DISSERTATION

# ELUCIDATING THE PERFORMANCE AND MECHANISMS OF MEMBRANE SEPARATION: THE USE OF ARTIFICIAL INTELLIGENCE AND A CASE STUDY OF PRODUCED WATER TREATMENT

Submitted by:

Nohyeong Jeong

Department of Civil and Environmental Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2023

Doctoral Committee:

Advisor: Tiezheng Tong

Kenneth Carlson
Sybil Sharvelle
Todd Bandhauer

ABSTRACT

ELUCIDATING THE PERFORMANCE AND MECHANISMS OF MEMBRANE SEPARATION: THE USE OF ARTIFICIAL INTELLIGENCE AND A CASE STUDY OF PRODUCED WATER TREATMENT

Pressure-driven membrane technologies such as nanofiltration (NF) and reverse osmosis (RO) have been widely used in water and wastewater treatment because of their effective removal of contaminants and exceptional energy efficiencies. The performance of NF and RO membranes is regulated by the well-documented permeability-selectivity tradeoff, in which an increase of membrane permeability typically occurs at the expense of membrane selectivity and vice versa. To break the upper bound of this tradeoff and further enhance the efficiency of NF and RO treatment, a mechanistic understanding of the solute transport across membranes with pore sizes at the nanometer- or angstrom-scale is required.

Current theoretical models relating to solute transport across membranes are limited as the models require precise acquisition of multiple parameters. Machine learning (ML) models, a data-driven approach, have been applied to predict membrane performance and elucidate the membrane separation mechanisms. However, whether the ML models possess appropriate knowledge on membrane separation mechanisms has not yet been studied. Probing knowledge of ML models on membrane separation mechanisms can enhance the reliability of the ML model, which is of great importance to the implementation of ML models for decision-making processes, such as membrane design and selection. Moreover, contrary to the well-controlled experiments for studying the mechanisms or models associated with solute transport, where a limited number of defined solutes are present, membrane treatment has been used to treat wastewater containing

diverse organic and inorganic compounds. Thus, along with fundamental research on predictive ML models for membrane performance, investigating the performance of membranes for treating wastewater with complex compositions is also valuable to provide knowledge of solute transport across membranes in practical applications.

In this thesis, I present both a fundamental study of probing solute transport across NF and RO membranes using ML models and an applied study that explores membrane treatment of unconventional oil and gas (UOG) produced water. First, the reliability of the ML model as a tool to predict membrane performance was investigated. Specifically, the influence of data leakage on the ML model performance, as well as the solution to prevent this issue, was explored to evaluate the prediction capability of the ML model objectively. I discovered that data leakage can lead to falsely high prediction accuracy of the ML model, and appropriate data splitting for the training, validation, and testing dataset is necessary to avoid data leakage.

Second, the underlying knowledge of ML models for organic and inorganic solute transport across polyamide membranes was investigated by using a model interpretation method (i.e., Shapley additive explanation, SHAP). I not only tested whether ML models are able to possess adequate knowledge on solute transport, but also utilized the SHAP method to reveal solute transport mechanisms that are typically obtained using tedious, well-controlled experiments. For the ML model applied to predicting the rejection of organic constituents by NF and RO membranes, I found that the ML model had proper knowledge of size exclusion, but its understanding of electrostatic interaction and adsorption remains rudimentary. By using ML to predict the rejection of inorganic constituents, I elucidated that explainable artificial intelligence (XAI) can capture the major governing mechanisms of ion/salt transport across polyamide membranes (i.e., size

exclusion and electrostatic interaction), which have different importance for the transport of single salt, cation, anion transport in mixture salt solution.

Lastly, the performance of RO/NF membranes for the treatment of UOG produced water was explored as a case study, which comprehensively investigated the chemical composition and toxicity level of the treated water. NF permeates, which still had high salinities and high boron concentrations, were found to be inappropriate for irrigation and livestock drinking water, while RO membranes effectively removed most pollutants and met most water quality standards for beneficial reuse (i.e., irrigation and livestock drinking water). However, the chloride concentrations and sodium adsorption ratio (SAR) values of RO permeates were still higher than the recommended thresholds for irrigation. Also, surfactants with molecular weights higher than the molecular weight cut-off of RO/NF membranes were able to traverse through the membrane, indicating that NF and RO are not complete barriers against organic contaminants. The toxicity test results of NF and RO permeates demonstrated that NF permeates were still toxic to *Daphnia*, while RO permeates showed less toxicity than NF permeates or no toxicity. The toxicity level of NF and RO permeates showed a correlation with salinity in the permeates, which might be the main driver of the toxicity.

I envision that my thesis provides a framework to evaluate the knowledge and reliability of ML model predictions, while presenting a comprehensive investigation on membrane performance and the potential risks associated with membrane treatment of UOG produced water for beneficial reuse. The knowledge gained in this thesis improves our capability for rational membrane material design and selection, which has the potential to lead to more efficient NF and RO technologies for sustainable water and wastewater treatment.

# ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my academic advisor, Dr. Tiezheng Tong, for his everlasting patience, support, and insightful advice throughout my Ph.D. study. Without his guidance, any academic achievements, including this thesis, would not have been possible. I have learned so many valuable lessons about research and life from Dr. Tong that I will keep in mind forever.

I would also like to express my gratitude to my committee members, Dr. Kenneth Carlson, Dr. Sybil Sharvelle, and Dr. Todd Bandhauer, for their valuable comments and support during my Ph.D. study. Their insightful advice and suggestions helped me to improve the quality of this thesis. Additionally, I sincerely appreciate my previous and current lab members, Yiming Yin, Xuewei Du, Cristian Robbins, Shinyun Park, Yiqun Yao, Xijia Ge, Connor Coolidge, and Tai-heng Chung, for their help and friendship during my Ph.D. journey.

I would like to acknowledge the generous funding I received from the Korean government scholarship program and the hatch project of the United State Department of Agriculture (via Agricultural Experiment Station of Colorado State University), which supported my Ph.D. study.

Lastly, I would like to express my deepest gratitude to my wife, Yein Choi, for her unconditional love and support during last six years of my graduate study in the US. I could not have successfully completed my Ph.D. without her help. Also, I am grateful for the support and love from my beloved sons, Woojin, Woochan, and Wooseong, who motivated and encouraged me to successfully complete my Ph.D. study.

# DEDICATION

I would like to dedicate this thesis to my sons, Woojin, Woochan, and Wooseong, who are eager to make Iron Man suits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# 1. Introduction

Nanofiltration (NF) and reverse osmosis (RO) membranes have been widely used to separate solutes from solvents for wastewater treatment, resource recovery, and seawater desalination.[1-5] The interfacial polymerization, which is the state-of-the-art membrane fabrication method, has been widely used to fabricate NF and RO membranes.[6] By using a chemical reaction between an amine monomer and an acyl chloride at the water-organic interface, a thin polyamide active layer is created, which provides efficient separation of solutes (e.g., salts and pollutants) from water.[6] The performance of polyamide membranes is governed by the well-documented permeability-selectivity trade-off.[7-9] The advancement of membrane fabrication techniques significantly improves the energy efficiency and permeability of NF and RO membranes, and further development may lead to a limited contribution to membrane permeability.[7] In recent years, there has been a rise in demand for high selectivity membranes that can selectively separate valuable or undesirable solutes from others.[7, 10, 11] To fabricate novel membranes with high selectivity, it is important to obtain proper knowledge of transport mechanisms across membrane pores at the nanoscale.

The membrane separation mechanisms have been studied by using theoretical models, such as the Donnan-steric pore model with dielectric exclusion (DSPM-DE).[12-14] In the DSPM-DE model, the solute partitioning is described by the theoretical equations pertaining to size exclusion, Donnan exclusion, and dielectric exclusion, while the solute transport along the membrane pore is described by the extended Nernst-Plank (ENP) equation.[12] The DSPM-DE model is based on many assumptions to make the description of the solute transport simple.[12, 14] Therefore, this model only considers membrane separation mechanisms that are included in its assumptions. If there are important components that are not considered by the theoretical model, the influences of those

components cannot be investigated. Moreover, the application of the DSPM-DE model requires the acquisition of membrane properties (e.g., effective membrane pore size, effective active layer thickness, surface charge density, and dielectric constant of the membranes).[12] These variables are typically obtained by fitting the experimental results of membrane performance.[15, 16] Because of their interdependent relationships, a small uncertainty of one variable can cause high uncertainties in the others, which affects the interpretation of membrane separation mechanisms.[12] Further, the transition-state theory (TST) model has also provided valuable insights into the molecular-level solute transport mechanisms.[17-20] According to TST, a solute must undergo an unstable energy transition state to translocate from one coordinate to another.[11] The transition state is related to entropic and enthalpic energy barriers, allowing the transport of only solutes with enough energy to overcome the energy barrier.[11] Although current research about the influence of entropic (size exclusion) and enthalpic (ion dehydration) energy barriers reveals important knowledge of membrane separation mechanisms, the TST model also requires precise acquisition of energy barriers for solute transport with experiments, limiting the application of the model. Hence, it is required to develop an alternative approach that is not limited to any assumptions and less sensitive to an uncertainty of certain variable. Such an approach is easy to use and allows comprehensive exploration of membrane separation mechanisms.

The development of computational models and the rising access to large datasets have opened the door to alternative data-driven models using the machine learning (ML) approach. ML has been utilized for comprehensive analyses in the field of environmental science and engineering where the identification of the nonlinear relationship between input and output is required.[21-23] As the ML models mathematically analyze the complex nonlinear problems, any variables related to membrane separation can be incorporated to the ML model without preoccupied knowledge of

certain mechanisms. The ML algorithms allow leveraging a wide variety of experimental data available in the literature as well as considering any factors that may be overlooked in the theoretical models. However, due to the 'black-box' nature of the ML model, the studies of membrane separation mechanisms have been focused on predicting the power of the ML models.[22, 24, 25] Although high prediction accuracies were reported in the literature,[24, 25] it is still unknown whether the predictions by ML models are based on the adequate knowledge of membrane separation mechanisms. In several fields, such as medical informatics and polymer synthesis, explainable artificial intelligence (XAI) has been applied to explain the algorithms of ML by discovering important factors that determine the model outcome.[26-28] XAI enables users to judge whether a prediction is reliable and provides insight into the reasoning of the prediction, thereby evaluating the reliability of the ML model. Applying XAI in investigating ion transport mechanisms allows investigating whether the knowledge of ML models is aligned with the domain knowledge. This can unveil the important variables for model predictions and facilitate the membrane design process tailored to fit-for-purpose applications.

In addition, the application of membrane technology to the treatment of real wastewater is essential to reveal the membrane performance in the presence of various organic and inorganic contaminants. There have been growing concerns on managing unconventional oil and gas (UOG) produced water, which is generated during UOG extraction.[29] Produced water contains high concentrations of salts, as well as hydrocarbons and hazardous organic compounds, rendering it difficult to treat and reuse produced water.[30] Current management of UOG produced water mainly relies on deep-well injection into Class II wells.[31] However, considering the high volume of produced water (20 billion barrels every year in the U.S.[32]) and limited volume of wells, deep-well injection cannot be a sustainable management method for produced water, and it is urgent to

develop a more sustainable management methods for produced water. Recently, there are more studies on the treatment of unconventional oil and gas (UOG) produced water using membrane technology as a potential solution for addressing the dual challenges of water scarcity and pollution posed by UOG production.[30, 33, 34] The comprehensive investigation on the membrane performance for treating UOG produced water can provide valuable knowledge on solute transport across membranes in the presence of complex compositions of contaminants, as well as potential risks and opportunities of using the treated produced water for beneficial reuse.

In this dissertation, the research background associated with mechanistic models of solute transport through membranes, machine learning, and the treatment of UOG produced water is provided in Chapter 2. The primary research objectives are listed in Chapter 3. In Chapter 4, the influence of data leakage on the model prediction that hinders proper evaluation of the ML model performance is investigated. I also apply XAI to reveal the knowledge of the ML model on the transport of organic pollutants across polyamide membranes in this chapter. In Chapter 5, XAI is utilized to understand the underlying knowledge of ML models for inorganic solute transport through membranes. Chapter 6 focuses on probing the performance of RO and NF employed in the treatment of UOG produced water as a case study, with the potential risks of reusing the treated produced water for agricultural applications evaluated. Finally, conclusions and the recommended future research are provided in Chapter 7.

**References**


1. Li, X.; Mo, Y.; Qing, W.; Shao, S.; Tang, C. Y.; Li, J., Membrane-based technologies for lithium recovery from water lithium resources: A review. *Journal of Membrane Science* 2019, *591*, 117317.
2. Amy, G.; Ghaffour, N.; Li, Z.; Francis, L.; Linares, R. V.; Missimer, T.; Lattemann, S., Membrane-based seawater desalination: Present and future prospects. *Desalination* 2017, *401*, 16-21.
3. Bunani, S.; Yörükoğlu, E.; Sert, G.; Yüksel, Ü.; Yüksel, M.; Kabay, N., Application of nanofiltration for reuse of municipal wastewater and quality analysis of product water. *Desalination* 2013, *315*, 33-36.
4. Liang, S.; Liu, C.; Song, L., Two-step optimization of pressure and recovery of reverse osmosis desalination process. *Environmental science & technology* 2009, *43*, (9), 3272-3277.
5. Li, K.; Wang, J.; Liu, J.; Wei, Y.; Chen, M., Advanced treatment of municipal wastewater by nanofiltration: Operational optimization and membrane fouling analysis. *Journal of environmental sciences* 2016, *43*, 106-117.
6. Paul, M.; Jons, S. D., Chemistry and fabrication of polymeric nanofiltration membranes: A review. *Polymer* 2016, *103*, 417-456.
7. Werber, J. R.; Deshmukh, A.; Elimelech, M., The Critical Need for Increased Selectivity, Not Increased Water Permeability, for Desalination Membranes. *Environmental Science & Technology Letters* 2016, *3*, (4), 112-120.
8. 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), eaab0530.
9. Labban, O.; Liu, C.; Chong, T. H.; Lienhard, J. H., Relating transport modeling to nanofiltration membrane fabrication: Navigating the permeability-selectivity trade-off in desalination pretreatment. *Journal of Membrane Science* 2018, *554*, 26-38.
10. Liang, Y.; Zhu, Y.; Liu, C.; Lee, K.-R.; Hung, W.-S.; Wang, Z.; Li, Y.; Elimelech, M.; Jin, J.; Lin, S., Polyamide nanofiltration membrane with highly uniform sub-nanometre pores for sub-1 Å precision separation. *Nature communications* 2020, *11*, (1), 2015.
11. Epsztein, R.; DuChanois, R. M.; Ritt, C. L.; Noy, A.; Elimelech, M., Towards single-species selectivity of membranes with subnanometre pores. *Nature Nanotechnology* 2020, *15*, (6), 426-436.
12. Wang, R.; Lin, S., Pore model for nanofiltration: History, theoretical framework, key predictions, limitations, and prospects. *Journal of Membrane Science* 2021, *620*, 118809.
13. Roy, Y.; Warsinger, D. M.; Lienhard, J. H., Effect of temperature on ion transport in nanofiltration membranes: Diffusion, convection and electromigration. *Desalination* 2017, *420*, 241-257.
14. Ali, F. A. A.; Alam, J.; Shukla, A. K.; Almutairi, Z. A.; Alhoshan, M., Assessing the properties of thin-film nanocomposite membrane embedded with GO nanosheets using the DSPM-DE model. *Journal of Materials Research and Technology* 2022, *19*, 74-90.
15. Hussain, A.; Abashar, M.; Al-Mutaz, I., Influence of ion size on the prediction of nanofiltration membrane systems. *Desalination* 2007, *214*, (1-3), 150-166.
16. Bowen, W. R.; Mohammad, A. W.; Hilal, N., Characterisation of nanofiltration membranes for predictive purposes—use of salts, uncharged solutes and atomic force microscopy. *Journal of membrane science* 1997, *126*, (1), 91-105.

17. Pavluchkov, V.; Shefer, I.; Peer-Haim, O.; Blotevogel, J.; Epsztein, R., Indications of ion dehydration in diffusion-only and pressure-driven nanofiltration. *Journal of Membrane Science* 2022, *648*, 120358.
18. Shefer, I.; Peer-Haim, O.; Leifman, O.; Epsztein, R., Enthalpic and Entropic Selectivity of Water and Small Ions in Polyamide Membranes. *Environmental Science & Technology* 2021, *55*, (21), 14863-14875.
19. Epsztein, R.; Cheng, W.; Shaulsky, E.; Dizge, N.; Elimelech, M., Elucidating the mechanisms underlying the difference between chloride and nitrate rejection in nanofiltration. *Journal of Membrane Science* 2018, *548*, 694-701.
20. Richards, L. A.; Richards, B. S.; Corry, B.; Schäfer, A. I., Experimental energy barriers to anions transporting through nanofiltration membranes. *Environmental science & technology* 2013, *47*, (4), 1968-1976.
21. Jeong, N.; Chung, T.-h.; Tong, T., Predicting Micropollutant Removal by Reverse Osmosis and Nanofiltration Membranes: Is Machine Learning Viable? *Environmental Science & Technology* 2021, *55*, (16), 11348-11359.
22. Lee, S.; Kim, J., Prediction of nanofiltration and reverse-osmosis-membrane rejection of organic compounds using random forest model. *Journal of Environmental Engineering* 2020, *146*, (11), 04020127.
23. Gupta, S.; Aga, D.; Pruden, A.; Zhang, L.; Vikesland, P., Data analytics for environmental science and engineering research. *Environmental Science & Technology* 2021, *55*, (16), 10895-10907.
24. Ammi, Y.; Khaouane, L.; Hanini, S., Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. *Korean Journal of Chemical Engineering* 2015, *32*, 2300-2310.
25. Khaouane, L.; Ammi, Y.; Hanini, S., Modeling the retention of organic compounds by nanofiltration and reverse osmosis membranes using bootstrap aggregated neural networks. *Arabian Journal for Science and Engineering* 2017, *42*, 1443-1453.
26. Gao, H.; Zhong, S.; Zhang, W.; Igou, T.; Berger, E.; Reid, E.; Zhao, Y.; Lambeth, D.; Gan, L.; Afolabi, M. A., Revolutionizing Membrane Design Using Machine Learning-Bayesian Optimization. *Environmental Science & Technology* 2021, *56*, (4), 2572-2581.
27. Song, X.; Yu, A. S.; Kellum, J. A.; Waitman, L. R.; Matheny, M. E.; Simpson, S. Q.; Hu, Y.; Liu, M., Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nature communications* 2020, *11*, (1), 5668.
28. Lauritsen, S. M.; Kristensen, M.; Olsen, M. V.; Larsen, M. S.; Lauritsen, K. M.; Jørgensen, M. J.; Lange, J.; Thiesson, B., Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature communications* 2020, *11*, (1), 1-11.
29. Al-Ghouti, M. A.; Al-Kaabi, M. A.; Ashfaq, M. Y.; Da'na, D. A., Produced water characteristics, treatment and reuse: A review. *Journal of Water Process Engineering* 2019, *28*, 222-239.
30. Chang, H.; Li, T.; Liu, B.; Vidic, R. D.; Elimelech, M.; Crittenden, J. C., Potential and implemented membrane-based technologies for the treatment and reuse of flowback and produced water from shale gas and oil plays: A review. *Desalination* 2019, *455*, 34-57.
31. Conrad, C. L.; Ben Yin, Y.; Hanna, T.; Atkinson, A. J.; Alvarez, P. J. J.; Tekavec, T. N.; Reynolds, M. A.; Wong, M. S., Fit-for-purpose treatment goals for produced waters in shale oil and gas fields. *Water Research* 2020, *173*, 115467.

32. Clark, C. E.; Veil, J. A. *Produced water volumes and management practices in the United States*; Argonne National Lab.(ANL), Argonne, IL (United States): 2009.
33. Riley, S. M.; Ahoor, D. C.; Oetjen, K.; Cath, T. Y., Closed circuit desalination of O&G produced water: An evaluation of NF/RO performance and integrity. *Desalination* 2018, *442*, 51-61.
34. Guo, C.; Chang, H.; Liu, B.; He, Q.; Xiong, B.; Kumar, M.; Zydney, A. L., A combined ultrafiltration–reverse osmosis process for external reuse of Weiyuan shale gas flowback and produced water. *Environmental Science: Water Research & Technology* 2018, *4*, (7), 942-955.

# 2. Background

## 2.1 NF and RO membrane

Membranes are semipermeable barriers that can separate organic and inorganic contaminants in the water by using the differences in permeability of contaminants and water.[1] When a membrane is placed between the feedwater and freshwater, water molecules preferentially diffuse from freshwater toward the feedwater side through a semipermeable membrane, generating osmotic pressure toward the feedwater side of the membrane.[1] Contrary to osmosis, the hydraulic pressure that is higher than the osmotic pressure is applied on the feedwater side of the membrane in RO to produce freshwater on the permeate side, while retaining organic and inorganic solutes in the feedwater.[2] RO membranes have been used to treat brackish water and seawater where rejection of monovalent ions is required.[3-9] Due to relatively larger membrane pore sizes, NF membranes have shown high rejections of small organic solutes, such as micropollutants, and divalent ions in the water, while requiring lower operation pressure and having higher water permeability compared to RO membranes.[10] The solute transport within membranes is governed by the well-documented permeability-selectivity trade-off: a higher water permeability of the membrane leads to a lower selectivity and vice versa.[11, 12] Advances in thin-film composite membranes have significantly improved the energy efficiency and permeability of NF and RO,[3] meaning that further developments of membrane permeability may lead to limited improvement in energy efficiency. Recently, there has been more demand for highly selective membranes that enable precise solute-solute separation.[13, 14] The fabrication of highly selective membranes requires a thorough understanding of solute transport mechanisms through membranes. The membrane separation mechanisms have been investigated by using theoretical models, such as the Donnan steric pore

model with dielectric exclusion (DSPM-DE)[15-19] and transition-state theory (TST).[20-23] These models are described in the following section.

## 2.2 Mechanistic models of solute transport

### *2.2.1 Donnan-steric pore model with dielectric exclusion*

The DSPM-DE model explains the role of size exclusion, Donnan exclusion, and dielectric exclusion during solute partitioning (Figure 2-1). Bowen et al.[19] proposed the Donnan-steric pore model (DSPM), which combines the equations pertaining to steric (size) exclusion and Donnan exclusion (electrostatic interaction). With the DSPM model, they successfully predicted the NF performance in dye-salt diafiltration and inorganic rejections.[17] In order to utilize the model, membrane properties such as the effective pore radius, effective ratio of membrane thickness to porosity, and effective membrane charge density should be extracted from the experimental data.[24] Although the DSPM model has been successful in predicting the transport of monovalent ions and multivalent co-ions (ions with the same charge as the membrane), its prediction accuracy was not good enough to apply for multivalent cations and mixture salt solutions.[17] To improve the description of separation processes, it is necessary to include an additional governing separation mechanism (i.e., dielectric exclusion). The inclusion of the dielectric effect reduces the membrane charge density to a realistic range as well as improves the predictions for multivalent ion transport.[24]

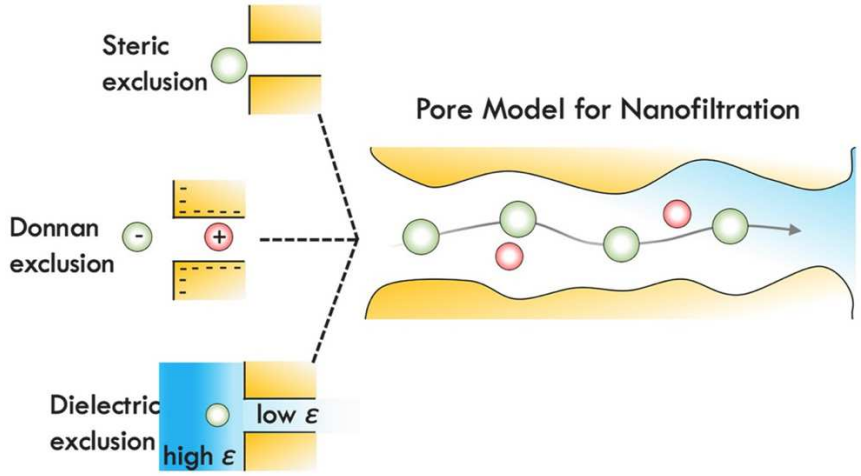


**Figure 2-1**. Solute exclusion mechanisms of the DSPM-DE model.[12]

2.2.1.1 Size exclusion

Size exclusion plays an important role in solute removal by NF and RO membranes. The membranes physically remove the solutes that are larger than the membrane pore size. For solutes smaller than the pore size, the entrance of the solute into the membrane pore depends on the ratio between the membrane pore size and solute size. The possibility of solute transport in the DSPM-DE model is described as:[24]

$$\varphi_{S,i} = \begin{cases} \left(1 - \frac{r_i}{r_p}\right)^2, \ for\ r_i < r_p \\ 0, \ for\ r_i \geq r_p \end{cases} \quad (2\text{-}1)$$

where $\varphi_{S,i}$ is Steric exclusion factor, $r_i$ is the Stokes radius of solute $i$, and $r_p$ is the average membrane pore radius.

2.2.1.2 Donnan exclusion

Donnan exclusion explains the influence of electrostatic interaction on the transport of solutes. The ion concentration difference between the membrane and solution creates Donnan potentials.[18] Due to electrostatic repulsion induced by the Donnan potential, the transport of co-ions (ions with

the same charge sign as the membrane) is hindered. Thus, co-ions with higher valence experience greater Donnan exclusion.[25] The Donnan exclusion factor for ion $i$, $\varphi_{D,i}$, can be described as followed:[24]

$$\varphi_{D,i} = \exp\left(-\frac{z_i e}{k_B T}\right)\Delta\phi_D \quad (2\text{-}2)$$

where $\Delta\phi_D$ is the Donnan potentials, $z_i$ is the solute valence, $e$ is an elemental charge, $k_B$ is the Boltzmann constant, $T$ is the solution temperature.

2.2.1.3 Dielectric exclusion

Compared to the bulk solution, the water in the confined membrane pore has a lower dielectric constant.[26] The decrease in the dielectric constant in the pore creates a solvation energy barrier, preventing ion transport from the bulk solution to the membrane pores. This means that it is energetically unfavorable for a hydrated ion to enter the pore where the dielectric constant is low. The dielectric exclusion factor is calculated as below:[24]

$$\varphi_{DE,i} = \exp\left(-\frac{\Delta W_i}{k_B T}\right) \quad (2\text{-}3)$$

The ion solvation energy barrier, $\Delta W$, is written as

$$\Delta W = \frac{z_i^2 e^2}{8\pi\varepsilon_0 r_i}\left(\frac{1}{\varepsilon_b} - \frac{1}{\varepsilon_p}\right) \quad (2\text{-}4)$$

where $\varepsilon_0$, $\varepsilon_b$, and $\varepsilon_p$ are dielectric constant of vacuum, the solvent in bulk, and the solvent in the pore.

Solute partitioning at the feed/membrane and membrane/permeate interfaces can be expressed as followed: [24]

$$\frac{r_i(x=0)c_i(x=0)}{r_{i,f}c_{i,f}} = \varphi_{D,i}\varphi_{DE,i}\varphi_{S,i} \quad (2\text{-}5)$$

where $r_i(x=0)$ and $r_{i,f}$ are the activity coefficient of ion $i$ at the entrance of membrane pore and in the feed solution, respectively. $c_i(x=0)$ and $c_{i,f}$ are corresponding concentrations of ions.

$$\frac{r_i(x=\delta)c_i(x=\delta)}{r_{i,p}c_{i,p}} = \varphi_{D,i}\varphi_{DE,i}\varphi_{S,i} \quad (2\text{-}6)$$

where $r_i(x=\delta)$ and $r_{i,p}$ are the activity coefficient of ion $i$ at the entrance of membrane pore and in the feed solution, respectively. $c_i(x=0)$ and $c_{i,p}$ are corresponding concentrations of ions.

*2.2.3 Transition-state theory*

TST has been utilized to explain the elementary reaction rate, such as diffusion and solute permeability in the membranes (Figure 2-2).[21, 22, 28] Molecular diffusion can be considered as a thermally activated process where species possessing sufficient energy are able to move from one basin to another. TST has been used to discover the role of hydration energy and ion-membrane interactions, including partitioning at the pore entrance and intrapore diffusion.[22, 29, 30]

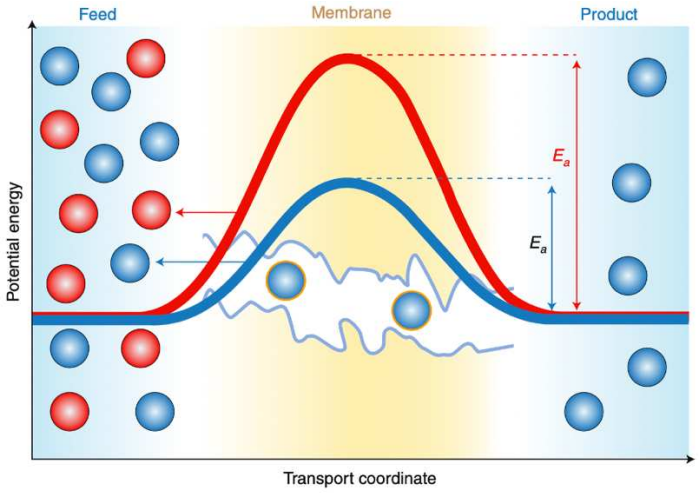


**Figure 2-2.** Energy barriers of solute during transport through membranes. The solutes that can overcome the energy barrier are able to traverse the membrane pores.[28]

The transmission coefficient, $k$, can be expressed with an activation energy of enthalpy and entropy as:[22]

$$k = \kappa \frac{k_B T}{h} \exp\left(\frac{\Delta S^{\ddagger}}{R}\right) \exp\left(-\frac{\Delta H^{\ddagger}}{RT}\right) \quad (2\text{-}10)$$

where $k$ is a rate constant, $\kappa$ is the transmission coefficient, $k_B$is the Boltzmann constant, $h$ is the Planck constant, $R$ is the universal gas constant, $T$ is the absolute temperature, $\Delta S^{\ddagger}$ is the entropy of activation, and $\Delta H^{\ddagger}$ is the enthalpy of activation.

By modifying eq (2-10), the linearized version of eq (2-11) can be obtained.

$$\ln\left(\frac{kh}{\kappa k_B T}\right) = \frac{\Delta S^{\ddagger}}{R} - \frac{\Delta H^{\ddagger}}{RT} \quad (2\text{-}11)$$

As shown in eq (11), $\Delta S^{\ddagger}$ and $\Delta H^{\ddagger}$ for water and solutes can be quantified by measuring the permeability (i.e., transmission coefficient) at different temperatures.

2.3 Machine learning

Machine learning (ML), a subset of artificial intelligence, is a powerful tool that can solve complicated multivariable problems. Supervised learning, which is able to find the relationships between input variables (e.g., information about target labels) and output (e.g., the value that an ML model wants to accurately estimate) without any preconceived knowledge, has been applied in many studies of environmental engineering to make predictions. [31-42] Thus far, the application of the ML models for research in environmental engineering has mainly been focused on improving the ML model performance due to their 'black-box' nature.[31-33] The ML model with high prediction accuracy is desirable. However, if the prediction accuracy of ML is the only metric to evaluate the ML, it is still unknown whether the machine learning models have a thorough understanding on the mechanisms. The consistency of the underlying knowledge of the ML with the domain knowledge is of great importance, especially when the ML is used in decision-making processes.[43, 44]

Explainable artificial intelligence (XAI), a tool to reveal mechanisms of the ML algorithms, has been employed in several fields, such as medical informatics and polymer synthesis, to understand the determining factors of the predictions.[45-47] XAI allows users to judge logic behind

the predictions, enabling them to evaluate whether the ML model is trustworthy. Furthermore, the application of XAI to complex multivariable problems can extract knowledge that would otherwise not be recognized.[48] For example, Gao et al. implemented ML to identify the influence of membrane fabrication conditions on polymeric membrane performance.[45] By combining XAI and Bayesian optimization, they were able to inversely design membranes that could overcome the current upper bound of permeability and selectivity trade-off.[45] Similarly, Yang et al. used XAI to establish the relationship between polymer synthesis and properties of membranes for gas separation.[49] They demonstrated that the application of XAI for screening promising polymers for gas separation membrane facilitated the membrane design process.[49] However, although XAI has been used in a few studies to optimize and facilitate the membrane fabrication process, whether the ML can capture the important membrane separation mechanisms has not yet been studied in the literature, which is important yet commonly neglected by membrane scientists and environmental engineers.

**2.4 Unconventional oil and gas produced water treatment**

The production of unconventional oil and gas (UOG) has been significantly increased due to the development in horizontal drilling and hydraulic fracturing.[50] To extract the UOG, a vertical drilling is performed to reach a shale layer, followed by drilling a hole horizontally in a target layer.[51] Once the hole is encased with a steel or cement casing, a fluid consisting of sand and chemical additives is pumped at a high pressure to fracture the shale layer, which increases UOG extraction.[51, 52] This process consumes a large amount of freshwater and generates a significant volume of produced water.[53, 54] The produced water contains naturally occurring compounds in the geologic shale formation and chemical additives used in hydraulic fracturing, which have the potential to cause adverse impacts on human health and the environment.[54-56] The management of

produced water has mainly been conducted via deep-well injection (i.e., injecting the produced water into Class II wells).[57] However, deep-well injection have several disadvantages, including induced seismicity, limited access to disposal wells, and risks of groundwater contamination, which raise concerns about the reliability and sustainability of this management method.[58, 59]

Recently, there has been an increasing focus on finding better methods to treat and recycle produced water using membrane technologies.[60-64] The main challenge of recycling produced water is the high salinity and the presence of various hazardous organic constituents, which potentially cause membrane fouling.[60, 65] Therefore, research on treating produced water by membranes has primarily focused on pretreatment,[53] membrane fouling,[66, 67] and membrane desalination.[60, 68-70] For example, Riley et al.[62] investigated the performance and fouling propensities of various membranes, such as NF90, ECO, and BW30 membranes, in closed circuit mode for long-term operation (over 400 hours). The results demonstrated that BW30 membrane was able to achieve up to 99.6% TDS rejection and 89% DOC rejection throughout the experiment with regular membrane cleaning.[62] Jang et al.[71] treated UOG produced water with RO membrane to investigate the cation rejection efficiency. They reported that the cation removal rates by the membrane was 97-99% and that the cation removal rate increased with the hydrated radius.[71] However, there are limited studies that focus on membrane treatment of organics in produced water.[60] It is crucial to probe the concentrations of the residual organic constituents in the treated produced water and their potential risks for reuse applications. A comprehensive investigation on the membrane performance and separation mechanisms for organic and inorganic solute rejections in polyamide membranes can provide insights on proper membrane selection for the treatment of produced water and the potential reuse of the treated produced water.

**References**


1. Crittenden, J. C.; Trussell, R. R.; Hand, D. W.; Howe, K. J.; Tchobanoglous, G., *MWH's water treatment: principles and design*. John Wiley & Sons: 2012.
2. Biesheuvel, P. M.; Porada, S.; Elimelech, M.; Dykstra, J. E., Tutorial review of reverse osmosis and electrodialysis. *Journal of Membrane Science* 2022, *647*, 120221.
3. Elimelech, M.; Phillip, W. A., The future of seawater desalination: Energy, technology, and the environment. *Science* 2011, *333*, (6043), 712-717.
4. Kimura, K.; Toshima, S.; Amy, G.; Watanabe, Y., Rejection of neutral endocrine disrupting compounds (EDCs) and pharmaceutical active compounds (PhACs) by RO membranes. *Journal of Membrane Science* 2004, *245*, (1), 71-78.
5. Kimura, K.; Amy, G.; Drewes, J. E.; Heberer, T.; Kim, T.-U.; Watanabe, Y., Rejection of organic micropollutants (disinfection by-products, endocrine disrupting compounds, and pharmaceutically active compounds) by NF/RO membranes. *Journal of membrane science* 2003, *227*, (1-2), 113-121.
6. Alghoul, M. A.; Poovanaesvaran, P.; Sopian, K.; Sulaiman, M. Y., Review of brackish water reverse osmosis (BWRO) system designs. *Renewable and Sustainable Energy Reviews* 2009, *13*, (9), 2661-2667.
7. Alsarayreh, A. A.; Al-Obaidi, M.; Al-Hroub, A.; Patel, R.; Mujtaba, I. M., Evaluation and minimisation of energy consumption in a medium-scale reverse osmosis brackish water desalination plant. *Journal of Cleaner Production* 2020, *248*, 119220.
8. Brehant, A.; Bonnelye, V.; Perez, M., Comparison of MF/UF pretreatment with conventional filtration prior to RO membranes for surface seawater desalination. *Desalination* 2002, *144*, (1), 353-360.
9. Shenvi, S. S.; Isloor, A. M.; Ismail, A., A review on RO membrane technology: Developments and challenges. *Desalination* 2015, *368*, 10-26.
10. Mohammad, A. W.; Teow, Y.; Ang, W.; Chung, Y.; Oatley-Radcliffe, D.; Hilal, N., Nanofiltration membranes review: Recent advances and future prospects. *Desalination* 2015, *356*, 226-254.
11. Werber, J. R.; Deshmukh, A.; Elimelech, M., The critical need for increased selectivity, not increased water permeability, for desalination membranes. *Environmental Science & Technology Letters* 2016, *3*, (4), 112-120.
12. 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), eaab0530.
13. Liang, Y.; Zhu, Y.; Liu, C.; Lee, K.-R.; Hung, W.-S.; Wang, Z.; Li, Y.; Elimelech, M.; Jin, J.; Lin, S., Polyamide nanofiltration membrane with highly uniform sub-nanometre pores for sub-1 Å precision separation. *Nature communications* 2020, *11*, (1), 2015.
14. Zhang, X., Selective separation membranes for fractionating organics and salts for industrial wastewater treatment: Design strategies and process assessment. *Journal of Membrane Science* 2022, *643*, 120052.
15. Bandini, S.; Vezzani, D., Nanofiltration modeling: the role of dielectric exclusion in membrane characterization. *Chemical Engineering Science* 2003, *58*, (15), 3303-3326.
16. Vezzani, D.; Bandini, S., Donnan equilibrium and dielectric exclusion for characterization of nanofiltration membranes. *Desalination* 2002, *149*, (1), 477-483.

17. Bowen, W. R.; Welfoot, J. S., Modelling the performance of membrane nanofiltration—critical assessment and model development. *Chemical engineering science* 2002, *57*, (7), 1121-1137.
18. Bowen, W.; Mohammad, A. W., Characterization and prediction of nanofiltration membrane performance—a general assessment. *Chemical Engineering Research and Design* 1998, *76*, (8), 885-893.
19. Bowen, W. R.; Mohammad, A. W.; Hilal, N., Characterisation of nanofiltration membranes for predictive purposes—use of salts, uncharged solutes and atomic force microscopy. *Journal of membrane science* 1997, *126*, (1), 91-105.
20. Shefer, I.; Peer-Haim, O.; Epsztein, R., Limited ion-ion selectivity of salt-rejecting membranes due to enthalpy-entropy compensation. *Desalination* 2022, *541*, 116041.
21. Pavluchkov, V.; Shefer, I.; Peer-Haim, O.; Blotevogel, J.; Epsztein, R., Indications of ion dehydration in diffusion-only and pressure-driven nanofiltration. *Journal of Membrane Science* 2022, *648*, 120358.
22. Shefer, I.; Peer-Haim, O.; Leifman, O.; Epsztein, R., Enthalpic and entropic selectivity of water and small ions in polyamide membranes. *Environmental Science & Technology* 2021, *55*, (21), 14863-14875.
23. Shefer, I.; Lopez, K.; Straub, A. P.; Epsztein, R., Applying transition-state theory to explore transport and selectivity in salt-rejecting membranes: A critical review. *Environmental Science & Technology* 2022.
24. Wang, R.; Lin, S., Pore model for nanofiltration: History, theoretical framework, key predictions, limitations, and prospects. *Journal of Membrane Science* 2021, *620*, 118809.
25. Seidel, A.; Waypa, J. J.; Elimelech, M., Role of charge (Donnan) exclusion in removal of arsenic from water by a negatively charged porous nanofiltration membrane. *Environmental engineering science* 2001, *18*, (2), 105-113.
26. Senapati, S.; Chandra, A., Dielectric constant of water confined in a nanocavity. *The Journal of Physical Chemistry B* 2001, *105*, (22), 5106-5109.
27. Dresner, L., Some remarks on the integration of the extended Nernst-Planck equations in the hyperfiltration of multicomponent solutions. *Desalination* 1972, *10*, (1), 27-46.
28. Epsztein, R.; DuChanois, R. M.; Ritt, C. L.; Noy, A.; Elimelech, M., Towards single-species selectivity of membranes with subnanometre pores. *Nature Nanotechnology* 2020, *15*, (6), 426-436.
29. Richards, L. A.; Richards, B. S.; Corry, B.; Schäfer, A. I., Experimental energy barriers to anions transporting through nanofiltration membranes. *Environmental science & technology* 2013, *47*, (4), 1968-1976.
30. Epsztein, R.; Cheng, W.; Shaulsky, E.; Dizge, N.; Elimelech, M., Elucidating the mechanisms underlying the difference between chloride and nitrate rejection in nanofiltration. *Journal of Membrane Science* 2018, *548*, 694-701.
31. Lee, S.; Kim, J., Prediction of nanofiltration and reverse-osmosis-membrane rejection of organic compounds using random forest model. *Journal of Environmental Engineering* 2020, *146*, (11), 04020127.
32. Ammi, Y.; Khaouane, L.; Hanini, S., Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. *Korean Journal of Chemical Engineering* 2015, *32*, 2300-2310.

33. Khaouane, L.; Ammi, Y.; Hanini, S., Modeling the retention of organic compounds by nanofiltration and reverse osmosis membranes using bootstrap aggregated neural networks. *Arabian Journal for Science and Engineering* 2017, *42*, 1443-1453.
34. Ai, H.; Zhang, K.; Sun, J.; Zhang, H., Short-term Lake Erie algal bloom prediction by classification and regression models. *Water Research* 2023, 119710.
35. Yang, H.; Huang, K.; Zhang, K.; Weng, Q.; Zhang, H.; Wang, F., Predicting heavy metal adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities. *Environmental Science & Technology* 2021, *55*, (20), 14316-14328.
36. Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S. S.; Lian, C.; Kwon, H.; Wong, B. M., A machine learning approach for predicting defluorination of per-and polyfluoroalkyl substances (PFAS) for their efficient treatment and removal. *Environmental Science & Technology Letters* 2019, *6*, (10), 624-629.
37. Li, L.; Rong, S.; Wang, R.; Yu, S., Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical Engineering Journal* 2021, *405*, 126673.
38. Cheng, W.; Ng, C. A., Using machine learning to classify bioactivity for 3486 per-and polyfluoroalkyl substances (PFASs) from the OECD list. *Environmental Science & Technology* 2019, *53*, (23), 13970-13980.
39. Libotean, D.; Giralt, J.; Rallo, R.; Cohen, Y.; Giralt, F.; Ridgway, H. F.; Rodriguez, G.; Phipps, D., Organic compounds passage through RO membranes. *Journal of Membrane Science* 2008, *313*, (1-2), 23-43.
40. Guo, H.; Jeong, K.; Lim, J.; Jo, J.; Kim, Y. M.; Park, J.-p.; Kim, J. H.; Cho, K. H., Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences* 2015, *32*, 90-101.
41. El-Rawy, M.; Abd-Ellah, M. K.; Fathi, H.; Ahmed, A. K. A., Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *Journal of Water Process Engineering* 2021, *44*, 102380.
42. Bernardelli, A.; Marsili-Libelli, S.; Manzini, A.; Stancari, S.; Tardini, G.; Montanari, D.; Anceschi, G.; Gelli, P.; Venier, S., Real-time model predictive control of a wastewater treatment plant based on machine learning. *Water Science and Technology* 2020, *81*, (11), 2391-2400.
43. Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I., From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2020, *2*, (1), 56-67.
44. Ribeiro, M. T.; Singh, S.; Guestrin, C. In *" Why should i trust you?" Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016; 2016; pp 1135-1144.
45. Gao, H.; Zhong, S.; Zhang, W.; Igou, T.; Berger, E.; Reid, E.; Zhao, Y.; Lambeth, D.; Gan, L.; Afolabi, M. A., Revolutionizing membrane design using machine learning-bayesian optimization. *Environmental Science & Technology* 2021, *56*, (4), 2572-2581.
46. Song, X.; Yu, A. S.; Kellum, J. A.; Waitman, L. R.; Matheny, M. E.; Simpson, S. Q.; Hu, Y.; Liu, M., Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nature communications* 2020, *11*, (1), 5668.
47. Lauritsen, S. M.; Kristensen, M.; Olsen, M. V.; Larsen, M. S.; Lauritsen, K. M.; Jørgensen, M. J.; Lange, J.; Thiesson, B., Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature communications* 2020, *11*, (1), 1-11.

48. Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-j.; Zhang, H., Machine learning: new ideas and tools in environmental science and engineering. *Environmental Science & Technology* 2021, *55*, (19), 12741-12754.
49. Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y., Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. *Science Advances* 2022, *8*, (29), eabn9545.
50. Tong, T.; Carlson, K. H.; Robbins, C. A.; Zhang, Z.; Du, X., Membrane-based treatment of shale oil and gas wastewater: The current state of knowledge. *Frontiers of Environmental Science & Engineering* 2019, *13*, 1-17.
51. Thiel, G. P.; Tow, E. W.; Banchik, L. D.; Chung, H. W.; Lienhard, J. H., Energy consumption in desalinating produced water from shale oil and gas extraction. *Desalination* 2015, *366*, 94-112.
52. Barati, R.; Liang, J.-T., A review of fracturing fluid systems used for hydraulic fracturing of oil and gas wells. *Journal of Applied Polymer Science* 2014, *131*, (16).
53. Sun, Y.; Wu, M.; Tong, T.; Liu, P.; Tang, P.; Gan, Z.; Yang, P.; He, Q.; Liu, B., Organic compounds in Weiyuan shale gas produced water: identification, detection and rejection by ultrafiltration-reverse osmosis processes. *Chemical Engineering Journal* 2021, *412*, 128699.
54. Stringfellow, W. T.; Domen, J. K.; Camarillo, M. K.; Sandelin, W. L.; Borglin, S., Physical, chemical, and biological characteristics of compounds used in hydraulic fracturing. *Journal of hazardous materials* 2014, *275*, 37-54.
55. Kondash, A.; Vengosh, A., Water footprint of hydraulic fracturing. *Environmental Science & Technology Letters* 2015, *2*, (10), 276-280.
56. Regnery, J.; Coday, B. D.; Riley, S. M.; Cath, T. Y., Solid-phase extraction followed by gas chromatography-mass spectrometry for the quantitative analysis of semi-volatile hydrocarbons in hydraulic fracturing wastewaters. *Analytical methods* 2016, *8*, (9), 2058-2068.
57. Conrad, C. L.; Ben Yin, Y.; Hanna, T.; Atkinson, A. J.; Alvarez, P. J. J.; Tekavec, T. N.; Reynolds, M. A.; Wong, M. S., Fit-for-purpose treatment goals for produced waters in shale oil and gas fields. *Water Research* 2020, *173*, 115467.
58. Ellsworth, W. L., Injection-induced earthquakes. *Science* 2013, *341*, (6142), 1225942.
59. Gregory, K.; Mohan, A. M., Current perspective on produced water management challenges during hydraulic fracturing for oil and gas recovery. *Environmental Chemistry* 2015, *12*, (3), 261-266.
60. Chang, H.; Li, T.; Liu, B.; Vidic, R. D.; Elimelech, M.; Crittenden, J. C., Potential and implemented membrane-based technologies for the treatment and reuse of flowback and produced water from shale gas and oil plays: A review. *Desalination* 2019, *455*, 34-57.
61. Dischinger, S. M.; Rosenblum, J.; Noble, R. D.; Gin, D. L.; Linden, K. G., Application of a lyotropic liquid crystal nanofiltration membrane for hydraulic fracturing flowback water: Selectivity and implications for treatment. *Journal of Membrane Science* 2017, *543*, 319-327.
62. Riley, S. M.; Ahoor, D. C.; Oetjen, K.; Cath, T. Y., Closed circuit desalination of O&G produced water: An evaluation of NF/RO performance and integrity. *Desalination* 2018, *442*, 51-61.
63. Kong, F.-x.; Sun, G.-d.; Chen, J.-f.; Han, J.-d.; Guo, C.-m.; Tong, Z.; Lin, X.-f.; Xie, Y. F., Desalination and fouling of NF/low pressure RO membrane for shale gas fracturing flowback water treatment. *Separation and Purification Technology* 2018, *195*, 216-223.

64. Wei, X.; Zhang, S.; Han, Y.; Wolfe, F. A., Treatment of petrochemical wastewater and produced water from oil and gas. *Water Environment Research* 2019, *91*, (10), 1025-1033.
65. Al-Ghouti, M. A.; Al-Kaabi, M. A.; Ashfaq, M. Y.; Da'na, D. A., Produced water characteristics, treatment and reuse: A review. *Journal of Water Process Engineering* 2019, *28*, 222-239.
66. Miller, D. J.; Huang, X.; Li, H.; Kasemset, S.; Lee, A.; Agnihotri, D.; Hayes, T.; Paul, D. R.; Freeman, B. D., Fouling-resistant membranes for the treatment of flowback water from hydraulic shale fracturing: A pilot study. *Journal of Membrane Science* 2013, *437*, 265-275.
67. Mondal, S.; Wickramasinghe, S. R., Produced water treatment by nanofiltration and reverse osmosis membranes. *Journal of Membrane Science* 2008, *322*, (1), 162-170.
68. Maguire-Boyle, S. J.; Huseman, J. E.; Ainscough, T. J.; Oatley-Radcliffe, D. L.; Alabdulkarem, A. A.; Al-Mojil, S. F.; Barron, A. R., Superhydrophilic functionalization of microfiltration ceramic membranes enables separation of hydrocarbons from frac and produced water. *Scientific Reports* 2017, *7*, (1), 12267.
69. Lester, Y.; Ferrer, I.; Thurman, E. M.; Sitterley, K. A.; Korak, J. A.; Aiken, G.; Linden, K. G., Characterization of hydraulic fracturing flowback water in Colorado: Implications for water treatment. *Science of The Total Environment* 2015, *512-513*, 637-644.
70. Butkovskyi, A.; Bruning, H.; Kools, S. A. E.; Rijnaarts, H. H. M.; Van Wezel, A. P., Organic pollutants in shale gas flowback and produced waters: Identification, potential ecological impact, and implications for treatment strategies. *Environmental Science & Technology* 2017, *51*, (9), 4740-4754.
71. Jang, E.; Jeong, S.; Chung, E., Application of three different water treatment technologies to shale gas produced water. *Geosystem Engineering* 2017, *20*, (2), 104-110.

# 3. Knowledge gap and research objectives

Based on the literature review, the following knowledge gaps on the use of ML models for predicting membrane performance and membrane treatment of UOG produced water will be addressed by my research:

- The reliability of ML models for predicting membrane performance has not been revealed.
- The knowledge of ML models on the transport of organic and inorganic solutes across polyamide membranes has not been investigated, and it is still unknown whether such knowledge is consistent with the domain knowledge of membrane science.
- There is a lack of comprehensive investigations on membrane performance for treating UOG produced water, with the chemical compositions and toxicity level of the treated produced water being rarely reported.

To close the aforementioned knowledge gaps, I conducted research with the following research objectives.

- **Developing a proper method to train and evaluate ML model for membrane performance prediction.**

Satisfactory predictive accuracies of ML models have been reported in the literature.[1, 2] However, the misuse of the ML models could cause falsely good predictive power. Given that the applications and understanding of ML in environmental engineering and science are still in the early stage, it is necessary to investigate proper methods for ML model training and evaluation for membrane performance predictions. Many ML models have been used for membrane performance prediction with improper data splitting for training, validation, and testing datasets, which potentially results in data leakage issues.[1-3] Data leakage can cause the falsely high prediction accuracy of the ML model, and it is hard to recognize the issue without domain expertise.[4] I

developed a stratified sampling method that avoids data leakage in ML model training. By comparing the ML model performance with and without data leakage, the influence of data leakage on model accuracy was also investigated.

