separations

MDPI

Review

Organic Solvent Nanofiltration and Data-Driven Approaches

Pieter-Jan Piccard [1,2,3], Pedro Borges [3], Bart Cleuren [1], Jef Hooyberghs [1,2,*] and Anita Buekenhoudt [3]

[1] Theory Lab., Faculty of Sciences, UHasselt—Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium
[2] Data Science Institute, Faculty of Sciences, UHasselt—Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium
[3] Unit Separation and Conversion Technology, VITO N.V.—Flemish Institute of Technological Research, Boeretang 200, 2400 Mol, Belgium
* Correspondence: jef.hooyberghs@uhasselt.be

Abstract: Organic solvent nanofiltration (OSN) is a membrane separation method that has gained much interest due to its promising ability to offer an energy-lean alternative for traditional thermal separation methods. Industrial acceptance, however, is held back by the slow process of membrane screening based on trial and error for each solute-solvent couple to be separated. Such time-consuming screening is necessary due to the absence of predictive models, caused by a lack of fundamental understanding of the complex separation mechanism complicated by the wide variety of solute and solvent properties, and the importance of all mutual solute-solvent-membrane affinities and competing interactions. Recently, data-driven approaches have gained a lot of attention due to their unprecedented predictive power, significantly outperforming traditional mechanistic models. In this review, we give an overview of both mechanistic models and the recent advances in data-driven modeling. In addition to other reviews, we want to emphasize the coherence of all mechanistic models and discuss their relevance in an increasingly data-driven field. We reflect on the use of data in the field of OSN and its compliance with the FAIR principles, and we give an overview of the state of the art of data-driven models in OSN. The review can serve as inspiration for any further modeling activities, both mechanistic and data-driven, in the field.

Keywords: organic solvent nanofiltration; data science; mathematical modeling; machine learning; data standardization

Citation: Piccard, P.-J.; Borges, P.; Cleuren, B.; Hooyberghs, J.; Buekenhoudt, A. Organic Solvent Nanofiltration and Data-Driven Approaches. *Separations* **2023**, *10*, 516. https://doi.org/10.3390/separations10090516

Academic Editor: Victoria Samanidou

Received: 30 August 2023
Revised: 15 September 2023
Accepted: 16 September 2023
Published: 19 September 2023

1. Introduction

Membranes are powerful, versatile separation tools, offering an energy-lean alternative for traditional thermal separation methods like the ubiquitous distillations and evaporations [1,2]. Their strength has been proven in water-based streams with numerous large-scale implementations [3]. In the last 15 years, membranes have also shown great potential in organic solvents, giving rise to a new field named organic solvent nanofiltration (OSN), very relevant in more sustainable chemistry [4]. Next to seriously declined energy use and correlated low CO_2 footprint, the strengths of OSN are gentle processing, avoidance of additives, and easy scalability due to modular buildup.

Unlike membrane processes in aqueous solutions, in organic solvents, the underlying transport mechanism is not well understood. Experimental OSN results have clearly shown the complexity of this membrane process, influenced by the wide variety of solute and solvent properties, and the importance of all mutual solute–solvent–membrane affinities and competing interactions. On top of this, solvent-dependent swelling further complicates separation for polymeric membranes, even for the membranes specifically developed for OSN. As a result, and despite its advantages, OSN is still not widely used in industry. An important bottleneck is this absence of a detailed fundamental understanding of the OSN process and the lack of an efficient predictive model [5,6]. To use OSN without such a model, membranes need to be screened based on trial and error for each solute–solvent

couple to be separated. Consequently, the development process for OSN is tedious and time-consuming, and needs to be repeated for each new separation case, slowing down general industrial acceptance.

There do exist mechanistic models that can be used to try and describe OSN, though originally developed for water filtration. However, due to the complex nature of OSN, these models have shown only limited predictive capacity and are not general for all membranes, conditions or maybe even solvents because a lot of interactions are not well described in the models. Nevertheless, these mechanistic models are still used to, for example, investigate the relative importance of different transport mechanisms (as diffusion or convection) and of non-idealities [7].

Nowadays, data-driven modeling is receiving more and more interest in process technology, and especially in complex processes. The combination of this general rise in interest in data-driven approaches and the urgent need for predictive models in OSN has resulted in the recent popularity of data-driven models in this field. These models are used to predict membrane performance and are already outperforming the mechanistic models that came before [8,9]. Using the predictive power that data-driven models provide, they can help fast track the industrial acceptance of OSN. Next to prediction, data-driven models can also be used to determine the key solute–solvent–membrane properties influencing the separation process [6,10]. Moreover, these models are being used in combination with mechanistic models as so-called hybrid models to provide extra insights and to reveal possible relationships between the physical parameters of a mechanistic model and specific solute–solvent–membrane properties [11,12]. Hence, data-driven approaches can still provide physical insight into the separation mechanism.

This review will first give an overview of all mechanistic models relevant for OSN, focusing on the governing parameters and assumptions of the physical context, as well as the relations between the different models. Compared to other reviews, we want to emphasize the coherence of all mechanistic models and we provide an overview of the free parameters in all models. The remaining part of the review will discuss data-driven modeling in OSN. We start with a discussion of the data itself and the importance of complying with the FAIR principles [13]. We reflect on how well the OSN community has been doing in this regard and discuss how and where to improve. Additionally, we discuss the representation of the data, namely the descriptors. Thereafter, we zoom in on data-driven approaches in general and their application to membranes in water filtration before we give an in-depth discussion on the state of the art of data-driven modeling in OSN and its results up to now. Finally, we conclude with an outlook on the future and the challenges the field will face on the way.

2. Mechanistic Transport Models

In the current scientific literature, a whole set of models can be found that describe transport through membranes based on physical principles. These models provide analytical expressions for the flux of solvent(s) and solute(s) through the membrane (or equivalently the rejection rate), along with a set of physical parameters. The choice of a particular model is driven by the properties of the separation mechanism, the type of membrane and, certainly in OSN [4], the solute and solvent. Different models can be distinguished by the assumptions made on the physical context, the approximations they apply, and the set of physical parameters they are represented by, which can differ both in meaning and amount.

Mechanistic membrane transport models are traditionally categorized in three major branches: solution–diffusion (SD), pore flow (PF), and irreversible thermodynamics (IT) models [14]. SD and PF models start from specific assumptions related to the physicochemical properties of the membrane. SD models describe separation as a process driven by the relative diffusion of the solute and solvent through the membrane. PF models describe the transport of solute and solvent as viscous flow through membrane pores, where the separation happens due to size exclusion. Hence, the choice for either SD or

PF typically depends on the membrane: SD models are often used to describe dense membranes (through which the permeating species can diffuse), and PF models are often used to describe porous membranes (through which the permeating species can flow). The third type of models, IT models, are, in that sense, less specific. They treat a membrane without any prior assumptions and the relative transport of solute and solvent through the membrane follows merely from energy dissipation and entropy production.

Compared to other reviews, we want to emphasize the coherence of all mechanistic models, and we provide an overview of the free parameters in all models. The different models were historically developed independently, but the resulting equations for solute and solvent flux are often mathematically very similar. The reason for this is that they can all be derived from the same general Maxwell–Stefan (MS) theory. The master model itself follows from fundamental statistical physics [15]. The set of models has a hierarchical structure, where a chain of assumptions leads to a specific model: the assumption of the dominating separation mechanism, the physical parameters and their interpretation, and the membrane-dependent assumptions on the pressure and concentration profiles. This shows that most models are not fundamental in nature, but at the same time, it underlines the coherence of the whole: although many models were developed with ad hoc concepts, by making the underlying assumption explicit, the differential equations can be explained. This is the unique, particular approach followed in this review and visualized in Figure 1. This figure displays the amount of parameters used in each model for a binary system (1 solute, 1 solvent) as well. Counting the parameters gives an idea of how much information is contained within a model, as well as what information is lost when going from a more general model to one with fewer parameters.

