Environmental Science & Technology

pubs.acs.org/est

Open Access

This article is licensed under CC-BY 4.0

Critical Review

# Machine Learning-Aided Inverse Design and Discovery of Novel Polymeric Materials for Membrane Separation

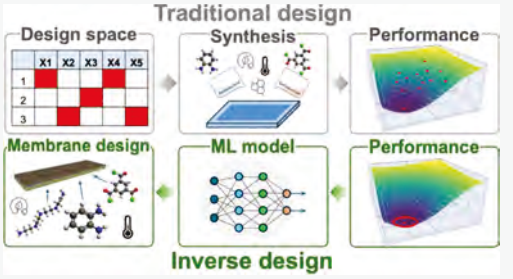
Raghav Dangayach,[#] Nohyeong Jeong,[#] Elif Demirel, Nigmet Uzal, Victor Fung, and Yongsheng Chen*

Cite This: *Environ. Sci. Technol.* 2025, 59, 993−1012

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Polymeric membranes have been widely used for liquid and gas separation in various industrial applications over the past few decades because of their exceptional versatility and high tunability. Traditional trial-and-error methods for material synthesis are inadequate to meet the growing demands for high-performance membranes. Machine learning (ML) has demonstrated huge potential to accelerate design and discovery of membrane materials. In this review, we cover strengths and weaknesses of the traditional methods, followed by a discussion on the emergence of ML for developing advanced polymeric membranes. We describe methodologies for data collection, data preparation, the commonly used ML models, and the explainable artificial intelligence (XAI) tools implemented in membrane research. Furthermore, we explain the experimental and computational validation steps to verify the results provided by these ML models. Subsequently, we showcase successful case studies of polymeric membranes and emphasize inverse design methodology within a ML-driven structured framework. Finally, we conclude by highlighting the recent progress, challenges, and future research directions to advance ML research for next generation polymeric membranes. With this review, we aim to provide a comprehensive guideline to researchers, scientists, and engineers assisting in the implementation of ML to membrane research and to accelerate the membrane design and material discovery process.




**KEYWORDS:** *machine learning, polymeric membrane, separation, inverse design, material discovery*

Downloaded via 99.7.135.49 on February 14, 2025 at 17:02:53 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

## 1. INTRODUCTION

The precise separation of nanoscale molecules and ions from diverse solutions has gained significant importance in various industries over the past few decades.[1] Membrane technology has emerged as an effective strategy to achieve this goal due to its high separation and energy efficiency, low capital cost, and easy scalability. The benefits associated with membrane technology has led to its utilization for a variety of applications such as wastewater treatment, water purification, gas separation, and resource recovery.[2,3] Polymeric materials are at the forefront of membrane manufacturing as a result of their outstanding processability, high versatility, as well as exceptional mechanical and chemical stability.[4] These polymers possess distinctive chemical and physical characteristics, which can be tailored to form multifunctional membranes.[5] This diversity allows engineers and scientists to fine-tune polymeric membranes according to their individual needs such as high permeability and selectivity.

With the growing use of polymeric membranes for different applications, a permeability-selectivity trade-off has been observed due to their intrinsic limitations.[6−8] This implies that polymeric membranes with high permeability typically possess low selectivity and vice versa. Investigating an exact property-process-structure relationship to balance this trade-off is a complex task due to three primary factors: the presence of numerous features such as material properties and synthesis variables, the vast material design space, and the lack of complete understanding of the physics and chemistry of sophisticated material systems.[9,10] Traditional membrane fabrication is based on a "bottom-up" approach wherein the polymeric membranes are selected and put through an iterative trial-and-error process of adaptation and testing to improve membrane performance metrics.[11] These approaches for membrane design are associated with being laborious and resource intensive. Alternatively, computational simulation tools such as molecular dynamics simulation (MD) and density functional theory (DFT) have shown good potential in the prediction of material structures and performance to varying degrees.[12,13] However, it is important to note that these models require high computational demands and thus their applications are often limited to simpler conditions.[14]

Instead of using a "bottom-up" approach for membrane design, where material properties and performance metrics are calculated after membrane synthesis, scientists need to revamp their design approach to an inverse design methodology.[15]

**Received:** August 10, 2024
**Revised:** December 3, 2024
**Accepted:** December 4, 2024
**Published:** December 16, 2024

© 2024 The Authors. Published by American Chemical Society

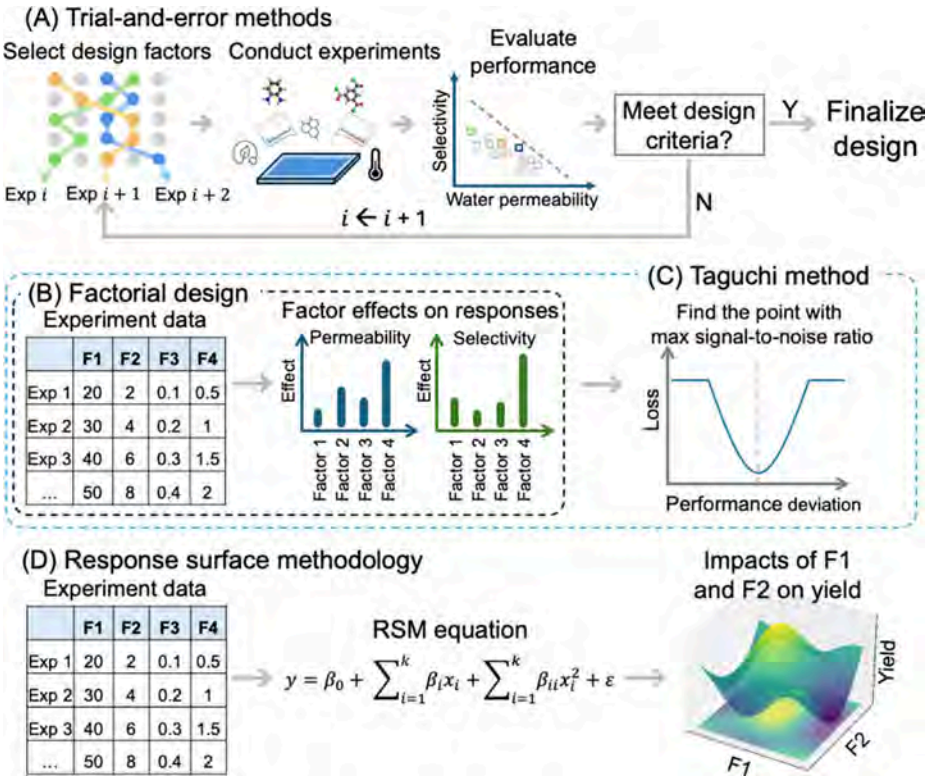



**Figure 1.** Design and discovery of polymeric membranes by using (A) trial-and-error method and systematic methods including (B) factorial design, (C) Taguchi method, and (D) response surface methodology.

This approach facilitates effective exploration of the membrane design space, enabling the identification of novel membrane materials and optimal fabrication conditions to achieve desired objectives.[16] In this context, artificial intelligence (AI) has facilitated groundbreaking advancements in the field of design and discovery of membrane materials.[17−19] AI refers to the study of computer programs or systems that can mimic human cognitive functions in data analysis, decision-making, and problem-solving in order to accomplish tasks, such as understanding natural language, learning from experience, and adapting to new situations.[20] Machine learning (ML) is a branch of AI focused on the development of algorithms that leverage data to make predictions and decisions.[21] ML has emerged as a viable alternative to conventional experimental approach or simulation methods because of its ability to analyze extensive and complex data patterns. ML can also be used to reveal insights into the underlying separation mechanism and find key features that may guide future membrane design for specific applications.[22,23] The utilization of these algorithms has not only enabled in accurate predictions of material properties, but also expedited the discovery of potential material candidates from a vast search space.[17,24] Thus, researchers have demonstrated the application of ML to design polymeric membranes for gas separation, nanofiltration (NF), and pervaporation with the shared objective of improving specific performance metrics.[25−27]

The application of AI in material discovery and design of membranes is a relatively new research area, and the number of studies in this domain remains limited to prediction, analysis, and optimization of numerous process conditions. The previously published review papers in this subfield have comprehensively highlighted the advancements in state-of-the-art tools and techniques that assist researchers in applying ML to membrane science and technology.[28−32] A review positioned at the intersection of the traditional direct design approach and the ML aided inverse design approach is currently absent. We aim to supplement the existing reviews by linking the various substeps involved in building ML models, from a different perspective that is focused specifically on the inverse design of membranes and polymeric material discovery. Researchers have faced several challenges while designing ML models, such as difficulties in formulating strategies to procure, clean, and treat data, as well as identifying key features necessary to support their hypothesis. Additionally, there is a need for adequate representation of categorical data (e.g., polymers, solvents, and ions) in a format that is readable by computers. The black-box nature of ML models requires the use of additional tools to understand the relationship between input features, such as experimental conditions and polymer characteristics, and output features, such as membrane performance. It is of vital importance to introduce comprehensive guidelines on ML methodologies to expedite the discovery of novel membrane materials, particularly from the perspectives of environmental science and engineering.

In this review, we illustrate the advantages and drawbacks of the Edisonian trial-and-error approach and statistical design of experiments (DoE) for synthesis of high-performance membranes. This leads to a discussion on the necessity for data-driven approaches such as ML, where we propose a comprehensive framework, highlighting the various steps involved in the development of ML models. As part of the ML blueprint, we cover the data collection and preprocessing steps, along with various feature generation methods. Subsequently, we briefly discuss various types of ML algorithms used for membrane research, implementation of explainable artificial intelligence (XAI) for model interpreta-


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

tion and validation of the membrane synthesis conditions obtained using ML via experimentation or computational tools. We then review relevant case studies on ML-assisted inverse design material discovery, discussing the research advances made using high-throughput virtual screening (HTVS), Bayesian optimization (BO), and generative ML within a structured framework. Finally, we conclude with a discussion on the constraints and difficulties of current AI applications in this domain, pinpointing existing deficiencies, ongoing progress, future research direction and its environmental implications.

## 2. TRADITIONAL APPROACHES OF POLYMERIC MEMBRANE MATERIAL DISCOVERY

The identification of materials and ideal synthesis conditions to fabricate polymeric membranes has been an important area of research for membrane scientists. Polymeric materials are crucial in the advancement of membranes due to their exceptional processability, and widespread availability. These materials will remain vital to membrane technology, as demonstrated by the chemistry-processing-structure-performance paradigm.[33] The pursuit of improved efficiency, performance, and environmental sustainability is the driving force behind the investigation and development of novel polymeric materials for membrane applications. This undertaking necessitates a comprehension of polymer physics and chemistry, as well as a perceptive awareness of the unique requirements of each application.[34] The utilization of trial-and-error methods and the DoE framework for experimental design and screening has been pivotal in advancing membrane technology. The selection between these approaches often depends on the specific goals, resources, and constraints of the investigation. The fundamental objective of material scientists is to enhance the membrane design efficiency to shorten the research and development cycle, enabling them to keep pace with rapid advancements in science and technology.

**2.1. Trial-and-Error Methods.** The trial-and-error approach has been the fundamental basis for the evolution of polymeric membrane technology. Researchers using this methodology combine or choose different polymers based on established knowledge or assumptions regarding certain polymer characteristics.[35] Subsequently, these polymeric membranes undergo a sequence of experiments to assess their appropriateness as potential candidates, with a specific emphasis on performance metrics, such as permeability, selectivity, and stability.[36] This practical experimentation-based process is extremely iterative, often requiring numerous cycles of synthesis and testing to discover a polymer with desired properties (Figure 1A).

The effectiveness of trial-and-error approaches is greatly dependent on the researcher's discernment and expertise. Researchers select materials and process conditions that show potential in accordance with their knowledge of polymer chemistry and the desired characteristics of the membrane.[37] While the trial-and-error approaches are characterized by their simplicity and directness, they are also recognized for being time-consuming, labor intensive, resulting in the squandering of resources. Furthermore, the task of predicting results of trials becomes challenging with a limited comprehension of the correlation between a material's structure and its performance. The selection of the appropriate approach is contingent upon the particular application, available resources, and desired level of efficiency. While this approach may require significant time and resources, it has the potential to produce unforeseen results, often resulting in significant advancements that might not be attainable through more systematic methods.

**2.2. Experimental Design and Screening.** Although trial-and-error method is still useful due to its simplicity, a systematic approach to design high performance membranes is necessary.[38,39] The utilization of experimental design methods provides an efficient strategy to synthesize membranes by leveraging statistical tools to optimize experiments, analyze data, and evaluate outcomes.[40] This enables in a focused exploration of synthesis conditions, pinpointing the key parameters that impact performance and their ideal values in membrane applications.[41] Although it demands sophisticated statistical expertise and possibly a larger initial investment, this method provides a more profound understanding of the relationships between structure and properties. Consequently, it guides the design of polymeric membranes in a more efficient manner.

The DoE is a comprehensive framework with multiple experimental designs. It provides variety and robustness in organizing, analyzing, and interpreting controlled tests to evaluate the factors that influence the value of a parameter or collection of parameters. It can be tailored to accommodate a broad spectrum of factors and enables a thorough examination of cause-and-effect relationships. Utilizing statistical methods and factorial designs, DoE can effectively minimize the number of required experiments, resulting in expedited and economically efficient research. The field of DoE encompasses a range of methodologies, including factorial designs, Taguchi methods, and response surface methodology (RSM).[42−44]

Factorial design is a fundamental technique in DoE that effectively investigates the impact of individual factors and their interactions, which is essential for comprehending intricate systems (Figure 1B). This approach is especially beneficial when examining a significant number of factors as it offers a thorough understanding of how these variables affect the desired outcome. Scientists frequently utilize factorial design approaches to methodically investigate the impacts of different polymers and additives on membrane characteristics.[45,46] Identifying viable material combinations and creating a baseline for further optimization is a critical step in this process. Scientists can determine the performance characteristics of membranes by changing elements, such as polymer type, pore-forming agents, and cross-linkers.[39,47] The Taguchi Method, well-known for its emphasis on robust and resilient design, aims to minimize variability and improve product quality by reducing susceptibility to external noise factors (Figure 1C). The utilization of orthogonal arrays optimizes the experimental procedure, reducing resource requirements while providing significant insights on the impacts of various parameters.[48] The RSM is utilized for comprehensive process optimization (Figure 1D). This method is highly effective when the relationship between the input factors and the output responses is not well understood. It involves a series of planned experiments to establish a mathematical model that accurately represents a response surface map. This assists in the investigation of the most favorable conditions to achieve the desired outcome.[49] These methods (i.e., Factorial design, Taguchi, and RSM) are highly valuable in optimizing the synthesis and processing conditions for polymeric membranes. By enabling researchers to pinpoint the most relevant elements and their interconnections, these methodologies enhance the efficiency of membrane develop-


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

ment.[50] Table S1 (Supporting Information) provides a brief overview of various traditional methods used for exploring optimal process parameters for membrane design.

