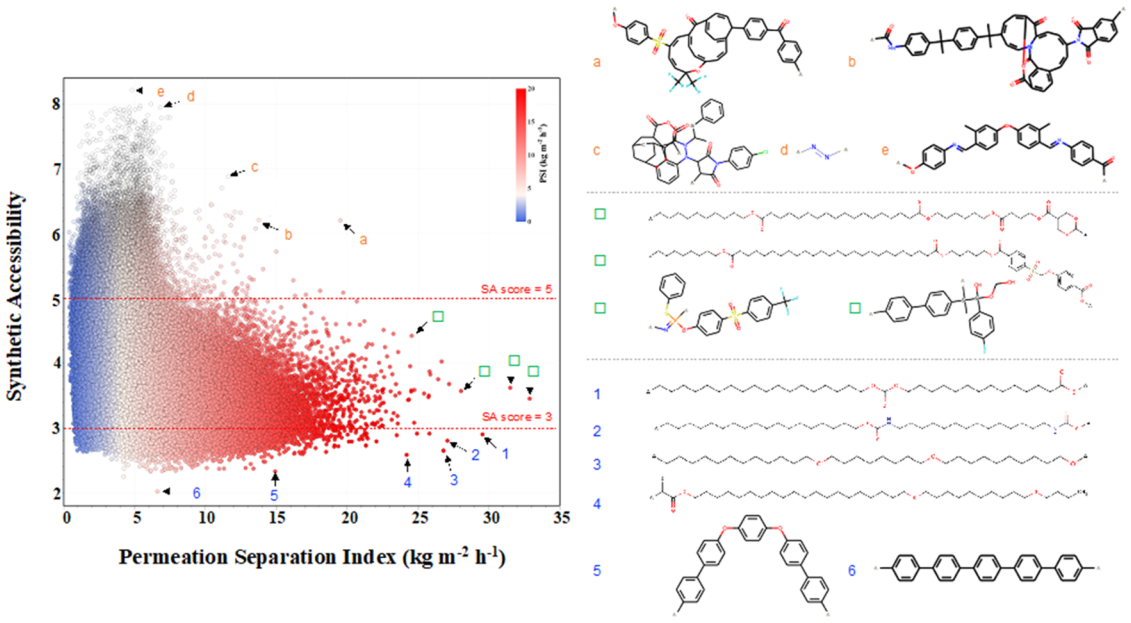


Figure 6. Evaluation of the synthetic accessibility and permeation separation index of correlated polymers in the PI1M data set, where "A" indicate polymerization points.

temperatures could result in higher total flux. The increase in temperature enhances the driving force for mass transfer due to the higher partial pressure and increased diffusion coefficient of a given solute.[47−49] Furthermore, thermal agitation, characterized by the frequency and amplitude of polymer chain movement, also increases with the temperature. This leads to the expansion of the free volume within the membrane and facilitates solute diffusion.[50] The selective layer thickness ranked second and showed a negative correlation with the J (Figure 4F), which aligns with previous studies and is consistent with theoretical models such as Fick's equation and the solution-diffusion model.[51]

The mass ratio ranked third and exhibited a negative correlation with J (Figure 4G). In the context of both organic dehydration and organic extraction, a higher feed concentration (lower mass ratio) indicates that more target molecules come into contact with the selective layer of the membrane. Consequently, due to membrane selectivity, more target molecules are adsorbed to the membrane, causing greater swelling in the top layer of the membrane. This allows more target molecules to pass through the swollen membrane, leading to increased permeability with higher feed content.[1,52,53] The effective area is ranked fourth (Figure 4H) and has a negative correlation with J. This means that, similar to the separation factor, increasing the effective area is not favorable for achieving a higher total flux. The relatively low total flux observed for membranes with larger effective areas may be attributed to the fact that the fabrication of advanced membranes requires novel technologies that have not yet been proven scalable. Also, bench-scale results are often obtained under different operating conditions pilot- and commercial-scale results; in particular, larger module areas can result in reduced driving force along the length of the module.[54,55] Thus, membrane scale-up should be a key research focus to improve PV membrane selectivity and permeability. Additionally, the filler concentration ranked fifth and exhibited a positive correlation with J. This suggests that the addition of fillers could facilitate solute permeation, as the thickness of composite membranes is typically smaller than that of homogeneous membranes or possibly due to the defects caused by membrane fabrication with fillers.

