5-11-2023

Membrane Science Meets Machine Learning: Future and Potential Use in Assisting Membrane Material Design and Fabrication

Musabbir J. Talukder

Ali Alshami
University of North Dakota, ali.alshami@und.edu

Arash Tayyebi

Nadhem Ismail

Xue Yu

Membrane Science Meets Machine Learning: Future and Potential Use in Assisting Membrane Material Design and Fabrication

Musabbir J. Talukder[a], Ali S. Alshami[a1], Arash Tayyebi[a], Nadhem Ismail[a], Xue Yu[b]

[a]University of North Dakota, Chemical Engineering, Grand Forks, ND, 58201, USA
[b]Energy & Environmental Research Center, University of North Dakota

Abstract

The evolving membrane technology integrated with machine learning (ML) algorithm can significantly advance the novel membrane material design and fabrication. Although several studies have reported ML-based assistance in membrane development, none of them have offered a complete analysis of existing ML-assisted membrane fabrication methods from material design perspective. A two-way information gateway is therefore necessary to achieve the desired objective, whereby experienced researchers and data scientists from both sides need to provide valuable insights into novel membrane development process. In this work, we offer a midway platform by providing an overall view and scopes of ML use in membrane science. This is accomplished by analyzing reported ML-assisted membrane fabrications via lensing through the overall ML development. This work culminates in identifying four crucial factors affecting ML-assisted membrane development: data mining, material functional description, selection of ML models, and model interpretation. A future direction is proposed by making specific ML models and descriptors suggestions, in addition to molecular similarity analysis technique and ML based Image processing. We believe the proposed approaches and analysis through our identified lens will prove crucial for the future of ML-assisted membrane material design and development.

[1]Corresponding author: ali.alshami@und.edu, Tel: +1 701 777 6838

1. Introduction

We cannot think of a sustainable world with our purification, treatment, and separation needs are maintained by environmentally harmful and energy-intensive processes. Large distillation columns, scrubbers, and chemical treatments still dictate the separation process industries. Membrane-based technologies offer a simple and elegant solution; however, a relatively slow transition is observed due primarily to a lack of broader applicability. This prospective technology has significantly impacted separation industries, especially Reverse Osmosis (RO) membranes, which are a mature technology and very attractive for water desalination and purification industries (Figure 1) [1]–[3]. Additionally, commercial gas separating membranes were developed over four decades ago.

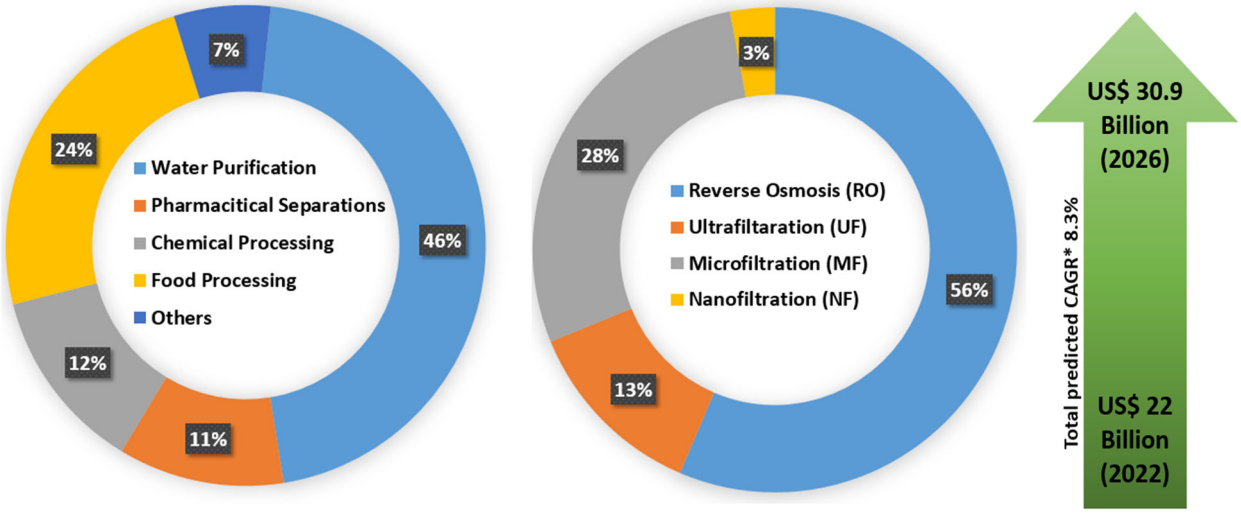

Figure 1: Global membrane separation market: by application (left); by technology (right) with projected compound annual growth rate or CAGR [1][2][3]

However, a more comprehensive and reliable application of this sustainable technology has not been achieved yet. It has been attributed mostly to a lack of significant advancements in membrane material development [4]–[6]. Most membrane materials fail to adapt and perform economically in diverse process conditions. Common failures occur due to long-term exposure to high pressure and temperature. As a result, the demand is being met by using harmful unsustainable technologies. For instance, there is an increased interest in post-combustion CO_2

capture, which uses harmful amine scrubbing. Membrane technology can offer a better, sustainable solution, mitigating the adverse effects. However, implementation of membrane technology in such fields requires discovery of novel materials capable of withstanding process conditions while exhibiting the desired transport properties. An extensive study in materials behavior and experimentation is necessary to meet these material innovation demands. It is often argued that unprecedented approaches need to be taken in developing high-performance membrane materials if we want to meet the increasing demands of separation industries[7]–[9][10].

Membrane material development has largely been an “Edisonian” process, sometimes taking decades to create a new efficient membrane material. Polymers are considered sustainable form ofmembrane materials. Therefore, membranes have attracted increased attention since inception. Polymeric membranes have been used in microfiltration (MF), ultrafiltration (UF), and nanofiltration (NF) processes. Polymer membranes are easy to develop, fabricate, and control, providing a sustainable and scalable platform for chemical separations. However, it remains subject to permeability-selectivity trade-offs. In addition, most used polymer membranes seldom show resistance to high temperature and pressure. Nonetheless, the polymer fabrication process has various optimizable fabrication parameters which control the flow transport [11][12]. The fabrication process is time-consuming with exhaustive trial and error-based progression. The advent of mixed-matrix membranes also introduces crucial decision points on filler and matrix material choice [13]. These parameters add variability to the membrane material selection and fabrication process.

The existing polymer database constitutes an ample sample space varying in mechanical and chemical resistance. The published literature on various membrane materials has also added transport properties to the database. However, this inherent high dimensionality of the polymer space makes discovery and fabrication a highly arduous task. The trial and error approach of novel membrane fabrication has historically been guided by subjective experience and intuition; henceforth, membrane technology for versatile separation operations has evolved comparatively sluggishly [14]. These traditional techniques have become ineffective for rapid high-performance membrane design, and there is an increasing drive for screening novel high-performance

materials as well as discovering new ones. Current efforts can often be described as tuning responsible chemical groups for efficiently transporting the desired chemical species instead of incorporating broader perspectives when designing novel membrane materials.

The continual shift toward data-driven science has made the advent of Artificial Intelligence (AI) or smart algorithms to convey better, faster and cheaper designs. The most basic concept includes categorization, learning and then prediction from a dataset. Theses predicting algorithms or Machine Learning (ML) models depends on the pattern formation in large data which requires a crucial step of training the model on given set of instructions. Going beyond this, in Deep learning (DL), data scientists are looking into analysis of unstructured data and automated identification of features instead of training on given set of conditions [15][16].

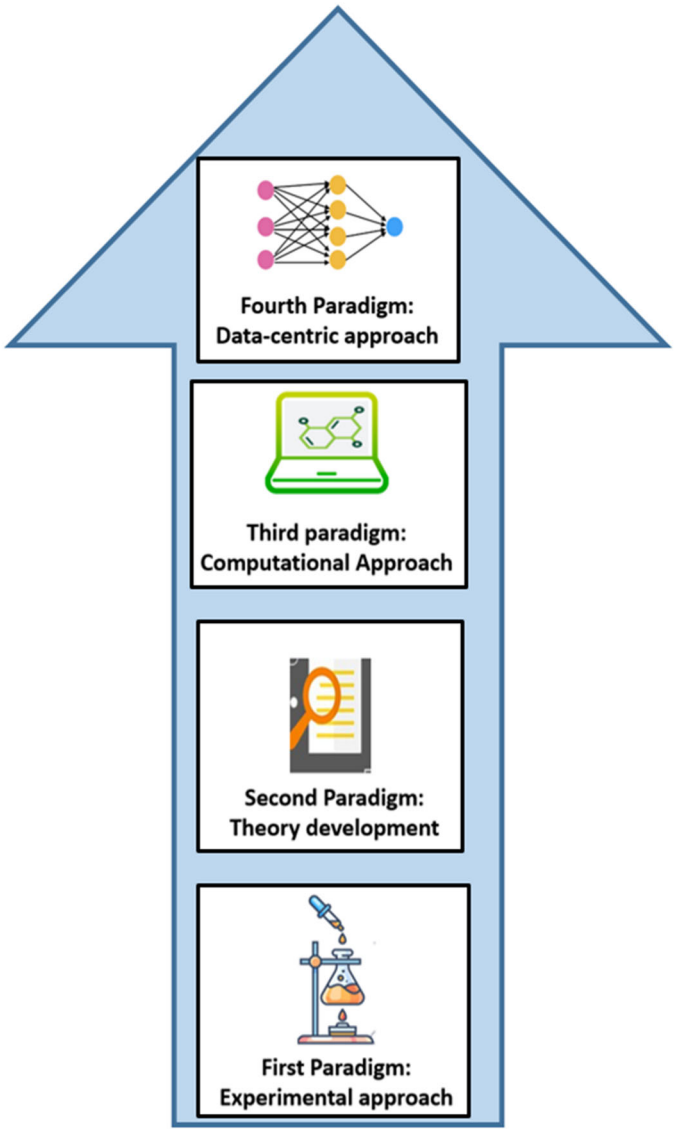

Figure 2: Advances of statistical techniques has led to the introduction of new paradigm for novel material development.