**2.3. Challenges and Limitations of Traditional Methods.** Traditional methods for discovering and designing polymeric membranes have largely depended on empirical approaches, particularly the trial-and-error testing of various material combinations and processing conditions. Researchers typically synthesize polymers and then evaluate their membrane-forming capabilities, focusing on properties like permeability, selectivity, mechanical strength, and chemical stability. This often leads to a repetitive cycle, where the synthesis stage is revisited to adjust polymer compositions or processing parameters, continuing until a material meets the desired criteria. On the other hand, DoE can pose challenges due to its complexity, requiring a high level of statistical competence for successful execution. Based on various experimental design methodologies (e.g., full factorial, central composite design, and Box-Behnken), researchers develop an experimental design matrix by fixing one variable and systematically varying the levels of other variables to generate the required experimental combinations. Conducting and evaluating experiments can be labor-intensive, and the cost associated with extensive experimental trials can be considerable, especially in intricate multifactorial designs.[50] In certain cases, researchers may integrate multiple design approaches to find the optimum response which can lead to increased operational difficulty.[51] A major drawback of these conventional methods is their unpredictability, making it difficult to predict which polymers will perform effectively as membranes. Moreover, both approaches can be environmentally and economically burdensome, particularly due to the extensive use of chemicals and solvents in polymer synthesis and processing, which could be a constraint for smaller research teams or institutions with limited budgets.[52,53]

The trial-and-error approach often relies on using theoretical equations to model transport and describe the underlying separation principles of membrane processes (e.g., Solution-diffusion, extended Nernst-Plank, Donnan-steric pore model with dielectric exclusion).[54−57] DoE, on the other hand, develops first- or second-order models to approximate the membrane performance based on the synthesis conditions or membrane properties.[58,59] These models are easy to interpret, but they struggle to capture the complex nonlinear relationships between synthesis conditions, membrane properties, and membrane performance. To address the limitations of traditional methods, ML algorithms are increasingly integrated into the discovery process of polymeric membrane materials, offering a potential to accelerate innovation and drive a paradigm shift in the field.[30,32,60,61]

## 3. MACHINE LEARNING FOR POLYMERIC MEMBRANE MATERIAL DISCOVERY

Leveraging massive data, ML uncovers the intricate relationships between input variables and outputs, enabling predictions without possessing an extensive domain knowledge.[62] This approach can surpass the traditional, time-consuming, and costly methods for membrane design, offering a more efficient and cost-effective pathway to innovation in membrane technology. It is impossible to screen all possible polymeric materials for optimizing membranes by using traditional methods given the huge number of polymer candidates, each possessing unique physical and chemical properties. ML has been successfully used to tackle this problem since it enables in rapid screening of materials for membranes, reducing the time and effort required for experimentation or computational assessment.[63] A typical workflow for the development of polymers using ML approaches is as follows: (1) Data collection: Acquisition of comprehensive, high-quality data is indispensable for building robust ML models that predict membrane performance. Data set for ML can be experimentally generated, extracted from existing literature, or obtained from simulations. (2) Data preprocessing and feature generation: Preprocessing the data collected to handle missing values and outliers. The data can be normalized or scaled to ensure uniformity within the ranges of input variables. Additionally, the structural information from categorical variables needs to be converted into machine-readable representations for ML model use. (3) ML model development: After the data is preprocessed, several ML algorithms can be used depending on the nature of the problem and data characteristics. (4) ML model interpretation: Interpreting ML models helps identify the features that are most important to the model's performance. This can narrow down the search space for new polymer candidates or find optimal experimental parameters. (5) Validation: Based on the fabrication conditions or polymeric candidates screened using the ML models, we can validate the model either experimentally or through computational tools (Figure 2).

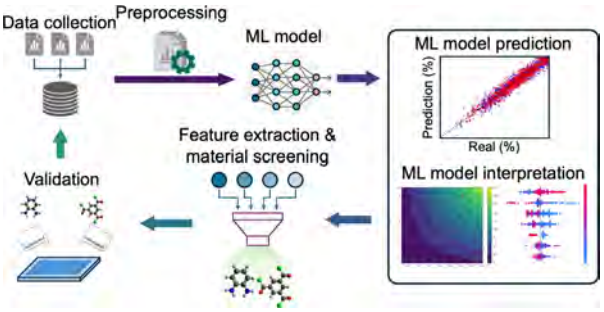



**Figure 2.** Machine learning pipeline for the inverse design and discovery of polymeric membranes.

**3.1. Data Collection.** In the context of employing ML for polymer membrane research, the significance of data cannot be overstated. The effectiveness of ML models is significantly impacted by the quality and quantity of data they are trained on. One of the more pronounced challenges in developing robust ML models is associated with the limitation in the available data, especially in maintaining high quality. Smaller, low-quality data sets fail to represent the full diversity (either in terms of the number of data points or the required number of features) of the entire data population, which can result in biased predictions. Training with small data sets often results in overfitting as the model learns noise in the data instead of intended underlying patterns, compromising its ability to generalize to new, unseen data. Larger, high-quality data sets can help overcome these shortcomings by reducing bias, improving feature representation and allowing more room for models to generalize beyond the training data. There is a lack of well-defined criteria to identify low quality data, which makes it difficult to evaluate data reliability. Cross-checking the data can improve the quality of the data set by identifying outliers or unreliable data points. A commonly accepted guideline to prevent overfitting and enhance the model


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

reliability is to ensure that the data set includes at least ten times more data points than the number of input variable types.[64] Common data collection methods are described in the following subsections.

*3.1.1. Traditional Method and Medium/High-Throughput Experimentation.* Manual experimentation is a common method to generate high quality data. Experimentation within a closed laboratory environment allows for quality control and standardization, which is critical to the data collection process. The biggest issue with this method is that it is extremely time-consuming and requires a lot of manpower. It is exceptionally difficult to conduct hundreds of experiments and generate a data set. Thus, using this experimental approach to develop a comprehensive data set could take months or years. Glass et al. devised a high-quality data set with only 50 data points to predict water permeance and zeta potential of polymeric membranes. They obtained high coefficients of determination (>0.9) using gradient boosting methods, showcasing the potential of in-house derived data sets in ML.[65]

Medium/high-throughput experimentation has emerged as an effective approach to conduct a vast array of experiments simultaneously or through automated methods. These methodologies are often coupled with combinatorial approaches, where various factors or variables are systematically combined in different permutations to identify optimal operational or synthesis conditions. For example, Cano-Odena et al. used a high-throughput setup (HTS) to find the optimum parameters for the removal of ibuprofen using cellulose acetate-based NF/RO membranes.[66] The authors were able to complete the entire experimental work within 3 months, including the time required for optimization using genetic algorithms. In another study, Ignacz et al. designed a medium-throughput setup (MTS) crossflow system that enabled them to collect around 1,938 rejection data points measured for 3 membranes and 10 green solvents. This comprehensive data set allowed for an in-depth analysis of membrane performance and selectivity, as well as an exploration of the correlation between the molecular weights and selectivity.[67] As demonstrated in the aforementioned studies, HTS/MTS has shown its potential in generating extensive data sets for ML, data analysis, and optimization processes.

*3.1.2. Data Collection from Literature.* The most widely used approach to curate a data set for ML is data mining from literature.[68,69] Using journal databases like Google Scholar, Scopus, and Web of Science, researchers can manually find publications and articles related to their problem statement for data extraction. This method is much more resource efficient as compared to performing new experiments since it leverages the studies conducted over the past few decades. However, the data found in research papers is often sparse and presented in various formats (e.g., tables, graphs, figures, or text), with no standardized reporting criteria available.

Researchers have used a combination of input variables representing the membrane and polymer properties to predict membrane performance metrics such as water/solvent/gas permeability, salt rejection, gaseous selectivity, membrane fouling characteristics, conductivity, ion selectivity, and ion transport rate for their desired application.[13,26,70,71] In Table S2 (Supporting Information), we have listed several examples of the different labels used to collect data. Extraction of data requires domain knowledge to navigate through the literature and find the relevant data points. This process is quite tedious and time-consuming since finding relevant articles with all the key features is required to build accurate ML models. The challenges associated with manual data collection have prompted researchers to seek computational (DFT and MD) tools for data generation and natural language processing (NLP) based methods to expedite the data collection process which will be explained in Sections 5.2 and 5.3.

**3.2. Data Preprocessing and Feature Generation.** After data collection, the raw data need to be cleaned and prepared to make it suitable for further analysis or employment in ML models. Data preprocessing or pretreatment has several aspects associated with it. Data cleaning involves the identification and rectification of errors, removal of duplicate entries and outliers from a data set. This is to improve the quality and integrity of data while avoiding any biases.[64] Data imputation is the process of filling in missing values in a data set with substituted values. This is typically done when a data set has missing values for some of the features, which is usually the case when data is mined from literature. Some researchers use either mean or median values to impute data sets, whereas others utilize data collected in its raw format (true missing values) since these features have physical and chemical properties.[72] Data normalization is the process of adjusting data distribution to make it suitable for ML analysis. This involves standardizing the data to have a mean of zero and a standard deviation of one and transforming the data to follow a normal distribution. For categorical features, it is important to convert them into computer readable forms as ML models typically cannot process categorical features (e.g., types of polymers, solutes, or solvents) directly.[73] Examples of these computer-readable forms include one-hot encoding (transforming categorical data into binary vectors), label encoding (assigning each category a unique integer), and binary encoding (a hybrid method that combines aspects of both to efficiently handle numerous categories with minimal data expansion).[16,74]

Feature generation, also known as feature engineering, is a strategic process of creating new features from both numerical and categorical features to improve ML model performance. The goal of this process is to transform (e.g., combination or decomposition) the existing features into more informative and insightful forms.[75] By using domain knowledge, we can incorporate attributes potentially relevant for predictive analysis, improving the capability of the ML models to discern complex patterns and relationships. For example, water composition such as types of ions in feedwater can influence charge effects, osmotic pressure, and scaling in membrane experiments. Instead of listing all cations and anions with their concentrations, utilizing ionic strength can be more informative and simpler feature for ML. Descriptors are fundamental representations of chemical structures, playing a crucial role in feature generation. Traditional descriptors encompass compositional, structural, and spectral information. In the context of polymer materials, the macromolecular chains consist of repeating units linked by covalent bonds. The properties of materials are directly influenced by the structures and compositions of these repeating units. As a result, the focus of research on polymer descriptors revolves around the characterization of these repeating units.[76] The following subsections discuss the most frequently used descriptors to represent categorical features.

*3.2.1. Line Notation.* Simplified molecular-input line-entry system (SMILES) is a widely used format to represent the



https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

structure of molecules.[77,78] SMILES strings consist of alphanumeric characters that represent atoms and bonds in a molecule.[79] It encodes all the atoms, bonds, rings, and branches of the molecule. An important aspect of SMILES strings is that they can be directly used as a descriptor in ML models or they can be transformed to other descriptors.[80] Thus, SMILES strings play a crucial role in representing molecules in a format that can be easily processed by ML algorithms, facilitating the application of ML in polymer design and development.

*3.2.2. Polymer Fingerprinting.* Polymer fingerprints are compact and lossless representations of polymer structures that capture important structural features, such as monomer sequences, branching patterns, and functional groups. Morgan fingerprints (MF) are the most widely used fingerprinting method in membrane design. MF captures the substructure around non-hydrogen atoms within a defined radius and converts the molecules into binary vectors which are suitable for ML model training. The formation of these binary vectors, however, can result in "bit collision" wherein the representation has a single encoding for the multiple functional groups. This means that random substructures without any contribution to the performance get included in the model training process and its subsequent interpretation. Increasing the number of bits (fingerprint size) is one way to tackle the issue, however, it requires higher computational power to process due to increased number of features. A newer method such as molecular access system (MACCS) key vectorization helps address "bit collision" by mapping the specific substructures to individual indices.[81] Additionally, binary morgan fingerprints can only capture presence/absence of a substructure, not how many times they occur. Count-based morgan fingerprints can tackle this problem since they can count the number of times each chemical bond occurs resulting in a more accurate representation of the polymer.[82] Molecular embedding and molecular graphs are two other commonly used methods for polymer fingerprinting. Molecular embedding generates continuous vector representations of substructures, allowing for the measurement of similarity between different structures, while molecular graph representations view a molecular structure as an undirected graph, making it particularly suitable for applications of deep learning in polymer materials.[83]

Apart from the standard methods used for polymer fingerprinting, group contribution method has been used to determine the structural groups present in the polymer sets. The basic assumption is that the polymeric properties can be represented as a weighted sum of the individual contribution by its constituents. The Bondi method, the Marrero and Gani method, and the Yampolskii's method are common approaches that have been used to predict gas permeability of polymeric membranes.[84,85] The newest toolbox in polymer fingerprinting is inspired by the transformer-based architecture for NLP. Kuenneth and Ramprasad present polyBERT, a novel method that outperforms the traditionally used approaches for polymer fingerprinting.[86] The authors trained a model on a vast data set of polymer SMILES (~100 million hypothetical strings), enabling it to understand the grammar and syntax of polymer chemical language. This fingerprinting approach has significant potential for polymer property prediction, polymer structure prediction, and the design of new polymers.

*3.2.3. Molecular Descriptors.* Molecular descriptors encompass a wide range of features derived from the molecular structure that can be used in ML models.[87] These descriptors are quantitative representations of chemical compounds that capture various aspects of their chemical and physical properties.[88] Molecular descriptors can be generated using professional descriptor generation software, such as Dragon, or open-source toolkits, such as Mordred or RDKit. Ritt et al. represented anions using molecular descriptors to study the thermodynamic mechanism associated with the design of single-species selective membrane. They were able to showcase the importance of various molecular features (i.e., structure, polarizability, interactions, electrostatics, macroscopic, confined polarizability, confined interactions, and confined electrostatics) that affected the transport of anion of through a cellulose acetate membrane.[89] In another study, Ignacz et al. obtained the best descriptors that showcase the effect of different functional groups present in solutes on their rejection in organic solvent nanofiltration (OSN).[90] Researchers have also converted chemical structure of cationic and anionic groups into extensive molecular descriptors to predict ion conductivity of polymer-based ion-exchange membranes.[91,92]

Quantitative structure activity-relationships (QSAR) and quantitative structure property-relationships (QSPR) have been widely utilized to model the physical and chemical properties of molecules on the basis of their chemical structures.[93,94] These generated properties can be used as molecular descriptors in ML models since they assist in the feature generation of (1) specific molecules we aim to separate or (2) polymers used to design membranes.[65,95,96] This extensive representation enhances the ML models and can potentially be used to design membranes for tailored applications.

**3.3. ML Model Development and Approaches.** Once the data set is preprocessed and the relevant descriptors are added to transform the data set, it is ready to be used for training ML models depending on the requirements. Supervised and unsupervised learning have been widely used in the research of polymer membrane design. In supervised learning, the algorithm is trained on a labeled data set to map the input feature to outputs (membrane performance). In unsupervised learning, the algorithm is provided with unlabeled data to find patterns or structures within the data without any predefined model output.[97] This is in contrast with the traditional design methods or supervised learning wherein researchers focus on performance metrics of the membrane to gain insights and develop a hypothesis. Researchers have obtained unforeseen results using unsupervised learning since they can group high-performance materials based on their properties allowing for efficient exploration of new candidates for their desired application.[98,99] ML algorithms that are commonly used to model polymeric membranes are provided in Table S3 (Supporting Information).