As stated earlier, the performance of the currently developed models is still limited due to the mixed-quality nature of the training data set, which could be possibly improved by future high-quality data sets. Despite the model accuracy, conducting SHAP on a model with DLM would be more reliable than a model without DLM since the model developed with DLM is more robust. Besides, it is worth noting that the SHAP analysis of the currently established models could be limited to drawing conclusions about feature importance. In addition, based on its nature, SHAP explains the importance of features based on ML models instead of real-life PV performance. Therefore, more experiments and simulations are essential to validate the importance of different influencing factors in the real PV process.

3.4. High-Throughput Polymer Screening and Identification from Hypothetical Polymers. We applied the established prediction models to identify promising polymers for separating acetic acid from water. The current upper-bound limit of homogeneous polymeric membranes for acetic acid extraction is summarized in Figure S13. Table S11 provides details of the input features, excluding polymers, with the effluent type set as 1 to indicate organic extraction as the separation objective. The chemical space of the studied polymers and PI1M polymers is shown in Figure S15, where PC1 and PC2 explained 70% data variance (Table S5). The similarity scores between 995,799 hypothetical polymers and 52 studied polymers were compiled in Figure S14. To ensure the feasibility of polymer synthesis, we screened out 183,570 hypothetical polymers with higher similarity (>0.9) to the previously studied polymers. Subsequently, we applied the best-established models from above for separation factor and total flux (Table S12) to these 183,570 hypothetical polymersthen the PSI was calculated for each polymer. As a result, we identified 27,344 polymers with predicted PSI values exceeding the current upper bond limit by 130% (Figure 5A). Polymers

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

with the highest predicted log(PSI) values are highlighted and expanded in Figure 5B. Clear patterns emerge in the two marked regions, indicating highly promising predicted separation performance (i.e., PSI > 25); see details in Table S13. Upon comparison of the two marked regions (Figure 5C), we observe that the polymers in the blue region exhibit higher total flux despite a relatively lower separation factor. The molecular structure of these polymers suggests that their high flux may be attributed to the presence of long carbon chains. These long chains are expected to result in a greater molar volume, which in turn increases the free volume within the membrane. This increased free volume facilitates solute transport, leading to higher total flux. Conversely, the polymers in the green region demonstrate a lower total flux compared to the blue region while exhibiting attractive separation factors, indicating greater selectivity for acetic acid over water. Upon analyzing their molecular structures, we found that all of the polymers in the green region contain more than two benzene rings. Benzene rings have been identified as important features in the model interpretation, as discussed in Section 3.2. These rings consist of relatively nonpolar C−C and C−H bonds, which are not effectively solvated by water. The increased hydrophobicity associated with benzene rings may contribute to the selectivity of these polymers for organic solutes over water.

After promising polymers were identified, their synthetic complexity and feasibility were assessed using SA scores. The details of the SA scores can be found in Text S6. Figure 6 illustrates the evaluation of PSI and SA scores of the 183,570 hypothetical polymers with higher similarity (>0.9) to previously studied polymers. We selected three marked regions based on SA score (<3, 3−5, and >5) to further analyze the relationship between PSI and SA score based on their distribution patterns (details in Table S14). A clear pattern emerged in Figure 6. In the region where the SA score is greater than 5, five polymers were selected (a−e), all of which exhibited unsatisfactory PSI. From their molecule structure, polymers a−c exhibit complex cyclic structures, making their synthesis inaccessible. Compound d has a relatively simple structure of a repeating unit, but a polymeric compound involving a recurring nitrogen−nitrogen double bond could be challenging. Compound e also consists of benzene rings and multiple C−N double bonds and ether bonds, which increase its synthetic complexity. From the SA region of 3−5, four compounds (α, β, γ, and δ) were selected. All of these four compounds have shown promising predicted PSI. For synthetic accessibility, compounds α and β consist of a long carbon chain and ring structure, thus exhibiting moderate synthetic complexity. As for compounds γ and δ, although they contain only one ring structure, the benzene ring, they also incorporate other elements besides C and O, such as fluorine (F), phosphorus (P), silicon (Si), etc. Additionally, they consist of various chemical bonds, such as Si−O, N−P, and S−P. From the region where the SA score is less than 3, six polymers were selected (1−6). Among these, polymers labeled 1, 2, 3, and 4 exhibited both satisfactory PSI and relatively low SA scores, making them promising candidates for future exploration of acetic acid recovery. However, polymers 5 and 6 sacrifice PSI despite their relatively low SA scores, which can be attributed to their symmetrical molecular structure and recurring benzene rings.