The material synthesis process and modern data collection techniques generate a large amount of data, making data-centric approaches a new paradigm for the discovery, rational design, and synthesis guide for novel high-performance materials (Figure 2) [17]. ML and DL models have been the flag-bearing methods to assist in the synthesis of stable inorganic materials based on a theoretically generated database [18]. ML

models have also successfully predicted polymer properties, such as glass transition temperature and dielectric constants [19]–[21]. Metal Organic Frameworks (MOF) with their vast array of potential combinations between metal nodes and organic ligands presents a significant challenge for predicting the structural properties and synthesizing novel MOFs. However, ML has demonstrated its utility in identifying the relationship between material properties and their structures. This has been exemplified in the study of metal-organic frameworks (MOFs) with electrically conductive structures[23],[24]. By leveraging ML, large datasets can be analyzed more efficiently, allowing researchers to uncover important patterns and features that may not be immediately apparent using traditional approaches. Recently, ML based predictions are aided by theoretical quantum chemical properties from Quantum MOF (QMOF) Database and resulted in a more accurate structural property predictions[24], [25]. Deep learning (DL), a recently popularized umbrella of algorithms is a subset of machine learning that uses artificial neural networks with multiple layers to learn and make predictions from complex data. Combining DL based Recurrent neural network (RNN) with Monte Carlo tree search, Zhang et al. introduced a modified approach to design novel MOFs with high density of adsorption sites for methane-storage and carbon-capture applications[26].

Dependence of large dataset for higher accuracy is often considered as a major limiting factor in ML implementation. However, synthetic data generation or simulated dataset could be a viable option for potential screening and property predictions as it is demonstrated by many researchers in MOF design[27]. Li et al. investigated the combine approach with large-scale computational screening and ML to accelerate MOF material screening[28]. Their study revealed the potential for ML study in performance metrics and their characteristic descriptors. Such explorations into the virtual MOF space ushers great potential in membrane applications such as gas storage, separation, catalysis, and sensing. Furthermore, with porous material descriptor i.e. a more detailed material description has led to the development of supervised ML algorithm. Which has led to the identification of 481 never investigated porous MOF structures [29]. However, screening large databases of MOFs to find well-performing materials is very time-consuming; hence ML techniques were used in recent years to predict novel MOFs design through a process call inverse design[30][31].

ML models for membrane science have been historically regarded as scientifically futile since they did not reveal much about the underlying physical phenomena [32]. A mathematical model's primary purpose in membrane science is to offer a description to the separation process using well-known analytical equations that represent and explain the physical process. In contrast, ML-based models are not identical to statistical models, where the underlying algorithms and governing equations do not directly reveal the physiochemical insights. However, ML models can exploit inter-variable relationships and their weights for predictions, which can be interpreted as a reliable prediction for material properties. Explored polymers as membrane materials create significant transport data to train and validate ML algorithms for membrane research. The global materials informatics initiative, including the “Materials Genome Initiative (USA)” [33], the “Materials Research by Information Integration” (Japan) [34], and the “NOMAD Laboratory” (EU)[35], have been assisting this data-centric research by providing a stable material database. The new insights made possible by the reliable database and advancement of ML models that can create the capability to achieve desired modeling goals. These tools are revolutionary in the field of novel material design. Several ML models have successfully initiated the design assistance of novel water and gas permeating membranes. In the breakthrough work, Barnett et al. fabricated two gas separation membranes based on the results they obtained from the ML algorithm [36]. For water purification, Gao et al. produced two polyimide nanofiltration membranes based on their ML-assisted predictions [37]. In a theoretical study, Yang et al. predicted more than 100 novel polymer constructs with superior performance for gas separations[38]. These notable works along with others paved the way and warrant a formal review of ML-assisted membrane design and fabrication methods. To the best of our knowledge, no such effort has been made on this research topic.

Furthermore, integrated with ML, the evolving membrane field can significantly advance the novel membrane material design process. ML models can be a crucial guiding tool for the next generation of experimentalists. However, ML method development for membrane fabrication is still currently in its infancy phase. A general guide on ML methods and development needs to be introduced from membrane research’s perspective to accelerate this process. Moreover, a two-way information gateway is necessary to achieve the desired objective. Experienced researchers

and data scientists from both sides need to provide valuable insights into novel membrane development process, and for that a midway platform is necessary. Henceforth, this work aims at creating a platform by providing a general picture and scopes of this field. As such, this paper explores recent ML-assisted membrane fabrication activities by lensing through the overall ML development process. In doing so, we also reviewed relevant works on ML assisted material design and identified four crucial factors affecting ML-assisted membrane design and fabrication. Additionally, we proposed a future research direction by making specific ML models and descriptors suggestions. To further enrich the membrane database for ML exploration, molecular similarity analysis technique and ML based Image processing is also offered. We believe the proposed approaches along with the analysis on the recent progress of ML assisted membrane design through our identified lens, could be crucial for designing and guiding the future of this field.

2. Current progress in Machine Learning (ML)-assisted membrane fabrication

ML-assisted membrane fabrication is a successor to the advances in polymer computational chemistry and predictive ML algorithms. ML methods have been successfully used from predicted glass transition temperatures [39][40][41] to photovoltaic properties [42][43] in polymer materials. In membrane processing, ML based approaches have been successfully applied in optimizing process parameters and performance evaluation [44][45][46][47][48][49] [50][51]. ML methods have been extensively utilized to decipher water quality[52], [53] and such attempts could greatly benefit the water purification by membranes processes. Based on the water quality data, Younes et al. employed two unsupervised ML model (Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA)) to optimize filter material selection[54]. With their small dataset approach, their HCA analysis was limited to classification without revealing the factors influencing this classification. However, their PCA provided a general view of correlation and dissimilarity among the investigated seven membrane types and the six factors.

In ML-assisted membrane design and fabrication, screening, and inverse designing are the two primary explored techniques. Several attempts have been made to screen potential polymer materials from a polymer space to identify and design new membrane materials. Proper screening can accelerate the complex material discovery process by creating a faster feedback process that allows for rapid material discovery cycle [55]. Conversely, ML-based novel material prediction has also been accelerated significantly by inverse modeling techniques. Inverse prediction allows the design of the desired material by identifying targeted functionalities [56], prompting researchers to shift from screening to novel membrane material design (Figure 3). Inversely membrane material design, which is free of the current trial-and-error approach, has been initiated by employing various statistical optimization. This process of optimization for a target performance can accelerate material design by reducing the number of experiments as possible[47][48][59]. Robeson et al. were first to implement inverse design techniques for gas membranes to establish a relationship between polymer chain units, subunits to gas permeation [60]. However, their work could not take advantage of currently available statistical tools nor ML algorithms and was limited to only identified relations. Decades later, Barnett et al. built on the "group contribution approach" to devise the first ML algorithm for identifying conducive patterns in gas separation membranes[36]. The authors screened and synthesized two novel polyamide membranes that exceeded the so-called "upper bound" for CO_2/CH_4 separation performance. In later works, Gao et al. utilized Bayesian Optimization(BO) driven inverse design frameworks [61] to fabricate polyamide NF membranes for water filtration. Their tree based ML models deduce

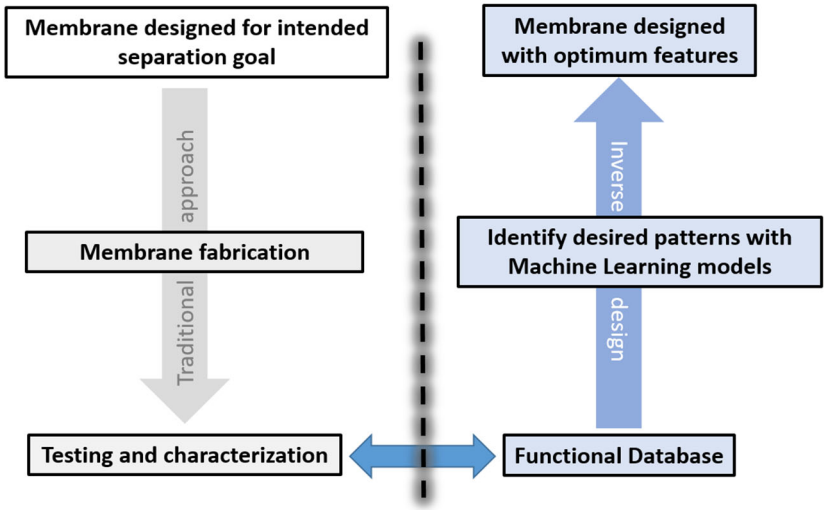

Figure 3: Inverse Design can reduce novel membrane design time with ML enabled approach.

relations with blocks of polymer structure (translated by functional description) known as features, and membrane performance. To train their supervised ML models, they constructed two datasets of total 567 polyamide based NF membranes from 218 published literatures. The interpreted relations then used to construct a virtual reference which is then used to screen potential monomers. They utilized Bayesian optimization (BO) for optimizing crucial fabrication parameters, such as casting temperature, thickness, and density to optimize fabrication conditions for the screened monomers as well as to produce membranes exceeding the upper bounds. The optimization process of BO enables the efficient design and optimization of membranes by exploring the parameter space to identify the optimal set of parameters that maximize or minimize the desired properties. This approach can ultimately result in the development of more efficient and cost-effective membranes suitable for diverse applications through inverse designing approach. Their work resulted in eight polyamide membranes exceeding the current performance limits in terms of higher water permeability and salt rejection [37]. Their Optimization allowed them to surpass conventional screening and build novel designs by optimizing precursor fabrication.