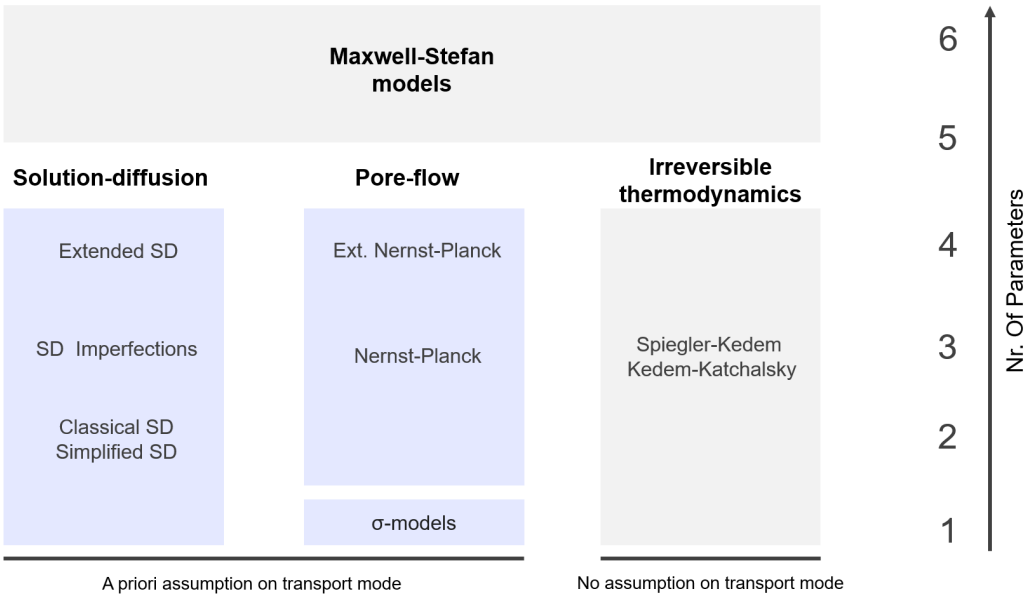

Figure 1. Overview of mechanistic models relevant for OSN. The parameter count is valid for binary systems only.

This section is organized as follows. First, the general MS theory is discussed in detail. Next each major branch is addressed together with the assumptions made to derive it from the MS model. For each branch, some representative models will be discussed together with the parameters needed to represent them. Specific attention is paid to the physical meaning of the parameters. Table 1 summarizes the symbols used in the different models.

Table 1. Nomenclature.

c [mol/m^3]	Molar concentration	D [m^2/s]	Fick's Diffusion coefficient
$\mathcal{D}$ [m^2/s]	MS diffusion coefficient	F [N/mol]	Force per mole
$\mathcal{F}$ [9.649 $\times$ 10^4 C/mol]	Faraday constant	J [mol/(s m^2)]	Molar flux
J_V [m/s]	Volumetric flux	K [−]	Sorption coefficient
K_c [−]	Conductive hindrance factor	K_d [−]	Diffusive hindrance factor
L	Solvent permeability parameter	N [−]	Number of species
p [Pa]	Pressure	P	Solute permeability parameter
R [8.314 J/(K mol)]	Gas constant	R_i [−]	Rejection of species i
T [K]	Temperature	u [m/s]	Diffusive velocity
x [−]	Mole fraction	z [m]	Spatial coordinate perpendicular to membrane surface
Z [−]	Charge number	α [−]	Viscous selectivity
γ [−]	Activity coefficient	ϵ [−]	Membrane porosity
ζ [kg/(s mol)]	MS Friction coefficient	η [Pa s]	Viscosity
λ [−]	Ratio of solute to pore radius	μ [J/mol]	Chemical potential
ν [m^3/mol]	Molar volume	π [Pa]	Osmotic pressure
σ [−]	Reflection Coefficient	τ [−]	Membrane tortuosity
χ [−]	Friction coefficient	ψ [V]	Electric potential
	Subscripts		
1	Solvent	2	Solute
i, j	Either solute or solvent	m	Membrane
	Superscripts		
$'$	Feedside	$''$	Permeate side
(0)	External side of membrane boundary	(m)	Membrane side of boundary

2.1. Maxwell–Stefan Theory

The Maxwell–Stefan equation is a model that describes the diffusion and, by expansion, viscous flow of a multi-component system [16–18]. The model generalizes Fick's law for diffusion, as it is possible to derive it from the MS model. Maxwell–Stefan theory relies on the assumption that any relative diffusive movement of species is caused by a deviation from equilibrium between molecular friction and the thermodynamic driving force. In the steady state, the governing equation is then found by balancing the thermodynamic driving force on a species i to the friction of i with all other species j [19]:

$$\begin{bmatrix} \text{driving force} \\ \text{on a species } i \end{bmatrix} = \begin{bmatrix} \text{sum of friction with} \\ \text{all other species } j \end{bmatrix}.$$

The friction force of species j on i is proportional to the difference in *diffusive* velocities $(u_i - u_j)$ of the species, with a proportionality constant $\zeta_{i,j}$ called the friction coefficient. The only further assumptions that the model makes are thermal equilibrium ($T = cst$) and that the friction force of species j on i is proportional to the mole fraction x_j of species j in the mixture. The driving force F_i contains contributions from gradients in the chemical potential μ_i (as suggested by irreversible thermodynamics [20]) and external forces F_i^{ext} as electric potentials. The gradient $\nabla\mu_i$ further contains contributions from the gradients of pressure p and mole fractions x_i (or equivalently concentrations $c_i = x_i c_{tot}$).

This then leads to the Maxwell–Stefan equation for species i in one dimension [19,21]:

$$-\frac{RT}{x_i\gamma_i}\frac{d(x_i\gamma_i)}{dz} - \nu_i\frac{dp}{dz} + F_i^{ext} = \sum_j x_j\zeta_{i,j}(u_i - u_j) + \zeta_{i,m}u_i \tag{1}$$

where z is the spatial coordinate perpendicular to the membrane surface, ν_i is the molar volume, and γ_i is the activity coefficient of species i, which accounts for the non-ideality of the mix. The membrane, denoted by subscript m, is considered a separate species with no absolute or diffusive velocity. Accordingly, it was taken out of the summation, which now

only sums over solutes and solvents. Note that there is a Maxwell–Stefan equation for all components of the mix, and there is no distinction between the solvent or solute.

It is useful to rewrite the MS equations in terms of molar fluxes J_i instead of the diffusive velocities since the flux is the actual observable. Neglecting any electrical forces, this gives

$$J_i = -\underbrace{\frac{RT\epsilon c_{tot}}{\zeta_{i,m}}\frac{dx_i}{dz}}_{\text{Diffusion}} - \underbrace{\frac{RT\epsilon c_{tot}x_i}{\zeta_{i,m}}\frac{d\ln\gamma_i}{dz}}_{\text{Non-ideality}} - \underbrace{\frac{x_i\epsilon c_{tot}v_i}{\zeta_{i,m}}\frac{dp}{dz}}_{\text{Pressure diffusion}} - \underbrace{\sum_j \frac{\zeta_{i,j}}{\zeta_{i,m}}(x_jJ_i - x_iJ_j)}_{\text{Friction}} - \underbrace{x_i\epsilon\alpha_i\frac{dp}{dz}}_{\text{Viscous flow}} \tag{2}$$

where ϵ is the porosity and α_i accounts for the separative character of the membrane on the viscous flow [15]. The parameters $\zeta_{i,j}$ and $\zeta_{i,m}$ account for friction between all components in the system: solute–solvent, solute–membrane and solvent–membrane. From this form of the MS equation, one can see that the flux gets contributions from (i) concentration diffusion, (ii) thermodynamic non-idealities, (iii) pressure diffusion, (iv) friction with other species, and (v) viscous flow. The term containing non-ideality is often neglected, which is acceptable when the gradient $d\ln\gamma_i/dz$ is negligible, i.e., when the concentrations within the membrane, or variations in concentration, are small [22]. However, several studies have shown that non-ideal thermodynamics can have a significant impact on the separation process [7,23,24].

