Capturing unmodelled phenomena: A hybrid approach for the prediction of the transport through ceramic membranes in organic solvent nanofiltration

Juan Pablo Gallo-Molina[a,*], Benjamin Claessens[b] , Anita Buekenhoudt[c], Arne Verliefde[d,e] ,and Ingmar Nopens[a,e]

[a] BIOMATH, Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium.
[b] Aix Marseille Univ, CNRS, MADIREL, UMR7246, F-13397, Marseille, France.
[c] Department of Separation and Conversion Technology, VITO (Flemish Institute of Technological Research), Mol, Belgium.
[d] PaInT, Particle and Interfacial Technology Group, Department of Green Chemistry and Technology, Ghent University, Ghent, Belgium.
[e] Centre for Advanced Process Technology for Urban Resource Recovery (CAPTURE), Ghent, Belgium.
[*] Corresponding author. Email address: JuanPablo.GalloMolina@UGent.be

Abstract

Organic Solvent Nanofiltration (OSN) is widely recognized as an interesting substitute to the thermal-based separation processes traditionally used in many industries. However, its industrial implementation has been hampered by the relatively poor understanding of the complex phenomena that drive its performance. Consequently, various model formulations have been proposed in the past with various degrees of success. Unfortunately, it seems that mechanistic methodologies tend to suffer from the mentioned lack of fundamental understanding, while data-driven approaches are typically bound to have very large data requirements and their predictions are often difficult to interpret and hard to extrapolate. Accordingly, this work presents a hybrid modelling methodology focused on OSN with ceramic membranes and aimed at the engineering practice. In this approach, a mechanistic component is married with a data-driven algorithm with the objective of obtaining accurate predictions that remain interpretable and bounded by the physics of the system. Two model architectures (parallel and serial) were formulated and subsequently tested with an experimental dataset. Both options were found to offer improved predictions of total flux and solute rejection in comparison with the widely used solution-diffusion framework. The models were found to produce accurate out-of-sample predictions and it was possible to obtain a picture of the influence of the effects driving model predictions. Additionally, adequate estimations of systems in which affinity interactions are dominant were obtained. This positions the hybrid methodology presented here as a promising alternative for the prediction of the performance of OSN.

Keywords: OSN, machine learning, hybrid modelling.

1. Introduction

The critical importance of separation processes for the chemical and pharmaceutical industries is a given, since these processes account for up to 15% of the world's energy consumption [1] and represent as much as 70% of said industries' capital and operating expenses [2]. As such, there is a strong economic and environmental argument in favour of challenging the current status quo. Organic Solvent Nanofiltration (OSN) is one emergent technology with the potential of accomplishing this feat in a wide range of applications. Since it is non-thermal, OSN promises energy-efficient separations with high selectivity, while remaining appropriate for temperature-sensitive compounds [1,3–5].

Unfortunately, OSN remains a relatively niche operation even though it has achieved large-scale technical feasibility in the past two decades [5,6]. There are multiple reasons for this but, among them, the comparatively poor fundamental understanding of the process is prominent. Unlike in aqueous solvent nanofiltration, OSN features an extended set of solvent/solute/membrane interactions, which limits its predictability [2,5,7–10]. As a result, expensive and time-consuming membrane screening experiments are required prior to every OSN industrial deployment [3,6].

One second issue limiting the reach of OSN is the fact that the typically used polymeric membranes have some significant disadvantages. In the first place, they swell when in contact with organic solvents, which can impact severely their performance in time. This swelling behaviour is also highly dependent on the solvent used, hampering even more both modelling and experimental exercises [11–14]. Secondly, aggressive solvents are frequently capable of degenerating this type of membrane which is, naturally, a highly undesirable property [14–17]. In contrast, ceramic membranes do not suffer from these problems and their performance can be fine-tuned via functionalization [18–21]. Consequently, the spotlight has been moving in the direction of said membrane class.

One challenge, however, remains the same: the performance of ceramic membranes is also dependent on various entangled phenomena, which greatly encumbers any process design and optimisation. This has led to an expanding corpus of literature devoted to unravelling the physicochemical mechanisms behind experimental observations [3,7,18,19,21,22]. Based on this, multiple mechanistic model structures have been proposed over the years with the eventual objective of obtaining a reliable way of predicting the behaviour of these systems [2,3,6,7,11,23–25]. This classical approach has the advantage of requiring relatively small amounts of data and, since it is anchored in first principles and/or phenomenological knowledge, predictions can be interpreted in a straightforward fashion and extrapolation is often possible. Conversely, one distinct disadvantage of these models is their large knowledge requirements which, as suggested before, cannot yet be fully satisfied in the case of OSN. This frequently leads to unsatisfactory results or extensive calibration demands [7,9,26].

Alternatively, purely data-driven approaches have been touted as a way of circumventing the mentioned gaps in fundamental knowledge, while providing predictive capabilities [5,9,27–31]. Unfortunately, data-driven models produce predictions which are difficult to link to physical phenomena, which limits interpretability. Moreover, the quality and quantity of the available data are of paramount importance, while extrapolation outside the original training space tends to be a risky proposition [26,27]

In view of the above, a middle ground can be envisioned: hybrid modelling. This method seeks to combine the mechanistic and data-driven paradigms in order to minimize their weaknesses and synergize their strengths. The technique has proven to be beneficial in the modelling of other membrane-based technologies and beyond [26,32–35]. In the field of OSN, Santos et al. [9] showed that a parallel hybrid structure can be useful to improve the predictability of solvent fluxes in polymeric membranes, but their approach retained the usual extensive calibration requirements associated with conventional mechanistic models and a relatively small range of solvents (in the absence of solutes) were considered.

Accordingly, we aim here to go one step further, while exploring less studied ceramic membranes. In this work, parallel and serial hybrid modelling structures were implemented with the aim of predicting both fluxes and rejections for different membrane/solvent/solute permutations. The focus is on the potential engineering application of the methodology rather than on the detailed study of the driving mechanisms. Both native and modified membranes were studied, and a wide range of industrially relevant solvents and solutes was considered. In order to capture affinity interactions, the thermodynamic non-idealities of the solvent/solute mixtures were included in the models, and several appropriate physicochemical properties of the systems at hand were used to inform their data-driven side. To the best of our knowledge, this approach has not been previously demonstrated for ceramic OSN applications and as such it can offer a promising path towards the robust prediction of these systems. This, in turn, should significantly improve the feasibility of model-based membrane selection, reducing experimental screening and leading to a more widespread industrial implementation of this technology.

2. Hybrid model development.

The general structure of the proposed models is portrayed in figure 1. The starting point is the data collected in a series of permeation experiments for different conditions and systems. Figure 1.a shows a simplified diagram of the setup used to obtain said data, while section 3 offers details regarding the experimental part of this work. As mentioned before, two hybrid model structures were proposed: parallel (figure 1.b) and serial (figure 1.c). In the former, a calibrated mechanistic component (solution-diffusion model, see section 2.1.) predicts the observables of interest (flux and/or rejection), while the data-driven component (XGBoost model, see section 2.2.) predicts the expected difference between reality and the values estimated by the mechanistic model (i.e., the residual error). Merging these two data streams results in the final prediction. In contrast, the data-driven component of the serial implementation is used to predict the values of the parameters of the mechanistic model, which is subsequently used to compute the expected fluxes or rejections. A more detailed assessment of these proposed architectures is made in section 4.

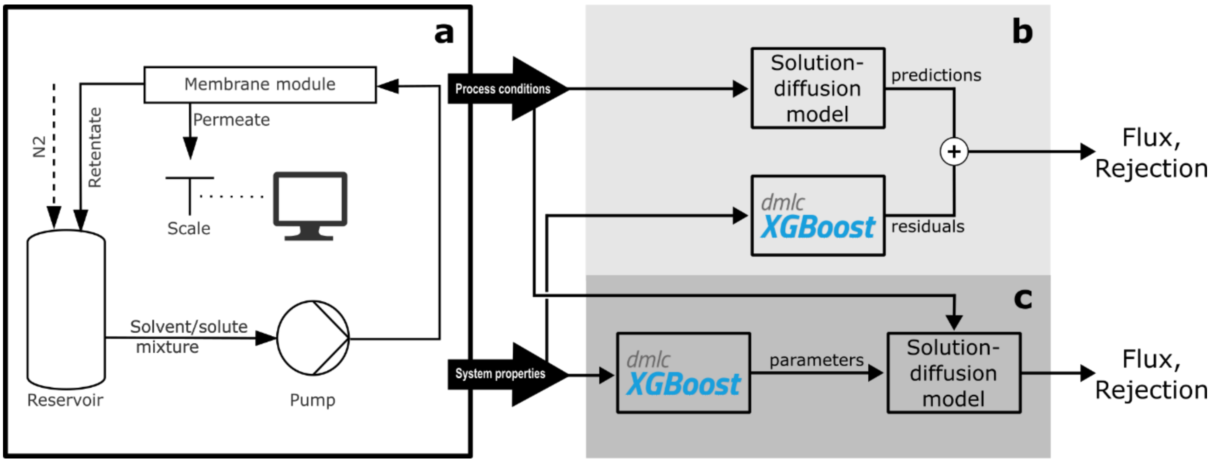

Figure 1. Structure of the implemented hybrid models. Permeation experiments (a) were performed under different conditions for different membrane/solvent/solute combinations. These conditions, as well as the properties of the studied systems were used to inform a parallel (b) and serial (c) hybrid model aimed at prediction fluxes and rejections.

2.1. Mechanistic component.

In a previous work, Claessens et al. [22] used the Maxwell-Stefan theory to analyse the transport in the same systems that concern this work. It was found that a transport model based on this theory can successfully capture the complexities typically observed in OSN experiments. Furthermore, it was observed that several factors, such as the thermodynamic non-ideality of the solvent/solute mixture and the partitioning coefficients play a large role in the determination of the resulting fluxes and rejections. This model is, however, challenging to apply in practice due to its reliance on the Hagen-Poiseuille's law and its intricate parametric structure, which severely limits calibration and identifiability.

Considering this, we chose the solution-diffusion model as the mechanistic description of the transport in the ceramic membranes studied here. This model emerges from the Maxwell-Stefan equations if cross-coupling effects are disregarded and diffusion coefficients are assumed to be constant [36]. In addition, this formulation does not consider swelling phenomena and provides steady-state estimations, which is adequate in the case of ceramic membranes. In the solution-diffusion model, transport is assumed to occur due to the diffusion of solvent and solutes through a homogenous active layer after an initial dissolution event (hence its name). This model has been found to describe OSN better than rival alternatives such as the pore flow model, which typically assumes viscous flow through well-defined pores [2,9,10,25,37]. Indeed, strong deviations from viscous flow have been observed in tight ceramic membranes [3,19], which suggests that the dominant mechanisms may be closer to the one described by the solution-diffusion model.