Mixed Matrix Membrane (MMM) development has yet to utilize the full ML capabilities, despite its scopes and potential. Fabrication hurdles with dispersion as well as agglomeration tendencies makes MMM development primarily an optimization process [62]. It can greatly benefit from the existing optimization process algorithms from ML. In literature, Fetanat et al. used Artificial Neural Networks (ANN) to predict and obtain insights into polymeric nanocomposite ultrafiltration membrane performance [63]. The authors' works included several fabrication conditions as independent variables [64]. Yeo et al. employed gradient boosting tree (GBT) model to gain insight of RO-MMM fabrication process[65]. Their result evaluated nanoparticles performance in terms of loading rate, pore size and hydrophilicity. However, their work was limited to performance insights rather than actual membrane design. In more recent work, Yang et al. focused on gas membrane design, using two major ML algorithms (random forest (RF) regression and deep neural networks (DNN)) and studying the two prominent descriptors of choice: Chemical Descriptor and Morgan Fingerprint with Frequency (MFF)[38]. Gao et al. and

Yang et al.'s work used SHapley Additive exPlanations, or SHAP analysis, to integrate their ML predictions into membrane design [38].

In the following sections, we delineate into four identified crucial factors (Figure 4) in ML assisted novel membrane material design and fabrication. These factors are identified in coherence with ML assisted general material design process as reported in published literature. Under these factors, recent progress of membrane material design is also captured. Moreover, these factors outline the general development process of ML-based models.

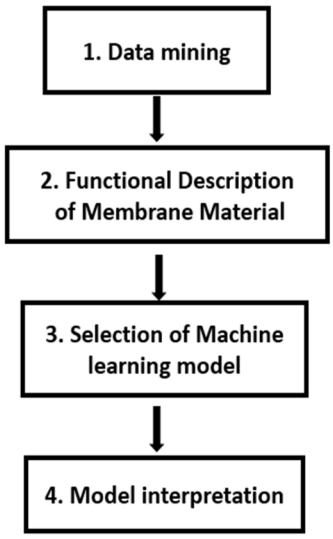

Figure 4: Four crucial factors identified in this work for ML-based novel material development.

2.1 Data mining

ML is a data-centric approach. Extensive data relating existing material properties to a particular objective are analyzed to obtain a numerical prediction on novel designs. Therefore, data selection in relation to the objective definition is crucial. The general argument is that quantity will precede quality in all successful predictions. However, an effective model is generally built using data representing the overall modeling goal with the best strategy for stratifying information related to species transport. Moreover, a homogenous and consistent dataset with no abrupt missing data is a major priority when designing novel high-performance membrane materials. The literature has already identified that mining a consistent database often creates the greatest hurdle, not a lack of ML algorithms [66]–[68].

Data variables for membrane material creation can take many forms. Therefore, mining a robust and diverse data set is a significant challenge. In addition, a large data set must be homogeneously distributed amongst the transport-defining parameters. Specific membrane

properties mined from the literature have primarily focused on parameters such as permeability, selectivity (gas separation), or salt rejection (Water purification)[36][37][38][69]. Adding to this, Table 1 lists descriptions of the most commonly explored membrane fabrication databases.

Table 1: Major databases related to material properties and crucial species transport

Name	Description	Ref.
Membrane Society of Australasia	Reported data on explored membrane materials	[70]
PolyInfo	Database for polymer material design	[71]
Materials Project	Theoretically generated polymer properties	[72]
NIST Interatomic Potential Repository	Database for interatomic force fields or potentials	[73], [74]
NIST Material Data Repository	Data reported on experimental materials	[75]
NIST Standard Reference Data	General material property data	[76]
PubChem	Chemical entity database	[77]
MatWeb	Database for material properties	[78]
NIMS Materials Database	Database dedicated to the development of new materials and the selection of materials	[79]
Nanomaterial Registry	Nanomaterial database	[80]
Nanoporous Material Explorer	A database of porous material properties	[81]
CoRE MOF	A database of metal-organic frameworks	[82]

The first constructed membrane database included gas permeability for 149 polymer membranes. Hasnaoui et al. created a database with N_2, CO_2, CH_4, and O_2 permeability to predict permeability in gas membranes [69]. However, this small amount of data made it unreliable for reverse predictions such as obtaining novel polymer backbone designs with superior gas transport properties. Most work on ML-based membrane performance predictions uses synthetic data since these data sets are large and easy to acquire. In their pioneering work, Barnett et al. departed from the trend of focusing only on the theoretically generated data by including experimental data sets [36]. The authors used a novel approach to mine data from the literature. They mined the permeation data for 500 to 1,000 polymers from literature and categorized them as CH_4, N_2, He, H_2, CO_2, and O_2 instead of progressing polymer by polymer. The

availability of data in the literature created a wide range of reported data points. Therefore, missing values significantly limit the power of their data set. To their credit, the authors chose not to fill their dataset with synthetically generated values despite apparent limitations. Many researchers have reported the addition of user bias with data curation. However, their data sets have failed to include Polymers with Intrinsic Microporosity (PIMs), an essential branch of polymers already established in the literature for efficient gas transport. Experimental studies on PIM structures have also agreed with these results [83][84]. The data mined from these studies are limited to existing reported polymers, which restricts their use for novel membrane designs. Therefore, Yang et al. used a more expanded approach to include both theoretical and experimental data from PolyInfo [85] and the Membrane Society of Australasia (MSA) [86]. The authors' dataset had potential ladder type and polyimide-based PIMs conducive to gas transport [38].

Fetanat et al. deviated from the trend of designing pristine polymeric membranes by exploring the domain of nanocomposite membranes, commonly known as Mixed Matrix Membranes (MMM)[63] . For that, the authors' strategy focused on mining the ultrafiltration polymeric membranes with nanomaterials. They surveyed the literature on ultrafiltration nanocomposite polymeric membranes and gathered 735 sample spaces with eight polymer supports. The authors then categorized the nanomaterial features into support, nanoparticle type, size, distribution, and concentration (%wt). Polymeric phase concentration (%wt), solvent type and concentration, operation pressure, contact angle, thin layer thickness, post-treatment temperature, and duration were also included in the input variables. The ML models developed using these data were then used to evaluate insights into solute rejection, pure water flux, and flux recovery for ultrafiltration membrane selection and fabrication.

Other researcher also attempted to include membrane compositions with these fabrication conditions into a design database. For example, Gao et al. used fabrication conditions to gain insights into the polyamide membrane fabrication process. They included features such as initial monomer concentration, polymerization time, heat curing time, additive and solvent type, and a nanomaterial dispersion medium to reveal the uncertainties of the polymer membrane fabrication process. However, their data set included missing data points due to the literature's

lack of reported fabrication data [37]. The authors mined 218 published works and gathered water permeability and salt rejection data. A normalized approach was taken by defining salt ions into five fundamental properties: valence, ionic radius, Stokes radius, hydrated radius, and hydration free energy. It allowed them to collect and model a broader dataset. A similar approach in gas separation membrane is forthcoming.

2.2 Functional Description of Membrane Material

Each functional material is defined on a set of specific-factors related to the material's function. In membrane application, these specific-factors can specify the material transport properties. Finer details must be identified and included in the material's functional description to increase model prediction efficacy. Therefore, developing an efficient membrane material functional description is a crucial step in the ML-based material design processes. Zhou et al. identified three critical characteristics of a good descriptor: (1) the descriptor should offer a unique definition for each sample space, representing a unique material characterization, (2) the descriptor should be sensitive to the design objective by having a unique sensitivity in relation to parts of the materials instead of just tabulating material descriptors versus properties, and (3) the descriptors should be easy to translate and produce [17]. Furthermore, polymer repeat units should be represented by a dynamic descriptor that includes chemical connectivity. Molecular fingerprints are remarkably reliable as polymer material descriptors compared to Molecular Descriptor (MD), Molecular Image (MI), and Molecular Graph (MG) [87][88]–[91][89].

Barnett et al. and Gao et al. in their work chose the Morgan fingerprint as their functional descriptor [37][36]. This technique allows chemical groups to be represented with greater flexibility due to their atomic group size and length definition flexibility [88]. Barnett et al. decided to use a more dynamic fingerprint approach for this process instead of a static group contribution. These fingerprints can evolve on limited capacity as new materials are designed and predicted. There are some overlaps as chemical entities are converted into a binary vector due to the limitation of their overlapping bit position. For example, Gao et al.'s work represented an amine group and sulfonic group using the same binary vector which is reflected in their

experimentally viable material prediction accuracy [37]. Yang et al. ran a comparative study between chemical descriptors generated by RDKit [92] and Morgan fingerprint with frequency (MFF) [93]. The authors' systematic study complied with other reported results [88] concluded that MFF performed comparatively better when used to define polymeric gas membranes [38].

In an orthodox approach, Fetanat et al. represented 14 polymers and the solvent types used in fabrication with numeric values to predict nanocomposite membrane performance. Though, they did not include any material properties of the filler materials. The filler materials can be described by material descriptors, summarized in the works of Curtarolo et al. as dimensional parameters [94]. One-dimensional parameters help define molecular weight, volume, connectivity, number of electrons and polarities, and surface area; however, two-dimensional (2D) or even three-dimensional (3D) parameters are preferable to describe functional material for transport.

2.3 Machine Learning Models

ML models can significantly improve membrane fabrication process. Developing an accurate and reliable model ensures better predictions. Recognizing the patterns for higher dimensional variables is essential for novel membrane design. Carefully designed ML model can give valuable insights into designing membrane materials and optimum fabrication conditions. ML models are broadly classified as either supervised or unsupervised [95]. Supervised modeling tries to predict functional descriptor blocks or features conducive to the desired transport based on the mined data. Broadly, two powerful techniques: regression and classification algorithms, fall under supervised models. Regression analysis uses parameters such as glass transition temperature, density, and strain behavior, which each can take free values, to make continuous variable predictions. Predictions for discreet targets or searching for the desired predicted function are known as classification. Among them, linear classifiers, support vector machines, decision trees and Random forest are all common types of classification algorithms in material science[96].