**3.4. ML Model Interpretation.** After ML algorithms are applied to our data set, the best performing model is selected based on the chosen model performance metric (e.g., mean absolute error, mean squared error, root mean squared error). Due to the "black box" nature of ML models, XAI tools have been employed to understand and interpret the model's predictions. These tools can identify the influence of features on the model output and present the contribution of each descriptor to polymeric membrane performance which can guide membrane design and material discovery. In the subsequent subsections, we discuss the two most commonly


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

used XAI tools for ML model interpretation, namely Shapley additive explanations (SHAP) and partial development plots (PDP)

*3.4.1. SHAP.* The SHAP method is a commonly used technique that provides insights into the decision-making process of a ML model. It can find the impact of the polymer features on the membrane performance.[100] The SHAP values for an input feature x gives the prediction $p$ as

$$\mathcal{O}_x(p) = \sum_{S \subseteq N \backslash x} \frac{|S|!(N - |S| - 1)!}{N!}[p(S \cup x) - p(S)]$$

where $S$ is the subsets of all features without feature $x$, $N$ is the set of all features, $p(S \cup x)$ are the predictions by the built ML model considering feature $x$, and $p(S)$ are the predictions without considering feature $x$. The differences among all possible subsets of $S \subseteq N\backslash x$ are calculated due to the dependency of the effect of withholding a feature on other features in the ML model.[101]

The SHAP values indicate the impact that a specific feature has on membrane performance. Positive and negative SHAP values indicate positive and negative contributions to membrane performance, respectively. Moreover, features with higher absolute SHAP values have greater contributions on the particular performance indicator (i.e., higher influences on the target variable).[102] Tao et al. correlated the polymer's functional groups to fractional free volume (FFV) of the membranes using SHAP analysis. They were able to validate the impact of experimentally verified concepts, such as the positive contribution of rigid aromatic rings or phenyl groups to the FFV of polymeric membranes. They also found that carbonyl group density had a positive influence on the transport properties of polymer membranes.[103] Jeong et al. evaluated whether the knowledge gained by ML for membrane separation aligned with the fundamental principles of membrane science. Using SHAP analysis, the authors were able to reveal the influence of several factors (i.e., the properties of membranes and solutes) on the membrane performance, demonstrating that ML was able to understand the complex mechanisms of membrane separation.[22] Gallage Dona et al. used SHAP analysis to rank the most important polymer and molecular descriptors for determining the ion-activity coefficients of polymer ion exchange membranes (IEMs).[104] These coefficients play a crucial role in the ion-transport process across the membrane, directly influencing selectivity in IEMs. In another study, Gao et al. developed an ML model that highlighted the fundamentals of ultrafiltration (UF) membrane design. The authors identified the most significant membrane features and fabrication conditions that need to be optimized for the design of high-performance UF membranes.[72]

*3.4.2. Partial Development Plots (PDP).* PDP is a graphical representation showing the relationship between specific features and the predicted outcome of a model while keeping all the other features constant. They help in understanding the marginal effects of a single feature on the output that assists in model interpretation. The average partial dependence function for a feature $S$, $f_s$ can be calculated using

$$f_s = E_{x_c}[f(x_s, x_c)] = \int f(x_s, x_c) dP(x_c)$$

where feature variable $C$ is the complement of $S$; $x_c$ and $x_s$ are their feature vectors, respectively.[105]

Wang et al. developed a ML model to screen polymeric materials for pervaporation application. They performed PDP for total flux and separation factor against water contact angle, membrane thickness, Hildebrand solubility parameter, and operational parameters. The PDP studies gave insight on the relationship between the features and their effects on the ML model predictions.[27] Guan et al. used PDPs to optimize pore size and BET values of MOFs for synthesizing mixed matrix membranes with $CO_2$ permeability and $CO_2/CH_4$ selectivity that surpasses 2008 Robeson upper bound.[106] Deng et al. applied PDP to identify the suitable fabrication candidate space for membranes that exhibited high mono/divalent ion selectivity. They were able to synthesize four new membranes that exceeded the present upper bounds of the permeability-selectivity trade-off.[107] Researchers also studied the importance of structural and operational features of polyamide nanofiltration (PA-NF) membranes using PDP plots to achieve high water/salt selectivity. The authors determined the ideal ranges of pore radius and zeta potential to achieve high mono/divalent salt selectivity for both anions and cations.[105,108] Li et al. developed a model to study the permeability-selectivity trade-off on thin film nanocomposite reverse osmosis (RO) membranes. Using bivariate PDPs, the authors identified optimal ranges for nanoparticle loading and nanoparticle size, enhancing both the relative water permeability and relative salt passage for RO membranes. They later used these results as input conditions in their ML model and found improved outcomes in their model-optimized experiments.[109]

Regardless of their results in visualizing the factors associated with exceptional membrane performance, the results from PDP should be carefully evaluated before they are directly applied to experimental design. These plots may produce misleading visuals while extrapolating to regions with little data, often resulting in artificial trends beyond the values at extremities for any specific feature.

**3.5. Validation.** The final step involves validation of the ML models using laboratory experimentation or computational tools. MD is a computational technique that helps in studying the dynamic behavior of molecules and materials at the molecular level. The simulation involves numerically solving the classical equations of motion for a system of interacting particles (atoms or molecules) over a specified time period. It is an important tool in the field of membrane science, facilitating the development of structure−property-performance relationships.[110]

Xu et al. validated an ML model developed to study the permeability and behavior of OSN membranes using both MD simulation and experimental studies. MD simulation was not only able to predict the methanol permeability of polymer intrinsic microporosity (PIM) membranes, but it also provides new insights into their swelling behavior. These membranes were also experimentally fabricated in the lab and tested for their methanol permeability to validate the ML model. The experiments indicated that the PIM-1 membranes had complete solute rejection with methanol permeability close to the predictions of the ML models.[111] Through their pioneering study, Barnett et al. developed a ML model with a gas permeation data set of 700 polymers. This model was used to predict the gas permeation behavior of over 11,000 polymers and discovered more than 100 polymers exceeding the Robeson upper bound line for $O_2/N_2$ and $CO_2/CH_4$ gas pairs. The researchers validated the results by fabricating two novel polymeric membranes and testing their performance for


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

the separation of $CO_2/CH_4$. The experimental results for the novel polymeric membranes closely matched the predicted values from the ML model, falling within the prediction error margin.[25] In another investigation, Tayyebi et al. identified the key chemical functional groups which can positively or negatively affect membrane performance using SHAP analysis. They leveraged this knowledge to synthesize grafted PA-RO membranes which was able to surpass the water permeability/salt selectivity trade-off. They further characterized their ML developed membranes using Fourier transform infrared spectroscopy (FTIR), scanning electron microscopy (SEM)-energy dispersive X-ray spectroscopy (EDS), thermogravimetric analysis (TGA) and contact angle measurements which aids in understanding the underlying physical and chemical properties of the ML designed membrane.[112]

## 4. ML-DRIVEN INVERSE DESIGN AND MATERIAL DISCOVERY OF POLYMERIC MEMBRANES

The acceleration of material discovery using data-driven approaches in polymeric membrane research marks a significant shift from traditional trial-and-error methodologies to more efficient design strategies. Using ML as an inverse design methodology will greatly reduce the time and effort required to design and explore new materials to synthesize high-performance membranes.[29,113] Unlike the traditional trial-and-error method, which is time-consuming and inefficient since it involves testing of candidate materials until those with the desired properties (materials → property) are obtained, inverse design begins by selecting the desired properties of material and then work backward to identify materials that can achieve those properties (property → materials).[114−116] Inverse-design approach can also be used to simultaneously optimize multiple target properties, a challenge that is often difficult to address using the traditional approaches.[117,118]

The first step of inverse design is to define the scope of the design problem, which involves identifying the input variables (e.g., chemical composition, molecular structure, and membrane fabrication conditions) and specifying the target properties (e.g., permeability, hydrophobicity, and mechanical strength). The next step is data collection and generation, as a comprehensive and high-quality data set is crucial for training accurate ML models. Considering the high computational demand for optimization process, it is important to focus on the most relevant characteristics of polymers and membrane fabrication conditions to define design space, while minimizing the inclusion of less important, irrelevant input variables. To use ML models effectively, the chemical structures of the polymers must be converted into a machine-processable format, as explained in Section 3.2. Recent advancement in polymer informatics plays a substantial role in facilitating these applications by allowing adequate representation of polymers that meet the desired design criteria. ML models which are carefully curated using novel algorithms and feature representations can be interpreted using XAI to provide insights into the underlying principles guiding the separation process. They are also used to identify the desirable physical and chemical properties or membrane fabrication conditions critical to the design of high-performance membranes. Wang et al. developed a ML model to predict the membrane performance of layer-by-layer (LbL) membranes and expedite the exploration of polymer candidates. They conducted SHAP analysis of Morgan Fingerprints which helped in identifying the key atomic groups conducive to high permeability and selectivity. This analysis generated a reference Morgan fingerprint which was mapped against PoLyInfo database using Tanimoto coefficient screen similar candidates.[119] The authors were able to find 23 potential polymers that can be used to synthesize LbL NF membranes.[120] In addition to using XAI as a tool for membrane design, scientists employ other methodologies to navigate the vast polymer candidate and fabrication space, such as (1) high-throughput virtual screening, (2) global optimization, and (3) generative ML.

**4.1. High-Throughput Virtual Screening.** Candidates may often be overlooked using the traditional approaches since researchers usually focus on evaluating previously reported high-performance materials or those with similar chemical structures. ML models with high prediction accuracy enable us to curate a high-throughput virtual screening (HTVS) setup that can be used to screen potential polymer candidates.[121] In a HTVS setup, researchers can define the ideal performance metrics, physical and chemical properties, and functionalities which can be used to screen several polymeric candidates at the same time without conducting any experiments. ML models can predict performance metrics for previously untested candidates, helping narrow down the number of potential candidates from a large selection pool. Yang et al. devised a HTVS setup to identify potential polymeric candidates with a high potential for acetic acid extraction from water using pervaporation. In the first stage, they screened ∼180,000 potential polymeric candidates from the PI1M database (which includes 1 million polymers from Polymer Informatics database) based on their similarity (>0.9) to the polymers used to train their ML model. This was followed up by further screening of the selected polymeric candidates based on predicted permeation separation index (indicating performance) and synthetic accessibility score (indicating ease of synthesis), ultimately identifying 10 potential polymer candidates for pervaporation.[122] It is also recommended for researchers to not only rely on predefined performance metrics, but also use their intuition, expertise, and knowledge to define better selection criteria and build more robust HTVS setups.[123]

**4.2. Bayesian Optimization.** Global optimization tools such as Bayesian Optimization (BO) have also demonstrated their great potential in a variety of inverse design problems in materials engineering.[124] BO is an iterative approach that allows the exploration of design conditions using surrogate functions and acquisition functions to build an optimization model to guide membrane design. A surrogate function is essentially a ML model trained on available experimental data, wherein the model can be used to approximate membrane performance metrics based on the input features. This surrogate model estimates membrane performance metrics on the chosen exploratory design space. The acquisition function can then be used to determine which experiments are most likely to be successful.[125] Gao et al. combined ML and BO to guide experimentation in discovering high performance PA-NF membranes capable of surpassing the current upper bound for permeability-selectivity trade-off.[126] The surrogate model was trained using the data obtained from synthesis conditions. The BO function was then used to identify new combinations of monomers and fabrication conditions within the input design space. Using these conditions, the authors synthesized 8 membranes that were able to surpass upper bound for the water permeability-salt selectivity trade-off. This validated the ML model's capability to discover new


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

monomers and synthesis conditions that enhance membrane performance. BO is not only limited to experimental design, but it can also assist in the material discovery process. Chen et al. used a Bayesian algorithm to modify existing polymers within a data set to discover 200 new polymers showing exceptional separation performance for $CO_2/CH_4$ and $CO_2/N_2$.[127]

Given the large number of influencing factors on membrane design, new methods have been developed to apply BO to these high dimensional parameter spaces, while minimizing the computational demands. Gui et al. proposed a taking-another-step BO (TAS-BO), which offers a simple-yet-effective strategy to tackle high dimensional BO problems. At each iteration, a local Gaussian process (GP) is trained using points neighboring the current candidate. This coarse-to-fine local search enables a more efficient exploration of the search space.[128] A strategic optimization approach can also enhance the efficiency of the discovery process. Dalal et al. used a batch BO to efficiently explore a vast design space of over 5,790 polymer formulations for optimizing pDNA and CRISPR-Cas9 ribonucleoprotein delivery.[129] The BO predicted the most promising formulations in sequential rounds, significantly reducing the design space. After three rounds of optimization, they sampled less than 10% of the design space while identifying the top-performing polymer combinations for delivery efficiency and cell viability.

**4.3. Generative ML.** Generative ML techniques, such as RNNs or graph-based design tools, can also be used to navigate the chemical space and accelerate material discovery by generating new data points based on previously trained data.[130−132] In general, high dimensional polymeric data is scaled down to a lower dimension to capture relevant features, which are then used to generate new polymeric candidates.[133] These candidates can be tested using accurate ML models built on training data, enabling us to screen newer high-performance materials. Yang et al. trained a ML model with 778 polymers mapping their Morgan fingerprints to their gas permeabilities.[134] This helped them develop an accurate ML model for predicting the permeabilities of 9 million new polymers that had never been tested before for gas separation. These 9 million polymers were generated using (a) RNNs trained on SMILES strings of existing polymers, (b) theoretical chemical reaction between binary pairs of dianhydride and diamine which yields polyimides, and (c) ladder polymers generated using a combination of monomer combinations and RNN generation. They used the ML model to predict the permeabilities of these polymeric membranes to identify new candidates surpassing the Robeson upper bound. In another study, Giro et al. developed an inverse-material design workflow to design new monomers for carbon capture with targeted property ranges for the permeability of $CO_2$, glass transition temperature, and half-decomposition temperature. They represented the input molecules as feature vectors (containing encoded information related to molecular building blocks of the monomer) whose features were extracted using regression modeling. These feature vectors were optimized using Particle Swarm Optimization and converted to molecular structures using a graph-based McKay's Canonical Construction Path Algorithm to generate new polymeric structures with desired properties.[135,136] Even though genetic algorithms (GA) are not traditionally classified under the generative ML umbrella, researchers generated new polymer compounds using them based on the chemical fragments present in the polymer data set.[137]

Despite the unforeseen performance showcased by the hypothetical candidates engineered using these generative techniques, their synthesis can be quite complex which limits its application. Including synthetic accessibility score within the screening or design process is one way to tackle this problem, however, more research regarding their synthesizability is desired.[123] Another major factor that governs the use of generative ML models is the requirement for high-quality data sets to capture the relevant features. Poor quality in training data would result in the formation of unrealistic samples.[28]

In Table 1, we explore a series of success case studies using ML to enable the discovery and design of new polymeric membranes, organized within a comprehensive framework. These case studies aim to provide an overview of the synergy between ML and polymeric membrane technology, highlighting current advancements and its future potential in our field.

## 5. RECENT PROGRESS, FUTURE DIRECTIONS, AND PERSPECTIVES

**5.1. ML Algorithms for Polymeric Membrane Technology.** The separation and purification performance of membranes depends on a variety of factors including synthesis conditions, operational conditions, and the structural, chemical, and functional properties of membranes, solvents, liquids and gases. Due to the complexity of this process, researchers are trying to develop new, robust ML algorithms to predict the properties and performance of the polymeric membranes, which can provide a better understanding of their separation mechanisms.