In addition, it is also valuable to evaluate the structural characteristics such as free volume for the identified polymers because for single-polymeric dense membranes, the mass transport mainly occurs within the free volume of the amorphous regions. Therefore, to quantify the FFV of the membranes formed by polymers, we also developed a CatBoost Regression model that could predict the FFV using the data set acquired by Tao et al. (Text S10).[56] As a result, our established model performance on the testing data set was R^2 = 0.88, and RMSE = 0.0092. Then, we employed the polymers from the PI1M polymer data set after the first screening (N = 183,570) as the testing data set and acquired their predicted FFVs (Figure S16). As a result, we listed the predicted FFV for the top 10 selected polymers with the best PV performance (Table S15).

Taking into account the SA score, we summarized the top 10 polymers with the highest PSI along with their respective SA scores in Table S15. Among the top 10 polymers, those with long carbon chains (specifically, 37 carbons) exhibited relatively low SA scores, indicating that their synthetic pathways are feasible. These polymers hold great promise for future experimental studies, which may involve validating their predicted performance through experiments or conducting process modeling to fully assess their suitability for PV applications.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.4c00060.

Data structure; data preprocessing; details about model selection; model evaluation metrics; model explanation; calculation of synthetic accessibility score; details about evaluating polymers' fractional free volume (FFV); detailed data structure of 2341 studied PV data set inventory of the 52 types of polymer membranes; and summary of screened promising polymers from PI1M polymer data set for acetic acid recovery (PDF)

Raw dataset (XLSX)

Raw dataset (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Zhiyong Jason Ren − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States; Andlinger Center for Energy and the Environment, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0001-7606-0331; Email: zjren@princeton.edu

Authors

Meiqi Yang − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States; Andlinger Center for Energy and the Environment, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0003-0913-6804

Jun-Jie Zhu − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States; Andlinger Center for Energy and the Environment, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0002-7546-2870

Allyson L. McGaughey − *Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey 08544, United States; Andlinger Center for*

https://doi.org/10.1021/acs.est.4c00060
Environ. Sci. Technol. 2024, 58, 10128−10139

Energy and the Environment and Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, United States; orcid.org/0000-0003-0841-3240

Rodney D. Priestley − *Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, United States;* orcid.org/0000-0001-6765-2933

Eric M. V. Hoek − *Department of Civil & Environmental Engineering, University of California Los Angeles, Los Angeles, California 90095, United States;* orcid.org/0000-0002-5748-6481

David Jassby − *Department of Civil & Environmental Engineering, University of California Los Angeles, Los Angeles, California 90095, United States;* orcid.org/0000-0002-2133-2536

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c00060

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the financial support from US Department of Energy's Advanced Materials and Manufacturing Technologies Office (DE-EE0009494) and National Science Foundation Princeton University Materials Research Science and Engineering Center (DMR-2011750).

■ REFERENCES

(1) Jyoti, G.; Keshav, A.; Anandkumar, J. Review on Pervaporation: Theory, Membrane Performance, and Application to Intensification of Esterification Reaction. *J. Eng.* **2015**, *2015*, No. 927068, DOI: 10.1155/2015/927068.

(2) Liu, G.; Wei, W.; Jin, W. Pervaporation Membranes for Biobutanol Production. *ACS Sustainable Chem. Eng.* **2014**, *2* (4), 546−560.

(3) Van Der Bruggen, B.; Luis, P. Pervaporation as a Tool in Chemical Engineering: A New Era? *Curr. Opin. Chem. Eng.* **2014**, *4*, 47−53.

(4) Chapman, P. D.; Oliveira, T.; Livingston, A. G.; Li, K. Membranes for the Dehydration of Solvents by Pervaporation. *J. Membr. Sci.* **2008**, *318* (1−2), 5−37.