- **Unveiling the knowledge of ML models on organic and inorganic solute transport.**

Unveiling the knowledge of ML models is an important step to validate the reliability of ML and provide opportunities to discover new knowledge on membrane separations.[4] By using XAI, the influences of various input variables (including membrane and solute properties as well as operational condition) on the transport of organic and inorganic solutes across NF and RO membranes were explored. Shapley additive explanation (SHAP), a type of XAI that is based on the cooperative game theory, was used in my study to quantify the contribution of each input variable to model predictions. The SHAP values of variables can show whether a certain variable increases or decreases the transport rates. In this way, the importance of different variables to model prediction to the transport of organic compounds, cations, and anions as predicted by ML models was investigated. I utilized the SHAP dependent plots that show the relationship between variables and their SHAP values to investigate whether the knowledge of ML is aligned with the fundamental principles of membrane science. I revealed that XAI could identify the mechanisms governing ion transport possess different relative importance to organic compound, single salt, cation, and anion transport during RO and NF filtration.

- **Characterizing the chemical compositions and toxicity levels of treated UOG produced water comprehensively.**

In the existing literatures, the experiments for membrane performance and separation mechanisms were generally conducted in well-controlled conditions, with relatively simple water compositions. Under such conditions, the membrane performance and the limitations of membrane treatment for

real wastewater may not be revealed. As a result, I applied NF and RO to the treatment of UOG produced water from the Niobrara shale play in Colorado containing high salinity and a diverse set of contaminants. The residual organic and inorganic contaminants in the permeate from NF and RO membranes with varied perm-selectivity were comprehensively investigated. Also, the toxicity level of the treated produced water was probed by a biological assay using *Daphnia magna*. The results demonstrated that RO permeates showed minor or no toxicity to *Daphnia*, while the toxicity level of NF permeates were similar to that of the MF filtrate, indicating that NF membranes could not effectively reduce the toxicity level of UOG produced water. Furthermore, the feasibility of using the treated produced water by NF and RO membranes for beneficial reuse (e.g., irrigation and livestock drinking water) was evaluated. The results showed that the NF permeates did not meet the water quality criteria for irrigation and livestock drinking water. Despite high removal rates for most contaminants in the produced water by RO, the concentrations of chloride and boron as well as sodium adsorption rate (SAR) in the RO permeates exceeded irrigation guidelines.

## References


1. Ammi, Y.; Khaouane, L.; Hanini, S., Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. *Korean Journal of Chemical Engineering* 2015, 32, (11), 2300-2310.
2. Khaouane, L.; Ammi, Y.; Hanini, S., Modeling the retention of organic compounds by nanofiltration and reverse osmosis membranes using bootstrap aggregated neural networks. *Arabian Journal for Science and Engineering* 2017, 42, (4), 1443-1453.
3. Lee, S.; Kim, J., Prediction of nanofiltration and reverse-osmosis-membrane rejection of organic compounds using random forest model. *Journal of Environmental Engineering* 2020, 146, (11), 04020127.
4. Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H., Machine learning: New ideas and tools in environmental science and engineering. *Environmental Science & Technology* 2021, 55, (19), 12741-12754.
5. Al-Rashdi, B.; Johnson, D.; Hilal, N., Removal of heavy metal ions by nanofiltration. *Desalination* 2013, 315, 2-17.
6. Boo, C.; Wang, Y.; Zucker, I.; Choo, Y.; Osuji, C. O.; Elimelech, M., High performance nanofiltration membrane for effective removal of perfluoroalkyl substances at high water recovery. *Environmental Science & technology* 2018, 52, (13), 7279-7288.

# 4. Predicting micropollutant removal by reverse osmosis and nanofiltration membranes: Is machine learning viable?[1]

## 4.1 Introduction

Micropollutants such as pharmaceutical and personal care products (PPCPs), endocrine-disrupting chemicals (EDCs), pesticides, and per- and polyfluoroalkyl substances (PFAS) are anthropogenic chemicals that threaten human and ecological health.[1-4] As current wastewater treatment plants are not designed to remove micropollutants, several micropollutants are able to escape from - conventional wastewater treatment procedures[5] and detected in the aquatic environment or even drinking water.[6] It has been reported that long-term exposure to trace concentrations of micropollutants imposed adverse effects on wildlife and human health.[7, 8] Therefore, there is a critical need for developing highly efficient treatment technologies for micropollutant removal.

Pressure-driven membrane technologies such as reverse osmosis (RO) and nanofiltration (NF) have been widely used in water and wastewater treatment because of their effective removal of contaminants and exceptional energy efficiencies.[9, 10] The performance of NF and RO membranes is regulated by the well-documented permeability-selectivity tradeoff, in which an increase of membrane permeability typically occurs at the expense of membrane selectivity and vice versa.[11-13] Therefore, membranes that possess higher efficiencies for micropollutant removal tend to have lower water fluxes, which correspond to higher energy consumption and economic

---

[1] This chapter has been accepted and published as a research article in the journal of *Environmental Science & Technology* with the following reference:

Jeong, N., Chung, T. H., & Tong, T. (2021). Predicting micropollutant removal by reverse osmosis and nanofiltration membranes: is machine learning viable?. *Environmental Science & Technology*, *55*(16), 11348-11359.

Cost. Selecting adequate membranes that maximize water permeability while maintaining satisfying removal rates of target compounds is of critical importance to achieving more energy- and cost-efficient NF and RO treatment.

Predictive models that enable the estimation of membrane removal of target compounds are valuable for the design and selection of suitable membranes as well as the optimization of membrane processes. Such models bridge the removal efficiencies for micropollutants with both membrane and compound properties, allowing estimation of micropollutant removal by NF and RO prior to experimental investigations. The mechanisms and corresponding variables that govern solute removal by membrane filtration have been explored extensively in the literature.[14-18] By identifying nonlinear correlations between input variables and target labels without governing equations, data-driven approaches have been used to address the complexity of establishing predictive models. Recently, machine learning (ML) has gained increasing popularity in solving multivariable problems in the field of environmental science, including membrane separation and materials development.[19-31]

Although satisfying prediction accuracy has been claimed in the literature,[26, 27] the prospects and limitations of applying ML to predicting membrane performance are still unclear. For example, inappropriate methods of data splitting may cause data leakage. Data leakage indicates the introduction of information about testing data, which should not be available to the training or validation datasets.[32] The data of ML models are typically divided into three groups for training, validation, and testing. Data leakage occurs when the same (or similar) data are randomly split into the training/validation datasets and the testing dataset. Such splitting method has been widely used to train ML models for predicting micropollutant removal by membrane filtration in previous studies.[25-27, 29] In such scenarios, the ML model is tested with the data used for model

training, leading to higher performance than its actual predicting capacity. Thus, proper methods of data splitting should be utilized to avoid data leakage and evaluate the prediction accuracy of ML models objectively. In fact, the true value of establishing predictive models should be to estimate membrane performance under *unknown* scenarios (e.g., for unknown compounds, membranes, or experimental conditions). However, whether ML models possess the capability to make reasonable predictions (e.g., on micropollutant removal rates in our study) for unfamiliar conditions have not been investigated in the literature. Further, a truly intelligent ML model should make predictions according to adequate knowledge on the mechanisms of membrane separation. So far, ML models for predicting membrane performance have been considered as "black box" models. No studies have tested whether the models are able to gain proper mechanistic knowledge on the micropollutant removal by membrane filtration.

In this work, we evaluate the capability of an ML model established using an XGBoost (XGB) model to predict micropollutant removal efficiencies of NF and RO membranes. We first randomly split the collected data into training, validation, and testing groups, in which the experimental conditions of some data are very similar (potential of data leakage). The prediction results were compared with those obtained by multivariable linear regression. Then we evaluated the accuracy of model prediction for specific compounds or membranes, which were completely excluded from the training and validation datasets. By doing so, the ML model was used to predict membrane performance under fully unknown conditions. The corresponding prediction accuracy was compared with that based on random data splitting in order to understand the effect of data leakage. A hybrid NGBoost-XGBoost model was also built to estimate the conditional probability distribution of the predicted values. The 95% confidence interval of the prediction was compared with the observed removal rates to check the reliability of the model predictions. Further, we

investigated whether the developed model understands the mechanisms that govern membrane removal of micropollutants by exploring the contribution of each input variable to model predictions. In addition, the prospects and limitations of using ML models to predict membrane performance are discussed. It is worth mentioning that the main goal of the current study is not to demonstrate or enhance the accuracy of predictive models using ML. Instead, we aim to objectively evaluate whether and how data-driven ML models can be utilized to improve the design and selection of NF and RO membranes for the effective removal of micropollutants.

## 4.2 Methods

### *4.2.1 Data collection*

The micropollutant removal rates of NF and RO membranes were obtained from the literature [14, 25, 33-85] using the reported data or 'Engauge Digitizer', a software designed to extract data points from graphs. We collected 1907 data points of commercial membranes and 61 data points of self-fabricated membranes, resulting in a total of 1968 data points for 231 micropollutants and 49 types of NF and RO membranes. To the best of our knowledge, our study contains the highest number of data points to develop predictive models for membrane removal of micropollutants. Most data were obtained in deionized water containing no or low concentrations of background electrolytes (typically no more than 50 mM NaCl and no divalent ions). Water quality variables such as ionic strength and co-existing ions are not included in the current study because of the insufficient data available in the literature (if the volume of available data is too small, the inclusion of these data could cause bias of the ML model).

Eight input variables were selected based on the interactions between membranes and micropollutants (i.e., steric effects, electrostatic effects, and hydrophobic interaction) as well as the experimental conditions. These variables include total charge (TC), membrane contact angle

(MCA, measured with water), membrane molecular weight cut-off (MWCO), compound size (CS), compound log $K_{ow}$, hydraulic pressure (P), initial concentration of the compound ($C_{in}$), and the measurement time (T). Particularly, TC and CS were calculated by the equations below:

TC = membrane zeta potential × compound charge (4-1)

CS = (compound maximum projection × compound minimum projection)$^{0.5}$ (4-2)

The details on acquiring or calculating the input variables and the correlation matrix of input variables (Figure A1, Appendix A) are described in Supporting Information.

4.2.2 XGBoost model for micropollutant removal predictions

In this study, XGBoost, a popular type of shallow learning, was used for predicting micropollutant removal by NF and RO due to its advantages. Compared to deep learning, XGBoost is less vulnerable to overfitting with a small dataset and requires lower computational demands.[86] Although deep learning models have shown high performance in many multivariable problems,[87-89] our research objective and dataset do not require the extraction of higher-level features from lower-level ones as deep learning models do.

The XGBoost model, a scalable tree boosting system, is widely used for data mining. XGBoost is an improved gradient boosting that converts many weak learners (decision trees) to a strong learner. Figure A2A (Appendix A) shows the structure of gradient boosting trees. Decision nodes (red and blue circles) of each tree are split into two sub-nodes based on the input variables (e.g., whether the MWCO or MCA is higher than certain thresholds) from the training dataset. Each node is divided into child nodes (sub-nodes) or leaf nodes (predicted values) for two possible answers. In Figure A2A, the red circles indicate the decision nodes where the input data are categorized, while the blue circles are the decision nodes that can be used for different input variables. The next decision tree is created to reduce the residual errors from the previous decision

tree. The products of weights and predicted values of each tree are summed to produce the final predictions. Unlike the traditional gradient boosting tree, XGBoost utilizes the second-order Taylor expansion for optimizing the objective function, a novel learning method for sparse data, parallel and distributed computing, and out-of-core computation. The state-of-art performance of XGBoost has been proven in many data science competitions.[90]

Data were divided into training, validation, and testing datasets. For model optimization, one of five groups in 5-fold cross-validation (see a detailed explanation on K-fold cross-validation in Supporting Information) was used for validation, and the other four groups were used for training (Figure A2B). During five times of training and validation with different combinations, the average root mean squared errors (RMSE) from five cases were calculated to find the optimal parameters and hyperparameters. After 23 iterations (3 steps of random explorations and 20 steps of optimization), the best hyperparameters with the lowest RMSE were determined. The list of hyperparameters and the associated ranges used in the optimization is shown in Table A1 (Appendix A). The test dataset was then utilized to evaluate the ML model performance on the predictions. The model was coded in Python with Scikit-Learn and XGBoost packages. All computational work was performed on Google Colab.

*4.2.3 Data splitting methods*

Two different methods of data splitting were applied in order to understand the potential effect of data leakage and investigate the membrane separation mechanisms. First, 1907 data points from commercial membranes were randomly split into 90% of training and validation data (5-fold cross-validation) and 10% of testing data. Figure 4-1A depicts the distribution of randomly split training, validation, and testing data. Data leakage can occur when data points with similar experimental conditions (e.g., data in the red dashed rectangles) are distributed within the training, validation,

and testing datasets. In our study, these data points are typically obtained from the same study and most of their input variables are identical, resulting in similar micropollutant removal rates.

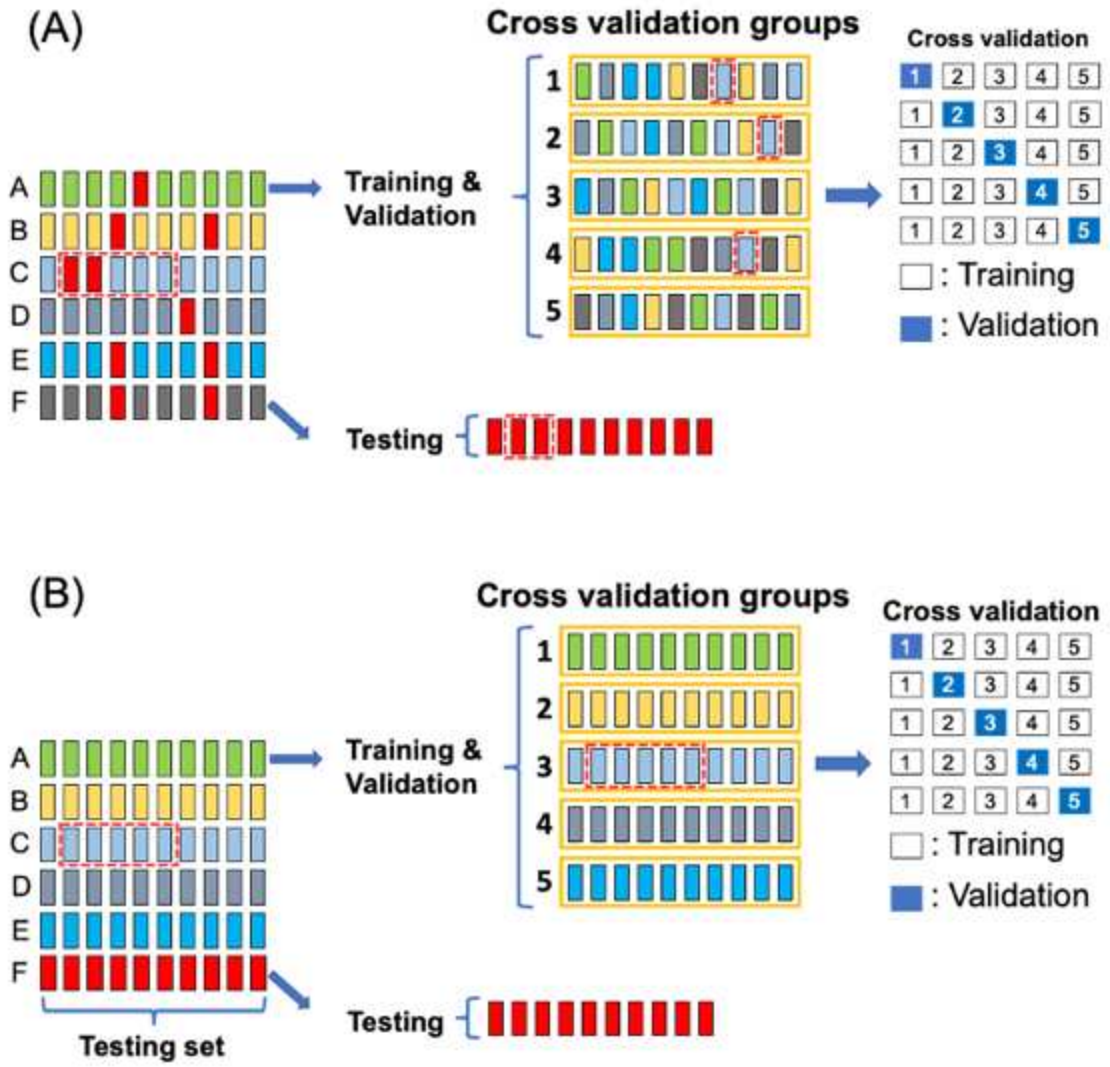


**Figure 4-1.** Data distribution of training, validation, and testing datasets for (A) prediction for known compounds with random splitting, (B) predictions for unknown compounds and self-fabricated membranes with grouping by compounds. The letters represent different compound types. Red blocks are used for testing, and the blocks of all other colors are used for training and validation. The data in the red dashed rectangle indicate those from similar experimental conditions (e.g., from one experiment). It is worth mentioning that those data are included in training, validation, and testing datasets in Figure (A), causing potential data leakage. Such potential of data leakage does not occur in Figure (B).

Second, we isolated each specific compound that had more than 20 data points for testing, with the rest of the data designated for training and validation (5-fold cross-validation). As shown in Figure 4-1B, once compound F was selected and isolated for model testing, the training and validation datasets only consisted of data for other compounds. Such a data splitting method, which is referred to as "grouping by compounds" in this study, enables genuine prediction for scenarios that the model is not familiar with (i.e., unknown compounds in this case) and does not cause data leakage. This is because, in this way, data with similar situations are not included in both the datasets for cross-validation and testing and not influenced by data leakage. We were able to make predictions for 25 compounds with sufficient data points, and 931 testing data points were generated in total. It is worth mentioning that each prediction has different training/validation data that were not used for the testing data, and the results of predicting unknown compounds from 25 predictions were accumulated to calculate the MAE of the model prediction. We also used this method to predict the performance of self-fabricated membranes by using 61 data associated with self-fabricated membranes as the testing dataset and 1907 data of commercial membranes as training and validation datasets.

*4.2.4 NGB-XGB hybrid model for probabilistic estimation of prediction*

The XGBoost model only produces one best guess for a single data point in the regression problems and thus has no knowledge about the probabilistic estimation on predictions. In order to estimate the conditional probability distribution of the predicted values, we employed the NGBoost-XGBoost (NGB-XGB) hybridized model for the prediction. The concept of this hybrid model has been proposed recently by Başağaoğlu et al.[91] for the prediction of evaporation and evapotranspiration in south-central Texas.

Natural Gradient Boost (NGBoost) is a gradient boosting tree that has probabilistic estimation capacity.[92] Different from most gradient boosting trees whose learning object is to minimize the loss function, the NGBoost model estimates the conditional probability distributions of the estimation, allowing users to obtain the confidence intervals of the predictions. The learning object of the NGBoost is to minimize the difference between the conditional probability distribution of the estimations and that of the target labels. The unique objective function of NGBoost allows users to estimate both means and probability distributions of the output values. The NGBoost users can flexibly specify any gradient boosting models, types of distributions, and scoring methods for model training.[92] After optimization, the XGB model was used as the base learner of NGB to build the NGB-XGB hybrid model in this study. It is worth mentioning that the XGB model was used for model predictions, and the NGB-XGB hybrid model was only used for estimating the probabilistic distribution of the prediction. The NGBoost model was built by using Python with NGBoost package.

*4.2.5 Shapley additive explanations (SHAP) for model interpretation*

Shapley additive explanations (SHAP) is based on the cooperative game theory that can explain the contribution of each player (i.e., input variable) to the output values by measuring the importance of each variable. In SHAP, the models are retrained with all possible variable subsets, and the prediction differences of models including and excluding a variable of interest are calculated. A weighted average of the prediction differences from all possible combinations of variables is defined as the SHAP value, which is the average marginal contribution of the variables to the predictions.[93] SHAP values for all the variables used in this study were measured using the 1907 data of commercial membranes in order to reveal the effects of variables on each training data point. This effort enabled us to investigate the contributions of three mechanisms (i.e., size

exclusion, electrostatic repulsion, and adsorption) and operating conditions associated with the membrane performance for micropollutant removal.

## 4.3 Results and Discussion

### *4.3.1 Prediction performance of XGB model with data leakage*

The prediction performance of the ML model using XGB model was first evaluated using random data splitting (potential of data leakage, Figure 4-2B) and compared with that of multilinear regression (MLR, Figure 4-2A). For the MLR model, principal component analysis (PCA) was used to reduce the dimension of the input data (the details of PCA are described in the Supporting Information). We used 90% of the commercial membrane data to establish the MLR model, which was tested with 10% of the commercial membrane data. As presented in Figure 4-2A, the MLR model exhibited low prediction accuracy, with a high MAE of 19.28%. This means that the capability of the MLR model is insufficient to accurately predict membrane removal of micropollutants. The limited prediction accuracy of the MLR model suggests that smarter models, which are able to better understand the correlations of input variables with the rates of micropollutant removal, are needed to improve the prediction performance.

To establish the XGB model, the data obtained from commercial membranes were first randomly split into training and validation datasets (90% of data, 1716 data points in total, 5-fold cross-validation) as well as a testing dataset (10% of data, 191 data points in total). It is noted that the same training and testing data were used for both MLR and XGB models. Compared with the MLR model, the XGB model displayed significantly improved prediction accuracy, with an MAE of only 6.25% (Figure 4-2B). We also performed ten replicate predictions, and the results showed that the prediction accuracy is highly reproducible (Figure A4, Appendix A). Such prediction

accuracy for membrane removal of micropollutants was similar to those of ML models in the literature

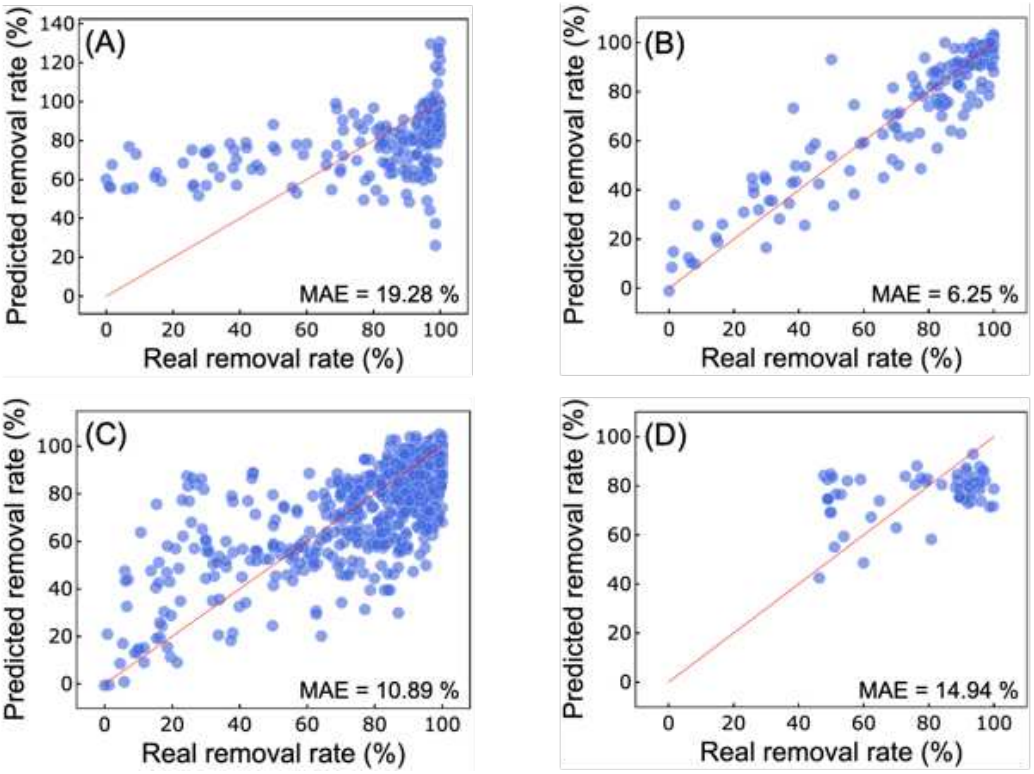


**Figure 4-2.** Predictions on the removal rates of micropollutants with random data splitting (potential data leakage) by (A) the multilinear regression model (MAE: 19.28%) and (B) the XGB model (MAE: 6.25%). Predictions by the XGB model with data grouping by compounds for (C) unknown compounds (MAE: 10.89%) and (D) self-fabricated membranes (MAE: 14.94%). Red lines represent the line where the predicted values and real removal rates are equal.

(Table A2, Appendix A), while our model contained the highest number of data points obtained from diverse experimental conditions.

To understand the impact of the data leakage, we compared the prediction accuracy of the ML model on data with and without leakage. When the training and testing data are associated with the same membrane and micropollutant with only one operating condition different (e.g., pressure, initial compound concentration, or measurement time), we consider that those data have similar experimental conditions with potential leakage issue. The testing data without leakage do not have a similar experimental condition compared to any data in the training dataset. Table A3

(Appendix A) shows the details of model prediction for testing data with and without potential leakage problem.

The results showed that the model predictions of data with and without leakage have MAE of 2.83% and 10.17%, respectively (Figure A5, Appendix A). When the dataset with leakage issue is tested, the model already knows the answers before it makes predictions because it is trained with similar data. Thus, falsely high prediction accuracy is obtained with data leakage. In such cases, the model is not trained objectively, which potentially leads to the low performance on predicting unseen data from unknown conditions.[94] We noticed that the model predictions in the absence and presence of data leakage had different data split ratios. In order to test the effect of data split fraction on model prediction accuracy, different data split ratios (90:10, 80:20, and 70:30) of training/validation data to testing data were used for prediction with data leakage (the data split ratio without data leakage could not be changed because compounds that have more than 20 data points were selected). These data split ratios resulted in similar MAE (difference less than 0.7%, Figure A6 and Table A4, Appendix A). These results clearly demonstrate that the prediction accuracy of the data set with leakage issue was not sensitive to data split ratio and the falsely high prediction accuracy in the presence of data leakage was a valid finding. Thus, alternative approaches of data splitting are required to prevent data leakage and to evaluate the prediction performance of ML models more accurately.

*4.3.2 Performance of the XGB model in predicting micropollutant removal pertaining to unknown compounds and membranes*

To evaluate the prediction capacity of the ML model objectively, we tested the XGB model with a different data splitting method where the leakage issue is avoided. The XGB model was employed to predict the membrane removal rates of unknown compounds. Among the 1907 data

for commercial membranes, the data for one specific compound was set apart as the testing dataset, and the rest of the data were used for training and validation (i.e., grouping by compounds, Figure 4-1B). During the training and optimization of the model, each compound was only included in one of the five cross-validation groups to avoid data leakage between training and validation datasets. Unlike random splitting, grouping by compounds ensures that the data with similar experimental conditions are assigned to the same group (either testing data or one of five cross-validation groups). By doing so, the ML model is not affected by data leakage. The predictions were made for 25 compounds (those compounds are referred to as unknown compounds in this study because the ML model is not trained with data associated with such compounds), each of which had more than 20 data points available. MAE was 10.89% for the prediction of membrane removal for unknown compounds (Figure 4-2C), which was higher than what was obtained via random splitting (Figure 4-2B). When the prediction accuracy was compared among data with different input variables, the absolute error was not a function of any input variable, indicating that no input variable had more weights in contributing to the higher errors (Figure A7, Appendix A). However, a majority of data with low absolute errors (i.e., absolute error <10%) are associated with membranes with relatively small MWCO or compounds with relatively large size (Figure A8, Appendix A), whereas data with high absolute errors (i.e., absolute error > 20%) are mainly related to small compound size or large MWCO. These results indicate that the XGB model generally makes good predictions when size exclusion is the dominant mechanism for micropollutant removal. However, when other mechanisms such as electrostatic effects or adsorption play important roles in regulating membrane separation, the model displayed reduced accuracy. Such limitations of the ML model will be investigated and discussed in the following sections.

For the model prediction of the performance of the self-fabricated membrane, 1907 data points for commercial membranes were used for training and validation, and 61 data points of self-fabricated membranes were used for the testing dataset. In such a scenario, the XGB model was also not constrained by data leakage because the testing data were obtained in different studies from those of training and validation datasets (i.e., avoiding similar input variables). Figure 4-2D shows that the MAE for self-fabricated membrane data was 14.94%, which was also higher than what was shown in Figure 4-2B. As in the case of prediction for the unknown compounds, the absolute errors did not have an explicit relationship with any input variable, as shown in Figure A9 (Appendix A). High absolute errors were observed from the studies that aimed to develop novel membranes for the removal of perfluorobutanesulfonic acid (PFBS, absolute error of 25.53%) and perfluorooctanesulfonic acid (PFOS, absolute error of 17.88%), which are hydrophobic and have small compound sizes (< 0.44 nm). This result was consistent with our above suggestion that the model had low prediction accuracy when mechanisms other than size exclusion contribute to determining membrane removal efficiency. For those micropollutants, electrostatic interaction and adsorption are likely to play important roles in regulating NF removal, which involve multiple variables pertaining to the membrane and compound properties (e.g., charge and hydrophobicity), and the knowledge of the XGBoost model on such mechanisms will be further discussed in this study.

*4.3.3 The probabilistic distribution of predictions using NGB-XGB model*

When ML models were used to predict micropollutant removal efficiencies of membrane filtration in the literature, the predicted values do not contain information about the prediction reliability. As the actual removal rates can be significantly different from the predicted values on a case-by-case basis (even when the average prediction error is low), it is valuable to understand the

reliability of model predictions by checking the confidence interval. Recently, NGBoost model has been developed to estimate the probabilistic distribution of the predictions by ML models.[92] After the optimization, the XGBoost model was used as the base learner for NGBoost to estimate the means and the logarithm of standard deviations of predictions. Normal distribution and continuous ranked probability score (CRPScore) were utilized for the prediction distribution and the scoring rule of the model, respectively.

The data for commercial membranes were split into training/validation datasets (90% of data, 1716 data points) and a testing dataset (10% of data, 191 data points). The same types of compounds were assigned to groups either for training/validation or testing to avoid data leakage. Training and validation datasets were used during the hyperparameter optimization of the XGB model, and the hyperparameters were utilized for the NGB-XGB model. The 95% confidence interval was obtained from the normal distribution and plotted with predicted and actual removal rates (Figure 4-3). The 95% confidence intervals indicate the range where 95% or more predicted values should be. It is worth mentioning that in our study the 95% confidence interval of the NGBoost was estimated without the information about testing data, and thus the percentage of testing data covered by the confidence interval varies depending on the predictive reliability of the model. The narrow range of the 95% prediction interval in Figure 4-3 shows that the NGB-XGB model had high confidence in its predictions. However, only 24% (46 testing data points) of actual removal rates were covered by the 95% confidence interval of prediction. This result, along with what was observed in Figures 4-2C and 4-2D, indicates that the prediction accuracy of the model is limited. It is desirable to have high-performance ML models whose confidence intervals of prediction can cover the majority of actual removal rates, while maintaining narrow ranges of the

confidence interval. We will discuss the reasons that cause such model inaccuracy in the next section.

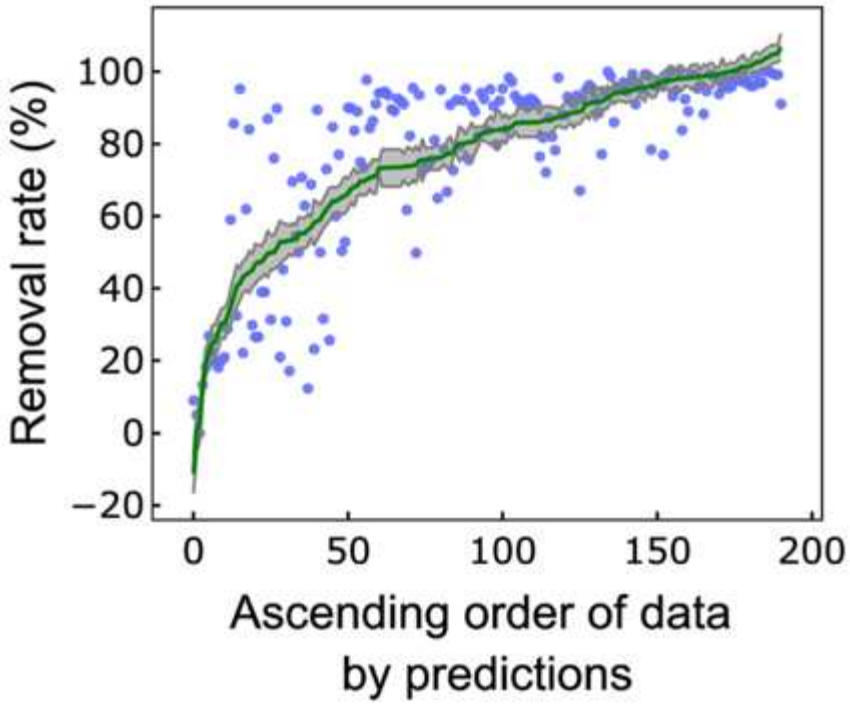


**Figure 4-3.** Predicted (green line) and actual (blue dots) micropollutant removal rates with 95% prediction interval (a grey area) of NGB-XGB model. The x-axis indicates ascending order of data by predictions.

*4.3.4 Does XGBoost model understand the mechanisms of membrane separation?*

In order to understand the effect of each variable on the model prediction, SHAP values of the variables were calculated using the XGB model with all the 1907 data obtained from commercial membranes (Figure 4-4). The SHAP values are the marginal contributions of variables to the model predictions, with a higher absolute score indicating more contribution of the variable to the model prediction. For instance, the SHAP values of 20 and -10 mean that the contributions of specific variables to the micropollutant removal rate are 20% (rejection of the micropollutant) and -10% (passage of the micropollutant), respectively. When the SHAP values of all the variables for a certain datum are summed, it results in the predicted removal rate of that datum.

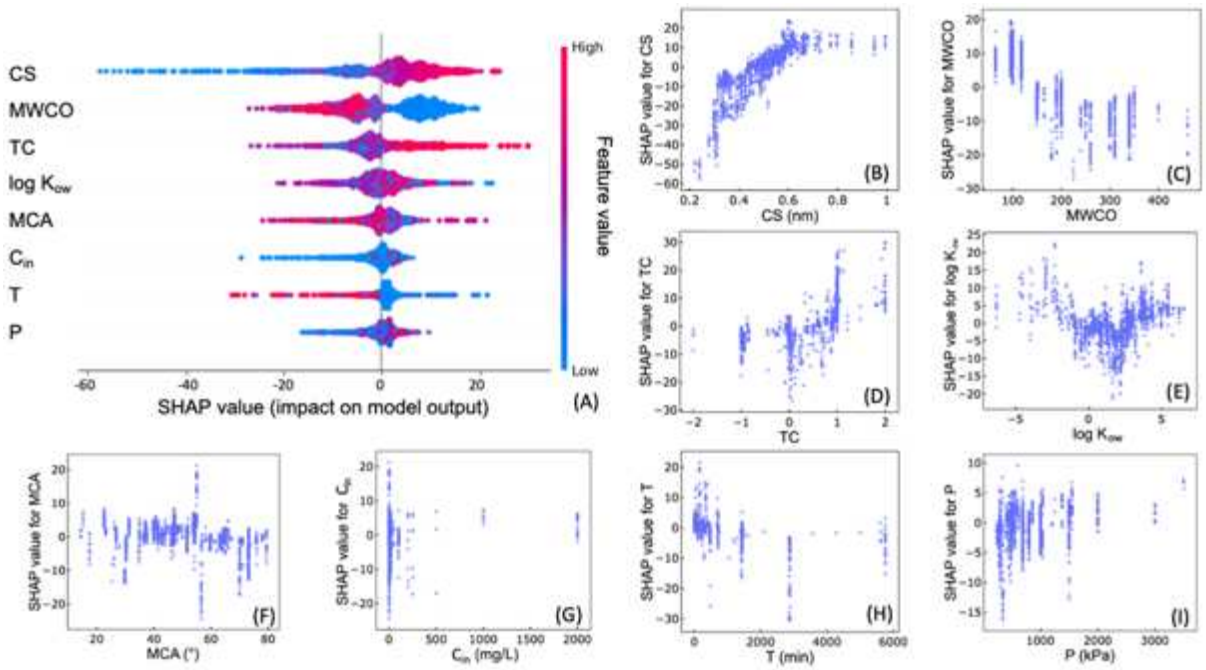


**Figure 4-4.** (A) SHAP summary plot of input variables. The higher absolute SHAP values represent more contribution of the variable to the model predictions. The color of points indicates the magnitude of the variable values. (B-I) The dependence plots of SHAP values as a function of (B) compound size (CS), (C) MWCO, (D) total charge (TC), (E) log $K_{ow}$, (F) membrane contact angle (MCA), (G) initial concentration of compound ($C_{in}$), (H) measurement time (T), and (I) pressure (P).

The SHAP summary plot (Figure 4-4A) displays the contributions of input variables to the prediction of all the data points. The variable values are presented in blue (low value) and red (high value), depending on their magnitude. Taking CS as an example, red and blue colors represent large and small compounds, respectively. Also, the SHAP value is able to indicate the positive or negative contribution of the corresponding variable to micropollutant removal efficiency. Positive and negative SHAP values indicate that considering the variable results in more and less micropollutant removal, respectively. Therefore, the SHAP summary plot shows the direction and distribution relating to the contribution of each variable.[93]

In the SHAP summary plot, the order of input variables on the y-axis indicates the hierarchy of variable importance, according to the average of absolute SHAP value.[93] As shown in Figure 4-

4A, CS and MWCO, both of which are related to size exclusion, are the most important variables for the ML model to predict membrane separation. Total charge has a higher importance score compared to log $K_{ow}$ and MCA, indicating that adsorption (relating to hydrophobic interaction) has a lower contribution to micropollutant removal rates than size exclusion and electrostatic interaction. The operating conditions, including $C_{in}$, T, and P, are placed in the lowest ranks. The above results suggest that the XGB model makes predictions mainly relying on the three common mechanisms of membrane separation (i.e., size exclusion, electrostatic interaction, and adsorption), although operational conditions are taken into consideration. However, the order of variable importance is created based on the average of absolute SHAP values, reflecting the general variable contribution used by the model. Thus, the hierarchy of variable importance is not always the same for all the data. Under certain conditions, for example, operating conditions such as measurement time T (e.g., some red dots with negative SHAP values) can make significant contributions.

In order to further test whether the XGB model genuinely understands the mechanisms of micropollutant removal by NF and RO membranes, we plotted the SHAP values as a function of each corresponding variable (Figures 4-4B to 4-4I). The compound size displays a positive correlation with its SHAP values within a CS range of 0.2 to 0.6 nm (Figure 4-4B). In this range, increasing the size of micropollutant molecules results in higher SHAP values and contributes positively to micropollutant removal (i.e., increased removal efficiency), consistent with the mechanism of size exclusion. When the CS value exceeds 0.6 nm, the SHAP value reaches a plateau of ~15, indicating that there is a negligible change in the contribution of CS to micropollutant removal in the given range. It is worth mentioning that CS contributes as low as -60% (CS of 0.22 nm) to the removal rates. This indicates that the micropollutants are easy to

penetrate across the membranes when the molecular sizes are very small. This is in accordance with the literature that NF and RO have low efficiencies in removing small molecules such as formaldehyde (0.22 nm) and 1,2-ethanediol (0.31 nm).[12, 95] The SHAP values change from negative (contribute to solute passage) to positive (contribute to solute removal) values at CS of ~0.5 nm. This transition point is close to the size of solute (0.48 nm) that corresponds to the smallest MWCO (65 Da) of the investigated membranes, calculated by the relationship between the molecular weight and Stokes radius.[96] Thus, the XGB model seems to recognize the molecular size that starts to result in efficient membrane removal. Contrary to those of CS, the SHAP values of MWCO (Figure 4-4C) decrease in general when the MWCO value increases (i.e., an increase of membrane pore size), which could also be explained by size exclusion. Therefore, the above results indicate that the ML model understands the mechanism of size exclusion correctly via training with the dataset of this study.

The impact of electrostatic interactions on the model prediction is shown in Figure 4-4D. The SHAP values are negative and positive in general when electrostatic attraction (i.e., TC < 0) and repulsion (i.e., TC > 0) exist, respectively. This is reasonable pertaining to the mechanism that electrostatic repulsion enhances micropollutant rejection, whereas electrostatic attraction facilitates micropollutant transport across the membrane. However, no clear trends were observed when TC values were above or below zero. In other words, an increase of electrostatic attraction or repulsion does not alter the contribution of electrostatic interactions to the model prediction. Further, the contributions of TC are from -27% to 5% when TC is 0, even though no electrostatic interactions exist between the membrane surface and micropollutants. Therefore, although the XGB model recognizes the mechanism of electrostatic interactions to some extent, the knowledge

of the model remains rudimentary with our dataset, which might contribute to the inaccuracy of prediction as depicted in Figures 4-2 and 4-3.

The SHAP dependence plot of log $K_{ow}$ (Figure 4-4E) has a V shape, with positive SHAP values (contribute to micropollutant rejection) displayed for highly hydrophilic (log $K_{ow}$ < -1) and highly hydrophobic (log $K_{ow}$ > 3) compounds. Negative SHAP values (contribute to micropollutant passage) are observed for moderately hydrophilic and hydrophobic compounds (-1 < log $K_{ow}$ < 3). The SHAP values of MCA, which is another variable that relates to hydrophobic interaction, have the range of -20 to 20, with no explicit trend shown in Figure 4-4F. It has been reported that hydrophilic surface coating increases the performance of NF membranes in removing EDCs due to the reduced hydrophobic interaction.[97, 98] However, higher membrane hydrophilicity (i.e., lower MCA) does not result in more positive SHAP values in our study. Given that the SHAP values did not display clear relationships with log $K_{ow}$ and MCA, the ML model does not understand the role of hydrophobic interaction, which closely relates to micropollutant adsorption onto membrane surface,[60, 97] in membrane separation. In fact, how hydrophobic interaction or adsorption regulates membrane removal of micropollutants is more sophisticated than size exclusion and electrostatic repulsion. Compared to other interactions, a higher number of variables are involved in regulating micropollutant adsorption. In addition to input variables related to the hydrophobicity of micropollutant molecules and membranes (e.g., log $K_{ow}$ and MCA), membrane pore size and operational conditions impact the adsorption of compounds. Adsorption occurs on the membrane surface as well as pore walls within the active and the support layers, rendering larger pores providing more internal adsorption sites accessible to micropollutants.[99-101] Also, operational conditions of NF and RO including applied hydraulic pressure and cross-flow velocity determine the extent of concentration polarization, which is correlated with the mass of compounds

adsorbed onto the membrane.[102] Further, the degree of adsorption changes until the micropollutant-membrane interactions reach equilibrium.[85] The adsorption typically results in the increase of removal rates until the adsorption sites are saturated; after saturation, there is no more removal by the adsorption but a higher passage of compounds through the membranes, resulting in the reduction of micropollutant removal.[85, 102, 103] Given such complexity of micropollutant adsorption, the ML model lacks appropriate knowledge on how this mechanism regulates micropollutant removal, as reflected by the unexplicit dependence of the SHAP values on log $K_{ow}$ and MCA (Figures 4-4E and 4-4F). This could also contribute to the low prediction accuracy of the ML model shown in Figures 4-2 and 4-3.

In addition, the contributions of operating conditions, including $C_{in}$, T, and P, to the model predictions are shown in Figures 4-4G to 4-4I. There was no trend observed for the dependence of SHAP value on $C_{in}$ (Figure 4-4G). As shown in Figure 4-4H, measurement time T contributes negatively to micropollutant rejection in general. Some data from neutral (TC = 0) and moderately hydrophobic (log $K_{ow}$ = 2.2) compounds at 3000 minutes have much more negative SHAP values compared to others. (Figure A10, Appendix A). This is consistent with the literature[85, 100, 102] that the removal rates decrease as a function of time for neutral micropollutants in the presence of hydrophobic interaction, which results from the reduction of the membrane adsorption sites. The SHAP values for pressure (Figure 4I) display moderate absolute values and variance (absolute SHAP values $\leq 5$). Even though the operating conditions occasionally have high impacts on model predictions, the relevant variables generally have lower contributions (i.e., lower absolute SHAP values) to micropollutant removal than those related to size exclusion, electrical repulsion, and adsorption, which are the main separation mechanisms of NF and RO.

## 4.4 Implications

In this study, we evaluate the prediction capability of the XGBoost model for micropollutant removal by NF and RO membranes objectively and investigate whether this ML model understands the mechanisms of membrane separation. Although the model outperformed multilinear regression in terms of prediction accuracy, we show that data leakage results in falsely high prediction accuracy and that the accuracy of the model (using 1968 data points, more than those used in the literature) is limited. We demonstrate that the ML model possesses an adequate understanding of the role that size exclusion plays in regulating membrane removal of micropollutants. However, its knowledge of electrostatic interactions and adsorption is incomplete, probably leading to the low accuracy of the ML model. The methods applied in this study to understand the mechanistic knowledge of ML models on membrane separation could be potentially used for other ML models or fields to investigate model validity.

The accuracy of the ML model is determined by both the quantity and quality of data. Input variables related to electrostatic interactions and adsorption (or hydrophobic interaction) were included in the model training, but the total dataset had limited experiment conditions that could not fully show these membrane separation mechanisms to the model. Since the ML model is highly influenced by the training and validation data, the data with limited experimental conditions lead to the low accuracy of the model. As shown in Figure A8, the prediction accuracy was low for the small compound size and high MWCO, where adsorption and electrical interactions have more significant impacts on removal rates. Thus, obtaining more data with such molecular and membrane features under diverse conditions could potentially improve the model accuracy. Further, it is of importance to obtain reliable data to achieve ML models with good prediction performance. We encountered challenges in collecting data for this work because sometimes certain input variables related to membrane properties were missing in the publications. For those

data, we had to use the properties (for the same type of membrane) from other literature or not include the data in our dataset. Thus, we encourage the researchers to report more membrane characteristics (MCA, zeta potential, and MWCO) and operational conditions to increase the accessible data for the training of predictive models. Also, some data with the same membrane and compounds show considerably different removal rates. In such cases, it is difficult to distinguish which data are more trustable (Table A5, Appendix A). The considerable differences in micropollutant removal rates might be caused by errors during measurements or reporting the incorrect membranes by the authors. Appropriate screening methods for identifying data that are likely incorrect are useful to improve the data reliability and the model accuracy.[22]

Furthermore, more informative variables can be included to help the ML model learn the mechanisms of membrane separation. Cha et al.[104] demonstrated that the ML model accuracy for the prediction of oxidant exposure and micropollutant abatement during ozonation could be enhanced by using more appropriate variables. As shown in the model explanation by the SHAP values, the mechanism of adsorption and its effect on micropollutant removal are very complex. These interactions are affected by many factors such as functional groups, molecular structures, and charges of surface and compounds.[100, 105, 106] The ML model needs more information to improve its knowledge of this mechanism. In addition to hydrophobic interaction, the supramolecular interactions (e.g., H-bonding and $\pi - \pi$ interactions) between membranes and compounds also contribute to micropollutant adsorption.[41, 101] These interactions are affected by the functional groups of membranes and compounds. Therefore, information about the functionalities of membrane and micropollutants can be used to teach the model to better understand the mechanism of adsorption. Also, applying integrated variables might be useful to reduce the complexity of the adsorption mechanism. For example, operating conditions such as

pressure, initial concentration of compound, and cross-flow rate determine the extent of adsorption.[102] These variables might be integrated as a single variable (e.g., the concentration of micropollutant at the membrane surface), using the concept of concentration polarization. In our future study, we will investigate the selection of more appropriate variables in order to improve the mechanistic knowledge and accuracy of the ML model for the prediction of micropollutant removal by membrane technologies.

## References


1. Corcoran, J.; Winter, M. J.; Tyler, C. R., Pharmaceuticals in the aquatic environment: A critical review of the evidence for health effects in fish. *Critical Reviews in Toxicology* 2010, 40, (4), 287-304.
2. Colborn, T.; Saal, F. S. v.; Soto, A. M., Developmental effects of endocrine-disrupting chemicals in wildlife and humans. *Environmental Health Perspectives* 1993, 101, (5), 378-384.
3. Kim, K.; Kabir, E.; Jahan, S. A., Exposure to pesticides and the associated human health effects. *Science of The Total Environment* 2017, 575, 525-535.
4. Stanifer, J. W.; Stapleton, H. M.; Souma, T.; Wittmer, A.; Zhao, X.; Boulware, L. E., Perfluorinated chemicals as emerging environmental threats to kidney health: A scoping review. *Clinical Journal of the American Society of Nephrology* 2018, 13, (10), 1479-1492.
5. Rogowska, J.; Cieszynska-Semenowicz, M.; Ratajczyk, W.; Wolska, L., Micropollutants in treated wastewater. *Ambio* 2020, 49, (2), 487-503.
6. Zafeiraki, E.; Costopoulou, D.; Vassiliadou, I.; Leondiadis, L.; Dassenakis, E.; Traag, W.; Hoogenboom, R. L.; van Leeuwen, S. P., Determination of perfluoroalkylated substances (PFASs) in drinking water from the Netherlands and Greece. *Food additives & contaminants: part A* **2015,** *32*, (12), 2048-2057.
7. Alavanja, M. C. R.; Hoppin, J. A.; Kamel, F., Health effects of chronic pesticide exposure: cancer and neurotoxicity. *Annual Review of Public Health* 2004, 25, (1), 155-197.
8. Galus, M.; Kirischian, N.; Higgins, S.; Purdy, J.; Chow, J.; Rangaranjan, S.; Li, H.; Metcalfe, C.; Wilson, J. Y., Chronic, low concentration exposure to pharmaceuticals impacts multiple organ systems in zebrafish. *Aquatic Toxicology* 2013, 132-133, 200-211.
9. Van Der Bruggen, B.; Vandecasteele, C.; Van Gestel, T.; Doyen, W.; Leysen, R., A review of pressure-driven membrane processes in wastewater treatment and drinking water production. *Environmental progress* 2003, 22, (1), 46-56.
10. Van der Bruggen, B.; Everaert, K.; Wilms, D.; Vandecasteele, C., Application of nanofiltration for removal of pesticides, nitrate and hardness from ground water: Rejection properties and economic evaluation. *Journal of Membrane Science* 2001, 193, (2), 239-248.
11. 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), eaab0530.
12. Werber, J. R.; Deshmukh, A.; Elimelech, M., The critical need for increased selectivity, not increased water permeability, for desalination membranes. *Environmental Science & Technology Letters* 2016, 3, (4), 112-120.
13. Yang, Z.; Guo, H.; Tang, C. Y., The upper bound of thin-film composite (TFC) polyamide membranes for desalination. *Journal of Membrane Science* 2019, 590, 117297.
14. Van der Bruggen, B.; Schaep, J.; Wilms, D.; Vandecasteele, C., Influence of molecular size, polarity and charge on the retention of organic molecules by nanofiltration. *Journal of Membrane Science* 1999, 156, (1), 29-41.
15. Wijmans, J. G.; Baker, R. W., The solution-diffusion model: a review. *Journal of Membrane Science* 1995, 107, (1), 1-21.
16. Deen, W. M., Hindered transport of large molecules in liquid-filled pores. *AICHE Journal* 1987, 33, (9), 1409-1425.