Graph neural networks (GNNs) are a type of neural network designed to operate on graph-structured data, where graph consists of nodes (representing entities) and edges (representing the connections or relationships between these entities). GNNs leverage the inherent structure of graphs to capture the relationships in the data and does not require extensive feature engineering or representation design.[140] Ignacz et al. used GNNs to model solute−solvent-membrane interactions and understand the structural impact of solutes and solvents on the performance of OSN membranes. With the help of GNNs, they were able to visualize the effects of functional groups and substructures and further extract the atomic and bond level information on the molecules of interest.[141] Queen et al. also used GNN to develop POLYMERGNN, a model that allows for the prediction of polymeric properties.[142] In an attempt to improve the standard deep learning model, Li et al. developed a three-component residual ANN (R-TNN) to study the water permeability and salt selectivity trade-off in TFN-RO membranes. Using this approach, the authors were able to adjust the model such that the first two networks emphasized on learning the data from relative water permeability and relative salt rejection, while the other layers focused on feature analysis and gaining knowledge from the previous networks. The authors demonstrated that this modified network outperformed neural networks and ML models (RF, K-nearest neighbor, XGBoost, and adaptive boosting trees) in predicting membrane performance.[109] Cui et al. combined MD simulations with density peak clustering algorithm based on unsupervised learning to model the ionic channels of membranes. They were able to visualize and


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

**Table 1. Case Studies Demonstrating the Application of ML for Polymeric Membrane Material Discovery**

| data collection | feature generation | ML algorithms | model interpretation | validation | inference | ref |
|---|---|---|---|---|---|---|
| Around 500 to 1000 data points for 6 gases ($CH_4$, $CO_2$, He, $H_2$, $N_2$, $O_2$) from the literature | Polymers represented as binary fingerprints | Gaussian process regression (GPR) | - | Synthesis of 2 polymers with performance surpassing the current upper bound for $CO_2/CH_4$ separations. | Predicted the gas permeabilities of ~11,000 potential polymers and screened candidate polymers whose predicted permeabilities lie above the Robeson upper bound. | 25 |
| Polymer chemistry and gas permeability data from PoLyInfo and Membrane Society of Australia (MSA) | Polymers represented as SMILES strings. Their molecular representations were performed using Morgan Fingerprinting and Chemical Descriptors. | Random Forest (RF) and Deep Neural Networks (DNNs) | SHAP analysis revealed the chemical basis required for overcoming permeability selectivity trade-off. | MD simulations were used to verify the permeabilities used to predict the ML model. | ML models were used to conduct high-throughput screening of over 9 million hypothetical polymers with unknown gas permeabilities designed using generative ML. | 134 |
| 152 different data points collected from the literature for OSN | Molecular descriptors (representing chemical structure of polymer and solvent) or gross descriptors (representing important measurable polymer and solvent properties) | Kernel Ridge Regression (KRR), Gradient Boosting Regression (GBR) and LASSO | SHAP analysis found that the presence of hydrophobic and amine fragments improved solvent permeability. | MD simulations were used to provide insights on swelling behavior while PIM membranes were synthesized for experimental validation. | Study showed the applicability of ML, combined with molecular simulation and experimentation, to predict solvent permeability and develop new polymeric membranes. | 111 |
| Data collected from 218 publications for PA-NF membranes | SMILES and Morgan Fingerprints to represent polymers. The salts were represented using molecular descriptors. | XGBoost and CatBoost | The authors developed a reference Morgan Fingerprint using positive contributions of the atomic structures obtained using SHAP to screen 20 new monomers. | The models were validated by synthesizing 8 PA-based NF membranes. | Showed the applicability of Bayesian optimization to design new experiments and explore the fabrication search space to design high-performance NF membrane materials. | 126 |
| Generated data (114 data points) from their own experiments | The synthesis conditions for PA-TFC membranes as input features | ANN | PDP plots defined the fabrication candidate space for desirable performance. | Synthesized 4 PA membranes according to the ML model showing high mono/divalent ion selectivity. | Devised a data driven methodology to synthesize high-performance PA-TFC membranes that exceeded the upper bound of the permeability selectivity trade-off. | 107 |
| Gas permeability data set consisting of 1,169 homopolymers collected from patents and publications | Polymers represented as SMILES. Topological, Geometrical and Structural descriptors were used for property predictions. | LASSO regression, Ridge regression, Elastic Net Regression, RF, KRR, SVR | - | MD simulations were used to model $CO_2$ permeability for ML model verification. It also gave insight on the filtration dynamics of materials. | Suggested a framework for ML-based generative monomer design for carbon capture. | 135 |
| The data was collected from 30 publications with 227 different types of LbL membranes | Numerical features, such as concentration of polyanions and polycations, reaction time, ionic strength, along with categorical features including LbL method type, polyanion and polycation types, and the name of substrate, serve as the input features. | RF, Boosted tree model, Linear Regression, and XGBoost | SHAP analysis of Morgan Fingerprints gave insights into the key chemical structures important for membrane performance. | - | Designed 2 reference Morgan Fingerprints on the basis of the contributions obtained from SHAP to give 23 potential polymers candidates for LbL application | 120 |
| 1,347 experimental data points collected from 41 journal articles. | The input features include SMILES along with 5 molecular fingerprinting techniques. | GBR | PDP and SHAP analysis revealed that operating conditions are the most important for permeance whereas molecular weight of solutes are the most important for rejection. | Synthesis of a TFC membrane with good agreement between actual and predicted permeance for 3 solvents (dimethylformamide, methanol, and acetone) | In-silico design of potential membranes synthesized by combining different monomers was conducted. The performance of these membranes was predicted to screen potential candidates for OSN application. | 138 |
| Data collected from[126] | SMILES notation and Morgan Fingerprinting | CatBoost and RF | SHAP analysis used to identify chemical functional groups important for membrane performance to select additives for membrane grafting. | Synthesized and characterized high-performance membranes using FTIR, SEM, TGA and EDS. | This paper showed the application of SHAP to guide the chemical modification process of RO membranes for performance greater than the upper bound for water permeability- salt selectivity. | 112 |
| ~2400 data points (consisting of 52 unique polymers and 32 types of organic solvents) collected from 264 articles | Morgan Frequency Fingerprint (MFF), which includes topological features in addition to the substructures present in the molecule. | CatBoost, extra trees regression, RF, LGBR | SHAP summary plots were used to reveal the importance of different chemical features which were transformed to principal components. | - | High-throughput screening to identify potential candidates from PI1M data set with the potential for acetic acid extraction from water having high permeation separation index and low synthetic accessibility score | 122 |



https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

**Table 1. continued**

| data collection | feature generation | ML algorithms | model interpretation | validation | inference | ref |
|---|---|---|---|---|---|---|
| Collected literature data containing over 780 unique polymers for $CO_2$, $N_2$ and $O_2$ | SMILES notation for polymer representation. Extended Connectivity Fingerprint (Morgan Fingerprint) and MACCS for polymer fingerprinting | SVR, K-Nearest Neighbors, Decision Tree and RF | - | Used polymer genome software to make permeability predictions for the hypothetical polymers[139] | Used genetic algorithms to develop new polymers which were screened using a HTVS approach with exceptional permeability and $CO_2/N_2$, $CO_2/O_2$ selectivity. | 137 |
| Data consisted of 749 Anion Exchange Membranes (AEMs) | SMILES notation was used to represent the copolymers and $OH^-$ conductivity, water uptake, swelling ratio (%) and tensile strength chosen as performance metrics. They calculated theconductivity- dimension stability trade-off (CDST) coefficient for AEMs to characterize performance. | XGBoost performed the best out of 12 algorithms used for training | SHAP analysis revealed that octanol–water partition coefficient of cations, number of rotatable bonds in backbones and BalabanJ are the most important parameters for $OH^-$ conduction and CDST. | - | The authors screened 2519 potential copolymers from ~172,000,000 candidates to synthesize high-performance AEMs | 123 |

quantify the properties of the ionic channels which helped them study water transport across proton exchange membrane fuel cells.[143] Transformer models have also gained huge attention for the property predictions of polymers. Xu et al. developed TransPolymer, a transformer-based language model to predict the various properties of polymers, which include their electrolyte conductivity, electron affinity, ionization energy, and refractive index.[144] Zhong et al. used generative pretrained transformer (GPT) based models to develop QSAR relationships for water contaminant properties using SMILES strings. These models outperformed CatBoost-based QSAR models.[145] The algorithms mentioned above can improve data visualization and membrane performance predictions allowing for better research outcomes.

Integrating ML models with physical and chemical principles can further improve the prediction accuracy of the models and deepen our understanding of membrane separation. Rehman et al. developed a physics-informed deep learning model to study ion transport across polyamide membranes. They integrated charge conservation laws into the deep learning model, which led to an improvement in the prediction of membrane performance.[146] Lee et al. also developed a physics-informed ML model to study the diffusion of gases through polymeric membranes. Using physical equations, the authors enforced the neural network to learn the physical relationships governing the diffusion process for the prediction of gas diffusivity.[147] In another study, Wang et al. used physics-informed performance metrics (fractional free volume and average void size) to assess the gas separation of polymeric membranes. They were able to screen polyamide membranes that exceeded the Robeson upper bound plots for several gaseous mixtures.[148] Researchers have used chemistry informed ML to find promising candidates for solid state polymer electrolytes for lithium-ion batteries.[149] Thus, incorporating the essential rules of physics and chemistry has the potential to enhance the "intelligence" of ML models and can be of significant importance in utilizing ML to identify new polymers.

There is a computational cost associated with obtaining data from different sources: High-fidelity data have better accuracy and are more expensive to obtain, whereas low-fidelity data are less accurate, but require a lower computational cost.[150] Multifidelity models use data from multiple sources to address the trade-off between fidelity and computational demands, helping develop accurate ML models with minimal resource use.[151] Rall et al. developed a multiscale optimization framework that integrates high-fidelity ion transport models with ML to optimize membrane processes for water treatment.[152] Using the data generated from the one-dimensional extended Nernst–Planck ion transport model, ANN was trained to predict the performance of NF membranes and served as a surrogate model of high-fidelity model. The authors integrated this surrogate ML with mechanistic process models to optimize membrane synthesis properties and overall process design for membrane plant, reducing computational resources and maintaining the accuracy of the physical model. Instead of using ML as a surrogate for high-fidelity models, models across different fidelity levels can also be integrated for inverse membrane design. Lazin et al. suggested an efficient multi-fidelity BO for solving inverse problems in the quantum control of time-dependent system. By combining low- (prior distribution) and high-fidelity (posterior distribution) models for GP, this method enables efficient exploration of the next


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993–1012

query point in BO, reducing computational time and maintaining high accuracy in the optimization process.[153]

Generative ML can innovate and accelerate material discovery by expanding the diversity of potential materials used to design membranes. Generative adversarial networks (GANs) can generate synthetic data with target properties by exploiting sequential or graph representations of organic materials.[130] Several researchers also used transformer based models to generate unique molecules using SMILES strings and desired property values as inputs.[154−156] Diffusion models are the latest advancement in generative ML for a variety of chemistry and drug design applications.[157,158] Inspired by nonequilibrated thermodynamics, these models are able to generate 3D molecular structures through forward and backward diffusion processes. These models captured the chemical and physical properties of molecules represented via graph structures to design target molecules.[159] Park et al. developed ZeoDiff relying on diffusion model architecture to generate porous materials with user-desired characteristics.[160]

A big challenge in the application of ML on polymeric membrane design is the adequate representation of polymers and additives. Polymer structures are highly complex that have dynamics ranging from various length and time scales.[161] ML for polymer design requires encoding polymers in formats that are interpretable by computers. The chemical representation technique based on SMILES uses a single molecular representation to extract all the features from the polymer.[162] The majority of research in this area is still primarily concerned with the topology of individual monomers or cross-linkers, whereas a polymer network might consist of topologies or structural features that are never observed in monomers or cross-linkers. For example, when two different monomers were considered to react to form a new polymer, the newly generated topologies may not be fully defined by the individual monomers alone, and thus the new structure requires further description.[63] BigSMILES has recently emerged as an extension to SMILES, which is tailored specifically to polymeric systems. It can help in better representation of homo-, co-, and block polymers and map out the nature of the polymer in terms of its branched, network, and terminal group information via bond descriptors, making it an ideal choice for polymer representation.[163] Researchers have often used nanomaterials such as graphene oxide, titanium dioxide, and carbon nanotubes as additives to modify membranes. These materials are highly complex in terms of their physical and chemical properties, yet researchers usually represent them as categorical features in ML models which may lead to oversimplification of their effects on membrane performance.[26,72,164] The development of more robust techniques to represent these additives can allow algorithms to learn more nuances associated with their impact, aiding in the development of more accurate ML models.

Another significant challenge in the process of developing ML models is data management, data preprocessing, and data scarcity. Certain models, including DNN, are incapable of processing data sets that have missing values. Filling in missing values with generated data, such as utilizing mean or median values derived from statistical distributions or ML models, can lead to a loss of some of its useful practical insights about the model's performance.[64] Yuan et al. imputed gas permeability data in the polymer gas separation membrane database using a multivariate imputation by chained equations (MICE) method, which predicts missing permeability values through an iterative process via predictive models. The imputed data was used to identify promising polymeric materials for applications different from those for that were initially intended.[165] Data augmentation can expand the size and diversity of training data sets to address data scarcity commonly observed in polymer and membrane science.[166] Tayyebi et al. generated 300 new SMILES strings by randomizing atom ordering, increasing their data set size from the original 583 points to a total of 17,500 data points.[112] This data augmentation strategy helped the model train on different representations of the same molecule, enhancing its ability to grasp the chemical space limitations present in the data set.[167] Transfer learning can also address data scarcity by transferring knowledge from a model pretrained on a larger, related data set for new tasks.[168] Using transfer learning for a small data set of 12 membrane electrode assemblies (MEA), Tan et al. investigated the influences of anode catalyst ink formation on low-iridium membrane assemblies. By combining transfer learning with the Harris Hawk optimization, the authors significantly reduced experimental cost and achieved high-performance MEA, demonstrating the potential of transfer learning for materials optimization with small data sets.[169−171]

**5.2. Data Generation Using Computational Tools.** A major obstacle in designing polymeric materials using ML for membrane application arises from the diversity of polymer properties, which requires a substantial and high-quality data set for precise modeling. MD simulations and DFT are commonly used techniques implemented to simulate the behavior of polymeric materials at the molecular scale, assisting in the development of membranes with customized properties.[172] Wei et al. investigated the diffusive response of water in cross-linked polyamide membranes using MD simulation. Their results were consistent with the experimentally obtained flux values, suggesting that MD simulations can reliably surrogate for laboratory-performed experiments.[173] On the other hand, DFT assists in quantum mechanical calculation of the interactions between polymeric surface and salts or gaseous molecules. This can provide fundamental insights into the separation process and further help in the synthesis of selective membranes.[174,175]

Owing to the improvements in computational power, it is possible to run parallel computational simulations to generate data, which can be used for ML modeling. High-throughput computations have been used to generate vast databases to determine the crystalline and optoelectrical properties of polymers.[176,177] A scalable modeling and rapid theoretical (SMART) calculation approach has been presented that aims to combine high-throughput calculations with ML for the development of superior carbon capture materials.[178] This approach can be translated toward the ML process for the development of membrane materials. Tao et al. utilized high-throughput MD to simulate a large data set consisting of over 6,500 homopolymers and 1,400 polyamides to develop a ML model that determines the FFV of polymers.[103] Researchers used grand canonical Monte Carlo (GCMC) and MD simulations to train ML models with the capability to model the gas separation behavior for binary gas mixtures in mixed matrix membranes.[179,180] Meng et al. generated a data set containing 2D graphene-based membranes using CALYPSO, a structural prediction tool based on particle swarm optimization. Using this data set, the authors were able to screen membranes for desalination with superior flux, salt rejection,


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

and mechanical properties.[181,182] Zhang et al. augmented a polyamide NF data set having $10^2$ points to $10^4$ points by using a combination of vibrational augmentation and DFT calculations. They considered the 3D geometry of the monomer structure along with the chemical coupling of functional groups to spatially represent the monomer groups.[183]

Membrane selectivity is impacted by the molecular interactions between polymeric membrane functional groups and targeted species (solutes or gases). DFT models are often used to calculate the binding or adsorption energies which are commonly used to study this interaction behavior. With the advancement of high-throughput DFT, these interactions can be calculated to be used as input features to train ML models. Inclusion of these interactions can potentially aid in the design of selective membranes that allow favorable transport of targeted species. DFT calculations can also assist in modeling energy barriers associated with multiple solute selectivity, further helping in the design of ion selective membranes.[89,184,185] Thus, data generation using high-throughput computations holds significant potential for enhancing ML process and assist in the discovery of a superior class of polymeric membranes for various engineering applications.