Notice that Fick's law is incorporated in the first term of Equation (2). In analogy, we define the Maxwell–Stefan diffusion coefficient $\mathcal{D}_{i,j}$ between species i and j as

$$\mathcal{D}_{i,j} = \frac{RT}{\zeta_{i,j}} \tag{3}$$

where j could be either the membrane ($j = m$) or any of the permeating species. In the literature, one may find the Maxwell–Stefan equation to be described with either diffusivities $\mathcal{D}_{i,j}$ or friction coefficients $\zeta_{i,j}$. It should be noted, however, that these $\mathcal{D}_{i,j}$ are not identical to the regular diffusion coefficients found in Fick's law.

The MS theory is a very general model for membrane separation as shown by the large number of parameters in Equation (2) describing all mutual interaction in the system. A bookkeeping of all parameters for a system with N different species of solutes and solvents combined is given in the following table:

Parameter	$\zeta_{i,m}$	$\zeta_{i,j}$	ϵ	α_i	Total
Amount	N	$N(N-1)/2$	1	N	$N^2/2+3N/2+1$

For a binary system (i.e., one solvent and one solute) $N = 2$, this yields a total of six parameters.

Alternatively, the MS equations can be written down without a separate term dedicated to describing viscous flow. In fact, in the original MS equations, the last term in (2) is absent. The model was only later altered by Mason and coworkers to account for viscous flow separately [21]. To distinguish the two approaches of the MS model, the method discussed above (Equation (2)) is referred to as the structured approach, while the one with no distinct viscous flow term is referred to as the overall approach [19]. The physical concepts behind the different approaches are, however, quite distinct. In the structured approach discussed above, the membrane is considered to be a completely distinct phase, for example, merely serving as a pore system through which the permeating species can flow. In the overall approach, the membrane and permeating species are instead considered to be one thermodynamic phase, making the membrane an active participant in transport. Hence, the driving force of flux is fully described by the gradients of the chemical potentials, and there is no need for an extra term describing the viscous flow of the mix through the membrane, as the membrane is itself considered a stationary component of this mix. Accordingly, this alternative overall form of the MS equations is easily obtained if one assumes that friction

results from a difference in the absolute velocity of different species (rather than just the diffusional velocity), i.e., by replacing u_i by $v_i = J_i/(x_i c_{tot}\epsilon)$ in Equation (1). Then, however, the friction coefficients, say $\zeta'_{i,j}$, appearing in this alternative form, are not identical to the $\zeta_{i,j}$ in (2), and neither is their physical interpretation [15]. Although this difference in approach may give the impression that the two models are completely distinct, they are in fact mathematically identical, as it has been shown that the two approaches can be obtained from one another by pure algebraic manipulations, without any physical assumptions, resulting in equations relating $\zeta_{i,j}$ to $\zeta'_{i,j}$ [19,21,25]. This means that in such models, viscous flow may indeed be present but not in an obvious way, as it is hidden in the parameters. The advantage of using the equations as formulated in the structured approach (Equation (2)) is that they make investigating the contributions from different transport mechanisms easier (see, for example, [7]).

In the structured approach, the parameters $\zeta_{i,j}$ only account for friction between species i and j, without any (hidden) contributions from viscous flow. Accordingly, since the friction that species i applies on j is the same as that of j on i, it makes sense that the $\zeta_{i,j}$ are symmetric:

$$\zeta_{i,j} = \zeta_{j,i} \tag{4}$$

which can be traced back to be exactly the Onsager symmetry found in linear irreversible thermodynamics [26]. However, in the overall approach, the Onsager symmetry might be broken ($\zeta'_{i,j} \neq \zeta'_{j,i}$) since the friction coefficients $\zeta'_{i,j}$ now also contain a contribution from viscous flow on species i that might influence the permeating species differently due to some separative behavior of the flow. To be more precise, the symmetry is not broken but merely hidden inside the coefficients because $\zeta_{i,j} = \zeta_{j,i}$ still holds. Note that in the overall approach, Onsager symmetry can still hold but only if the viscous flow affects all permeating species in the same way, i.e., when $\alpha_i = \alpha_j$ ($\forall\ i,\ j$). Consequently, the symmetry also holds in a pure diffusive system when there is no viscous flow at all. This can also be confirmed from the equations relating $\zeta'_{i,j}$ to $\zeta_{i,j}$, which show that indeed $\zeta'_{i,j} = \zeta'_{j,i}$ when $\alpha_i = \alpha_j$ ($\forall\ i,\ j$) [19].

2.2. How to Solve the Differential Equations

2.2.1. Binary Systems

In practice, the approximation to binary systems (1 solvent, 1 solute) can often be made. For these systems, the differential equations simplify significantly, and it allows us to write them in a more elegant way. Many of the important transport theories mentioned above were originally written down for binary systems and only later altered to apply for multi-component systems.

Writing the differential equations in their more elegant binary form is not only possible for a 1 solvent–1 solute system but it may also be an acceptable approximation for a multi-component system if (i) one can treat multiple solvents as one new solvent mix, and if (ii) one could at the same time assume that different solutes move independently from each other (i.e., non-interacting), thus allowing one to describe solutes by multiple mutually uncoupled differential equations of exactly the same form, only differing in the values of their physical parameters. Under these assumptions, any binary model can be trivially extended to a multi-component system: simply adopt the equation for solvent/volumetric flux, which now describes the solvent mix as a whole, and make M copies of the binary solute equation, one for each of the M solutes, which only differ by the values of their physical parameters. Below, the solute flux J_i and all parameters $\{X_i\}$ appearing in the binary solute equation are labeled with a subscript i so that upon extension of the binary model to a multi-component system, each extra solute j is represented by an identical equation for J_j while having an associated set of parameters $\{X_j\}$, differing from $\{X_i\}$. An example of how this can be performed in practice is given in Section 2.3.1. Due to the elegance and simplicity of binary models, and their trivial extension to a multi-component system, the majority of models discussed below are formatted to binary solutions.

For a binary system, the MS Equation (2) reduce to

$$J_1 + \frac{\zeta_{1,2}}{\zeta_{1,m}}(x_2 J_1 - x_1 J_2) = -\frac{\epsilon c_{tot} x_1}{\zeta_{1,m}} \frac{d\mu_1}{dz} - x_1 \epsilon \alpha_1 \frac{dp}{dz}, \tag{5}$$

$$J_2 + \frac{\zeta_{2,1}}{\zeta_{2,m}}(x_1 J_2 - x_2 J_1) = -\frac{\epsilon c_{tot} x_2}{\zeta_{2,m}} \frac{d\mu_2}{dz} - x_2 \epsilon \alpha_2 \frac{dp}{dz} \tag{6}$$

where the gradients of the chemical potentials $\frac{d\mu_i}{dz} = RT\frac{d}{dz}\ln(x_i\gamma_i) + \nu_i\frac{dp}{dz}$ are related to one another via the Gibbs–Duhem relation $x_1\frac{d\mu_1}{dz} + x_2\frac{d\mu_2}{dz} = \frac{1}{c_{tot}}\frac{dp}{dz}$, and where additionally, due to mass balance, the mole fractions are related to the molar volumes as $x_1\nu_1 + x_2\nu_2 = 1/c_{tot}$. Finally, in the mathematically identical overall approach of the MS model, the last terms, separately describing viscous flow, do not appear.