17. Chaabane, T.; Taha, S.; Taleb Ahmed, M.; Maachi, R.; Dorange, G., Coupled model of film theory and the Nernst–Planck equation in nanofiltration. *Desalination* 2007, 206, (1), 424-432.
18. Wang, X.; Tsuru, T.; Nakao, S.; Kimura, S., The electrostatic and steric-hindrance model for the transport of charged solutes through nanofiltration membranes. *Journal of Membrane Science* 1997, 135, (1), 19-32.
19. Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S. S. R. K. C.; Lian, C.; Kwon, H.; Wong, B. M., A machine learning approach for predicting defluorination of per- and polyfluoroalkyl substances (PFAS) for their efficient treatment and removal. *Environmental science & technology letters* 2019, 6, (10), 624-629.
20. Li, L.; Rong, S.; Wang, R.; Yu, S., Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical Engineering Journal* 2021, 405, 126673.
21. Cheng, W.; Ng, C. A., Using machine learning to classify bioactivity for 3486 per- and polyfluoroalkyl substances (PFASs) from the OECD List. *Environmental Science & Technology* 2019, 53, (23), 13970-13980.
22. Gharagheizi, F.; Tang, D.; Sholl, D. S., Selecting adsorbents to separate diverse near-azeotropic chemicals. The Journal of Physical Chemistry C 2020, 124, (6), 3664-3670.
23. Richard Bowen, W.; Jones, M. G.; Welfoot, J. S.; Yousef, H. N. S., Predicting salt rejections at nanofiltration membranes using artificial neural networks. *Desalination* 2000, 129, (2), 147-162.
24. Libotean, D.; Giralt, J.; Rallo, R.; Cohen, Y.; Giralt, F.; Ridgway, H. F.; Rodriguez, G.; Phipps, D., Organic compounds passage through RO membranes. *Journal of Membrane Science* 2008, 313, (1), 23-43.
25. 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. *Journal of Membrane Science* 2009, 342, (1), 251-262.
26. Khaouane, L.; Ammi, Y.; Hanini, S., Modeling the retention of organic compounds by nanofiltration and reverse osmosis membranes using bootstrap aggregated neural networks. *Arabian Journal for Science and Engineering* 2017, 42, (4), 1443-1453.
27. Ammi, Y.; Khaouane, L.; Hanini, S., Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. *Korean Journal of Chemical Engineering* 2015, 32, (11), 2300-2310.
28. Hu, J.; Kim, C.; Halasz, P.; Kim, J. F.; Kim, J.; Szekely, G., Artificial intelligence for performance prediction of organic solvent nanofiltration membranes. *Journal of Membrane Science* 2021, 619, 118513.
29. Lee, S.; Kim, J., Prediction of nanofiltration and reverse-osmosis-membrane rejection of organic compounds using random forest model. *Journal of Environmental Engineering* 2020, 146, (11), 04020127.
30. Hardian, R.; Liang, Z.; Zhang, X.; Szekely, G., Artificial intelligence: the silver bullet for sustainable materials development. *Green Chemistry* 2020, 22, (21), 7521-7528.
31. Shi, Z.; Yang, W.; Deng, X.; Cai, C.; Yan, Y.; Liang, H.; Liu, Z.; Qiao, Z., Machine-learning-assisted high-throughput computational screening of high performance metal–organic frameworks. *Molecular Systems Design & Engineering* 2020, 5, (4), 725-742.

32. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O., Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 2012, 6, (4), 1-21.
33. Yangali-Quintanilla, V.; Kim, T. U.; Kennedy, M.; Amy, G., Modeling of RO/NF membrane rejections of PhACs and organic compounds: A statistical analysis. *Drinking Water Engineering and Science* 2008, 1, (1), 7-15.
34. Comerton, A. M.; Andrews, R. C.; Bagley, D. M.; Hao, C., The rejection of endocrine disrupting and pharmaceutically active compounds by NF and RO membranes as a function of compound and water matrix properties. *Journal of Membrane Science* 2008, 313, (1), 323-335.
35. Boussu, K.; Vandecasteele, C.; Van der Bruggen, B., Relation between membrane characteristics and performance in nanofiltration. *Journal of Membrane Science* 2008, 310, (1), 51-65.
36. Kim, T.; Amy, G.; Drewes, J. E., Rejection of trace organic compounds by high-pressure membranes. *Water Science and Technology* 2005, 51, (6-7), 335-344.
37. Kim, T.; Drewes, J. E.; Scott Summers, R.; Amy, G. L., Solute transport model for trace organic neutral and charged compounds through nanofiltration and reverse osmosis membranes. *Water Research* 2007, 41, (17), 3977-88.
38. Yoon, Y.; Lueptow, R. M., Removal of organic contaminants by RO and NF membranes. *Journal of Membrane Science* 2005, 261, (1), 76-86.
39. Bellona, C.; Drewes, J. E., The role of membrane surface charge and solute physico-chemical properties in the rejection of organic acids by NF membranes. *Journal of Membrane Science* 2005, 249, (1), 227-234.
40. Verliefde, A.; Van Vliet, N.; Amy, G.; Van der Bruggen, B.; van Dijk, J. C., A semi-quantitative method for prediction of the rejection of uncharged organic micropollutants with nanofiltration. *Water Practice and Technology* 2006, 1, (4), wpt2006084.
41. Dolar, D.; Drašinac, N.; Košutić, K.; Škorić, I.; Ašperger, D., Adsorption of hydrophilic and hydrophobic pharmaceuticals on RO/NF membranes: Identification of interactions using FTIR. *Journal of Applied Polymer Science* 2017, 134, (5), 44426.
42. Xu, R.; Zhou, M.; Wang, H.; Wang, X.; Wen, X., Influences of temperature on the retention of PPCPs by nanofiltration membranes: Experiments and modeling assessment. *Journal of Membrane Science* 2020, 599, 117817.
43. Kimura, K.; Amy, G.; Drewes, J. E.; Heberer, T.; Kim, T.-U.; Watanabe, Y., Rejection of organic micropollutants (disinfection by-products, endocrine disrupting compounds, and pharmaceutically active compounds) by NF/RO membranes. *Journal of Membrane Science* 2003, 227, (1), 113-121.
44. Dolar, D.; Vuković, A.; Ašperger, D.; Košutić, K., Effect of water matrices on removal of veterinary pharmaceuticals by nanofiltration and reverse osmosis membranes. *Journal of Environmental Sciences* 2011, 23, (8), 1299-1307.
45. Lin, Y.; Chiou, J.; Lee, C., Effect of silica fouling on the removal of pharmaceuticals and personal care products by nanofiltration and reverse osmosis membranes. *Journal of Hazardous Materials* 2014, 277, 102-109.
46. Yang, L.; She, Q.; Wan, M. P.; Wang, R.; Chang, V. W.; Tang, C. Y., Removal of haloacetic acids from swimming pool water by reverse osmosis and nanofiltration. *Water Research* 2017, 116, 116-125.

47. Benitez, F. J.; Acero, J. L.; Real, F. J.; Garcia, C., Removal of phenyl-urea herbicides in ultrapure water by ultrafiltration and nanofiltration processes. *Water Research* 2009, 43, (2), 267-276.
48. Ozaki, H.; Li, H., Rejection of organic compounds by ultra-low pressure reverse osmosis membrane. *Water Research* 2002, 36, (1), 123-130.
49. Azaïs, A.; Mendret, J.; Gassara, S.; Petit, E.; Deratani, A.; Brosillon, S., Nanofiltration for wastewater reuse: Counteractive effects of fouling and matrice on the rejection of pharmaceutical active compounds. *Separation and Purification Technology* 2014, 133, 313-327.
50. Ozaki, H.; Ikejima, N.; Shimizu, Y.; Fukami, K.; Taniguchi, S.; Takanami, R.; Giri, R. R.; Matsui, S., Rejection of pharmaceuticals and personal care products (PPCPs) and endocrine disrupting chemicals (EDCs) by low pressure reverse osmosis membranes. *Water Science and Technology* 2008, 58, (1), 73-81.
51. Guo, H.; Yao, Z.; Yang, Z.; Ma, X.; Wang, J.; Tang, C. Y., A one-step rapid assembly of thin film coating using green coordination complexes for enhanced removal of trace organic contaminants by membranes. *Environmental Science & Technology* 2017, 51, (21), 12638-12643.
52. Werber, J. R.; Porter, C. J.; Elimelech, M., A path to ultraselectivity: Support layer properties to maximize performance of biomimetic desalination membranes. *Environmental Science & Technology* 2018, 52, (18), 10737-10747.
53. Lin, Y.; Lee, C., Elucidating the rejection mechanisms of PPCPs by nanofiltration and reverse osmosis membranes. *Industrial & Engineering Chemistry Research* 2014, 53, (16), 6798-6806.
54. Yangali-Quintanilla, V.; Maeng, S. K.; Fujioka, T.; Kennedy, M.; Amy, G., Proposing nanofiltration as acceptable barrier for organic contaminants in water reuse. *Journal of Membrane Science* 2010, 362, (1), 334-345.
55. Nghiem, L. D.; Schäfer, A. I.; Elimelech, M., Pharmaceutical retention mechanisms by nanofiltration membranes. *Environmental Science & Technology* 2005, 39, (19), 7698-7705.
56. Nikbakht Fini, M.; Madsen, H. T.; Muff, J., The effect of water matrix, feed concentration and recovery on the rejection of pesticides using NF/RO membranes in water treatment. *Separation and Purification Technology* 2019, 215, 521-527.
57. Comerton, A. M.; Andrews, R. C.; Bagley, D. M., The influence of natural organic matter and cations on the rejection of endocrine disrupting and pharmaceutically active compounds by nanofiltration. *Water Research* 2009, 43, (3), 613-22.
58. Zhu, L., Rejection of organic micropollutants by clean and fouled nanofiltration membranes. *Journal of Chemistry* 2015, 2015, 9.
59. Zazouli, M. A.; Susanto, H.; Nasseri, S.; Ulbricht, M., Influences of solution chemistry and polymeric natural organic matter on the removal of aquatic pharmaceutical residuals by nanofiltration. *Water Research* 2009, 43, (13), 3270-3280.
60. Guo, H.; Deng, Y.; Yao, Z.; Yang, Z.; Wang, J.; Lin, C.; Zhang, T.; Zhu, B.; Tang, C. Y., A highly selective surface coating for enhanced membrane rejection of endocrine disrupting compounds: Mechanistic insights and implications. *Water Research* 2017, 121, 197-203.
61. Toure, H.; Anwar Sadmani, A. H. M., Nanofiltration of perfluorooctanoic acid and perfluorooctane sulfonic acid as a function of water matrix properties. *Water Supply* 2019, 19, (8), 2199-2205.
62. Dolar, D.; Kosutic, K.; Asperger, D.; Babic, S., Removal of glucocorticosteroids and anesthetics from water with RO/NF membranes. *Chemical and Biochemical Engineering Quarterly* 2013, 27, 1-6.

63. Kimura, K.; Toshima, S.; Amy, G.; Watanabe, Y., Rejection of neutral endocrine disrupting compounds (EDCs) and pharmaceutical active compounds (PhACs) by RO membranes. *Journal of Membrane Science* 2004, 245, (1), 71-78.
64. Steinle-Darling, E.; Zedda, M.; Plumlee, M. H.; Ridgway, H. F.; Reinhard, M., Evaluating the impacts of membrane type, coating, fouling, chemical properties and water chemistry on reverse osmosis rejection of seven nitrosoalklyamines, including NDMA. *Water Research* 2007, 41, (17), 3959-3967.
65. Heo, J.; Boateng, L. K.; Flora, J. R. V.; Lee, H.; Her, N.; Park, Y.-G.; Yoon, Y., Comparison of flux behavior and synthetic organic compound removal by forward osmosis and reverse osmosis membranes. *Journal of Membrane Science* 2013, 443, 69-82.
66. Zhao, Y.; Kong, F.; Wang, Z.; Yang, H.; Wang, X.; Xie, Y. F.; Waite, T. D., Role of membrane and compound properties in affecting the rejection of pharmaceuticals by different RO/NF membranes. *Frontiers of Environmental Science & Engineering* 2017, 11, (6), 20.
67. Zeng, C.; Tanaka, S.; Suzuki, Y.; Fujii, S., Impact of feed water pH and membrane material on nanofiltration of perfluorohexanoic acid in aqueous solution. *Chemosphere* 2017, 183, 599-604.
68. Verliefde, A. R. D.; Cornelissen, E. R.; Heijman, S. G. J.; Petrinic, I.; Luxbacher, T.; Amy, G. L.; Van der Bruggen, B.; van Dijk, J. C., Influence of membrane fouling by (pretreated) surface water on rejection of pharmaceutically active compounds (PhACs) by nanofiltration membranes. *Journal of Membrane Science* 2009, 330, (1), 90-103.
69. Plakas, K. V.; Karabelas, A. J., A systematic study on triazine retention by fouled with humic substances NF/ULPRO membranes. *Separation and Purification Technology* 2011, 80, (2), 246-261.
70. Steinle-Darling, E.; Litwiller, E.; Reinhard, M., Effects of sorption on the rejection of trace organic contaminants during nanofiltration. *Environmental Science & Technology* 2010, 44, (7), 2592-2598.
71. Xu, R.; Zhang, P.; Wang, Q.; Wang, X.; Yu, K.; Xue, T.; Wen, X., Influences of multi influent matrices on the retention of PPCPs by nanofiltration membranes. *Separation and Purification Technology* 2019, 212, 299-306.
72. Mahlangu, T. O.; Msagati, T. A. M.; Hoek, E. M. V.; Verliefde, A. R. D.; Mamba, B. B., Rejection of pharmaceuticals by nanofiltration (NF) membranes: Effect of fouling on rejection behaviour. *Physics and Chemistry of the Earth, Parts A/B/C* 2014, 76-78, 28-34.
73. Soriano, Á.; Gorri, D.; Urtiaga, A., Selection of high flux membrane for the effective removal of short-chain perfluorocarboxylic acids. *Industrial & Engineering Chemistry Research* 2019, 58, (8), 3329-3338.
74. Tang, C. Y.; Fu, Q. S.; Criddle, C. S.; Leckie, J. O., Effect of flux (transmembrane pressure) and membrane properties on fouling and rejection of reverse osmosis and nanofiltration membranes treating perfluorooctane sulfonate containing wastewater. *Environmental Science & Technology* 2007, 41, (6), 2008-2014.
75. Licona, K. P. M.; Geaquinto, L. R. d. O.; Nicolini, J. V.; Figueiredo, N. G.; Chiapetta, S. C.; Habert, A. C.; Yokoyama, L., Assessing potential of nanofiltration and reverse osmosis for removal of toxic pharmaceuticals from water. *Journal of Water Process Engineering* 2018, 25, 195-204.
76. de Souza, D. I.; Dottein, E. M.; Giacobbo, A.; Siqueira Rodrigues, M. A.; de Pinho, M. N.; Bernardes, A. M., Nanofiltration for the removal of norfloxacin from pharmaceutical effluent. *Journal of Environmental Chemical Engineering* 2018, 6, (5), 6147-6153.

77. Zhao, C.; Zhang, J.; He, G.; Wang, T.; Hou, D.; Luan, Z., Perfluorooctane sulfonate removal by nanofiltration membrane the role of calcium ions. *Chemical Engineering Journal* 2013, 233, 224-232.
78. Appleman, T. D.; Dickenson, E. R. V.; Bellona, C.; Higgins, C. P., Nanofiltration and granular activated carbon treatment of perfluoroalkyl acids. *Journal of Hazardous Materials* 2013, 260, 740-746.
79. Puspasari, T.; Pradeep, N.; Peinemann, K.-V., Crosslinked cellulose thin film composite nanofiltration membranes with zero salt rejection. *Journal of Membrane Science* 2015, 491, 132-137.
80. Zhao, Y.; Wang, X.; Yang, H.; Xie, Y. F., Effects of organic fouling and cleaning on the retention of pharmaceutically active compounds by ceramic nanofiltration membranes. *Journal of Membrane Science* 2018, 563, 734-742.
81. He, B.; Peng, H.; Chen, Y.; Zhao, Q., High performance polyamide nanofiltration membranes enabled by surface modification of imidazolium ionic liquid. *Journal of Membrane Science* 2020, 608, 118202.
82. Sun, S. P.; Hatton, T. A.; Chan, S. Y.; Chung, T., Novel thin-film composite nanofiltration hollow fiber membranes with double repulsion for effective removal of emerging organic matters from water. *Journal of Membrane Science* 2012, 401-402, 152-162.
83. Wang, J.; Wang, L.; Xu, C.; Zhi, R.; Miao, R.; Liang, T.; Yue, X.; Lv, Y.; Liu, T., Perfluorooctane sulfonate and perfluorobutane sulfonate removal from water by nanofiltration membrane: The roles of solute concentration, ionic strength, and macromolecular organic foulants. *Chemical Engineering Journal* 2018, 332, 787-797.
84. Weng, X.; Ji, Y.; Ma, R.; Zhao, F.; An, Q.; Gao, C., Superhydrophilic and antibacterial zwitterionic polyamide nanofiltration membranes for antibiotics separation. *Journal of Membrane Science* 2016, 510, 122-130.
85. Kimura, K.; Amy, G.; Drewes, J.; Watanabe, Y., Adsorption of hydrophobic compounds onto NF/RO membranes: An artifact leading to overestimation of rejection. *Journal of Membrane Science* 2003, 221, (1), 89-101.
86. Pasupa, K.; Sunhem, W. In A comparison between shallow and deep architecture classifiers on small dataset, 2016 8th *International Conference on Information Technology and Electrical Engineering (ICITEE)*, 5-6 Oct. 2016, pp 1-6.
87. Krizhevsky, A.; Sutskever, I.; Hinton, G. E., ImageNet classification with deep convolutional neural networks. Commun. ACM 2017, 60, (6), 84–90.
88. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; Hassabis, D., Mastering the game of Go without human knowledge. *Nature* 2017, 550, (7676), 354-359.
89. Abiodun, O. I.; Jantan, A.; Omolara, A. E.; Dada, K. V.; Umar, A. M.; Linus, O. U.; Arshad, H.; Kazaure, A. A.; Gana, U.; Kiru, M. U., Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access* 2019, 7, 158820-158846.
90. Chen, T.; Guestrin, C. In XGBoost: A scalable tree boosting system, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016; 2016; pp 785-794.
91. Başağaoğlu, H.; Chakraborty, D.; Winterle, J., Reliable evapotranspiration predictions with a Probabilistic Machine Learning Framework. *Water* 2021, 13, (4), 557.

92. Duan, T.; Anand, A.; Ding, D. Y.; Thai, K. K.; Basu, S.; Ng, A.; Schuler, A. In Ngboost: Natural gradient boosting for probabilistic prediction, *International Conference on Machine Learning*, 2020; PMLR: 2020; pp 2690-2700.
93. Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2020, 2, (1), 56-67.
94. Chollet, F., Deep Learning with Python. Manning Publications: New York, USA, 2017; p 68–100.
95. Košutić, K.; Kunst, B., Removal of organics from aqueous solutions by commercial RO and NF membranes of characterized porosities. *Desalination* 2002, 142, (1), 47-56.
96. Bowen, W. R.; Mohammad, A. W., Characterization and prediction of nanofiltration membrane performance—A general Assessment. *Chemical Engineering Research and Design* 1998, 76, (8), 885-893.
97. Guo, H.; Deng, Y.; Tao, Z.; Yao, Z.; Wang, J.; Lin, C.; Zhang, T.; Zhu, B.; Tang, C. Y., Does hydrophilic polydopamine coating enhance membrane rejection of hydrophobic endocrine-disrupting compounds? *Environmental Science & Technology Letters* 2016, 3, (9), 332-338.
98. Dai, R.; Guo, H.; Tang, C. Y.; Chen, M.; Li, J.; Wang, Z., Hydrophilic selective nanochannels created by metal organic frameworks in nanofiltration membranes enhance rejection of hydrophobic endocrine-disrupting compounds. *Environmental Science and Technology* 2019, 53, (23), 13776-13783.
99. Comerton, A. M.; Andrews, R. C.; Bagley, D. M.; Yang, P., Membrane adsorption of endocrine disrupting compounds and pharmaceutically active compounds. *Journal of Membrane Science* 2007, 303, (1), 267-277.
100. Semião, A. J. C.; Schäfer, A. I., Removal of adsorbing estrogenic micropollutants by nanofiltration membranes. Part A—Experimental evidence. *Journal of Membrane Science* 2013, 431, 244-256.
101. Schäfer, A. I.; Akanyeti, I.; Semião, A. J., Micropollutant sorption to membrane polymers: a review of mechanisms for estrogens. *Advances in Colloid Interface Science* 2011, 164, (1-2), 100-17.
102. Semião, A. J. C.; Schäfer, A. I., Estrogenic micropollutant adsorption dynamics onto nanofiltration membranes. *Journal of Membrane Science* 2011, 381, (1), 132-141.
103. McCallum, E. A.; Hyung, H.; Do, T. A.; Huang, C.; Kim, J., Adsorption, desorption, and steady-state removal of 17β-estradiol by nanofiltration membranes. *Journal of Membrane Science* 2008, 319, (1), 38-43.
104. Cha, D.; Park, S.; Kim, M. S.; Kim, T.; Hong, S. W.; Cho, K. H.; Lee, C., Prediction of oxidant exposures and micropollutant abatement during ozonation using a machine learning method. *Environmental Science & Technology* 2021, 55, (1), 709-718.
105. Su, Y.; Rao, U.; Khor, C. M.; Jensen, M. G.; Teesch, L. M.; Wong, B. M.; Cwiertny, D. M.; Jassby, D., Potential-driven electron transfer lowers the dissociation energy of the C–F bond and facilitates reductive defluorination of perfluorooctane sulfonate (PFOS). *ACS Applied Materials & Interfaces* 2019, 11, (37), 33913-33922.
106. Rao, U.; Su, Y.; Khor, C. M.; Jung, B.; Ma, S.; Cwiertny, D. M.; Wong, B. M.; Jassby, D., Structural dependence of reductive defluorination of linear PFAS compounds in a UV/Electrochemical System. *Environmental Science & Technology* 2020, 54, (17), 10668-10677.

# 5. Exploring the knowledge attained by machine learning on ion transport across polyamide membranes using explainable artificial intelligence[2]

## 5.1 Introduction

Reverse osmosis (RO) and nanofiltration (NF) are pressure-driven membrane processes that have the capabilities of separating a variety of solutes from water.[1] NF and RO membranes are commonly fabricated via interfacial polymerization between an amine monomer and an acyl chloride, which forms a very thin polyamide active layer on a porous substrate.[2] NF and RO have been applied to the desalination of seawater and brackish water as well as the purification of drinking water and reclamation of municipal wastewater, contributing to freshwater augmentation and resource recovery.[3-7] The performance of NF and RO membranes, including water permeability and the selectivity between different species, are highly influenced by the active layer properties.[8-11] Since the advent of thin-film composite (TFC) polyamide membrane, the energy consumption of NF and RO has been significantly reduced; therefore, further improvement of membrane permeability may result in a limited benefit to energy efficiency.[12-14] Recently, there have been more demands on achieving high membrane selectivity that enables the separation of valuable or undesirable solutes with high precision.[13, 15] The design and fabrication of highly selective membranes for fit-for-purpose applications such as irrigation, wastewater reuse, and resource recovery[3, 13] require accurate predictions and an in-depth understanding of molecular

[2] This chapter was submitted as a manuscript under review as

Jeong, N., Epsztein, R., Wang, R., Park, S., Lin, S., & Tong, T. (2023). Exploring the knowledge attained by machine learning on ion transport across polyamide membranes using explainable artificial intelligence. *Environmental Science & Technology*.

transport across membranes with pore sizes of nanometer or angstrom scale.

Ion transport across membranes has been investigated and predicted by using theoretical models such as the Donnan-steric pore model with dielectric exclusion (DSPM-DE).[16, 17] Based on the extended Nernst-Planck equation with the combination of steric, dielectric, and Donnan exclusion,[18] DSPM-DE has demonstrated its capability of predicting salt rejection.[16, 17] To utilize this model, however, a variety of membrane properties (e.g., effective membrane pore size and thickness, volumetric charge density, and dielectric constant within the membrane matrix) need to be calculated experimentally.[19] Even a small uncertainty on one property may lead to high uncertainties on other parameters, potentially resulting in an inappropriate interpretation of membrane separation mechanisms.[20] Although theoretical models have provided valuable knowledge on solute transport in membrane pores, the utilization of these models is limited due to the complexity of obtaining unknown parameters (e.g., multiple membrane properties) experimentally.

Machine learning (ML) algorithms have been recently applied to the field of membrane separation.[21-28] Such powerful tools, which can handle complex, multivariable problems and reveal the relationships among diverse variables, have successfully predicted membrane performance (e.g., water permeability and solute rejection) and identified key features for improved membrane design.[21-24] In this regard, recent studies using ML for membrane design have demonstrated that ML models are able to extract important features of polymers and optimum conditions for membrane fabrication, which were used to break the current upper bounds of membrane perm-selectivity.[22, 29] This approach has the potential to not only leverage experimental data obtained from a variety of conditions (e.g., different membrane materials, solutes, and operational parameters), but also investigate diverse governing factors of membrane separation

comprehensively. However, the capabilities of ML with a limited number of experimental data to properly learn fundamental principles of membrane science are essential to the model reliability but have not received sufficient attention. So far, it is still unknown whether the performance predictions made by ML models trained with data available from the literature are based on appropriate intelligence that captures membrane separation mechanisms. Understanding the knowledge behind ML model predictions on solute transport, therefore, is of vital importance to validating model reliability and facilitating the applications of ML to membrane separation.

Explainable artificial intelligence (XAI), which is a tool that unveils the complicating decision-making process of ML, has been employed in several fields, such as medical informatics and polymer synthesis, to understand the determining factors of the model predictions.[22, 30, 31] XAI assists users in judging the reasoning of the predictions and produces insight on whether the ML model is trustworthy, resulting in better reliability of the model and its predictions.[32, 33] Recently, we applied XAI to probe the knowledge obtained by ML model on organic micropollutant transport in NF and RO membranes.[23] We discovered that the ML model has a proper knowledge of size exclusion as a key membrane separation mechanism, but its understanding of electrostatic interactions and adsorption is incomplete.[23] Compared to the transport of organic solutes across membranes, the transport of inorganic salts is even more complicated, with the behaviors and mechanisms of cation and anion transport potentially different.[34, 35] To the best of our knowledge, a systematic investigation of the knowledge gained by ML on ion transport across membranes has not been performed so far. As Zhong et al. stated,[36] it is of great importance to investigate whether the predictions of ML models are consistent with the domain science; such investigation is important yet commonly neglected in environmental science and engineering. Applying XAI to investigating the intelligence of ML related to ion transport will unveil whether ML models are

able to correctly learn complex mechanisms of membrane separation that underlie the model predictions.

In this study, we utilize XAI to predict the performance and explore the knowledge attained by ML associated with ion transport across NF and RO membranes. We collected 1,585 data pertaining to the rejections of 463 cations and 478 anions in mixed salt solutions, as well as 644 single salts, by 26 types of commercial polyamide membranes from the literature and our experiments. We explore the knowledge behind the ML predictions by utilizing the Shapley additive explanation (SHAP) method that is based on the cooperative game theory. The importance of multiple variables (including various membrane and ion properties, and operational conditions) to determining inorganic ion transport predicted by ML is revealed and compared among different ion types. We then pair the knowledge learned by the ML model with membrane separation mechanisms that have been discovered in the literature. It is worth mentioning that our work does not focus on developing new ML models for membrane performance predictions. Instead, our results demonstrate the implications of XAI in exploring ion transport in NF and RO membranes and provide a framework to evaluate the knowledge underlying ML model predictions, which can facilitate more reliable and explainable ML applications to membrane selection and design.

## 5.2 Materials and methods

### *5.2.1 Materials*

Sodium perchlorate ($NaClO_4$), sodium fluoride (NaF), and sodium bromide (NaBr) were purchased from Sigma-Aldrich (St. Louis, MO). Sodium nitrate ($NaNO_3$), hydrochloric acid (HCl, 36%), and sodium hydroxide (NaOH) were provided by Fisher Chemical (Hampton, NH). Commercial NF and RO membranes (BW30, NF270, NF, and NF90) were acquired from DuPont

FILMTEC (Wilmington, DE). Deionized (DI) water was produced by a commercial water purification system (>18MΩ, Millipore, Burlington, MA).

*5.2.2 Data collection*

The data for ion rejection of NF and RO membranes were collected from both the literature and the experiments. For the literature where the exact ion rejection rates were not mentioned in the text, we utilized "Engauge Digitizer" to extract the values from the graphs. We obtained 1,382 data points for membrane rejection of single salts, cations, and anions from 61 journal articles (Supporting Data). Compared to the data sets of single salt (644 data points) and cation (463 data points) rejections, the number of anion rejection data was relatively small (only 275 data points). Thus, we performed experiments to obtain data for fluoride, nitrate, perchlorate, and bromide rejections (different membranes and experimental conditions were used), which resulted in additional 203 data points. The detailed experimental conditions are described in the following section, and the obtained experimental data are available from the Supplementary Data.

*5.2.3 Anion rejection tests.*

We conducted anion rejection experiments to augment the data set for anion rejection. A customized membrane filtration system equipped with cross-flow cells with a membrane area of 20.02 $cm^2$ was used for the experiments.[37] Before the experiments, the membranes were stored in DI water for 24 hours. Then the membranes were compacted for at least 12 hours at 300 psi using DI water until the permeate flux was stable. The temperature and crossflow velocity were maintained at 23.0±1.0 °C and 1.0±0.1 L/min throughout the experiments, respectively. After membrane compaction, the feed water was replaced with the mixture solution of anions, which consisted of 1 mM of each $NaClO_4$, NaF, NaBr, and $NaNO_3$. The anion rejection by the membranes was measured at different pH values (i.e., from 5 to 8) and pressures (i.e., 100, 200, and 300 psi).

The concentrations of those anions in the feed and permeate solutions were measured with ion chromatography (Dionex Integrion Rfic, Thermofisher Scientific) to calculate the rejection rate for each anion.

*5.2.4 Variable selection*

To avoid overfitting of the ML model and reduce computational demands, it is important to select appropriate input variables and exclude those that are redundant.[38] The input variables we initially considered included ion properties (ionic radius, hydrated radius, and hydration energy), membrane properties (molecular weight cut-off (MWCO) and water contact angle), charge product (i.e., the product of membrane charge index and ionic potential[39]), and operational conditions (initial solute concentration, ionic strength, measurement time, and the difference between transmembrane pressure and feedwater osmotic pressure). As shown in Figures B1 and B2 (Appendix B), the hydration energy and hydrated radius are highly correlated with each other for all the scenarios (i.e., single salt solutions, cations and anions in mixture salt solutions), indicating that the ML model considers these two variables providing the same information for prediction (i.e., information redundancy). Therefore, we removed hydration energy from our input variables (this removal was done only from the perspective of ML model training; that is, the authors do not ignore the importance of the strength of hydration shell to ion transport). As demonstrated in Figure B3 (Appendix B), there were no variables showing a clear correlation with other variables across all scenarios after excluding hydration energy. Therefore, a total of 11 variables were used in our ML model for single salt solutions (Table B1, Appendix B), including MWCO, ionic radius of the cation (Ionic_radius+), ionic radius of the anion (Ionic_radius-), hydrated radius of the cation (Hyd_radius+), hydrated radius of the anion (Hyd_radius-), charge product, ionic strength (IS), water contact angle (WCA), the difference between transmembrane pressure and feedwater

osmotic pressure ($\Delta p$-$\Delta\pi$), initial solute concentration ($C_{in}$), and measurement time (T). For the mixture salt solutions, we only considered the properties of either cations or anions, thereby reducing the total number of variables to 9. The ion properties were mainly obtained from Marcus (1991) and Nightingale (1959). [40, 41] Especially, $\Delta p$-$\Delta\pi$ and charge product were calculated as follows:

$\Delta p$-$\Delta\pi$ = transmembrane pressure – C×R×T (5-1)

Charge product = membrane charge index × ionic potential of solute (5-2)

Ionic potential = ionic charge$^2$ / ionic radius × (+1 for cation or -1 for anion)[39] (5-3)

where C is the total molar concentration of all solutes in the solution, R is the ideal gas coefficient, and T is the absolute temperature. Also, we defined the membrane charge index as the difference between the isoelectric point of the membrane and experimental pH (i.e., isoelectric point minus experimental pH) rather than using the membrane zeta potential because the background electrolytes for zeta potential measurements are not the same for different types of membranes. We used the charge product to represent the electrostatic interaction between membranes and ions in ML model training, instead of using multiple separate variables (i.e., membrane charge index and ionic potential). This choice was made because charge product directly reflects the electrostatic interaction, thereby enhancing the efficiency of model learning process.[42]

*5.2.5 Machine learning models and their interpretations*

We examined the performance of four ML algorithms including Random Forest (RF), LightGBM (LGB), XGBoost (XGB), and Catboost (CAT), which have been commonly used in the literatures pertaining to ML,[22-24, 43] for the predictions of ion or salt rejections. As shown in Figure 5-1, the data collected were divided into training and testing data sets. To avoid data leakage (i.e., the intrusion of testing data information into the training and validation datasets),[23] the

training/validation and testing data set was split by using stratified sampling.[44] The data with identical type of salt/ion from the same set of experiments were designated to either one of the five groups in the cross-validation or the testing data set. As a result, the data with similar input variables and output were distributed to data for either model training or testing, avoiding the undesirable release of the information of testing data to training and validation datasets. The prediction results were accumulated (153, 156, and 108 ML models with different training/validation and testing data sets for single salts, and cations and anions in mixture salt solutions, respectively) to evaluate the ML model performance. The performance of the model during training was evaluated with the average root mean square error (RMSE) of validation datasets, and the combination of hyper-parameters with the lowest error was chosen for the best model. The hyper-parameters are the parameters, (e.g., learning rate, the number of decision trees, and the maximum depth of the decision trees) that determine the performance of the ML models.[45] They are not updated during ML model training and are generally optimized by finding the combinations of hyper-parameters showing the best performance of ML models.[45] RMSE was calculated as below:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2} \quad (4)$$

where $n$ is the number of samples, $y$ is the real salt/ion rejection, $\hat{y}$ is the predicted values by the model. The code was written in Python with Scikit-learn for Random Forest, LightGBM, XGBoost, Catboost, and SHAP package using Google Colab. Bayesian optimization was used to find the optimum hyper-parameters for the ML model. The lists of hyper-parameters and their ranges for different ML models are described in Tables B2 to B5 (Appendix B).

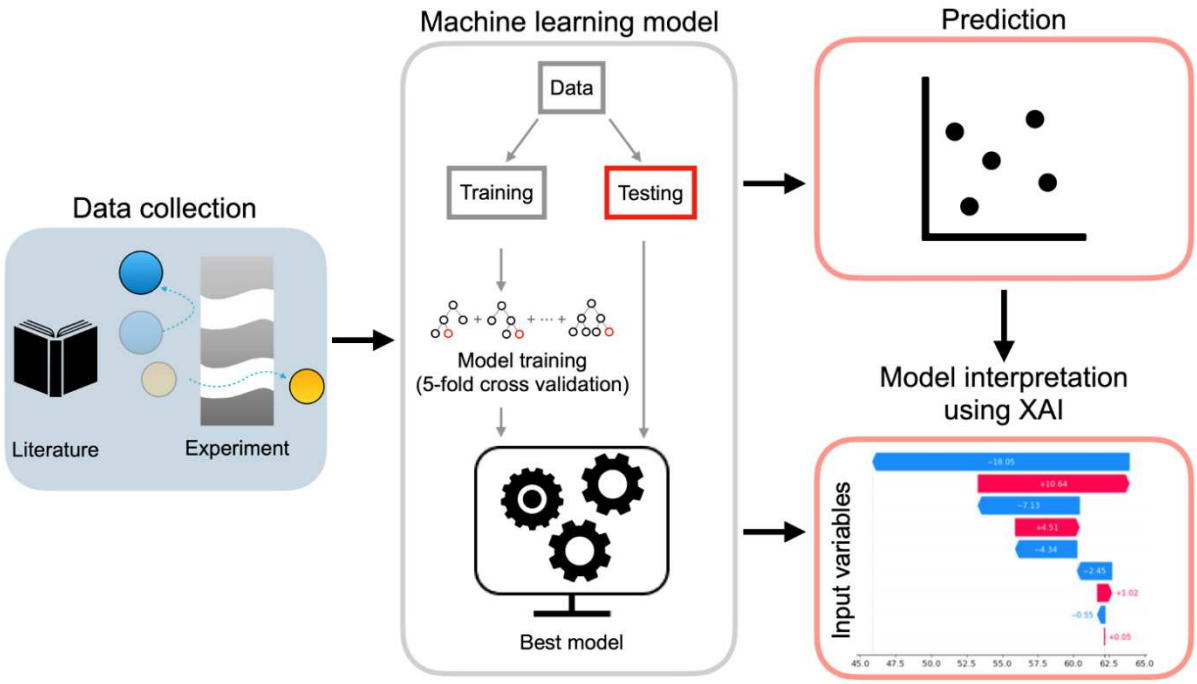


**Figure 5-1.** A schematic diagram of the machine learning models for ion rejection prediction and model interpretation using XAI.

We utilized Shapley additive explanation (SHAP) to interpret the underlying knowledge learned by the ML model on the mechanisms of ion rejection by polyamide membranes. SHAP is based on cooperative game theory that quantifies the contributions of each player (variable) to model predictions.[32] As a type of XAI, the SHAP method itself does not make predicts or impose influence on the predicting capability of ML models, but it can enhance the trust on the models by showing the underlying knowledge attained by ML.[46] SHAP values can be either positive or negative, which indicates whether the variable increases (positive SHAP value) or decreases (negative SHAP value) the average ion or salt rejection as predicted by the ML model, respectively. SHAP dependence plots were made to investigate the relationship of each variable with its SHAP values.

## 5.3 Results and discussion

### *5.3.1 Predictive performance of the ML model*

The data for our ML model were obtained from 61 publications and our own experiments (Supporting Data). The data for single salt solutions (644 data points) as well as cations (463 data points) and anions (478 data points) in mixture solutions were separately collected and utilized to build the ML model, because of the different ion transport behaviors for different solution compositions. Specifically, in a mixed salt solution, when one type of ion is not able to enter the membrane pore but the counter-ion permeates through the membrane, the transport of another less hindered co-ion may be facilitated to maintain electroneutrality.[47, 48] Further, even for the same ion at the same concentration, ion rejections can be different between single and mixture salt solutions due to the charge screening effect.[35] In brief, the higher ionic strengths of mixture salt solutions screen the membrane charge and diminish electrostatic exclusion of co-ions, promoting ion passage.[35, 47] Therefore, we consider ion transports in single and mixture salt solutions independently in this study.

The prediction accuracy of the ML model was evaluated using mean absolute error (i.e., the average of the absolute errors, MAE) and RMSE (i.e., the square root of the average squared values of the errors). The MAE values of ML model predictions are ~10%, ~11%, and ~12%, while the RMSE values are ~16%, ~15%, and ~16% for salt rejection in single salt solutions and cation and anion rejections in mixture salt solutions, respectively (Table 5-1). Such model prediction accuracy is comparable to that of ML models for the prediction of micropollutant rejections by NF and RO in our previous study,[23] in which we demonstrate that data leakage causes falsely high prediction accuracy of the ML model. As presented in Figures B4 to B6 (Appendix B), the MAEs of the ML model predictions are as low as ~4% and RMSEs are less than 10% with

the influence of data leakage (using XGB as an example). When data leakage occurs, the ML models make prediction outputs not only from the knowledge gained during model training, but also from the information of testing data, which should not be obtainable from the training dataset. This type of knowledge processing results in higher prediction accuracy of the ML models than their real prediction capabilities, preventing us from understanding the true model knowledge on membrane separation. Therefore, although the prediction accuracy of ML models is lower without data leakage, avoiding data leakage is essential to investigating the performance and analyzing the knowledge acquired by ML models objectively. In addition, as shown in Figures B7-B11 (Appendix B), the predicting capability of ML model was not vulnerable to data with certain ion or membrane types, and no clear trends were observed between any input variables and absolute error. This observation indicates that the predictions of ML model do not depend on certain variables and their corresponding membrane separation mechanisms.

**Table 5-1.** Evaluation of the performance of Random Forest, LightGBM, XGBoost, and Catboost models for predicting single salt rejection as well as cation and anion rejection in mixture salt solutions.

| | | Random Forest | LightGBM | XGBoost | Catboost |
|---|---|---|---|---|---|
| Single salt | MAE (%) | 10.20 | 11.00 | 9.85 | 10.95 |
| | RMSE (%) | 16.01 | 16.23 | 14.97 | 15.99 |
| | $R^2$ | 0.55 | 0.54 | 0.59 | 0.54 |
| Cation | MAE (%) | 10.69 | 11.59 | 11.10 | 9.27 |
| | RMSE (%) | 14.63 | 15.65 | 15.14 | 12.32 |
| | $R^2$ | 0.68 | 0.64 | 0.65 | 0.78 |
| Anion | MAE (%) | 11.74 | 11.27 | 11.10 | 13.16 |
| | RMSE (%) | 17.47 | 15.71 | 16.65 | 17.14 |
| | $R^2$ | 0.72 | 0.76 | 0.74 | 0.74 |

*5.3.2 Explainable artificial intelligence (XAI) for unveiling the knowledge learned by machine learning on ion transport across polyamide membranes.*

The ML algorithms have been utilized to understand nonlinear relationships between input variables and outputs in the field of environmental science and engineering, including membrane desalination.[22, 24, 25, 28, 49, 50] Due to the "black box" nature of ML models, the explainability of these models is essential to investigating the reliability of the decision-making process. XAI is able to reveal the decision-making process of ML models and enhances the trust on the models.[46] In this section, we applied XAI to quantify the contributions of multiple variables (e.g., membrane and ion properties, and operational conditions) to the ML model prediction simultaneously, in order to reveal the knowledge gained by ML regarding the mechanisms underlying ion transport across polyamide membranes in NF and RO.

Variable importance has been used to explore the influence of each input variable on ML model prediction.[24, 28, 51] However, this approach does not reveal detailed contributions of the variables to the model prediction (e.g., whether the change of certain variable values leads to an increase or decrease of output values predicted by the model). To better probe the knowledge learned by the ML model and quantify the detailed contribution of each variable to salt/ion rejection predictions, we implemented the Shapley additive explanation (SHAP), which is based on the cooperative game theory. Briefly, the cooperative game theory describes the payoffs (i.e., salt or ion rejection prediction in our study) of players (i.e., input variables) in the game where cooperation with other players enhances mutual benefits.[52] After building the ML model, the SHAP values are calculated by using a weighted average of the changes in the predictions with and without a certain variable, resulting from exploring every possible combination of the variables,[32] The SHAP value of a specific variable indicates whether this specific variable

increases (i.e., positive SHAP value) or decreases (i.e., negative SHAP value) the average rejection in the current study. The summation of SHAP values for all the variables and base rate (i.e., the average output of the training dataset) in each datum results in the model prediction.[32] We also presented SHAP dependence plots, which demonstrate the relationship between each variable and the SHAP values. Such information reveals whether the ML model gains proper knowledge on the role of each variable (and the corresponding membrane separation mechanism) in regulating ion transport during NF and RO, given the limited number of membrane performance data in the literature. We explored the variable importance to salt/ion transport as judged by ML models (i.e., RF, LGB, XGB, and CAT) using the SHAP method and presented the SHAP dependence plots mainly for XGB. As each variable under each scenario (i.e., single salt solutions, cations and anions in mixture salt solutions) generates one SHAP dependence plot, we selected XGB model, which exhibits comparable prediction accuracy to other ML models (Table 5-1) and was used to investigate micropollutant transport within polyamide membranes in our previous work,[23] as the major algorithm for discussing the SHAP dependence plots to reduce redundancy of this article.

5.3.2.1 Single salt solutions.

The SHAP summary displays the order of the variable importance (with decreasing importance from top to bottom) and the SHAP values calculated for the different variables. To explore the relative importance of the variables and the influence of each variable on the salt rejection, we quantify the SHAP importance (i.e., the average absolute SHAP values) and construct the SHAP dependence plots of the four highest-ranked variables. As shown in Figures 5-2A and 5-2B, MWCO, the difference between transmembrane pressure and feedwater osmotic pressure $(\Delta p\text{-}\Delta\pi)$, charge product (defined as the product of membrane charge index and ionic potential; the exact definition of membrane charge index and ionic potential have been described in the Materials and

Methods section), and the hydrated radius of the anion component (Hyd_radius-) have higher importance than other variables to the XGB model predictions (for ease of comparison, the SHAP importance score of each variable was normalized with that of the highest-ranked variable, whose score is defined as 100). It is worth mentioning that MWCO, $\Delta p$-$\Delta\pi$, and charge product also have high importance scores across all the ML models (Figure B12, Appendix B), demonstrating that ML models consistently recognize the importance of those parameters in governing single salt rejection regardless of the algorithm used.

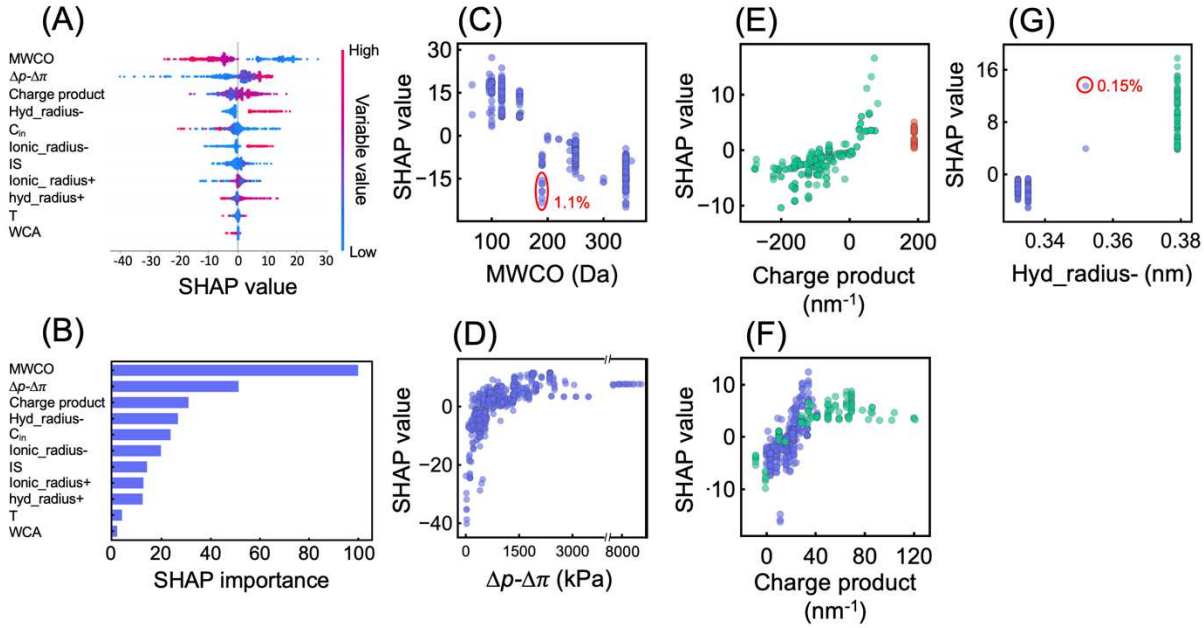


**Figure 5-2.** (A) The SHAP summary plot and (B) SHAP importance of the XGB model for single salt rejection prediction. The scale of the variable value is presented by red (high) and blue (low) colors. The number of data points sampled for all the variables is the same (644 data points) and the data points can be overlapped. The SHAP importance of each variable is normalized by the SHAP importance of the first-rank variable (defined as 100) for ease of comparison. The SHAP dependence plots of (C) MWCO (1.1% of data as outliers to the SHAP-MWCO relationship are indicated within the red circle), (D) $\Delta p$-$\Delta\pi$, (E) charge product (where cation is the dominant component determining the value of charge product), (F) charge product (where anion is the dominant component determining the value of charge product), and (G) hydrated radius of the anion (Hyd_radius-, 0.15% of data as outliers to the SHAP-Hyd_radius- relationship are indicated within the red circle) are shown. These four variables are the four highest-ranking variables in the SHAP summary plots, which significantly affect the prediction by the ML model. The blue, green, and red dots in (E), (F) and (G) indicate monovalent, divalent, and trivalent ions, respectively.

The most important variable for all the ML models to predict single salt rejection is MWCO, which is directly related to the mechanism of size exclusion. The SHAP dependence plot reveals that the SHAP values increase with a decrease of MWCO (Figure 5-2C, except for 1.1% of data as outliers), consistent with the fact that a smaller MWCO typically leads to a higher salt removal efficiency of polyamide membranes. For $\Delta p$-$\Delta\pi$ that is the second-ranked variable, the SHAP values increase with $\Delta p$-$\Delta\pi$ until 500 kPa and reach a plateau afterwards (Figure 5-2D). According to the solution-diffusion model,[53] the rate of water transport is proportional to $\Delta p$-$\Delta\pi$, and the increase in hydraulic pressure is able to cause higher water fluxes and consequently, higher ion rejections. While the water molecules are transported through the membrane, the solutes are accumulated near the membrane surfaces (i.e., a phenomenon referred to as concentration polarization). Concentration polarization enhances solute transport and limits further increase of water flux (the diminishing effect of hydraulic pressure on water flux across membranes as pressure increases was also shown by Wijmans and Baker),[53] which explains the plateau observed in the plot of SHAP value as a function of $\Delta p$-$\Delta\pi$. Similar results have been observed from the literature, demonstrating that salt rejection increased with pressure until certain thresholds and was maintained afterwards.[54, 55] Therefore, the knowledge attained by the ML model on how hydraulic pressure regulates salt rejection in NF and RO also agrees with the literature.

Figures 5-2E and 5-2F show the influence of electrostatic interaction between the salt ions and the membrane, as manifested by the parameter charge product. Charge product is calculated as the product of membrane charge index and ionic potential (detailed definition of charge product has been described in the Materials and Methods section). Thus, negative and positive values of charge product indicate electrostatic attraction and repulsion, respectively. As ionic potential considers the charge density of the ion, which affects the strength of ion-membrane or ion-ion

interactions from an electrostatic perspective,[39, 56] we employed ionic potential as a more proper variable than ionic charge to calculate charge product. In a single salt solution, the effect of electrostatic interactions on salt rejection varies depending on the valence ratio between the cation and anion components. The electrostatic repulsion between the co-ions and membrane surface regulates the transport of symmetric salts (e.g., NaCl and $MgSO_4$), whereas the electrostatic interactions between the membrane and asymmetric salts is governed by the ion with higher valence (e.g., $Mg^{2+}$ for $MgCl_2$ and $SO_4^{2-}$ for $Na_2SO_4$).[57] Therefore, for the data pertaining to single salt, we calculated the charge product by using ionic potential of the ions that determine the electrostatic interaction.

The plots of SHAP values as a function of charge product are shown in Figures 5-2E and 5-2F, which present the data where cations (for asymmetric salts with the cations having higher valence) and anions (for asymmetric salts with the anion having higher valence and symmetric salts) are the dominant components that determine salt transport, respectively. The SHAP values increase generally when charge product becomes more positive where anions regulate single salt rejection (Figure 5-2F), whereas the increase of electrostatic attraction between divalent cations and the membrane surface (charge product < 0) causes the decrease of the SHAP values (Figure 5-2E). This indicates that the ML model correctly predicts higher or lower salt rejections from electrostatic repulsion and attraction, respectively. However, no clear trends are observed when trivalent cations determine salt transport, probably due to the lack of sufficient data. Thus, ML captures the relationship between electrostatic interactions and salt rejection, identifying the regulating roles of membrane-ion interactions in single salt transport across polyamide membranes.