Equations 1-3 therefore constitute the selected mechanistic component. Considering their potentially substantial effect [22,25,38–40], the non-ideal thermodynamic interactions between solvent and solutes were captured by considering the activity coefficients on both sides of the membrane. The Dortmund-modified version of the UNIFAC group contribution method [41,42] was used to estimate these coefficients in all cases. All other dynamics and interactions are lumped in the permeability coefficients which are often treated as fitted parameters. In contrast, this work aims to complement the physics-based solution diffusion-model with a data-driven approach in order to better capture the phenomena that are not explicitly considered in said model.

$$J_{1,P} = P_1 \left[x_{1,F} - \frac{\gamma_{1,P}}{\gamma_{1,F}} \frac{J_{1,P}}{J_{1,P} + J_{2,P}} \exp\left(-\frac{v_1 p}{RT}\right) \right] \tag{1}$$

$$J_{2,P} = P_2 \left[x_{2,F} - \frac{\gamma_{2,P}}{\gamma_{2,F}} \frac{J_{2,P}}{J_{1,P} + J_{2,P}} \exp\left(-\frac{v_2 p}{RT}\right) \right] \tag{2}$$

$$rej = 1 - \frac{J_2}{J_{2,F}} \tag{3}$$

In the above expressions the subscript *1* refers to the solvent, while *2* points to the solute (note that a generalization towards more components is possible). In turn, the subscripts *F* and *P* refer to quantities on the feed and permeate sides, respectively. The solute rejection is denoted by *rej*. Moreover, *J* is the component flux, *x* is its mole fraction, and ν is its molar volume. The transmembrane pressure (including the osmotic pressure) is denoted by *p,* the operating temperature is *T,* and R signifies the universal gas constant. Finally, γ and *P* correspond to the activity coefficients and permeabilities, respectively. When applicable, the flux of the solute on the feed side ($J_{2,F}$) was deduced from its known concentration on said side.

Since equations 1 and 2 are coupled, the component fluxes were calculated numerically with the hybrid Powell method [43]. Additionally and when required, the model parameters were obtained after minimizing the sum of square errors with the Trust Region Reflective algorithm formulated by Branch et al. (1999). Due to the limited number of parameters and the assumed uncoupled diffusion, no significant identifiability problems were encountered.

2.2. Data-driven component.

The XGBoost (which stands for eXtreme Gradient Boosting) algorithm [45] was selected as the data-driven component of the hybrid models proposed here. This and similar tree-based techniques have been observed to outperform even state-of-the-art deep-learning algorithms when the available data set is tabular [45–48]. XGBoost is an ensemble learning algorithm that works by iteratively adding decision trees[1] (figure 2.a) to the model, with each tree trained to correct the errors made by the previous trees (figure 2.b). Mathematically, the model can be described by equation 4 [45]. This expression describes the estimation ($\hat{y}_i$) of a real response variable (y_i) with a function ϕ, which depends on a *m* number of features/descriptors (*x*) selected to predict *y* in the dataset at hand.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathcal{F} \tag{4}$$

The tree ensemble model described by equation 4 uses *K* independent decision trees (denoted by *f)*. In turn, the space of regression trees, $\mathcal{F}$, is given by equation 5. There, *q* corresponds to the structure of each tree, while *T* is the number of leaves. In turn, the leaf weights are denoted by *w.*

$$\mathcal{F} = \{f(x) = w_{q(x)}\}(q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^{\mathrm{T}}) \tag{5}$$

The algorithm is based on the gradient boosting framework and, as such, it implements the gradient descent algorithm to minimize the loss function by iteratively updating the model parameters to find the direction of steepest descent [49]. Figure 2b illustrates schematically the functioning of the algorithm.

---

[1] See, for example, the work of Quinlan [49] for additional information.

In order to prevent overfitting and improve generalization performance, XGBoost uses the regularized loss function given by equation 6 [45].

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{6}$$

Where:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{7}$$

In the equations above, *l* refers to a selected loss function that gives the differences between predictions and reality (squared loss, in this case). Moreover, the function Ω is a regularization term that penalizes complex trees and therefore forces the algorithm to prefer simpler yet predictive structures. This prevents overfitting but introduces the necessity to tune two additional hyperparameters, namely γ and λ.

Shrinkage [50] and feature subsampling are two additional strategies used by the implemented algorithm to avoid overfitting and improve generalization performance. The former reduces the weight of each new tree, decreasing the influence of individual trees and allowing new ones to better improve the overall model's predictive capability.

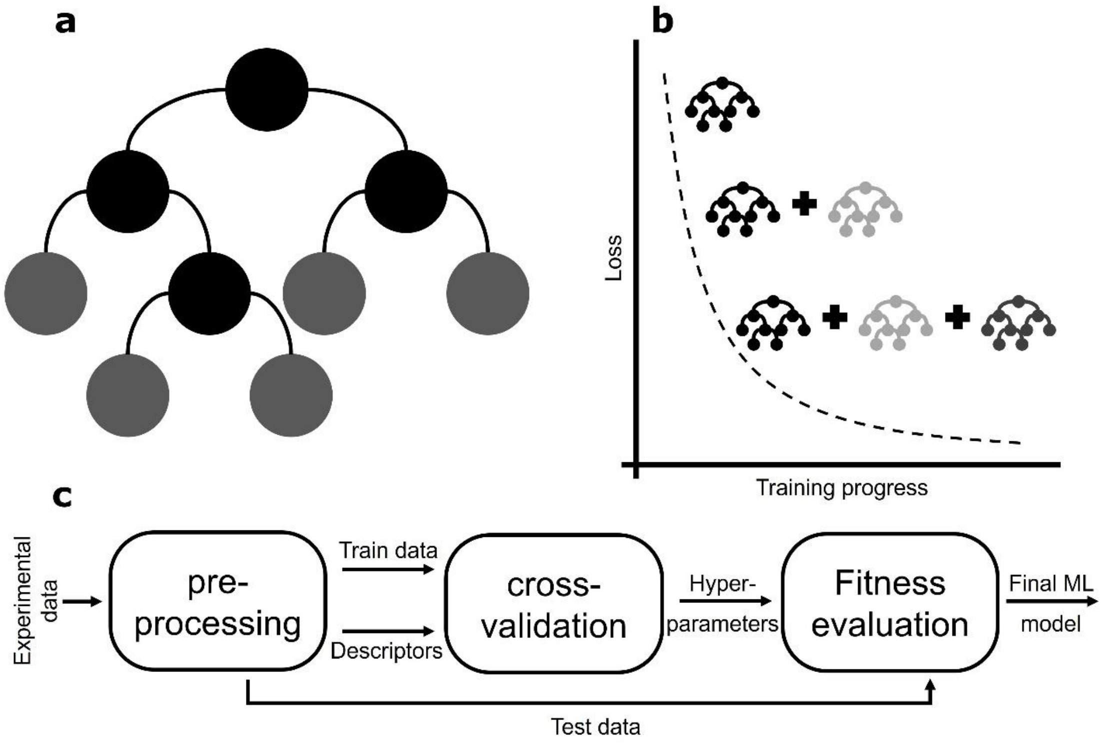

Figure 2. Schematic representation of the main concepts behind the chosen XGBoost algorithm. a) A decision tree consisting of various decision nodes (black) and leaf nodes (grey). At each decision node a condition based on a descriptor is tested till a final estimation is reached in a leaf node. b) The final model is an ensemble of decision trees. These structures are added iteratively in order to correct for the errors of previous trees. The gradient decent algorithm is used to minimize a loss function. c) The final model structure is obtained with the shown workflow.

The latter technique forces the model to choose different descriptors for each tree (or node within a tree), which allows the algorithm to better use the information carried by all features instead of overly relying on a limited number of important variables. Furthermore, trees are constructed in such a way that data sparsity can be properly considered, which further enhances generalizability. Additional details can be found in the work of Chen and Guestrin [45] and references within.

The workflow used during the development of the data-driven components of both hybrid models is schematically shown ion figure 2c. As illustrated in the figure, the experimental data is pre-processed to obtain train and test sets as well as the descriptors (see section 2.2.1.) selected to produce the desired predictions. Subsequently, the model is calibrated with cross-validation [51,52]. During this procedure, the internal structure of the model is determined and optimised, overfitting is reduced, and the model´s hyperparameters (e.g., γ and λ) are determined. Finally, the test data is used to obtain an evaluation of the model´s performance.

2.2.1. *Model features.*

A set of physicochemical properties was chosen to act as the descriptors used by the data-driven component of the models. Two criteria were taken into account when selecting these parameters. In the first place, it was sought that the selected features contain meaningful information regarding the affinity interactions among the components of the studied systems. Secondly, the selection was limited to variables that are either available in the literature or can be calculated with relative ease. This was done with the aim of maximizing the applicability of the proposed model. Table 1 summarizes the descriptors used in this work. As can be seen there, properties describing membranes, solvents, and solutes were identified. Note that some of the selected variables are more informationally rich than others and that correlations among some of them exist. While the XGBoost algorithm is capable of coping with this situation, additional feature selection was performed during model training. Further discussion is offered in section 4.

The characterization of the properties of the membranes was achieved primarily with their surface free energies (derived from contact angles) and Hansen solubility parameters. The latter are often used to quantify solute/solvent/membrane interaction effects, and they have been demonstrated to contain useful information in ceramic OSN systems [3]. In this work, the modified group contribution method developed by Andecochea Saiz et al. [53] was implemented to estimate the Hansen parameters of both native and functionalized membranes. The surface free energy of the membrane also conveys information regarding the interactions present on the membrane's surface and it is a useful measure of its hydrophobicity. Measured values reported in the literature were used in this case [15,54–58].

Additionally, the nominal pore sizes of the membranes were used to allow the model to discern the point in which viscous flow may start playing a larger role. Since it is known that the real pore size in modified membranes ought to be smaller due to the presence of grafted functional groups, a binary variable codifying whether functionalization was performed or not was included in the model. This allowed the algorithm to place different weights on the pore size information depending on the type of system to be predicted.

Table 1. Summary of features used by data-driven models.