It may also be useful to rewrite the flux equations for J_1 and J_2 to equations for solute flux J_2 and volumetric flux $J_V = \nu_1 J_1 + \nu_2 J_2$, which contains, of course, the same amount of information, only written in a different form. For example, the irreversible thermodynamics models are formatted this way. Mason and Lonsdale derived a simple expression for the binary MS model in terms of the volumetric flux J_V, given by [15]

$$J_V = -L_p\left(\frac{dp}{dz} - \sigma_v\frac{d\pi}{dz}\right), \tag{7}$$

$$J_2 = -\omega c_1 \nu_1 \frac{d\pi}{dz} + c_2[1 - \sigma_s c_1 \nu_1] J_V \tag{8}$$

where $\pi = RTc_2\frac{d}{dz}\ln(c_2\gamma_2)$ is the osmotic pressure and parameters L_p, σ_v, σ_s, ω are all in function of the MS parameters $\zeta_{i,j}$, $\zeta_{i,m}$, α_i, ϵ and concentrations c_1, c_2 [15]. Note that these four parameters are less than the six found in (5) and (6), so the algebraic manipulations result in some loss of information of the model. An example of this loss of information is that it is no longer possible, from L_p, to determine how much of the volumetric flux J_V originates from viscous flow, and how much from diffusion. Additionally, the porosity ϵ, and its contribution to the flux, is completely hidden and contained within the parameters L_p and ω. From the volumetric MS Equations (7) and (8), it is easy to see that the irreversible thermodynamics models (see below) are a special case of Maxwell–Stefan theory.

We finally note that if a given model provides equations for solute and solvent flux that differ in form, the user needs to be able to point out which of the particle species is solvent and which is solute. This distinction is not necessary in, for example, the MS model (5) and (6) where the solute and solvent are treated alike, but it is, however, necessary for the volumetric MS Equations (7) and (8).

2.2.2. Membrane Boundary and Non-Ideal Thermodynamics

To solve the differential equations, membrane boundary assumptions are needed. An assumption made in almost all models is that each boundary of the membrane is in thermodynamic equilibrium, with the exception of, for example, transport models that include chemical reactions [27]. Denoting quantities q on the membrane-phase side of the membrane–liquid boundary as $q^{(m)}$ and on the external liquid side of the membrane–liquid boundary as $q^{(0)}$ allows us to write this thermodynamic equilibrium condition as

$$\mu^{(0)} = \mu^{(m)} \tag{9}$$

which holds on either side of the membrane. Note that the distance between (0) and (m) is infinitesimal. To express this condition in terms of concentrations, one can integrate the differential of the chemical potential $d\mu = RTd\ln(\gamma c) + \nu dp$ over just the membrane–liquid boundary. Use (9), and rewrite for $c^{(m)}$ to obtain

$$c^{(m)} = Kc^{(0)}e^{\nu(p^{(m)} - p^{(0)})/(RT)} \tag{10}$$

where the sorption coefficient $K = \gamma^{(0)}/\gamma^{(m)}$ is defined as the ratio between the activity coefficients of the internal and external phases. Note that Equation (10) holds on either of the two membrane borders. Since the distance between (0) and (m) is infinitesimal, $(p^{(m)} - p^{(0)})$ only differs from zero when the pressure changes discontinuously at the membrane boundary. If the pressure changes continuously across the boundary, the exponential in (10) vanishes, leaving only the condition

$$c^{(m)} = Kc^{(0)}. \tag{11}$$

This condition becomes important when non-ideal thermodynamics are considered (captured by K), which can result in a significant effect on the solute rejection [7,24]. A continuously changing pressure is, however, not always the case, like in the solution–diffusion model, where the pressure is assumed to decrease discontinuously on the permeate side of the membrane boundary, see Section 2.3.3.

The boundary conditions (10) and (11) relate the concentration in the external liquid to the concentration inside the membrane at either of the two borders of the membrane. The concentration at some point within the membrane is obtained from solving the flux differential equations for the boundary conditions at $z = 0$ and $z = \Delta z$ for a membrane with thickness Δz and starting at $z = 0$.

2.2.3. Rejection Calculation

Together with the flux, solute rejection is the most important quantity that one wishes to predict for membrane processes. The rejection quantifies what fraction of solute has successfully been held back by the membrane. While a transport model typically only provides an equation for solute and solvent flux, the rejection usually has to be calculated from these equations by solving them for the appropriate boundary conditions. In particular, the rejection R_i of solute i is calculated by comparing the solute concentration at the permeate and feed side of the membrane according to

$$R_i = 1 - \frac{c_i''}{c_i'} \tag{12}$$

where $c_i'' = c_i(z = \Delta z)$ and $c_i' = c_i(z = 0)$ are, respectively, the permeate and feed side concentrations of the solute, and are calculated by solving the differential flux equations for the appropriate boundary conditions at $z = 0$ and $z = \Delta z$ before applying Equations (10) and (11).

2.3. More Specific Models

As mentioned before, there is a vast landscape of models (SD, PF, and IT) trying to describe transport across the membrane, all with their own assumptions. Figure 2 displays the assumptions made on profiles of the chemical potential μ_i and activity $c_i\gamma_i$ of species i and the pressure for all different model approaches in a one-component solution. A summary of all model parameters is given in Table 2. We note that neither in the MS model, nor any more specific model discussed below, are there any known explicit relations of these parameters to the characteristics of solutes, solvents and membranes. To be complete, we also mention the development of a variety of semi-empirical models, mostly elaborations of one of the mechanistic models described here [4].

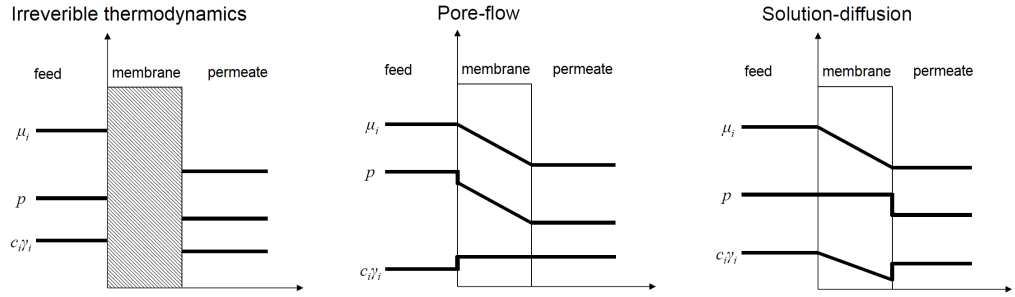

Figure 2. Profiles of the chemical potential μ_i, activity $c_i\gamma_i$, and the pressure in a one-component solution, displayed in the format familiarly used in the literature [4,14,27,28].

Table 2. Parameters in different binary models.

Model	Parameters	Amount
MS	$\zeta_{1,2}$, $\zeta_{1,m}$, $\zeta_{2,m}$, α_1, α_2, ϵ	6
MS—Volumetric form	L_p, σ_v, σ_s, ω	4
IT—Kedem–Katchalsky	P_i, σ_i, L	3
IT—Spiegler–Kedem	P_i, σ_i, L	3
PF—ext. Nernst–Planck	$K_{c,1}$, $K_{d,1}$, $K_{c,2}$, $K_{d,2}$	4
PF—Nernst–Planck	K_c, K_d, L	3
PF—σ-models	λ	1
SD—Imperfections	L_{sd}, L_{imp}, $P_{i,dif}$	3
SD—Classical	P_1, P_2	2
SD—Simplified	L, P_i	2

2.3.1. Irreversible Thermodynamics Models

Historically, these models were the first to use irreversible thermodynamics to describe membrane transport. Later, it was shown that they can also be derived from the Maxwell–Stefan model [15]. Unlike the SD and PF models, these models do not make any a priori assumptions regarding membrane properties and transport modes (see also Figure 2). The separation mechanism is purely based on energy dissipation and entropy production.