The hydrated radius of the anion, which also relates to size exclusion, is the 4th ranked variable for XGB model prediction (Figure 5-2G). The increase in the hydrated radius of the anion

generally leads to greater salt rejection (higher SHAP values, except for 0.15% of data as outliers). Although this result is aligned with the mechanism of size exclusion, we notice that the importance of ion size to model prediction is not consistent among ML algorithms for predicting single salt rejection. The parameters relating to ion size (i.e., hydrated radius or ionic radius) are not always among the top tanked variables in terms of SHAP importance (Figure B12, Appendix B), indicating that ML models have a weaker understanding on the role of ion size in regulating single salt rejection compared to the variables discussed above. For symmetric salt solutions, salt transport is governed by the co-ion (typically anions due to the negatively charged polyamide membrane surface at operational pH values), whereas the ion with higher valence is the dominant factor that determines salt rejection for asymmetrical salt solutions. Therefore, the effect of ion size on salt rejection is not straightforward for the ML models to learn, explaining the lower importance of ion size than MWCO (another variable regarding size exclusion) to determining model prediction and the inconsistency of its importance among ML algorithms. Additionally, salt transport for 74.5% of our dataset is governed by anions (i.e., symmetric salts or asymmetrical salts with anion having higher valence). This explains why the ion size of anion, rather than that of cation, have more significant contributions to the model prediction (Figure 5-2G and Figure B12).

5.3.2.2 Cations in mixture salt solutions

The importance of variables in determining cation rejection for mixture salt solutions is shown in Figure 5-3 and Figure B13 (Appendix B). Different from the results for single salt solutions, hydrated radius has much higher importance than other variables for cation rejection predictions by the XGB model, followed by MWCO, $\Delta p$-$\Delta \pi$, and charge product (Figures 5-3A and 5-3B). Such an order of variable importance is consistently observed for RF and LGB models (for the

CAT model, hydrated radius, MWCO, and $\Delta p$-$\Delta\pi$ also ranked the highest, except for a relatively lower rank for charge product), showing that the underlying knowledge of ML models on cation transport across polyamide membranes are generally consistent among different type of ML models (Figure B13, Appendix B).

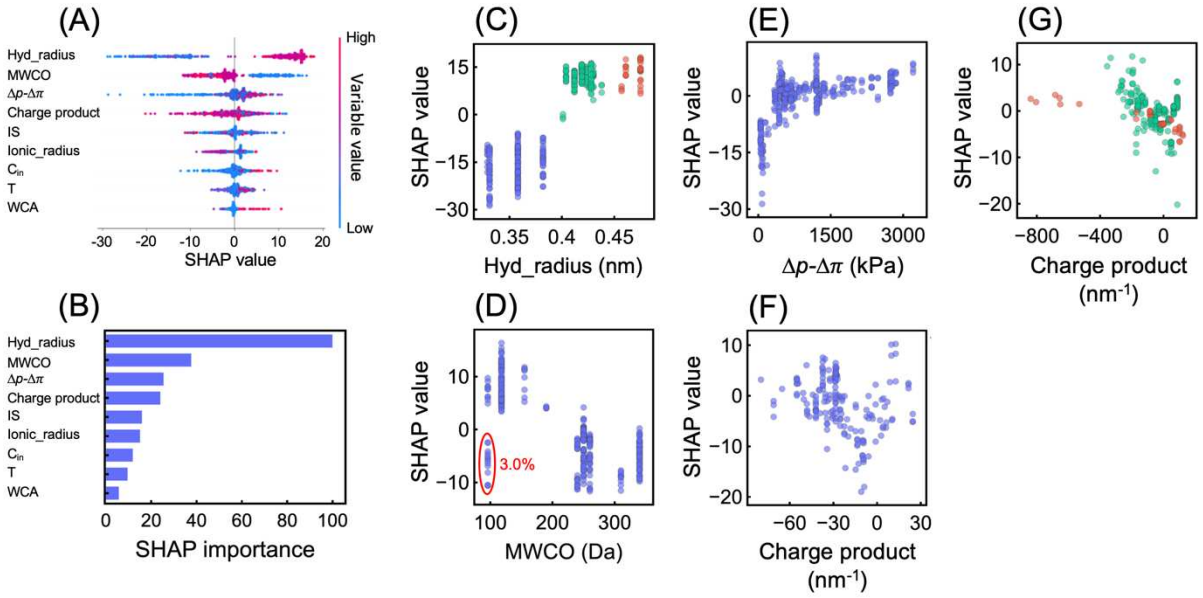


**Figure 5-3.** (A) The SHAP summary plot and (B) SHAP importance of the ML model for cation rejection prediction. The scale of the variable value is presented by red (high) and blue (low) colors. The number of data points sampled for all the variables is the same (463 data points) and the data points can be overlapped. The SHAP importance of each variable is normalized by the SHAP importance of the first-rank variable (defined as 100) for ease of comparison. The SHAP dependence plots of (C) hydrated radius (Hyd_radius), (D) MWCO (3.0% of data as outliers to the SHAP-MWCO relationship are indicated within the red circle), (E) $\Delta p$-$\Delta\pi$, (F) charge product for monovalent cations, and (G) charge product for multivalent cations are shown. These four variables are the four highest-ranking variables in the SHAP summary plots, which significantly affect the prediction of the ML model. The blue, green, and red dots in (C), (F), and (G) indicate monovalent, divalent, and trivalent cations, respectively.

As shown in Figure 5-3C, negative and positive SHAP values are generally obtained for monovalent (hydrated radius < 0.39 nm) and multivalent (hydrated radius > 0.39 nm) cations, respectively. This trend of SHAP values as a function of hydrated radius is in good agreement with experimental results from the literature where cation rejection increases with hydrated radii for

monovalent and divalent cations, but the difference in ion transport rate between divalent and trivalent cations is relatively minor.[58-61] It is worth mentioning that hydrated radius ranks much higher than ionic radius for the predictions of cation removal (which is consistently observed for all the ML models, Figure B13). Along with the fact that hydrated radius is highly correlated to hydration energy of ions (Figure B2; thus we only keep hydrated radius when selecting model variables, in order to avoid variable redundancy), the high importance of hydrated radius and the trend of SHAP values (i.e., higher SHAP values of multivalent cations with larger hydrated radius than monovalent cations) indicate that ML captures the importance of the hydration shell and its potential rearrangement (or partial dehydration) in cation transport across polyamide membranes, which has been recently revealed by well-controlled experiments.[60, 62] Indeed, when we replace hydrated radius with hydration energy in ML model training, hydration energy is also found to be the most important variable for model prediction (Figure B14, Appendix B), justifying both our variable selection and the aforementioned conclusion on ML knowledge.

The impact of MWCO on model predictions of cation rejection is displayed in Figure 5-3D. The SHAP value generally decreases with an increase of MWCO (except for 3.0% of data as outliers), indicating the capability of the ML model to identify the mechanism of size exclusion. However, the importance of MWCO in the prediction of cation rejection is relatively low compared to that of hydrated radius (Figures 5-3B and B13), which is different from what is observed for single salt and anion rejections (Figures 5-2B, 5-4B, B12 and B15, Appendix B). The prominently high contribution of hydrated radius to the ML prediction of cation rejection will be discussed in the next subsection.

The role of $\Delta p$-$\Delta\pi$ in predicting cation rejection is shown in Figure 5-3E. Similar to what is observed for single salt rejection (Figure 5-2D), the SHAP value increases with $\Delta p$-$\Delta\pi$ until a

threshold is reached. The increase of $\Delta p$-$\Delta\pi$ results in higher water fluxes and consequently enhances cation rejection when the influence of concentration polarization is relatively minor (i.e., when $\Delta p$-$\Delta\pi$ is low).[63] As the hydraulic pressure increases, the contribution of hydraulic pressure to ion rejection becomes limited because the accumulation of solutes close to the membrane surface enhances the osmotic pressure (which limits further increase of water flux) and the permeabilities of cations.[60, 64] The fact explains the observed threshold of $\Delta p$-$\Delta\pi$ after which the SHAP values are kept relatively constant.

The SHAP values as a function of charge product for the model prediction of cation rejection are displayed in Figures 5-3F and 5-3G. The SHAP-charge product plots for monovalent and divalent cations show a V shape, with the lowest SHAP values obtained around the charge product value of 0. The rejections of monovalent and divalent cations are enhanced (i.e., a general increase of SHAP value) when the electrostatic repulsion (charge product > 0) gets stronger. Accordingly, the ML model identifies that electrostatic repulsion hampers the approaches of ions to the membrane pores[65] and the subsequent ion transport across membranes. However, the SHAP value also increases when charge product becomes more negative, indicating that enhanced electrostatic attraction between membranes and cations also increases the rejection of monovalent and divalent cations. This trend of SHAP values for cation in mixture salt solution is different from what is shown in Figure 5-2E where the cation is the dominant ion determining electrostatic interactions between single salts and membrane surface. For cations in mixture salt solutions, the anions (typically the co-ions due to the negatively charged membrane surface), which experience electrostatic repulsion, also determine the impact of electrostatic interactions on ion rejections.[57, 66] When charge product becomes more negative for cations, the negative charge of membrane surface typically increases, resulting in stronger electrostatic repulsion of the anions. The

consequently improved anion rejection could lead to an increase of cation rejection to maintain electroneutrality. Such a phenomenon is not applicable to single salt and anion rejections, because the ions used to calculate charge product are usually those that dominate salt permeation. As a result, the V shape of SHAP-charge product plots is only observed for cation rejection. In addition, no clear relationship between SHAP and charge product values is observed for trivalent cations, which might be due to the relatively small number of trivalent cations within the training data that is not enough to reveal the pattern.

#### 5.3.2.3 Anions in mixture salt solutions

Figure 5-4 and Figure B15 show the contributions of variables to anion rejection prediction. The four most important contributors to anion rejection prediction by the XGB model are MWCO, charge product, WCA, and hydrated radius. Although MWCO, charge product, and hydrated radius are highly ranked for predicting both cation and anion transport, the relative importance of the variables differs (Figures 5-3, 5-4, B13, and B15). Especially, charge product, which is related to electrostatic interactions, plays a more significant contribution to determining the model predictions of anion rejection than cation rejection. Moreover, we notice that MWCO ranks as the most important variable for model prediction, which is contrasting to the scenario of cation rejection prediction (where hydrated radius ranks the highest) but consistent with that of single salt rejection prediction. Such trends of SHAP importance (i.e., the high importance of MWCO and charge product) are consistent across different ML models for anion rejection predictions (Figures 5-4 and B15).

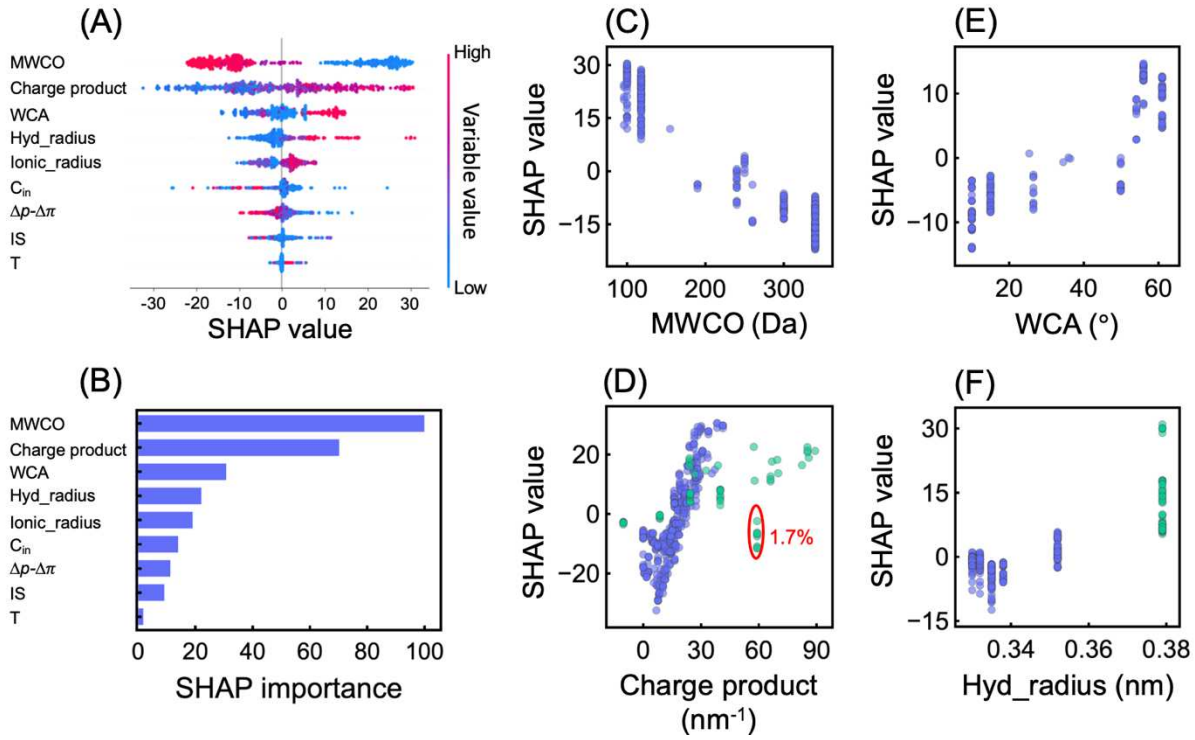


**Figure 5-4.** (A) The SHAP summary plot and (B) SHAP importance of the ML model for anion rejection prediction. The scale of the variable value in (A) is presented by red (high) and blue (low) colors. The number of data points sampled for all the variables is the same (478 data points) and the data points can be overlapped. The SHAP importance indicates the average of the absolute SHAP value in the data. The SHAP importance of each variable was normalized by the SHAP importance of the first rank variable for ease of comparison. The SHAP dependence plots of (C) MWCO, (D) charge product (1.7% of data as outliers to the SHAP-charge product relationship are indicated within the red circle), (E) water contact angle (WCA), and (F) hydrated radius (Hyd_radius) are shown. These four variables are the four highest-ranking variables in the SHAP summary plots, which significantly affect the prediction of the ML model. The blue and green dots in (D) and (F) indicate monovalent and divalent anions, respectively.

We explain the lower importance of hydrated radius to ML prediction for anions than for cations by examining the relative sizes of hydrated radius and membrane pores. The range of hydrated radii for anions (0.33 nm – 0.38 nm) is narrower compared to those for cations (0.33 nm – 0.48 nm).[41] Such a difference suggests that there are more chances for cations to have hydrated radii larger than membrane pores. For example, the average pore radii of NF270 and NF90 membranes, which are the most commonly used NF membrane in our dataset, are estimated to be 0.34-0.38 nm and 0.29-0.34 nm in the literature.[60] Compared to the scenarios for anions, hydrated radius plays a more important role in determining whether cations are able to partition into the

membrane pores. As a result, hydrated radius imposes a more significant contribution to membrane rejection of cations than anions, and the ML models recognize this difference when making predictions (for single salts, the salt transport is mostly regulated by the anion components as we discussed above).

The negative correlation of MWCO with its SHAP values indicates higher ion rejection for smaller MWCO (Figure 5-4C), consistent with the mechanism of size exclusion. Electrostatic repulsion between membrane and anion (i.e., charge product > 0) was observed in most of the anion rejection data (Figure 5-4D). Charge product ranks higher for its contribution to ML prediction of anion rejection than cation rejection, indicating that XAI identifies a greater impact of electrostatic interaction on anion transport. This is consistent with the literature that the membrane selectivity of monovalent to divalent anions, which experience electrostatic repulsion with the membrane surface, is higher than the selectivity of monovalent to divalent cations (e.g., the $Na^+/Ca^{2+}$ selectivity is typically 1.3–4.0, whereas the $Cl^-/SO_4^{2-}$ selectivity is 11.1–68.0 for polyamide membranes).[67-69] The lowest SHAP value of monovalent anions is observed at charge product of ~7, with the contribution of charge product to ion rejection increasing afterwards as electrostatic repulsion gets stronger. Such a phenomenon is observed across all ML models (Figure B16, Appendix B) and is indeed consistent with the fact that the highest anion transport rate was reported at pH values slightly higher than the isoelectric points of membranes.[70, 71] However, a few low SHAP values (1.7% of the data) for divalent anions were observed at charge product around ~60 for XGB model. These irregular SHAP values might be due to the relatively smaller number of data for divalent anions compared to monovalent anions, leading to incomplete knowledge acquired by the XGB model on electrostatic interaction. As a very limited number of data points exist for trivalent ions, no clear trend is observed.

Interestingly, WCA of the membrane, which has low importance for single salt and cation rejections, ranks the third for anion rejection prediction by the XGB model (Figures 5-4A and 5-4B). The SHAP values of WCA indicate that more hydrophobic membrane surfaces result in higher anion rejection (Figure 5-4E). This phenomenon might be due to the negative correlation of WCA with MWCO for the membranes that were used to generate our experimental results (i.e., NF270, NF, NF90, and BW30 membranes, which were used to generate 43% of the data for anion rejection, Figure B17, Appendix B), rather than reflecting physical contribution of the hydrophobicity of the membrane to ion transport (we still include WCA as an input variable to make the variable types the same across all the scenarios).

Further, the SHAP values of hydrated radius for divalent anions are higher than those for monovalent anions (Figure 5-4F), consistent with the literature that energy barriers of anions during transport across membranes are higher for those with larger hydrated radius and stronger hydration energy.[59, 65, 72] For divalent anions ($SO_4^{2-}$), positive contributions of hydrated radius to ion rejection (i.e., SHAP values > 0) are always observed, reflective of the fact that divalent anions with larger hydrated radius experience a hindered transport across membranes.[73] These results suggest that the ML has gained a generally proper knowledge of the role that hydrated radius plays in regulating anion transport.

Additionally, we also examine the contribution of $\Delta p$-$\Delta \pi$ to ML prediction of anion rejection (Figure B18, Appendix B). Although the SHAP values increase until $\Delta p$-$\Delta \pi$ reaches ~700 kPa for divalent anions, they decrease as $\Delta p$-$\Delta \pi$ increases for monovalent anions. This phenomenon, which is valid for all the ML models, contrasts what is observed for single salt and cation rejections, and is also inconsistent with experimental results from the literature that the increase of hydraulic pressure enhances the anion rejection.[74] This inconsistency indicates that the

ML model is constrained from understanding the role of hydraulic pressure in regulating anion transport in our study, and that it is important to examine the SHAP dependence plots (rather than only the SHAP importance) to test whether the understanding of ML is aligned with domain knowledge.

### 5.4 Implications

In this study, we utilized XAI to probe the underlying knowledge attained by ML trained with data available from the literature on the mechanisms of membrane separation. XAI unveils that ML is able to capture the important roles of size exclusion and electrostatic interaction in governing ion transport across polyamide NF and RO membranes. XAI also reveals that the mechanisms of membrane separation pose different relative importance in determining the rejection of cations and anions by NF and RO. Electrostatic interactions (manifested by charge product) possess a heavier weight in determining anion rejection than cation rejection, whereas the hydrated radius (related to size exclusion and ion dehydration) contributes more to the prediction of cation rejection by the ML models.[57, 75] The membrane separation mechanisms learned by ML are generally aligned with the results from the literature. For example, the high importance of MWCO and hydrated radius to model prediction, as well as their roles of determining salt/ion rejections (as indicated by the SHAP dependence plots), correctly reflect how membrane pore size and ion size regulate salt/ion permeation through NF and RO membranes.[15, 76] It has been shown that the rate of cation transport across polymeric membranes decreases with an increase of hydrated radius,[59] and a recent study by Lu et al.[62] reveals that ion dehydration governs membrane separation of NF when the hydrated radius of cation is larger than membrane pores. These results are consistent with the ML knowledge that hydrated radius plays a significant role in determining cation rejection. Also, the higher SHAP importance of charge product for anion rejection agrees with the fact that higher membrane

selectivity of monovalent to divalent anions (e.g., $Cl^-/SO_4^{2-}$) has been achieved by tuning the surface charge of NF membranes than that of monovalent to divalent cations (e.g., $Na^+/Ca^{2+}$).[15, 67] The consistency between the knowledge of ML models unveiled by XAI and the solute transport mechanisms revealed by experimental results in the literature indicates that ML has the capability of capturing the fundamental principles of membrane separation. To the best of knowledge, our work represents the first effort that thoroughly scrutinizes the knowledge of ML models underlying their predictions of ion/salt transport across polyamide membranes.

Recently, there have been an increasing number of applications of ML to the prediction and optimization associated with membrane performance and design. For example, Gao et al.[22] combined ML with Bayesian optimization to investigate the key variables for membrane fabrications and to inversely identify optimal conditions for membrane fabrication. Also, Yang et al.[29] utilized ML to find new polymers for fabricating gas separation membranes and confirmed the high performance of the resultant polymeric membranes using molecular dynamics simulations. Both studies were able to develop novel membranes that broke the current upper boundary of membrane perm-selectivity trade-off.[22, 29] Although the use of ML to improve rational membrane selection and design is promising, the reliability of ML is largely dependent on whether the prediction made by ML models, which are trained with a limited number of available data, can be properly aligned with fundamental domain principles of membrane science. The comprehensive investigations using XAI in this study provides a framework for evaluating the knowledge learned by ML and evidence that ML models can learn and understand membrane separation mechanisms to guide their predictions.

However, ML is not able to identify the roles of variables in governing solute transport across membranes when the volume of data is insufficient. For example, the relationship between

charge product and its SHAP value is not clearly observed for trivalent cations due to the smaller number of available data compared to those of monovalent and divalent cations. There is no criterion on the sufficient amount of data for ML model training, but a relatively smaller number of data points for a certain group deteriorates the ML model performance. It is worth mentioning that such a limitation is not due to the quality of ML model itself and should be addressed if more membrane performance data of high quality are available from diverse experimental conditions. One of the major challenges is to find available data containing all the required variables from the literature. In order to augment the data volume for future research, we encourage researchers to provide more comprehensive details of their experiments, such as the properties of membranes and operational conditions in their publications. A standardized, cross-laboratory criterion on the reporting of experimental procedure, combined with publicly accessible databases such as the recently published Open Membrane Database (OMD; it is worth mentioning that this database is currently focusing on perm-selectivity of membranes but not reporting detailed data on specific salt/ion rejection),[77] will significantly enhance the knowledge and performance of ML for a better, data-driven paradigm of membrane selection and design.

**References**


1. Yoon, Y.; Lueptow, R. M., Removal of organic contaminants by RO and NF membranes. *Journal of Membrane Science* 2005, 261, (1), 76-86.

2. Paul, M.; Jons, S. D., Chemistry and fabrication of polymeric nanofiltration membranes: A review. *Polymer* 2016, 103, 417-456.

3. Li, X.; Mo, Y.; Qing, W.; Shao, S.; Tang, C. Y.; Li, J., Membrane-based technologies for lithium recovery from water lithium resources: A review. *Journal of Membrane Science* 2019, 591, 117317.

4. Amy, G.; Ghaffour, N.; Li, Z.; Francis, L.; Linares, R. V.; Missimer, T.; Lattemann, S., Membrane-based seawater desalination: Present and future prospects. *Desalination* 2017, 401, 16-21.

5. Bunani, S.; Yörükoğlu, E.; Sert, G.; Yüksel, Ü.; Yüksel, M.; Kabay, N., Application of nanofiltration for reuse of municipal wastewater and quality analysis of product water. *Desalination* 2013, 315, 33-36.

6. Li, K.; Wang, J.; Liu, J.; Wei, Y.; Chen, M., Advanced treatment of municipal wastewater by nanofiltration: Operational optimization and membrane fouling analysis. *Journal of Environmental Sciences* 2016, 43, 106-117.

7. Liang, S.; Liu, C.; Song, L., Two-step optimization of pressure and recovery of reverse osmosis desalination process. *Environmental Science & Technology* 2009, 43, (9), 3272-3277.

8. Hermans, S.; Bernstein, R.; Volodin, A.; Vankelecom, I. F. J., Study of synthesis parameters and active layer morphology of interfacially polymerized polyamide–polysulfone membranes. *Reactive and Functional Polymers* 2015, 86, 199-208.

9. Yan, F.; Chen, H.; Lü, Y.; Lü, Z.; Yu, S.; Liu, M.; Gao, C., Improving the water permeability and antifouling property of thin-film composite polyamide nanofiltration membrane by modifying the active layer with triethanolamine. *Journal of Membrane Science* 2016, 513, 108-116.

10. Gong, G.; Wang, P.; Zhou, Z.; Hu, Y., New insights into the role of an interlayer for the fabrication of highly selective and permeable thin-film composite nanofiltration membrane. *ACS Applied Materials & Interfaces* 2019, 11, (7), 7349-7356.

11. Culp, T. E.; Khara, B.; Brickey, K. P.; Geitner, M.; Zimudzi, T. J.; Wilbur, J. D.; Jons, S. D.; Roy, A.; Paul, M.; Ganapathysubramanian, B.; Zydney, A. L.; Kumar, M.; Gomez, E. D., Nanoscale control of internal inhomogeneity enhances water transport in desalination membranes. *Science* 2021, 371, (6524), 72-75.

12. Elimelech, M.; Phillip, W. A., The future of seawater desalination: Energy, technology, and the environment. *Science* 2011, 333, (6043), 712-717.

13. Werber, J. R.; Deshmukh, A.; Elimelech, M., The critical need for increased selectivity, not increased water permeability, for desalination membranes. *Environmental Science & Technology Letters* 2016, 3, (4), 112-120.

14. Werber, J. R.; Osuji, C. O.; Elimelech, M., Materials for next-generation desalination and water purification membranes. *Nature Reviews Materials* 2016, 1, (5), 16018.

15. Zhao, Y.; Tong, T.; Wang, X.; Lin, S.; Reid, E. M.; Chen, Y., Differentiating Solutes with Precise Nanofiltration for Next Generation Environmental Separations: A Review. *Environmental Science & Technology* 2021, 55, (3), 1359-1376.

16. Bowen, W. R.; Welfoot, J. S., Modelling the performance of membrane nanofiltration - critical assessment and model development. *Chemical Engineering Science* 2002, 57, 1121-1137.
17. Bowen, W. R.; Mohammad, A. W., Characterization and prediction of nanofiltration membrane performance—A general Assessment. *Chemical Engineering Research and Design* 1998, 76, (8), 885-893.
18. Bowen, W. R.; Mukhtar, H., Characterisation and prediction of separation performance of nanofiltration membranes. *Journal of Membrane Science* 1996, 112, (2), 263-274.
19. Bowen, W. R.; Mohammad, A. W.; Hilal, N., Characterisation of nanofiltration membranes for predictive purposes — use of salts, uncharged solutes and atomic force microscopy. *Journal of Membrane Science* 1997, 126, (1), 91-105.
20. Wang, R.; Lin, S., Pore model for nanofiltration: History, theoretical framework, key predictions, limitations, and prospects. *Journal of Membrane Science* 2021, 620, 118809.
21. 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. *Science Advances* 2022, 8, (2), eabl5771.
22. 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. *Environmental Science & Technology* 2022, 56, (4), 2572-2581.
23. Jeong, N.; Chung, T.-h.; Tong, T., Predicting micropollutant removal by reverse osmosis and nanofiltration membranes: Is machine learning viable? *Environmental Science & Technology* 2021, 55, (16), 11348-11359.
24. Lee, S.; Kim, J., Prediction of nanofiltration and reverse-osmosis-membrane rejection of organic compounds using random forest model. *Journal of Environmental Engineering* 2020, 146, (11), 04020127.
25. Hu, J.; Kim, C.; Halasz, P.; Kim, J. F.; Kim, J.; Szekely, G., Artificial intelligence for performance prediction of organic solvent nanofiltration membranes. *Journal of Membrane Science* 2021, 619, 118513.
26. Khaouane, L.; Ammi, Y.; Hanini, S., Modeling the retention of organic compounds by nanofiltration and reverse osmosis membranes using bootstrap aggregated neural networks. *Arabian Journal for Science and Engineering* 2017, 42, (4), 1443-1453.
27. Ammi, Y.; Khaouane, L.; Hanini, S., Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. *Korean Journal of Chemical Engineering* 2015, 32, (11), 2300-2310.
28. 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. *Journal of Membrane Science* 2009, 342, (1), 251-262.
29. Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y., Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. *Science Advances* 2022, 8, (29), eabn9545.
30. Song, X.; Yu, A. S. L.; Kellum, J. A.; Waitman, L. R.; Matheny, M. E.; Simpson, S. Q.; Hu, Y.; Liu, M., Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nature communications* 2020, 11, (1), 5668.

31. Lauritsen, S. M.; Kristensen, M.; Olsen, M. V.; Larsen, M. S.; Lauritsen, K. M.; Jørgensen, M. J.; Lange, J.; Thiesson, B., Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications* 2020, 11, (1), 3852.

32. Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S., From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2020, 2, (1), 56-67.

33. Ribeiro, M. T.; Singh, S.; Guestrin, C. In " Why should i trust you?" Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016; 2016; pp 1135-1144.

34. 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. *Science Advances* 2020, 6, (48), eabd9045.

35. Wang, R.; Zhang, J.; Tang, C. Y.; Lin, S., Understanding selectivity in solute–solute separation: Definitions, measurements, and comparability. *Environmental Science & Technology* 2022, 56, (4), 2605–2616.

36. Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-j.; Zhang, H., Machine learning: new ideas and tools in environmental science and engineering. *Environmental Science & Technology* 2021, 55, (19), 12741-12754.

37. Yin, Y.; Kalam, S.; Livingston, J. L.; Minjarez, R.; Lee, J.; Lin, S.; Tong, T., The use of anti-scalants in gypsum scaling mitigation: Comparison with membrane surface modification and efficiency in combined reverse osmosis and membrane distillation. *Journal of Membrane Science* 2022, 643, 120077.

38. Zhao, Z.; Wang, L.; Liu, H., Efficient spectral feature selection with minimum redundancy. *Proceedings of the AAAI Conference on Artificial Intelligence* 2010, 24, (1), 673-678.

39. Tansel, B., Significance of thermodynamic and physical characteristics on permeation of ions during membrane separation: Hydrated radius, hydration free energy and viscous effects. *Separation and Purification Technology* 2012, 86, 119-126.

40. Marcus, Y., Thermodynamics of solvation of ions. Part 5.—Gibbs free energy of hydration at 298.15 K. Journal of the Chemical Society, *Faraday Transactions* 1991, 87, (18), 2995-2999.

41. Nightingale, E. R., Phenomenological Theory of Ion Solvation. Effective Radii of Hydrated Ions. *The Journal of Physical Chemistry* 1959, 63, (9), 1381-1387.

42. Chollet, F., Deep learning with Python. In Manning Publications: 2018; p 103.

43. Ibrar, I.; Yadav, S.; Braytee, A.; Altaee, A.; HosseinZadeh, A.; Samal, A. K.; Zhou, J. L.; Khan, J. A.; Bartocci, P.; Fantozzi, F., Evaluation of machine learning algorithms to predict internal concentration polarization in forward osmosis. *Journal of Membrane Science* 2022, 646, 120257.

44. Liberty, E.; Lang, K.; Shmakov, K., Stratified Sampling Meets Machine Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, Maria Florina, B.; Kilian, Q. W., Eds. PMLR: Proceedings of Machine Learning Research, 2016; Vol. 48, pp 2320--2329.

45. Yu, T.; Zhu, H., Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv:2003.05689 2020.

46. Burkart, N.; Huber, M. F., A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 2021, 70, 245-317.

47. Yaroshchuk, A. E., Negative rejection of ions in pressure-driven membrane processes. *Advances in Colloid and Interface Science* 2008, 139, (1), 150-173.

48. Nicolini, J. V.; Borges, C. P.; Ferraz, H. C., Selective rejection of ions and correlation with surface properties of nanofiltration membranes. *Separation and Purification Technology* 2016, 171, 238-247.

49. Kerckhoffs, J.; Hoek, G.; Portengen, L.; Brunekreef, B.; Vermeulen, R. C. H., Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. *Environmental Science & Technology* 2019, 53, (3), 1413-1421.

50. Zhang, Z.; Luo, Y.; Peng, H.; Chen, Y.; Liao, R.-Z.; Zhao, Q., Deep spatial representation learning of polyamide nanofiltration membranes. *Journal of Membrane Science* 2021, 620, 118910.

51. Ignacz, G.; Szekely, G., Deep learning meets quantitative structure–activity relationship (QSAR) for leveraging structure-based prediction of solute rejection in organic solvent nanofiltration. *Journal of Membrane Science* 2022, 646, 120268.

52. Chalkiadakis, G.; Elkind, E.; Wooldridge, M., Cooperative game theory: Basic concepts and computational challenges. *IEEE Intelligent Systems* 2012, 27, (3), 86-90.

53. Wijmans, J. G.; Baker, R. W., The solution-diffusion model: a review. *Journal of Membrane Science* 1995, 107, (1), 1-21.

54. Jang, E.-S.; Mickols, W.; Sujanani, R.; Helenic, A.; Dilenschneider, T. J.; Kamcev, J.; Paul, D. R.; Freeman, B. D., Influence of concentration polarization and thermodynamic non-ideality on salt transport in reverse osmosis membranes. *Journal of Membrane Science* 2019, 572, 668-675.

55. Koyuncu, I.; Topacik, D., Effect of cross flow velocity, feed concentration, and pressure on the salt rejection of nanofiltration membranes in reactive dye having two sodium salts and NaCl mixtures: Model application. *Journal of Environmental Science and Health, Part A* 2004, 39, (4), 1055-1068.

56. Li, R.; Yang, W.; Su, Y.; Li, Q.; Gao, S.; Shang, J. K., Ionic potential: A general material criterion for the selection of highly efficient arsenic adsorbents. *Journal of Materials Science & Technology* 2014, 30, (10), 949-953.

57. Luo, J.; Wan, Y., Effects of pH and salt on nanofiltration—a critical review. *Journal of Membrane Science* 2013, 438, 18-28.

58. Huang, Q.; Liu, S.; Guo, Y.; Liu, G.; Jin, W., Separation of mono-/di-valent ions via charged interlayer channels of graphene oxide membranes. *Journal of Membrane Science* 2022, 645, 120212.

59. Wang, P.; Wang, M.; Liu, F.; Ding, S.; Wang, X.; Du, G.; Liu, J.; Apel, P.; Kluth, P.; Trautmann, C.; Wang, Y., Ultrafast ion sieving using nanoporous polymeric membranes. *Nature Communications* 2018, 9, (1), 569.

60. Pavluchkov, V.; Shefer, I.; Peer-Haim, O.; Blotevogel, J.; Epsztein, R., Indications of ion dehydration in diffusion-only and pressure-driven nanofiltration. *Journal of Membrane Science* 2022, 648, 120358.

61. Wen, Q.; Yan, D.; Liu, F.; Wang, M.; Ling, Y.; Wang, P.; Kluth, P.; Schauries, D.; Trautmann, C.; Apel, P.; Guo, W.; Xiao, G.; Liu, J.; Xue, J.; Wang, Y., Highly selective ionic transport through subnanometer pores in polymer films. *Advanced functional materials* 2016, 26, (32), 5796-5803.

62. Lu, C.; Hu, C.; Ritt, C. L.; Hua, X.; Sun, J.; Xia, H.; Liu, Y.; Li, D.-W.; Ma, B.; Elimelech, M.; Qu, J., In situ characterization of dehydration during ion transport in polymeric nanochannels. *Journal of the American Chemical Society* 2021, 143, (35), 14242-14252.

63. Song, L.; Elimelech, M., Theory of concentration polarization in crossflow filtration. Journal of the Chemical Society, *Faraday Transactions* 1995, 91, (19), 3389-3398.

64. Crittenden, J. C.; Trussell, R. R.; Hand, D. W.; Howe, K. J.; Tchobanoglous, G., MWH's water treatment: principles and design. John Wiley & Sons: 2012; p 1368.

65. Epsztein, R.; Cheng, W.; Shaulsky, E.; Dizge, N.; Elimelech, M., Elucidating the mechanisms underlying the difference between chloride and nitrate rejection in nanofiltration. *Journal of Membrane Science* 2018, 548, 694-701.

66. Shefer, I.; Peer-Haim, O.; Leifman, O.; Epsztein, R., Enthalpic and entropic selectivity of water and small ions in polyamide membranes. *Environmental Science & Technology* 2021, 55, (21), 14863-14875.

67. Nativ, P.; Fridman-Bishop, N.; Gendel, Y., Ion transport and selectivity in thin film composite membranes in pressure-driven and electrochemical processes. *Journal of Membrane Science* 2019, 584, 46-55.

68. Fridman-Bishop, N.; Tankus, K. A.; Freger, V., Permeation mechanism and interplay between ions in nanofiltration. *Journal of Membrane Science* 2018, 548, 449-458.

69. Nativ, P.; Birnhack, L.; Lahav, O., DiaNanofiltration-based method for inexpensive and selective separation of Mg2+ and Ca2+ ions from seawater, for improving the quality of soft and desalinated waters. *Separation and Purification Technology* 2016, 166, 83-91.

70. Epsztein, R.; Shaulsky, E.; Dizge, N.; Warsinger, D. M.; Elimelech, M., Role of ionic charge density in Donnan exclusion of monovalent anions by nanofiltration. *Environmental Science & Technology* 2018, 52, (7), 4108-4116.

71. Rho, H.; Chon, K.; Cho, J., Surface charge characterization of nanofiltration membranes by potentiometric titrations and electrophoresis: Functionality vs. zeta potential. *Desalination* 2018, 427, 19-26.

72. Richards, L. A.; Schäfer, A. I.; Richards, B. S.; Corry, B., The importance of dehydration in determining ion transport in narrow pores. *Small* 2012, 8, (11), 1701-1709.

73. Kim, J.; Lee, S. E.; Seo, S.; Woo, J. Y.; Han, C.-S., Near-complete blocking of multivalent anions in graphene oxide membranes with tunable interlayer spacing from 3.7 to 8.0 angstrom. *Journal of Membrane Science* 2019, 592, 117394.

74. Al-Zoubi, H.; Omar, W., Rejection of salt mixtures from high saline by nanofiltration membranes. *Korean Journal of Chemical Engineering* 2009, 26, (3), 799-805.

75. Richards, L. A.; Richards, B. S.; Corry, B.; Schäfer, A. I., Experimental energy barriers to anions transporting through nanofiltration membranes. *Environmental Science & Technology* 2013, 47, (4), 1968-1976.

76. Epsztein, R.; DuChanois, R. M.; Ritt, C. L.; Noy, A.; Elimelech, M., Towards single-species selectivity of membranes with subnanometre pores. *Nature Nanotechnology* 2020, 15, (6), 426-436.

77. 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. *Journal of Membrane Science* 2022, 641, 119927.

# 6. Comprehensive characterizations of oil-field produced water treated by nanofiltration and reverse osmosis membranes[3]

## 6.1 Introduction

The technology development in horizontal drilling and hydraulic fracturing has increased unconventional oil and gas (UOG) production and enhanced energy security of the U.S.[1] The extraction of UOG requires an extensive amount of freshwater and produces a significant volume of wastewater.[2,3] The UOG produced water contains high concentrations of salinity and other hazardous chemicals,[2, 4] many of which cause health issues and adverse impacts on the environment.[3] Moreover, considering that many shale plays are located within or close to areas that suffer from severe water scarcity,[5-8] the intensive water usage of UOG production can lead to conflict against agricultural, municipal, and industrial water demands.[7] As a result, there is an urgent need of developing proper produced water management strategies to address the dual challenges of water scarcity and pollution associated with UOG production.

Deep-well injection is the current business-as-usual strategy of produced water management due to its maturity.[9] However, the cost of deep-well injection increases with the distance for the transport of produced water,[8] especially when the storage capacities of nearby wells are reached. Further, the disadvantages of deep-well injection, such as induced seismicity, restrained access to disposal wells, and the potential risk of contaminating groundwater resources,

---

[3] This chapter will be submitted as manuscript as

Jeong, N., Wiltse, M., Boyd, A., Blewett, T., Park, S., Broeckling, C., Borch, T., & Tong, T. (2023). Comprehensive characterizations of oil-field produced water treated by nanofiltration and reverse osmosis membranes. Submitted to *ACS ES&T Engineering*.

cast doubt on its reliability and sustainability.[10, 11] Internal reuse of UOG produced water for hydraulic fracturing has been applied to reduce the consumption of freshwater in some areas such as Marcellus, Barnett, and Fayetteville shale plays.[12, 13] However, as the volume of produced water exceeds the water demand for hydraulic fracturing in many shale plays,[12] external reuse of treated produced water to handle the large volume of produced water needs to be considered.

Membrane technologies such as nanofiltration (NF) and reverse osmosis (RO) have been applied to the treatment of UOG produced water for external beneficial reuse.[14-24] Due to the high salinity and fouling potential of produced water, the performance of membranes is hindered by membrane fouling.[2, 4] Existing studies pertaining to produced water treatment using NF and RO membranes have mainly focused on pretreatment,[15, 22] membrane fouling,[18, 20] and desalination performance[16, 21] to improve the efficiency of membrane treatment. Although several studies have characterized the chemical constituents of untreated produced water,[25-27] only a few studies provide chemical characterizations of produced water after membrane treatment.[19, 28] Regnery et al.[19] quantitatively analyzed semi-volatile aliphatic hydrocarbons and polycyclic aromatic hydrocarbons (PAH) in produced water treated by a hybrid forward osmosis-RO process using solid-phase extraction followed by gas chromatography-mass spectrometry (GC-MS). Also, Riley et al.[28] characterized organic matters in the produced water after a treatment train consisting of biological active filtration, ultrafiltration, and NF using multi-analytical methods, such as liquid chromatography – organic carbon detection (LC-OCD), liquid chromatography–high-resolution mass spectrometry (LC-HRMS), parallel factor analysis (PARAFAC), and gas chromatography–mass spectrometry (GC–MS).[28] The main focuses of the aforementioned works were to develop and evaluate analytical methods for quantifying organic pollutants. However, to the best of our knowledge, there is still a lack of comprehensive chemical analyses that reveal the efficiencies of

different technologies for produced water treatment comparatively (e.g., for NF and RO with different types of membranes). Further, the toxicity level of the treated produced water, which indicates potential ecological and health impacts, is of essential significance for risk assessment but has rarely been discussed in the literature pertaining to produced water treatment. Indeed, the toxic effects of untreated produced water have been demonstrated for various organisms,[25, 26, 29, 30] rendering toxicity a key indicator for ensuring sufficient treatment.

In this study, we compared the treatment efficacies of different technologies for the treatment of produced water sampled from the Niobrara shale play in Colorado. By performing detailed chemical and toxicological characterizations of both untreated and treated produced water, we evaluated the treatment efficacies of coagulation and microfiltration (MF, as pretreatment), as well as NF and RO equipped with membranes possessing varied permeability and selectivity. The efficiency of each technology to remove a diverse set of inorganic (6 cations, 4 anions, and boron) and organic (i.e., volatile compounds, total petroleum hydrocarbons, non-purgeable organic carbons, PAH, and surfactant) constituents were quantified. The toxicity of the produced water before and after treatment was tested by a biological assay using *Daphnia magna*. *Daphnia magna* has been shown to be one of the most sensitive species to UOG effluents,[31] which could serve as an ideal candidate to assess the efficacy of treatment technologies in addressing public health concerns for beneficial reuse of produced water. Based on the characterization results, we also discuss the transport of different chemical constituents across NF and RO membranes, as well as the effectiveness of NF and RO in mitigating toxicity of produced water. Additionally, the suitability of treated produced water for different beneficial reuse purposes, such as being applied for irrigation and livestock drinking water, is examined. Our efforts provide regulators, policy makers, engineers, and the UOG industry with important quantitative information for pollutant

removal efficiency and risk assessment associated with produced water treatment, representing a key step towards the shift of produced water management paradigm towards treatment and reuse.

## 6.2 Materials and methods

### *6.2.1 Materials, chemicals, and produced water*

Aluminum sulfate octadecahydrate (Al2(SO4)3·18H2O) was purchased from VWR BDH Chemicals (Radnor, PA). MF (grade 4) membranes were supplied from Whatman (Maidstone, UK). Commercial NF (NF270 and NF) and RO (XLE and BW30) membranes were purchased from DuPont (Wilmington, DE) and used for produced water treatment. In order to prevent the confusion between the commercial NF membrane (as a brand name) from DuPont and NF membrane as a type of membranes used in pressure-driven NF, we name the DuPont NF membrane as NFD membrane in this study. The produced water samples were collected from the Niobrara formation of the Denver-Julesburg basin, Colorado in December 2020. The produced water was stored at 4 °C until it was used for chemical analyses or treatment.

### *6.2.2 Produced water treatment*

Coagulation ($Al_2(SO_4)_3 \cdot 18H_2O$) was conducted by following the procedures reported by Rosenblum et al.[32] The produced water was transferred into jar test containers (Phipps & Bird 7790-901B, Richmond, VA) and placed in the fume hood until the produced water reached room temperature. To optimize the concentration of coagulants, 0–120 mg/L of $Al_2(SO_4)_3 \cdot 18H_2O$ were tested, and the optimal dose was found to be 40 mg/L, which resulted in a turbidity removal efficiency exceeding 87% while minimizing the use of coagulants. The solution was mixed at 300 revolutions per minute (rpm) for 1 minute, followed by a three-stage tapered flocculation procedure (55 rpm for 10 minutes, 35 rpm for 10 minutes, and 15 rpm for 10 minutes). After 30 minutes of settling, the supernatant was collected, and the turbidity of each sample was measured

with a Hach 2100N turbidimeter (Loveland, CO). Then, the collected supernatants were used for further treatment. For MF treatment, a MF membrane (diameter of 9 cm) was placed on a Buchner funnel on top of a suction flask. A small amount of deionized (DI) water was poured onto the MF membrane to ensure that the membrane attached to the funnel seamlessly. The suction flask was then connected to a vacuum pump using a rubber tube. With a vacuum applied, the MF filtrates were collected and used as the feedwater for NF and RO treatment.

The NF and RO treatment was performed by using a customized crossflow membrane filtration system (the effective membrane area was 20.02 $cm^2$).[33] Membrane coupons were stored in DI water for at least 24 hours prior to the experiments. The membranes were compacted with DI water under a hydraulic pressure of 300 psi until the permeate flux was stable. The temperature of feed water and the crossflow velocity were set to 23.0±0.5 °C and 1.0±0.1 L/min, respectively. After the compaction of membranes, the membranes were used for either membrane characterization (e.g., measuring the water and salt permeability) or the treatment of produced water. For the treatment of produced water, the initial permeate flux was set to 30.0±3.0 L/($m^2$ h). The permeates were collected for > 20 hours until the cumulative volume reached 1.6 L. Both the feed water and the permeate were collected after NF and RO membrane treatments for chemical analyses.

*6.2.3 Membrane characterization*

The pure water permeability and NaCl rejection (with a feed solution containing 20 mM NaCl) of the membranes were obtained after membrane compaction as described above. The pressure and crossflow velocity were maintained at 200±20 psi and 1.0±0.1 L/min for the measurements of the pure water permeability and NaCl rejection. The water contact angles of the membranes were measured by a contact angle goniometer (KRÜSS DSA10, Hamburg, Germany) with the sessile

drop method.[34] The zeta potentials of the NF and RO membranes were measured by using an electrokinetic analyzer (SurPASS, Anton-Paar, Ashland, VA), with 1 mM of KCl as the background electrolyte. The summary of main membrane properties is presented in Table 6-1.

**Table 6-1**. The properties of NF and RO membranes

| Membrane | Type | MWCO [Da] | Water permeability [L/(m$^2$ h bar)] | NaCl rejection [%] | Contact angle [°] | Zeta potential at pH 8 [mV] |
|---|---|---|---|---|---|---|
| NF270 | NF | 340[35] | 15.8 (± 1.8) | 24.8 (± 4.9) | 10 (± 1) | -104.7 |
| NFD | NF | 305[36] | 8.7 (± 0.7) | 66.3 (± 9.5) | 15 (± 1) | -64.3 |
| BW30 | RO | 100[37] | 5.9 (± 0.6) | 98.3 (± 1.3) | 61 (± 3) | -75 |
| XLE | RO | 96[37] | 7.4 (± 0.9) | 97.7 (± 0.6) | 52 (± 3) | -51.6 |

*6.2.4. Analytical methods*

6.2.4.1. Inorganic constituent analyses

Electrical conductivity (EC) and pH were measured by using an Accumet XL60 Dual Channel pH/Ion/Conductivity/DO meter (Fisher Scientific, Hampton, NH). Total dissolved solid (TDS) was calculated by converting the measured EC of each sample to the approximate concentration of TDS.[38] The concentrations of sodium ($Na^+$), potassium ($K^+$), magnesium ($Mg^{2+}$), calcium ($Ca^{2+}$), barium ($Ba^{2+}$), iron (Fe), and boron (B) were quantified using EPA-200.7 and EPA-6010B procedure by a certified laboratory (Technology Laboratory, Inc., Fort Collins, CO). Ion chromatography (IC, Dionex ICS-2100, Sunnyvale, CA) was used to quantify the concentrations of anions including chloride ($Cl^-$), nitrate ($NO_3^-$), bromide ($Br^-$), and sulfate ($SO_4^{2-}$). Each sample was filtered through a 0.22 μm polyethersulfone (PES) membrane filter before the measurement

with IC. We used a Dionex IonPac AS11-HC RFIC 4 x 250 mm analytical column (Sunnyvale, CA) for the analyses with a gradient anion method. The reported values for cation and anion concentrations were an average of three replicates.

6.2.4.2. Organic constituent analyses

Non-purgeable organic carbon (NPOC) and total nitrogen (TN) were measured using a Shimadzu TOC analyzer (Kyoto, Japan), with the quality control performed every 10 samples by checking that the standard was within 10% of the expected concentration of NPOC and TN. Total petroleum hydrocarbons (TPH), benzene, toluene, ethylbenzene, and total xylenes (BTEX) were analyzed using EPA-8015B and modified EPA-8260B procedures by a certified laboratory (Technology Laboratory, Inc., Fort Collins, CO), which performed stringent quality controls for TPH and BTEX measurement. To analyze the presence of organic surfactants, the produced water samples (both untreated and treated) were filtered through a 0.7 μm glass fiber filter and then a 0.22 μm PES membrane filter. Solid-phase extraction (SPE) was conducted by using Supel Select HLB cartridges (200mg/3mL, Supelco, Bellefonte, PA) to concentrate the surfactants and remove salts in 250 mL of samples. The samples were stored at -20°C before the analyses. The surfactants were detected using Agilent 1100 series liquid chromatograph (Santa Clara, CA) coupled with an Agilent G3250AA time-of-flight (Santa Clara, CA) mass spectrometer (LC/TOF/MS) equipped with ZORBAX RR Eclipse XDB-C8 (4.6 mm x 150 mm, 3.5 μm) columns (Agilent Technologies, Santa Clara, CA). The temperature was maintained at 25±1 °C, and the sample injection volume was 20 μL. Formic acid (0.1%) and acetonitrile (0.1%) were used as mobile phases A and B, respectively. The chromatographic method was developed with initial mobile phase composition (10% B) constant for 5 min and a linear gradient to 100% B after 30 min. The flow rate was set to 0.6 mL/min. After each measurement, a 10-min post-run time was used. The Kendrick mass defect

(KMD) bubble charts as a function of the mass to charge ratio (m/z) for surfactants were plotted by using Microsoft Excel (Microsoft, Redmond, WA). For PAH analysis, samples were prepared according to the sample preparation procedure from Dórea et al.[39] and measured by using atmospheric pressure gas chromatography (APGC) coupled with a Waters Xevo G2 quadrupole time-of-flight (QTOF) mass spectrometer (Milford, MA). One µl of samples was injected into an Agilent DB-5MS column (30m x 250µm x 0.25µm, Agilent Technologies, Santa Clara, CA). The chromatography temperature gradient was set to 18 °C/min and the final temperature was 330 °C resulting in a run time of 17.0±0.1 min. The inlet was split/splitless using a pulsed splitless mode. The pulse time was 0.50 min. The mass spectrometer was run in positive mode with 0.1 second of scan time at a collision energy of 4V. The analyzer was set to sensitivity mode with an API source. The scan range is 50-1200Da. Quality control was performed with an internal standard of nitrobenzene-d5, sample randomization, and method blanks being run after every 5 injections.