**5.3. Data Extraction Using NLP and Large Language Models (LLMs).** Scientific research is growing at an unprecedented rate with thousands of publications, reports, and papers being published every year related to polymer, material, and membrane science. It is extremely difficult for scientists to keep up to date with these advances. NLP has emerged as a facilitative approach for information retrieval from literature in the past couple of years.[186,187] It enables a computer to understand, interpret, and generate human language, bridging the gap between human communication and digital data processing.[188] NLP facilitates the extraction of data from written texts in diverse formats, enabling efficient analysis and interpretation of large volumes of information. Information retrieval is the first step of NLP, which involves collection of papers of interest as PDF, HTML, or XML files.[189] Once the papers are obtained, the next step is to process the documents by cleaning up their text to remove irrelevant content and special characters. This step is followed by tokenization wherein the cleaned text is converted into individual names or broad concepts, which are suitable for NLP. Following tokenization, several methods can be used to extract data from the input text. Named entity recognition (NER) technique is utilized to focus on the identification and classification of entities within the text. This process is followed by word embedding wherein the text is transformed into word vectors that can be used for further information extraction.[190] LLMs such as GPT, BERT, or LLaMA are the latest advancement in this field and have shown great potential for information retrieval. Unlike NER, which is a multistep process requiring intermediate processing and classification of links between entities, LLMs can directly be used to transform input text into structured output data (as JSON documents or other hierarchical structures), allowing for ease in the data mining process.[191]

With the rapid advancement of NLP in the field of material science, there have been multiple instances where it is used for information retrieval.[192−195] Shetty et al. leveraged this growth in NLP and applied this knowledge to the field of polymer science as they extracted and processed data from ∼0.5 million publications.[196] They trained word vector models on the polymer literature corpus, encoding polymer domain knowledge in the vector space. The data extracted from the literature can facilitate the generation of training data for downstream ML models. In this study, unsupervised ML was used to identify application trends and generate meaningful information. The authors were able to establish relationships between monomer-polymer as well as property-polymer which helped them cluster polymers based on their properties (e.g., conductivity, biodegradability, and adhesiveness). Additionally, the authors demonstrated the capability of the model to predict polymers with new functions. For example, a model trained on a subset of data for a particular year could accurately predict the occurrence of the polymer for a completely different application in the subsequent years. This highlights the potential of NLP in polymer science, allowing researchers to uncover new insights and applications from a vast corpus of literature. Thus, by leveraging the data extracted and processed through NLP techniques, researchers can build more sophisticated ML models, enhancing the ability to derive meaningful insights and prediction from large data sets in polymer science.

Developing a fully automated, end-to-end ML model that regularly updates its data set from recently published literature through NLP will revolutionize the field. However, this comes with its own set of challenges, the biggest of which is to develop a pipeline that can effectively embed textual, written, or graphical data points into the correct features, especially in cases with large dimensionality.

**5.4. Collaborative Efforts and Open Data Sharing Initiatives.** Given that data serves as the fundamental basis for ML, it is crucial to address data-related challenges, such as the scarcity of adequate high-quality data or metadata. Various techniques can be employed to acquire data, including manual collection, high-throughput experiments or simulations, NLP, curated databases, and user populated databases. Each of these methods has its own specific factors to be considered.[197] Manual data collection is tedious in nature, while high-throughput experiments/simulations require specific expertise and can be time-consuming, resource-intensive, and interdependent.[198] Open access databases for polymeric membranes can offer a solution to the issues of data acquisition. These platforms play a crucial role in enabling the prediction and analysis of polymer properties, addressing the data scarcity problem by organizing extensive data sets from various sources.

The Open Membrane Database (OMD) is a great step for a collaboratory database in the field of membrane science, which allows researchers to create a centralized archive for thin-film composite RO membranes for water purification and desalination purposes. It has data from over 1,000 different types of polymeric membranes from peer reviewed journals and patents.[199] MSA and PoLyInfo consist of gas permeability data for at least one of the gases among He, $H_2$, $O_2$, $N_2$, $CO_2$, and $CH_4$ for around 800 homopolymers. These databases have already been used in ML studies to model the gas separation of polymeric membranes.[134,165] Additionally, OSN database is a repository which contains a collection of publications with their data sets for membrane applications such as NF, RO, and gas separation.[200] In the future, it is encouraged that researchers upload their data sets along with their publications as supplemental files or web sites (e.g., GitHub, Zenodo, and Figshare) facilitating convenient access and retrieval of experimental data.[30] By sharing data, resources, and expertise, authors can develop comprehensive databases containing


https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

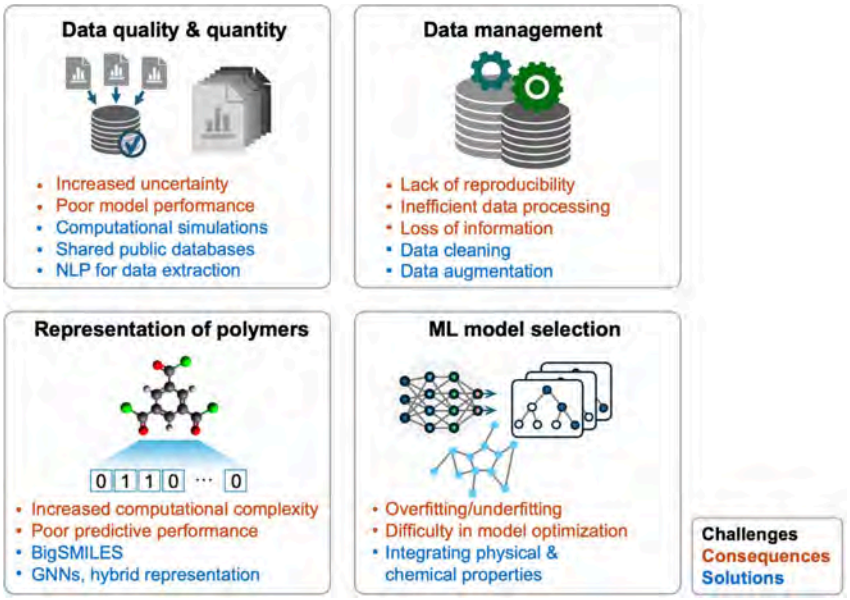



**Figure 3.** Challenges, consequences, and solutions related to the application of AI for polymeric membrane discovery.

information on material properties, synthesis methods, and performance metrics. Open data sharing initiatives foster innovation, reproducibility, and transparency by granting access to high-quality data sets for training and validating AI/ML models to researchers worldwide. Figure 3 gives a summary of the commonly faced challenges during the application of ML tools for membrane design, how they impact model development process, and potential approaches of dealing with these problems.

## 6. ENVIRONMENTAL IMPLICATIONS

The current research paradigm for membrane material discovery and development is largely driven by the direct design approach. Experimentally testing new polymeric materials is costly, resource-consuming, and challenging, which significantly slows down novel membrane design process. ML-aided design strategy largely relies on capturing key chemical structures with positive contributions to performance, which are used to screen top polymer candidates. This general-purpose framework can be applied to discover materials for environmental applications such as gas separation, water purification, energy generation, solvent and other resource recovery, carbon capture, which has significant real-world impact.

One of the biggest open challenges in the field is the development of polymeric membranes with high selectivity. ML aided-inverse design approach has provided a fit-for-purpose strategy to synthesize polymeric membranes with exceptional selectivity. The ability to tailor membrane selectivity depending on the application can assist in the removal of contaminants and micropollutants from wastewater, or the recovery of critical industry essential resources such as expensive solvents, nutrients, and minerals.[201,202] Using high-performance materials for pollutant removal from wastewater can improve process efficiency, reducing the reliance on chemical treatments and minimizing the release of contaminants into aquatic ecosystems.[203,204] With the advancement of computational data generation tools, researchers have the potential to tackle pressing issues such as the recovery of plant essential nutrients or critical metals with minimal reliance on experimental data. AI-driven membrane design enhances the efficiency of carbon capture and industrial gas purification, contributing to lower greenhouse gas emissions and improving sustainability efforts across energy-intensive industries.[205] Leveraging ML to screen candidates and design membranes can minimize material waste and reduce energy consumption through process optimization, which reduces the environmental footprint of membrane design and operation. Developing robust quantitative metrics to relate the chemical structure of polymers to its biodegradability can aid in greener synthesis routes; however, research in this area is still in its infancy.[206,207] Researchers need to ensure that the environmental footprint of membrane processes is minimized throughout their lifecycle, from production to end-of-life disposal. These advancements align with the broader goals of the circular economy by optimizing resource recovery and minimizing waste.[208]

It is not implied through this review that the direct-design approach is inferior to the ML-aided inverse-design approach. ML research benefits from the data generated through extensive experimentation as much as traditional experimentalists can benefit from a guided approach. The computational resources required for ML model development and high-throughput data generation are energy-intensive, potentially offsetting the environmental gains realized through optimized membrane performance and design. Broader adoption of ML in membrane design may require comprehensive life cycle assessments to ensure that the energy and material savings during membrane operation outweigh the carbon footprint incurred through AI computation.[209] Interdisciplinary cooperation among material scientists, chemists, physicists, computer scientists, and environmental engineers is crucial for addressing the issues and promoting innovation in the design and discovery of polymeric membranes in an environmentally sustainable way.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.4c08298.



https://doi.org/10.1021/acs.est.4c08298
Environ. Sci. Technol. 2025, 59, 993−1012

Details on traditional methods used for polymer design for membranes, examples of input and output features used for ML in polymeric membrane research, and commonly used ML algorithms in polymeric membrane research (PDF)

## AUTHOR INFORMATION


### Corresponding Author

**Yongsheng Chen** − *School of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0002-9519-2302; Email: yongsheng.chen@ce.gatech.edu

### Authors

**Raghav Dangayach** − *School of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

**Nohyeong Jeong** − *School of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

**Elif Demirel** − *School of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

**Nigmet Uzal** − *School of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States; Department of Civil Engineering, Abdullah Gul University, 38039 Kayseri, Turkey*

**Victor Fung** − *School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c08298

### Author Contributions

#R.D. and N.J. contributed equally to the work


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was partially supported by the U.S. Department of Agriculture (Award Nos. 2018-68011-28371, 2021-67021-34499, 2021-67021-38585, and 2024-67021-41534), National Science Foundation (Award Nos. 2112533, 2345543, and 2419122), and US Environmental Protection Agency (Award no. 840080010).

## REFERENCES


(1) Liu, M.-L.; Zhang, C.-X.; Tang, M.-J.; Sun, S.-P.; Xing, W.; Lee, Y. M. Evolution of Functional Nanochannel Membranes. *Prog. Mater. Sci.* **2023**, *139*, 101162.

(2) Sidhikku Kandath Valappil, R.; Ghasem, N.; Al-Marzouqi, M. Current and Future Trends in Polymer Membrane-Based Gas Separation Technology: A Comprehensive Review. *J. Ind. Eng. Chem.* **2021**, *98*, 103−129.

(3) Adam, M. R.; Othman, M. H. D.; Kurniawan, T. A.; Puteh, M. H.; Ismail, A. F.; Khongnakorn, W.; Rahman, M. A.; Jaafar, J. Advances in Adsorptive Membrane Technology for Water Treatment and Resource Recovery Applications: A Critical Review. *J. Environ. Chem. Eng.* **2022**, *10* (3), 107633.

(4) Ulbricht, M. Advanced Functional Polymer Membranes. *Polymer* **2006**, *47* (7), 2217−2262.

(5) Wu, J.; Xu, F.; Li, S.; Ma, P.; Zhang, X.; Liu, Q.; Fu, R.; Wu, D. Porous Polymers as Multifunctional Material Platforms toward Task-Specific Applications. *Adv. Mater.* **2019**, *31* (4), 1802922.

(6) Park, H. B.; Kamcev, J.; Robeson, L. M.; Elimelech, M.; Freeman, B. D. Maximizing the Right Stuff: The Trade-off between Membrane Permeability and Selectivity. *Science* **2017**, *356* (6343), No. eaab0530.

(7) Geise, G. M.; Park, H. B.; Sagle, A. C.; Freeman, B. D.; McGrath, J. E. Water Permeability and Water/Salt Selectivity Tradeoff in Polymers for Desalination. *J. Membr. Sci.* **2011**, *369* (1), 130−138.

(8) Kitto, D.; Kamcev, J. Predicting the Conductivity-Selectivity Trade-Off and Upper Bound in Ion-Exchange Membranes. *ACS Energy Lett.* **2024**, *9* (4), 1346−1352.

(9) Liu, Y.; Wang, K.; Zhou, Z.; Wei, X.; Xia, S.; Wang, X.; Xie, Y. F.; Huang, X. Boosting the Performance of Nanofiltration Membranes in Removing Organic Micropollutants: Trade-Off Effect, Strategy Evaluation, and Prospective Development. *Environ. Sci. Technol.* **2022**, *56* (22), 15220−15237.

(10) Li, J.; Lim, K.; Yang, H.; Ren, Z.; Raghavan, S.; Chen, P.-Y.; Buonassisi, T.; Wang, X. AI Applications through the Whole Life Cycle of Material Discovery. *Matter* **2020**, *3* (2), 393−432.

(11) Nunes, S. P.; Culfaz-Emecen, P. Z.; Ramon, G. Z.; Visser, T.; Koops, G. H.; Jin, W.; Ulbricht, M. Thinking the Future of Membranes: Perspectives for Advanced and New Membrane Materials and Manufacturing Processes. *J. Membr. Sci.* **2020**, *598*, 117761.

(12) Chen, H.; Zheng, Y.; Li, J.; Li, L.; Wang, X. AI for Nanomaterials Development in Clean Energy and Carbon Capture, Utilization and Storage (CCUS). *ACS Nano* **2023**, *17* (11), 9763−9792.

(13) Maleki, R.; Shams, S. M.; Chellehbari, Y. M.; Rezvantalab, S.; Jahromi, A. M.; Asadnia, M.; Abbassi, R.; Aminabhavi, T.; Razmjou, A. Materials Discovery of Ion-Selective Membranes Using Artificial Intelligence. *Commun. Chem.* **2022**, *5* (1), 1−13.

(14) Ridgway, H. F.; Orbell, J.; Gray, S. Molecular Simulations of Polyamide Membrane Materials Used in Desalination and Water Reuse Applications: Recent Developments and Future Prospects. *J. Membr. Sci.* **2017**, *524*, 436−448.

(15) Wang, J.; Wang, Y.; Chen, Y. Inverse Design of Materials by Machine Learning. *Materials* **2022**, *15* (5), 1811.

(16) Patra, T. K. Data-Driven Methods for Accelerating Polymer Design. *ACS Polym. Au* **2022**, *2* (1), 8−26.

(17) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Materiomics* **2017**, *3* (3), 159−177.

(18) Shetty, P.; Adeboye, A.; Gupta, S.; Zhang, C.; Ramprasad, R. Accelerating Materials Discovery for Polymer Solar Cells: Data-Driven Insights Enabled by Natural Language Processing. *Chem. Mater.* **2024**, *36* (16), 7676−7689.

(19) Jain, A.; Armstrong, C. D.; Joseph, V. R.; Ramprasad, R.; Qi, H. J. Machine-Guided Discovery of Acrylate Photopolymer Compositions. *ACS Appl. Mater. Interfaces* **2024**, *16* (14), 17992−18000.

(20) Sarker, I. H. AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. *SN Comput. Sci.* **2022**, *3* (2), 158.