Kedem–Katchalsky model

Kedem and Katchalsky [29] were the first to create a model based on these ideas. In their model, the volumetric flux J_V and solute flux J_i are given by

$$J_V = L(\Delta p - \sigma_i \Delta\pi), \tag{13}$$

$$J_i = -P_i \Delta c_i + J_V\ (1 - \sigma_i)\bar{c}_i \tag{14}$$

where $\Delta\pi$ is the osmotic pressure difference over the membrane and $\bar{c}_i$ is the mean concentration of solute i inside the membrane. Note that the solute flux consists of a diffusion term, proportional to Δc, and a convection term, proportional to J_V. The binary model provides a total of three parameters, namely L, σ_i, and P_i. Note that there is a pair of parameters σ_i and P_i for every species of solute i in the system. The parameter L, called the solvent permeability coefficient, is the proportionality factor between the solvent flux and the pressure driving force. The parameter P_i is the permeability coefficient of solute i. It is the proportionality factor between the solute flux and its concentration gradient, describing the strength of the solute diffusion as a transport mechanism.

The parameter σ_i is called the reflection coefficient of solute i. If $\sigma = 0$, the convection term is at its maximum value, meaning that the solute can become easily dragged trough the membrane by the bulk motion of the solvent. Instead, one expects a well-performing membrane to retain the solute or, in other words, one expects the viscous flow of solute, caused by the bulk flow of solvent, to be reflected. Thus, $\sigma = 0$ corresponds to a non-selective membrane, as both the solute and solvent can perfectly permeate it. If $\sigma = 1$, the membrane is perfectly semi-permeable, i.e., the solvent can pass through the membrane,

but the solute cannot, and the viscous flow of solute particles (caused by the solvent flow) is reflected. However, there can still be the transport of the solute through the membrane via diffusion. Alternatively, one can also view σ as the maximum value that the rejection can take on; see the Spiegler–Kedem model below. Because of this, it may be helpful to rescale the rejection R to R/σ, which always has theoretical values between 0 and 1. We mention here that in the pore flow models, σ can be calculated from λ, the ratio of the solute size to the pore size. For different pore structure assumptions, different $\sigma(\lambda)$ equations apply.

Assuming no interactions between the solutes, this binary model can be trivially extended to non-binary systems, as will be the case for all models that follow. In a non-binary system, there exists still one equation for volumetric flux, as the solvent is modeled as a solvent mixture, while there are M solute flux equations, one for each of the M species of solute i, only differing in values of their physical parameters σ_i and P_i. Non-binary systems containing M solutes will thus have $2M + 1$ parameters.

Spiegler–Kedem model

In the Spiegler–Kedem (SK) version of the model, the equations for solute and solvent flux J_V and J_i are very similar to those in the KK model, be it in differential form, and given by [30]

$$J_V = L(\Delta p - \sigma_i \Delta \pi), \tag{15}$$

$$J_i = -P_i \Delta z \frac{dc_i}{dz} + J_V \, (1 - \sigma_i) c_i. \tag{16}$$

This model has the same parameters of L, P_i and σ_i as the Kedem–Katchalsky model, and can similarly be extended to multiple solutes. Note the similarity between these equations and the binary volumetric MS Equations (7) and (8). It can be shown that the SK equations can in fact be derived from this volumetric Maxwell–Stefan theory by assuming dilute ideal solutions, and taking $\sigma_s = \sigma_v = \sigma$ [15]. These approximations result in some loss of information and a decrease in the amount of parameters from four to three [15].

The solute rejection $R_i \equiv (1 - c_i''/c_i')$ can be derived by solving these differential equations for the boundary conditions $c_i'' = c_i(\Delta z)$ and $c_i' = c_i(0)$ (the solute concentration in the permeate and feed respectively):

$$R_i = \frac{1 - f}{1 - \sigma_i f} \sigma_i \tag{17}$$

with the function $f \equiv \exp(-J_V[1 - \sigma_i]/P_i)$. The equation for rejection R_i clearly shows that it has a maximum value $R_{i,max} = \sigma_i$ (if $f = 0$), which provides the interpretation of the reflection coefficient σ_i as the maximal obtainable rejection for solute i.

2.3.2. Pore Flow Models

In pore flow (PF) models, transport occurs by viscous flow through the membrane pores [4,14]. The membrane is not an active participant but rather a pass way, where the permeating species can flow through, unlike in SD models. The actual separation mechanism is assumed to be size exclusion, i.e., if the solute is larger than the pore diameter, it is rejected. Accordingly, PF models are used to describe transport through porous membranes, whereas OSN membranes are generally between porous and dense. Such membranes are traditionally, though not necessarily, assumed to have a constant bulk concentration throughout the inside of the membrane (see Figure 2), but it need not be constant on the membrane–permeate boundary, where an instant and discontinuous decrease might occur caused by rejection due to size exclusion.

Nernst–Planck equation

To describe the solute flux, a great part of PF models use the Nernst–Planck (NP) equation or a derivation thereof, assuming that the solutes are smaller than the pore diameter and have in fact already passed the size excluding membrane–permeate boundary.

In one dimension and electrostatic conditions, the NP equation describes the solute flux as a sum of Fickian diffusion, viscous flow (or rather advection, solute transport carried by bulk motion of solvent) and electromigration:

$$J_i = -D_i \frac{dc_i}{dz} + c_i \alpha_i J_V - \frac{D_i Z_i \mathcal{F}}{RT} c_i \frac{d\psi}{dz} \tag{18}$$

where α_i accounts, like in the MS model, for the selective character of the membrane on the viscous flow. The concentration of solutes is allowed to vary through the membrane. For a calculation of rejection, we refer to Bowen and Welfoot [22]. It is important to note that the Nernst–Planck equation can be completely derived from the Maxwell–Stefan equation. This means that all models that are derived from the Nernst–Planck equations below are simplifications of the MS model. In PF models, the volumetric flux is only affected by a pressure gradient due to the assumed constant bulk concentration inside the membrane (see Figure 2):

$$J_V = -L \frac{dp}{dz} \tag{19}$$

which could be derived from either hydrodynamic analysis or irreversible thermodynamics, while any temperature dependence can be absorbed by the proportionality factor because thermal equilibrium is assumed. This equation provides just one parameter, L, denoting the permeability coefficient. However, the equation for J_V is of course very general, giving no information on the permeability L. There are several models that predict specific values for L, most notably, when the membrane can be modeled as consisting of parallel cylindrical pores, i.e., the well-known Hagen–Poiseuille model, or consisting of closely packed spherical beads, as in the Carman–Kozeny model [31]. In both cases, L depends on the pore diameter, porosity, tortuosity and is inversely proportional to the viscosity.

In the context of membrane transport models, the Nernst–Planck equation is often presented as the 'Donnan steric pore flow model', which is mathematically completely equivalent to the Nernst–Planck model, apart from the physical interpretation of its parameters. The NP equation was in this way first adapted by Bowen et al. [22,32] to be applied to membrane transport models governed by steric effects. In particular, the diffusion coefficient is now accompanied by the diffusive hindrance factor $K_{d,i}$, altering the strength of diffusion, while α_i is replaced by the convective hindrance factor $K_{c,i}$, which yields the solute flux [22,32]

$$J_i = -K_{i,d} D_i \frac{dc_i}{dz} + K_{i,c} c_i J_V \tag{20}$$

where the electromigration term is neglected here. As mentioned before, the separation mechanism in pore flow models is mainly size exclusion. Accordingly, size exclusion and steric effects are incorporated into the physical parameters $K_{d,i}$ and $K_{c,i}$, which are expected to depend on λ, the ratio between the solute and pore radius [32].

In total, these models count three parameters, namely two describing the solute flux ($K_{d,i}$ and $K_{c,i}$), and one describing the solvent flux (L in Equation 19). When extending the model to M solutes, it is described by $M + 1$ flux equations (M for the solutes and 1 for a solvent-mix), accounting for $2M + 1$ parameters (M times $\{K_{d,i}, K_{c,i}\}$ and L). The decrease in parameters compared to the MS model shows that the NP models are less general. This loss of information is evident from, for example, the fact that only one parameter governs the whole of the volumetric flux, without any regard for the diffusive transport of bulk fluid, and the absence of any considerations of non-idealities (though this could be added for the solute by replacing its concentration c_i by activity $\gamma_i c_i$).