Category	Feature	Symbol
Membrane	Functionalization [-]	f_M
	Hansen solubility parameters (dispersive, polar, H-bonds, total) [MPa$^{1/2}$]	$\delta_{d,M}$, $\delta_{p,M}$, $\delta_{h,M}$, $\delta_{t,M}$
	Nominal pore size [nm]	r_M
	Surface free energy [mN/m]	γ_M
Solvent	Density [kg/m^3]	ρ_1
	Dielectric constant [-]	ϵ_1
	Dipole moment [D]	μ_1
	Geometric radii (average, perpendicular] [nm]	$r_{a,1}$, $r_{p,1}$
	Hansen solubility parameters (dispersive, polar, H-bonds, total) [MPa$^{1/2}$]	$\delta_{d,1}$, $\delta_{p,1}$, $\delta_{h,1}$, $\delta_{t,1}$
	Molar mass [g/mol]	M_1
	Molar volume [cm^3/mol]	v_1
	Polarity [-]	p_1
	Shape indices (ellipsoid, geometric, petitjean) [-]	EI_1, GI_1, PI_1
	Surface tension [mN/m]	γ_1
	Viscosity [mPa s]	η_1
Solute	Dipole moment [D]	μ_2
	Geometric radii (average, perpendicular] [nm]	$r_{a,2}$, $r_{p,2}$
	Hansen solubility parameters (dispersive, polar, H-bonds, total) [MPa$^{1/2}$]	$\delta_{d,2}$, $\delta_{p,2}$, $\delta_{h,2}$, $\delta_{t,2}$
	Molar mass [g/mol]	M_2
	Molar volume [cm^3/mol]	v_2
	Feed molar concentration [mM]	c_2
	Shape indices (ellipsoid, geometric, petitjean) [-]	EI_2, GI_2, PI_2

Concerning solvent characterization, several properties were considered. In the first place, density and viscosity were added as general descriptors of the substance bulk characteristics. The latter being also important if viscous flow is significantly present. In the second place, we sought to capture the solvent solubility with its dielectric constant, dipole moment, surface tension and Hansen solubility parameters. These variables have appeared in the past in various phenomenological models applied to similar cases [5,7,11–13,23,59]. Thirdly, the normalized scale presented by Reichardt and Welton [60] was used as a relatively direct measure of the solvent's polarity. These descriptors were collected from the existing literature [60–63]. One additional factor that is known to play a role in the performance of organic solvent nanofiltration is the size of the solvent molecules [3,64–66]. This information was incorporated via the molar mass and volume.

Size does not tell the complete story though. It has been observed that the shape of the molecule can play a critical role [10,66–70]. Consequently, the classic geometric and Petitjean shape indices [71] were considered here. Furthermore and following on more recent advances in topological molecular characterization, the ratio between the largest and smallest axes of the minimum volume ellipsoid capable of enclosing a given molecule was used as well. This descriptor was found by Santos et al.[66] to be a well-rounded measure of molecular shape for membrane applications. However, unlike in the cited reference, the smallest enclosing ellipsoid was found here after solving the optimisation problem described by equation 8.

$$\begin{gathered}\min_{A,C} \log[\det(A^{-1})] \\ s.t.\ (x-C)^T A (x-C)\end{gathered} \tag{8}$$

In the above expression, the constraint corresponds to the matrix form equation describing a three-dimensional ellipsoid in an arbitrary cartesian coordinate system, whose volume is to be minimized while encompassing a group of points corresponding to the positions of the atoms composing the solvent molecule. The length of the axes was found after performing the singular value decomposition of the matrix A.

The estimated enclosing ellipsoids were subsequently used to calculate the effective molecular radii, which unlike the frequently used Stokes radius does not assume sphericity but rather incorporates shape information as well. Since no reliable information regarding preferential orientation is available, radii were calculated after assuming the average of possible orientations or a perpendicular orientation of the major axis with respect to the membrane surface. The calculation method described by Santos et al. [66] was followed in all cases. These parameters as well as the previously described shape indices were estimated from optimised molecular geometries, which were obtained with density functional theory computations performed in the ORCA quantum chemistry package [72].

As seen in table 1 a similar group of descriptors were chosen for the solutes. The rationale behind their selection is analogous to the one discussed for solvents in the previous paragraphs. However, strictly bulk properties were not considered as it would be inappropriate given that solute molecules are indeed solvated and thus far from a bulk-like state. Moreover, the molar concentrations in the feed were included to correct for the potential error introduced by the differences in actual concentration induced by the utilization of solutes with disparate molecular weights.

Finally, a series of derived features were included to induce the data-driven algorithm to explicitly evaluate some relationships that have been found useful in the modelling of OSN. These are the differences between: the surface energy of the membrane and the surface tension of the solvent; the total Hansen solubility parameters of the solvent and the membrane: the total Hansen solubility parameters of the solute and the membrane; and the total Hansen solubility parameters of the solute and the membrane. These total Hansen parameters were computed as follows:

$$\delta_t = \left(\delta_d^2 + \delta_p^2 + \delta_h^2\right)^{\frac{1}{2}} \tag{9}$$

The radius of the interaction sphere (*ra*) described by Hansen [62] to quantify the affinity between substances was considered as well. This variable was calculated for the corresponding interaction between solvent and membrane, solute and membrane, and solvent and solute. The computation was done using equation 10, in which X and Y correspond to the possible interacting components.

$$ra_{X,Y} = \left[4(\delta_{d,X} - \delta_{d,Y})^2 + \left(\delta_{p,X} - \delta_{p,Y}\right)^2 + (\delta_{h,X} - \delta_{h,Y})^2\right]^{\frac{1}{2}} \tag{10}$$

3. Materials and Methods.

We used a combination of literature permeation values and original, in-house measurements for model development. Permeation experiments were performed in a lab-scale crossflow setup with tubular ceramic membranes (area: 65.03 cm^2), under different transmembrane pressures (in a range of 2 – 20 bar), and at a room temperature of approximately 20 °C. Crossflow velocities of approximately 2 m/s were used to avoid concentration polarization effects. Both pure solvent and binary solvent/solute mixtures were studied. Fluxes were deduced from permeate mass measurements and the geometric specifications of the membranes. When a solute was included, its concentration was kept as 0.1 g/L in the feed size. Rejections were obtained after measuring the concentrations in the permeate side via UV/Vis spectroscopy. Steady state values were used in all cases. In order to expand the experimental space available to train and test the models, additional data was obtained from the literature (see table 2 and supplementary information). While the experimental methods used in the chosen publications are similar to the ones being described here, the reader is referred to the original references for additional details [3,7,15,19,73,74].

Table 2. Summary of the membranes used in this work.

Number of membranes	37
Active layer material	SiO_2 TiO_2. ZrO_2
Pore size [nm]	0.9 1.0 1.4 2.5 3.0 5.0
Functionalization	Native Methyl groups Pentyl groups Octyl groups Dodecyl groups Phenyl groups Silanated
Reference	This work [3] [7] [15] [19] [73] [74]

Commercially available TiO_2 ceramic membranes provided by Inopor GmbH (Germany) were used in their native (unmodified) state or they were functionalized prior to experimentation. Data for different top layer materials was obtained from the literature in order to expose the developed model to the possible variations in surface chemistry given by different ceramic materials. Functionalization was performed in order to include different levels of hydrophobicity in the active separation layer, and to mimic the fact that modifications are to be expected in an industrial setting with the aim of augmenting the affinity for any given target separation. When applicable, the membranes' surface chemistries were altered by grafting functional groups to their titania surfaces using Grignard chemistry [75] or by alternative methods as described in the relevant reference (see table 2). In most cases, the reported pore size of the selected membranes is firmly within the nanofiltration range [2], but larger pore sizes (up to 5 nm) were considered to expand the applicability of the proposed methods to the fuzzy limit between nano and ultrafiltration.

Industrially relevant solvents were chosen with the objective of obtaining a broad range of physicochemical properties. All solvents used were of analytical grade. In a similar manner, solutes were selected in such a way that their property space was wide, although experimental limitations precluded the inclusion of a larger sample. Table 3 show a selection of the properties of the solvents and solutes studied here. These were obtained from the literature or estimated with the methods described in the previous section.

Table 3. Solvent and solute properties

	Compound	**Formula**	**Molar mass [g/mol]**	**Geometric radius [nm]**[a]	**Total Hansen solubility parameter [MPa$^{1/2}$]**[b]	**Viscosity [mPa s]**[c]	**Polarity** [d]
Solvents	1-Propanol	C_3H_8O	60.10	0.226	24.6	1.72	61.7
	2-Propanol	C_3H_8O	60.10	0.199	23.6	2.00	54.6
	Acetone	C_3H_6O	58.08	0.191	19.9	0.33	35.5
	Acetonitrile	C_2H_3N	41.05	0.171	24.4	0.38	46.0
	Cyclohexane	C_6H_{12}	84.16	0.228	16.8	0.98	0.60
	Dichloromethane	CH_2Cl_2	84.93	0.154	20.2	0.44	30.9
	Dimethyl formamide	C_3H_7NO	73.09	0.217	24.9	0.82	40.4
	Dimethyl Sulfoxide	C_2H_6OS	78.13	0.203	26.7	2.00	44.4
	Ethanol	C_2H_6O	46.07	0.191	26.5	1.08	65.4
	Ethyl acetate	C_4H_8O2	88.11	0.268	18.2	0.46	23.0
	Heptane	C_7H_{16}	100.21	0.405	15.3	0.41	1.20
	Hexane	C_6H_{14}	88.18	0.331	14.9	0.31	0.90
	Methanol	CH_3OH	32.04	0.142	29.6	0.60	76.2
	Methyl cyclohexane	C_7H_{14}	98.19	0.284	16.0	0.68	2.50
	Methyl ethyl ketone	C_4H_8O	72.11	0.245	19.1	0.41	32.7
	Methyl isobutyl ketone	$C_6H_{12}O$	100.16	0.294	17.0	0.61	27.0
	N-Methyl-2-Pyrrolidone	C_5H_9NO	99.13	0.251	23.0	1.80	36.0
	Tetrahydrofuran	C_4H_8O	72.11	0.218	19.5	0.55	21.0
	Toluene	C_7H_8	92.14	0.289	18.2	0.59	9.90
	Water	H_2O	18.02	0.098	47.8	0.89	100
Solutes	9,10-Diphenylanthracene	$C_{26}H_{18}$	330.42	0.600	20.13	-	-
	BINAP	$C_{44}H_{32}P_2$	622.67	0.723	19.65	-	-
	Bromothymol Blue	$C_{27}H_{28}Br_2O_5S$	624.38	0.579	21.25	-	-
	$C_{64}H_{36}$ (cas. 175606-05-0)	$C_{64}H_{36}$	804.90	0.853	22.43	-	-
	Erythrosine-b	$C_{20}H_8I_4O_5$	835.89	0.525	26.7	-	-
	Fluorescein	$C_{20}H_{12}O_5$	332.30	0.536	24.38	-	-
	Fully methyl-capped PEG	$CH_3(OC_2H_4)_9C_4H_5O_2$	495.56	0.635	17.99	-	-
	Partially methyl-capped PEG	$CH_3(OC_2H_4)_{12}OH$	560.63	0.612	18.37	-	-
	Polyethylene glycol 200	$H(OC_2H_4)_4OH$	194.20	0.492	24.94	-	-
	Polyethylene glycol 600	$H(OC_2H_4)_{13}OH$	590.70	0.725	18.81	-	-
	Polyethylene glycol 1500	$H(OC_2H_4)_{34}OH$	1515.8	0.957	18.01	-	-
	Polystyrene 580	$(C_8H_8)_6$	614.80	0.805	20.12	-	-
	Polystyrene 1370	$(C_8H_8)_{13}$	1353.95	1.334	20.62	-	-
	Polystyrene 1530	$(C_8H_8)_{14}$	1431.80	1.485	20.62	-	-
	Tridodecylamine	$C_{36}H_{75}N$	522.00	1.050	15.97	-	-
	Trioctylamine	$C_{24}H_{51}N$	353.70	0.769	15.92	-	-

[a] Estimated from molecular geometries (see section 2.2.1.).
[b] Calculated from the Hansen parameters reported by Barton and Hansen [61,62].
[c] Values at 25 C reported by Smallwood [63].
[d] Normalized scale (water is 100) based on long-wavelength UV/VIS charge-transfer absorption measurements [60].