(21) Jordan, M. I.; Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349* (6245), 255−260.

(22) Jeong, N.; Epsztein, R.; Wang, R.; Park, S.; Lin, S.; Tong, T. Exploring the Knowledge Attained by Machine Learning on Ion Transport across Polyamide Membranes Using Explainable Artificial Intelligence. *Environ. Sci. Technol.* **2023**, *57*, 17851.

(23) Jeong, N.; Park, S.; Mahajan, S.; Zhou, J.; Blotevogel, J.; Ying, L.; Tong, T.; Chen, Y. Elucidating Governing Factors of PFAS Removal by Polyamide Membranes Using Machine Learning and Molecular Simulations. *Nat. Commun.* **2024**, in press.

(24) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3* (1), 2810.

(25) 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.

(26) 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.
(27) 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.
(28) Ignacz, G.; Bader, L.; Beke, A. K.; Ghunaim, Y.; Shastry, T.; Vovusha, H.; Carbone, M. R.; Ghanem, B.; Szekely, G. Machine Learning for the Advancement of Membrane Science and Technology: A Critical Review. *J. Membr. Sci.* **2025**, *713*, 123256.
(29) Cao, Z.; Barati Farimani, O.; Ock, J.; Barati Farimani, A. Machine Learning in Membrane Design: From Property Prediction to AI-Guided Optimization. *Nano Lett.* **2024**, *24* (10), 2953−2960.
(30) Yin, H.; Xu, M.; Luo, Z.; Bi, X.; Li, J.; Zhang, S.; Wang, X. Machine Learning for Membrane Design and Discovery. *Green Energy Environ.* **2024**, *9* (1), 54−70.
(31) Osman, A. I.; Nasr, M.; Farghali, M.; Bakr, S. S.; Eltaweil, A. S.; Rashwan, A. K.; Abd El-Monaem, E. M. Machine Learning for Membrane Design in Energy Production, Gas Separation, and Water Treatment: A Review. *Environ. Chem. Lett.* **2024**, *22* (2), 505−560.
(32) Wang, J.; Tian, K.; Li, D.; Chen, M.; Feng, X.; Zhang, Y.; Wang, Y.; Van der Bruggen, B. Machine Learning in Gas Separation Membrane Developing: Ready for Prime Time. *Sep. Purif. Technol.* **2023**, *313*, 123493.
(33) Lin, H.; Ding, Y. Polymeric Membranes: Chemistry, Physics, and Applications. *J. Polym. Sci.* **2020**, *58* (18), 2433−2434.
(34) Geise, G. M.; Lee, H.-S.; Miller, D. J.; Freeman, B. D.; McGrath, J. E.; Paul, D. R. Water Purification by Membranes: The Role of Polymer Science. *J. Polym. Sci., Part B: Polym. Phys.* **2010**, *48* (15), 1685−1718.
(35) Wu, L.; Sun, J.; He, C. Effects of Solvent Sort, PES and PVP Concentration on the Properties and Morphology of PVDF/PES Blend Hollow Fiber Membranes. *J. Appl. Polym. Sci.* **2010**, *116* (3), 1566−1573.
(36) Webb, M. T.; Condes, L. C.; Ly, H. G.; Galizia, M.; Razavi, S. Rational Design, Synthesis, and Characterization of Facilitated Transport Membranes Exhibiting Enhanced Permeability, Selectivity and Stability. *J. Membr. Sci.* **2023**, *685*, 121910.
(37) Paul, M.; Jons, S. D. Chemistry and Fabrication of Polymeric Nanofiltration Membranes: A Review. *Polymer* **2016**, *103*, 417−456.
(38) Khayet, M.; Cojocaru, C.; García-Payo, C. Application of Response Surface Methodology and Experimental Design in Direct Contact Membrane Distillation. *Ind. Eng. Chem. Res.* **2007**, *46* (17), 5673−5685.
(39) Khayet, M.; Cojocaru, C.; García-Payo, M. C. Experimental Design and Optimization of Asymmetric Flat-Sheet Membranes Prepared for Direct Contact Membrane Distillation. *J. Membr. Sci.* **2010**, *351* (1), 234−245.
(40) Cojocaru, C.; Zakrzewska-Trznadel, G.; Jaworska, A. Removal of Cobalt Ions from Aqueous Solutions by Polymer Assisted Ultrafiltration Using Experimental Design Approach. Part 1: Optimization of Complexation Conditions. *J. Hazard. Mater.* **2009**, *169* (1), 599−609.
(41) Zhao, C.; Xu, X.; Chen, J.; Yang, F. Optimization of Preparation Conditions of Poly(Vinylidene Fluoride)/Graphene Oxide Microfiltration Membranes by the Taguchi Experimental Design. *Desalination* **2014**, *334* (1), 17−22.
(42) Suhaimi, N. H.; Yeong, Y. F.; Jusoh, N.; Chew, T. L.; Bustam, M. A.; Mubashir, M. RSM Modeling and Optimization of CO2 Separation from High CO2 Feed Concentration over Functionalized Membrane. *Polymers* **2022**, *14* (7), 1371.
(43) Mohammadi, T.; Safavi, M. A. Application of Taguchi Method in Optimization of Desalination by Vacuum Membrane Distillation. *Desalination* **2009**, *249* (1), 83−89.
(44) Onsekizoglu, P.; Savas Bahceci, K.; Acar, J. The Use of Factorial Design for Modeling Membrane Distillation. *J. Membr. Sci.* **2010**, *349* (1), 225−230.
(45) Mahmud, N. A. C.; Abu Seman, M. N.; Takriff, M. S.; Ang, W. L.; Saufi, S. M. Influence of Dope Composition of Polyethersulfone Membrane Blend with Cellulose Nanocrystal and Carboxylated Multi-Walled Carbon Nanotube on Humic Acid Rejection. *Mater. Today Proc.* **2023**, DOI: 10.1016/j.matpr.2023.07.006.
(46) Khayet, M.; Cojocaru, C.; Essalhi, M.; García-Payo, M. C.; Arribas, P.; García-Fernández, L. Hollow Fiber Spinning Experimental Design and Analysis of Defects for Fabrication of Optimized Membranes for Membrane Distillation. *Desalination* **2012**, *287*, 146−158.
(47) Valtcheva, I. B.; Marchetti, P.; Livingston, A. G. Crosslinked Polybenzimidazole Membranes for Organic Solvent Nanofiltration (OSN): Analysis of Crosslinking Reaction Mechanism and Effects of Reaction Parameters. *J. Membr. Sci.* **2015**, *493*, 568−579.
(48) Azizi, J.; Sharif, A. Optimization of Water Flux and Salt Rejection Properties of Polyamide Thin Film Composite Membranes. *J. Appl. Polym. Sci.* **2020**, *137* (28), 48858.
(49) Khayet, M.; Cojocaru, C.; Zakrzewska-Trznadel, G. Response Surface Modelling and Optimization in Pervaporation. *J. Membr. Sci.* **2008**, *321* (2), 272−283.
(50) Lee, B. C. Y.; Mahtab, M. S.; Neo, T. H.; Farooqi, I. H.; Khursheed, A. A Comprehensive Review of Design of Experiment (DOE) for Water and Wastewater Treatment Application - Key Concepts, Methodology and Contextualized Application. *J. Water Process Eng.* **2022**, *47*, 102673.
(51) Mohammadi, M.; Mohammadi, N.; Mehdipour-Ataei, S. On the Preparation of Thin Nanofibers of Polysulfone Polyelectrolyte for Improving Conductivity of Proton-Exchange Membranes by Electrospinning: Taguchi Design, Response Surface Methodology, and Genetic Algorithm. *Int. J. Hydrog. Energy* **2020**, *45* (58), 34110−34124.
(52) Malykh, O. V.; Golub, A. Yu.; Teplyakov, V. V. Polymeric Membrane Materials: New Aspects of Empirical Approaches to Prediction of Gas Permeability Parameters in Relation to Permanent Gases, Linear Lower Hydrocarbons and Some Toxic Gases. *Adv. Colloid Interface Sci.* **2011**, *164* (1), 89−99.
(53) Alqaheem, Y.; Alomair, A. A. Microscopy and Spectroscopy Techniques for Characterization of Polymeric Membranes. *Membranes* **2020**, *10* (2), 33.
(54) Wijmans, J. G.; Baker, R. W. The Solution-Diffusion Model: A Review. *J. Membr. Sci.* **1995**, *107* (1), 1−21.
(55) Luo, T.; Abdu, S.; Wessling, M. Selectivity of Ion Exchange Membranes: A Review. *J. Membr. Sci.* **2018**, *555*, 429−454.
(56) Wang, L.; Cao, T.; Dykstra, J. E.; Porada, S.; Biesheuvel, P. M.; Elimelech, M. Salt and Water Transport in Reverse Osmosis Membranes: Beyond the Solution-Diffusion Model. *Environ. Sci. Technol.* **2021**, *55* (24), 16665−16675.
(57) Wang, R.; Lin, S. Pore Model for Nanofiltration: History, Theoretical Framework, Key Predictions, Limitations, and Prospects. *J. Membr. Sci.* **2021**, *620*, 118809.
(58) Tavakolmoghadam, M.; Mokhtare, A.; Rekabdar, F.; Esmaeili, M.; khaneghah, A. h. k. A Predictive Model for Tuning Additives for the Fabrication of Porous Polymeric Membranes. *Mater. Res. Express* **2020**, *7* (1), 015312.
(59) Karunakaran, A.; Chaturvedi, A.; Ali, J.; Singh, R.; Agarwal, S.; Garg, M. C. Response Surface Methodology-Based Modeling and Optimization of Chromium Removal Using Spiral-Wound Reverse-Osmosis Membrane Setup. *Int. J. Environ. Sci. Technol.* **2022**, *19* (7), 5999−6010.
(60) Xu, Q.; Jiang, J. Recent Development in Machine Learning of Polymer Membranes for Liquid Separation. *Mol. Syst. Des. Eng.* **2022**, *7* (8), 856−872.
(61) Tayyebi, A.; Alshami, A. S.; Yu, X.; Kolodka, E. Can Machine Learning Methods Guide Gas Separation Membranes Fabrication? *J. Membr. Sci. Lett.* **2022**, *2* (2), 100033.
(62) Talukder, M. J.; Alshami, A. S.; Tayyebi, A.; Ismail, N.; Yu, X. Membrane Science Meets Machine Learning: Future and Potential Use in Assisting Membrane Material Design and Fabrication. *Sep. Purif. Rev.* **2024**, *53* (2), 216−229.

(63) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer Informatics: Current Status and Critical next Steps. *Mater. Sci. Eng. R Rep.* **2021**, *144*, 100595.
(64) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57* (46), 17671−17689.
(65) Glass, S.; Schmidt, M.; Merten, P.; Abdul Latif, A.; Fischer, K.; Schulze, A.; Friederich, P.; Filiz, V. Design of Modified Polymer Membranes Using Machine Learning. *ACS Appl. Mater. Interfaces* **2024**, *16* (16), 20990−21000.
(66) Cano-Odena, A.; Spilliers, M.; Dedroog, T.; De Grave, K.; Ramon, J.; Vankelecom, I. F. J. Optimization of Cellulose Acetate Nanofiltration Membranes for Micropollutant Removal via Genetic Algorithms and High Throughput Experimentation. *J. Membr. Sci.* **2011**, *366* (1), 25−32.
(67) Ignacz, G.; Beke, A. K.; Szekely, G. Data-Driven Investigation of Process Solvent and Membrane Material on Organic Solvent Nanofiltration. *J. Membr. Sci.* **2023**, *674*, 121519.
(68) Fetanat, M.; Keshtiara, M.; Keyikoglu, R.; Khataee, A.; Daiyan, R.; Razmjou, A. Machine Learning for Design of Thin-Film Nanocomposite Membranes. *Sep. Purif. Technol.* **2021**, *270*, 118383.
(69) Fetanat, M.; Keshtiara, M.; Low, Z.-X.; Keyikoglu, R.; Khataee, A.; Orooji, Y.; Chen, V.; Leslie, G.; Razmjou, A. Machine Learning for Advanced Design of Nanocomposite Ultrafiltration Membranes. *Ind. Eng. Chem. Res.* **2021**, *60* (14), 5236−5250.
(70) Zheng, W.; Chen, Y.; Xu, X.; Peng, X.; Niu, Y.; Xu, P.; Li, T. Research on the Factors Influencing Nanofiltration Membrane Fouling and the Prediction of Membrane Fouling. *J. Water Process Eng.* **2024**, *59*, 104876.
(71) Yao, L.; Zhang, Z.; Li, Y.; Zhuo, J.; Chen, Z.; Lin, Z.; Liu, H.; Yao, Z. Precise Prediction of CO2 Separation Performance of Metal-Organic Framework Mixed Matrix Membranes Based on Feature Selection and Machine Learning. *Sep. Purif. Technol.* **2024**, *349*, 127894.
(72) Gao, H.; Zhong, S.; Dangayach, R.; Chen, Y. Understanding and Designing a High-Performance Ultrafiltration Membrane Using Machine Learning. *Environ. Sci. Technol.* **2023**, *57*, 17831.
(73) Li, H.; Zeng, B.; Qiu, T.; Huang, W.; Wang, Y.; Sheng, G.-P.; Wang, Y. Deep Learning Models for Assisted Decision-Making in Performance Optimization of Thin Film Nanocomposite Membranes. *J. Membr. Sci.* **2023**, *687*, 122093.
(74) Li, H.; Wang, Y.; Wang, Y. Machine Learning for Predicting the Dynamic Extraction of Multiple Substances by Emulsion Liquid Membranes. *Sep. Purif. Technol.* **2023**, *313*, 123458.
(75) Chollet, F. *Deep Learning with Python, Second ed.*; Simon and Schuster, 2021.
(76) Patel, R. A.; Borca, C. H.; Webb, M. A. Featurization Strategies for Polymer Sequence or Composition Design by Machine Learning. *Mol. Syst. Des. Eng.* **2022**, *7* (6), 661−676.
(77) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.
(78) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97−101.
(79) Cencer, M. M.; Moore, J. S.; Assary, R. S. Machine Learning for Polymeric Materials: An Introduction. *Polym. Int.* **2022**, *71* (5), 537−542.
(80) Saini, V. Machine Learning Prediction of Empirical Polarity Using SMILES Encoding of Organic Solvents. *Mol. Divers.* **2023**, *27* (5), 2331−2343.
(81) Shastry, T.; Basdogan, Y.; Wang, Z.-G.; Kumar, S. K.; Carbone, M. R. Machine Learning-Based Discovery of Molecular Descriptors That Control Polymer Gas Permeation. *J. Membr. Sci.* **2024**, *697*, 122563.
(82) 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* (46), 18193−18202.
(83) Xu, P.; Chen, H.; Li, M.; Lu, W. New Opportunity: Machine Learning for Polymer Materials Design and Discovery. *Adv. Theory Simul.* **2022**, *5* (5), 2100565.
(84) Ismaeel, H.; Gibson, D.; Ricci, E.; De Angelis, M. G. Estimating Gas Sorption In Polymeric Membranes From The Molecular Structure: A Machine Learning Based Group Contribution Method For The Non-Equilibrium Lattice Fluid Model (ML-GC-NELF). *J. Membr. Sci.* **2024**, *691*, 122220.
(85) Zhao, M.; Zhang, C.; Weng, Y. Improved Artificial Neural Networks (ANNs) for Predicting the Gas Separation Performance of Polyimides. *J. Membr. Sci.* **2023**, *681*, 121765.
(86) Kuenneth, C.; Ramprasad, R. polyBERT: A Chemical Language Model to Enable Fully Machine-Driven Ultrafast Polymer Informatics. *Nat. Commun.* **2023**, *14* (1), 4099.
(87) Gong, W.; Xu, H.; Lu, J.; Kim, J.; Zhao, Y.; Li, N.; Zhang, Y.; Yang, J.; Xu, D.; Liang, H. Gradient Boosting Decision Tree Algorithms for Accelerating Nanofiltration Membrane Design and Discovery. *Desalination* **2024**, *592*, 118072.
(88) Zhao, Y.; Mulder, R. J.; Houshyar, S.; Le, T. C. A Review on the Application of Molecular Descriptors and Machine Learning in Polymer Design. *Polym. Chem.* **2023**, *14* (29), 3325−3346.
(89) Ritt, C. L.; Liu, M.; Pham, T. A.; Epsztein, R.; Kulik, H. J.; Elimelech, M. Machine Learning Reveals Key Ion Selectivity Mechanisms in Polymeric Membranes with Subnanometer Pores. *Sci. Adv.* **2022**, *8* (2), No. eabl5771.
(90) Ignacz, G.; Yang, C.; Szekely, G. Diversity Matters: Widening the Chemical Space in Organic Solvent Nanofiltration. *J. Membr. Sci.* **2022**, *641*, 119929.
(91) Zhai, F.-H.; Zhan, Q.-Q.; Yang, Y.-F.; Ye, N.-Y.; Wan, R.-Y.; Wang, J.; Chen, S.; He, R.-H. A Deep Learning Protocol for Analyzing and Predicting Ionic Conductivity of Anion Exchange Membranes. *J. Membr. Sci.* **2022**, *642*, 119983.
(92) Phua, Y. K.; Fujigaya, T.; Kato, K. Predicting the Anion Conductivities and Alkaline Stabilities of Anion Conducting Membrane Polymeric Materials: Development of Explainable Machine Learning Models. *Sci. Technol. Adv. Mater.* **2023**, *24* (1), 2261833.
(93) Shahmansouri, A.; Bellona, C. Application of Quantitative Structure-Property Relationships (QSPRs) to Predict the Rejection of Organic Solutes by Nanofiltration. *Sep. Purif. Technol.* **2013**, *118*, 627−638.
(94) Yangali-Quintanilla, V.; Verliefde, A.; Kim, T.-U.; Sadmani, A.; Kennedy, M.; Amy, G. Artificial Neural Network Models Based on QSAR for Predicting Rejection of Neutral Organic Compounds by Polyamide Nanofiltration and Reverse Osmosis Membranes. *J. Membr. Sci.* **2009**, *342* (1), 251−262.
(95) 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.
(96) Webb, M. T.; Condes, L. C.; Box, W. J.; Ly, H. G.; Razavi, S.; Galizia, M. Revisiting Experimental Techniques and Theoretical Models for Estimating the Solubility Parameter of Rubbery and Glassy Polymer Membranes. *J. Membr. Sci. Lett.* **2023**, *3* (2), 100060.
(97) Bagheri, M.; Akbari, A.; Mirbagheri, S. A. Advanced Control of Membrane Fouling in Filtration Systems Using Artificial Intelligence and Machine Learning Techniques: A Critical Review. *Process Saf. Environ. Prot.* **2019**, *123*, 229−252.
(98) Phua, Y. K.; Terasoba, N.; Tanaka, M.; Fujigaya, T.; Kato, K. Unsupervised Machine Learning-Derived Anion-Exchange Membrane Polymers Map: A Guideline for Polymers Exploration and Design. *ChemElectroChem.* **2024**, *11* (14), No. e202400252.
(99) Zou, X.; Xu, G.; Fang, P.; Li, W.; Jin, Z.; Guo, S.; Hu, Y.; Li, M.; Pan, J.; Sun, Z.; Yan, F. Unsupervised Learning-Guided Accelerated Discovery of Alkaline Anion Exchange Membranes for Fuel Cells. *Angew. Chem.* **2023**, *135* (19), No. e202300388.