Supervised modeling trains under a defined dataset and predicts based on recognized relationships with the trained models. On the other hand, unsupervised ML aims to identify functional relationships from input data in a process called “clustering” based on the modeling objective. Clustering algorithms can group together similar samples or data points based on their features, allowing researchers to identify groups or clusters of samples with similar properties or behaviors. This can be useful in identifying new membrane materials or optimizing membrane properties from material. It can be beneficial for categorizing hidden conducive patterns in material design from a large comparative dataset. Unsupervised ML models are useful in reducing dimensionality of a large dataset to reduces the number of data inputs to a manageable size while also preserving the data integrity. Although in membrane field, we seldom encounter with a large enough dataset, however, the material itself can be represented in high dimensional blocks or features. Unsupervised ML can be beneficial in such scenarios to visualize and extract the most relevant features that are responsible for the variations in the dataset. Furthermore, some unsupervised ML models like Principal component analysis (PCA) can be used to detect outliers in a dataset [97]. These can help us mine a better dataset with reduced homogeneity issues. Other unsupervised ML models that include dimensionality reduction techniques such as t-Distributed Stochastic Neighbor Embedding (t-SNE) can help to visualize high-dimensional data and identify patterns and relationships, as well as anomaly detection algorithms that can identify unusual or anomalous data points that may be indicative of faulty or poorly performing membranes[98]. Furthermore, t-SNE can be used to identify the optimal combination of membrane components that will yield the desired properties. By visualizing the relationships between different membrane components, researchers can design and optimize membranes that have improved performance and selectivity. Returning to the supervised modeling, the model needs to be “trained” and later “validated” against a defined dataset, creating the need to choose a ratio between training and validation data. Barnett et al. split their datasets into a 3:1 ratio for their Gaussian Process Regression (GPR) models. This nonparametric model computes the distribution of all possible functions over the feed data. Each data point is related based on the original non-linear observations translated into a higher-dimensional space known

as kernel functions [99]. The authors observed stable predictions after adding approximately 400 data points into the model when training it on variable-sized datasets[36].

Missing values in the dataset is a significant challenge when developing ML models. Some models, such as deep neural networks (DNN), cannot process datasets with missing values. Therefore, replacing missing values in the dataset using generated data, such as using mean or median values, estimated from statistical distribution or ML models, reduces the dataset's ability to provide valuable practical insights into the model's performance [64] [100]. Gao et al. developed an ensemble algorithm from multiple decision tree regression, Random Forest (RF) to mitigate this problem. The authors' XGBoost and CatBoost models managed missing values in the dataset. They trained models using 80% of the dataset, attaining an R^2 value of 0.78 for water permeability with 567 data points, and 0.84 for salt rejection with 1,524 data points for the validation data sets.

On another work, Yang et al. chose two supervised models, RF regression and DNN, to discover novel polymeric gas membranes. Previously, Tao et al. reported that these two ML models could be potentially used for material predictions [41]. However, Yang et al. reported an average 0.74 R^2 value for RF and 0.90 for DNN due to the RF model's inability to handle data with homogeneity issues. Furthermore, a dataset with abrupt high and low-performance features also creates fitting concerns, resulting in a low R^2 value. More importantly, the dataset's quality, quantity, and

modeling objectives dictate ML model choice for membrane material design.In order to make an assessment on trained models and experimental interpretation of model's efficiency, we remodeled CO_2 permeability with Multi-Layer Perceptron (MLP) model. To compare our model with the efficient RF, we used Morgan fingerprint on Yang et al.'s gas permeability database. These values are greater than the accuracy of the RF model at 0.99 and 0.85 for the training and validation sets, respectively (Figure 5). However, in terms of experimental prediction, the larger disparities of MLP performance between training and testing dataset means there is significant overfitting problems in MLP than RF [101]. Therefore, RF still dictates in experimental prediction conforming established literature's preference.

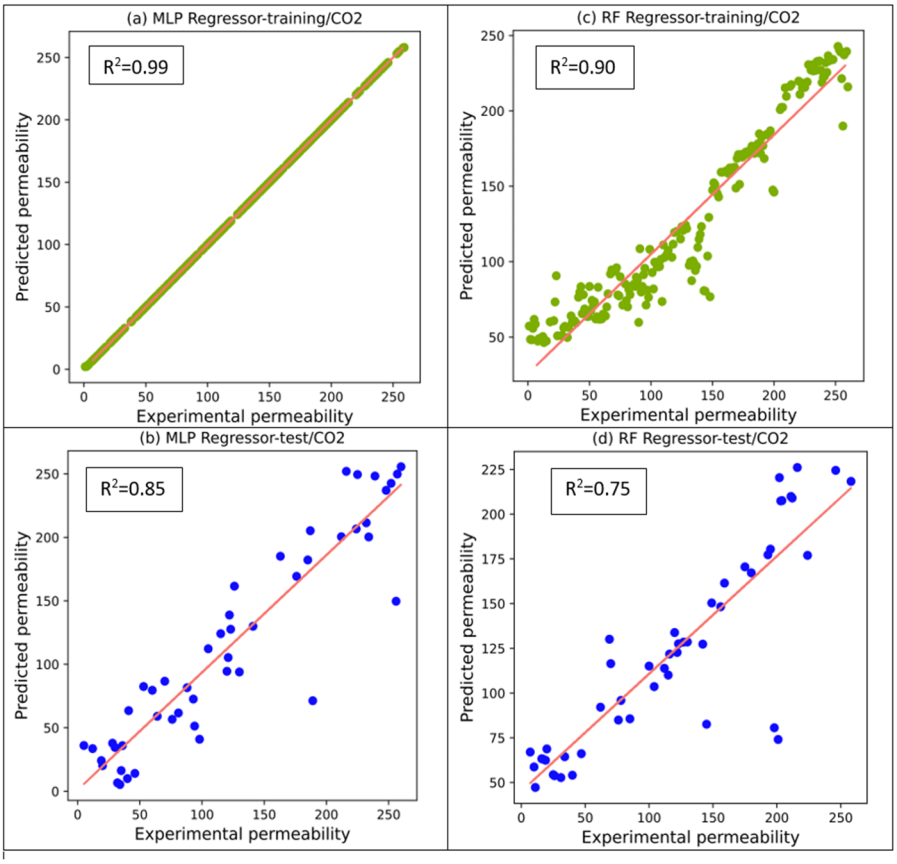

Figure 5: Linear regression and model assessment for CO_2 permeability (a) & (b) training and test with MLP regression; (c) &(d) with RF regression.

2.4 Model Interpretation:

The trained and validated ML algorithm finally steps into the interpretation part. From the group contribution approach to modern statistical tools, membrane scientists have invested considerable efforts into identifying novel polymers for membrane applications [60]. Screening over a potential space has been a primary means for interpreting material discovery predictions from ML models. This process reduces the material space to a manageable dataset. It is productive as Barnett et al. explored a similar screening technique by training the ML model with a gas permeation dataset of 500 to 1,000 polymers. The authors trained and validated model screened over 11,000 known polymers. Initial screening resulted in more than 100 novel polymers surpassing the current CO_2/CH_4 selectivity-permeability upper bound [102]. Later, the authors' experimental work successfully fabricated two novel polymeric membranes with higher performance.

For a better visualization, SHAP or SHapley Additive exPlanations method helps researchers to calculate the Shapley value for each transporting "feature," providing a comparative insight into the process [59]. This unique prediction approach uses differences between features, allowing researchers to interpret the data visually and assist new, novel membrane design. Also, this process gives the relationships in terms of positive or negative contributions to evaluate the influence of atomic groups on membrane transport (Figure 6).

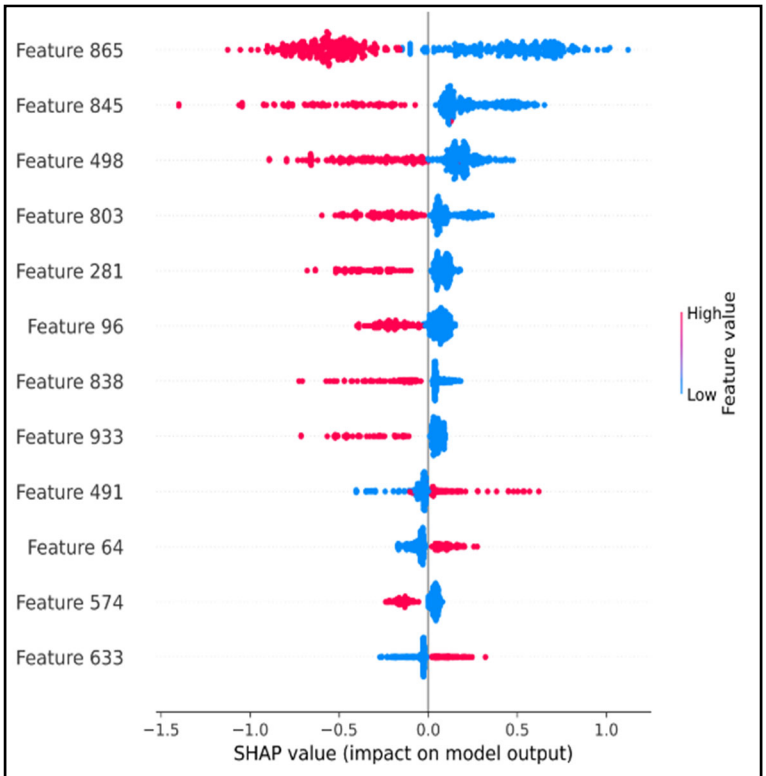

Figure 6: A visual representation of a SHAP analysis; positive and negative interactions are assigned to pre-defined features evaluated by ML model based on their respective contributions to the predicting outcome.