4. Results and Discussion.

4.1. Model assessment.

As mentioned before, two model architectures were explored (Figure 1). In the first, serial configuration, the Machine Learning (ML) component was trained to predict the solvent and solute permeabilities needed by the solution-diffusion model. Alternatively, in the parallel configuration a separate data-driven model was implemented to estimate the residual errors between the values obtained from a calibration of the mechanistic component and the empirically observed permeabilities. Although it is possible in principle to formulate a single XGBoost model for both solvent and solute transport, here we chose to implement separate models. While this implies the generation of potentially redundant structures and increases the required computational effort, it allows for a simpler training process, and it eases the interpretation of the obtained predictions.

The compiled experimental information was divided into train and tests sets. The former comprised 80% of the data and was used to calibrate the XGBoost models as well as to tune their hyperparameters (see figure 2c). These two tasks were performed in tandem with 5-fold cross-validation, which allows for a more efficient usage of the available data when compared to the classical train-validation-test split, while maintaining bias at acceptable levels [51,52]. Hyperparameters were searched by initially sampling a large candidate space and subsequently performing a thorough grid search in a more limited subset of promising permutations. The remaining test set comprising 20% of the data was only used at the end of the process to evaluate the predictive quality of the models.

Figure 3 shows the observed performance of the two assessed model architectures. To facilitate comparisons, the estimated total permeabilities (instead of the residual errors) are being plotted for the parallel model. Note that these permeabilities are the parameters of the solution diffusion model (the proportionality factor *P* present in equations 1 and 2) instead of the commonly used homonym quantity expressing a given flux per unit pressure. As seen in the figure, a satisfactory agreement between observations and predictions was obtained in all cases. More importantly, it can be argued that the out-of-sample predictive power of both models, as exemplified by the obtained error scores in the test set, is adequate (see the figure and table 4). Perhaps more interestingly, it was

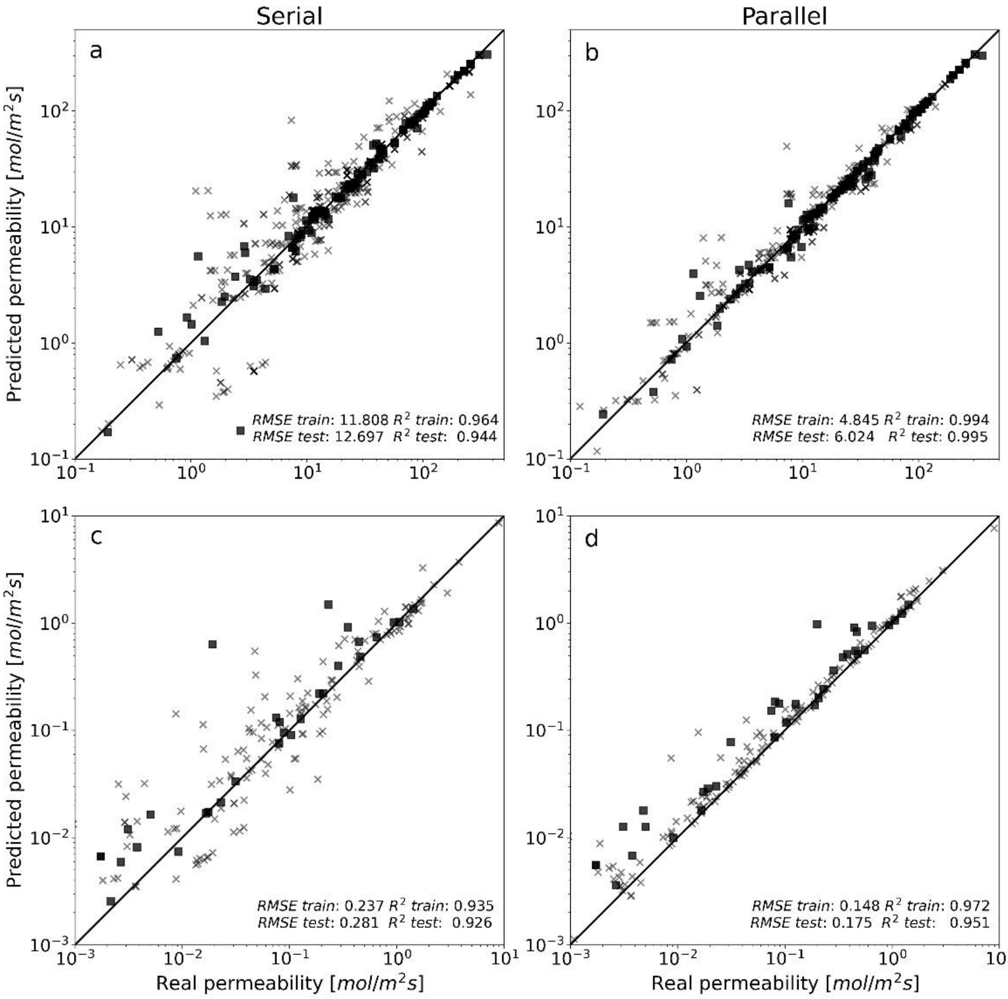

Figure 3. Predicted vs observed permeabilities for the two proposed hybrid model architectures. The top panels (a, b) show solvent permeabilities. The bottom panels (c, d) show solute permeabilities. The left and right panels illustrate the performance of the serial and parallel models, respectively. Cross symbols correspond to the train set, while square symbols denote the test set. The units of the RMSE are mol/m^{2}s.

observed that the parallel model (Figure 3b, d) offers an overall improvement in predictive quality when compared to its serial counterpart (Figure 3a, c). This can be attributed to the fact that in the parallel approach, the calibrated mechanistic model is able to collaborate with the data-driven component, which allows for a more comprehensive assimilation of the relevant affinity interactions driving both solvent and solute transport. Conversely, in the serial model the ML algorithm is tasked with learning all the lumped phenomena being described by the solution-diffusion model's permeabilities, which somewhat reduces its predictive ability if regularization is applied to prevent overfitting.

Table 4 presents a summary of the observed error (Root Mean Squared Error, RMSE; Mean Average Error, MAE; and coefficient of determination, R^2) of both models. Furthermore, the standard deviation of the RMSE scores obtained during cross validation and the size of the 95% confidence intervals of the test RMSE are presented as a proxy for the uncertainty of model performance. These intervals were estimated with the bootstrap method [76] by resampling the train-test split a large number of times and subsequently calculating the 2.5% and 97.5% percentiles of the distributions obtained for the test set errors. As suggested above, both models showed a good capability of predicting both solvent and solute transport, with the parallel approximation being generally better. It was observed that the solute transport models presented a reduced level of competency when the expected permeability was abnormally large (these points are not being displayed in figure 3 to improve its clarity but were included in all relevant computations). This corresponds to extreme cases in which the solute affinity for the membrane is comparable or higher to that of the solvent, which leads to very low or even negative rejections. Since the available data set contains a limited number of these situations and the uncertainty of the calculated activity coefficients is likely higher, the models lost some quantitative predictive power on this edge of the experimental space.

Table 4. Summary of model errors

Model	**Component**	**Set**	**RMSE [mol/m^{2}s]**	**SD / 95%CI – RMSE [mol/m^{2}s]^a**	**MAE [mol/m^{2}s]**	**R^2 [-]**
Serial	Solvents	Train	11.808	0.533	3.664	0.964
		Test	12.697	3.958	6.208	0.944
	Solutes	Train	0.237	0.013	0.100	0.935
		Test	0.281	0.081	0.109	0.926
Parallel	Solvents	Train	4.845	0.194	2.119	0.994
		Test	6.024	1.137	1.964	0.995
	Solutes	Train	0.148	0.008	0.048	0.972
		Test	0.175	0.067	0.081	0.951

[a] The standard deviation of the cross-validation scores is reported for the train set.
The size of the 95% confidence intervals is reported for the test set.

A further evaluation of the models can be made after inspecting figure 4. There, the predictions of total flux (panels a-c) and rejection (panels d-f) can be compared with their corresponding experimental values for all the available data points. The performance of the standalone solution-diffusion model as well as that of the two hybrid models are being displayed. As could be expected, the alluded mechanistic model seems to be capable of explaining a large part of the variance present in the data set, which lends support to the previously made assertion of diffusive transport being the main mechanism driving transport in the studied membranes. However, a significant improvement in the predictions was accomplished when permeabilities were estimated in the serial model, and it was again improved with the collaborative estimations made by the parallel paradigm. This strongly suggests that additional mechanisms (e.g., cross coupling, viscous flow, etc.) do play a role in the observed permeation in the analysed membranes, as has been theorized in previous studies [2,7].

Although the general performance of the analysed models is good, a somewhat reduced performance for intermediate values of rejection was observed. This suggests an increased complexity of the competing affinity interactions in comparison to those present in both extremes of the rejection spectrum, where size exclusion or more clear-cut differences in affinity tend to dominate. In any case, the predictions of the proposed models seem to be of comparable or even higher quality than those of, for example, the purely data-driven algorithms presented by Hu et al. (2021) for polymeric membranes (note, however, that the presence of swelling in polymeric membranes can make modelling more difficult). This suggests that the hybrid methodology presented in this work represents a valuable middle-ground in which satisfactory predictive power can be obtained without the large data requirements of a purely machine learning based solution. These reduced data requirements are associated with the fact that the model is physically grounded by the presence of its mechanistic component. The explicit calculation of solvent-solute interactions via the UNIFAC method is also believed to reduce the amount of information that must be sought in the experimental results. It must be noted, however, that making more data points available for training can be valuable for increasing the reliability of the models, particularly when rejection predictions are the main interest. Moreover, a certain amount of residual variability is expected to always remain for the reason that history effects may be present and variations between membrane batches (in, for example, their pore size distributions) produce differences in the final solvent and solute transport [3]. Additionally, some polymeric solutes consist of a distribution of differently sized chains centred in a reported average. Moreover, the influence of the structure of the functional groups added to the active layer during functionalization was not considered directly, which negates a source of valuable information. These effects could not be considered since experimental data with enough resolution is not currently available.

Although out of the scope of this work, the proposed models are expected to be applicable to tight polymeric membranes and, after appropriate training, their performance should be satisfactory. This is related with the fact that similar mechanisms are present in the aforementioned polymeric systems. However, an additional consideration must be given to the presence of swelling. This phenomenon often impacts the behaviour of polymer membranes, but it was not considered here since it is not present in ceramic OSN. Accordingly, additional mechanistic components and/or further features are likely necessary to fully account for said phenomenon.