*6.2.5 Toxicity assessments*

6.2.5.1. Daphnia colony maintenance

Gravid *Daphnia magna* was purchased from Aquatic Research Organisms (USA, October 2021) and used to culture a colony within the University of Alberta Biological Science Department's aquatics facilities according to the guidelines outlined by the Organization for Economic Cooperation and Development.[40] Briefly, 30 daphnids were housed in 1 L of OECD water prepared by adding salts at 294 mg/L $CaCl_2$, 123 mg/L $MgSO_4$, 64.8 mg/L $NaHCO_3$, and 5.80 mg/L KCl to dechlorinated water from the City of Edmonton (pH ≈ 7.6).[35] The chemical characteristics of the dechlorinated City of Edmonton water has been previously described in Delompré et al.[41] Each colony was fed with a diet of 3 mL YCT (yeast, cereal leaf, trout chow) and 3 mL of $3 \times 10^7$ cells/mL freshwater green algae (*Raphidocelis subcapitata*) daily from Aquatic Research

Organisms and supplemented with 100 µL of Roti-Rich invertebrate food (VWR, Edmonton, Alberta, Canada) once per week. Complete water changes were performed every 2-3 days. The colonies and all experimental exposures were maintained on a 14 h light: 10 h dark photoperiod at 20.0±1.0 °C.

6.2.5.2. Acute toxicity assays

Acute 48-hour median lethal concentrations ($LC_{50}$) and effect concentrations ($EC_{50}$) were determined according to OECD guidelines.[35] An absence of gill respiratory motion when viewed under a dissecting microscope was used to determine mortality for $LC_{50}$ determination. Immobilization as defined by the inability for a daphnid to move over a period of 15 s under gentle agitation was used as the designated effect for $EC_{50}$ determination. For all experiments, 5 neonates <24 h old, which were housed in 20 mL of the appropriate dilution of experimental test solution, were used for each tested concentration and fasted for the duration of the experiment. A total of 6 replicates were performed for each test, scored by a researcher who was blind to the treatment identities. The OECD water without additions was utilized as the control to verify a lack of mortality in *Daphnia* not exposed to produced water. These exposures were housed in uncapped 20 mL glass scintillation vials, which were washed in 10% nitric acid (24 h), 10% ethanol (24 h), and thorough rinsing with DI water prior to experimentation.

6.2.5.3. Statistics

All LC50 and EC50 estimates were calculated using R version 3.6.2,[42] using the "ecotox" package to determine the median concentration as well as the 95% confidence intervals for each tested solution.

## 6.3 Results and discussion

### *6.3.1 The removal of inorganic constituents after treatment*

#### 6.3.1.1 The removal of cation constituents and boron

The pretreatments (i.e., coagulation and MF), NF, and RO were used to treat the produced water samples, and the efficiency of each treatment step in removing cations and B was evaluated (Figure 6-1). The detailed information on cation concentrations in samples collected from each step is provided in Table C1 (Appendix C). As presented in Figures 6-1A, 6-1B, 6-1H and Table C2 (Appendix C), minor changes in cation concentrations after coagulation were observed except for Fe and $Ba^{2+}$ whose removals were 96.9% and 79.9%, respectively. These cations are either a coagulating agent (Fe) that facilitates the formation of flocs or a reactant with sulfate to form precipitates ($BaSO_4$) during coagulation.[15, 43] After MF, 81.1% of Fe was rejected (Figure 6-1C and Table C2), which was likely due to the further removal of iron-containing flocs by MF. The rejections of $Ba^{2+}$ and B after MF were 15.2% and 16.5%, respectively, which were probably because of removals of precipitants (i.e., barium sulfate[15] and calcium borate[44, 45]) after their reactions with sulfate and calcium, respectively.

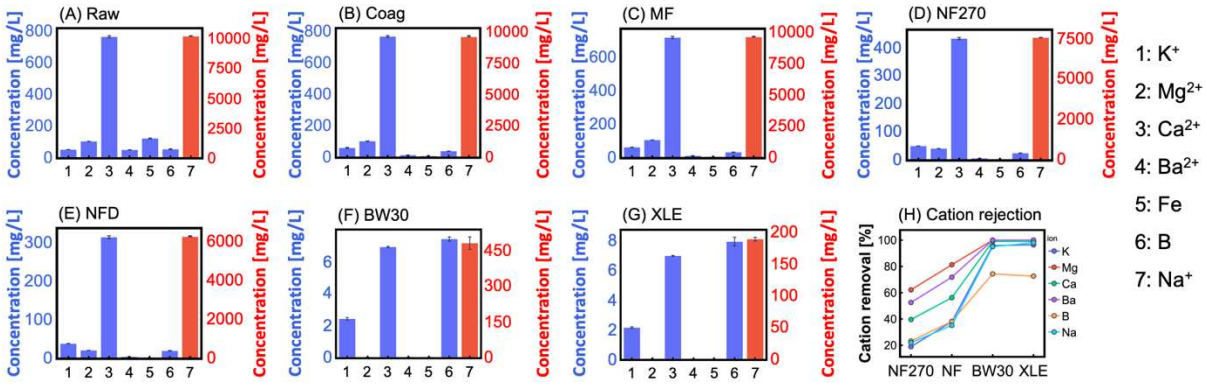


**Figure 6-1.** The concentrations of potassium ($K^+$), magnesium ($Mg^{2+}$), calcium ($Ca^{2+}$), barium ($Ba^{2+}$), iron (Fe), boron (B), and sodium ($Na^+$) in (A) raw produced water, samples after (B) coagulation (coag), (C) microfiltration (MF), coagulation and microfiltration followed by nanofiltration using (D) NF270

membrane and (E) NFD membrane, reverse osmosis using (F) BW30 membrane and (G) XLE membrane. (H) Cation rejections of different types of membranes. The error bars represent the standard deviations calculated from three replicates. As Fe could be in the form of $Fe^{2+}$ or $Fe^{3+}$, no valence is indicated for Fe.

For NF270 membrane that is typically considered as a loose membrane for NF treatment, the rejections of multivalent cations were higher than those of monovalent cations. The range of multivalent cation ($Mg^{2+}$, $Ca^{2+}$, $Ba^{2+}$, and Fe) rejections was 39.6%–100%, while the rejections of $Na^+$ and $K^+$ were 22.0% and 18.9%, respectively (Figures 6-1D, 6-1H and Table C2). For comparison, the rejections of multivalent cations by NFD membrane ranged from 56.2% to 100%, while those of $Na^+$ and $K^+$ were 35.2% and 38.0%, respectively (Figures 6-1E, 6-1H and Table C2). Due to the lower MWCO of NFD membrane than NF270 membrane, higher rejections of cations were observed for NFD membrane, whereas water permeability of NF270 membrane was 1.8 times higher than NFD membrane (Table 6-1). Such an observation was consistent with the trade-off between permeability and selectivity of polyamide membranes.[46] For both NF270 and NFD membranes, the lowest rejection was observed for boron (the rejections of boron by NF270 and NFD membranes were 23.3% and 38.2%, respectively). As shown in Table C1, the boron concentrations in the treated produced water by NF270 and NFD membranes were 22.1 mg/L and 17.8 mg/L, respectively, which are well above the threshold of boron content recommended for irrigation (<0.5–5 mg/L depending on the crop types[47]) and livestock drinking water quality (<5 mg/L).[47] Meanwhile, the EC of the permeates from NF270 (41.3 mS/cm) and NFD (30.7 mS/cm) membranes are classified as not recommended for livestock drinking water (> 5 mS/cm)[48] and subject to severe restriction on irrigation water use (> 3 mS/cm)[47] (Table C1). This shows that the NF treatment using NF270 and NFD membranes was not appropriate for directly producing water suitable for irrigation or livestock drinking purposes. To meet the requirement for those beneficial reuse purposes, the NF270 and NFD permeates can be potentially diluted with freshwater. When

the NF270 and NFD permeates are diluted by 10 to 20 times with freshwater, their concentrations of inorganic constituents meet the salinity and cation concentration requirements for irrigation and livestock drinking water. For example, after diluting NF270 and NFD permeates by 14 and 11 times with freshwater, the EC of the permeates are 2.95 mS/cm and 2.79 mS/cm, respectively, which meet the water quality requirements of EC for both irrigation and livestock drinking water.[47, 48] Moreover, the boron concentration of NF270 and NFD permeates after the same dilution factors would be both ~1.6 mg/L, which is below the requirements for irrigating moderately boron-tolerant (2–4 mg/L) crops and livestock drinking water.[47, 48] Another alternative of improving the quality of the treated produced water is to apply RO after NF treatment.[49-51] If so, these NF membranes can be used to remove multivalent cations (e.g., $Mg^{2+}$, $Ca^{2+}$, and $Ba^{2+}$) and alleviate mineral scale formation on RO membrane surfaces, improving the membrane performance and reducing operation and maintenance cost for the membrane treatment train.

For the two RO membranes (i.e., BW30 and XLE membranes), the cation rejections were generally greater than 94% (Figures 6-1F, 6-1G, 6-1H, and Table C2). However, the boron rejections of BW30 and XLE membranes were only 72-74%, which are similar to boron rejection of 70%-80% reported in the literature.[52, 53] The relatively low rejection of boron was because of the uncharged and poorly hydrated nature of boron molecules at near neutral pH (the pKa of boric acid is ~9.4),[54] making it challenge to be removed by RO membranes.[55] Steric exclusion is not enough to effectively remove boron at the given pH of produced water (pH of 7.9). Boron concentrations in BW30 and XLE permeates were 7.4 and 7.9 mg/L, respectively, which are appropriate for boron-tolerant crops such as asparagus and cotton (6-15 mg/L).[47] However, multi-stage RO treatments or selective boron removal (e.g., using adsorption[56, 57]) would be necessary if

the treated produced water is used for irrigating agricultural crops that are sensitive to boron (< 1 mg/L, such as for lemon and blackberry[58, 59]) as well as potable water for livestock (< 5 mg/L).[48]

Furthermore, the EC of produced water samples after treatment with BW30 and XLE membranes were 1.6 and 1.4 mS/cm, which satisfy the livestock drinking water quality (<5 mS/cm) and irrigation water quality (<3 mS/cm).[47, 48] However, the sodium adsorption ratio (SAR) of RO permeates (50.0 meq/L and 19.7 meq/L for permeates from BW30 and XLE membranes, respectively) are higher than or very close to the recommended water quality for irrigation (SAR < 20 meq/L). The higher rejections of $Ca^{2+}$ (99%) and $Mg^{2+}$ (100%) than that of $Na^{+}$ (95%–98%) by BW30 and XLE membranes, along with the higher $Na^{+}$ concentration (9,640 mg/L) compared to $Ca^{2+}$ (715 mg/L) and $Mg^{2+}$ (102 mg/L) concentrations in the MF filtrates, result in the high SAR values for BW30 and XLE permeates. The high SAR values lead to low infiltration rates of water in the soil, causing water unable to reach the depth of the soil where the plants obtain water.[47, 60] To reduce the SAR when using the RO permeate for irrigation, it is necessary to supplement calcium directly (e.g., by adding gypsum) or indirectly (e.g., by adding acid that dissolves calcium from lime in the soil) in order to meet the irrigation water quality.[47, 60, 61]

6.3.1.2 The removal of anion constituents

The anion concentrations in the produced water samples before and after coagulation, MF, and membrane treatments, as well as the anion rejections by different membranes are presented in Figure 6-2. Coagulation did not lead to significant removal of monovalent anions (Figures 6-2A and 6-2B), with 3.7–6.6% of removal rates observed (Table C2). Such moderate removal rates might be due to adsorption of anions to precipitates formed during coagulation.[62-65] Also, the concentration of sulfate increased from 61.4 mg/L to 68.9 mg/L after coagulation, probably due to the addition of sulfate coagulant (Al2(SO4)3 · 18H2O) to the raw produced water (Table C1).

Additionally, slight rejections (<4%) of anions were observed by MF (Figures 6-2B and 6-2C, Table C2), consistent with the pore size of MF membrane that well exceeds the sizes of hydrated anions.[66]

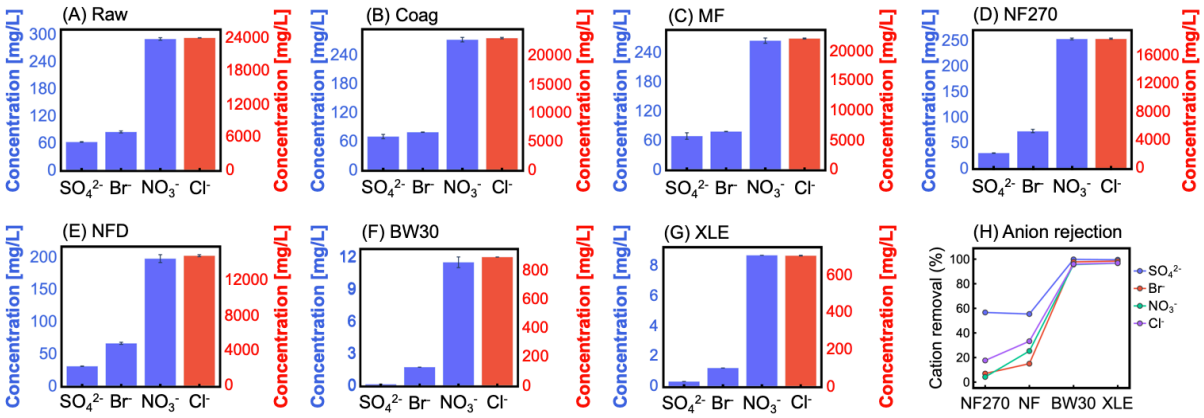


**Figure 6-2.** The concentrations of sulfate ($SO_4^{2-}$), bromide ($Br^-$), nitrate ($NO_3^-$), and chloride ($Cl^-$) in (A) raw produced water, samples after (B) coagulation (coag), (C) microfiltration (MF), coagulation and microfiltration followed by nanofiltration using (D) NF270 membrane and (E) NFD membrane, reverse osmosis using (F) BW30 membrane and (G) XLE membrane. (H) Anion rejections of different types of membranes. The error bars represent the standard deviations calculated from three replicates.

Among the NF and RO membranes tested in the current study, NF270 membrane showed the lowest rejections of monovalent anions (4.2%–17.6%, Figure 6-2H and Table C2), consistent with its highest MWCO (Table 6-1). However, NF270 membrane still possessed high $SO_4^{2-}$ rejection (~60%, Figures 6-2D and 6-2H). The higher rejection of $SO_4^{2-}$ compared to those of monovalent anions by NF270 membrane stems from the stronger electrostatic repulsion between the divalent $SO_4^{2-}$ ions and the negative charged membrane surface (Table 6-1).[67, 68] Also, the hydrated radius and hydration energy of $SO_4^{2-}$ are much greater than those of monovalent anions,[69, 70] resulting in a higher extent of steric exclusion.[71]

The anion rejections by both RO membranes (i.e., BW30 and XLE membranes) were greater than 95% for all anions. The smaller MWCO (and also the smaller pore sizes) of RO membranes than those of NF membranes results in a higher degree of steric exclusion and consequently, anion rejections. We further examined the order of anion rejection to understand the factors that regulate anion selectivity of the NF and RO membranes. For RO membranes, the anion rejection was better correlated with ionic radius ($R^2$=0.92-0.97 using linear regression) than with hydrated radius ($R^2$=0.80-0.87) and hydration energy of the ions ($R^2$=0.67-0.76), as shown in Figures C1-C3 (Appendix C). However, the order of anion rejection by NF270 membrane (the loosest NF membrane) follows $SO_4^{2-}$ > $Cl^-$ > $Br^-$ > $NO_3^-$, which is negatively correlated with hydration energy (i.e., higher rejection of anion with stronger hydration energy, Figure C1A). Such anion rejection trends might be explained by the transition-state theory (TST), according to which ion transport across membranes is regulated by entropic (e.g., steric exclusion) and enthalpic (e.g., ion dehydration) energy barriers.[72] Generally, an increase of enthalpy energy barrier leads to a decrease of entropy energy barrier, and vice versa.[73] As shown by Shefer et al.,[73] the anion transport within loose NF membranes are highly influenced by the enthalpic energy barrier, suggesting higher energy penalties for anions with stronger hydration energy, whereas the influence of entropic energy barrier becomes higher when membrane pore sizes decrease.[74] This indicates that when ions are transporting across membranes with low MWCO (e.g., RO membranes), the ions with smaller ionic radii have more opportunities to enter membrane pores than larger bare ions,[73] enhancing the permeability of entropically favorable anions (i.e., anions with small ionic radii). This explains the anion selectivity for RO membranes observed in this study.

In addition, the $Cl^-$ concentrations after membrane treatment with RO membranes (891.1 mg/L and 701.2 mg/L for permeates from BW30 and XLE membranes) are still higher than the allowable limits of $Cl^-$ concentration for irrigation (<354.5 mg/L),[47] requiring two-stage RO treatments or 2-3 times dilution with freshwater to meet the criteria (Table C1). Also, the livestock drinking water quality requires the total concentrations of nitrate and nitrite to be less than 100 mg/L.[48] All the permeates from NF and RO membranes meet this requirement (26.8 mg-N/L, 22.9 mg-N/L, 4.0 mg-N/L, and 5.3 mg-N/L for NF270, NFD, BW30, and XLE permeates, respectively, Table C1).

*6.3.2 The removal of organic constituents after treatment*

6.3.2.1 The removal of total xylenes, TPH, NPOC, and PAH

The concentrations of total xylenes, TPH, NPOC, and total PAH in the produced water before and after treatment are presented in Figure 6-3. For BTEX, only the concentration of total xylenes was measurable in the raw produced water, whose concentrations of benzene, toluene, and ethylbenzene were lower than the detection limit (i.e., <0.001 mg/L, Table C3, Supporting Information). This might be due to the volatilization of benzene, toluene, and ethylbenzene during sample transport and storage. The concentration of total xylenes in the raw produced water was 0.24 mg/L, and it was effectively removed after coagulation (79.2% removal) and MF (86.0% removal, Figure 6-3A and Table C3). The residual concentration of total xylenes was lower than the detection limit after NF or RO treatment (Figures 6-3A, 6-3E, and Table C3).

As shown in Figure 6-3B and Table C3, TPH was removed by 58.9% and 98.5% after coagulation and MF, respectively, with the residual TPH concentration after pretreatment at 5.2 mg/L for the MF permeate. The order of TPH rejections by membranes was BW30 (90.4%) > NF270 (82.7%) > NFD (76.9%) > XLE (59.6%), which was not consistent with the ascending

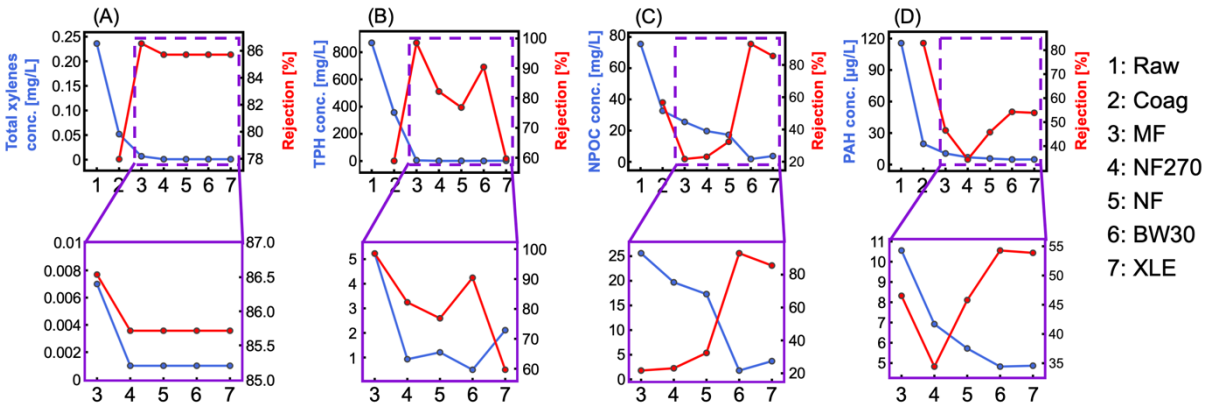


**Figure 6-3.** The concentrations of (A) total xylenes, (B) total petroleum hydrocarbons (TPH), (C) non-purgeable organic carbon (NPOC), and (D) total polycyclic aromatic hydrocarbons (PAH) in raw produced water, samples after coagulation (coag), microfiltration (MF), the treatment with NF270, NFD, BW30, and XLE membranes. The concentrations (left y-axis) and the corresponding rejections (right y-axis) of total xylenes, TPH, NPOC, and PAH after each treatment step are displayed in blue and red lines, respectively.

order of MWCO of the membranes (i.e., XLE < BW30 < NFD < NF270, Table 6-1). We measured the TPH concentration for the XLE permeate multiple times and confirmed that the TPH concentration (2.1 mg/L) of XLE permeate was higher than those of NF270, NFD, and BW30 permeates (Figure 6-3F). We acknowledge that the reason for the inconsistency between TPH removal efficiency and MWCO of the membranes is still unknown, and future studies need to be performed to further understand the regulating factors for membrane removal of TPH. Also, the NPOC removal rates by coagulation and MF were 56.7% and 21.7%, respectively (Figure 6-3C and Table C3). The rejections of NPOC by RO membranes were higher than 85%, while the NF membranes rejected up to around 33% of NPOC. There is no specific guideline pertaining to TPH and NPOC concentrations for irrigation and livestock drinking water.[47, 48] However, the influences of TPH on the crop have been reported in the literature. The TPH present in soil forms a thin film around crop seeds and prevents oxygen intake, resulting in mortality or reduced emergence of crop embryo.[75]

Further, the concentration of total PAH was measured by AP/GC-QTOF, as presented in Figure 6-3D and Table C3. The total PAH rejections increased as the MWCO of the membranes decreased. After 84.6% and 44.1% of total PAH removal by coagulation and MF, the total PAH rejections by NF or RO membranes were 37.4%, 48.5%, 57.6%, and 57.6% for NF270, NFD, BW30, and XLE membranes, respectively (Table C3). Such PAH rejection efficiencies by NF or RO membranes were similar to or lower than those reported in the literature (the average PAH rejection by RO membrane was 90.5 % in the study by Gong et al.,[76] while Smol et al.[77] reported that the average PAH rejection by RO membrane was 64.1%). The incomplete removal of PAH, even by RO membranes, could be attributed to the saturated adsorption sites of membranes after collecting permeates for more than 20 hours of membrane treatment. As shown by Li et al.,[78] adsorption is the dominant mechanism for PAH rejection by polyamide membranes at the initial stage of filtration. After this stage, the rejection could decrease significantly once the adsorption capacity is reached, and the main PAH rejection mechanisms is size exclusion.[78] The concentrations of total PAHs in the NF and RO permeates decreased with the MWCO of membranes, supporting that the contribution of size exclusion to PAH rejection. Similar to TPH and NPOC, there is no guideline pertaining to PAH for irrigation and livestock drinking water.

6.3.2.2 Surfactant analysis in the treated produced water

By using LC/TOF/MS, we identified five surfactant classes in the untreated and treated produced water, including C11 alcohol ethoxylate (C11-EO), C12 alcohol ethoxylate (C12-EO), C18 alcohol ethoxylate (C18-EO), polyethylene glycol (PEG), and polyethylene glycol-carboxylates (PEG-COOH), which are commonly present in the produced water.[79-81] The semi-quantitative analysis was performed for those surfactants because of the difficulties of quantification of each surfactant as explained below. In the presence of various matrix components, the coelutions of analytes

disrupt quantitative analysis of surfactants, leading to suppression of signals in LC/MS.[82] Depending on the chemical compositions of samples, ion enhancement or ion suppression may occur due to the competition for ionization between target analytes and matrix components.[83] The ionization state of the surfactants may also vary according to the salinity of the produced water samples. In addition, the quantification of each surfactant requires standards for all the detected surfactants with various molecular weights in the solution, which are difficult to obtain (if not impossible). The Kendrick mass defect (KMD) analysis was used to identify and visualize the homologous series of surfactants (Figure 6-4). It is worth mentioning that the sizes of bubbles represent the relative abundance in the produced water sample rather than the concentrations of surfactants.

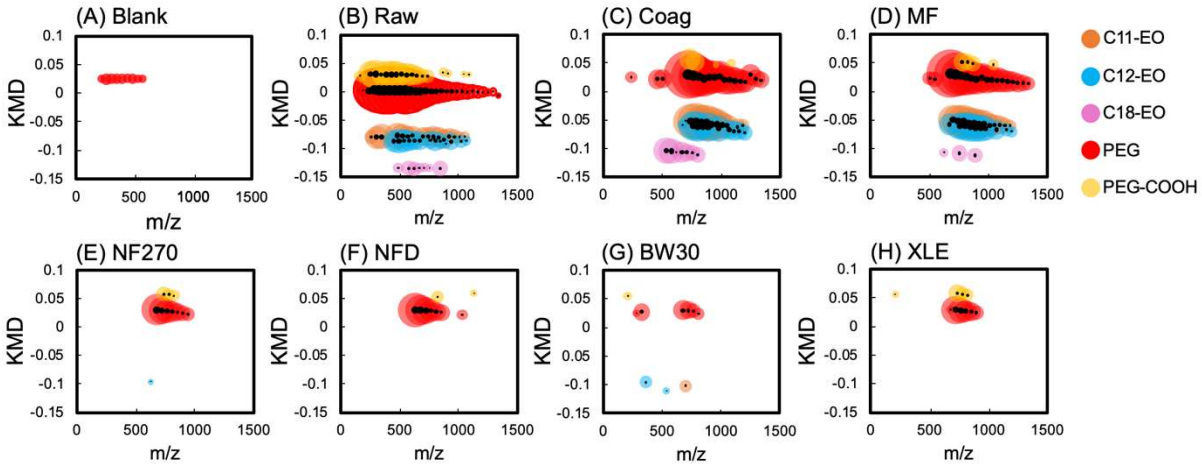


**Figure 6-4.** Identification of the specific surfactant species in (A) blank, (B) raw produced water, samples after (C) coagulation (coag), (D) microfiltration (MF), coagulation and microfiltration followed by nanofiltration using (E) NF270 membrane and (F) NFD membrane, reverse osmosis using (G) BW30 membrane and (H) XLE membrane. The areas of bubbles indicate the relative abundance of surfactants. The orange, blue, pink, red, and yellow bubbles represent C11 alcohol ethoxylate (C11-EO), C12 alcohol ethoxylate (C12-EO), C18 alcohol ethoxylate (C18-EO), polyethylene glycol (PEG), and polyethylene glycol-carboxylates (PEG-COOH).

Method blank was conducted to identify residual surfactants in the membrane system. The membrane system was run with DI water for 10 min, and then the permeate sample was collected for the analysis. As displayed in Figure 6-4A, PEGs with mass-to-charge ratios (m/z) of 210–570 were detected. Such residual PEGs were likely due to the ubiquitous presence of surfactants in the environment[84, 85] and were effectively removed by NF and RO membrane treatment (Figures 6-4 E-H; thus, they did not affect the surfactant analysis for NF and RO permeates).

After coagulation, C11-EO, C12-EO, PEG, and PEG-COOH with low molecular weight (MW < 700 Da) were mostly removed, and surfactants with high molecular weight (MW > 700 Da) still existed in the produced water samples (Figures 6-4 B and C). However, this phenomenon was different from the results reported by Beltrán-Heredia et al.,[86] who discovered higher removals for long-chain surfactants than short-chain surfactants after coagulation. One possibility is that compared to short-chain surfactants, the long-chain surfactants were more strongly adsorbed on the surface of petroleum hydrocarbons and remained with petroleum hydrocarbons as coagulation could not effectively remove TPH (TPH removal rate after coagulation was only 58.9%, Table C3). The removal rate of PEG-COOH was higher than PEG, which might be due to the removal of anionic surfactant (PEG-COOH) with the positively charged coagulant.[87-89] Compared to produced water after coagulation, there was only a minor change in the types of surfactants in MF filtrates (Figure 6-4D) as the pore sizes (20–25 μm as reported by the manufacturer) of MF membrane are larger than the size of surfactants.

As demonstrated in Figures 6-4E and 6-4F, NF270 and NFD membranes removed most of C11-EO, C12-EO, and C18-EO, as well as PEG with high m/z (>800 m/z). Interestingly, despite the low MWCO of RO membranes (~100 Da), some surfactants, such as C11-EO (m/z of 700), C12-EO (m/z of 362 and 538), and PEG-COOH (m/z of 208) that were not observed from the

produced water sample after NF270 and NFD treatment, were detected in the BW30 and XLE permeates (Figures 6-4 E-H). This could be attributed to the different ionization efficiencies of surfactants between the NF and RO permeates. As shown in Figures 6-1H, 6-2H, and Table C3, the permeates of BW30 and XLE membranes contained less organic and inorganic constituents than those after treatment using NF270 and NFD membranes, resulting in less competition between the target analytes and matrix components for ionization[83] and consequent detections of more surfactants. However, it is still counterintuitive to observe surfactants whose MWs are higher than the MWCO of membranes in RO filtrates. This could be due to the defects in the active layers of polyamide membranes. Song et al.[90] studied the presence of defects in commercial polyamide membranes such as BW30 and NF90 membranes, as well as the formation mechanisms of defective nodular layers on membrane surfaces. They demonstrated that forceful degassing during interfacial polymerization could lead to the formation of defects within the polyamide active layer. Such defects can cause the passage of organic and inorganic constituents that are larger than average membrane pore sizes, hampering the membrane performance.[90, 91] It is worth mentioning that our results only demonstrate the presence, rather than the concentrations, of the surfactants in NF and RO permeates. Therefore, we are not able to evaluate the associated environmental and health risks. Even though, our data suggest that membrane treatment should not be considered as an absolute barrier for organic contaminants such as surfactants. Future research is needed to understand the long-term effects of organic contaminants at trace levels in NF and RO permeates, as well as to develop novel membrane fabrication techniques that reduce the number of defects in the polyamide layer.

*6.3.3 The toxicity level after treatments*

The toxicity level of produced water after each treatment step was evaluated to investigate the potential impact of treated produced water on the aquatic environment by measuring LC50 and EC50 for Daphnia. The 48-h LC50 of raw produced water was 5.2%, which is similar to previously tested UOG produced water samples from four Montney formation wells in British Columbia, Canada (7.4%–13.6%) with the same D. magna colony.[26] A progressive decrease of toxicity was observed as the produced water was treated with coagulation (12.6%) and MF (15.3%), as shown in Figures 6-5 A-C. The treatment with NF membranes led to an increase of LC50 to 19.5%–24.3% (Figures 6-5D and 6-5E). The largest decreases in toxicity were observed in the BW30 permeate, where <40% mortality occurred with undiluted sample (estimated LC50 = 115.0%), and the XLE permeate, where no mortality was observed at any dilution (Figures 6-5F and 6-5G).

Blewett et al. has demonstrated that the acute toxicity of UOG produced water to Daphnia was largely determined by salt contents.[92] The corresponding decreases in toxicity across treated produced water samples closely correlate with decreasing salt concentrations, with the least toxic BW30 and XLE permeates having salinity comparable to freshwater.[93] As shown in Figure C4A (Appendix C), the 48-h LC50 showed a negative correlation with the salinity. Given that NaCl was the dominant compound in the untreated and treated produced water, we attributed the main driver of the toxicity of the treated produced water to salinity. Moreover, the concentrations of total xylenes and PAHs in the raw sample are both several orders of magnitude below their respective 48-h LC50 values,[94-96] indicating that minimal toxicity is expected to occur due to these organic fractions of the samples.

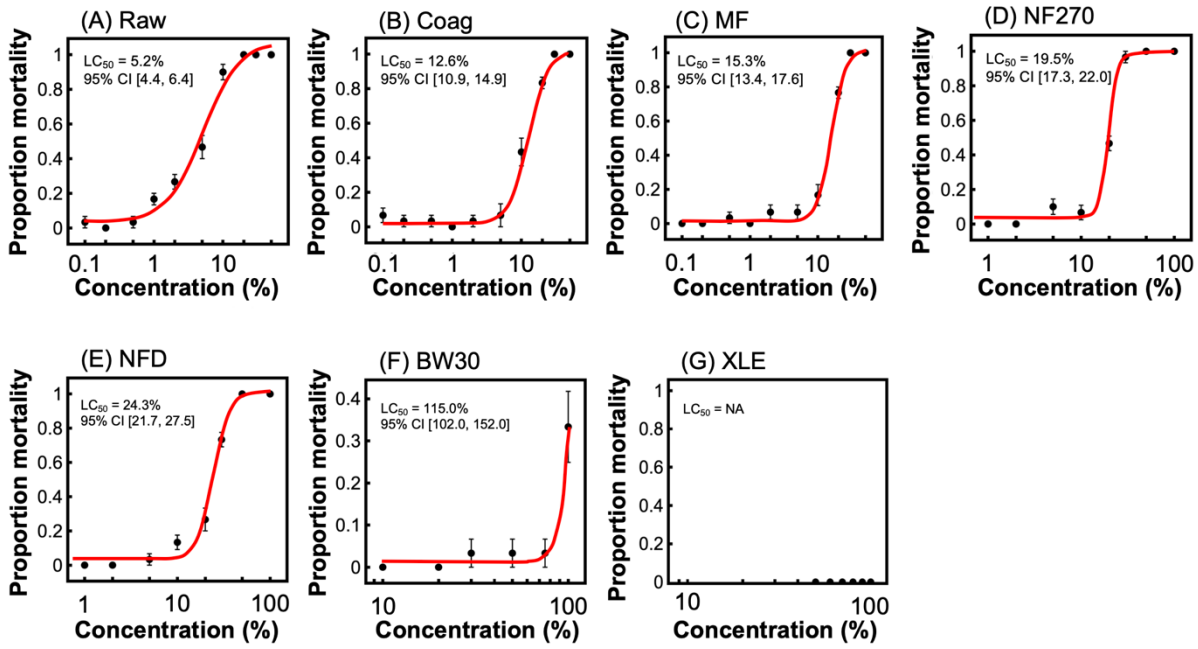


**Figure 6-5.** The 48-h median lethal concentration ($LC_{50}$) for *Daphnia* after exposing to (A) raw water, samples after (B) coagulation (coag), (C) microfiltration (MF), coagulation and microfiltration followed by nanofiltration using (D) NF270 membrane and (E) NFD membrane, reverse osmosis using (F) BW30 membrane and (G) XLE membrane. The concentration indicates the fraction of each treated produced water for $LC_{50}$ analysis.

The estimated EC50 values followed a similar pattern as the LC50 values, with membranes with lower MWCO resulting in less toxic samples (Figure 6-6). All EC50 values were approximately 5–10% lower than their corresponding LC50 concentrations, ranging from 0.5% for the raw produced water sample, 1.2–6.0% for samples after coagulation and MF, and 13.9–18.2% for NF permeates. Although the EC50 of the produced water after MF was slightly higher than coagulation sample, such a difference was minor, meaning that the toxicity levels of the samples after coagulation and coagulation-MF were similar. The undiluted permeate of BW30 membrane resulted in approximately 50% of immobilized daphnids (estimated LC50 = 103.0%), while the undiluted permeate of XLE membrane immobilized < 10% of daphnids, indicating very minimal toxicity.

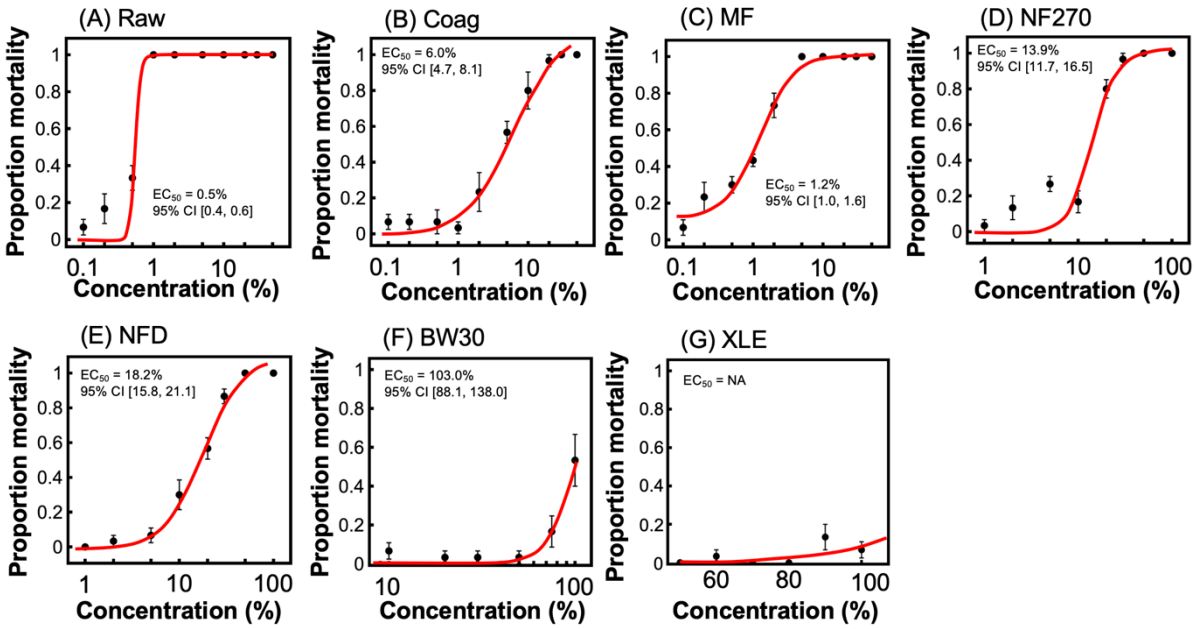


**Figure 6-6.** The 48-h median effect concentration causing immobilization ($EC_{50}$) for *Daphnia* after exposing to (A) raw water, samples after (B) coagulation (coag), (C) microfiltration (MF), coagulation and microfiltration followed by nanofiltration using (D) NF270 membrane and (E) NFD membrane, reverse osmosis using (F) BW30 membrane and (G) XLE membrane. The concentration indicates the fraction of each treated produced water for $EC_{50}$ analysis.

Immobilization represents a more ecologically sensitive measure of toxicity than mortality. The immobilized daphnids are less capable of feeding, may not be able to molt (impairing growth), and are more vulnerable to predation due to their lack of mobility.[97, 98] As shown in Figure 6-6F, the BW30 permeate immobilized ~50% population of Daphnia. The salinity level of the BW30 permeate (~1,200 mg/L of TDS, Table C1) was close to the NaCl concentration (1,600 mg/L) that immobilized 50% of Daphnia reported by Okamoto et al.[99] The XLE permeate (~1050 mg/L TDS) was the only sample that exhibited no toxicity for Daphnia, which was consistent with its lowest TDS concentration among the treated produced water samples. As shown in Figure C4B, an inverse correlation between EC50 and TDS was obtained, supporting that salinity plays an important role in determining the EC50 of Daphnia. The presence of organic species is also a potential concern; however, we were unable to make explicit conclusions on their contribution to

toxicity. Although we could not quantify the surfactants and investigate their influences on the toxicity of the treated produced water, the effects of surfactants on the Daphnia have been reported by Blewett et al.[97] who suggested that the immobilization of D. magna exposed to UOG effluents could be largely driven by surfactants, and to a lesser degree, other organic substances. The toxicity induced by surfactant are largely indirect; by reducing the surface tension of the water, daphnids may become trapped at the surface indefinitely, frequently succumbing to mortality.[97] Also, the influence of other organic compounds such as PAH could not be excluded, and future work needs to be conducted to understand the impacts of specific organic compounds in the raw and treated produced water on the toxicity level of the water samples.

### 6.4 Conclusions

We assessed the performance of pretreatment (including coagulation and MF) and different NF/RO membranes with varied permeability and selectivity for treating UOG produced water from the Niobrara shale play in Colorado. Our comprehensive investigation revealed detailed compositions of organic and inorganic constituents, as well as the toxicity level of the treated produced water. The results demonstrate that pretreatment only resulted in minor reduction of chemical constituents and toxicity level of the produced water. For NF treatment, the permeates of both NF270 and NFD membranes (representative of loose and tight NF membranes) did not meet the water quality requirements for irrigation and livestock drinking water in terms of salinity level and boron concentration, and their ecological risks could not be neglected. Although RO membranes were able to remove a majority of pollutants from the produced water, the $Cl^-$ concentrations of the RO permeates were higher than what are regulated in the guideline for irrigation water.[47] Moreover, the SAR values of the RO permeates were higher than or close to the recommended SAR for irrigation water.[47] The relatively low removal efficiency of boron was a

major problem that plagued both NF and RO treatment. Even after RO treatment, the treated produced water contained high boron contents that could impose toxicity to crops. This finding emphasizes the importance of developing novel membranes with improved boron removal efficiency[100, 101] or other selective boron removal technologies[102, 103] if industrial wastewater such as UOG produced water is treated for irrigation applications. In addition, we discovered that surfactants with molecular weights much higher than the MWCO of NF and RO membranes were able to escape into the permeates. Although the environmental and health implications associated with trace concentrations of surfactants are still unknown, our results suggest that membranes should not be considered as an absolute barrier for contaminant removal. Future work that investigates the ecological and health impacts of NF/RO permeates more thoroughly and fabricates NF/RO membranes with less defects is needed to enhance the reliability and reduce uncertainty associated with produced water treatment and reuse using membrane processes. The comprehensive analysis of treated produced water in this study provide stakeholders (e.g., regulators, policy makers, wastewater treatment practitioners, and the UOG industry) with quantitative information on the feasibility and potential risk of reusing UOG produced water for agricultural applications.

**References**


1. Tong, T.; Carlson, K. H.; Robbins, C. A.; Zhang, Z.; Du, X., Membrane-based treatment of shale oil and gas wastewater: The current state of knowledge. *Frontiers of Environmental Science & Engineering* 2019, 13, (4), 63.
2. Sun, Y.; Wu, M.; Tong, T.; Liu, P.; Tang, P.; Gan, Z.; Yang, P.; He, Q.; Liu, B., Organic compounds in Weiyuan shale gas produced water: identification, detection and rejection by ultrafiltration-reverse osmosis processes. *Chemical Engineering Journal* 2021, 412, 128699.
3. Stringfellow, W. T.; Domen, J. K.; Camarillo, M. K.; Sandelin, W. L.; Borglin, S., Physical, chemical, and biological characteristics of compounds used in hydraulic fracturing. *Journal of Hazardous Materials* 2014, 275, 37-54.
4. Chang, H.; Li, T.; Liu, B.; Vidic, R. D.; Elimelech, M.; Crittenden, J. C., Potential and implemented membrane-based technologies for the treatment and reuse of flowback and produced water from shale gas and oil plays: A review. *Desalination* 2019, 455, 34-57.
5. Nicot, J.-P.; Scanlon, B. R., Water use for shale-gas production in Texas, US. *Environmental science & technology* 2012, 46, (6), 3580-3586.
6. Scanlon, B. R.; Reedy, R. C.; Nicot, J. P., Will water scarcity in semiarid regions limit hydraulic fracturing of shale plays? *Environmental Research Letters* 2014, 9, (12), 124011.
7. Du, X.; Carlson, K. H.; Tong, T., The water footprint of hydraulic fracturing under different hydroclimate conditions in the Central and Western United States. *Science of The Total Environment* 2022, 840, 156651.
8. Robbins, C. A.; Du, X.; Bradley, T. H.; Quinn, J. C.; Bandhauer, T. M.; Conrad, S. A.; Carlson, K. H.; Tong, T., Beyond treatment technology: Understanding motivations and barriers for wastewater treatment and reuse in unconventional energy production. *Resources, Conservation and Recycling* 2022, 177, 106011.
9. Conrad, C. L.; Ben Yin, Y.; Hanna, T.; Atkinson, A. J.; Alvarez, P. J. J.; Tekavec, T. N.; Reynolds, M. A.; Wong, M. S., Fit-for-purpose treatment goals for produced waters in shale oil and gas fields. *Water Research* 2020, 173, 115467.
10. Ellsworth, W. L., Injection-Induced Earthquakes. *Science* 2013, 341, (6142), 1225942.
11. Gregory, K.; Mohan, A. M., Current perspective on produced water management challenges during hydraulic fracturing for oil and gas recovery. Environmental chemistry 2015, 12, (3), 261-266.
12. Scanlon, B. R.; Ikonnikova, S.; Yang, Q.; Reedy, R. C., Will water issues constrain oil and gas production in the United States? *Environmental Science & Technology* 2020, 54, (6), 3510-3519.
13. Clark, C. E.; Horner, R. M.; Harto, C. B., Life cycle water consumption for shale gas and conventional natural gas. *Environmental science & technology* 2013, 47, (20), 11829-11836.
14. Riley, S. M.; Oliveira, J. M. S.; Regnery, J.; Cath, T. Y., Hybrid membrane bio-systems for sustainable treatment of oil and gas produced water and fracturing flowback water. *Separation and Purification Technology* 2016, 171, 297-311.
15. Chang, H.; Liu, B.; Yang, B.; Yang, X.; Guo, C.; He, Q.; Liang, S.; Chen, S.; Yang, P., An integrated coagulation-ultrafiltration-nanofiltration process for internal reuse of shale gas flowback and produced water. *Separation and Purification Technology* 2019, 211, 310-321.

16. Riley, S. M.; Ahoor, D. C.; Oetjen, K.; Cath, T. Y., Closed circuit desalination of O&G produced water: An evaluation of NF/RO performance and integrity. *Desalination* 2018, 442, 51-61.
17. Tang, P.; Liu, B.; Zhang, Y.; Chang, H.; Zhou, P.; Feng, M.; Sharma, V. K., Sustainable reuse of shale gas wastewater by pre-ozonation with ultrafiltration-reverse osmosis. *Chemical Engineering Journal* 2020, 392, 123743.
18. Miller, D. J.; Huang, X.; Li, H.; Kasemset, S.; Lee, A.; Agnihotri, D.; Hayes, T.; Paul, D. R.; Freeman, B. D., Fouling-resistant membranes for the treatment of flowback water from hydraulic shale fracturing: A pilot study. *Journal of Membrane Science* 2013, 437, 265-275.
19. Regnery, J.; Coday, B. D.; Riley, S. M.; Cath, T. Y., Solid-phase extraction followed by gas chromatography-mass spectrometry for the quantitative analysis of semi-volatile hydrocarbons in hydraulic fracturing wastewaters. *Analytical methods* 2016, 8, (9), 2058-2068.
20. Kong, F.-x.; Sun, G.-d.; Chen, J.-f.; Han, J.-d.; Guo, C.-m.; Tong, Z.; Lin, X.-f.; Xie, Y. F., Desalination and fouling of NF/low pressure RO membrane for shale gas fracturing flowback water treatment. *Separation and Purification Technology* 2018, 195, 216-223.
21. Dischinger, S. M.; Rosenblum, J.; Noble, R. D.; Gin, D. L.; Linden, K. G., Application of a lyotropic liquid crystal nanofiltration membrane for hydraulic fracturing flowback water: Selectivity and implications for treatment. *Journal of Membrane Science* 2017, 543, 319-327.
22. Michel, M. M.; Reczek, L.; Granops, M.; Rudnicki, P.; Piech, A., Pretreatment and desalination of flowback water from the hydraulic fracturing. *Desalination and Water Treatment* 2016, 57, (22), 10222-10231.
23. Guo, C.; Chang, H.; Liu, B.; He, Q.; Xiong, B.; Kumar, M.; Zydney, A. L., A combined ultrafiltration–reverse osmosis process for external reuse of Weiyuan shale gas flowback and produced water. *Environmental Science: Water Research & Technology* 2018, 4, (7), 942-955.
24. Jang, E.; Jeong, S.; Chung, E., Application of three different water treatment technologies to shale gas produced water. *Geosystem Engineering* 2017, 20, (2), 104-110.
25. Hu, L.; Jiang, W.; Xu, X.; Wang, H.; Carroll, K. C.; Xu, P.; Zhang, Y., Toxicological characterization of produced water from the Permian Basin. *Science of The Total Environment* 2022, 815, 152943.
26. Boyd, A.; Myers, S. P.; Luu, I.; Snihur, K.; Alessi, D. S.; Freitag, K.; Blewett, T. A., A common well pad does not imply common toxicity: Assessing the acute and chronic toxicity of flowback and produced waters from four Montney Formation wells on the same well pad to the freshwater invertebrate Daphnia magna. *Science of The Total Environment* 2022, 807, 150986.
27. Maguire-Boyle, S. J.; Barron, A. R., Organic compounds in produced waters from shale gas wells. *Environmental Science: Processes & Impacts* 2014, 16, (10), 2237-2248.
28. Riley, S. M.; Ahoor, D. C.; Regnery, J.; Cath, T. Y., Tracking oil and gas wastewater-derived organic matter in a hybrid biofilter membrane treatment system: A multi-analytical approach. *Science of The Total Environment* 2018, 613-614, 208-217.
29. Weinrauch, A. M.; Folkerts, E. J.; Alessi, D. S.; Goss, G. G.; Blewett, T. A., Changes to hepatic nutrient dynamics and energetics in rainbow trout (Oncorhynchus mykiss) following exposure to and recovery from hydraulic fracturing flowback and produced water. *Science of The Total Environment* 2021, 764, 142893.

30. Aghababaei, M.; Luek, J. L.; Ziemkiewicz, P. F.; Mouser, P. J., Toxicity of hydraulic fracturing wastewater from black shale natural-gas wells influenced by well maturity and chemical additives. *Environmental Science: Processes & Impacts* 2021, 23, (4), 621-632.

31. Folkerts, E. J.; Blewett, T. A.; Delompré, P.; Mehler, W. T.; Flynn, S. L.; Sun, C.; Zhang, Y.; Martin, J. W.; Alessi, D. S.; Goss, G. G., Toxicity in aquatic model species exposed to a temporal series of three different flowback and produced water samples collected from a horizontal hydraulically fractured well. *Ecotoxicology and Environmental Safety* 2019, 180, 600-609.

32. Rosenblum, J. S.; Sitterley, K. A.; Thurman, E. M.; Ferrer, I.; Linden, K. G., Hydraulic fracturing wastewater treatment by coagulation-adsorption for removal of organic compounds and turbidity. *Journal of environmental chemical engineering* 2016, 4, (2), 1978-1984.

33. Yin, Y.; Kalam, S.; Livingston, J. L.; Minjarez, R.; Lee, J.; Lin, S.; Tong, T., The use of anti-scalants in gypsum scaling mitigation: Comparison with membrane surface modification and efficiency in combined reverse osmosis and membrane distillation. *Journal of Membrane Science* 2022, 643, 120077.

34. Tong, T.; Zhao, S.; Boo, C.; Hashmi, S. M.; Elimelech, M., Relating silica scaling in reverse osmosis to membrane surface properties. *Environmental Science & Technology* 2017, 51, (8), 4396-4406.

35. OECD, Test No. 202: Daphnia sp. Acute Immobilisation Test. 2004.

36. Rohani, R.; Hyland, M.; Patterson, D., A refined one-filtration method for aqueous based nanofiltration and ultrafiltration membrane molecular weight cut-off determination using polyethylene glycols. *Journal of Membrane Science* 2011, 382, (1), 278-290.

37. Huang, H.; Cho, H.; Schwab, K.; Jacangelo, J. G., Effects of feedwater pretreatment on the removal of organic microconstituents by a low fouling reverse osmosis membrane. *Desalination* 2011, 281, 446-454.

38. Rusydi, A. F., Correlation between conductivity and total dissolved solid in various type of water: A review. *IOP Conference Series: Earth and Environmental Science* 2018, 118, (1), 012019.

39. Dórea, H. S.; Bispo, J. R.; Aragão, K. A.; Cunha, B. B.; Navickiene, S.; Alves, J. P.; Romão, L. P.; Garcia, C. A., Analysis of BTEX, PAHs and metals in the oilfield produced water in the State of Sergipe, Brazil. *Microchemical journal* 2007, 85, (2), 234-238.

40. OECD, Test No. 211: Daphnia magna Reproduction Test. 2008.

41. Delompré, P. L. M.; Blewett, T. A.; Snihur, K. N.; Flynn, S. L.; Alessi, D. S.; Glover, C. N.; Goss, G. G., The osmotic effect of hyper-saline hydraulic fracturing fluid on rainbow trout, Oncorhynchus mykiss. *Aquatic Toxicology* 2019, 211, 1-10.