There are several more examples of Nernst–Planck-derived models for membrane transport, for example, surface force PF models [33] (and variations) and the finely porous model [34]. Both examples use the same equation but different physical parameters, replacing $K_{d,i}$ and $K_{c,i}$ by $1/(\chi_{i,V} + \chi_{i,m})$ and $\chi_{i,V}/(\chi_{i,V} + \chi_{i,m})$, respectively, where $\chi_{i,V}$ $\chi_{i,m}$ are interpreted as friction coefficients between solute–solute and solvent–membrane. The difference, however, between the Donnan steric pore flow model and these example models

are that the examples specify the solute distribution at membrane–solution boundary and the radial dependence of the solute concentration within the pores (so these models are necessarily 3D). They do this to account for interactions between the membrane and solute. Similar to these examples are the space charge models, which also specify some radial dependence of the solute concentration and additionally assume a surface charge. More details can also be found in the following reviews [4,14].

Extended Nernst–Planck equation

Several authors also discuss the so-called extended Nernst–Planck equation, for example, used by Bowen et al., as a model for membrane transport [22,32], given by

$$J_i = K_{c,i} c_i J_V - \frac{K_{d,i} D_i c_i}{RT} \frac{d\mu_i}{dz} - \frac{K_{d,i} D_i Z_i c_i}{RT} \mathcal{F} \frac{d\psi}{dz} \tag{21}$$

where $\frac{d\mu_i}{dz} = RT \frac{d}{dz} \ln(\gamma_i c_i) + \nu_i \frac{dp}{dz}$ and $J_V = \sum_{i=1}^{N} \nu_i J_i$. Equation (21) holds for both the solvent and solute, which results in a total of four parameters for a binary system ($K_{d,1}$, $K_{c,1}$, $K_{d,2}$, and $K_{c,2}$), making the model more general than the previous NP models (see Table 2). The main difference between the extended and regular NP model is that both the solute and solvent are now described by the same equation (compared to the simplified flux in Equation (19)). To a lesser extent, they also differ in the way that the extended NP model includes non-idealities and contains a separate term describing pressure diffusion (included in $d\mu/dz$). The latter, however, could just as well be added to the regular NP model (in analogy to the structured and overall approach of the MS equation), like one may find in the literature, e.g., [22]. Similarly, non-idealities can be included by replacing concentration c by activity γc. Also notice that the concentration of both the solute and solvent is now not necessarily constant throughout the whole of the membrane.

Upon inspection, the extended Nernst–Planck equation is very similar to the binary Maxwell–Stefan flux equation. Some algebraic manipulations can show that they are, in fact, mathematically identical, up to a rescaling of parameters. Therefore, it can be useful to think of this model as an effective implementation of the MS flux equation in the PF formalism. That is, the MS equation, but assuming the same separation mechanism, physical parameters, and concentration and pressure profiles as in the PF models.

Empirical σ-models

Next to the Nernst–Planck models, there exists a whole set of empirical models that try to use arguments based on size exclusion and steric effects to describe the reflection coefficient σ from the IT models, referred to as σ models in this review. Accordingly, they are most appropriately classified as PF models due to the pore flow nature of the arguments used by these models. Examples of σ models include the Verniory model, Ferry model, steric hindrance model and log-normal model, which all describe σ as a function of λ, the ratio of the solute to the pore radius. For a detailed discussion, we refer to the excellent review of Marchetti et al. [4].

2.3.3. Solution–Diffusion Models

Solution–diffusion models are a class of transport models, where the transport mechanism is dominated by diffusion, originally developed by Lonsdale in 1965 [35]. These models assume the permeating species to dissolve in the membrane material and molecularly diffuse through it as a consequence of the concentration gradient. The membrane material is thus assumed to be an active participant on the molecular level. Think of the membrane as being part of the solution, a liquid component with zero velocity. This is not the case for the PF models, where the membrane is composed of open pores through which the permeating species can flow. Next to the concentration gradient, a pressure gradient is also assumed not to be smooth or continuous as in the PF models but rather discontinuous at the permeate–membrane boundary, while it is assumed that the pressure is uniform within the membrane (see Figure 2). This internal constant pressure is a consequence of the assumption that the membrane phase is an active participant of the solution so that the

membrane transmits pressure in the same way as liquids. Accordingly, the membranes on which these models apply are usually dense membranes.

Next to the generic profiles of concentration and pressure, the main assumptions of SD models are that the flux is driven by a chemical potential gradient $d\mu/dz$, set by a pressure and concentration gradient, and that the flux of the solute and solvent are independent from each other [27]. In the binary MS equation, the latter assumption would correspond to assuming that there is no friction between the solute and solvent, or $\zeta_{1,2} = 0$. The most general equation for flux driven by a chemical potential and independent of other species is

$$J_i = -\frac{L_i}{RT}\frac{d\mu_i}{dz} = -L_i\frac{d\ln(\gamma_i c_i)}{dz} - \frac{L_i v_i}{RT}\frac{dp}{dz} \tag{22}$$

which holds for every species present in the system, be it a solute or solvent. Such an equation is also suggested by linear irreversible thermodynamics, where the proportionality factor L_i between the flux and drivings force is referred to as an Onsager coefficient. This equation then defines regular SD models. Since there is one parameter per equation of flux, the regular solution–diffusion model has a mere two parameters in a binary system, one for each permeating species i (see Table 2). In much of the literature, the factor RT is absorbed in the proportionality factor L_i, which is acceptable because thermal equilibrium is assumed. For a rejection calculation, we refer to [28].

We note that assuming that the solute flux is uncoupled from the solvent flux is a strong approximation. For a perfect diffusive system, this can make sense since we then assume that there is no viscous bulk motion of the solvent that drags the solute along. But even then, one might expect some coupling, for example, originating from friction, exactly like in the Maxwell–Stefan equation. Instead, in the regular SD models, there is only a diffusive transport mechanism without any mutual interaction, and hence, no solute–solvent cross coupling in the flux equations. However, for OSN, and in particular for ceramic membranes, this cannot be accurate, as the solute flux is known to very much be dependent on the solvent flux due to its bulk motion [36,37]. To asses these shortcomings, more advanced SD models were developed (see further).

The classical and simplified SD models

To obtain the classical SD models, we assume $\gamma_i = cst$ and a constant pressure within the membrane (justified above, see Figure 2), which simplifies Equation (22) to

$$J_i = -\frac{L_i}{c_i}\frac{dc_i}{dz} \tag{23}$$

where $D_i = L_i/c_i$ can be recognized as a diffusion coefficient [27]. Simplification can in this case go further by actually solving the differential equation. Integrating both sides of Equation (23) by separating variables and assuming constant flux throughout the membrane, as is appropriate in steady-state, gives

$$J_i\Delta z = -D_i(c_i'' - c_i')$$

with membrane thickness Δz, and where we, again, use the notation $c'' = c(\Delta z)$ and $c' = c(0)$. Finally, to obtain the flux J_i in terms of concentrations outside the membrane, we use boundary condition (11) for the membrane surface on the feed side (where there is no discontinuous change in pressure) and (10) for the membrane surface on the permeate side (where there is a discontinuous decrease in pressure). This finally gives the **classical solution–sdiffusion model**

$$J_i = P_i\left(c_i' - c_i''e^{-v_i\Delta p/(RT)}\right) \tag{24}$$

where $P_i \equiv D_iK_i/\Delta z$ is called the permeability coefficient. For the solvent, experiencing osmotic pressure, the flux equation can be simplified further. Consider that the pressure difference equals the osmotic pressure $\Delta p = \Delta\pi$, which means that there is no solvent flux

across the membrane, $J_1 = 0$, leading to $c_1'' = c_1' \, e^{\nu_1 \Delta\pi/(RT)}$. Filling this back into the flux equation for the solvent gives

$$J_1 = P_1 c_1' \left(1 - e^{-\nu_1(\Delta p - \Delta\pi)/RT}\right). \tag{25}$$

Finally, the **simplified solution–diffusion model** is obtained when the exponents in Equations (24) and (25) are very small so that the solute and solvent fluxes become [27]

$$J_1 = P_1 c_1' \frac{\nu_1(\Delta p - \Delta\pi)}{RT} \equiv L_1(\Delta p - \Delta\pi), \tag{26}$$

$$J_i = P_i(c_i' - c_i'') \equiv -P_i \Delta c_i \tag{27}$$

where, in the first equation, a first-order Taylor series is used, and in the second equation, the exponent is approximated to unity. The non-linear classical SD model is reduced to a linear one. This implies that species with a large molar volume may induce non-linear effects. One can also notice that this simplified SD model has the same mathematical form as the Kedem–Katchalsky model for an ideal semi-permeable membrane, so one with $\sigma_i = 1$.