After considering these possible pitfalls, both hybrid methodologies offer a promising path towards a fully predictive model of OSN. Moreover, it can be argued that each methodology could be leveraged at different stages of a hypothetical experimental screening or industrial deployment of OSN. Indeed, after training, the serial architecture should be able to offer good qualitative (and reasonable quantitative) predictions of the expected performance of a given system without the need for additional experiments. In turn, the parallel model could be used to obtain more reliable quantitative predictions after a relatively simple experimental calibration of the underling solution-diffusion model for a more limited number of candidate or interesting systems. These superior estimations could then be used to fine-tune systems of interest or to further inform a screening process.

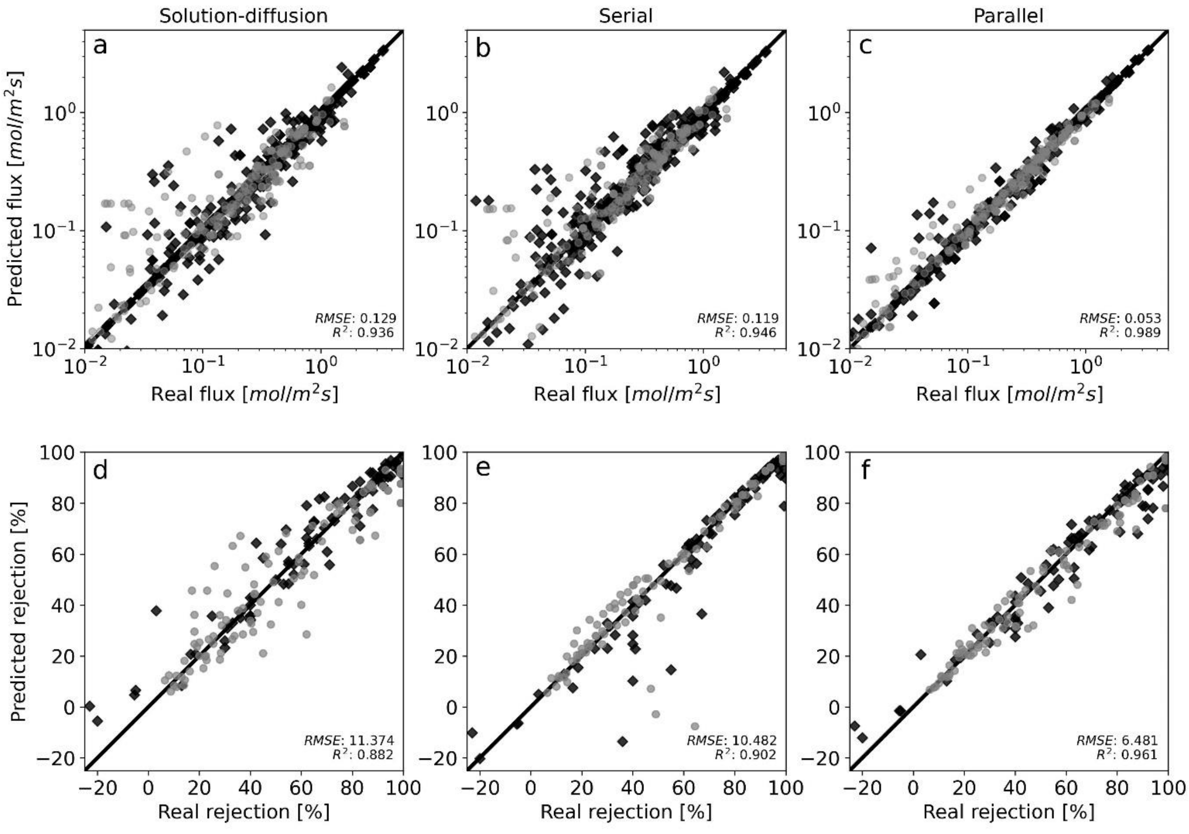

Figure 4. Comparison of flux (top, a-c) and rejection (bottom, d-f) predictions. The left panels correspond to the solution-diffusion model, the centre panels to the serial model, and the left panels to the parallel architecture. The units of the RMSE of the flux predictions are mol/m^{2}s. The RMSE of rejection predictions is expressed in percentage units. Both train and test sets are being displayed in all cases. Grey circles denote data points measured by the authors, while black squares indicate measurements obtained from the literature (see table 2).

4.2. Feature selection and importance.

In order to assess the relevance of the chosen features for the predictions being made by the proposed models, the intrinsic feature selection capability of the XGBoost algorithm in conjunction with its subsampling subroutines were utilized. In addition, the Recursive Feature Elimination (RFE) algorithm was used to explore which parameters were more valuable to reach the predictions generated by the hybrid models. For the sake of conciseness and considering that these relative contributions of the serial model should shed light on all the affinity interactions present in the studied systems (as opposed to only those not captured by the calibrated permeabilities of the solution-diffusion model), we restrict ourselves to the analysis of the behaviour of that model. Figure 5 therefore shows the normalized average contributions of the ten most important features used by the serial model.

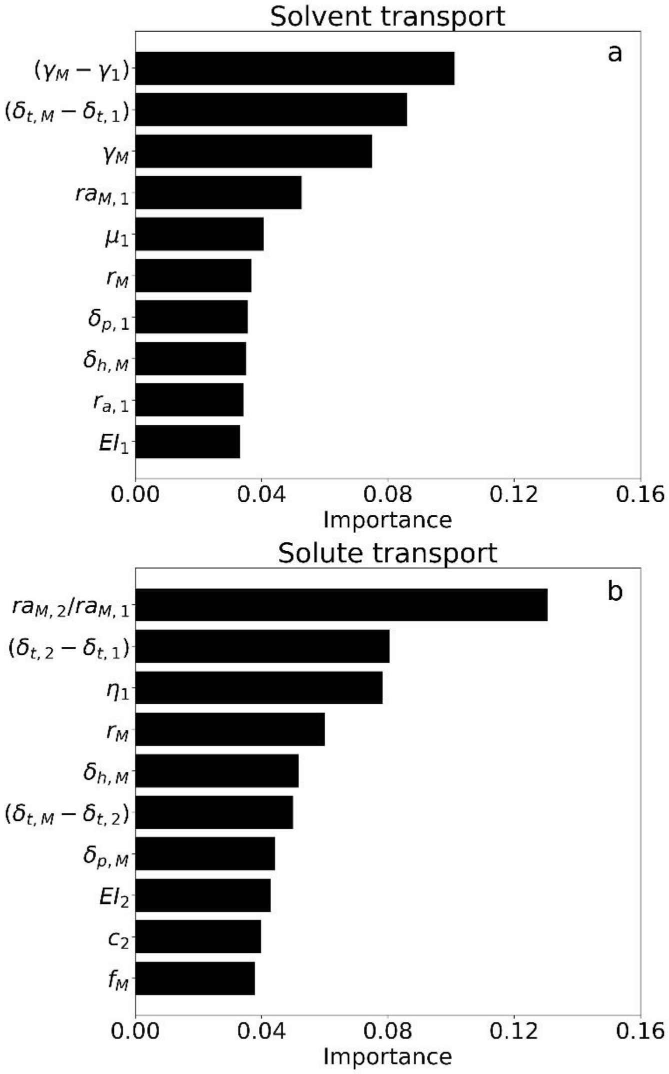

Figure 5. Summary of feature importance for solvent (a) and solute (b) transport. Only the ten most important descriptors for each model are being shown.

As could be expected, a one-to-one match between the main drivers of solvent (Figure 5a) and solute (Figure 5b) transport was not observed, which is consistent with previous results showing that different mechanisms affect the permeation of solvent and solutes [22]. More specifically, the model was found to make extensive use of the difference in surface tension between membrane and solvent. Interestingly, this term appears prominently in the model developed by Machado et al. [23], who included it to account for the energy change associated with the solvent transfer into the membrane surface. Similarly, the difference between Hansen solubility parameters also seems to play an important role. This descriptor has been found to be a well-rounded measure of the degree of the total interaction between membrane and solvent and consequently it has made repeated appearances in previous attempts to characterize solvent permeation in OSN [5,6,19,66]. Other parameters such as the dielectric constant and the hydrogen-bond and polar components of the solubility parameters were also predicted to be relevant for the prediction of the overall solvent permeabilities. This indicates that solvent polarity has a high apparent score for the predictive model, which seems consistent with the known physical relevance of this variable [12,70]. Moreover, the size and shape of the solvent molecules as well as the nominal membrane pore size were also extensively used by the model, which is again consistent with previous observations that ascribe a critical role to the degree of hindrance experimented by the molecules during transport through the membrane [10,66,77]. Finally, it must be noted that the solvent viscosity was not universally used by the algorithm to produce permeability predictions (hence its absence from the top panel of figure 5), but its importance tended to increase with the pore size. This suggests an increasing influence of viscous transport, which fits the current understanding that OSN is affected by both diffusive and viscous transport [6,7,77].

The picture becomes more complex when solutes are present. As evidenced by figure 5b, it is not only the affinity between solute and membrane or solute and solvent, but also the relative interactions between these three components (as approximated by the ratio of radii of the interaction spheres in the Hansen space) that determine how the model predicts solute transport through a membrane. This situation was qualitatively explored by Hosseinabadi et al. [18] based on observed fluxes, but the models proposed in this work are able to make effective use of a quantification of this phenomenon to obtain predictions. In addition, the surface tension, Hansen parameters and size related parameters appear again in the feature importance graph, which once again hints that overall permeation is strongly driven by solubility and degree of hindrance. Intriguingly, the viscosity of the solvent was also consistently leveraged by the models to estimate solute permeations. This may be an indication that a degree of solvent-solute coupled transport is present in most cases, as discussed by Schmidt and Lutze [6] and expected from the Maxwell-Stefan theory [22].

As shown above, the proposed models were found to offer satisfactory predictions by using the chosen descriptors. Their observed importance, furthermore, do not appear inconsistent with the known physical phenomena present in the explored systems. However, other features may contain useful information that could potentially improve model performance. Consequently, an expanded feature set may be worth exploring in a future work after an adequate assessment of the potentially higher data requirements.

4.3. Case studies.

In order to provide the reader with a more concrete evaluation of the capabilities of the developed models, in this section we present two case studies. In the first (section 4.3.1.), the extrapolation capabilities of both hybrid models are briefly explored, while the second (section 4.3.2.) is devoted to studying the potential prediction improvements with respect to the solution-diffusion model for systems where affinity interactions are of importance.

4.3.1. Nanofiltration of raffinose and linoleic acid.

As suggested before, one promising attribute of the hybrid methodology used in this work is its higher expected capacity of extrapolating outside the feature space used for training. To explore this, the models were challenged with the prediction of the nanofiltration of raffinose and linoleic acid under the experimental conditions reported by Tsuru at al. [78,79]. The former compound is a trisaccharide with a polarity significantly higher (δ_t = 46.51 MPa$^{1/2}$) than that of the solutes used during training. In turn, the latter is a fatty acid containing a polar carboxyl group and an apolar aliphatic chain. Both solutes then belong to compound groups not used during model training and, as could be expected, their chemical structure differs as well. In the case of the raffinose experiments, water was used as the solvent and a 1nm TiO_2 membrane was used. For linoleic acid, the solvent was n-hexane, and titania membranes with different pore sizes were utilized. Additional details are available elsewhere [78,79].