(5) Liu, G.; Jin, W. Pervaporation Membrane Materials: Recent Trends and Perspectives. *J. Membr. Sci.* **2021**, *636*, No. 119557.

(6) Bhat, S. D.; Aminabhavi, T. M. Pervaporation Separation Using Sodium Alginate and Its Modified Membranes - A Review. *Sep. Purif. Rev.* **2007**, *36* (3), 203−229.

(7) Upadhyay, D. J.; Bhat, N. V. Pervaporation Studies of Gaseous Plasma Treated PVA Membrane. *J. Membr. Sci.* **2004**, *239* (2), 255−263.

(8) Castro-Muñoz, R.; González-Valdez, J.; Ahmad, M. Z. High-Performance Pervaporation Chitosan-Based Membranes: New Insights and Perspectives. *Rev. Chem. Eng.* **2020**, *37* (8), 959−974, DOI: 10.1515/revce-2019-0051.

(9) Li, L.; Xiao, Z.; Tan, S.; Pu, L.; Zhang, Z. Composite PDMS Membrane with High Flux for the Separation of Organics from Water by Pervaporation. *J. Membr. Sci.* **2004**, *243* (1−2), 177−187.

(10) González-Velasco, J.; González-Marcos, J. A.; López-Dehesa, C. Pervaporation of Ethanol-Water Mixtures through Poly(1-Trimethylsilyl-1-Propyne) (PTMSP) Membranes. *Desalination* **2002**, *149* (1−3), 61−65.

(11) Zhu, X.; Leininger, A.; Jassby, D.; Tsesmetzis, N.; Ren, Z. J. Will membranes break barriers on volatile fatty acid recovery from anaerobic digestion? *ACS EST Eng.* **2021**, *1* (1), 141−153.

(12) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J. J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55* (19), 12741−12754.

(13) Luis, P.; Van der Bruggen, B. *Pervaporation Modeling: State of the Art and Future Trends*; Elsevier Ltd, 2015.

(14) Shan, H.; Li, S.; Zhang, X.; Meng, F.; Zhuang, Y.; Si, Z.; Cai, D.; Chen, B.; Qin, P. Molecular Dynamics Simulation and Preparation of Vinyl Modified Polydimethylsiloxane Membrane for Pervaporation Recovery of Furfural. *Sep. Purif. Technol.* **2021**, *258*, No. 118006.

(15) Nalaparaju, A.; Zhao, X. S.; Jiang, J. W. Biofuel Purification by Pervaporation and Vapor Permeation in Metal-Organic Frameworks: A Computational Study. *Energy Environ. Sci.* **2011**, *4* (6), 2107−2116.

(16) Pollice, R.; Dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54* (4), 849−860.

(17) Jeong, N.; Chung, T. H.; Tong, T. Predicting Micropollutant Removal by Reverse Osmosis and Nanofiltration Membranes: Is Machine Learning Viable? *Environ. Sci. Technol.* **2021**, *55* (16), 11348−11359.

(18) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing Exceptional Gas-Separation Polymer Membranes Using Machine Learning. *Sci. Adv.* **2020**, *6* (20), No. eaaz4301, DOI: 10.1126/sciadv.aaz4301.

(19) Yang, J.; Tao, L.; He, J.; Mccutcheon, J. R.; Li, Y. Machine Learning Enables Interpretable Discovery of Innovative Polymers for Gas Separation Membranes. *Sci. Adv.* **2022**, *8*, No. eabn9545, DOI: 10.1126/sciadv.abn9545.

(20) Zunger, A. Inverse Design in Search of Materials with Target Functionalities. *Nat. Rev. Chem.* **2018**, *2* (4), No. 0121, DOI: 10.1038/s41570-018-0121.

(21) Hou, D.; Jassby, D.; Nerenberg, R.; Ren, Z. J. Hydrophobic Gas Transfer Membranes for Wastewater Treatment and Resource Recovery. *Environ. Sci. Technol.* **2019**, *53* (20), 11618−11635.

(22) Kim, B.; Lee, S.; Kim, J. Inverse Design of Porous Materials Using Artificial Neural Networks. *Sci. Adv.* **2020**, *6* (1), No. eaax9324, DOI: 10.1126/sciadv.aax9324.