Extended SD models

Paul noticed the shortcomings of the solution–diffusion model due to, for example, the uncoupled solute and solvent fluxes [28]. To overcome these shortcomings of SD models but still keep the same separation mechanism and SD-specific assumption, like their concentration and pressure profiles, more advanced extended SD models, directly based on the MS equation, are used. Mathematically, these models are effectively incorporations of the Maxwell–Stefan equation in the SD ideology, just like the extended Nernst–Planck equation for PF models. The most prominent example of such incorporation is the extended SD model developed by Paul, specifically for polymer membranes [28].

Solution–Diffusion with imperfections

These models combine pore flow and solution–diffusion models to account for both diffusion, viscous flow, and interactions between the permeating species [4,15]. In particular, the model assumes a SD transport mechanism of the solute and solvent diffusing through dense membrane but extends it by allowing for the convective transport of solute and solvent particles through parallel physical pathways or pores (called imperfections) that are larger than both the solute and solvent. Mathematically, this amounts to adding a new term to the SD flux that describes the convective flow through pores, or imperfections:

$$J = J_{SD} + J_{PF}. \tag{28}$$

For example, in the simplified SD formalism, extended by Nernst–Planck (PF) models, one would obtain

$$J_V = L_{sd}(\Delta p - \Delta\pi) + L_{imp}\Delta p, \tag{29}$$

$$J_i = -P_{i,dif}\frac{dc_i}{dz} + c_i J_V \tag{30}$$

where L_{sd} is the permeability coefficient of the membrane matrix, L_{imp} is the permeability coefficient of the imperfections, and $P_{i,dif}$ is the diffusive permeability coefficient. Notice that the solute flux is again coupled to the solvent flux via a convection term $J_V c_i$ like in the SK, KK, and, in particular, the Nernst–Planck models. Thus, the resulting model assumes the transport mechanism of SD models, while dodging possibly their biggest drawback, namely the uncoupled (i.e., non-interacting) solute and solvent particles, at the cost of adding an extra parameter.

SD-imperfection models were originally made to account for swelling, which may create free volumes and thus allow room for viscous flow, or similarly account for defects in the membrane. For a rejection calculation, we refer to Mason and Lonsdale [15].

2.4. Challenges of Mechanistic Models in OSN

All of the mechanistic models described above were actually developed for the membrane transport of aqueous streams. The strength of the SD approach to model and predict the aqueous reverse osmosis process is well known [4]. However, all described models have also extensively been used to model and understand OSN performance, with varying success [4]. Experimental OSN results have clearly shown the complexity of this membrane process, influenced by the wide variety of solute and solvent properties, and the importance of all mutual solute–solvent–membrane affinities and competing interactions, not well covered in many of the simpler mechanistic models. Other challenges in using mechanistic models for fitting or performance prediction are the solvent-dependent swelling of polymeric membranes, even for membranes specifically developed for OSN, and the occurrence of negative rejections in a variety of mixtures. All this leads to fluxes and rejections that are strongly solvent-dependent, even for a fixed set of solute and membrane. Even for the non-swelling ceramic membranes, OSN performance is shown to be solvent dependent, and appears to be influenced by solvent–membrane affinity and the solubility of the solute in the solvent. Due to the limited success of the mechanistic models for OSN, there is a tendency in the field to switch to data-driven modeling. The data-driven approach aims for better prediction power but can also help to reveal the solute–solvent–membrane descriptors governing the performance, something that is hardly possible with the mechanistic models, as they provide no direct relation between the solute–solvent–membrane properties and the model parameters. The mechanistic models are, however, still valuable for retrieving the relative importance of different transport mechanisms, like diffusion or convection, and of non-idealities [7]. Moreover, in combination with data-driven approaches, so-called hybrid modeling, extra insights can be derived, and possible relationships of the model parameters with specific solute–solvent–membrane properties can be revealed. Additionally, since every model assumes some kind of ruling separation mechanism and assumes a certain physical context, understanding what mechanistic model describes the process best means understanding the separation process.

For completion, we remark that in real application situations, the solute concentration can be relatively high, leading to concentration polarization effects, not covered in the described physical models. This can, however, be assessed by using film theory in combination with one of the physical models [4]. On the contrary, the mixtures used in lab testing are mostly highly diluted, avoiding any concentration polarization and the need for using film theory.

3. Data Collection

3.1. Data Availability

The general rise of data-driven modeling is in part fueled by the increase in data availability, attributed to the steep increase in storing capabilities, and rise of the internet, which allows for the collection and distribution of large datasets [38]. While this a general trend, there are large differences in the data management maturity between scientific disciplines and communities. Historically, the scientific community has had a strong focus on the reuse of the findings of other researchers, while in the context of data-driven modeling, there is a need for their reuse by machines. The challenges associated with this goal are summarized in the four FAIR principles: findability, accessibility, interoperability, and reusability of data by machines [13].

Implementing these principles in practice is challenging and requires both technological solutions and community-agreed standardization. As made clear by recent publications in the field of OSN, there is still a lot of room for improvement for the FAIRification of its data and their subsequent use. In their survey on scientific publications within the field

of OSN, Le Phuong et al. [39] note the general lack of standardization of measurement setups, like dead-end and cross-flow processing or temperature and pressure, leading to difficulties in comparing results as well as complicating the industrial implementation. Moreover, there is no established, standardized membrane characterization test as molecular weight cut-off (MWCO) measurements in aqueous filtrations. In [10], the authors run into the lack of standardization and realize that membrane–solute and solute–solute interaction assessment is limited, owing to insufficient comparable datasets in the literature. Additionally, they note the lack of the reproducibility of results in the field of membranes, typically caused by the insufficient availability of datasets [10]. In [40], it is stated that there is no comprehensive study nor large dataset on the rejection behavior of solutes in a wide range of solvents, which results in a general uncertainty in terms of the solvent effect on solute rejection. The observations of these authors and the way they are dealing with the challenges are important drivers to push things forward. An initiative specifically aimed at this goal was made in 2021 when Ignacz and coworkers established an online library (www.osndatabase.com) to collect and share OSN data [41].

3.2. Data as Driver for Models

The input of data-driven modeling is obviously the data themselves. In their raw form, data can be considered as points in a high-dimensional descriptor space (see Section 3.2.2 for more detail). In data-driven research, an important concept is the manifold hypothesis, which states that all natural data lay on a lower-dimensional manifold within the high-dimensional descriptor space and that it is possible to interpolate between two points on this manifold [38]. This is a very general principle, holding for all data-driven modeling. This implies that the ability of a data-driven model to generalize to new data is a consequence of the structure of the data more than being a property of the model itself. From this perspective, there are two crucial conditions for a model to make good interpolations. On the one hand, one needs to find a good data structure, i.e., find a limited set of descriptors containing the necessary information about the flux and solute rejection of the separation process. On the other hand, the input space should be sufficiently densely filled. An overview of these two points is provided in the following subsections.