Gao et al. identified the positive Shapley values for water permeability in the presence of a specific amine group. Similar conclusions led the authors to build a Morgan fingerprint record of all atomic groups with positive Shapley values. These virtual references (two) were then compared to the 310 hand-picked amine monomers from the National Institute for Materials Science (NIMS) database [85]. The authors then used a sophisticated Bayesian optimization tool to construct novel polyamide membranes by training the model with fabrication conditions from literature. This work is the first to include such optimization approach novel for ML assisted membrane fabrication. They employed Bayesian Optimization or BO on their supervised models to identify conducive fabrication conditions. Bayesian Optimization is the best fit for these tasks since it uses Bayesian probability theory to balance exploitation and exploration [103]. Fabrication process of membranes can often be objectified as such optimization for high-performance materials. It has the potential to facilitate novel chemical and functional material design for drug discovery, molecular modeling, electrolyte design, and additive manufacturing [104]. BO can accelerate material design since the feedback loop provides a beneficial interplay between the inexpensive stochastic surrogate model and expensive computational acquisition functions to provide an optimal decision for future descriptors [105]. This statistical technique is also built on Gaussian process regression (GPR) capable of predicting the performance at novel conditions based on previously tested designs[57] therefore significantly reducing number experiments and resource cost. For identifying the optimal fabrication condition, they trained ML model with fabrication conditions from literature as input variables. Further optimization was done to a significantly smaller space to produce a membrane exceeding upper bound. This novel optimization approach can be utilized after creating a reliable virtual reference point through ML models, which is believed to be the key to next-generation membrane development. Such inverse design techniques, using Bayesian optimization with a well-developed supervised ML model could produce novel, unprecedented membrane materials that otherwise would not be possible with conventional screening techniques.

3. Future Direction for ML-Assisted Membrane Design and fabrication

ML methods enable efficient and reliable screening, novel designs, and optimization for potential membrane material fabrications. As a result, membrane material development cost and time can be reduced significantly by generating or screening from the material space. Moreover, potential predictions will result directly in experimental transport studies, saving time and opening new opportunities.

The general *MATLAB* platform has some ML tools that are used for some of these tasks. Nonetheless, many recently available open-source software packages, such as *TensorFlow* developed by Google's AI department, *Scikit-learn* in the Python package, and *Chainer*, have easy-to-use online guides, tutorials, and books non-specialists can utilize to implement ML models in their research.

The primary challenge in membrane science begins with creating a reliable and robust database. Membrane transport data are, in many cases, reported differently. Current approaches do not adequately translate species' transport properties against membrane materials. Gao et al. tried to incorporate species (salt) radius and free energy as a normalization approach, which resulted in a broader and better predicting ML model; however, a more encompassing normalization approach for membrane material transport would result in a more robust data set.

Standardizing data collection protocols and experimental procedures can help to reduce variability in the data and improve the homogeneity of the datasets. Which has already been initiated by the Open Membrane Database for RO and NF membranes[59]. They have a robust membrane database with a standardized data reporting protocol. Furthermore, creating such dynamic membrane data repository system will encourage self-reporting membrane data in a standardized form for other separation systems. Additional information or domain knowledge related to polymer chemistry, such as crosslinking, swelling, and branching, can also be a key to novel membrane design and aid the prepossessing of data in the future. By leveraging domain knowledge, researchers can make more informed decisions about data selection and preprocessing, and develop more accurate and interpretable ML models. Also, unsupervised ML

models can be used to detect outliers or unusual data points in a material dataset[97], [106]. By identifying the data points that lie outside the normal range, it is possible to identify potential errors or anomalies in the data, thereby creating homogenous dataset. Furthermore, data augmentation techniques such as data synthesis, data augmentation, and transfer learning can be used to increase the size and diversity of the available data. These techniques rely on generating new data by transforming existing data or by transferring knowledge from other datasets. For membrane research, the advances in other ML based material developments, like MOF database can be an excellent external source for this data augmentation.

To get a quantitative measure of the data set's variability, we performed a pairwise molecular similarity analysis on the gas permeation dataset Yang et al. [38] (Figure 7). Molecular similarity analysis allows comparing the similarity of any two molecules using their molecular fingerprints. It is represented by the "Tanimoto similarity score" given by:

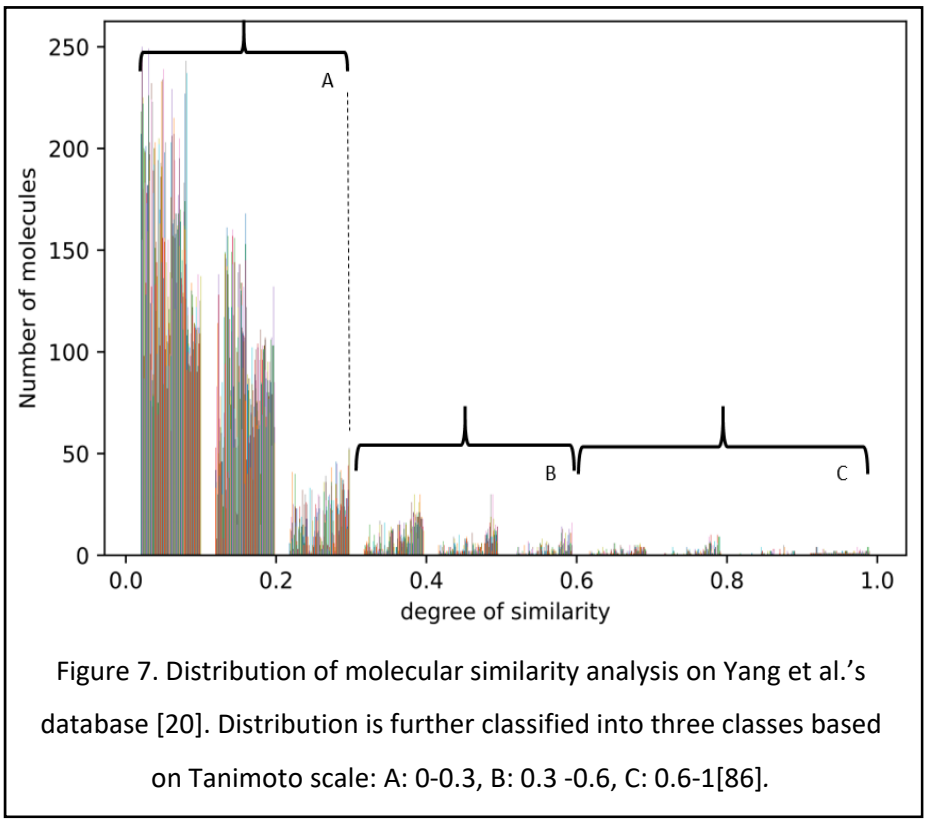

Figure 7. Distribution of molecular similarity analysis on Yang et al.'s database [20]. Distribution is further classified into three classes based on Tanimoto scale: A: 0-0.3, B: 0.3 -0.6, C: 0.6-1[86].

$$S(M,N)=(M\cap N)/(M\cup N)$$

Here $(M\cap N)$ represents the size of the intersection, i.e., the number of 1-bits common to M and N, and $(M\cup N)$ represents the size of the union, i.e., the number of 1-bits in A or B[107]. A less varied database results in limited and unreliable predictions. So in a Tanimoto scale, a more varied database will have most date within 0 to 0.3 degree. Which is desirable as a more varied

database ensure a reliable experimental prediction. Molecular similarity analysis will ensure easy evaluation of database on similarity issues.

Molecular interactions between the transporting species and membrane material can create a point to reduce dimensionality of material's functional descriptions. Polymer membranes present an inherent, significant challenge in functional description. Polymers are not a single entity; therefore, they cannot be described as a simple static material. They can be best described as a distribution. The distributions have a significant effect on the membrane properties. For example, Polyvinyl alcohol membrane-based separations depend largely on polymer weight distribution[108]. So, a detailed picture of the polymer distribution should be included in functional descriptions to unlock unknown factors in polymer membrane transport.

Complex relationships exist between a polymer matrix and fillers in a mixed matrix membrane (MMM). The highly selective filler phase is generally dispersed in a polymer matrix to overcome the trade-off between permeability and selectivity in a homogeneous polymeric membrane[109], [110]. However, this hybrid separation films exhibit non-ideal interactions between the matrix and filler phase. Moore et al. classified these non-ideal interactions into five categories, which later translated into the three most common cases: non-transporting rigidified polymer layer, limited caged transporting, and clogged transport (Figure 8) [111]. Identification is tedious and, in many cases, makes it difficult to evaluate these interactions based solely on conventional characterization techniques. ML based image processing can reveal these interactions and classify

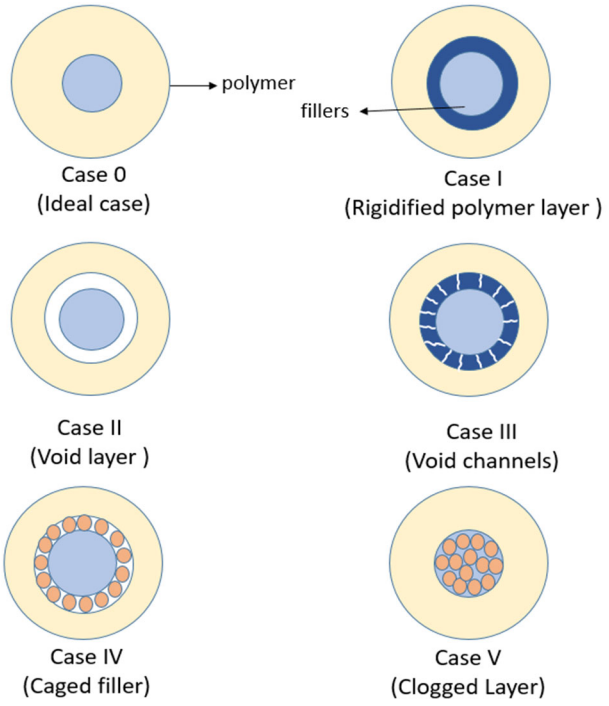

Figure 8: Summary of filler and matrix interactions as identified in Moore et al.'s work [90].

them as transport features in database by analyzing through the reported Scanning Electron Microscope (SEM) images from literature.