(23) Kostiuchenko, T.; Körmann, F.; Neugebauer, J.; Shapeev, A. Impact of Lattice Relaxations on Phase Transitions in a High-Entropy Alloy Studied by Machine-Learning Potentials. *npj Comput. Mater.* **2019**, *5* (1), No. 55, DOI: 10.1038/s41524-019-0195-y.

(24) Azoulay, J. D.; Koretz, Z. A.; Wong, B. M.; Bazan, G. C. Bridgehead Imine Substituted Cyclopentadithiophene Derivatives: An Effective Strategy for Band Gap Control in Donor-Acceptor Polymers. *Macromolecules* **2013**, *46* (4), 1337−1342.

(25) Martin, T. B.; Audus, D. J. Emerging Trends in Machine Learning: A Polymer Perspective. *ACS Polym. Au* **2023**, *3* (3), 239−258.

(26) Ouyang, Y.; Yu, C.; Yan, G.; Chen, J. Machine Learning Approach for the Prediction and Optimization of Thermal Transport Properties. *Front. Phys.* **2021**, *16* (4), No. 43200, DOI: 10.1007/s11467-020-1041-x.

(27) Huang, X.; Ma, S.; Zhao, C. Y.; Wang, H.; Ju, S. Exploring High Thermal Conductivity Polymers via Interpretable Machine Learning with Physical Descriptors. *npj Comput. Mater.* **2023**, *9* (1), No. 191, DOI: 10.1038/s41524-023-01154-w.

(28) Li, H.; Zeng, B.; Tuo, J.; Wang, Y.; Sheng, G. P.; Wang, Y. Development of an Improved Deep Network Model as a General Technique for Thin Film Nanocomposite Reverse Osmosis Membrane Simulation. *J. Membr. Sci.* **2024**, *692* (2023), No. 122320.

(29) 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* (2021), No. 120268.

(30) Sawada, S. ichi.; Sakamoto, Y.; Funatsu, K.; Maekawa, Y. Toward the Design of Graft-Type Proton Exchange Membranes with