42. Team, R. C., R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www. R-project. org/ 2013.

43. Kondash, A. J.; Warner, N. R.; Lahav, O.; Vengosh, A., Radium and barium removal through blending hydraulic fracturing fluids with acid mine drainage. *Environmental science & technology* 2014, 48, (2), 1334-1342.

44. Yilmaz, A. E.; Boncukcuoğlu, R.; Bayar, S.; Fil, B. A.; Kocakerim, M. M., Boron removal by means of chemical precipitation with calcium hydroxide and calcium borate formation. *Korean Journal of Chemical Engineering* 2012, 29, (10), 1382-1387.

45. Remy, P.; Muhr, H.; Plasari, E.; Ouerdiane, I., Removal of boron from wastewater by precipitation of a sparingly soluble salt. *Environmental Progress* 2005, 24, (1), 105-110.

46. Werber, J. R.; Deshmukh, A.; Elimelech, M., The critical need for increased selectivity, not increased water permeability, for desalination membranes. *Environmental Science & Technology Letters* 2016, 3, (4), 112-120.
47. Martin, D. L.; Gilley, J. R., Irrigation Water Requirements. Chapter 2. The SCS national engineering handbook 1993.
48. Soltanpour, P. N.; Raley, W. L., Livestock drinking water quality. Colorado State University Extension Service: 1999.
49. Song, Y.; Su, B.; Gao, X.; Gao, C., Investigation on high NF permeate recovery and scaling potential prediction in NF–SWRO integrated membrane operation. *Desalination* 2013, 330, 61-69.
50. Eriksson, P.; Kyburz, M.; Pergande, W., NF membrane characteristics and evaluation for sea water processing applications. *Desalination* 2005, 184, (1), 281-294.
51. Al-Amoudi, A.; Lovitt, R. W., Fouling strategies and the cleaning system of NF membranes and factors affecting cleaning efficiency. *Journal of Membrane Science* 2007, 303, (1), 4-28.
52. Koseoglu, H.; Kabay, N.; Yüksel, M.; Sarp, S.; Arar, Ö.; Kitis, M., Boron removal from seawater using high rejection SWRO membranes — impact of pH, feed concentration, pressure, and cross-flow velocity. *Desalination* 2008, 227, (1), 253-263.
53. Redondo, J.; Busch, M.; De Witte, J.-P., Boron removal from seawater using FILMTECTM high rejection SWRO membranes. *Desalination* 2003, 156, (1), 229-238.
54. Kabay, N.; Bryjak, M., Chapter 9 - Boron removal from seawater using reverse osmosis integrated processes. In boron separation processes, Kabay, N.; Bryjak, M.; Hilal, N., Eds. Elsevier: Amsterdam, 2015; pp 219-235.
55. Tu, K. L.; Nghiem, L. D.; Chivas, A. R., Boron removal by reverse osmosis membranes in seawater desalination applications. *Separation and Purification Technology* 2010, 75, (2), 87-101.
56. Kluczka, J.; Trojanowska, J.; Zolotajkin, M.; Ciba, J.; Turek, M.; Dydo, P., Boron Removal from Wastewater Using Adsorbents. *Environmental Technology* 2007, 28, (1), 105-113.
57. Harada, A.; Takagi, T.; Kataoka, S.; Yamamoto, T.; Endo, A., Boron adsorption mechanism on polyvinyl alcohol. Adsorption 2011, 17, (1), 171-178.
58. Ayers, R.; Westcot, D., Water quality for agriculture; FAO irrigation and drainage paper; Food and Agriculture Organization of the United Nations: Rome, Italy, 1985.
59. Bauder, T. A.; Waskom, R.; Sutherland, P.; Davis, J.; Follett, R.; Soltanpour, P., Irrigation water quality criteria. Service in action; no. 0.506 2011.
60. Oster, J., Irrigation with poor quality water. *Agricultural water management* 1994, 25, (3), 271-297.
61. Oster, J., Gypsum usage in irrigated agriculture: a review. *Fertilizer research* 1982, 3, (1), 73-89.
62. Goldberg, S.; J. Kabengi, N., Bromide adsorption by reference minerals and soils. *Vadose Zone Journal* 2010, 9, (3), 780-786.
63. Bottero, J.; Axelos, M.; Tchoubar, D.; Cases, J.; Fripiat, J.; Fiessinger, F., Mechanism of formation of aluminum trihydroxide from Keggin Al13 polymers. *Journal of Colloid and Interface Science* 1987, 117, (1), 47-57.
64. Bottero, J.; Tchoubar, D.; Cases, J.; Fiessinger, F., Investigation of the hydrolysis of aqueous solutions of aluminum chloride. 2. Nature and structure by small-angle X-ray scattering. *The Journal of Physical Chemistry* 1982, 86, (18), 3667-3673.

65. Serna, C. J.; White, J.; Hem, S. L., Anion-aluminum hydroxide gel interactions. *Soil Science Society of America Journal* 1977, 41, (5), 1009-1013.

66. Werber, J. R.; Osuji, C. O.; Elimelech, M., Materials for next-generation desalination and water purification membranes. *Nature Reviews Materials* 2016, 1, (5), 1-15.

67. Vezzani, D.; Bandini, S., Donnan equilibrium and dielectric exclusion for characterization of nanofiltration membranes. *Desalination* 2002, 149, (1), 477-483.

68. Luo, J.; Wan, Y., Effects of pH and salt on nanofiltration—a critical review. *Journal of Membrane Science* 2013, 438, 18-28.

69. Nightingale Jr, E., Phenomenological theory of ion solvation. Effective radii of hydrated ions. *The Journal of Physical Chemistry* 1959, 63, (9), 1381-1387.

70. Marcus, Y., Thermodynamics of solvation of ions. Part 5.—Gibbs free energy of hydration at 298.15 K. Journal of the Chemical Society, *Faraday Transactions* 1991, 87, (18), 2995-2999.

71. Epsztein, R.; Shaulsky, E.; Dizge, N.; Warsinger, D. M.; Elimelech, M., Role of ionic charge density in Donnan exclusion of monovalent anions by nanofiltration. *Environmental Science & Technology* 2018, 52, (7), 4108-4116.

72. Epsztein, R.; DuChanois, R. M.; Ritt, C. L.; Noy, A.; Elimelech, M., Towards single-species selectivity of membranes with subnanometre pores. *Nature Nanotechnology* 2020, 15, (6), 426-436.

73. Shefer, I.; Peer-Haim, O.; Epsztein, R., Limited ion-ion selectivity of salt-rejecting membranes due to enthalpy-entropy compensation. *Desalination* 2022, 541, 116041.

74. Shefer, I.; Lopez, K.; Straub, A. P.; Epsztein, R., Applying transition-state theory to explore transport and selectivity in salt-rejecting membranes: A critical review. *Environmental Science & Technology* 2022.

75. Kisic, I.; Mesic, S.; Basic, F.; Brkic, V.; Mesic, M.; Durn, G.; Zgorelec, Z.; Bertovic, L., The effect of drilling fluids and crude oil on some chemical characteristics of soil and crops. *Geoderma* 2009, 149, (3-4), 209-216.

76. Gong, C.; Huang, H.; Qian, Y.; Zhang, Z.; Wu, H., Integrated electrocoagulation and membrane filtration for PAH removal from realistic industrial wastewater: effectiveness and mechanisms. *Rsc Advances* 2017, 7, (83), 52366-52374.

77. Smol, M.; Włodarczyk-Makuła, M.; Mielczarek, K.; Bohdziewicz, J.; Włóka, D., The use of reverse osmosis in the removal of PAHs from municipal landfill leachate. *Polycyclic Aromatic Compounds* 2016, 36, (1), 20-39.

78. Li, S.; Luo, J.; Hang, X.; Zhao, S.; Wan, Y., Removal of polycyclic aromatic hydrocarbons by nanofiltration membranes: rejection and fouling mechanisms. *Journal of Membrane Science* 2019, 582, 264-273.

79. Robbins, C. A.; Yin, Y.; Hanson, A. J.; Blotevogel, J.; Borch, T.; Tong, T., Mitigating membrane wetting in the treatment of unconventional oil and gas wastewater by membrane distillation: A comparison of pretreatment with omniphobic membrane. *Journal of Membrane Science* 2022, 645, 120198.

80. Van Houghton, B. D.; Acharya, S. M.; Rosenblum, J. S.; Chakraborty, R.; Tringe, S. G.; Cath, T. Y., Membrane bioreactor pretreatment of high-salinity O&G produced water. *ACS ES&T Water* 2022, 2, (3), 484-494.

81. McAdams, B. C.; Carter, K. E.; Blotevogel, J.; Borch, T.; Hakala, J. A., In situ transformation of hydraulic fracturing surfactants from well injection to produced water. *Environmental Science: Processes & Impacts* 2019, 21, (10), 1777-1786.

82. de Alda, M. a. J. L.; Díaz-Cruz, S.; Petrovic, M.; Barceló, D., Liquid chromatography–(tandem) mass spectrometry of selected emerging pollutants (steroid sex hormones, drugs and alkylphenolic surfactants) in the aquatic environment. *Journal of Chromatography a* 2003, 1000, (1-2), 503-526.

83. Freitas, L. G.; Götz, C. W.; Ruff, M.; Singer, H. P.; Müller, S. R., Quantification of the new triketone herbicides, sulcotrione and mesotrione, and other important herbicides and metabolites, at the ng/l level in surface waters using liquid chromatography–tandem mass spectrometry. *Journal of Chromatography A* 2004, 1028, (2), 277-286.

84. Palmer, M.; Hatley, H., The role of surfactants in wastewater treatment: Impact, removal and future techniques: A critical review. *Water Research* 2018, 147, 60-72.

85. Petrovic, M.; Barceló, D., Fate and removal of surfactants and related compounds in wastewaters and sludges. In series anthropogenic compounds: Emerging organic pollution in waste waters and sludge, Vol. 1, Barceló, D., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2004; pp 1-28.

86. Beltrán-Heredia, J.; Sánchez-Martín, J.; Solera-Hernández, C., Anionic surfactants removal by natural coagulant/flocculant products. *Industrial & Engineering Chemistry Research* 2009, 48, (10), 5085-5092.

87. Holland, P. M.; Rubingh, D. N., Mixed surfactant systems: an overview. 1992.

88. Ogino, K.; Abe, M., Mixed surfactant systems. CRC Press: 1992.

89. Rodriguez, C. H.; Scamehorn, J. F., Kinetics of precipitation of surfactants. II. Anionic surfactant mixtures. *Journal of Surfactants and Detergents* 2001, 4, (1), 15-26.

90. Song, X.; Gan, B.; Qi, S.; Guo, H.; Tang, C. Y.; Zhou, Y.; Gao, C., Intrinsic nanoscale structure of thin film composite polyamide membranes: Connectivity, defects, and structure–property correlation. *Environmental Science & Technology* 2020, 54, (6), 3559-3569.

91. Li, Y.; Kłosowski, M. M.; McGilvery, C. M.; Porter, A. E.; Livingston, A. G.; Cabral, J. T., Probing flow activity in polyamide layer of reverse osmosis membrane with nanoparticle tracers. *Journal of Membrane Science* 2017, 534, 9-17.

92. Blewett, T. A.; Delompré, P. L.; He, Y.; Folkerts, E. J.; Flynn, S. L.; Alessi, D. S.; Goss, G. G., Sublethal and reproductive effects of acute and chronic exposure to flowback and produced water from hydraulic fracturing on the water flea Daphnia magna. *Environmental science & technology* 2017, 51, (5), 3032-3039.

93. Martínez-Jerónimo, F.; Martínez-Jerónimo, L., Chronic effect of NaCl salinity on a freshwater strain of Daphnia magna Straus (Crustacea: Cladocera): a demographic study. *Ecotoxicology and Environmental Safety* 2007, 67, (3), 411-416.

94. Honda, M.; Suzuki, N., Toxicities of polycyclic aromatic hydrocarbons for aquatic animals. *International Journal of Environmental Research and Public Health* 2020, 17, (4), 1363.

95. Barata, C.; Baird, D. J., Determining the ecotoxicological mode of action of chemicals from measurements made on individuals: results from instar-based tests with Daphnia magna Straus. *Aquatic Toxicology* 2000, 48, (2-3), 195-209.

96. Crookes, M.; Dobson, S.; Howe, P., Environmental hazard assessment: Xylenes. Building Research Establishment: 1993.

97. Blewett, T. A.; Delompré, P. L. M.; Glover, C. N.; Goss, G. G., Physical immobility as a sensitive indicator of hydraulic fracturing fluid toxicity towards Daphnia magna. *Science of The Total Environment* 2018, 635, 639-643.

98. Boyd, A.; Stewart, C. B.; Philibert, D. A.; How, Z. T.; El-Din, M. G.; Tierney, K. B.; Blewett, T. A., A burning issue: The effect of organic ultraviolet filter exposure on the behaviour and physiology of Daphnia magna. *Science of the Total Environment* 2021, 750, 141707.

99. Okamoto, A.; Yamamuro, M.; Tatarazako, N., Acute toxicity of 50 metals to Daphnia magna. *Journal of applied toxicology* 2015, 35, (7), 824-830.

100. Wang, S.; Zhou, Y.; Gao, C., Novel high boron removal polyamide reverse osmosis membranes. *Journal of Membrane Science* 2018, 554, 244-252.

101. Zhang, X.; Wei, M.; Zhang, Z.; Shi, X.; Wang, Y., Boron removal by water molecules inside covalent organic framework (COF) multilayers. *Desalination* 2022, 526, 115548.

102. Chen, M.; Dollar, O.; Shafer-Peltier, K.; Randtke, S.; Waseem, S.; Peltier, E., Boron removal by electrocoagulation: Removal mechanism, adsorption models and factors influencing removal. *Water Research* 2020, 170, 115362.

103. Zhang, J.; Cai, Y.; Liu, K., Extremely effective boron removal from water by stable metal organic framework ZIF-67. *Industrial & Engineering Chemistry Research* 2019, 58, (10), 4199-4207.

# 7. Conclusions and recommended future research

## 7.1 Conclusions

In my research, I investigated the applications of ML models to predicting membrane performance and the use of NF and RO membranes for UOG produced water treatment. To achieve this goal, I built ML models to predict membrane performance and investigated appropriate data splitting methods to avoid data leakage. I also analyzed and compared the underlying knowledge of ML models revealed by XAI with the domain knowledge of membrane separation. Meanwhile, the feasibility of NF and RO membranes in treating UOG produced water and the potential risks of applying treated produced water for beneficial reuse were investigated.

First, 1907 data pertaining to the removal of organic constituents using NF and RO membranes were collected from the literature to build ML models. By examining the influence of data splitting methods for training, validation, and testing data on the prediction accuracy of the ML models, I discovered that an improper data splitting method can lead to falsely high performance of ML models. As a result, I developed a stratified sampling method for data pertaining to membrane performance, enabling objective evaluation of the ML model performance.

Second, the underlying knowledges of ML models were probed using the SHAP method. My results showed that ML models could capture the important role of size exclusion in regulating the transport of organic solutes; however, the understandings of the models on electrostatic interaction and adsorption remained rudimentary. To investigate the knowledge of ML on inorganic solute transport, four different ML models (i.e., Random Forest, XGBoost, LightGBM, and Catboost) were built to predict the rejection of inorganic constituents by NF and RO membranes. The data for single salt solutions, cations, and anions in mixture salt solutions were

separately collected to build ML models. I revealed that the ML models captured the different importance of size exclusion and electrostatic interaction to solute transport for single salt solutions, cations, and anions in mixture salt solutions, which was aligned with the inorganic solute transport mechanisms reported in the literature. However, the ML models were not able to identify the governing mechanisms of inorganic solute transport correctly when the volume of data was not sufficient. Therefore, my work provides a framework for evaluating the knowledges of ML models, having the potential of facilitating more reliable and understandable ML applications to membrane selection and design.

Lastly, UOG produced water sampled from Niobrara, Colorado was treated by NF and RO membranes to investigate the performance of membrane possessing different perm-selectivity. The concentrations of organic and inorganic constituents in the permeates were compared to the water quality standards for irrigation and livestock drinking water. It was found that the NF permeates could not meet the water quality criteria for irrigation and livestock drinking water, while RO permeates generally met the standards, except for boron and sodium adsorption ratio (SAR). Furthermore, surfactants with molecular weight much higher than the molecular weight cut-off of membranes could pass through NF and RO membranes, suggesting that membrane could not completely prevent the passage of organic compounds. The toxicity tests using *Daphnia magna* showed that NF permeates were still toxic, whereas very low or no toxicity was observed by RO permeates. The main driver of toxicity was likely to be salinity, which showed a clear correlation with toxicity of the treated produced water.

## 7.2 Recommended future research

Activities for future research are suggested regarding to the application of ML model for membrane performance prediction and the improvement of membrane design and selection based on the findings of my thesis.

- More efforts are needed to find better variable representations of solute and membrane properties for ML models. In this thesis, simplified variables were used to represent certain properties of membranes and solutes. For example, the sizes of organic compounds were obtained by assuming all the compounds are spherical in Chapter 4, regardless of their chemical structures. This means that the information of actual chemical structures of organic compounds were not properly provided to ML models. The limitation of using simplified variables to represent certain properties of membranes and solutes may contribute to a deterioration in the prediction accuracy. Using line notation methods, such as Morgan fingerprints and simplified molecular-input line-entry system (SMILES) strings, for input variables has the potential to reflect the properties of compounds more accurately and lead to better performance of ML models.
- ML has been used to investigate the influence of certain polymers or functional groups on membrane performance.[1, 2] However, there have been insufficient efforts to understand the influence of polymers on membrane performance from the atomic-level information. This can be achieved using attention-based transformer models. Attention-based transformer models, which mimic the natural cognitive process of attention, selectively learn important information from building blocks (e.g., to find optimal polymer for membrane design, building blocks can be chemical elements consisting of polymers) and correlate the information with the output.[3-5] The application of transformer models to search potentially

promising polymers for highly selective membrane will pave a new way toward the fit-for-purpose membrane design.

**References**

1. Gao, H.; Zhong, S.; Zhang, W.; Igou, T.; Berger, E.; Reid, E.; Zhao, Y.; Lambeth, D.; Gan, L.; Afolabi, M. A., Revolutionizing membrane design using machine learning-Bayesian optimization. *Environmental Science & Technology* 2021, *56*, (4), 2572-2581.
2. Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y., Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. *Science Advances* 2022, *8*, (29), eabn9545.
3. Buehler, M. J., FieldPerceiver: Domain agnostic transformer model to predict multiscale physical fields and nonlinear material properties through neural ologs. *Materials Today* 2022, *57*, 9-25.
4. Rahardja, S.; Wang, M.; Nguyen, B. P.; Fränti, P.; Rahardja, S., A lightweight classification of adaptor proteins using transformer networks. *BMC bioinformatics* 2022, *23*, (1), 1-14.
5. Buehler, M. J., Multiscale modeling at the interface of molecular mechanics and natural language through attention neural networks. *Accounts of Chemical Research* 2022, *55*, (23), 3387-3403.

# Appendix A

**Details about input variables**

The properties of compounds (log $K_{ow}$, maximum projection, minimum projection, and charge) were acquired from the Chemicalize database.[1] As the percentages of charged and uncharged compounds change as a function of experimental pH, the compound charge used in this study refers to the net charge of the compound. For example, if 50% of a certain compound is neutral and the remaining 50% of this compound is positively charged (1+ for monovalent compound), the net charge of the compound is 0.5**.** The input variables associated with membrane properties and experimental conditions were obtained from the corresponding literature. Membrane zeta potential for TC was modified to 1 for positively charged surfaces and -1 for negatively charged surfaces because the membrane zeta potentials were typically measured in different conditions from the micropollutant rejection experiments. When some variables of membrane properties were absent in certain references, variables from other references using the same membrane were used. All data used in this study can be found from the excel files of Supporting Information. Pearson correlation coefficients were calculated in order to check the redundancy of input variables, using Python package Pandas. As shown in Figure A1, the correlation values between the selected inputs are less than 0.3, indicating that these variables are not redundant variables for model training. Although additional variables could be used in the model, the above eight input variables cover the major mechanisms of micropollutant removal by NF and RO membranes, and an increase of variable number increases the computational demand and may lead to overfitting.[2]

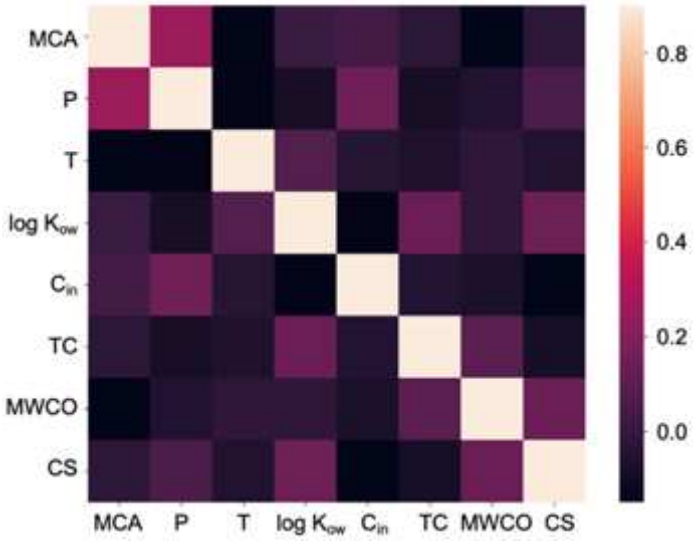


**Figure A1.** The correlation matrix of input variables. Variables in the matrix are membrane contact angle (MCA), pressure (P), measurement time (T), log $K_{ow}$, initial concentration of compound ($C_{in}$), total charge (TC), MWCO, and compound size (CS).

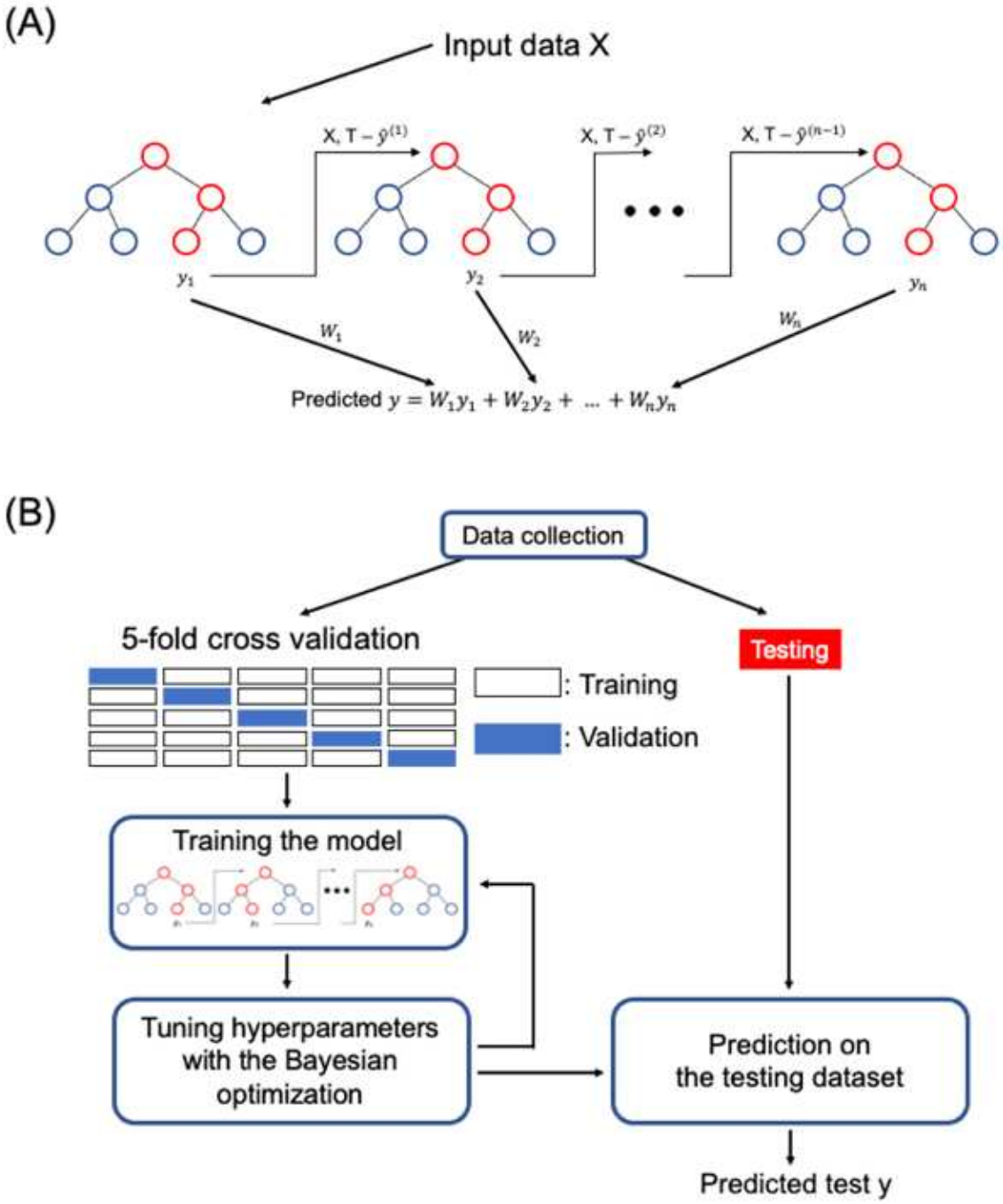


**Figure A2.** (A) The structure of gradient boosting tree with decision nodes (red and blue circles) and leaf nodes ($y_n$). X and T indicate the input data and target labels, respectively. $W$ is the weight of each tree. $\hat{y}^{(t)}$ is defined as $\sum_{i=1}^{t} y_t$. (B) The workflow of training the XGBoost model.

**K-fold cross-validation**

K-fold cross-validation was implemented during the training and validation of the model. K-fold cross-validation is a resampling method that increases the ambiguity of input data and reduces the generalization errors (i.e., prediction errors on unseen data).[3] Data for training and validation are divided into K groups that are mutually exclusive. One group is used for validation, and the others are assigned as training datasets. This procedure repeats K times, and the groups for training and validation change each time. During the optimization process, the average root mean squared errors (RMSE) of K groups are calculated to find the best hyperparameters and parameters.

**Principal component analysis**

Principal component analysis (PCA) is a method to find the new variables, which have linear relationship with those of the original data and the maximum variance, in order to reduce the dimensions of data while maintaining the data information. Based on the correlation structure of the variables, principal components, can be expressed as followed:

$$X = \sum_{i=1}^{n} s_i \cdot p_i + E$$

where $X$ is the matrix of original data with $n$ number of data, $S$ is the scores of the principal components, $p$ is the principal component and $E$ represents the noise of data that is not explained by the principal components.[4] When certain principal components account for relatively small portions in the total variances, these components are considered as less important information and can be removed to lower the computational demands. In our study, 14 variables including pH, compound charge, compound log $K_{ow}$, compound minimum projection, compound maximum projection, compound molecular weight (Mw), MWCO, membrane contact angle (MCA), membrane zeta potential, total charge (TC), pressure (P), measurement time (T), initial concentration of the compound ($C_{in}$), and compound size (CS), were normalized and used for PCA. After PCA, 12 principal components were extracted, which covers 98.7% of the total variances (Figure A1). PCA was conducted by using Python with Scikit-Learn package.

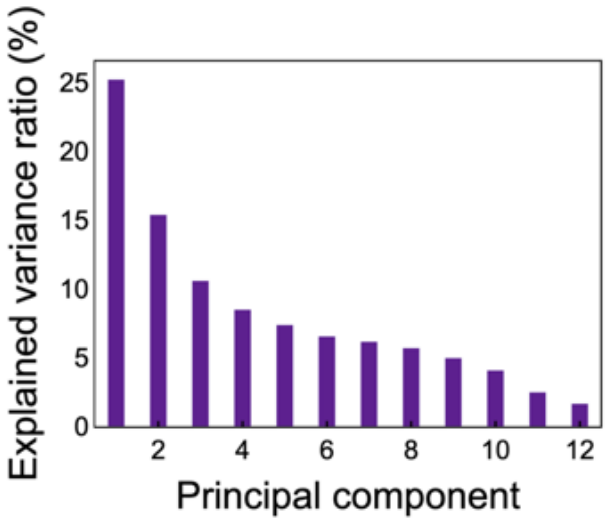


**Figure A3.** Explained variance ratio of principal components after the extraction of principal components.

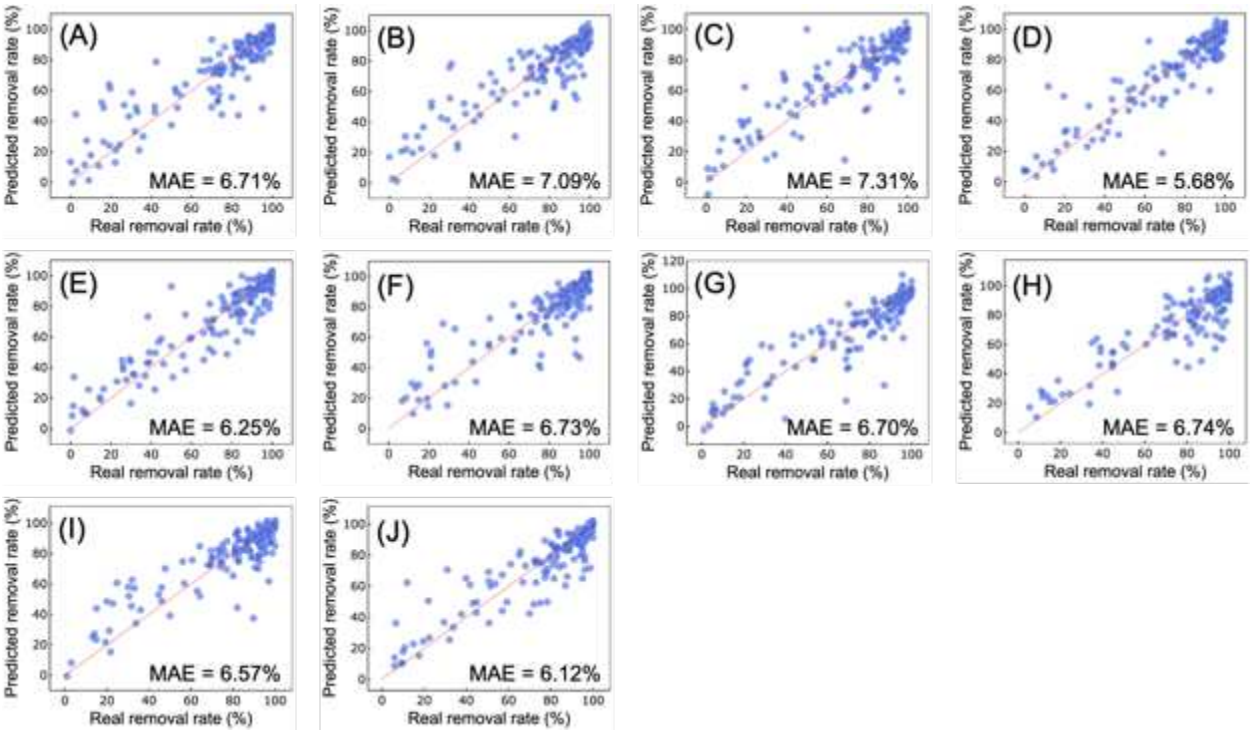


**Figure A4.** Predictions of the model with random data splitting. Ten replicates of the predictions with data split ratio of 90:10 were conducted with different training/validation and testing datasets.

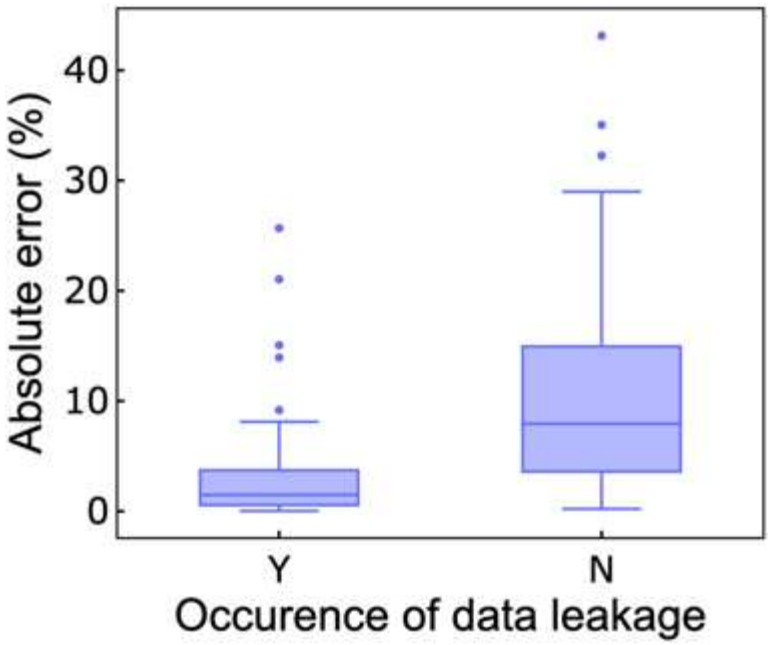


**Figure A5.** Absolute errors of the model predictions using the data splitting method of random mixing with (Y) and without (N) data leakage. Mean absolute errors with and without data leakage are 2.83% and 10.17 %, respectively. Detailed data are presented in Table A3.

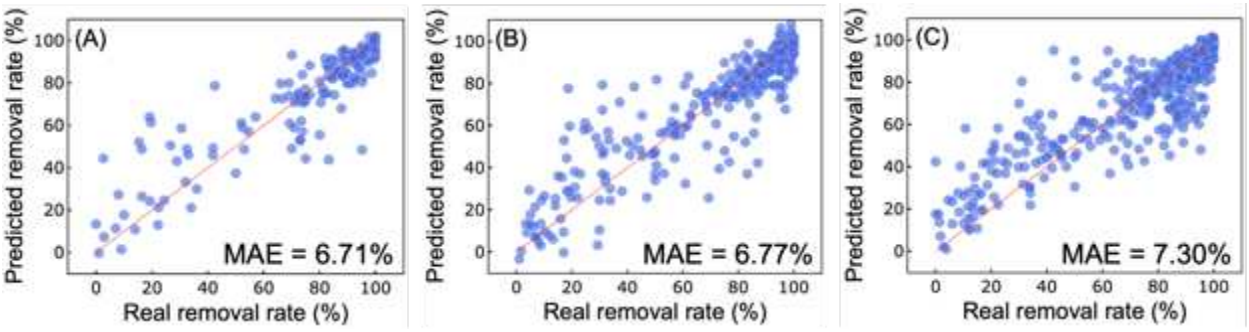


**Figure A6.** Predictions of the model with random data splitting. The data split ratios are (A) 90:10 (MAE: 6.71%), (B) 80:20 (MAE: 6.77%), and (C) 70:30 (MAE: 7.30%).

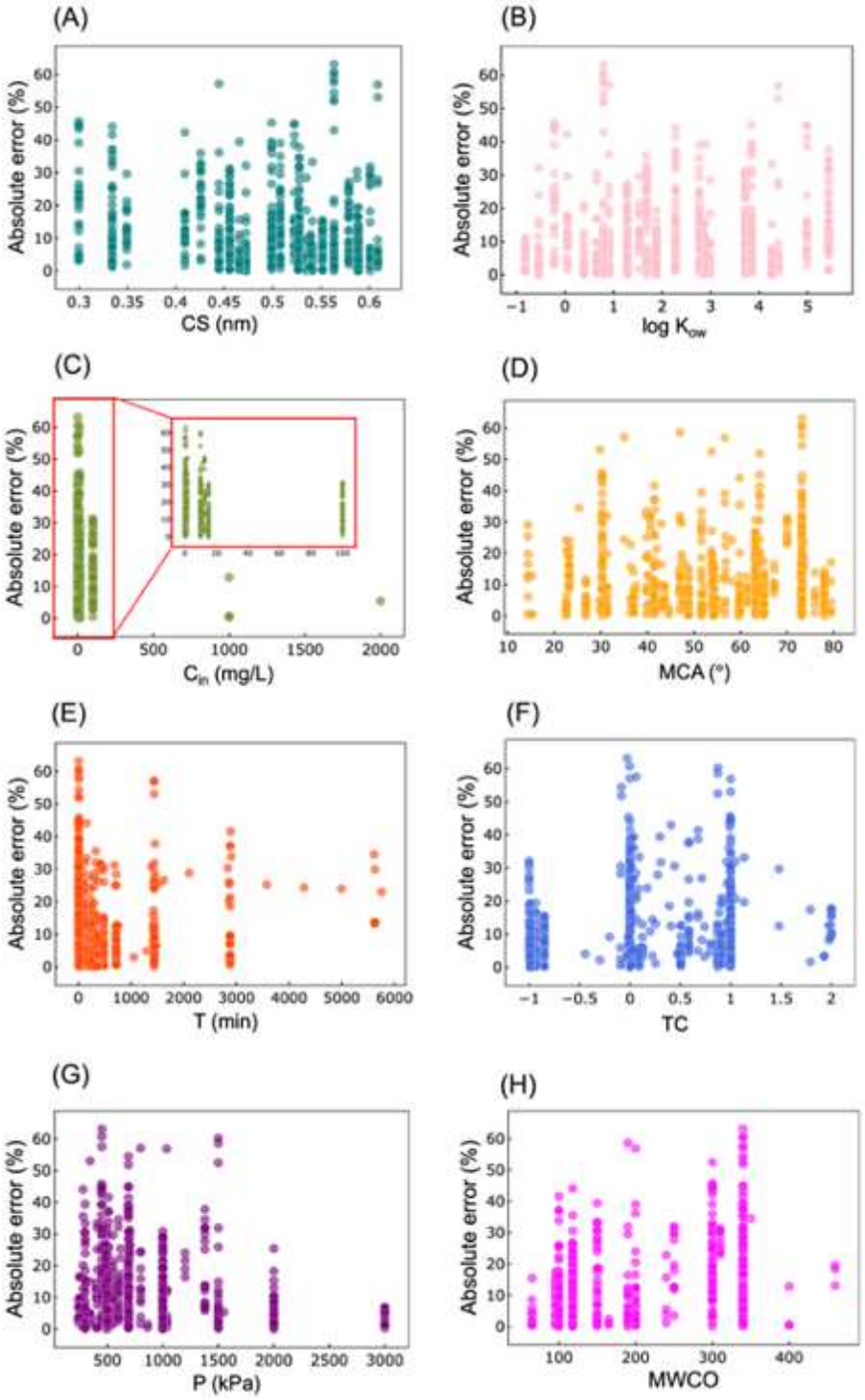


**Figure A7.** Absolute errors of the model predictions as a function of (A) compound size (CS), (B) log $K_{ow}$, (C) initial concentration of compounds ($C_{in}$), (D) membrane contact angle (MCA), (E) measurement time (T), (F) total charge (TC), (G) pressure (P), and (H) MWCO for model predictions for unknown compounds.

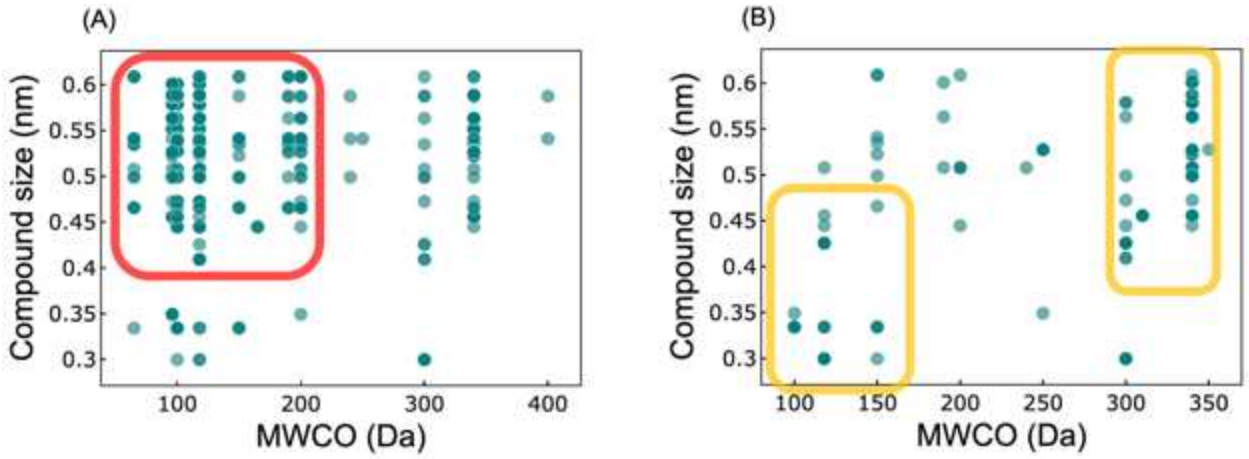


**Figure A8.** Comparison of data points in terms of compound size and MWCO for model predictions for unknown compounds. The predictions shown in Figure (A) have low absolute errors (< 10%), whereas those shown in Figure (B) have high absolute errors (> 20%). Compared to Figure (B), the majority of data points in Figure (A) are associated with small MWCO and high compound size (e.g., data within red rectangle, accounting for 78.1% of data with low absolute errors), while the majority of data points in Figure (B) are associated with either small compounds size or high MWCO (e.g., data within yellow rectangle, accounting for 81.0% of data with high absolute errors). It is worth to mention that each point shown in this figure might represent more than one data points, due to the overlap of the data.

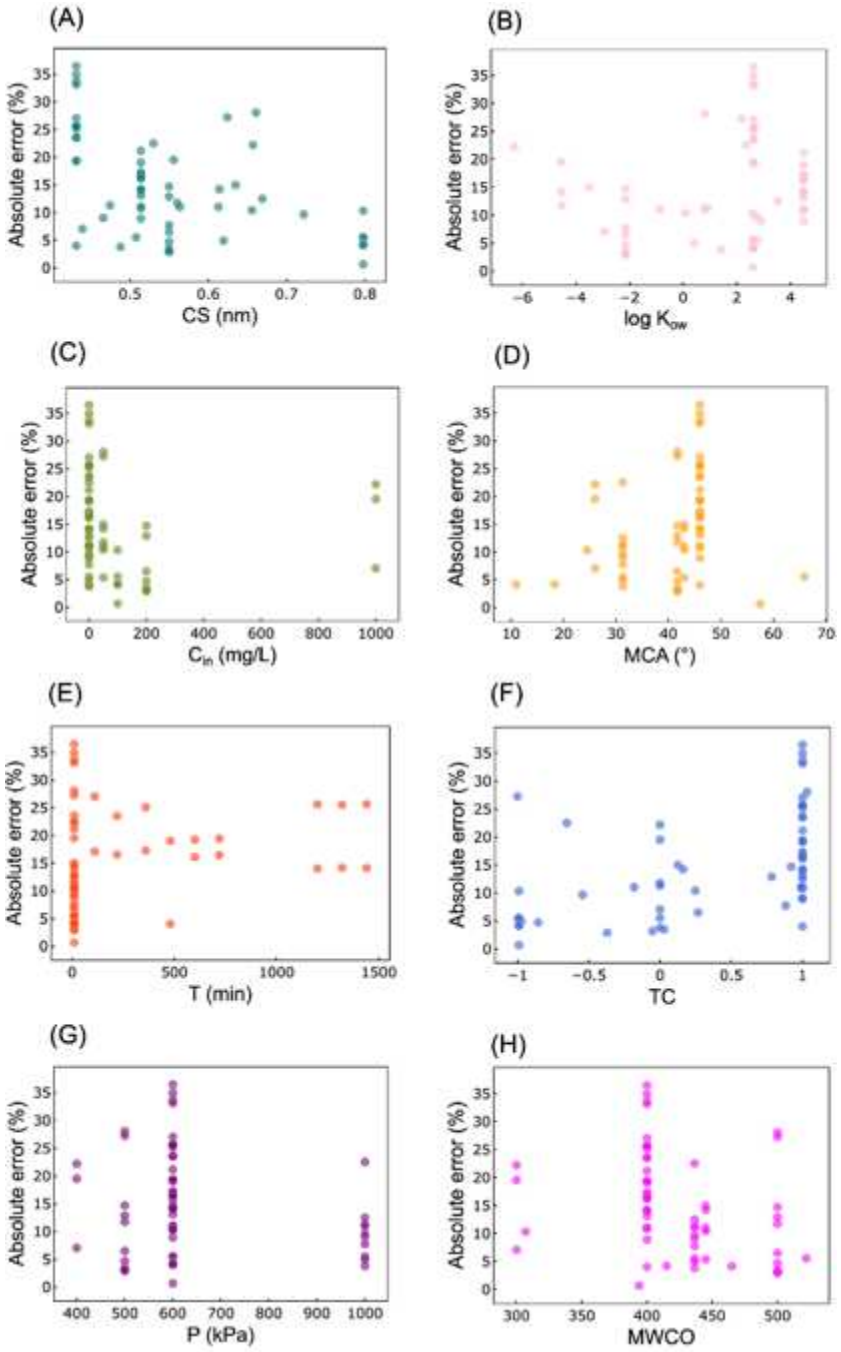


**Figure A9.** Absolute errors of the model predictions as a function of (A) compound size, (B) log $K_{ow}$, (C) initial concentration of compounds ($C_{in}$), (D) membrane contact angle (MCA), (E) measurement time (T), (F) total charge (TC), (G) pressure (P), and (H) MWCO for the predictions for self-fabricated membranes.

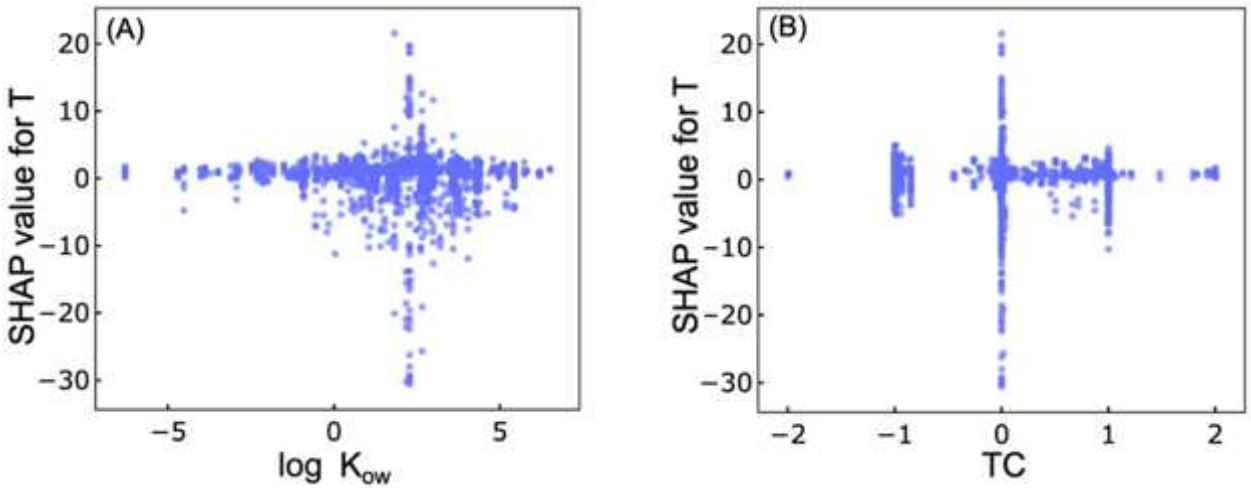


**Figure A10.** The SHAP values for measurement time (T) as a function of (A) log $K_{ow}$ and (B) total charge (TC).

**Table A1**. Hyperparameters and their ranges for Bayesian optimization.

| **Hyperparameters** | **Range** |
|---|---|
| Colsample_bytree | 0.5 – 0.9 |
| Gamma | 0 – 1 |
| Learning_rate | 0.01 – 0.3 |
| Max_depth | 3 – 10 |
| Min_child_weight | 0.0001 – 5 |
| Reg_alpha | 0 – 1 |
| Reg_lambda | 0 – 1 |
| Subsample | 0.5 – 0.9 |

**Table A2.** Mean absolute errors (MAE) of different models used for the prediction of micropollutant removal by membranes in this study and the literature. The table only includes the literatures with clearly presented MAE values. Models in the table randomly split the training, validation, and test dataset, except for reference 7 (106 training data from internal experiment and 89 test data from the literature). When multiple models were tested in the same literature, the best model with the lowest MAE was chosen.

| Model | Number of data | MAE (%) | Year published | Reference |
|---|---|---|---|---|
| Random mixing XGBoost | 1907 | 6.25 | 2021 | This paper |
| Multilinear regression | 1907 | 19.28 | 2021 | This paper |
| Bootstrap Aggregated Neural network (BANN) | 436 | 4.96 | 2017 | [5] |
| Artificial neural network (ANN) | 701 | 6.1 | 2015 | [6] |
| Multilinear regression | 195 | 11.7 | 2010 | [7] |
| Artificial neural network (ANN) | 161 | 4.56 | 2009 | [8] |

**Table A3.** Testing data for predictions associated with in Figure 4-2B. Features in the tables are identification (ID), type of membrane (MB), compound name (C), pH, membrane contact angle (MCA), pressure (P), measurement time (T), compound log $K_{ow}$, initial concentration of compound ($C_{in}$), total charge (TC), MWCO, compound size (CS), removal rates, predictions (Pred), errors, absolute errors (abs error), and occurrence of data leakage (N = no leakage, Y = with leakage).