Furthermore, topological information often contains the means to define transport in membrane fillers. 2D nanomaterial structures, one of the most promising potential fillers in membrane research, topological descriptors have the potential to define transport properties. Information regarding symmetry, branching, and atom connectivity often reflects features that are hard to define and utilize with the usual processes [91] [92]. The major drawback of general material descriptors is that they do not contain stereochemistry information. This problem does not affect membrane material predictions since most general fillers used for membrane transport are not affected by the filler backbone's spatial distribution.

ML model development is a dynamic process. Therefore, the choice at each development point is crucial for subsequent steps. The appropriate choice for functional descriptors can help reveal new materials. For this, the selected descriptor should be predictive of the target properties and species transport. Accurate prediction based on properly selected descriptors overshadows the accuracy, regardless of the choice of ML models [114]. Unfortunately, there are no logical means by which current descriptors can be used for predicting novel arbitrary bonds using current “fingerprint” methods. Gao et al. used virtual referencing to create a fingerprint with optimal features to mitigate this issue; however, the feature flexibility depends on descriptor dynamics. So, to overcome this issue and describe transport properties related to chemical structure the simplified molecular-input line-entry system, or SMILES descriptors are introduced. Other novel material design initiatives also used a formal investigation with a Quantitative structure-property relationship (QSPR) descriptor in the membrane field [115].

The hidden pattern searching though unsupervised modeling has been proven beneficial into understanding PFAS C−F bond dissociation energies in relation to the functional chemical trends [15]. Unsupervised t-SNE can be utilized in similar ways to categorize conducive atomic patterns from a more robust high dimensional representation of membrane materials into clusters/families to understand which chemical functional groups are responsible for intended separation performance.Similar to membrane technology, research into novel Anti-Microbial Peptides (AMP) development is also an “Edisonian” process. The recent success of the Support

Vector Machine (SVM) model in designing novel AMP gained attention in other material designing fields [116]. The SVM algorithm projects the data points in a Euclidian space, where the position defines the data. It means curation is necessary to manage the missing points to implement this algorithm. This linear classification technique can be beneficial for novel membrane design since the data point transport features are classified with an intuitive understanding of the classification rationale and the relative importance and relationships [117] [116]. However, conventional ML models require a large data set to implement uncertainty models in data science. Building such a database is difficult for many scenarios; therefore, new developments in ML models should consider using small and non-homogenous data sets.

There have been significant advancements in ML algorithms. Nonetheless, in membrane science interpretation technique of these models is mainly limited to screening. Hasnaoui et al. collected a database of 149 polymers and modeled it using an Artificial Neural Network (ANN) algorithm [69]. The authors' work reported high R^2 values for the validated model. This work can highly benefit from the advent of SHAP analysis. In addition, several other potential interpretation techniques can bring fruitful outcomes in membrane material predictions. Lately, Local interpretable model-agnostic explanations (LIME)[118] and Yellowbrick visualization[119] are gaining attention among ML researchers.

Furthermore, ML-based inverse design models have recently utilized supervised modeling to design novel molecular structures. For example, Yao et al. used a Supramolecular Variational Autoencoder (SmVAE) Model to develop a novel Metal-Organic Framework (MOF) to increase CO_2/CH_4 and CO_2/N_2 separation [120]. Similar computationally efficient approaches have yielded never-seen-before chemical structure designs [121]. A future collaborative investigation from membrane experts and data scientists can benefit from such activities for novel membrane discovery.

4. Conclusion

Membrane technology is at a crucial development stage. Advanced data-centric techniques, such as Machine Learning (ML), are creating new opportunities to replace the age-old “Edisonian” approach. The concealed information in the large data has the potential to solve many pressing issues, including discovering high-performance membranes. Integrating ML methods to guide and enhance the experimentation process through material screening and optimization has already shown its promises. The development of chemometrics and computer science advances have also aided the process. We anticipate that the use of ML techniques in membrane research will continually increase due to the crucial demand of separation membranes along with developments in effective algorithms and computational tools. Creating a reliable database with suitable membrane material representations and implementing a reliable ML model with proper interpretation is vital in ML-assisted membrane fabrication. Balancing this dynamic approach shall result in significantly accelerating separation membranes material discovery cycle to guide membrane design while saving money and time.

Conflict of Interest

The authors declare that no conflict of interests exists in this review paper.

Acknowledgements

The presented work was supported in-part by the City of Grand Forks grant offered to Dr. Ali Alshami through Project number: UND0023658.

Authors Responsibilities

Musabbir Talukder: Writing - Original Draft, Writing - Review & Editing; **Ali Alshami**: Conceptualization Writing - Review & Editing, Supervision; **Arash Tayyebi**: Writing - Review & Editing, **Nadhem Ismail**: Writing - Review & Editing; **Xue Yu**: Writing - Review & Editing.

References:

[1] "Global Membrane Market 2022-2026 - Research and Markets." https://www.researchandmarkets.com/reports/5304710/global-membrane-market-2022-2026 .

[2] "Membrane Separation Technology Market Report, 2021-2028." https://www.grandviewresearch.com/industry-analysis/membrane-separation-technology-market.

[3] "Global Membrane Separation Technologies Market to Reach." https://www.globenewswire.com/en/news-release/2022/04/20/2425854/0/en/Global-Membrane-Separation-Technologies-Market-to-Reach-US-30-9-Billion-by-the-Year-2026.html.

[4] B. Assfour and S. Dawahra, "Separation of noble gases through nano porous material membranes," *Ann. Nucl. Energy*, 2020, doi: 10.1016/J.ANUCENE.2020.107730.

[5] J. Cai, H. Yin, and F. Guo, "Transport analysis of material gap membrane distillation desalination processes," *Desalination*, 2020, doi: 10.1016/J.DESAL.2020.114361.

[6] J. Hu, W. Chen, Y. Qu, and D. Yang, "Safety and serviceability of membrane buildings: A critical review on architectural, material and structural performance," *Eng. Struct.*, 2020, doi: 10.1016/J.ENGSTRUCT.2020.110292.

[7] Q. Yuan *et al.*, "Imputation of missing gas permeability data for polymer membranes using machine learning," *J. Memb. Sci.*, 2021, doi: 10.1016/J.MEMSCI.2021.119207.

[8] Z. Zhang *et al.*, "Machine learning aided high-throughput prediction of ionic liquid@MOF composites for membrane-based CO2 capture," *J. Memb. Sci.* 2022, doi: 10.1016/J.MEMSCI.2022.120399.

[9] M. Zhou, A. Vassallo, and J. Wu, "Toward the inverse design of MOF membranes for efficient D2/H2 separation by combination of physics-based and data-driven modeling," *J.*

Memb. Sci., 2020, doi: 10.1016/J.MEMSCI.2019.117675.

[10] A. Tayyebi, A. Alshami, Z. Rabiei, X. Yu, M. J. Talukder, and J. Power, "Prediction of Organic Compound Aqueous Solubility Using Interpretable Machine Learning- A Comparison Study of Descriptor-Based and Topological Models," 2022, doi: 10.21203/RS.3.RS-2155283/V1.

[11] B. S. Lalia, V. Kochkodan, R. Hashaikeh, and N. Hilal, "A review on membrane fabrication: Structure, properties and performance relationship," *Desalination*, 2013, doi: 10.1016/J.DESAL.2013.06.016.

[12] C. Algieri, S. Chakraborty, and U. Pal, "Efficacy of Phase Inversion Technique for Polymeric Membrane Fabrication," *J. Phase Chang. Mater.*, 2021, doi: 10.6084/JPCM.V1I1.10.

[13] J. Lewis, M. A. Q. Al-sayaghi, C. Buelke, and A. Alshami, "Activated carbon in mixed-matrix membranes", 2019, doi: 10.1080/15422119.2019.1609986.

[14] "A Research Agenda for Transforming Separation Science (Technical Report) | OSTI.GOV." https://www.osti.gov/biblio/1637445-research-agenda-transforming-separation-science.

[15] A. Raza *et al.*, "A Machine Learning Approach for Predicting Defluorination of Per- And Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal," *Environ. Sci. Technol.Lett.*,2019,doi:10.1021

[16] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[17] T. Zhou, Z. Song, and K. Sundmacher, "Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design," *Engineering*, 2019, doi: 10.1016/J.ENG.2019.02.011.

[18] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," *JOM*, 2013, doi: 10.1007/S11837-013-0755-4.

[19] B. G. Sumpter and D. W. Noid, "Neural networks and graph theory as computational tools for predicting polymer properties," *Macromol. Theory Simulations*, 1994, doi:

10.1002/MATS.1994.040030207.

[20] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, “Machine Learning Strategy for Accelerated Design of Polymer Dielectrics,” *Sci. Reports* 2016, doi: 10.1038/srep20952.

[21] F. Jabeen, M. Chen, B. Rasulev, M. Ossowski, and P. Boudjouk, “Refractive indices of diverse data set of polymers: A computational QSPR based study,” *Comput. Mater. Sci.*, 2017, doi: 10.1016/J.COMMATSCI.2017.05.022.

[22] Y. He, E. D. Cubuk, M. D. Allendorf, and E. J. Reed, “Metallic Metal-Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations,” *J. Phys. Chem. Lett.*, 2018, doi: 10.1021/ACS.JPCLETT.8B01707.

[23] A. Rosen *et al.*, “Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery with a New Electronic Structure Database,” 2020, doi: 10.26434/CHEMRXIV.13147616.V1.

[24] A. S. Rosen *et al.*, “High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration,” *npj Comput. Mater.* 2022, doi: 10.1038/s41524-022-00796-6.