Figure 6a shows a comparison between the observed values of flux (left axis) and rejection (right axis) for different pressures. It was observed that both models were capable of producing good predictions of the former, with the parallel architecture offering a higher accuracy, which is in line with its observed performance with the test set (see section 4.1). In contrast, rejection predictions had a similar degree of accuracy for both models, which suggests that the proposed combination of the solution diffusion model with ML components can adequately capture the affinity interactions that lead to relatively small rejections with size-exclusion effects are not overwhelming.

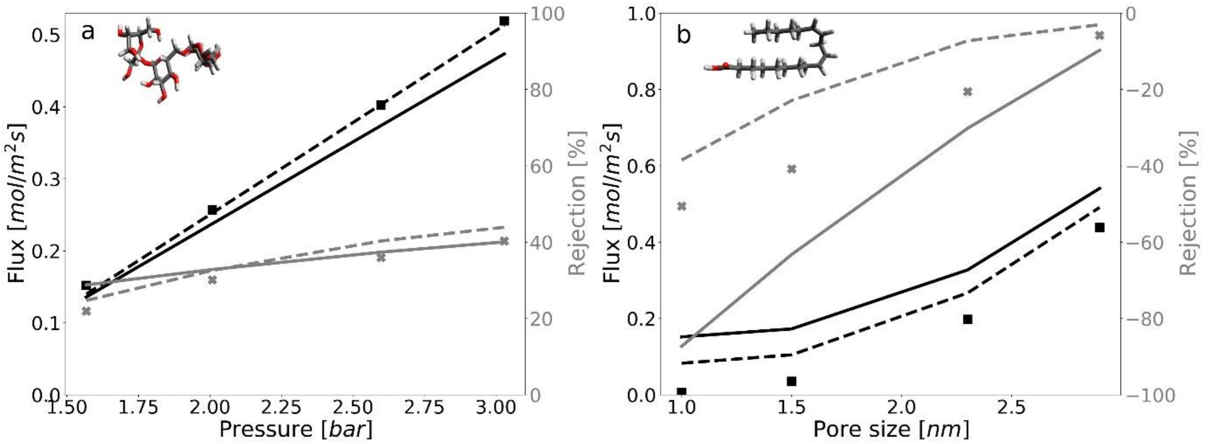

Figure 6. Fluxes and rejections for the nanofiltration of raffinose and linoleic acid. a) Nanofiltration of raffinose in water under different pressures [80]. b) Nanofiltration of linoleic acid in hexane for different pore sizes [79]. Black denotes fluxes (left axes), and grey corresponds to rejections (right axes). Square symbols illustrate flux measurements, while crosses are the corresponding rejections. Solid and dashed lines are the predictions made by the serial and parallel models, respectively.

The picture is less flattening when one looks at the predictions made for systems containing linoleic acid in hexane (figure 6b). While flux estimations were somewhat close to the experimental observations, both models overestimated its values, which was likely caused by the relative scarcity of information available to the models: the only changing variable being the pore size. Rejection predictions were less accurate, but it is noteworthy that both models predicted negative rejections for all the studied sizes, which indicates that the models succeeded in estimating the very high affinity of the solute for the membrane (this is likely caused by the interaction between carboxyl groups and the active layer [78]). Nevertheless, more work is required. Training the models with more data points for systems exhibiting this kind of behaviour and including additional features will probably improve predictions. Alternatively, a better methodology for estimating activity coefficients could be explored.

4.3.2. Rejection of fluorescein and diphenylanthracene in a native 1 nm TiO2 membrane.

One interesting feature of OSN is its potential ability of providing affinity-based separations. Consequently here we analyse the modelled permeation of fluorescein and 9,10-diphenylanthracene (DPA) in dimethyl formamide through a native ceramic membrane. Both solutes have very similar molar masses (332.3 g/mol for fluorescein and 330.4 g/mol for DPA), which translates in similar molecular sizes. On the other hand, DPA is an apolar solute and fluorescein is polar. Membrane-solute interactions could thus be expected to play a role in the rejection mechanism. Indeed, as seen in figure 7a, fluorescein was observed to be substantially more rejected than DPA. This type of situation, where said interactions are important, has traditionally encumbered modelling efforts. This is evidenced by the relatively poor performance of the solution-diffusion model (see the dashed lines in the top panel of the figure), which seems to not always be completely capable of reproducing observed OSN data. In contrast, it was found that the hybrid methodology discussed here can offer a substantial improvement when predicting (the data points were part of the test set) solute permeation in these complex systems (see the solid lines in figure 7a).

Although more work is required, this can open the door to meaningful comparative predictions of OSN systems. For instance, the membrane selectivity (defined as the ratio between the permeabilities of solute and solvent) can be obtained directly from the model. This allows for a fast estimation of the expected level of preferential permeation (considering all relevant effects) offered by any given membrane, which can be very valuable during a preliminary screening. In the case at hand, the selectivity of the DPA and fluorescein systems were calculated as 0.051 and 0.037, respectively. Although it is likely that the actual values differ, these numbers tell an interesting story when put together: the affinity of the analysed TiO_2 membrane is higher for DPA, leading to smaller rejections. This, again, has the potential of offering useful insights to practitioners and it is even possible to envisage the implementation of optimisation schemes aimed at maximizing, for example, the permeation of one species over another.

Figure 7b shows in turn the decision paths followed by the model to arrive to the estimated values of solute permeabilities, which were subsequently used to calculate the predicted rejection profiles. The panel can be interpreted as follows: starting from a base level, different features (in the y axis) are assigned positive or negative effects that, when added, allow the model to reach its final estimation. The magnitudes and directions of the paths thus indicate the overall influence of a given parameter. Even though it is not necessarily possible to ascribe a physical meaning to these relative movements, it is nonetheless noteworthy that the model appears to be able to make effective use of the information that is actually relevant. As showed by the plot, the model initially maintained similar estimates for both solutes given that certain important factors such as their molecular radii and the properties of the membrane were close or identical. However, the differences in the parameters linked to affinity specifically (i.e., the radii of the interaction spheres and the Hansen solubilities) and the slight variations in molecule shape prompted the model to generate a divergence, which eventually led to markedly different permeabilities.

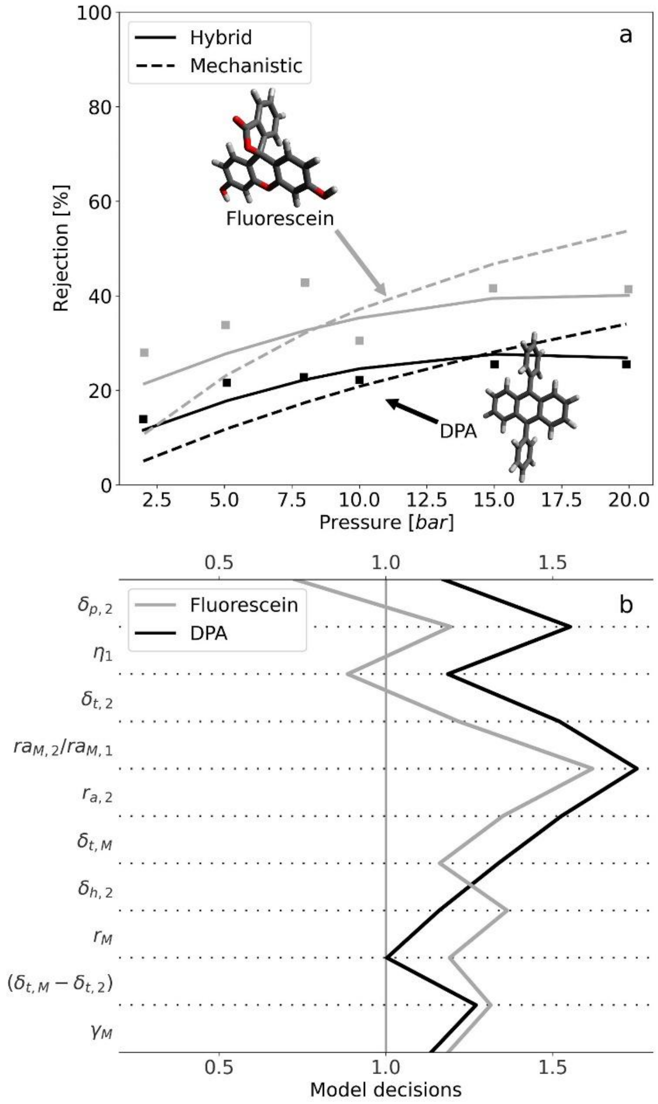

Figure 7. Modelled permeation of two solutes (fluorescein and 9,10-diphenylanthracene, DPA) in dimethyl formamide through a 1 nm TiO_2 native membrane. The top panel (a) shows the measured experimental data (squares) and a comparison between the standalone solution-diffusion model (dashed lines) and the hybrid model (solid lines). The bottom panel (b) shows the relative decision paths made by the model to reach its permeability predictions.

5. Conclusions

In this work a novel hybrid modelling methodology was explored as a way of obtaining predictions of the performance of ceramic organic solvent nanofiltration. This strategy combined a physics-based modelling component in the form of the solution-diffusion model and a data-driven module (XGBoost). In addition, two architectures were explored: 1) a serial, in which the data-driven algorithm is used to predict the parameters of the mechanistic model; and 2) a parallel in which both components produce complimentary predictions. An ample number of experimental points were used to calibrate and evaluate the models. They were observed to be capable of producing satisfactory estimates of the expected fluxes and solute rejections.

The proposed models performed better than the well-known solution-diffusion model, which suggests that they are able to extract relevant information regarding unmodelled dynamics from the experimental data used during training. However, the usage of said mechanistic model at the heart of the studied methodology allows for a reduction in the overall data requirements and produced predictions consistent with the known physical reality. As a result, it can be argued that a hybrid modelling approach is a good middle ground between exclusively phenomenological models and purely machine learning approaches. This is expected to be of benefit for industrial practitioners and it has the potential of clearing the way to a more widespread implementation of OSN. Nevertheless, it must be noted that the optimal choice of model (mechanistic, hybrid or data-driven) is likely dependent on the application, and the associated data availability; specialist knowledge present; and the needs for extrapolation and/or interpretation.

The proposed serial architecture is particularly well suited to discern the affinity interactions driving observed behaviours. Consequently, the importance of the physical descriptors used to characterize the studied systems was evaluated. As expected, different phenomena appear to drive solvent and solute transport, but it was generally found that a combination of surface characteristics, molecular properties and polarity effects play a large role in the obtained model predictions. More specifically, the surface tensions of membrane and solvent, the Hansen solubility parameters, the shape of the involved molecules, and the relative affinities between solute and membrane and solvent and membrane showed consistently a high importance.