https://doi.org/10.1021/acs.est.4c08298
*Environ. Sci. Technol.* 2025, 59, 993−1012

(100) Lee, S.; Shirts, M. R.; Straub, A. P. Molecular Fingerprint-Aided Prediction of Organic Solute Rejection in Reverse Osmosis and Nanofiltration. *J. Membr. Sci.* **2024**, *705*, 122927.
(101) Shi, F.; Lu, S.; Gu, J.; Lin, J.; Zhao, C.; You, X.; Lin, X. Modeling and Evaluation of the Permeate Flux in Forward Osmosis Process with Machine Learning. *Ind. Eng. Chem. Res.* **2022**, *61* (49), 18045−18056.
(102) Zhu, T.; Zhang, Y.; Tao, C.; Chen, W.; Cheng, H. Prediction of Organic Contaminant Rejection by Nanofiltration and Reverse Osmosis Membranes Using Interpretable Machine Learning Models. *Sci. Total Environ.* **2023**, *857*, 159348.
(103) 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*, 121131.
(104) Gallage Dona, H. K.; Olayiwola, T.; Briceno-Mena, L. A.; Arges, C. G.; Kumar, R.; Romagnoli, J. A. Determining Ion Activity Coefficients in Ion-Exchange Membranes with Machine Learning and Molecular Dynamics Simulations. *Ind. Eng. Chem. Res.* **2023**, *62* (24), 9533−9548.
(105) Ma, X.; Lu, D.; Lu, J.; Qian, Y.; Zhang, S.; Yao, Z.; Liang, L.; Sun, Z.; Zhang, L. Revealing Key Structural and Operating Features on Water/Salts Selectivity of Polyamide Nanofiltration Membranes by Ensemble Machine Learning. *Desalination* **2023**, *548*, 116293.
(106) Guan, J.; Huang, T.; Liu, W.; Feng, F.; Japip, S.; Li, J.; Wu, J.; Wang, X.; Zhang, S. Design and Prediction of Metal Organic Framework-Based Mixed Matrix Membranes for CO2 Capture via Machine Learning. *Cell Rep. Phys. Sci.* **2022**, *3* (5), 100864.
(107) Deng, H.; Luo, Z.; Imbrogno, J.; Swenson, T. M.; Jiang, Z.; Wang, X.; Zhang, S. Machine Learning Guided Polyamide Membrane with Exceptional Solute-Solute Selectivity and Permeance. *Environ. Sci. Technol.* **2023**, *57*, 17841.
(108) Lu, D.; Ma, X.; Lu, J.; Qian, Y.; Geng, Y.; Wang, J.; Yao, Z.; Liang, L.; Sun, Z.; Liang, S.; Zhang, L. Ensemble Machine Learning Reveals Key Structural and Operational Features Governing Ion Selectivity of Polyamide Nanofiltration Membranes. *Desalination* **2023**, *564*, 116748.
(109) 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*, 122320.
(110) Mollahosseini, A.; Abdelrasoul, A. Molecular Dynamics Simulation for Membrane Separation and Porous Materials: A Current State of Art Review. *J. Mol. Graph. Model.* **2021**, *107*, 107947.
(111) Xu, Q.; Gao, J.; Feng, F.; Chung, T.-S.; Jiang, J. Synergizing Machine Learning, Molecular Simulation and Experiment to Develop Polymer Membranes for Solvent Recovery. *J. Membr. Sci.* **2023**, *678*, 121678.
(112) Tayyebi, A.; Alshami, A. S.; Tayyebi, E.; Buelke, C.; Talukder, M. J.; Ismail, N.; Al-Goraee, A.; Rabiei, Z.; Yu, X. Machine Learning - Driven Surface Grafting of Thin-Film Composite Reverse Osmosis (TFC-RO) Membrane. *Desalination* **2024**, *579*, 117502.
(113) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360−365.
(114) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6* (1), 20952.
(115) Huang, X.; Zhao, C. Y.; Wang, H.; Ju, S. AI-Assisted Inverse Design of Sequence-Ordered High Intrinsic Thermal Conductivity Polymers. *Mater. Today Phys.* **2024**, *44*, 101438.
(116) Kumar, J. N.; Li, Q.; Tang, K. Y. T.; Buonassisi, T.; Gonzalez-Oyarce, A. L.; Ye, J. Machine Learning Enables Polymer Cloud-Point Engineering via Inverse Design. *Npj Comput. Mater.* **2019**, *5* (1), 1−6.
(117) Mannodi-Kanakkithodi, A.; Pilania, G.; Ramprasad, R.; Lookman, T.; Gubernatis, J. E. Multi-Objective Optimization Techniques to Design the Pareto Front of Organic Dielectric Polymers. *Comput. Mater. Sci.* **2016**, *125*, 92−99.
(118) Hanaoka, K. Bayesian Optimization for Goal-Oriented Multi-Objective Inverse Material Design. *iScience* **2021**, *24* (7), 102781.
(119) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminformatics* **2015**, *7* (1), 20.
(120) Wang, C.; Wang, L.; Yu, H.; Soo, A.; Wang, Z.; Rajabzadeh, S.; Ni, B.-J.; Shon, H. K. Machine Learning for Layer-by-Layer Nanofiltration Membrane Performance Prediction and Polymer Candidate Exploration. *Chemosphere* **2024**, *350*, 140999.
(121) 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.
(122) Yang, M.; Zhu, J.-J.; McGaughey, A. L.; Priestley, R. D.; Hoek, E. M. V.; Jassby, D.; Ren, Z. J. Machine Learning for Polymer Design to Enhance Pervaporation-Based Organic Recovery. *Environ. Sci. Technol.* **2024**, *58* (23), 10128−10139.
(123) Liu, L.; Li, Y.; Zheng, J.; Li, H. Expert-Augmented Machine Learning to Accelerate the Discovery of Copolymers for Anion Exchange Membrane. *J. Membr. Sci.* **2024**, *693*, 122327.
(124) Jin, Y.; Kumar, P. V. Bayesian Optimisation for Efficient Material Discovery: A Mini Review. *Nanoscale* **2023**, *15* (26), 10975−10984.
(125) Wang, K.; Dowling, A. W. Bayesian Optimization for Chemical Products and Functional Materials. *Curr. Opin. Chem. Eng.* **2022**, *36*, 100728.
(126) Gao, H.; Zhong, S.; Zhang, W.; Igou, T.; Berger, E.; Reid, E.; Zhao, Y.; Lambeth, D.; Gan, L.; Afolabi, M. A.; Tong, Z.; Lan, G.; Chen, Y. Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization. *Environ. Sci. Technol.* **2022**, *56* (4), 2572−2581.
(127) Chen, L.; Liu, G.; Zhang, Z.; Wang, Y.; Yang, Y.; Li, J. Machine Learning and Molecular Design Algorithm Assisted Discovery of Gas Separation Membranes Exceeding the CO2/CH4 and CO2/N2 Upper Bounds. *Chem. Eng. Sci.* **2024**, *291*, 119952.
(128) Gui, Y.; Zhan, D.; Li, T. Taking Another Step: A Simple Approach to High-Dimensional Bayesian Optimization. *Inf. Sci.* **2024**, *679*, 121056.
(129) Dalal, R. J.; Oviedo, F.; Leyden, M. C.; Reineke, T. M. Polymer Design via SHAP and Bayesian Machine Learning Optimizes pDNA and CRISPR Ribonucleoprotein Delivery. *Chem. Sci.* **2024**, *15* (19), 7219−7228.
(130) Menon, D.; Ranganathan, R. A Generative Approach to Materials Discovery, Design, and Optimization. *ACS Omega* **2022**, *7* (30), 25958−25973.
(131) Peng, J.; Schwalbe-Koda, D.; Akkiraju, K.; Xie, T.; Giordano, L.; Yu, Y.; Eom, C. J.; Lunger, J. R.; Zheng, D. J.; Rao, R. R.; Muy, S.; Grossman, J. C.; Reuter, K.; Gómez-Bombarelli, R.; Shao-Horn, Y. Human- and Machine-Centred Designs of Molecules and Materials for Sustainability and Decarbonization. *Nat. Rev. Mater.* **2022**, *7* (12), 991−1009.
(132) Jiang, S.; Dieng, A. B.; Webb, M. A. Property-Guided Generation of Complex Polymer Topologies Using Variational Autoencoders. *Npj Comput. Mater.* **2024**, *10* (1), 1−13.
(133) Kadulkar, S.; Sherman, Z. M.; Ganesan, V.; Truskett, T. M. Machine Learning-Assisted Design of Material Properties. *Annu. Rev. Chem. Biomol. Eng.* **2022**, *13* (1), 235−254.
(134) Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine Learning Enables Interpretable Discovery of Innovative Polymers for Gas Separation Membranes. *Sci. Adv.* **2022**, *8* (29), No. eabn9545.
(135) Giro, R.; Hsu, H.; Kishimoto, A.; Hama, T.; Neumann, R. F.; Luan, B.; Takeda, S.; Hamada, L.; Steiner, M. B. AI Powered, Automated Discovery of Polymer Membranes for Carbon Capture. *Npj Comput. Mater.* **2023**, *9* (1), 1−11.
(136) Takeda, S.; Hama, T.; Hsu, H.-H.; Piunova, V. A.; Zubarev, D.; Sanders, D. P.; Pitera, J. W.; Kogoh, M.; Hongo, T.; Cheng, Y.; Bocanett, W.; Nakashika, H.; Fujita, A.; Tsuchiya, Y.; Hino, K.; Yano, K.; Hirose, S.; Toda, H.; Orii, Y.; Nakano, D. Molecular Inverse-Design Platform for Material Industries. *KDD '20: Proceedings of the*


https://doi.org/10.1021/acs.est.4c08298
*Environ. Sci. Technol.* 2025, 59, 993−1012