[25] S. Tamakloe, “Machine learning improves metal–organic frameworks design and discovery,” *MRS Bull.*, 2022, doi: 10.1557/S43577-022-00427-X/METRICS.

[26] X. Zhang, K. Zhang, and Y. Lee, “Machine Learning Enabled Tailor-Made Design of Application-Specific Metal-Organic Frameworks,” *ACS Appl. Mater. Interfaces*, 2020, doi: 10.1021/ACSAMI.9B17867/SUPPL_FILE/AM9B17867_SI_002.ZIP.

[27] H. Daglar and S. Keskin, “Combining Machine Learning and Molecular Simulations to Unlock Gas Separation Potentials of MOF Membranes and MOF/Polymer MMMs,” *ACS Appl. Mater. Interfaces*, 2022, doi: 10.1021/ACSAMI.2C08977/.

[28] H. Li *et al.*, “Combining Computational Screening and Machine Learning to Predict Metal–Organic Framework Adsorbents and Membranes for Removing CH4 or H2 from Air,”

Membranes (Basel)., 2022, doi: 10.3390/MEMBRANES12090830/S1.

[29] J. D. Evans, D. M. Huang, M. Haranczyk, A. W. Thornton, C. J. Sumby, and C. J. Doonan, "Computational identification of organic porous molecular crystals," *CrystEngComm*, 2016, doi: 10.1039/C6CE00064A.

[30] Z. Wang, T. Zhou, and K. Sundmacher, "Interpretable machine learning for accelerating the discovery of metal-organic frameworks for ethane/ethylene separation," *Chem. Eng. J.*, 2022, doi: 10.1016/J.CEJ.2022.136651.

[31] C. Altintas, O. F. Altundal, S. Keskin, and R. Yildirim, "Machine Learning Meets with Metal Organic Frameworks for Gas Storage and Separation," *J. Chem. Inf. Model.*, doi: 10.1021/ACS.JCIM.1C00191.

[32] C. F. Galinha and J. G. Crespo, "From Black Box to Machine Learning: A Journey through Membrane Process Modelling," *Membr. 2021,* , doi: 10.3390/MEMBRANES11080574.

[33] "Materials Genome Initiative | WWW.MGI.GOV." https://www.mgi.gov/.

[34] "Top page | Materials Research by Information Integration Initiative." https://www.nims.go.jp/MII-I/en/.

[35] "NOMAD Repository & Archive - NOMAD Lab." https://nomad-lab.eu/services/repo-arch.

[36] J. W. Barnett *et al.*, "Designing exceptional gas-separation polymer membranes using machine learning," *Sci. Adv.*, 2020, doi: 10.1126/SCIADV.AAZ4301.

[37] H. Gao *et al.*, "Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization," *Environ. Sci. Technol.*, 2021, doi: 10.1021/ACS.EST.1C04373.

[38] J. Yang, L. Tao, J. He, J. R. Mccutcheon, and Y. Li, "Discovery of Innovative Polymers for Next-Generation Gas-Separation Membranes using Interpretable Machine Learning," 2021, doi: 10.26434/CHEMRXIV-2021-P4G7Z.

[39] L. Tao, G. Chen, and Y. Li, "Machine learning discovery of high-temperature polymers," *Patterns (New York, N.Y.)*, 2021, doi: 10.1016/J.PATTER.2021.100225.

[40] G. Chen, L. Tao, and Y. Li, “Predicting Polymers’ Glass Transition Temperature by a Chemical Language Processing Model,” *Polym. 2021,* doi: 10.3390/POLYM13111898.

[41] L. Tao, V. Varshney, and Y. Li, “Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature,” *J. Chem. Inf. Model.*, 2021, doi: 10.1021/ACS.JCIM.1C01031/SUPPL_FILE/CI1C01031_SI_003.XLSX.

[42] W. Sun *et al.*, “Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials,” *Sci. Adv.*, 2019, doi: 10.1126/SCIADV.AAY4275/SUPPL_FILE/AAY4275_SM.PDF.

[43] R. Gómez-Bombarelli *et al.*, “Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach,” *Nat. Mater.*, 2016, doi: 10.1038/NMAT4717.

[44] I. Ibrar *et al.*, “Evaluation of machine learning algorithms to predict internal concentration polarization in forward osmosis,” *J. Memb. Sci.*, 2022, doi: 10.1016/J.MEMSCI.2022.120257.

[45] D. J. Kovacs *et al.*, “Membrane fouling prediction and uncertainty analysis using machine learning: A wastewater treatment plant case study,” *J. Memb. Sci.*, 2022, doi: 10.1016/J.MEMSCI.2022.120817.

[46] D. Rall, A. M. Schweidtmann, M. Kruse, E. Evdochenko, A. Mitsos, and M. Wessling, “Multi-scale membrane process optimization with high-fidelity ion transport models through machine learning,” *J. Memb. Sci.*, 2020, doi: 10.1016/J.MEMSCI.2020.118208.

[47] A. V. Dudchenko and M. S. Mauter, “Neural networks for estimating physical parameters in membrane distillation,” *J. Memb. Sci.*, 2020, doi: 10.1016/J.MEMSCI.2020.118285.

[48] M. T. Gaudio, G. Coppola, L. Zangari, S. Curcio, S. Greco, and S. Chakraborty, “Artificial Intelligence-Based Optimization of Industrial Membrane Processes,” *Earth Syst. Environ.*, 2021, doi: 10.1007/S41748-021-00220-X.

[49] M. Bagheri, A. Akbari, and S. A. Mirbagheri, “Advanced control of membrane fouling in

filtration systems using artificial intelligence and machine learning techniques: A critical review," *Process Saf. Environ. Prot.*, 2019, doi: 10.1016/J.PSEP.2019.01.013.

[50] T. Cawte and A. Bazylak, "Accurately predicting transport properties of porous fibrous materials by machine learning methods," *Electrochem. Sci. Adv.*, 2022, doi: 10.1002/ELSA.202100185.

[51] N. Jeong, T. H. Chung, and T. Tong, "Predicting Micropollutant Removal by Reverse Osmosis and Nanofiltration Membranes: Is Machine Learning Viable?," *Environ. Sci. Technol.*, 2021, doi: 10.1021/ACS.EST.1C04041/SUPPL_FILE/ES1C04041_SI_003.XLSX.

[52] Z. Fu, "Water Quality Prediction Based on Machine Learning Techniques," *UNLV Theses, Diss. Prof. Pap. Capstones*, 2020, doi: 10.34917/22110053.

[53] D. Venkata Vara Prasad *et al.*, "Automating water quality analysis using ML and auto ML techniques," *Environ. Res.*, 2021, doi: 10.1016/J.ENVRES.2021.111720.

[54] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*, 1st ed. 2020. Cham: Springer International Publishing.

[55] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science (80-.).*, 2018, doi: 10.1126/SCIENCE.AAT2663.

[56] A. Zunger, "Inverse design in search of materials with target functionalities," *Nat. Rev. Chem.,* 2018, doi: 10.1038/s41570-018-0121.

[57] P. I. Frazier and J. Wang, "Bayesian optimization for materials design," *Springer Ser. Mater. Sci.*, doi: 10.1007/978-3-319-23871-5_3/COVER.

[58] Q. Liang *et al.*, "Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains," *npj Comput. Mater.,* 2021, doi: 10.1038/s41524-021-00656-9.

[59] K. Wang and A. W. Dowling, "Bayesian optimization for chemical products and functional materials," *Curr. Opin. Chem. Eng.*, 2022, doi: 10.1016/J.COCHE.2021.100728.

[60] L. M. Robeson, C. D. Smith, and M. Langsam, “A group contribution approach to predict permeability and permselectivity of aromatic polymers,” *J. Memb. Sci.*, 1997, doi: 10.1016/S0376-7388(97)00031-8.

[61] A. Zunger, “Inverse design in search of materials with target functionalities,” *Nat. Rev. Chem. 2018,* doi: 10.1038/s41570-018-0121.

[62] J. Lewis *et al.*, “Agglomeration tendency and activated carbon concentration effects on activated carbon-polysulfone mixed matrix membrane performance: A design of experiment formulation study,” *J. Appl. Polym. Sci.*, 2022, doi: 10.1002/APP.52875.

[63] M. Fetanat, M. Keshtiara, R. Keyikoglu, A. Khataee, R. Daiyan, and A. Razmjou, “Machine learning for design of thin-film nanocomposite membranes,” *Sep. Purif. Technol.*, 2021, doi: 10.1016/J.SEPPUR.2021.118383.

[64] M. Fetanat *et al.*, “Machine Learning for Advanced Design of Nanocomposite Ultrafiltration Membranes,” *Ind. Eng. Chem. Res.*, 2021, doi: 10.1021/ACS.IECR.0C05446/SUPPL_FILE/IE0C05446_SI_001.PDF.

[65] C. S. H. Yeo, Q. Xie, X. Wang, and S. Zhang, “Understanding and optimization of thin film nanocomposite membranes for reverse osmosis with machine learning,” *J. Memb. Sci.*, 2020, doi: 10.1016/J.MEMSCI.2020.118135.

[66] T. Le, V. C. Epa, F. R. Burden, and D. A. Winkler, “Quantitative structure-property relationship modeling of diverse materials properties,” *Chem. Rev.*, 2012, doi: 10.1021/CR200066H/SUPPL_FILE/CR200066H_SI_001.PDF.

[67] J. J. De Pablo, B. Jones, C. L. Kovacs, V. Ozolins, and A. P. Ramirez, “The Materials Genome Initiative, the interplay of experiment, theory and computation,” *Curr. Opin. Solid State Mater. Sci.*, 2014, doi: 10.1016/J.COSSMS.2014.02.003.

[68] D. J. Audus and J. J. De Pablo, “Polymer Informatics: Opportunities and Challenges,” *ACS Macro Lett.*, 2017, doi: 10.1021

[69] H. Hasnaoui, M. Krea, and D. Roizard, “Neural networks for the prediction of polymer

permeability to gases," *J. Memb. Sci.*, 2017, doi: 10.1016/J.MEMSCI.2017.07.031.