Additional work can lead to further improvements in the proposed methodology. For example, an extension towards multicomponent mixtures can be envisioned, which should increase even more the industrial relevance of the tool. This, for instance, could be achieved by formulating single models that do not distinguish between solute and solvent. Additionally, different mechanistic components could be integrated to extend its applicability to looser membranes. Moreover, the model can be continuously updated as additional experimental data is made available, which can improve its capacity to tackle complex interactions.

The developed models have the potential to provide valuable tools for the design of membrane separation operations. Moreover, after training with a larger data set, flux and rejection predictions with a good accuracy could be obtained from the proposed serial architecture without the need of additional experiments. This should prove valuable when a large number of candidate systems must be assessed. Alternatively, a higher degree of accuracy can be obtained if the discussed parallel configuration is implemented. This comes with the price of the relatively limited number of experiments needed to calibrate the mechanistic part of the model. Both approaches, however, are expected to bring closer the coveted appearance of a fully predictive generic description of organic solvent nanofiltration.

6. Funding

The authors gratefully acknowledge the support of VLAIO under the EASICHEM cSBO project (HBC.2018.0484).

7. Nomenclature

J	Molar flux [mol/m^2 s]
P	Solution-diffusion model permeability [mol/m^{2}s]
p	Transmembrane pressure [Pa]
R	Gas constant [J/mol K]
rej	Rejection [-]
T	Temperature [K]
v	Molar volume [m^3/mol]
x	Mole fraction [-]
γ	Activity coefficient [-]
$\mathbb{R}$	Real numbers [-]

Subscripts

1	Solvent
2	Solute
F	Feed side
M	Membrane
P	Permeate side

8. References

[1] R.P. Lively, D.S. Sholl, From water to organics in membrane separations, Nat. Mater. 16 (2017) 276–279. https://doi.org/10.1038/nmat4860.

[2] P. Marchetti, M.F. Jimenez Solomon, G. Szekely, A.G. Livingston, Molecular separation with organic solvent nanofiltration: A critical review, Chem. Rev. 114 (2014) 10735–10806. https://doi.org/10.1021/cr500006j.

[3] A. Buekenhoudt, F. Bisignano, G. De Luca, P. Vandezande, M. Wouters, K. Verhulst, Unravelling the solvent flux behaviour of ceramic nanofiltration and ultrafiltration membranes, J. Membr. Sci. 439 (2013) 36–47. https://doi.org/10.1016/j.memsci.2013.03.032.

[4] J. Geens, B. De Witte, B. Van der Bruggen, Removal of API's (Active Pharmaceutical Ingredients) from Organic Solvents by Nanofiltration, Sep. Sci. Technol. 42 (2007) 2435–2449. https://doi.org/10.1080/01496390701477063.

[5] R. Goebel, T. Glaser, M. Skiborowski, Machine-based learning of predictive models in organic solvent nanofiltration: Solute rejection in pure and mixed solvents, Sep. Purif. Technol. 248 (2020). https://doi.org/10.1016/j.seppur.2020.117046.

[6] P. Schmidt, P. Lutze, Characterisation of organic solvent nanofiltration membranes in multi-component mixtures: Phenomena-based modelling and membrane modelling maps, J. Membr. Sci. 445 (2013) 183–199. https://doi.org/10.1016/j.memsci.2013.05.062.

[7] P. Marchetti, A. Butté, A.G. Livingston, An improved phenomenological model for prediction of solvent permeation through ceramic NF and UF membranes, J. Membr. Sci. 415–416 (2012) 444–458. https://doi.org/10.1016/j.memsci.2012.05.030.

[8] P. Marchetti, A. Butté, A.G. Livingston, NF in organic solvent/water mixtures: Role of preferential solvation, J. Membr. Sci. 444 (2013) 101–115. https://doi.org/10.1016/j.memsci.2013.04.069.

[9] J.L.C. Santos, A.M. Hidalgo, R. Oliveira, S. Velizarov, J.G. Crespo, Analysis of solvent flux through nanofiltration membranes by mechanistic, chemometric and hybrid modelling, J. Membr. Sci. 300 (2007) 191–204. https://doi.org/10.1016/j.memsci.2007.05.024.

[10] L.S. White, Transport properties of a polyimide solvent resistant nanofiltration membrane, 2002.

[11] S. Darvishmanesh, A. Buekenhoudt, J. Degrève, B. Van der Bruggen, General model for prediction of solvent permeation through organic and inorganic solvent resistant nanofiltration membranes, J. Membr. Sci. 334 (2009) 43–49. https://doi.org/10.1016/j.memsci.2009.02.013.

[12] S. Darvishmanesh, J. Vanneste, E. Tocci, J.C. Jansen, F. Tasselli, J. Degrève, E. Drioli, B. Van der Bruggen, Physicochemical Characterization of Solute Retention in Solvent Resistant Nanofiltration: the Effect of Solute Size, Polarity, Dipole Moment, and Solubility Parameter, J. Phys. Chem. B. 115 (2011) 14507–14517. https://doi.org/10.1021/jp207569m.

[13] J. Geens, B. Van Der Bruggen, C. Vandecasteele, Transport model for solvent permeation through nanofiltration membranes, Sep. Purif. Technol. 48 (2006) 255–263. https://doi.org/10.1016/j.seppur.2005.07.032.

[14] E.S. Tarleton, J.P. Robinson, M. Salman, Solvent-induced swelling of membranes — Measurements and influence in nanofiltration, J. Membr. Sci. 280 (2006) 442–451. https://doi.org/10.1016/j.memsci.2006.01.050.

[15] A. Dobrak, B. Verrecht, H. Van den Dungen, A. Buekenhoudt, I.F.J. Vankelecom, B. Van der Bruggen, Solvent flux behavior and rejection characteristics of hydrophilic and hydrophobic mesoporous and

microporous TiO2 and ZrO2 membranes, J. Membr. Sci. 346 (2010) 344–352. https://doi.org/10.1016/j.memsci.2009.09.059.

[16] J. Geens, B. Van der Bruggen, C. Vandecasteele, Characterisation of the solvent stability of polymeric nanofiltration membranes by measurement of contact angles and swelling, Chem. Eng. Sci. 59 (2004) 1161–1164. https://doi.org/10.1016/j.ces.2004.01.003.

[17] Y.H. See Toh, F.W. Lim, A.G. Livingston, Polymeric membranes for nanofiltration in polar aprotic solvents, J. Membr. Sci. 301 (2007) 3–10. https://doi.org/10.1016/j.memsci.2007.06.034.

[18] S. Hosseinabadi, K. Wyns, V. Meynen, R. Carleer, P. Adriaensens, A. Buekenhoudt, B. Van der Bruggen, Organic solvent nanofiltration with Grignard functionalised ceramic nanofiltration membranes, J. Membr. Sci. 454 (2014) 496–504. https://doi.org/10.1016/j.memsci.2013.12.032.

[19] S.R. Hosseinabadi, K. Wyns, V. Meynen, A. Buekenhoudt, B. Van der Bruggen, Solvent-membrane-solute interactions in organic solvent nanofiltration (OSN) for Grignard functionalised ceramic membranes: Explanation via Spiegler-Kedem theory, J. Membr. Sci. 513 (2016) 177–185. https://doi.org/10.1016/j.memsci.2016.04.044.

[20] J. Kujawa, S. Cerneaux, W. Kujawski, M. Bryjak, J. Kujawski, How To Functionalize Ceramics by Perfluoroalkylsilanes for Membrane Separation Process? Properties and Application of Hydrophobized Ceramic Membranes, ACS Appl. Mater. Interfaces. 8 (2016) 7564–7577. https://doi.org/10.1021/acsami.6b00140.

[21] R.B. Merlet, M.-A. Pizzoccaro-Zilamy, A. Nijmeijer, L. Winnubst, Hybrid ceramic membranes for organic solvent nanofiltration: State-of-the-art and challenges, J. Membr. Sci. 599 (2020) 117839. https://doi.org/10.1016/j.memsci.2020.117839.

[22] B. Claessens, I. Hitsov, A. Verliefde, I. Nopens, Analyzing transport in ceramic membranes for organic solvent nanofiltration using Maxwell-Stefan theory, Chem. Eng. Sci. 264 (2022). https://doi.org/10.1016/j.ces.2022.118133.

[23] D.R. Machado, D. Hasson, R. Semiat, Effect of solvent properties on permeate flow through nanofiltration membranes, J. Membr. Sci. 166 (2000) 63–69. https://doi.org/10.1016/S0376-7388(99)00251-3.

[24] P. Marchetti, A.G. Livingston, Predictive membrane transport models for Organic Solvent Nanofiltration: How complex do we need to be?, J. Membr. Sci. 476 (2015) 530–553. https://doi.org/10.1016/j.memsci.2014.10.030.

[25] L.G. Peeva, E. Gibbins, S.S. Luthra, L.S. White, R.P. Stateva, A.G. Livingston, Effect of concentration polarisation and osmotic pressure on flux in organic solvent nanofiltration, J. Membr. Sci. 236 (2004) 121–136. https://doi.org/10.1016/j.memsci.2004.03.004.

[26] M.Y. Schneider, W. Quaghebeur, S. Borzooei, A. Froemelt, F. Li, R. Saagi, M.J. Wade, J.-J. Zhu, E. Torfs, Hybrid modelling of water resource recovery facilities: status and opportunities, Water Sci. Technol. 85 (2022) 2503–2524. https://doi.org/10.2166/wst.2022.115.

[27] C.F. Galinha, J.G. Crespo, From black box to machine learning: A journey through membrane process modelling, Membranes. 11 (2021). https://doi.org/10.3390/membranes11080574.

[28] J. Hu, C. Kim, P. Halasz, J.F. Kim, J. Kim, G. Szekely, Artificial intelligence for performance prediction of organic solvent nanofiltration membranes, J. Membr. Sci. 619 (2021) 118513. https://doi.org/10.1016/j.memsci.2020.118513.

[29] G. Ignacz, N. Alqadhi, G. Szekely, Explainable machine learning for unraveling solvent effects in polyimide organic solvent nanofiltration membranes, Adv. Membr. 3 (2023) 100061. https://doi.org/10.1016/j.advmem.2023.100061.

[30] G. Ignacz, G. Szekely, Deep learning meets quantitative structure–activity relationship (QSAR) for leveraging structure-based prediction of solute rejection in organic solvent nanofiltration, J. Membr. Sci. 646 (2022) 120268. https://doi.org/10.1016/j.memsci.2022.120268.

[31] C. Wang, L. Wang, A. Soo, N. Bansidhar Pathak, H. Kyong Shon, Machine learning based prediction and optimization of thin film nanocomposite membranes for organic solvent nanofiltration, Sep. Purif. Technol. 304 (2023) 122328. https://doi.org/10.1016/j.seppur.2022.122328.

[32] C.F. Galinha, G. Guglielmi, G. Carvalho, C.A.M. Portugal, J.G. Crespo, M.A.M. Reis, Development of a hybrid model strategy for monitoring membrane bioreactors, J. Biotechnol. 164 (2013) 386–395. https://doi.org/10.1016/j.jbiotec.2012.06.026.