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | AK | CPL | 7 | 50.0 | 800 | 10 | 0.31 | 0.1132 | 0.00 | 150 | 0.40 | 82.9 | 75.1 | 7.7 | 7.74 | N |
| 16 | BW30 | ACT | 4 | 59.8 | 2000 | 10 | 0.91 | 0.5000 | 0.00 | 100 | 0.44 | 94.3 | 93.1 | 1.3 | 1.27 | Y |
| 49 | BW30 | DPR | 7 | 59.8 | 1500 | 10 | -0.82 | 0.5000 | 1.00 | 100 | 0.54 | 99.0 | 99.3 | -0.3 | 0.32 | Y |
| 57 | BW30 | IBP | 5 | 59.8 | 1000 | 10 | 3.84 | 0.5000 | 0.58 | 100 | 0.50 | 99.3 | 99.3 | 0.0 | 0.02 | Y |
| 58 | BW30 | IBP | 7 | 59.8 | 1000 | 10 | 3.84 | 0.5000 | 0.99 | 100 | 0.50 | 99.1 | 97.2 | 1.9 | 1.92 | Y |
| 62 | BW30 | IBP | 4 | 59.8 | 2000 | 10 | 3.84 | 0.5000 | 0.12 | 100 | 0.50 | 99.2 | 99.9 | -0.7 | 0.70 | Y |
| 70 | BW30 | NMEA | 5.6 | 76.0 | 1550 | 10 | 0.06 | 0.2000 | 0.00 | 100 | 0.36 | 85.1 | 84.7 | 0.4 | 0.38 | N |
| 85 | CPA3 | DXT | 7 | 73.0 | 1000 | 240 | 1.68 | 15.0000 | 0.00 | 100 | 0.60 | 96.6 | 96.1 | 0.5 | 0.51 | Y |
| 122 | ES20 | UREA | 7 | 47.0 | 294 | 10 | -1.36 | 10.0000 | 0.00 | 150 | 0.30 | 29.9 | 44.1 | -14.3 | 14.26 | N |
| 137 | ESNA | PMD | 8 | 47.3 | 552 | 480 | 1.12 | 0.1000 | 0.00 | 250 | 0.47 | 86.9 | 64.1 | 22.8 | 22.77 | N |
| 139 | ESNA | TCA | 8 | 47.3 | 552 | 59 | 1.53 | 0.1000 | 1.00 | 250 | 0.35 | 97.2 | 97.3 | -0.2 | 0.18 | Y |
| 146 | ESPA | 22BET | 7 | 47.0 | 800 | 10 | 0.71 | 0.1622 | 0.00 | 200 | 0.51 | 85.2 | 75.0 | 10.2 | 10.23 | N |
| 161 | HL | CPF | 7 | 47.0 | 1500 | 10 | -0.86 | 10.0000 | -0.05 | 190 | 0.61 | 98.2 | 94.9 | 3.3 | 3.31 | N |
| 178 | HL | STL | 8 | 43.0 | 500 | 10 | -0.40 | 0.0020 | -0.95 | 190 | 0.58 | 75.5 | 78.2 | -2.7 | 2.65 | N |
| 189 | LE440 | BMF | 10 | 41.5 | 510 | 1416 | 2.28 | 0.1000 | 0.00 | 100 | 0.33 | 66.1 | 45.1 | 21.0 | 21.02 | Y |
| 192 | LE440 | BMF | 8 | 41.5 | 510 | 352 | 2.28 | 0.1000 | 0.00 | 100 | 0.33 | 83.5 | 87.3 | -3.8 | 3.80 | Y |
| 219 | LE440 | CTB | 8 | 41.5 | 510 | 721 | 3.60 | 0.1000 | 0.00 | 100 | 0.34 | 93.9 | 100.0 | -6.1 | 6.05 | Y |
| 229 | LE440 | NPX | 8 | 42.0 | 550 | 2880 | 2.99 | 0.1000 | 1.00 | 100 | 0.53 | 75.0 | 86.2 | -11.2 | 11.15 | N |
| 241 | LE440 | TCE | 8 | 41.5 | 510 | 1434 | 2.18 | 0.1000 | 0.00 | 100 | 0.35 | 31.8 | 35.5 | -3.7 | 3.73 | Y |
| 251 | LFC1 | DXT | 7 | 78.0 | 1000 | 1440 | 1.68 | 15.0000 | 0.00 | 100 | 0.60 | 99.8 | 99.9 | -0.1 | 0.15 | Y |
| 256 | LFC1 | HCT | 7 | 78.0 | 1000 | 360 | 1.28 | 15.0000 | 0.00 | 100 | 0.58 | 99.5 | 99.2 | 0.3 | 0.27 | Y |

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 261 | LFC1 | LDC | 7 | 78.0 | 1000 | 120 | 2.84 | 15.0000 | -0.85 | 100 | 0.55 | 100.0 | 99.5 | 0.5 | 0.46 | Y |
| 263 | LFC1 | LDC | 7 | 78.0 | 1000 | 360 | 2.84 | 15.0000 | -0.85 | 100 | 0.55 | 100.0 | 98.6 | 1.4 | 1.38 | Y |
| 267 | LFC1 | PFOS | 4 | 78.0 | 1379 | 134 | 5.43 | 10.0000 | -1.00 | 100 | 0.53 | 99.9 | 99.9 | 0.0 | 0.01 | Y |
| 275 | LFC1 | PRC | 7 | 78.0 | 1000 | 480 | 1.88 | 15.0000 | -0.99 | 100 | 0.59 | 99.4 | 97.7 | 1.7 | 1.74 | Y |
| 288 | NF | ABD | 7 | 53.8 | 1000 | 10 | 3.21 | 10.0000 | 0.00 | 300 | 0.58 | 84.0 | 77.5 | 6.5 | 6.52 | Y |
| 305 | NF | PHC | 7 | 53.8 | 1000 | 60 | 1.88 | 10.0000 | -0.99 | 300 | 0.59 | 76.9 | 80.2 | -3.3 | 3.29 | Y |
| 314 | NF | SFU | 7 | 53.8 | 1000 | 240 | -0.92 | 10.0000 | -0.01 | 300 | 0.49 | 46.2 | 42.5 | 3.7 | 3.69 | Y |
| 316 | NF | SFX | 7 | 53.8 | 1500 | 10 | 0.79 | 10.0000 | 0.87 | 300 | 0.56 | 29.4 | 45.5 | -16.1 | 16.06 | N |
| 324 | NF200 | 17AED | 7 | 30.3 | 483 | 10 | 4.33 | 0.0050 | 0.00 | 300 | 0.57 | 89.0 | 82.9 | 6.1 | 6.14 | N |
| 329 | NF200 | 24DBA | 4 | 30.3 | 448 | 10 | 1.67 | 1.6000 | 0.89 | 300 | 0.43 | 37.9 | 42.9 | -5.0 | 4.98 | Y |
| 331 | NF200 | 24DBA | 5 | 30.3 | 448 | 10 | 1.67 | 1.6000 | 0.99 | 300 | 0.43 | 70.2 | 70.7 | -0.5 | 0.53 | Y |
| 338 | NF200 | 24DBA | 8.5 | 30.3 | 448 | 10 | 1.67 | 1.6000 | 1.05 | 300 | 0.43 | 89.0 | 87.9 | 1.2 | 1.17 | Y |
| 348 | NF200 | ACA | 7 | 30.3 | 448 | 10 | -0.22 | 12.5000 | 1.00 | 300 | 0.30 | 77.1 | 63.2 | 13.9 | 13.92 | Y |
| 350 | NF200 | ACA | 8 | 30.3 | 448 | 10 | -0.22 | 12.5000 | 1.00 | 300 | 0.30 | 82.9 | 82.3 | 0.6 | 0.63 | Y |
| 384 | NF200 | SFX | 7 | 30.3 | 483 | 10 | 0.79 | 0.0050 | 0.87 | 300 | 0.56 | 84.0 | 70.1 | 13.9 | 13.92 | N |
| 391 | NF270 | ABD | 7 | 73.2 | 1000 | 300 | 3.21 | 10.0000 | 0.00 | 340 | 0.58 | 42.0 | 49.5 | -7.5 | 7.53 | Y |
| 399 | NF270 | HCT | 7 | 73.2 | 1000 | 240 | 1.28 | 10.0000 | 0.00 | 340 | 0.58 | 81.8 | 83.6 | -1.8 | 1.82 | Y |
| 401 | NF270 | PHC | 7 | 73.2 | 1000 | 120 | 1.88 | 10.0000 | -0.99 | 340 | 0.59 | 37.0 | 34.5 | 2.5 | 2.48 | Y |
| 413 | NF270 | TMP | 7 | 73.2 | 1000 | 120 | 1.28 | 10.0000 | -0.93 | 340 | 0.59 | 50.0 | 53.9 | -3.9 | 3.89 | Y |
| 431 | NF270 | DMA | 3.5 | 73.2 | 689 | 10 | 0.70 | 0.1000 | 0.99 | 340 | 0.35 | 27.6 | 31.9 | -4.3 | 4.30 | N |
| 454 | NF270 | PFBS | 6.5 | 73.2 | 523 | 10 | 2.63 | 0.0010 | 1.00 | 340 | 0.43 | 96.0 | 96.2 | -0.2 | 0.20 | Y |
| 455 | NF270 | PFBS | 6.5 | 73.2 | 695 | 10 | 2.63 | 0.0010 | 1.00 | 340 | 0.43 | 96.0 | 93.7 | 2.3 | 2.32 | Y |
| 461 | NF270 | PFHxA | 6.5 | 73.2 | 691 | 10 | 3.71 | 0.0010 | 1.00 | 340 | 0.46 | 95.0 | 95.5 | -0.5 | 0.52 | Y |
| 469 | NF270 | PFOA | 6.5 | 73.2 | 520 | 10 | 5.11 | 0.0010 | 1.00 | 340 | 0.51 | 95.0 | 96.2 | -1.2 | 1.16 | Y |
| 481 | NF270 | SFD | 5 | 64.1 | 690 | 10 | 0.39 | 0.8000 | 0.01 | 340 | 0.53 | 65.9 | 62.7 | 3.2 | 3.21 | N |
| 488 | NF270 | SFX | 6 | 73.2 | 450 | 10 | 0.79 | 0.5000 | 0.41 | 340 | 0.56 | 43.6 | 56.7 | -13.0 | 13.04 | N |

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 498 | NF270 | HCT | 7 | 73.2 | 1000 | 120 | 1.28 | 15.0000 | 0.00 | 340 | 0.58 | 94.9 | 90.8 | 4.1 | 4.07 | Y |
| 503 | NF270 | HCT | 7 | 73.2 | 1000 | 1440 | 1.28 | 15.0000 | 0.00 | 340 | 0.58 | 95.7 | 95.1 | 0.6 | 0.58 | Y |
| 508 | NF270 | LDC | 7 | 73.2 | 1000 | 480 | 2.84 | 15.0000 | -0.85 | 340 | 0.55 | 59.0 | 58.7 | 0.3 | 0.27 | Y |
| 530 | NF270 | SMT | 7 | 29.8 | 345 | 1440 | 0.21 | 0.0014 | 0.66 | 340 | 0.52 | 14.5 | 20.5 | -6.0 | 6.00 | N |
| 536 | NF270 | ACT | 7 | 29.8 | 345 | 1440 | 0.91 | 0.0011 | 0.00 | 340 | 0.44 | 1.3 | 14.9 | -13.6 | 13.56 | N |
| 543 | NF270 | DBT | 7 | 29.8 | 345 | 1440 | 5.19 | 0.0011 | 0.00 | 340 | 0.58 | 41.8 | 25.6 | 16.2 | 16.19 | N |
| 551 | NF270 | SFD | 10 | 64.1 | 690 | 10 | 0.39 | 0.8000 | 1.00 | 340 | 0.53 | 90.0 | 86.4 | 3.6 | 3.60 | N |
| 565 | NF70 | GLT | 7 | 54.1 | 1500 | 10 | -2.93 | 500.0000 | 0.00 | 350 | 0.43 | 95.1 | 75.4 | 19.7 | 19.73 | N |
| 576 | NF70 | PFOA | 6.5 | 25.3 | 600 | 10 | 5.11 | 0.0010 | 1.00 | 350 | 0.51 | 70.8 | 50.1 | 20.7 | 20.67 | N |
| 581 | NF70 | TLE | 7 | 54.1 | 1500 | 10 | 2.49 | 0.4114 | 0.00 | 350 | 0.37 | 77.6 | 48.6 | 29.0 | 28.96 | N |
| 604 | NF90 | ACA | 5 | 63.2 | 483 | 10 | -0.22 | 12.5000 | 0.74 | 118 | 0.30 | 55.7 | 47.8 | 7.9 | 7.88 | N |
| 606 | NF90 | ACA | 6 | 63.2 | 483 | 10 | -0.22 | 12.5000 | 0.97 | 118 | 0.30 | 67.5 | 70.7 | -3.2 | 3.18 | Y |
| 609 | NF90 | ACA | 7.5 | 63.2 | 483 | 10 | -0.22 | 12.5000 | 1.00 | 118 | 0.30 | 80.8 | 82.1 | -1.3 | 1.25 | Y |
| 616 | NF90 | ACT | 4 | 41.4 | 1000 | 10 | 0.91 | 0.5000 | 0.00 | 118 | 0.44 | 91.7 | 89.7 | 2.0 | 1.97 | Y |
| 618 | NF90 | ACT | 7 | 41.4 | 1000 | 10 | 0.91 | 0.5000 | 0.00 | 118 | 0.44 | 91.9 | 89.7 | 2.2 | 2.19 | Y |
| 619 | NF90 | ACT | 4 | 41.4 | 1500 | 10 | 0.91 | 0.5000 | 0.00 | 118 | 0.44 | 91.0 | 91.4 | -0.4 | 0.36 | Y |
| 628 | NF90 | ATZ | 7 | 62.0 | 500 | 10 | 2.20 | 0.0100 | 0.00 | 118 | 0.55 | 81.0 | 87.7 | -6.7 | 6.67 | N |
| 637 | NF90 | BMF | 8 | 59.8 | 280 | 2859 | 2.28 | 0.1000 | 0.00 | 118 | 0.33 | 0.8 | 8.5 | -7.7 | 7.69 | Y |
| 646 | NF90 | CAF | 5 | 41.4 | 500 | 10 | -0.55 | 0.5000 | 0.00 | 118 | 0.47 | 93.8 | 94.5 | -0.7 | 0.68 | Y |
| 675 | NF90 | DCF | 5 | 41.4 | 2000 | 10 | 4.26 | 0.5000 | 0.91 | 118 | 0.54 | 97.8 | 96.4 | 1.4 | 1.43 | Y |
| 676 | NF90 | DCF | 7 | 41.4 | 2000 | 10 | 4.26 | 0.5000 | 1.00 | 118 | 0.54 | 97.8 | 96.4 | 1.4 | 1.35 | Y |
| 677 | NF90 | PFOS | 4 | 54.1 | 1379 | 422 | 5.43 | 10.0000 | -1.00 | 118 | 0.53 | 98.6 | 99.1 | -0.4 | 0.44 | Y |
| 683 | NF90 | DPR | 5 | 41.4 | 500 | 10 | -0.82 | 0.5000 | 1.00 | 118 | 0.54 | 88.5 | 91.2 | -2.7 | 2.66 | Y |
| 702 | NF90 | SFX | 9 | 54.1 | 600 | 10 | 0.79 | 0.5000 | 1.00 | 118 | 0.56 | 100.0 | 103.4 | -3.4 | 3.37 | Y |
| 714 | NF90 | GTA | 6 | 63.2 | 483 | 10 | 0.05 | 11.0000 | 1.98 | 118 | 0.41 | 90.0 | 91.2 | -1.2 | 1.18 | Y |
| 716 | NF90 | GTA | 7 | 63.2 | 483 | 10 | 0.05 | 11.0000 | 2.00 | 118 | 0.41 | 91.9 | 92.3 | -0.3 | 0.35 | Y |

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 732 | NF90 | IBP | 5 | 41.4 | 1000 | 10 | 3.84 | 0.5000 | 0.58 | 118 | 0.50 | 97.7 | 97.1 | 0.6 | 0.60 | N |
| 737 | NF90 | IBP | 4 | 41.4 | 2000 | 10 | 3.84 | 0.5000 | 0.12 | 118 | 0.50 | 96.5 | 96.8 | -0.3 | 0.34 | N |
| 750 | NF90 | FBT | 7 | 54.1 | 1500 | 10 | 4.80 | 10.0000 | -0.87 | 118 | 0.73 | 100.0 | 99.3 | 0.7 | 0.66 | N |
| 754 | NF90 | HCT | 7 | 54.1 | 1000 | 360 | 1.28 | 15.0000 | 0.00 | 118 | 0.58 | 99.4 | 99.5 | -0.1 | 0.11 | Y |
| 758 | NF90 | LDC | 7 | 54.1 | 1000 | 10 | 2.84 | 15.0000 | -0.85 | 118 | 0.55 | 99.3 | 98.8 | 0.5 | 0.52 | Y |
| 759 | NF90 | LDC | 7 | 54.1 | 1000 | 120 | 2.84 | 15.0000 | -0.85 | 118 | 0.55 | 99.4 | 99.2 | 0.2 | 0.25 | Y |
| 780 | NF90 | PRC | 7 | 54.1 | 1000 | 360 | 1.88 | 15.0000 | -0.99 | 118 | 0.59 | 98.6 | 98.9 | -0.3 | 0.34 | Y |
| 799 | NF90 | CTB | 8 | 60.0 | 280 | 2880 | 3.60 | 0.1000 | 0.00 | 118 | 0.34 | 70.0 | 65.2 | 4.8 | 4.79 | N |
| 801 | NF90 | PCE | 8 | 60.0 | 280 | 2880 | 2.62 | 0.1000 | 0.00 | 118 | 0.35 | 57.0 | 38.3 | 18.7 | 18.72 | N |
| 821 | NTR7450 | PHN | 7 | 70.0 | 1500 | 10 | 1.67 | 200.0000 | 0.00 | 310 | 0.37 | 0.0 | -1.1 | 1.1 | 1.13 | N |
| 824 | NTR7450 | XLS | 7 | 70.0 | 1500 | 10 | -2.30 | 250.0000 | 0.00 | 310 | 0.41 | 26.1 | 38.8 | -12.7 | 12.72 | N |
| 827 | RE-BLR | CTB | 8 | 47.0 | 480 | 2880 | 3.60 | 0.1000 | 0.00 | 100 | 0.34 | 99.0 | 101.8 | -2.8 | 2.80 | N |
| 833 | SB50 | BMD | 3.5 | 63.0 | 689 | 10 | 1.68 | 0.1000 | 0.99 | 152 | 0.36 | 84.8 | 76.7 | 8.1 | 8.05 | N |
| 851 | SR2 | IDM | 7.5 | 40.1 | 1000 | 10 | 3.53 | 0.5000 | 1.00 | 460 | 0.67 | 82.0 | 87.4 | -5.4 | 5.40 | N |
| 859 | SR3 | CPX | 7.5 | 44.6 | 1000 | 10 | -2.14 | 0.5000 | 0.65 | 165 | 0.55 | 88.0 | 97.6 | -9.6 | 9.60 | N |
| 867 | SWC1 | DXT | 7 | 48.8 | 1000 | 360 | 1.68 | 15.0000 | 0.00 | 100 | 0.60 | 99.6 | 99.8 | -0.2 | 0.16 | Y |
| 898 | TS80 | AMP | 8 | 48.0 | 500 | 10 | 1.15 | 0.0020 | 0.00 | 200 | 0.52 | 90.6 | 97.6 | -7.0 | 7.05 | N |
| 899 | TS80 | ATL | 8 | 48.0 | 500 | 10 | 0.43 | 0.0020 | -0.98 | 200 | 0.62 | 78.4 | 78.2 | 0.2 | 0.22 | N |
| 901 | TS80 | BPA | 7 | 56.6 | 1034 | 1440 | 4.05 | 0.0021 | 0.00 | 200 | 0.50 | 25.6 | 44.8 | -19.2 | 19.20 | N |
| 905 | TS80 | CBP | 7 | 56.6 | 1034 | 1440 | 2.77 | 0.0011 | 0.00 | 200 | 0.51 | 8.9 | 25.6 | -16.7 | 16.68 | N |
| 908 | TS80 | CBX | 7 | 56.6 | 1034 | 1440 | -0.32 | 0.0013 | 1.00 | 200 | 0.56 | 1.7 | 33.9 | -32.2 | 32.23 | N |
| 909 | TS80 | CCP | 8 | 48.0 | 500 | 10 | 0.10 | 0.0020 | 0.00 | 200 | 0.54 | 87.8 | 91.5 | -3.7 | 3.68 | N |
| 948 | UTC20 | MHM | 7 | 34.8 | 1500 | 10 | 0.72 | 1.5645 | 0.03 | 180 | 0.39 | 29.9 | 16.5 | 13.4 | 13.38 | N |
| 951 | UTC20 | PHN | 7 | 34.8 | 1500 | 10 | 1.67 | 200.0000 | 0.00 | 180 | 0.37 | 6.8 | 10.3 | -3.5 | 3.51 | N |
| 953 | UTC70 | BMF | 8 | 54.4 | 340 | 2880 | 2.28 | 0.1000 | 0.00 | 65 | 0.33 | 34.0 | 28.2 | 5.8 | 5.81 | N |
| 967 | UTC70 | TCE | 8 | 54.4 | 340 | 2880 | 2.18 | 0.1000 | 0.00 | 65 | 0.35 | 6.0 | 12.4 | -6.4 | 6.41 | N |

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 979 | X20 | DEET | 7 | 55.0 | 1034 | 1440 | 2.50 | 0.0010 | 0.00 | 200 | 0.49 | 96.1 | 81.2 | 14.9 | 14.91 | N |
| 986 | X20 | SCP | 7 | 55.0 | 1034 | 1440 | 0.85 | 0.0007 | 0.72 | 200 | 0.56 | 96.3 | 84.0 | 12.3 | 12.28 | N |
| 991 | XLE | 2NPT | 7 | 65.0 | 500 | 10 | 2.66 | 0.1000 | 0.00 | 96 | 0.45 | 57.0 | 74.7 | -17.7 | 17.65 | N |
| 996 | XLE | 2NPT | 8 | 46.9 | 552 | 489 | 2.66 | 0.1000 | 0.02 | 96 | 0.45 | 82.7 | 57.0 | 25.7 | 25.66 | Y |
| 1002 | XLE | BCA | 3.5 | 39.8 | 689 | 10 | 0.79 | 0.1000 | 0.97 | 96 | 0.34 | 93.0 | 90.7 | 2.2 | 2.21 | N |
| 1009 | XLE | CPF | 7 | 65.0 | 1500 | 10 | -0.86 | 10.0000 | -0.05 | 96 | 0.61 | 100.0 | 100.7 | -0.7 | 0.69 | N |
| 1010 | XLE | DBM | 3.5 | 39.8 | 689 | 10 | 1.83 | 0.1000 | 1.00 | 96 | 0.35 | 92.3 | 90.1 | 2.2 | 2.22 | N |
| 1031 | XLE | HCT | 7 | 65.0 | 1000 | 1440 | 1.28 | 15.0000 | 0.00 | 96 | 0.58 | 98.8 | 99.8 | -1.0 | 1.01 | Y |
| 1061 | XLE | TCA | 8 | 46.9 | 552 | 482 | 1.53 | 0.1000 | 1.00 | 96 | 0.35 | 89.9 | 81.8 | 8.1 | 8.09 | Y |
| 1062 | XLE | TCA | 8 | 46.9 | 552 | 1443 | 1.53 | 0.1000 | 1.00 | 96 | 0.35 | 78.8 | 93.8 | -15.0 | 15.03 | Y |
| 1064 | XLE | TMP | 7 | 65.0 | 1500 | 10 | 1.28 | 10.0000 | -0.93 | 96 | 0.59 | 94.8 | 97.9 | -3.1 | 3.11 | N |
| 1068 | XLE | ABD | 7 | 59.6 | 1000 | 180 | 3.21 | 10.0000 | 0.00 | 96 | 0.58 | 80.0 | 74.1 | 5.9 | 5.92 | Y |
| 1069 | XLE | ABD | 7 | 59.6 | 1000 | 240 | 3.21 | 10.0000 | 0.00 | 96 | 0.58 | 69.0 | 66.3 | 2.7 | 2.72 | Y |
| 1090 | XLE | TMP | 7 | 59.6 | 1000 | 60 | 1.28 | 10.0000 | -0.93 | 96 | 0.59 | 94.0 | 93.7 | 0.3 | 0.26 | Y |
| 1094 | XLE | TMP | 7 | 59.6 | 1000 | 300 | 1.28 | 10.0000 | -0.93 | 96 | 0.59 | 96.0 | 93.5 | 2.5 | 2.48 | Y |
| 1101 | XLE440 | BMF | 8 | 39.8 | 410 | 2880 | 2.28 | 0.1000 | 0.00 | 150 | 0.33 | 16.4 | 26.0 | -9.6 | 9.62 | N |
| 1115 | TS80 | ETR | 7 | 37.5 | 500 | 5760 | 4.31 | 0.0020 | 0.00 | 200 | 0.53 | 89.9 | 62.9 | 26.9 | 26.94 | N |
| 1116 | TS80 | LDE | 7 | 37.5 | 500 | 5760 | 4.35 | 0.0020 | 0.00 | 200 | 0.44 | 90.7 | 70.4 | 20.3 | 20.33 | N |
| 1123 | TS80 | GLC | 7 | 37.5 | 500 | 5760 | -2.93 | 0.0020 | 0.00 | 200 | 0.44 | 98.6 | 78.6 | 20.0 | 19.98 | N |
| 1124 | TS80 | SCR | 7 | 37.5 | 500 | 5760 | -4.53 | 0.0020 | 0.00 | 200 | 0.56 | 98.6 | 82.3 | 16.3 | 16.30 | N |
| 1136 | NF270 | DBP | 7 | 26.1 | 500 | 480 | 4.63 | 0.6500 | 0.00 | 340 | 0.64 | 50.0 | 93.1 | -43.1 | 43.11 | N |
| 1153 | VNF1 | DCF | 7 | 36.4 | 500 | 10 | 4.26 | 0.1000 | 1.00 | 240 | 0.54 | 83.8 | 89.8 | -6.1 | 6.07 | N |
| 1161 | VNF2 | DCF | 7 | 79.4 | 500 | 10 | 4.26 | 0.1000 | 1.00 | 150 | 0.54 | 86.2 | 90.3 | -4.1 | 4.12 | N |
| 1169 | Desal51HL | XLS | 7 | 47.0 | 800 | 120 | -2.30 | 0.3003 | 0.00 | 190 | 0.41 | 85.0 | 100.0 | -15.0 | 14.96 | N |
| 1187 | NTR7450 | BZT | 7 | 70.0 | 800 | 120 | 1.83 | 0.2062 | 0.00 | 310 | 0.39 | 8.0 | 9.9 | -1.9 | 1.95 | N |
| 1190 | NTR7450 | BZD | 7 | 70.0 | 800 | 120 | 2.47 | 0.2924 | 0.00 | 310 | 0.46 | 15.0 | 18.8 | -3.8 | 3.77 | N |

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1201 | UTC60 | 24DP | 9 | 51.6 | 300 | 10 | 2.88 | 10.0000 | 0.97 | 150 | 0.41 | 71.0 | 71.5 | -0.6 | 0.59 | N |
| 1225 | Desal51HL | DLA | 3 | 46.0 | 800 | 120 | 2.35 | 0.0499 | -1.00 | 190 | 0.54 | 74.0 | 61.6 | 12.4 | 12.43 | N |
| 1231 | Desal51HL | EOS | 3 | 46.0 | 800 | 120 | 6.20 | 0.1342 | -0.45 | 190 | 0.66 | 100.0 | 88.2 | 11.8 | 11.80 | N |
| 1259 | Desal51HL | ISL | 10 | 35.0 | 800 | 120 | -1.51 | 0.0262 | 0.72 | 190 | 0.43 | 92.0 | 74.1 | 17.9 | 17.93 | N |
| 1260 | Desal5DL | ISL | 10 | 47.0 | 800 | 120 | -1.51 | 0.0262 | 0.72 | 260 | 0.43 | 39.0 | 49.9 | -10.9 | 10.87 | N |
| 1297 | UTC70 | NPX | 5 | 54.4 | 300 | 10 | 2.99 | 1.0000 | 0.87 | 65 | 0.53 | 96.7 | 99.1 | -2.4 | 2.43 | Y |
| 1308 | Desal5DL | PLL | 3 | 46.0 | 800 | 120 | -1.18 | 0.0330 | 0.23 | 260 | 0.45 | 39.0 | 43.4 | -4.4 | 4.45 | N |
| 1315 | UTC60 | CFA | 7 | 51.6 | 300 | 10 | 2.90 | 10.0000 | 1.00 | 150 | 0.47 | 96.9 | 95.1 | 1.7 | 1.74 | Y |
| 1324 | UTC70 | GFB | 7 | 54.4 | 300 | 10 | 4.39 | 10.0000 | 1.00 | 65 | 0.61 | 97.6 | 95.9 | 1.7 | 1.74 | N |
| 1326 | Desal5DL | SRE | 3 | 46.0 | 800 | 120 | -3.89 | 0.0210 | 0.23 | 260 | 0.36 | 26.0 | 41.2 | -15.2 | 15.25 | N |
| 1333 | UTC70 | CFA | 7 | 54.4 | 300 | 10 | 2.90 | 10.0000 | 1.00 | 65 | 0.47 | 97.2 | 96.0 | 1.3 | 1.26 | Y |
| 1342 | UTC60 | CBP | 7 | 51.6 | 300 | 10 | 2.77 | 10.0000 | 0.00 | 150 | 0.51 | 71.0 | 62.2 | 8.8 | 8.77 | N |
| 1378 | XLE | 2BTL | 7 | 65.0 | 860 | 10 | 0.77 | 2000.0000 | 0.00 | 96 | 0.34 | 97.2 | 92.5 | 4.7 | 4.70 | N |
| 1388 | SW30XLE | ETN | 7 | 55.0 | 1550 | 10 | -0.16 | 2000.0000 | 0.00 | 100 | 0.28 | 68.7 | 52.4 | 16.3 | 16.30 | N |
| 1390 | NTR7450 | TPT | 3 | 66.0 | 800 | 120 | -1.09 | 0.0408 | 0.26 | 310 | 0.51 | 45.0 | 58.8 | -13.8 | 13.80 | N |
| 1420 | NF90 | PFHxA | 7.1 | 73.5 | 2000 | 90 | 3.71 | 100.0000 | 1.00 | 118 | 0.46 | 99.8 | 95.5 | 4.4 | 4.35 | Y |
| 1422 | XLE | PFHxA | 7.1 | 67.2 | 500 | 90 | 3.71 | 100.0000 | 1.00 | 96 | 0.46 | 99.7 | 97.2 | 2.5 | 2.48 | Y |
| 1448 | BW30 | PFHxA | 3.5 | 43.7 | 2000 | 90 | 3.71 | 100.0000 | -1.00 | 100 | 0.46 | 99.2 | 100.5 | -1.3 | 1.33 | Y |
| 1449 | SW30XLE | PFHxA | 3.5 | 61.4 | 250 | 90 | 3.71 | 100.0000 | 1.00 | 100 | 0.46 | 96.0 | 98.6 | -2.5 | 2.54 | Y |
| 1464 | NF270 | DTT | 7 | 35.0 | 800 | 1440 | 2.89 | 0.3000 | 1.00 | 340 | 0.61 | 93.0 | 78.3 | 14.7 | 14.65 | N |
| 1467 | NF270 | SCR | 7 | 35.0 | 800 | 1440 | -4.53 | 0.3000 | 0.00 | 340 | 0.56 | 95.0 | 93.1 | 1.9 | 1.89 | N |
| 1469 | NF90 | ACT | 7 | 58.1 | 800 | 1440 | 0.91 | 0.3000 | 0.00 | 118 | 0.44 | 95.0 | 92.3 | 2.7 | 2.75 | N |
| 1486 | NF270 | PFOS | 7 | 23.2 | 1000 | 10 | 5.43 | 0.1000 | 1.00 | 340 | 0.53 | 96.1 | 96.1 | -0.1 | 0.06 | Y |
| 1490 | NF270 | PFOS | 7 | 23.2 | 800 | 10 | 5.43 | 0.0500 | 1.00 | 340 | 0.53 | 95.0 | 95.3 | -0.3 | 0.34 | Y |
| 1491 | NF270 | PFOS | 7 | 23.2 | 1000 | 10 | 5.43 | 0.0500 | 1.00 | 340 | 0.53 | 95.3 | 96.2 | -0.9 | 0.92 | Y |
| 1509 | NTR7450 | PFHxA | 2.8 | 70.0 | 700 | 3575 | 3.71 | 100.0000 | 1.00 | 310 | 0.46 | 96.1 | 95.4 | 0.7 | 0.65 | Y |

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1510 | NTR7450 | PFHxA | 2.6 | 70.0 | 700 | 4285 | 3.71 | 100.0000 | 1.00 | 310 | 0.46 | 96.9 | 96.2 | 0.6 | 0.63 | Y |
| 1529 | NF270 | SFX | 3 | 73.2 | 690 | 10 | 0.79 | 0.8000 | -0.08 | 340 | 0.56 | 23.0 | 30.9 | -7.9 | 7.92 | N |
| 1552 | XLE | SFD | 8 | 65.0 | 690 | 10 | 0.39 | 0.8000 | 0.91 | 96 | 0.53 | 89.0 | 92.3 | -3.3 | 3.30 | N |
| 1566 | NF270 | IBP | 5 | 73.2 | 690 | 10 | 3.84 | 0.8000 | 0.58 | 340 | 0.50 | 84.0 | 74.1 | 9.9 | 9.87 | N |
| 1569 | NF270 | SFX | 8 | 73.2 | 690 | 10 | 0.79 | 0.8000 | 0.99 | 340 | 0.56 | 86.0 | 75.7 | 10.3 | 10.31 | N |
| 1571 | NF270 | TCS | 3 | 73.2 | 690 | 10 | 4.98 | 0.8000 | 0.00 | 340 | 0.52 | 60.0 | 59.4 | 0.6 | 0.64 | Y |
| 1573 | XLE | CBP | 8 | 65.0 | 690 | 10 | 2.77 | 0.8000 | 0.00 | 96 | 0.51 | 82.0 | 88.4 | -6.4 | 6.42 | N |
| 1580 | XLE | TCS | 10 | 65.0 | 690 | 10 | 4.98 | 0.8000 | 1.00 | 96 | 0.52 | 97.0 | 92.8 | 4.2 | 4.18 | N |
| 1604 | NF200 | 15NSA | 3 | 30.3 | 448 | 10 | 1.33 | 1.6000 | -2.00 | 300 | 0.56 | 50.7 | 33.6 | 17.1 | 17.10 | N |
| 1606 | NF200 | 15NSA | 7 | 30.3 | 448 | 10 | 1.33 | 1.6000 | 2.00 | 300 | 0.56 | 87.3 | 86.9 | 0.4 | 0.40 | Y |
| 1614 | ES20 | PTP | 7 | 47.0 | 294 | 10 | 4.69 | 10.0000 | 0.99 | 150 | 0.47 | 98.5 | 91.3 | 7.2 | 7.19 | N |
| 1636 | BW30 | MCPA | 7 | 79.4 | 1000 | 10 | 2.41 | 1.0000 | 1.00 | 100 | 0.52 | 96.2 | 98.6 | -2.4 | 2.39 | Y |
| 1637 | BW30 | MCPA | 7 | 79.4 | 1000 | 10 | 2.41 | 5.0000 | 1.00 | 100 | 0.52 | 99.1 | 98.2 | 0.8 | 0.85 | Y |
| 1639 | BW30 | MCPP | 7 | 79.4 | 1000 | 10 | 2.56 | 1.0000 | 1.00 | 100 | 0.48 | 96.7 | 98.6 | -1.9 | 1.92 | Y |
| 1643 | NF270 | 26DBA | 7 | 17.5 | 1000 | 10 | 2.03 | 1.0000 | 0.00 | 340 | 0.45 | 30.9 | 35.6 | -4.7 | 4.71 | Y |
| 1657 | NF90 | 17BED | 7 | 59.8 | 280 | 10 | 3.75 | 0.0530 | 0.00 | 118 | 0.54 | 95.5 | 95.6 | -0.2 | 0.19 | N |
| 1711 | TS80 | MTC | 7 | 48.0 | 500 | 10 | 1.05 | 0.0020 | 0.00 | 200 | 0.49 | 38.3 | 73.3 | -35.0 | 35.02 | N |
| 1717 | LE440 | PCE | 8 | 41.5 | 510 | 10 | 2.62 | 0.1000 | 0.00 | 100 | 0.35 | 76.0 | 83.0 | -7.0 | 7.00 | N |
| 1718 | LE440 | CTC | 8 | 41.5 | 510 | 10 | 3.00 | 0.1000 | 0.00 | 100 | 0.32 | 69.0 | 81.6 | -12.6 | 12.60 | N |
| 1748 | HL | AMC | 7 | 26.8 | 500 | 30 | -2.01 | 0.0200 | 0.37 | 190 | 0.60 | 92.8 | 97.9 | -5.1 | 5.14 | Y |
| 1754 | HL | CFA | 7 | 26.8 | 500 | 1440 | 2.90 | 0.0200 | 1.00 | 190 | 0.47 | 91.7 | 86.9 | 4.8 | 4.78 | Y |
| 1759 | HL | GFB | 7 | 26.8 | 500 | 1440 | 4.39 | 0.0200 | 1.00 | 190 | 0.61 | 95.9 | 90.5 | 5.4 | 5.39 | Y |
| 1760 | HL | SFX | 7 | 26.8 | 500 | 1440 | 0.79 | 0.0200 | 0.87 | 190 | 0.56 | 89.0 | 83.1 | 5.9 | 5.92 | Y |
| 1774 | HL | ETM | 7 | 26.8 | 500 | 1440 | 2.60 | 0.0200 | -1.00 | 190 | 0.80 | 99.5 | 90.3 | 9.1 | 9.14 | Y |
| 1782 | NF270 | SFX | 7 | 31.3 | 500 | 30 | 0.79 | 0.0200 | 0.87 | 340 | 0.56 | 93.0 | 91.1 | 1.9 | 1.91 | Y |
| 1798 | NF270 | CFA | 7 | 31.3 | 500 | 1440 | 2.90 | 0.0200 | 1.00 | 340 | 0.47 | 93.5 | 93.5 | 0.0 | 0.05 | Y |

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Pred (%) | Errors (%) | Abs Errors (%) | Data leakage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1812 | NF270 | NAT | 7 | 31.3 | 500 | 1440 | 0.77 | 0.0200 | -0.26 | 340 | 0.68 | 91.0 | 89.7 | 1.3 | 1.28 | Y |
| 1813 | NF270 | SPR | 7 | 31.3 | 500 | 1440 | 0.22 | 0.0200 | -0.96 | 340 | 0.63 | 82.1 | 77.3 | 4.8 | 4.82 | Y |
| 1819 | NF270 | RXT | 7 | 31.3 | 500 | 1440 | 3.00 | 0.0200 | -0.99 | 340 | 0.99 | 97.2 | 96.3 | 0.8 | 0.83 | Y |
| 1847 | NF90 | GFB | 7 | 37.0 | 500 | 1440 | 4.39 | 0.0200 | 1.00 | 118 | 0.61 | 97.4 | 94.9 | 2.4 | 2.42 | Y |
| 1851 | NF90 | SFZ | 7 | 37.0 | 500 | 1440 | 0.65 | 0.0200 | 0.50 | 118 | 0.59 | 99.0 | 98.0 | 1.1 | 1.06 | Y |
| 1852 | NF90 | DCF | 7 | 37.0 | 500 | 1440 | 4.26 | 0.0200 | 1.00 | 118 | 0.54 | 98.4 | 95.7 | 2.7 | 2.70 | Y |
| 1862 | NF90 | ETM | 7 | 37.0 | 500 | 1440 | 2.60 | 0.0200 | -1.00 | 118 | 0.80 | 99.4 | 97.9 | 1.5 | 1.50 | Y |
| 1864 | ESPA1 | CFA | 7 | 22.6 | 500 | 30 | 2.90 | 0.0200 | 1.00 | 200 | 0.47 | 97.1 | 97.0 | 0.1 | 0.13 | Y |
| 1865 | ESPA1 | NPX | 7 | 22.6 | 500 | 30 | 2.99 | 0.0200 | 1.00 | 200 | 0.53 | 100.0 | 97.1 | 2.9 | 2.86 | Y |
| 1871 | ESPA1 | PPL | 7 | 22.6 | 500 | 30 | 2.58 | 0.0200 | -1.00 | 200 | 0.59 | 100.0 | 102.6 | -2.6 | 2.64 | Y |
| 1876 | ESPA1 | CPN | 7 | 22.6 | 500 | 30 | 0.88 | 0.0200 | 0.02 | 200 | 0.57 | 96.9 | 92.9 | 4.0 | 4.03 | Y |
| 1891 | ESPA1 | GFB | 7 | 22.6 | 500 | 1440 | 4.39 | 0.0200 | 1.00 | 200 | 0.61 | 100.0 | 93.8 | 6.2 | 6.24 | Y |
| 1894 | ESPA1 | MTR | 7 | 22.6 | 500 | 1440 | 1.76 | 0.0200 | -1.00 | 200 | 0.66 | 93.8 | 95.2 | -1.4 | 1.41 | Y |
| 1905 | ESPA1 | DTZ | 7 | 22.6 | 500 | 1440 | 2.73 | 0.0200 | -0.94 | 200 | 0.72 | 98.4 | 97.0 | 1.4 | 1.38 | Y |

**Table A4.** Predictions on the removal rates of micropollutants with different data split fractions. Data were randomly split into training/validation and testing set (potential leakage issue).

| No. | MAE of predictions with different data splitting (training/validation : testing) ratios | | |
|---|---|---|---|
| | 90:10 | 80:20 | 70:30 |
| 1 | 6.71 | 6.72 | 7.15 |
| 2 | 7.09 | 6.93 | 7.69 |
| 3 | 7.31 | 7.25 | 7.3 |
| 4 | 5.68 | 7.05 | 7.23 |
| 5 | 6.25 | 7.36 | 7.70 |
| 6 | 6.73 | 6.35 | 6.90 |
| 7 | 6.70 | 6.77 | 7.24 |
| 8 | 6.74 | 6.17 | 6.63 |
| 9 | 6.57 | 7.23 | 7.34 |
| 10 | 6.12 | 6.77 | 7.37 |
| Average | $6.59 \pm 0.47$ | $6.86 \pm 0.39$ | $7.26 \pm 0.32$ |

**Table A5.** Removal rates of micropollutants by commercial membranes. Removal rates vary depending on the reference. Features in the tables are identification (ID), type of membrane (MB), compounds (C), pH, membrane contact angle (MCA), pressure (P), measurement time (T), compound log $K_{ow}$, initial concentration of compound ($C_{in}$), total charge (TC), MWCO, compound size (CS), removal rates, and reference (Ref).

| ID | MB | C | pH | MCA (°) | P (kPa) | T (min) | log $K_{ow}$ | $C_{in}$ (mg/L) | TC | MWCO | CS (nm) | Removal Rates (%) | Ref |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 449 | NF270 | SFX | 7 | 73.16 | 1500 | 10 | 0.791 | 10 | 0.90 | 340 | 0.6 | 15.4 | [9] |
| 518 | NF270 | SFX | 7 | 73.16 | 450 | 10 | 0.791 | 0.5 | 0.90 | 340 | 0.6 | 82.13 | [10] |
| 519 | NF270 | SFX | 7 | 73.16 | 450 | 10 | 0.791 | 0.5 | 0.90 | 340 | 0.6 | 81.89 | [10] |
| 1366 | NF270 | SFX | 7 | 73.16 | 1000 | 10 | 0.791 | 200 | 0.90 | 340 | 0.6 | 82 | [11] |
| 1457 | NF270 | SFX | 7 | 73.16 | 340 | 10 | 0.791 | 0.1 | 0.90 | 340 | 0.6 | 87 | [12] |
| 1461 | NF270 | CBP | 7 | 35 | 800 | 1440 | 2.766 | 0.3 | 0 | 340 | 0.51 | 31 | [13] |
| 1565 | NF270 | CBP | 8 | 73.16 | 690 | 10 | 2.766 | 0.8 | 0 | 340 | 0.51 | 36 | [14] |
| 1779 | NF270 | CBP | 7 | 31.3 | 500 | 30 | 2.766 | 0.02 | 0 | 340 | 0.51 | 87.11 | [15] |
| 1801 | NF270 | CBP | 7 | 31.3 | 500 | 1440 | 2.766 | 0.02 | 0 | 340 | 0.51 | 87.78 | [15] |
| 78 | BW400 | BMF | 8 | 56.8 | 620 | 104.38 | 2.276 | 0.1 | 0 | 100 | 0.33 | 95.08 | [16] |
| 79 | BW400 | BMF | 8 | 56.8 | 620 | 340.07 | 2.276 | 0.1 | 0 | 100 | 0.33 | 86.89 | [16] |
| 80 | BW400 | BMF | 8 | 56.8 | 620 | 707.07 | 2.276 | 0.1 | 0 | 100 | 0.33 | 70.08 | [16] |
| 81 | BW400 | BMF | 8 | 56.8 | 620 | 1427.61 | 2.276 | 0.1 | 0 | 100 | 0.33 | 38.12 | [16] |
| 82 | BW400 | BMF | 8 | 56.8 | 620 | 2865.32 | 2.276 | 0.1 | 0 | 100 | 0.33 | 15.16 | [16] |

## References


1. Chemicalize https://chemicalize.com
2. Jović, A.; Brkić, K.; Bogunović, N. In A review of feature selection methods with applications, *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 25-29 May, 2015; 2015; pp 1200-1205.
3. Krogh, A.; Vedelsby, J., Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems* 1994**,** *7*.
4. Wold, S.; Esbensen, K.; Geladi, P., Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 1987, 2, (1), 37-52.
5. Khaouane, L.; Ammi, Y.; Hanini, S., Modeling the retention of organic compounds by nanofiltration and reverse osmosis membranes using bootstrap aggregated neural networks. *Arabian Journal for Science and Engineering* 2017, 42, (4), 1443-1453.
6. Ammi, Y.; Khaouane, L.; Hanini, S., Prediction of the rejection of organic compounds (neutral and ionic) by nanofiltration and reverse osmosis membranes using neural networks. *Korean Journal of Chemical Engineering* 2015, 32, (11), 2300-2310.
7. Yangali-Quintanilla, V.; Sadmani, A.; McConville, M.; Kennedy, M.; Amy, G., A QSAR model for predicting rejection of emerging contaminants (pharmaceuticals, endocrine disruptors) by nanofiltration membranes. *Water Research* 2010, 44, (2), 373-384.
8. 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. *Journal of Membrane Science* 2009, 342, (1), 251-262.
9. Dolar, D.; Vuković, A.; Ašperger, D.; Košutić, K., Effect of water matrices on removal of veterinary pharmaceuticals by nanofiltration and reverse osmosis membranes. *Journal of Environmental Sciences* 2011, 23, (8), 1299-1307.
10. Nghiem, L. D.; Schäfer, A. I.; Elimelech, M., Pharmaceutical retention mechanisms by nanofiltration membranes. *Environmental Science & Technology* 2005, 39, (19), 7698-7705.
11. Guo, H.; Yao, Z.; Yang, Z.; Ma, X.; Wang, J.; Tang, C. Y., A one-step rapid assembly of thin film coating using green coordination complexes for enhanced removal of trace organic contaminants by membranes. *Environmental Science & Technology* 2017, 51, (21), 12638-12643.
12. Steinle-Darling, E.; Litwiller, E.; Reinhard, M., Effects of sorption on the rejection of trace organic contaminants during nanofiltration. *Environmental Science & Technology* 2010, 44, (7), 2592-2598.
13. Azaïs, A.; Mendret, J.; Gassara, S.; Petit, E.; Deratani, A.; Brosillon, S., Nanofiltration for wastewater reuse: Counteractive effects of fouling and matrice on the rejection of pharmaceutical active compounds. *Separation and Purification Technology* 2014, 133, 313-327.
14. Lin, Y.; Lee, C., Elucidating the rejection mechanisms of PPCPs by nanofiltration and reverse osmosis membranes. *Industrial & Engineering Chemistry Research* 2014, 53, (16), 6798-6806.
15. Zhao, Y.; Kong, F.; Wang, Z.; Yang, H.; Wang, X.; Xie, Y. F.; Waite, T. D., Role of membrane and compound properties in affecting the rejection of pharmaceuticals by different RO/NF membranes. *Frontiers of Environmental Science & Engineering* 2017, 11, (6), 20.
16. Huang, H.; Cho, H.; Schwab, K.; Jacangelo, J. G., Effects of feedwater pretreatment on the removal of organic microconstituents by a low fouling reverse osmosis membrane. Desalination 2011, 281, 446-454.

# Appendix B

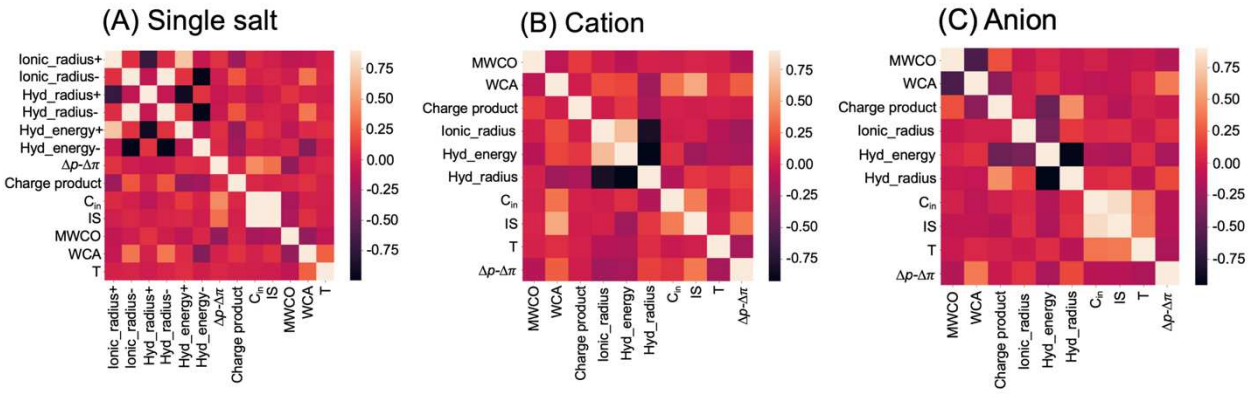


**Figure B1.** The correlation matrices of input variables for (A) ion rejection in single salt solutions, (B) cation, and (C) anion rejections in mixture salt solutions. The input variables in the single salt rejection are ionic radius of cation (Ionic_radius+), ionic radius of anion (Ionic_radius-), hydrated radius of cation (Hyd_radius+), hydrated radius of anion (Hyd_radius-), hydration energy of cation (Hyd_energy+), hydration energy of anion (Hyd_energy-), the difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$), charge product, the initial solute concentration ($C_{in}$), ionic strength (IS), molecular weight cut-off (MWCO), water contact angle (WCA), and measurement time (T). For cation and anion rejections for mixture salt solutions, all the input variables are identical with the single salt dataset, except for only considering properties (ionic radius, hydrated radius, and hydration energy) of either cation or anion.

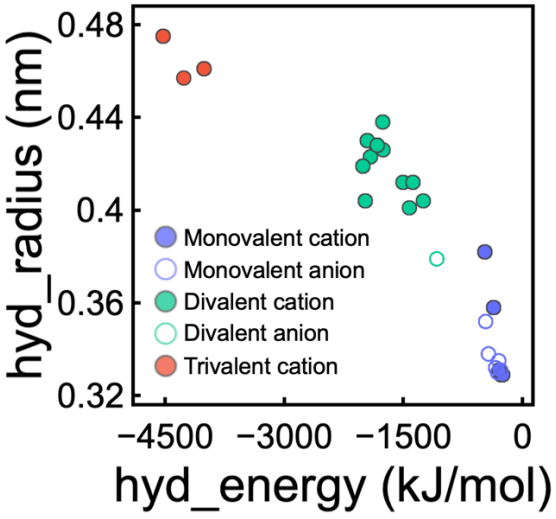


**Figure B2.** Hydrated radii of cations and anions in our dataset as a function of hydration energy. The hydrated radii and hydration energies have a clear negative correlation for cations and anions, thereby providing the same information to the ML model.

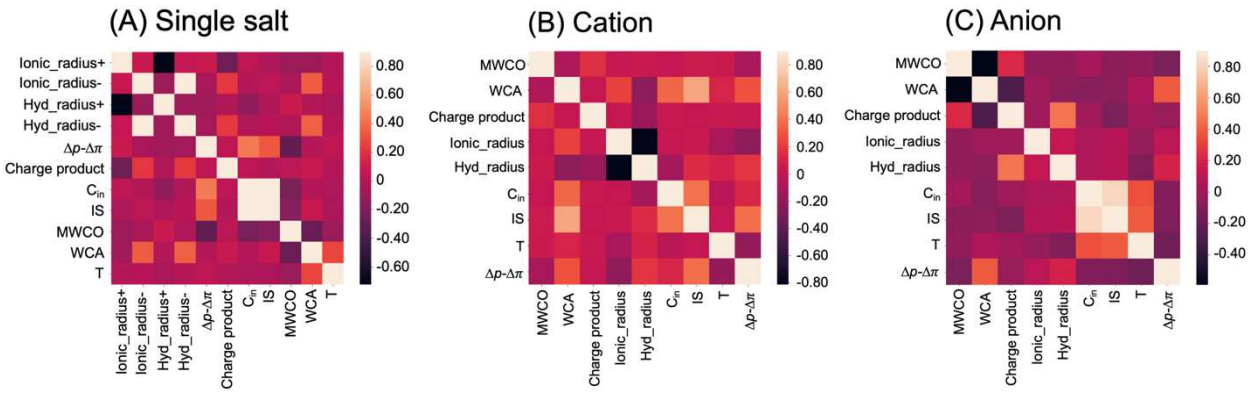


**Figure B3.** The correlation matrices of input variables for (A) ion rejection in single salt solutions, (B) cation, and (C) anion rejections in mixture salt solutions after excluding hydration energy. The input variables in the single salt rejection are ionic radius of cation (Ionic_radius+), ionic radius of anion (Ionic_radius-), hydrated radius of cation (Hyd_radius+), hydrated radius of anion (Hyd_radius-), the difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$), charge product, the initial solute concentration ($C_{in}$), ionic strength (IS), molecular weight cut-off (MWCO), water contact angle (WCA), and measurement time (T). For cation and anion rejections for mixture salt solutions, all the input variables are identical with the single salt dataset, except for considering only properties (ionic radius and hydrated radius) of either cation or anion.