[70] "Membrane Society of Australasia | Membrane Science & Technology | Membrane Research." https://www.membrane-australasia.org/

[71] "Polymer Database(PoLyInfo) - DICE :: National Institute for Materials Science." https://polymer.nims.go.jp/en/.

[72] "Materials Project - Home." https://materialsproject.org/.

[73] C. A. Becker, F. Tavazza, Z. T. Trautt, and R. A. Buarque De Macedo, "Considerations for choosing and using force fields and interatomic potentials in materials science and engineering," *Curr. Opin. Solid State Mater. Sci.*, 2013, doi: 10.1016/J.COSSMS.2013.10.001.

[74] L. M. Hale, Z. T. Trautt, and C. A. Becker, "Evaluating variability with atomistic simulations: The effect of potential and calculation methodology on the modeling of lattice and elastic constants," *Model. Simul. Mater. Sci. Eng.*,2018, doi: 10.1088/1361-651X/AABC05.

[75] "Materials Data Repository | NIST." https://www.nist.gov/programs-projects/materials-data-repository.

[76] "Standard Reference Data | NIST." https://www.nist.gov/srd.

[77] "PubChem." https://pubchem.ncbi.nlm.nih.gov.

[78] "Online Materials Information Resource - MatWeb." https://www.matweb.com.

[79] "NIMS Materials Database(MatNavi) - DICE :: National Institute for Materials Science." https://mits.nims.go.jp/en.

[80] "NanomaterialRegistry|re3data.org." https://www.re3data.org/repository/r3d100011129.

[81] "Nanoporous Material Explorer." https://materialsproject.org/porous.html

[82] Y. G. Chung *et al.*, "Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal-Organic Framework Database: CoRE MOF 2019," *J. Chem. Eng. Data*,

2019, doi: 10.1021/ACS.JCED.9B00835/SUPPL_FILE/JE9B00835_SI_004.ZIP.

[83] M. D. Guiver *et al.*, "Gas Transport in a Polymer of Intrinsic Microporosity (PIM-1) Substituted with Pseudo-Ionic Liquid Tetrazole-Type Structures," *Macromolecules*, 2020, doi: 10.1021/ACS.MACROMOL.0C01321.

[84] R. A. Kirk, M. Putintseva, A. Volkov, and P. M. Budd, "The potential of polymers of intrinsic microporosity (PIMs) and PIM/graphene composites for pervaporation membranes," *BMC Chem. Eng. 2019*, doi: 10.1186/S42480-019-0018-4.

[85] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, and M. Yamazaki, "PoLyInfo: Polymer database for polymeric materials design," *Proc. - 2011 Int. Conf. Emerg. Intell. Data Web Technol. EIDWT 2011*,

[86] "Polymer Gas Separation Membrane Database – Membrane Society of Australasia." https://membrane-australasia.org/msa-activities/polymer-gas-separation-membrane-database.

[87] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *J. Comput. Chem.*, 2011, doi: 10.1002/JCC.21707.

[88] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, 2010, doi: 10.1021/CI100050T/ASSET/IMAGES/MEDIUM/CI-2010-00050T_0018.GIF.

[89] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *J. Comput. Aided. Mol. Des.*, 2016, doi: 10.1007/S10822-016-9938-8/FIGURES/11.

[90] S. Zhong, J. Hu, X. Yu, and H. Zhang, "Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation," *Chem. Eng. J.*, 2021, doi: 10.1016/J.CEJ.2020.127998.

[91] H. Moriwaki, Y. S. Tian, N. Kawashita, and T. Takagi, "Mordred: A molecular descriptor calculator," *J. Cheminform.*, 2018, doi: 10.1186/S13321-018-0258-Y/FIGURES/6.

[92] “RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling - PDF Free Download.” https://docplayer.net/11897218-Rdkit-a-software-suite-for-cheminformatics-computational-chemistry-and-predictive-modeling.html.

[93] D. Rogers and M. Hahn, “Extended-Connectivity Fingerprints,” *J. Chem. Inf. Model.*, 2010, doi: 10.1021/CI100050T.

[94] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, “The high-throughput highway to computational materials design,” *Nat. Mater.* 2013, doi: 10.1038/nmat3568.

[95] D. T. Larose, “Discovering Knowledge in Data: An Introduction to Data Mining.”

[96] A. Dardzinska and M. Zdrodowska, “Classification algorithms in the material science and engineering data mining techniques,” *IOP Conf. Ser. Mater. Sci. Eng.*, 2020, doi: 10.1088/1757-899X/770/1/012096.

[97] H. Hu, N. Nguyen, C. He, and P. Li, “Advanced Outlier Detection Using Unsupervised Learning for Screening Potential Customer Returns,” 2020, doi: 10.1109/ITC44778.2020.9325225.

[98] M. Balamurali, K. L. Silversides, and A. Melkumyan, “A comparison of t-SNE, SOM and SPADE for identifying material type domains in geological data,” *Comput. Geosci.*, 2019, doi: 10.1016/J.CAGEO.2019.01.011.

[99] L. Tao, V. Varshney, and Y. Li, “Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature,” *J. Chem. Inf. Model.*, 2021, doi: 10.1021/ACS.JCIM.1C01031/SUPPL_FILE/CI1C01031_SI_003.XLSX.

[100] T. Liu, L. Liu, F. Cui, F. Ding, Q. Zhang, and Y. Li, “Predicting the performance of polyvinylidene fluoride, polyethersulfone and polysulfone filtration membranes using machine learning,” *J. Mater. Chem. A*, 2020, doi: 10.1039/D0TA07607D.

[101] “Model Fit: Underfitting vs. Overfitting - Amazon Machine Learning.” https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-

overfitting.html.

[102] L. M. Robeson, “The upper bound revisited,” *J. Memb. Sci.*, 2008, doi: 10.1016/J.MEMSCI.2008.04.030.

[103] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, “Taking the human out of the loop: A review of Bayesian optimization,” *Proc. IEEE*, doi: 10.1109/JPROC.2015.2494218.

[104] K. Wang and A. W. Dowling, “Bayesian optimization for chemical products and functional materials,” 2021.

[105] M. Plock, S. Burger, and P.-I. Schneider, “Recent advances in Bayesian optimization with applications to parameter reconstruction in optical nano-metrology,” *Model. Asp. Opt. Metrol. VIII*, 2021, doi: 10.1117/12.2592266.

[106] “Unsupervised Learning For Anomaly Detection | by Vardaan Bajaj | Towards Data Science.”-https://towardsdatascience.com/unsupervised-learning-for-anomaly-detection-44c55a96b8c1.

[107] P. Baldi and R. Nasr, “When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values,” *J. Chem. Inf. Model.*, 2010, doi: 10.1021/CI100010V.

[108] B. Bolto, T. Tran, M. Hoang, and Z. Xie, “Crosslinked poly(vinyl alcohol) membranes,” *Prog. Polym. Sci.*, 2009, doi: 10.1016/J.PROGPOLYMSCI.2009.05.003.

[109] J. Yin and B. Deng, “Polymer-matrix nanocomposite membranes for water treatment,” *J. Memb. Sci.*, 2015, doi: 10.1016/J.MEMSCI.2014.11.019.

[110] J. Wang, A. Cahyadi, B. Wu, W. Pee, A. G. Fane, and J. W. Chew, “The roles of particles in enhancing membrane filtration: A review,” *J. Memb. Sci.*, 2020, doi: 10.1016/J.MEMSCI.2019.117570.

[111] T. T. Moore and W. J. Koros, “Non-ideal effects in organic–inorganic materials for gas separation membranes,” *J. Mol. Struct.*, 2005, doi: 10.1016/J.MOLSTRUC.2004.05.043.

[112] R. Gozalbes, J. P. Doucet, and F. Derouin, “Application of topological descriptions in QSAR and drug design: History and new trends,” *Curr. Drug Targets - Infect. Disord.*, 2002, doi: 10.2174/1568005024605909.

[113] J. Gálvez, R. Garcia, M. T. Salabert, and R. Soler, “Charge Indexes. New Topological Descriptors,” *J. Chem. Inf. Comput. Sci.*, 1994, doi: 10.1021/CI00019A008/ASSET/CI00019A008.FP.PNG_V03.

[114] C. Altintas, O. F. Altundal, S. Keskin, and R. Yildirim, “Machine Learning Meets with Metal Organic Frameworks for Gas Storage and Separation,” *J. Chem. Inf. Model.*, 2021, doi: 10.1021/ACS.JCIM.1C00191/ASSET/IMAGES/LARGE/CI1C00191_0005.JPEG.

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

[116] E. Y. Lee, G. C. L. Wong, and A. L. Ferguson, “Machine learning-enabled discovery and design of membrane-active peptides,” *Bioorg. Med. Chem.*, 2018, doi: 10.1016/J.BMC.2017.07.012.

[117] E. Y. Lee, B. M. Fulan, G. C. L. Wong, and A. L. Ferguson, “Mapping membrane activity in undiscovered peptide sequence space using machine learning,” *Proc. Natl. Acad. Sci. U. S. A.*, 2016, doi: 10.1073/PNAS.1609893113/SUPPL_FILE/PNAS.1609893113.SAPP.PDF.

[118] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” *NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess.*, 2016, doi: 10.48550/arxiv.1602.04938.

[119] B. Bengfort and R. Bilbro, “Yellowbrick: Visualizing the Scikit-Learn Model Selection Process,” *J. Open Source Softw.*, 2019, doi: 10.21105/JOSS.01075.

[120] Z. Yao *et al.*, “Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models,” 2020, doi: 10.26434/CHEMRXIV.12186681.V1.

[121] R. Gómez-Bombarelli *et al.*, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Cent. Sci.*, 2016, doi: 10.1021/acscentsci.7b00572.