[33] L. Luo, S. Bao, Knowledge-data-integrated sparse modeling for batch process monitoring, Chem. Eng. Sci. 189 (2018) 221–232. https://doi.org/10.1016/j.ces.2018.05.055.

[34] A.R. Ricardo, R. Oliveira, S. Velizarov, M.A.M. Reis, J.G. Crespo, Hybrid modeling of counterion mass transfer in a membrane-supported biofilm reactor, Biochem. Eng. J. 62 (2012) 22–33. https://doi.org/10.1016/j.bej.2011.12.010.

[35] M. von Stosch, R. Oliveira, J. Peres, S. Feyo de Azevedo, Hybrid semi-parametric modeling in process systems engineering: Past, present and future, Comput. Chem. Eng. 60 (2014) 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008.

[36] D.R. Paul, Reformulation of the solution-diffusion theory of reverse osmosis, J. Membr. Sci. 241 (2004) 371–386. https://doi.org/10.1016/j.memsci.2004.05.026.

[37] S.J. Han, S.S. Luthra, L. Peeva, X.J. Yang, A.G. Livingston, Insights into the Transport of Toluene and Phenol Through Organic Solvent Nanofiltration Membranes, Sep. Sci. Technol. 38 (2003) 1899–1923. https://doi.org/10.1081/SS-120020126.

[38] K.P. Bye, M. Galizia, Fundamental origin of flux non-linearity in organic solvent nanofiltration: Formulation of a thermodynamic/diffusion framework, J. Membr. Sci. 603 (2020) 118020. https://doi.org/10.1016/j.memsci.2020.118020.

[39] B.D. Marshall, J.W. Allen, R.P. Lively, A model for the separation of complex liquid mixtures with glassy polymer membranes: A thermodynamic perspective, J. Membr. Sci. 647 (2022) 120316. https://doi.org/10.1016/j.memsci.2022.120316.

[40] R. Mathias, D.J. Weber, K.A. Thompson, B.D. Marshall, M.G. Finn, J.K. Scott, R.P. Lively, Framework for predicting the fractionation of complex liquid feeds via polymer membranes, J. Membr. Sci. 640 (2021) 119767. https://doi.org/10.1016/j.memsci.2021.119767.

[41] D. Constantinescu, J. Gmehling, Further Development of Modified UNIFAC (Dortmund): Revision and Extension 6, J. Chem. Eng. Data. 61 (2016) 2738–2748. https://doi.org/10.1021/acs.jced.6b00136.

[42] J.W. Kang, V. Diky, M. Frenkel, New modified UNIFAC parameters using critically evaluated phase equilibrium data, Fluid Phase Equilibria. 388 (2015) 128–141. https://doi.org/10.1016/j.fluid.2014.12.042.

[43] M.J.D. Powell, A New Algorithm for Unconstrained Optimization, in: Nonlinear Program., Academic Press, 1970: pp. 31–65. https://doi.org/10.1016/B978-0-12-597050-1.50006-3.

[44] M.A. Branch, T.F. Coleman, Y. Li, A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems, SIAM J. Sci. Comput. 21 (1999) 1–23. https://doi.org/10.1137/S1064827595289108.

[45] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, New York, NY, USA, 2016: pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[46] C.J.C. Burges, From RankNet to LambdaRank to LambdaMART: An Overview, Microsoft, 2010. https://www.researchgate.net/publication/228936665.

[47] P. Li, Robust Logitboost and adaptive base class (ABC) Logitboost, in: Proc. Twenty-Sixth Conf. Annu. Conf. Uncertain. Artif. Intell., 2010.

[48] J. Van Der Donckt, J. Van Der Donckt, E. Deprost, N. Vandenbussche, M. Rademaker, G. Vandewiele, S. Van Hoecke, Do not sleep on traditional machine learning, Biomed. Signal Process. Control. 81 (2023) 104429. https://doi.org/10.1016/j.bspc.2022.104429.

[49] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat. 29 (2001) 1189–1232.

[50] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (2002) 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

[51] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, 1st ed. 2013, Corr. 7th printing 2017 edition, Springer, New York, 2013.

[52] M. Stone, Cross-Validatory Choice and Assessment of Statistical Predictions, J. R. Stat. Soc. Ser. B Methodol. 36 (1974) 111–147.

[53] C. Andecochea Saiz, S. Darvishmanesh, A. Buekenhoudt, B. Van der Bruggen, Shortcut applications of the Hansen Solubility Parameter for Organic Solvent Nanofiltration, J. Membr. Sci. 546 (2018) 120–127. https://doi.org/10.1016/j.memsci.2017.10.016.

[54] T. Bodner, A. Behrendt, E. Prax, F. Wiesbrock, Correlation of surface roughness and surface energy of silicon-based materials with their priming reactivity, Monatshefte Für Chem. - Chem. Mon. 143 (2012) 717–722. https://doi.org/10.1007/s00706-012-0730-8.

[55] E. Conte, A. Martinho, H.A. Matos, R. Gani, Combined Group-Contribution and Atom Connectivity Index-Based Methods for Estimation of Surface Tension and Viscosity, Ind. Eng. Chem. Res. 47 (2008) 7940–7954. https://doi.org/10.1021/ie071572w.

[56] M.L. González-Martín, L. Labajos-Broncano, B. Jańczuk, J.M. Bruque, Wettability and surface free energy of zirconia ceramics and their constituents, J. Mater. Sci. 34 (1999) 5923–5926. https://doi.org/10.1023/A:1004767914895.

[57] P. Król, B. Król, Determination of free surface energy values for ceramic materials and polyurethane surface-modifying aqueous emulsions, J. Eur. Ceram. Soc. 26 (2006) 2241–2248. https://doi.org/10.1016/j.jeurceramsoc.2005.04.011.

[58] C. Sun, J.C. Berg, The effective surface energy of heterogeneous solids measured by inverse gas chromatography at infinite dilution, J. Colloid Interface Sci. 260 (2003) 443–448. https://doi.org/10.1016/S0021-9797(02)00238-2.

[59] D. Bhanushali, S. Kloos, C. Kurth, D. Bhattacharyya, Performance of solvent-resistant membranes for non-aqueous systems: solvent permeation results and modeling, 2001.

[60] C. Reichardt, T. Welton, Solvents and Solvent Effects in Organic Chemistry, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2010. https://doi.org/10.1002/9783527632220.

[61] A.F.M. Barton, CRC Handbook of Solubility Parameters and Other Cohesion Parameters, Routledge, 2017. https://doi.org/10.1201/9781315140575.

[62] C.M. Hansen, Hansen Solubility Parameters, CRC Press, 2007. https://doi.org/10.1201/9781420006834.

[63] I. Smallwood, Handbook of Organic Solvent Properties, Elsevier, 1996. https://doi.org/10.1016/C2009-0-23646-4.

[64] P. Berg, G. Hagmeyer, R. Gimbel, Removal of pesticides and other micropollutants by nanofiltration, Desalination. 113 (1997) 205–208. https://doi.org/10.1016/S0011-9164(97)00130-6.

[65] J.C. Giddings, E. Kucera, C.P. Russell, M.N. Myers, Statistical theory for the equilibrium distribution of rigid molecules in inert porous networks. Exclusion chromatography, J. Phys. Chem. 72 (1968) 4397–4408. https://doi.org/10.1021/j100859a008.

[66] J.L.C. Santos, P. de Beukelaar, I.F.J. Vankelecom, S. Velizarov, J.G. Crespo, Effect of solute geometry and orientation on the rejection of uncharged compounds by nanofiltration, Sep. Purif. Technol. 50 (2006) 122–131. https://doi.org/10.1016/j.seppur.2005.11.015.

[67] K.O. Agenson, J.-I. Oh, T. Urase, Retention of a wide variety of organic pollutants by different nanofiltration/reverse osmosis membranes: controlling parameters of process, J. Membr. Sci. 225 (2003) 91–103. https://doi.org/10.1016/j.memsci.2003.08.006.

[68] Y. Kiso, K. Muroshige, T. Oguchi, T. Yamada, M. Hhirose, T. Ohara, T. Shintani, Effect of molecular shape on rejection of uncharged organic compounds by nanofiltration membranes and on calculated pore radii, J. Membr. Sci. 358 (2010) 101–113. https://doi.org/10.1016/j.memsci.2010.04.034.

[69] P. Silva, S. Han, A.G. Livingston, Solvent transport in organic solvent nanofiltration membranes, J. Membr. Sci. 262 (2005) 49–59. https://doi.org/10.1016/j.memsci.2005.03.052.

[70] B. Van der Bruggen, J. Schaep, D. Wilms, C. Vandecasteele, Influence of molecular size, polarity and charge on the retention of organic molecules by nanofiltration, J. Membr. Sci. 156 (1999) 29–41. https://doi.org/10.1016/S0376-7388(98)00326-3.

[71] M. Petitjean, Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds, J. Chem. Inf. Comput. Sci. 32 (1992) 331–337. https://doi.org/10.1021/ci00008a012.

[72] F. Neese, F. Wennmohs, U. Becker, C. Riplinger, The ORCA quantum chemistry program package, J. Chem. Phys. 152 (2020) 224108. https://doi.org/10.1063/5.0004608.

[73] S.R. Hosseinabadi, K. Wyns, A. Buekenhoudt, B. Van der Bruggen, D. Ormerod, Performance of Grignard functionalized ceramic nanofiltration membranes, Sep. Purif. Technol. 147 (2015) 320–328. https://doi.org/10.1016/j.seppur.2015.03.047.

[74] K. Lechner, G. Brösigke, J.-U. Repke, Reproducible Measurement Results in Organic Solvent Nanofiltration with Ceramic Membranes, Chem. Ing. Tech. 91 (2019) 55–59. https://doi.org/10.1002/cite.201800062.

[75] P. Van Heetvelde, E. Beyers, K. Wyns, P. Adriaensens, B.U.W. Maes, S. Mullens, A. Buekenhoudt, V. Meynen, A new method to graft titania using Grignard reagents, Chem. Commun. 49 (2013) 6998. https://doi.org/10.1039/c3cc43695k.

[76] A.C. Davison, D.V. Hinkley, Bootstrap Methods and their Application, Cambridge University Press, Cambridge, 1997.

[77] L. Wang, J. He, M. Heiranian, H. Fan, L. Song, Y. Li, M. Elimelech, Water transport in reverse osmosis membranes is governed by pore flow, not a solution-diffusion mechanism, Sci. Adv. 9 (2023) eadf8488. https://doi.org/10.1126/sciadv.adf8488.

[78] T. Tsuru, M. Narita, R. Shinagawa, T. Yoshioka, Nanoporous titania membranes for permeation and filtration of organic solutions, Desalination. 233 (2008) 1–9. https://doi.org/10.1016/j.desal.2007.09.021.

[79] T. Tsuru, K. Ogawa, M. Kanezashi, T. Yoshioka, Permeation Characteristics of Electrolytes and Neutral Solutes through Titania Nanofiltration Membranes at High Temperatures, Langmuir. 26 (2010) 10897–10905. https://doi.org/10.1021/la100791j.