High Proton Conductivity and Low Water Uptake: A Machine Learning Study. *J. Membr. Sci.* **2024**, *692* (2023), No. 122169.
(31) Wang, M.; Xu, Q.; Tang, H.; Jiang, J. Machine Learning-Enabled Prediction and High-Throughput Screening of Polymer Membranes for Pervaporation Separation. *ACS Appl. Mater. Interfaces* **2022**, *14* (6), 8427−8436, DOI: 10.1021/acsami.1c22886.
(32) Yang, M.; Zhu, J. J.; McGaughey, A.; Zheng, S.; 5946Priestley, 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, DOI: 10.1021/acs.est.2c06382.
(33) Zhu, J. J.; Dressel, W.; Pacion, K.; Ren, Z. J. ES&T in the 21st Century: A Data-Driven Analysis of Research Topics, Interconnections, and Trends in the Past 20 Years. *Environ. Sci. Technol.* **2021**, *55* (6), 3453−3464.
(34) Ma, R.; Luo, T. PI1M: A Benchmark Database for Polymer Informatics. *J. Chem. Inf. Model.* **2020**, *60* (10), 4684−4690.
(35) Tao, L.; Chen, G.; Li, Y. Machine Learning Discovery of High-Temperature Polymers. *Patterns* **2021**, *2* (4), No. 100225.
(36) Zhong, S.; Guan, X. Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties. *Environ. Sci. Technol.* **2023**, *57*, 18193−18202, DOI: 10.1021/acs.est.3c02198.
(37) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57*, 17671−17689, DOI: 10.1021/acs.est.3c00026.
(38) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 3147−3155.
(39) Mateo, J. R. S. C. The Shapley Value. *Green Energy Technol.* **2012**, *83*, 95−101.
(40) Painer, D.; Lux, S.; Siebenhofer, M. Recovery of Formic Acid and Acetic Acid from Waste Water Using Reactive Distillation. *Sep. Sci. Technol.* **2015**, *50* (18), 2930−2936.
(41) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21* (6), 311−325.
(42) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1* (1), No. 8, DOI: 10.1186/1758-2946-1-8.
(43) Smuleac, V.; Wu, J.; Nemser, S.; Majumdar, S.; Bhattacharyya, D. Novel Perfluorinated Polymer-Based Pervaporation Membranes for the Separation of Solvent/Water Mixtures. *J. Membr. Sci.* **2010**, *352* (1−2), 41−49.
(44) Pulyalina, A.; Polotskaya, G.; Goikhman, M.; Podeshvo, I.; Chernitsa, B.; Kocherbitov, V.; Toikka, A. Novel Approach to Determination of Sorption in Pervaporation Process: A Case Study of Isopropanol Dehydration by Polyamidoimideurea Membranes. *Sci. Rep.* **2017**, *7* (1), No. 8415, DOI: 10.1038/s41598-017-08420-0.
(45) Jaimes, J. H. B.; Alvarez, M. E. T.; de Moraes, E. B.; Maciel, M. R. W.; Filho, R. M. Separation and Semi-Empiric Modeling of Ethanol—Water Solutions by Pervaporation Using PDMS Membrane. *Polymers* **2021**, *13* (1), No. 93, DOI: 10.3390/polym13010093.
(46) Liu, G.; Xiangli, F.; Wei, W.; Liu, S.; Jin, W. Improved Performance of PDMS/Ceramic Composite Pervaporation Membranes by ZSM-5 Homogeneously Dispersed in PDMS via a Surface Graft/Coating Approach. *Chem. Eng. J.* **2011**, *174* (2−3), 495−503.
(47) Kuhn, J.; Castillo-Sanchez, J. M.; Gascon, J.; Calero, S.; Dubbeldam, D.; Vlugt, T. J. H.; Kapteijn, F.; Gross, J. Adsorption and Diffusion of Water, Methanol, and Ethanol in All-Silica DD3R: Experiments and Simulations. *J. Phys. Chem. C* **2010**, *114* (14), 6877−6878.
(48) Hillaire, A.; Favre, E. Isothermal and Nonisothermal Permeation of an Organic Vapor through a Dense Polymer Membrane. *Energy* **1999**, *38*, 211−217.
(49) Vane, L. M. Review of Pervaporation and Vapor Permeation Process Factors Affecting the Removal of Water from Industrial Solvents. *J. Appl. Chem. Biotechnol.* **2020**, *95*, 495−512, DOI: 10.1002/jctb.6264.
(50) Ye, H.; Yu, J.; Zhang, Z.; Song, B.; Liao, Y.*Preparation and Modification of PU Membrane and Its Swelling and Pervaporation Properties*, Proceedings of the 2015 International Conference on Environmental Engineering and Remote Sensing; Atlantis Press, 2015.
(51) Kwon, Y.; Chaudhari, S.; Kim, C.; Son, D.; Park, J.; Moon, M.; Shon, M.; Park, Y.; Nam, S. Ag-Exchanged NaY Zeolite Introduced Polyvinyl Alcohol/Polyacrylic Acid Mixed Matrix Membrane for Pervaporation Separation of Water/Isopropanol Mixture. *RSC Adv.* **2018**, *8* (37), 20669−20678.
(52) Vane, L. M. A Review of Pervaporation for Product Recovery from Biomass Fermentation Processes. *J. Chem. Technol. Biotechnol.* **2005**, *80* (6), 603−629.
(53) Qin, F.; Li, S.; Qin, P.; Karim, M. N.; Tan, T. A PDMS Membrane with High Pervaporation Performance for the Separation of Furfural and Its Potential in Industrial Application. *Green Chem.* **2014**, *16* (3), 1262−1273.
(54) Hardikar, M.; Marquez, I.; Achilli, A. Emerging Investigator Series: Membrane Distillation and High Salinity: Analysis and Implications. *Environ. Sci.: Water Res. Technol.* **2020**, *6* (6), 1538−1552.
(55) Rezakazemi, M.; Shahverdi, M.; Shirazian, S.; Mohammadi, T.; Pak, A. CFD Simulation of Water Removal from Water/Ethylene Glycol Mixtures by Pervaporation. *Chem. Eng. J.* **2011**, *168* (1), 60−67.
(56) Tao, L.; He, J.; Arbaugh, T.; McCutcheon, J. R.; Li, Y. Machine Learning Prediction on the Fractional Free Volume of Polymer Membranes. *J. Membr. Sci.* **2023**, *665* (2022), No. 121131.