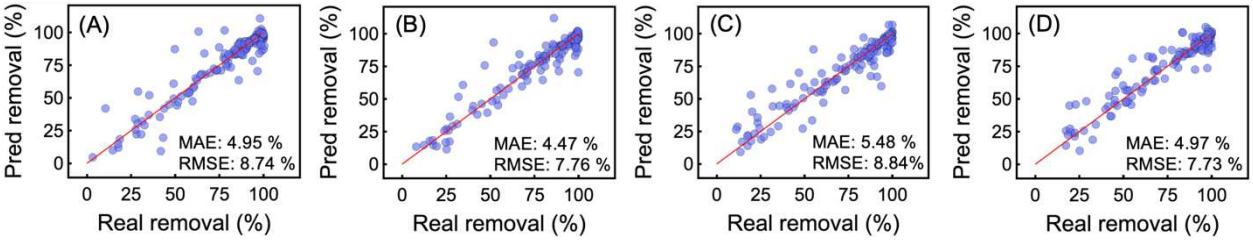


**Figure B4.** The prediction accuracy of the XGBoost model for the prediction of single salt rejection when data leakage occurs. The data split ratio for training/validation and testing data is 80/20. Different training/validation and testing data were used in Figures B4A–D to train the ML model and predict the single salt rejections. The red lines indicate the points where the real and predicted removal rates are equal.

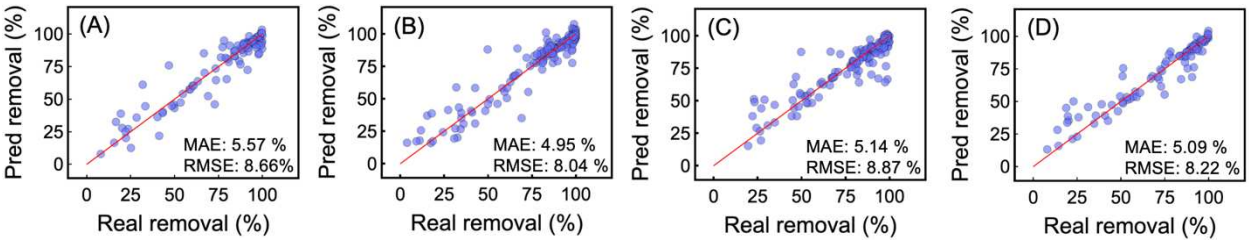


**Figure B5.** The prediction accuracy of the XGBoost model for the prediction of cation rejection for mixture salt solutions when data leakage occurs. The data split ratio for training/validation and testing data is 80/20. Different training/validation and testing data were used in Figures B5A–D to train the ML model and predict cation rejections in mixture salt solutions. The red lines indicate the points where the real and predicted removal rates are equal.

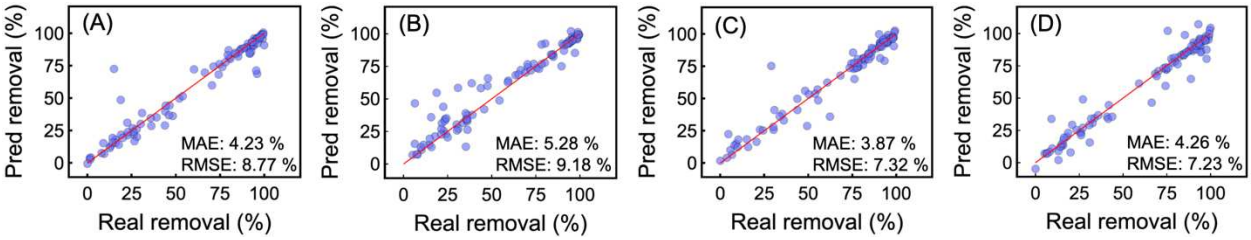


**Figure B6.** The prediction accuracy of the XGBoost model for the prediction of anion rejection for mixture salt solutions when data leakage occurs. The data split ratio for training/validation and testing data is 80/20. Different training/validation and testing data were used in Figures B6A–D to train the ML model and predict anion rejections in mixture salt solutions. The red lines indicate the points where the real and predicted removal rates are equal.

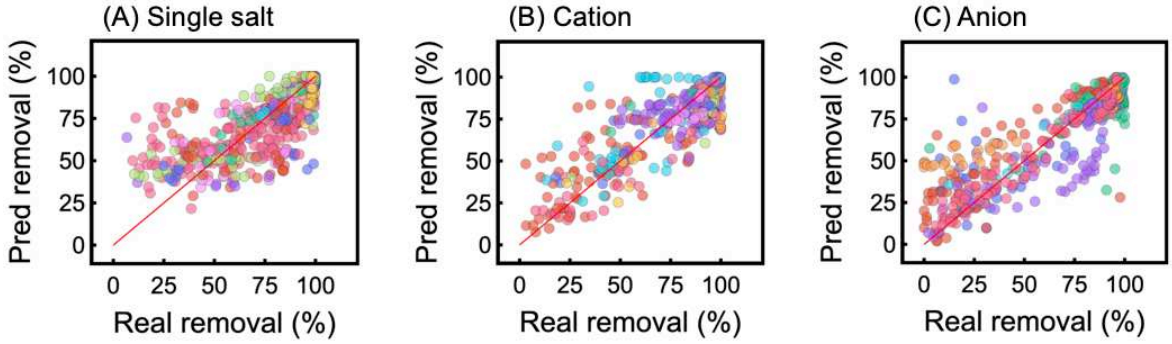


**Figure B7.** The prediction accuracy of XGBoost model for (A) single salt rejection in single salt solution, (B) cation, and (C) anion rejection in mixture salt solutions. The different types of salts/ions are presented in different colors. Due to a high number of single salt/ion type in each figure, the single salt/ion type is not indicated (the numbers of single salts/ions in Figure B7A, B7B, and S7C were 24, 20, and 7, respectively).

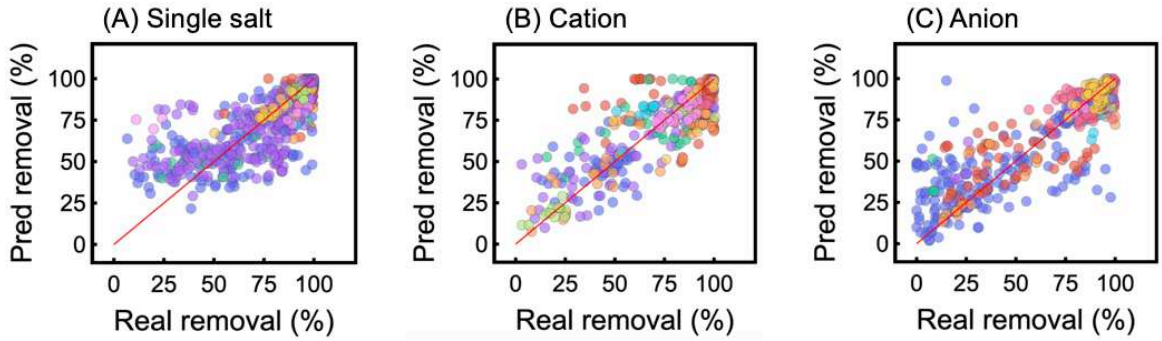


**Figure B8.** The prediction accuracy of XGBoost model for (A) single salt rejection in single salt solution, (B) cation, and (C) anion rejection in mixture salt solutions. The different types of membranes are presented in different colors. Due to a high number of membrane type in each figure, the membrane type is not indicated (the numbers of membrane types in Figure B8A, B8B, and B8C were 24, 10, and 11, respectively).

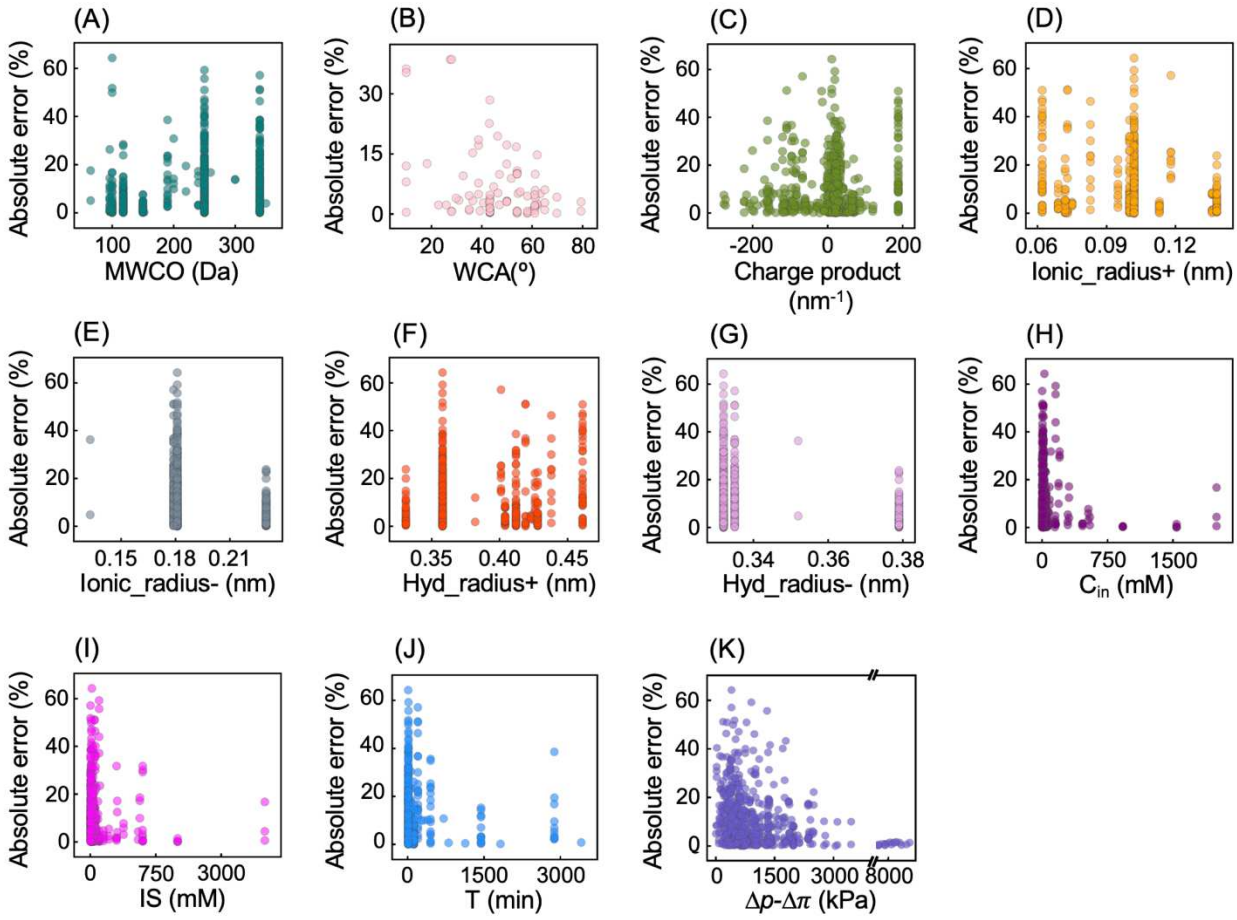


**Figure B9.** The absolute errors of XGBoost model for the prediction of single salt rejection as a function of (A) molecular weight cut-off (MWCO), (B) water contact angle (WCA), (C) charge product, (D) ionic radius of cation (Ionic_radius+), (E) ionic radius of anion (Ionic_radius-), (F) hydrated radius of cation (Hyd_radius+), (G) hydrated radius of anion (Hyd_radius-), (H) initial solute concentration ($C_{in}$), (I) ionic strength (IS), (J) measurement time (T), and (K) the difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$).

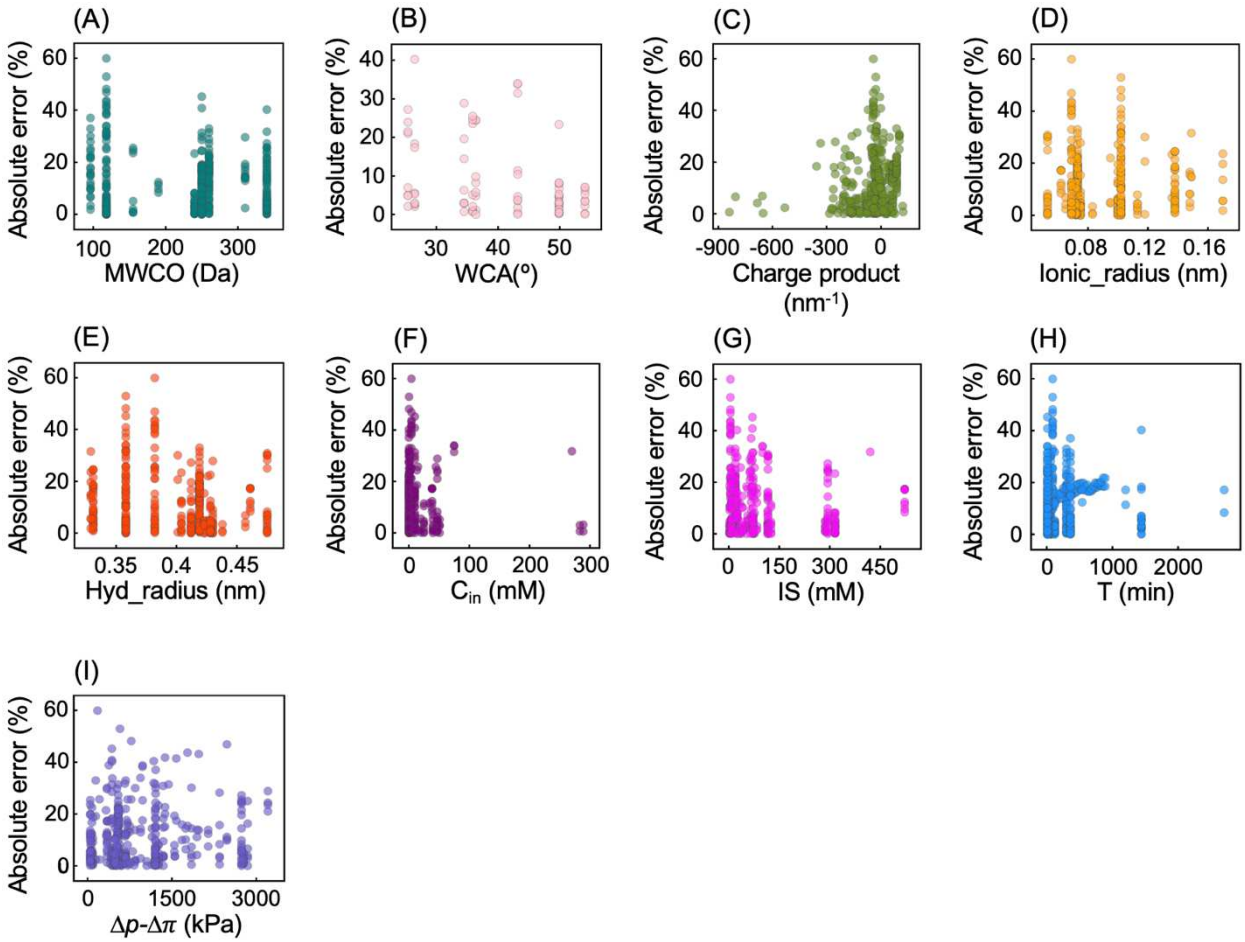


**Figure B10.** The absolute errors of the XGBoost model for the prediction of cation rejection in mixture salt solutions as a function of (A) molecular weight cut-off (MWCO), (B) water contact angle (WCA), (C) charge product, (D) ionic radius, (E) hydrated radius (Hyd_radius), (F) initial solute concentration ($C_{in}$), (G) ionic strength (IS), (H) measurement time (T), and (I) the difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$).

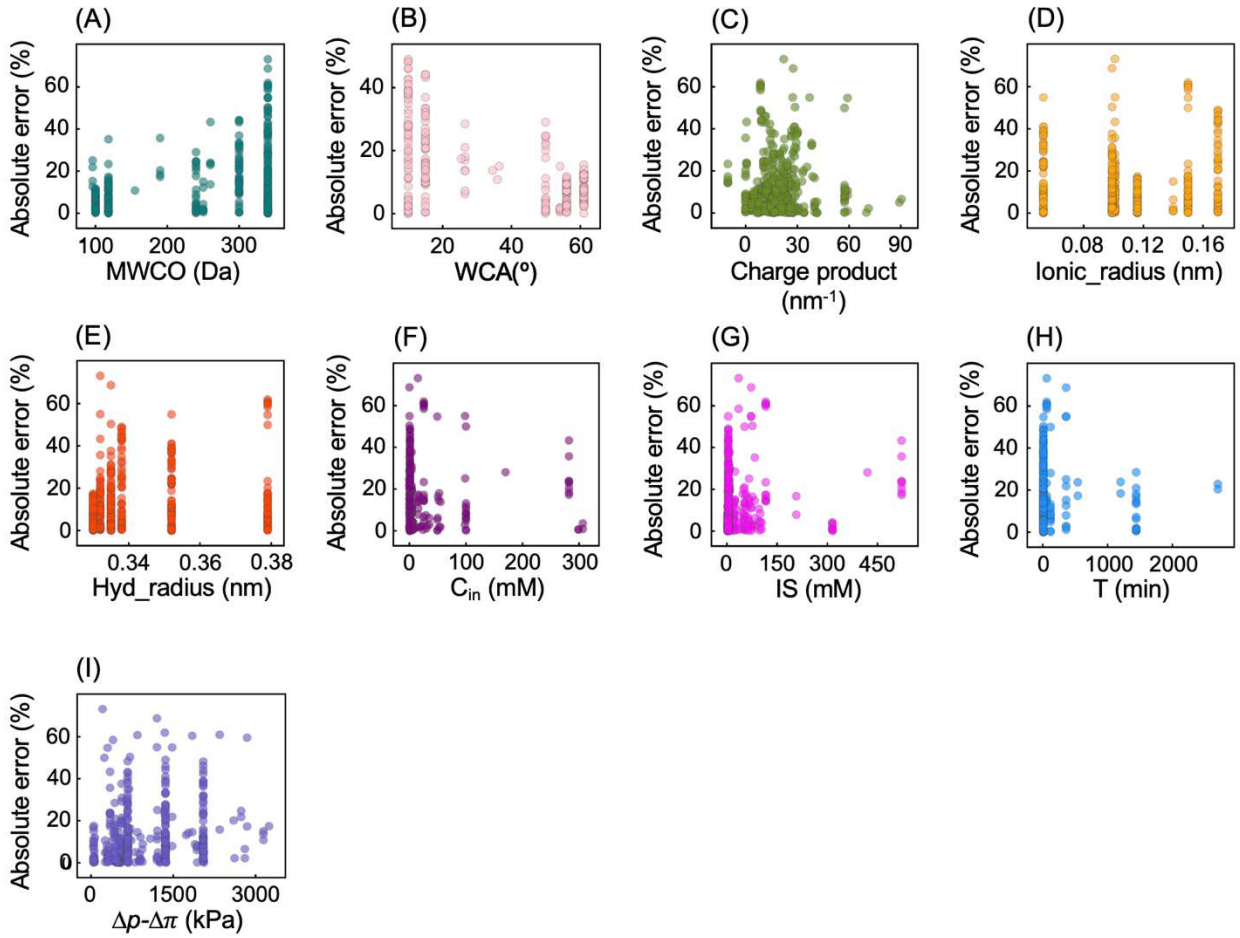


**Figure B11.** The absolute errors of the XGBoost model for the prediction of anion rejection in mixture salt solutions as a function of (A) molecular weight cut-off (MWCO), (B) water contact angle (WCA), (C) charge product, (D) ionic radius, (E) hydrated radius (Hyd_radius), (F) initial solute concentration ($C_{in}$), (G) ionic strength (IS), (H) measurement time (T), and (I) the difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$).

-

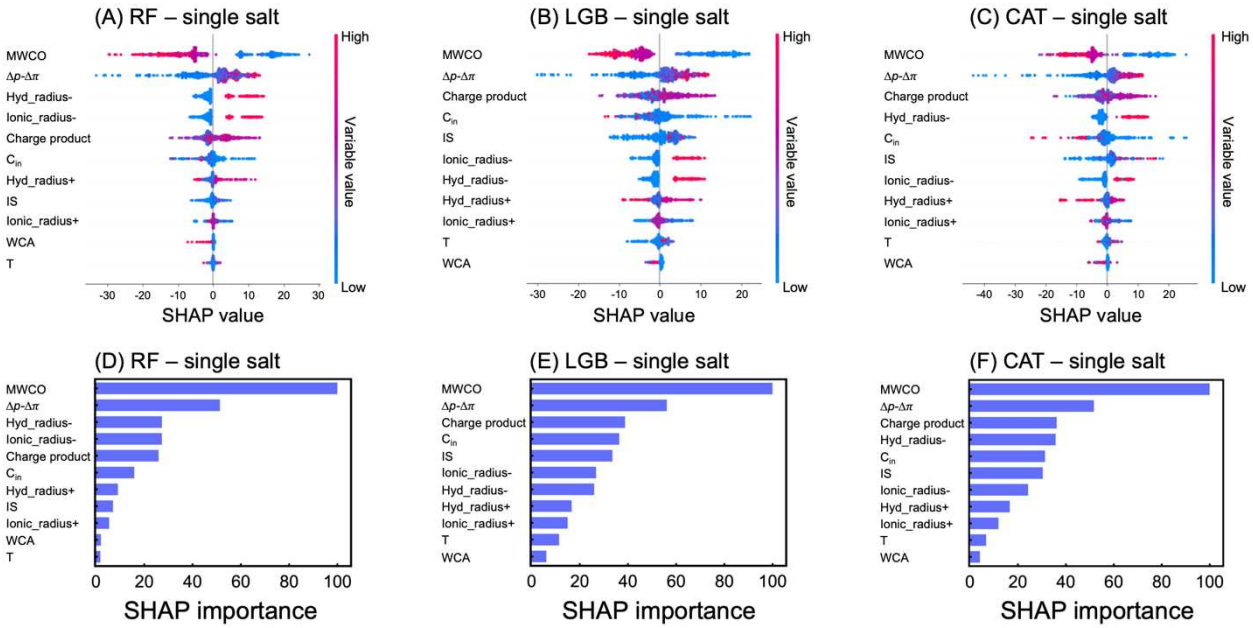


**Figure B12.** (A-C) The SHAP summary plots and (D-F) the SHAP importance for single salt rejection prediction using (A and D) Random Forest (RF), (B and E) LightGBM (LGB), and (C and F) Catboost (CAT) algorithms. The scale of the variable value for the SHAP summary plot is presented by red (high) and blue (low) colors. The SHAP importance of each variable is normalized by the SHAP importance of the first-rank variable (defined as 100) for ease of comparison.

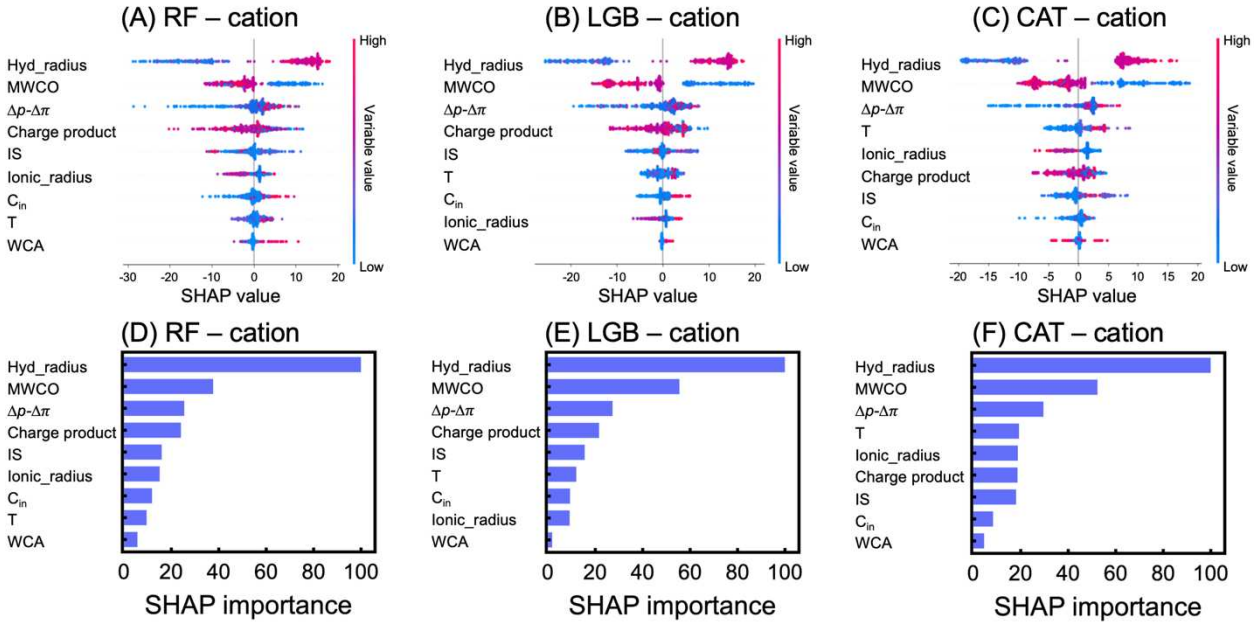


**Figure B13.** (A-C) The SHAP summary plots and (D-F) the SHAP importance for cation rejection prediction in mixture salt solution using (A and D) random forest (RF), (B and E) LightGBM (LGB), and (C and F) Catboost (CAT) algorithms. The scale of the variable value for the SHAP summary plot is presented by red (high) and blue (low) colors. The SHAP importance of each variable is normalized by the SHAP importance of the first-rank variable (defined as 100) for ease of comparison.

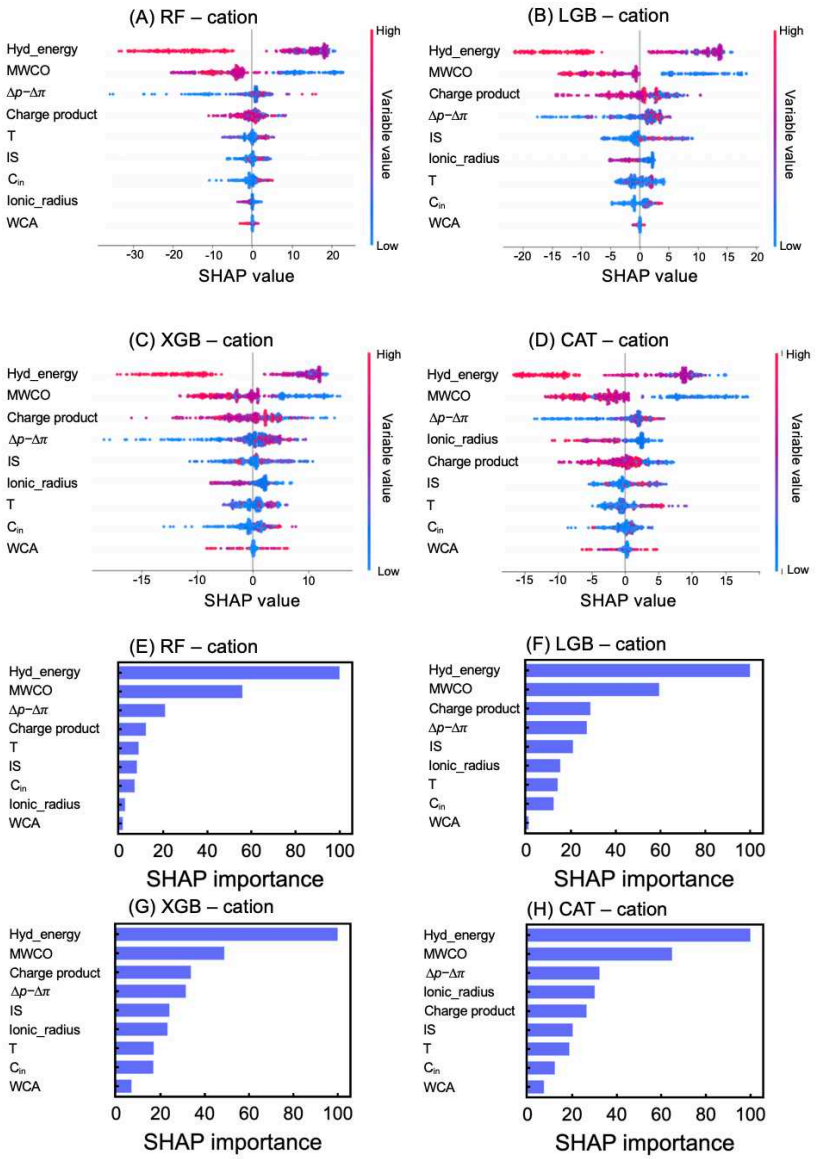


**Figure B14.** (A-D) The SHAP summary plots and (E-H) the SHAP importance for cation rejection prediction in mixture salt solution using (A and E) random forest (RF), (B and F) LightGBM (LGB), (C and G) XGBoost (XGB), and (D and H) Catboost (CAT) algorithms when hydrated radius (hyd_radius) is replaced by hydration energy (Hyd_energy) for ML model training. The scale of the variable value for the SHAP summary plot is presented by red (high) and blue (low) colors. The SHAP importance of each variable is normalized by the SHAP importance of the first-rank variable (defined as 100) for ease of comparison.

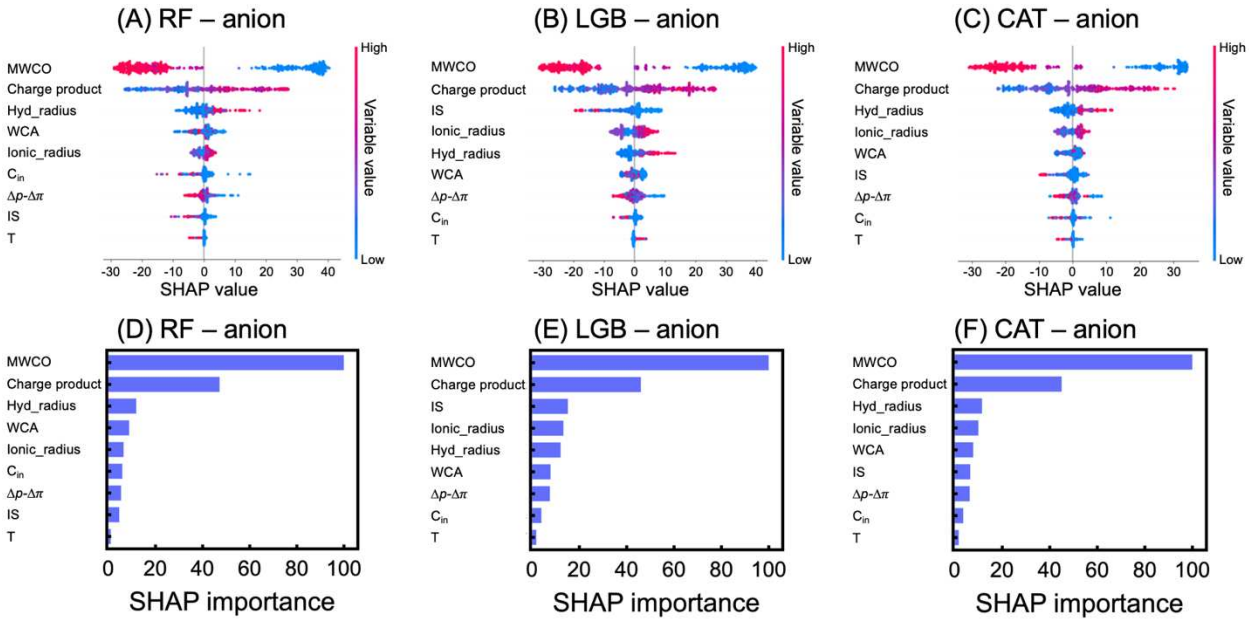


**Figure B15.** (A-C) The SHAP summary plots and (D-F) the SHAP importance for anion rejection prediction in mixture salt solution using (A and D) random forest (RF), (B and E) LightGBM (LGB), and (C and F) Catboost (CAT) algorithms. The scale of the variable value for the SHAP summary plot is presented by red (high) and blue (low) colors. The SHAP importance of each variable is normalized by the SHAP importance of the first-rank variable (defined as 100) for ease of comparison.

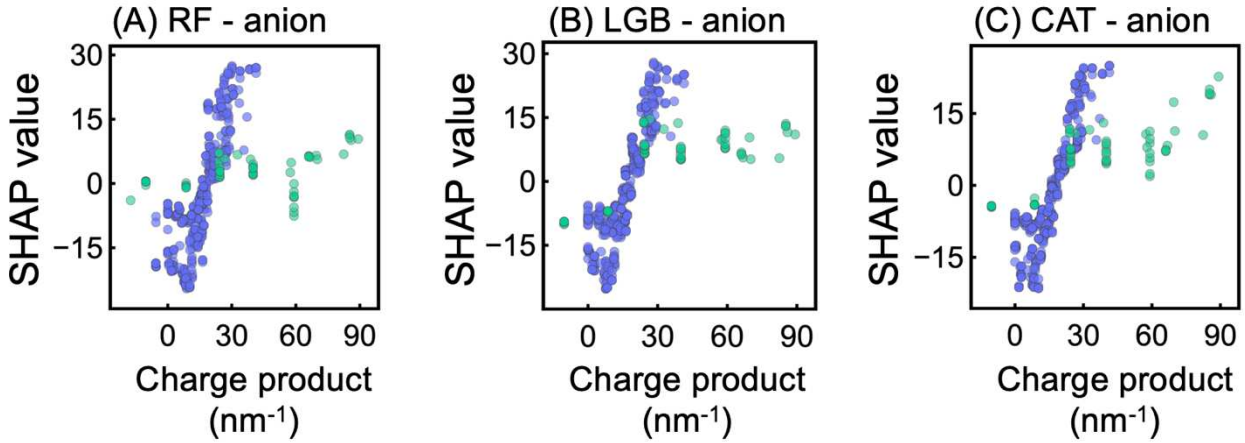


**Figure B16.** The SHAP dependence plots of charge product for anion rejection prediction by (A) Random Forest (RF), (B) LightGBM (LGB), and (C) Catboost (CAT) algorithms. The blue and green dots represent monovalent and divalent anions.

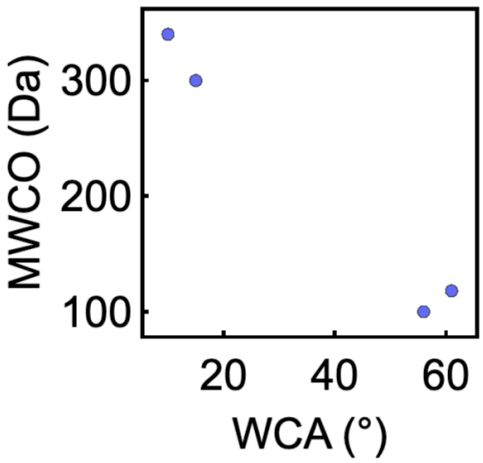


**Figure B17.** The MWCO values as a function of contact angle for anion rejection data obtained from our experiment.

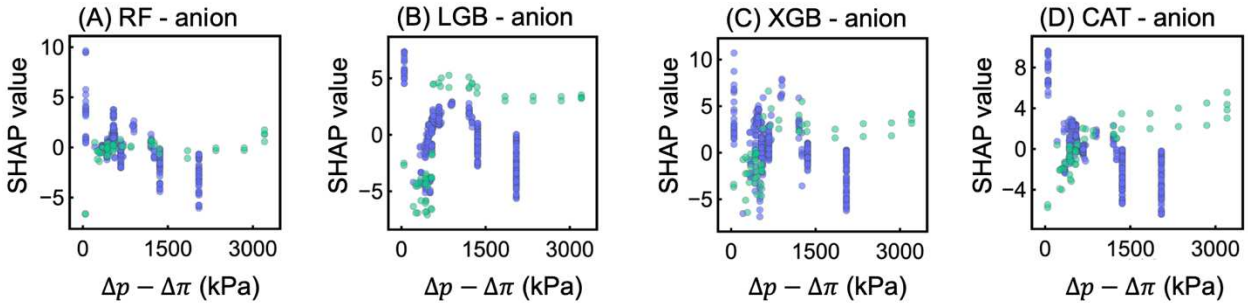


**Figure B18.** The SHAP dependence plot of $\Delta p$-$\Delta \pi$ for anion rejection prediction by (A) random forest (RF), (B) LightGBM (LGB), (C) XGBoost (XGB), and (D) Catboost (CAT) algorithms. The blue and green dots represent monovalent and divalent anions.

**Table B1.** The lists of input variables for the predictions of single salt rejection, cation, and anion rejection in mixture salt solutions. The variables labeled in blue and pink represent those related to size exclusion and electrostatic interaction. The variable of hydrated radius is related to both size exclusion and ion dehydration, which is labeled in green. The variables labeled in white are those that are difficult to be categorized in specific membrane separation mechanisms.

| **Single salt** | **Cation** | **Anion** |
|---|---|---|
| Molecular weight cut-off (MWCO) | Molecular weight cut-off (MWCO) | Molecular weight cut-off (MWCO) |
| Ionic radius of cation (Ionic_radius+) | Ionic radius (Ionic_radius) | Ionic radius (Ionic_radius) |
| Ionic radius of anion (Ionic_radius-) | Hydrated radius (Hyd_radius) | Hydrated radius (Hyd_radius) |
| Hydrated radius of cation (Hyd_radius+) | Charge product | Charge product |
| Hydrated radius of anion (Hyd_radius-) | Ionic strength (IS) | Ionic strength (IS) |
| Charge product | Water contact angle (WCA) | Water contact angle (WCA) |
| Ionic strength (IS) | The difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$) | The difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$) |
| Water contact angle (WCA) | Initial solute concentration ($C_{in}$) | Initial solute concentration ($C_{in}$) |
| The difference between transmembrane pressure and feedwater osmotic pressure ($\Delta p$-$\Delta\pi$) | Measurement time (T) | Measurement time (T) |
| Initial solute concentration ($C_{in}$) | | |
| Measurement time (T) | | |

**Table B2.** The list of hyper-parameters and their ranges that were used for optimizing the random forest model.

| Hyper-parameters | Range |
|---|---|
| N_estimators | 10 – 1000 |
| Max_depth | 3 – 10 |
| Min_samples_split | 2 – 10 |

**Table B3.** The list of hyper-parameters and their ranges that were used for optimizing the LightGBM model.

| Hyper-parameters | Range |
|---|---|
| N_estimators | 10 – 1000 |
| Learning_rate | 0.0001 – 0.1 |
| Max_depth | 3 – 10 |
| Lambda_l1 | 0 – 100 |
| Lambda_l2 | 0 – 100 |
| Bagging_fraction | 0.1 – 1 |
| Feature_fraction | 0.1 – 1 |

**Table B4.** The list of hyper-parameters and their ranges that were used for optimizing the XGBoost model.

| Hyper-parameters | Range |
|---|---|
| Colsample_bytree | 0.5 – 0.9 |
| Gamma | 0 – 1 |
| Learning_rate | 0.01 – 0.3 |
| Max_depth | 3 – 10 |
| Min_child_weight | 0.0001 – 5 |
| Reg_alpha | 0 – 1 |
| Reg_lambda | 0 – 1 |
| Subsample | 0.5 – 0.9 |

**Table B5.** The list of hyper-parameters and their ranges that were used for optimizing the Catboost model.

| Hyper-parameters | Range |
|---|---|
| Learning_rate | 0.0001 – 0.1 |
| Max_depth | 3 – 10 |
| L2_leaf_reg | 2 – 10 |
| Random_strength | 0 – 10 |

# Appendix C

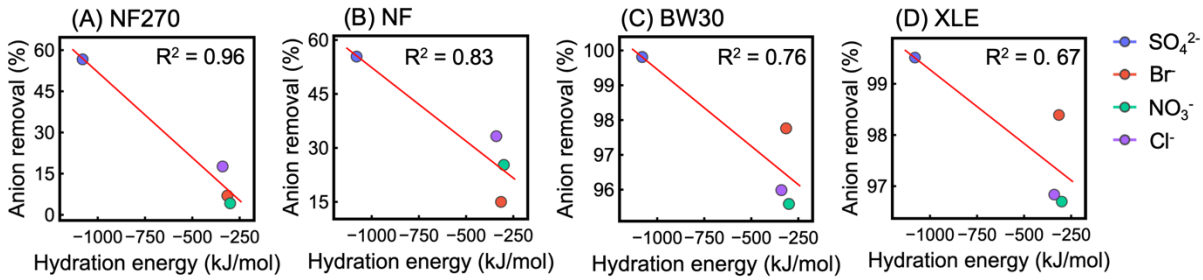


**Figure C1.** Anion rejections as a function of hydration energy after coagulation and microfiltration followed by treatment using (A) NF270 membrane, (B) NF membrane, (C) BW30 membrane and (D) XLE membrane.

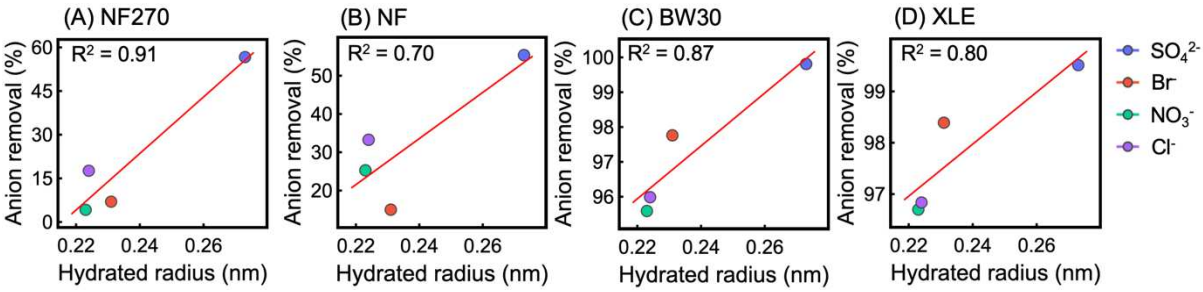


**Figure C2.** Anion rejections as a function of hydration radius after coagulation and microfiltration followed by treatment using (A) NF270 membrane, (B) NF membrane, (C) BW30 membrane and (D) XLE membrane.

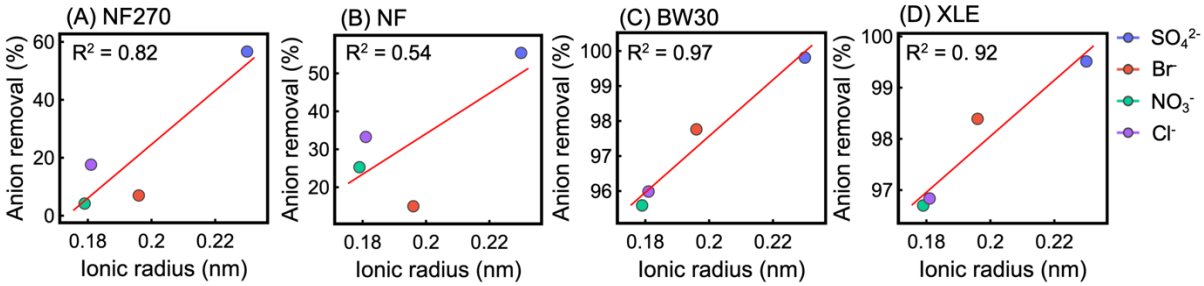


**Figure C3.** Anion rejections as a function of ionic radius after coagulation and microfiltration followed by treatment using (A) NF270 membrane, (B) NF membrane, (C) BW30 membrane and (D) XLE membrane.

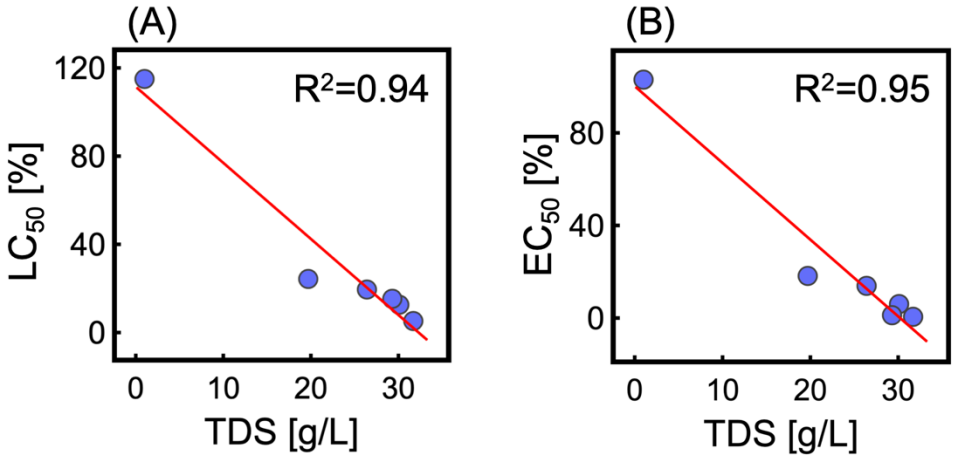


**Figure C4.** (A) The 48-h median lethal concentration ($LC_{50}$) and (B) the 48-h median effect concentration ($EC_{50}$) for *Daphnia* as a function of total dissolved solids (TDS).

**Table C1.** The concentrations of inorganic constituents (unit: mg/L unless specified) in the samples before and after pretreatments and membrane filtration. The standard deviations were calculated from three replicates.

| Parameter | Raw | Coagulation | Microfiltration | NF270 | NFD | BW30 | XLE |
|---|---|---|---|---|---|---|---|
| $Na^+$ | 10200.0±45.9 | 9580.0±100.6 | 9640.0±55.9 | 7520.0±18.0 | 6250.0±47.5 | 478.0±26.0 | 188.0±2.9 |
| $K^+$ | 47.6±0.3 | 55.2±2.4 | 58.7±0.5 | 47.6±0.1 | 36.4±0.2 | 2.4±0.1 | 2.1±0.0 |
| $Mg^{2+}$ | 99.4±0.3 | 97.5±1.5 | 102±0.4 | 38.6±0.4 | 19.1±0.1 | 0.0±0.0 | 0.0±0.0 |
| $Ca^{2+}$ | 760.0±6.7 | 763.0±5.3 | 715.0±7.4 | 432.0±4.1 | 313.0±3.8 | 6.9±0.0 | 6.9±0.0 |
| $Ba^{2+}$ | 45.7±0.9 | 9.2±0.1 | 7.8±0.2 | 3.7±0.0 | 2.2±0.1 | 0.0±0.0 | 0.0±0.0 |
| Fe | 118±2.3 | 3.7±0.1 | 0.7±0.0 | 0.0±0.0 | 0.0±0.0 | 0.0±0.0 | 0.0±0.0 |
| B | 50.1±2.3 | 34.5±0.5 | 28.8±1.3 | 22.1±0.4 | 17.8±0.7 | 7.4±0.1 | 7.9±0.3 |
| $Cl^-$ | 23903.1±23.9 | 23022.4±120.8 | 22161.5±85.1 | 18258.1±96.4 | 14791.3±116.1 | 891.1±0.6 | 701.2±1.8 |
| $Br^-$ | 83.6±2.2 | 78.0±0.4 | 77.2±0.3 | 71.8±3.4 | 65.6±1.6 | 1.7±0.0 | 1.2±0.0 |
| $NO_3^-$ | 289.1±2.8 | 271.1±4.7 | 263.6±5.4 | 252.5±1.7 | 197.0±6.1 | 11.5±0.5 | 8.7±0.0 |
| $SO_4^-$ | 61.4±0.6 | 68.9±4.5 | 67.9±6.7 | 29.2±0.4 | 30.1±0.3 | 0.1±0.0 | 0.3±0.0 |
| Total nitrogen (mg-N/L) | 31.9 | 32.3 | 32.4 | 26.8 | 22.9 | 4.0 | 5.3 |
| Electrical conductivity (mS/cm) | 41.8 | 47.0 | 45.8 | 41.3 | 30.7 | 1.6 | 1.4 |
| Total dissolved solid | 31350 | 35250 | 34350 | 30975 | 23025 | 1200 | 1050 |
| Sodium adsorption ratio (SAR, meq/L) | 92.3 | 86.7 | 89.2 | 92.9 | 92.6 | 50.0 | 19.7 |
| pH | 7.9 | 7.4 | 7.9 | 7.8 | 8.1 | 7.2 | 7.2 |

**Table C2.** The removal rates (%) of cations, anions, and boron in the produced water samples before and after pretreatments and membrane filtration. As Fe could be in the form of $Fe^{2+}$ or $Fe^{3+}$, no valence is indicated for Fe.

| Parameter | Coagulation | Microfiltration | NF270 | NFD | BW30 | XLE |
|---|---|---|---|---|---|---|
| $Na^+$ | 6.1 | -0.6 | 22.0 | 35.2 | 95.0 | 98.0 |
| $K^+$ | -16.0 | -6.3 | 18.9 | 38.0 | 95.9 | 96.4 |
| $Mg^{2+}$ | 1.9 | -4.6 | 62.2 | 81.3 | 100.0 | 100.0 |
| $Ca^{2+}$ | -0.4 | 6.3 | 39.6 | 56.2 | 99.0 | 99.0 |
| $Ba^{2+}$ | 79.9 | 15.2 | 52.6 | 71.8 | 100.0 | 100.0 |
| Fe | 96.9 | 81.1 | 100.0 | 100.0 | 100.0 | 100.0 |
| B | 31.1 | 16.5 | 23.3 | 38.2 | 74.3 | 72.6 |
| $Cl^-$ | 3.7 | 3.7 | 17.6 | 33.3 | 96.0 | 96.8 |
| $Br^-$ | 6.7 | 1.0 | 7.0 | 15.0 | 97.8 | 98.4 |
| $NO_3^-$ | 6.2 | 2.8 | 4.2 | 25.3 | 95.6 | 96.7 |
| $SO_4^{2-}$ | -12.2 | 1.5 | 57.0 | 55.7 | 99.9 | 99.6 |

**Table C3.** The concentrations and rejections of BTEX (benzene, toluene, ethylbenzene, and total xylenes), total petroleum hydrocarbons (TPH), non-purgeable organic carbons (NPOC), and polycyclic aromatic hydrocarbons (PAH) after coagulation (coag), microfiltration (MF), as well as coagulation and microfiltration followed by treatment using NF270, NF, BW30, or XLE membrane.

| Sample | Benzene [mg/L] | Toluene [mg/L] | Ethyl-benzene [mg/L] | Total Xylenes [mg/L] | Total Xylenes rejection [%] | TPH [mg/L] | TPH rejection [%] | NPOC [mg/L] | NPOC rejection [%] | Total PAH [µg/L] | Total PAH rejection [%] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Raw | < LOD | < LOD | < LOD | 0.240 | - | 869.0 | - | 75.5 | - | 114.7 | - |
| Coag | < LOD | < LOD | < LOD | 0.050 | 79.2 | 357.0 | 58.9 | 32.7 | 56.7 | 18.7 | 83.7 |
| MF | < LOD | < LOD | < LOD | 0.007 | 86.0 | 5.2 | 98.5 | 25.6 | 21.7 | 9.8 | 47.6 |
| NF270 | < LOD | < LOD | < LOD | < LOD | - | 0.9 | 82.7 | 19.7 | 23.0 | 6.1 | 37.6 |
| NFD | < LOD | < LOD | < LOD | < LOD | - | 1.2 | 76.9 | 17.3 | 32.4 | 5.0 | 48.9 |
| BW30 | < LOD | < LOD | < LOD | < LOD | - | <0.5 | 90.4 | 1.8 | 93.0 | 4.2 | 57.1 |
| XLE | < LOD | < LOD | < LOD | < LOD | - | 2.1 | 59.6 | 3.7 | 85.5 | 4.0 | 59.1 |

* < LOC: below limit of detection.