*26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* **2020**, 2961−2969.
(137) Basdogan, Y.; Pollard, D. R.; Shastry, T.; Carbone, M. R.; Kumar, S. K.; Wang, Z.-G. Machine Learning-Guided Discovery of Polymer Membranes for CO2 Separation with Genetic Algorithm. *J. Membr. Sci.* **2024**, *712*, 123169.
(138) Wang, M.; Shi, G. M.; Zhao, D.; Liu, X.; Jiang, J. Machine Learning-Assisted Design of Thin-Film Composite Membranes for Solvent Recovery. *Environ. Sci. Technol.* **2023**, *57*, 15914.
(139) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122* (31), 17575−17585.
(140) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20* (1), 61−80.
(141) Ignacz, G.; Alqadhi, N.; Szekely, G. Explainable Machine Learning for Unraveling Solvent Effects in Polyimide Organic Solvent Nanofiltration Membranes. *Adv. Membr.* **2023**, *3*, 100061.
(142) Queen, O.; McCarver, G. A.; Thatigotla, S.; Abolins, B. P.; Brown, C. L.; Maroulas, V.; Vogiatzis, K. D. Polymer Graph Neural Networks for Multitask Property Learning. *Npj Comput. Mater.* **2023**, *9* (1), 1−10.
(143) Cui, R.; Zhang, Z.; Yu, C.; Zhou, Y. The Nanostructure of Ion Channels of Thin PFSA Membrane in the Catalyst Layer: A Molecular Dynamics Simulation Study Combined with Unsupervised Machine Learning. *J. Membr. Sci.* **2024**, *705*, 122904.
(144) Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: A Transformer-Based Language Model for Polymer Property Predictions. *Npj Comput. Mater.* **2023**, *9* (1), 1−14.
(145) Zhong, S.; Guan, X. Developing Quantitative Structure-Activity Relationship (QSAR) Models for Water Contaminants' Activities/Properties by Fine-Tuning GPT-3 Models. *Environ. Sci. Technol. Lett.* **2023**, *10* (10), 872−877.
(146) Rehman, D.; Lienhard, J. H. Physics-Informed Deep Learning for Multi-Species Membrane Separations. *Chem. Eng. J.* **2024**, *485*, 149806.
(147) 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* (1), 4931.
(148) Wang, M.; Jiang, J. Accelerating Discovery of Polyimides with Intrinsic Microporosity for Membrane-Based Gas Separation: Synergizing Physics-Informed Performance Metrics and Active Learning. *Adv. Funct. Mater.* **2024**, *34* (23), 2314683.
(149) Bradford, G.; Lopez, J.; Ruza, J.; Stolberg, M. A.; Osterude, R.; Johnson, J. A.; Gomez-Bombarelli, R.; Shao-Horn, Y. Chemistry-Informed Machine Learning for Polymer Electrolyte Discovery. *ACS Cent. Sci.* **2023**, *9* (2), 206−216.
(150) Palizhati, A.; Torrisi, S. B.; Aykol, M.; Suram, S. K.; Hummelshøj, J. S.; Montoya, J. H. Agents for Sequential Learning Using Multiple-Fidelity Data. *Sci. Rep.* **2022**, *12* (1), 4694.
(151) Phan, B. K.; Shen, K.-H.; Gurnani, R.; Tran, H.; Lively, R.; Ramprasad, R. Gas Permeability, Diffusivity, and Solubility in Polymers: Simulation-Experiment Data Fusion and Multi-Task Machine Learning. *Npj Comput. Mater.* **2024**, *10* (1), 1−11.
(152) Rall, D.; Schweidtmann, A. M.; Kruse, M.; Evdochenko, E.; Mitsos, A.; Wessling, M. Multi-Scale Membrane Process Optimization with High-Fidelity Ion Transport Models through Machine Learning. *J. Membr. Sci.* **2020**, *608*, 118208.
(153) Lazin, M. F.; Shelton, C. R.; Sandhofer, S. N.; Wong, B. M. High-Dimensional Multi-Fidelity Bayesian Optimization for Quantum Control. *Mach. Learn. Sci. Technol.* **2023**, *4* (4), 045014.
(154) Wang, Y.; Zhao, H.; Sciabola, S.; Wang, W. cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation. *Molecules* **2023**, *28* (11), 4430.
(155) Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.* **2022**, *62* (9), 2064−2076.
(156) Xu, Z.; Lei, X.; Ma, M.; Pan, Y. Molecular Generation and Optimization of Molecular Properties Using a Transformer Model. *Big Data Min. Anal.* **2024**, *7* (1), 142−155.
(157) Huang, L.; Xu, T.; Yu, Y.; Zhao, P.; Chen, X.; Han, J.; Xie, Z.; Li, H.; Zhong, W.; Wong, K.-C.; Zhang, H. A Dual Diffusion Model Enables 3D Molecule Generation and Lead Optimization Based on Target Pockets. *Nat. Commun.* **2024**, *15* (1), 2657.
(158) Lyngby, P.; Thygesen, K. S. Data-Driven Discovery of 2D Materials by Deep Generative Models. *Npj Comput. Mater.* **2022**, *8* (1), 1−8.
(159) Alverson, M.; Baird, S. G.; Murdock, R.; Ho, E. S.-H.; Johnson, J.; Sparks, T. D. Generative Adversarial Networks and Diffusion Models in Material Discovery. *Digit. Discovery* **2024**, *3* (1), 62−80.
(160) Park, J.; Gill, A. P. S.; Moosavi, S. M.; Kim, J. Inverse Design of Porous Materials: A Diffusion Model Approach. *J. Mater. Chem. A* **2024**, *12* (11), 6507−6514.
(161) Amamoto, Y. Data-Driven Approaches for Structure-Property Relationships in Polymer Science for Prediction and Understanding. *Polym. J.* **2022**, *54* (8), 957−967.
(162) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *Npj Comput. Mater.* **2017**, *3* (1), 1−13.
(163) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5* (9), 1523−1531.
(164) Liu, T.; Liu, L.; Cui, F.; Ding, F.; Zhang, Q.; Li, Y. Predicting the Performance of Polyvinylidene Fluoride, Polyethersulfone and Polysulfone Filtration Membranes Using Machine Learning. *J. Mater. Chem. A* **2020**, *8* (41), 21862−21871.
(165) Yuan, Q.; Longo, M.; Thornton, A. W.; McKeown, N. B.; Comesaña-Gándara, B.; Jansen, J. C.; Jelfs, K. E. Imputation of Missing Gas Permeability Data for Polymer Membranes Using Machine Learning. *J. Membr. Sci.* **2021**, *627*, 119207.
(166) Addis, B.; Castel, C.; Macali, A.; Misener, R.; Piccialli, V. Data Augmentation Driven by Optimization for Membrane Separation Process Synthesis. *Comput. Chem. Eng.* **2023**, *177*, 108342.
(167) Lo, S.; Seifrid, M.; Gaudin, T.; Aspuru-Guzik, A. Augmenting Polymer Datasets by Iterative Rearrangement. *J. Chem. Inf. Model.* **2023**, *63* (14), 4266−4276.
(168) Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M. A. Transfer Learning: A Friendly Introduction. *J. Big Data* **2022**, *9* (1), 102.
(169) Briceno-Mena, L. A.; Romagnoli, J. A.; Arges, C. G. PemNet: A Transfer Learning-Based Modeling Approach of High-Temperature Polymer Electrolyte Membrane Electrochemical Systems. *Ind. Eng. Chem. Res.* **2022**, *61* (9), 3350−3357.
(170) Gong, Z.; Wang, B.; Benbouzid, M.; Li, B.; Xu, Y.; Yang, K.; Bao, Z.; Amirat, Y.; Gao, F.; Jiao, K. Cross-Domain Diagnosis for Polymer Electrolyte Membrane Fuel Cell Based on Digital Twins and Transfer Learning Network. *Energy AI* **2024**, *17*, 100412.
(171) Kebede, G. A.; Lo, S.-C.; Wang, F.-K.; Chou, J.-H. Transfer Learning-Based Deep Learning Models for Proton Exchange Membrane Fuel Remaining Useful Life Prediction. *Fuel* **2024**, *367*, 131461.
(172) Zhai, C.; Li, T.; Shi, H.; Yeo, J. Discovery and Design of Soft Polymeric Bio-Inspired Materials with Multiscale Simulations and Artificial Intelligence. *J. Mater. Chem. B* **2020**, *8* (31), 6562−6587.
(173) Wei, T.; Zhang, L.; Zhao, H.; Ma, H.; Sajib, M. S. J.; Jiang, H.; Murad, S. Aromatic Polyamide Reverse-Osmosis Membrane: An Atomistic Molecular Dynamics Simulation. *J. Phys. Chem. B* **2016**, *120* (39), 10311−10318.
(174) Tong, X.; Liu, S.; Zhao, Y.; Huang, L.; Crittenden, J.; Chen, Y. MXene Composite Membranes with Enhanced Ion Transport and


https://doi.org/10.1021/acs.est.4c08298
*Environ. Sci. Technol.* 2025, 59, 993−1012

Regulated Ion Selectivity. *Environ. Sci. Technol.* **2022**, *56* (12), 8964−8974.
(175) Noorani, N.; Mehrdad, A. Adsorption, Permeation, and DFT Studies of PVC/PVIm Blends for Separation of CO2/CH4. *J. Mol. Liq.* **2019**, *292*, 111410.
(176) Wilbraham, L.; Berardo, E.; Turcani, L.; Jelfs, K. E.; Zwijnenburg, M. A. High-Throughput Screening Approach for the Optoelectronic Properties of Conjugated Polymers. *J. Chem. Inf. Model.* **2018**, *58* (12), 2450−2459.
(177) Xu, D.; Zhang, Q.; Huo, X.; Wang, Y.; Yang, M. Advances in Data-Assisted High-Throughput Computations for Material Design. *Mater. Genome Eng. Adv.* **2023**, *1* (1), No. e11.
(178) Lei, Q.; Li, L.; Chen, H.; Wang, X. Emerging Directions for Carbon Capture Technologies: A Synergy of High-Throughput Theoretical Calculations and Machine Learning. *Environ. Sci. Technol.* **2023**, *57* (45), 17189−17200.
(179) Xin, B.; Feng, M.; Cheng, M.; Dai, Z.; Ye, S.; Zhou, L.; Dai, Y.; Ji, X. Combining Interpretable Machine Learning and Molecular Simulation to Advance the Discovery of COF-Based Membranes for Acid Gas Separation. *Ind. Eng. Chem. Res.* **2024**, *63* (18), 8369−8382.
(180) Cheng, X.; Liao, Y.; Lei, Z.; Li, J.; Fan, X.; Xiao, X. Multi-Scale Design of MOF-Based Membrane Separation for CO2/CH4Mixture via Integration of Molecular Simulation, Machine Learning and Process Modeling and Simulation. *J. Membr. Sci.* **2023**, *672*, 121430.
(181) Meng, K.; Niu, Y.; Zhao, X.; Zhang, Y.; Zhao, Y.; Yu, X.; Rong, J.; Hou, H. Data-Driven Design of High-Performance Graphene-Based Seawater Desalination Membranes. *ACS Appl. Nano Mater.* **2023**, *6* (7), 5889−5900.
(182) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Crystal Structure Prediction via Particle-Swarm Optimization. *Phys. Rev. B* **2010**, *82* (9), 094116.
(183) Zhang, Z.; Luo, Y.; Peng, H.; Chen, Y.; Liao, R.-Z.; Zhao, Q. Deep Spatial Representation Learning of Polyamide Nanofiltration Membranes. *J. Membr. Sci.* **2021**, *620*, 118910.
(184) Zhou, X.; Wang, Z.; Epsztein, R.; Zhan, C.; Li, W.; Fortner, J. D.; Pham, T. A.; Kim, J.-H.; Elimelech, M. Intrapore Energy Barriers Govern Ion Transport and Selectivity of Desalination Membranes. *Sci. Adv.* **2020**, *6* (48), No. eabd9045.
(185) Yue, S.; Nandy, A.; Kulik, H. J. Discovering Molecular Coordination Environment Trends for Selective Ion Binding to Molecular Complexes Using Machine Learning. *J. Phys. Chem. B* **2023**, *127* (49), 10592−10600.
(186) Lee, J. H.; Lee, M.; Min, K. Natural Language Processing Techniques for Advancing Materials Discovery: A Short Review. *Int. J. Precis. Eng. Manuf.-Green Technol.* **2023**, *10* (5), 1337−1349.
(187) Olivetti, E. A.; Cole, J. M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.-J.; Hiszpanski, A. M. Data-Driven Materials Research Enabled by Natural Language Processing and Information Extraction. *Appl. Phys. Rev.* **2020**, *7* (4), 041317.
(188) Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural Language Processing: State of the Art, Current Trends and Challenges. *Multimed. Tools Appl.* **2023**, *82* (3), 3713−3744.
(189) Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and Challenges of Text Mining in Materials Research. *iScience* **2021**, *24* (3), 102155.
(190) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894−1904.
(191) Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A. Structured Information Extraction from Scientific Text with Large Language Models. *Nat. Commun.* **2024**, *15* (1), 1418.
(192) Huang, S.; Cole, J. M. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. *J. Chem. Inf. Model.* **2022**, *62* (24), 6365−6377.
(193) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145* (32), 18048−18062.
(194) Gupta, T.; Zaki, M.; Krishnan, N. M. A.; Mausam. MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction. *Npj Comput. Mater.* **2022**, *8* (1), 1−11.
(195) Choudhary, K.; Kelley, M. L. ChemNLP: A Natural Language-Processing-Based Library for Materials Chemistry Text Data. *J. Phys. Chem. C* **2023**, *127* (35), 17545−17555.
(196) Shetty, P.; Ramprasad, R. Automated Knowledge Extraction from Polymer Literature Using Natural Language Processing. *iScience* **2021**, *24* (1), 101922.
(197) Martin, T. B.; Audus, D. J. Emerging Trends in Machine Learning: A Polymer Perspective. *ACS Polym. Au* **2023**, *3* (3), 239−258.
(198) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A Polymer Dataset for Accelerated Property Prediction and Design. *Sci. Data* **2016**, *3* (1), 160012.
(199) Ritt, C. L.; Stassin, T.; Davenport, D. M.; DuChanois, R. M.; Nulens, I.; Yang, Z.; Ben-Zvi, A.; Segev-Mark, N.; Elimelech, M.; Tang, C. Y.; Ramon, G. Z.; Vankelecom, I. F. J.; Verbeke, R. The Open Membrane Database: Synthesis-Structure-Performance Relationships of Reverse Osmosis Membranes. *J. Membr. Sci.* **2022**, *641*, 119927.
(200) OSN Database. https://osndb.kaust.edu.sa/en-US/#/datasets (accessed 2024-11-20).
(201) Zhao, Y.; Tong, X.; Chen, Y. Fit-for-Purpose Design of Nanofiltration Membranes for Simultaneous Nutrient Recovery and Micropollutant Removal. *Environ. Sci. Technol.* **2021**, *55* (5), 3352−3361.
(202) Hu, A.; Liu, Y.; Wang, X.; Xia, S.; Van der Bruggen, B. A Machine Learning Based Framework to Tailor Properties of Nanofiltration and Reverse Osmosis Membranes for Targeted Removal of Organic Micropollutants. *Water Res.* **2025**, *268*, 122677.
(203) Cairone, S.; Hasan, S. W.; Choo, K.-H.; Li, C.-W.; Zarra, T.; Belgiorno, V.; Naddeo, V. Integrating Artificial Intelligence Modeling and Membrane Technologies for Advanced Wastewater Treatment: Research Progress and Future Perspectives. *Sci. Total Environ.* **2024**, *944*, 173999.
(204) Reid, E.; Igou, T.; Zhao, Y.; Crittenden, J.; Huang, C.-H.; Westerhoff, P.; Rittmann, B.; Drewes, J. E.; Chen, Y. The Minus Approach Can Redefine the Standard of Practice of Drinking Water Treatment. *Environ. Sci. Technol.* **2023**, *57* (18), 7150−7161.
(205) Osman, A. I.; Chen, Z.; Elgarahy, A. M.; Farghali, M.; Mohamed, I. M. A.; Priya, A. k.; Hawash, H. B.; Yap, P.-S. Membrane Technology for Energy Saving: Principles, Techniques, Applications, Challenges, and Prospects. *Adv. Energy Sustain. Res.* **2024**, *5* (5), 2400011.
(206) Tran, H.; Gurnani, R.; Kim, C.; Pilania, G.; Kwon, H.-K.; Lively, R. P.; Ramprasad, R. Design of Functional and Sustainable Polymers Assisted by Artificial Intelligence. *Nat. Rev. Mater.* **2024**, *9*, 866−886.
(207) Kern, J.; Su, Y.; Gutekunst, W.; Ramprasad, R. An Informatics Framework for the Design of Sustainable, Chemically Recyclable, Synthetically-Accessible and Durable Polymers. *arXiv*, September 13, 2024, 2409.15354, ver. *1*. DOI: 10.48550/arXiv.2409.15354.
(208) He, Y.; Liu, G.; Li, C.; Yan, X. Reaching the Full Potential of Machine Learning in Mitigating Environmental Impacts of Functional Materials. *Rev. Environ. Contam. Toxicol.* **2022**, *260* (1), 21.
(209) Al-Sakkari, E. G.; Ragab, A.; Dagdougui, H.; Boffito, D. C.; Amazouz, M. Carbon Capture, Utilization and Sequestration Systems Design and Operation Optimization: Assessment and Perspectives of Artificial Intelligence Opportunities. *Sci. Total Environ.* **2024**, *917*, 170085.