3.2.1. Data Density

While it is perhaps intuitive that data with higher density lead to better model performance, this is also supported by the OSN literature. As an example, we refer to [41], where it was found that closer chemical similarity between two molecules leads to closer observed rejections. So for data-driven models to work in OSN, i.e., to interpolate the data space to new solute–solvent couples, the large input space of many possible solute–solvent couples needs to be filled to a sufficiently dense degree. However, it turns out that the available data in OSN are currently not meeting this goal [41]. This lack of diversity in OSN data is perhaps not surprising since most data are produced with specific applications in mind and not aimed at filling the data space. Several authors have clearly stressed the need for implementing fast and reliable solute-testing methods, making the parallel with the pharmaceutical industry where high-throughput screening techniques allow the assessment of large amounts of molecules in single systems [42]. Medium throughput systems (MTS) have been used in OSN in the past to generate a large quantity of high-quality data [43,44]. More recently, and in light of the rising interest in data-driven approaches in OSN, Ignacz et al. [41] proposed MTS as a robust method to efficiently fill in the chemical space in OSN with new measurements. They proved the strength of MTS in achieving the goal of a dense data space by measuring 336 different solute molecules with an optimized MTS method, having a theoretical weekly throughput of over 100 compounds, doubling the amount of unique solute molecules present in the OSN database at that point in time [41]. Such efforts to fill in the chemical data space are essential for the successful data-driven future of OSN.

Next to the call for denser data, the literature also contains questions about the fundamental structure of the data. It is hypothesized that the dependence of the separation process on the membrane- and solvent properties might lead to a data space which consists of smaller islands, complicating the generalization of data-driven models. The distinction between data collected from cross-flow versus dead-end processing in studies by Hu et al. [6] and Kim et al. [45] is an example of this. In their studies, the distinction between these two was captured by one of the descriptors, indicating that they are essentially described by two different models. A similar issue is at stake for ceramic membranes versus polymeric membranes [46], where the latter experiences swelling, impacting the parameters needed in a model design. Hence, narrowing down the data space may allow for less general but simplified models. Note that, from a data science point of view, making different classes of models reduces the dimension of the descriptor space since it avoids the use of a descriptor addressing the distinct models.

3.2.2. Data Structure: Descriptors of the Input Space

The input space of OSN processes is represented by a combination of descriptors. An active research question is determining what descriptors are most important in describing the separation process. Often, data-driven techniques, either supervised or unsupervised, are used to determine these key descriptors, see also Section 4.2. For example, unsupervised techniques could be used to determine the descriptors with the largest possible data coverage, while supervised techniques could be used to determine which descriptors have the largest influence on the flux or solute rejection. Note that while this is crucial for data-driven modeling, this information is clearly also relevant for mechanistic or hybrid modeling. These key descriptors are then used to represent the input space of the membrane–solute–solvent triplet with as much relevant information as possible while keeping the dimension of input space low.

There exists a distinction in the nature of the descriptors that are used in the literature. They can be divided into the so-called hand-crafted descriptors and theoretical molecular descriptors. The hand-crafted descriptors refer to the variables that have classically been used to directly describe certain physico-chemical observables. The viscosity of a fluid is such an example. One could imagine instead more abstract theoretical molecular descriptors, calculated at the molecular level. These can be built from descriptors as simple as the number of atoms, their type, coordinates, bonds and electrostatic interactions, to 2D molecular graphs, or even based on calculations from density functional theory (dft) [47]. A popular technique for finding such theoretical molecular descriptors are QSAR (quantitative structure–activity relationship) models. The descriptors found by QSAR models could, in theory, be any mathematical relationship that can be calculated from the information at the molecular level. The most useful theoretical molecular descriptors are then determined by supervised data-driven analysis, relating this abstract molecular-level information to some measurable property of the molecule (like rejection, but based on the goal, it could as well be polarity, toxicity, etc.). QSAR descriptors have been used in the past for membranes in general, see e.g., [48,49], while Ignacz and coworkers were the first to apply them for data-driven performance prediction in OSN [10].

Both kinds of descriptors have their advantages. Hand-crafted descriptors describe high-level information that relate one descriptor directly to an observable property, which results in a great degree of interpretability. However, for every new molecule or membrane considered, the value of these descriptors often needs to be measured in wet-lab experiments, which can be time-consuming, labor intensive, and expensive, certainly considering the near infinite amount of possible solute–solvent couples in OSN [10]. Molecular descriptors encompass more abstract but also more fundamental molecular-level information, which reduces the risk of neglecting important properties. For example, they may help to take properties, like solute geometry, into account, which are harder to cover using traditional hand-crafted descriptors. This approach is advocated by Ignacz et al. [10]; they hypothesized solute rejection to be dependent on its molecular structure. A downside

of descriptors determined by QSAR techniques is the inability to verify whether they represent the data to a sufficient degree [10].

4. Data-Driven Modeling

Nowadays, data-driven modeling is receiving more and more interest in process technology, motivated by cheaper, smarter sensor technologies allowing extensive online and real-time monitoring, supplemented by increased computational storage and much faster computation, as well as data transmission capacity. The current evolution in membrane technology shows the same trend. Although a high level of skepticism exists among membrane scientists, in the last 25 years, the use of data-driven, non-mechanistic modeling has grown also in this field.

As further exemplified below, the applied methods are usually supervised classification or regression techniques, such as artificial neural networks (ANNs), support vector machines (SVMs), genetic programming, or variants of decision trees. To improve the generalization of the trained models, they are often combined with techniques for (unsupervised) dimensionality reduction, like principal component analysis (PCA). This is related to the manifold hypothesis, mentioned in the previous section, which means that, in general, the input data can be represented in a lower-dimensional space without significant loss of information. Such techniques reduce the amount of data needed to train the model, or equivalently reduce the risk of overfitting. This improves generalization, the ability to make accurate predictions when fed with new unseen data. Next to pure data-driven approaches, also hybrid modeling is increasingly utilized: combining a mechanistic model with data-driven techniques.

Before we turn to OSN, we first provide a brief overview of the topic in the field of water membrane separation in Section 4.1. It is less complex, and the field has higher maturity but contains, from a data-driven point of view, some very similar challenges. In Section 4.2, we provide an extensive review of data-driven modeling in OSN and also highlight the differences with water separation.

4.1. Data-Driven Modeling in Water Membrane Separation

Regardless of the specific model used, data-driven approaches typically work with physico-chemical properties and molecular size/geometry parameters of the solutes, descriptors of the membrane and of the water matrices, and parameters related to the operating conditions, as input. For a recent review, we refer to Galinha, Crespo, 2021 [50]. Crespo underlines that this data-driven modeling experience has shown that a dedicated analysis of the functions used in the computation algorithms, often considered "black box", can definitely lead to a sensitivity assessment of the mixture–membrane–operation properties that dominate the performance or that are not captured well in the existing mechanistic models, and thus add to a better understanding of the physical phenomena involved. Crespo et al. have particularly been active in developing (hybrid) modeling approaches to predict the performance of membrane bio reactors (MBR) and the membrane-based algae harvesting system. For example, Galinha et al. [51] successfully combined a known activated sludge model (mechanistic model) strengthened by the input of on-line 2D fluorescence spectroscopy to model MBR results. A similar strategy was followed by Sà et al. [52] to model the production of carotenoid-rich Dunaliella salina. Other complexities in pressure-driven processes have also been the focus of data-driven modeling. We present some examples selected from the vast literature. Teodosiu et al. [53] used ANN to predict the complex phenomena of flux decline and restoration during the fouling and backwashing of ultrafiltration membranes. About the same time, R. Bowen, one of the experts in mechanistic model development for salt removal in nanofiltration (NF), also successfully explored the use of ANN to predict single and mixed salt rejections [54]. Yangali-Quintanilla et al. [48] were some of the first researchers to explore machine learning by ANN based on QSARs to predict the rejection of a wide range of neutral organic compounds in polyamide NF and RO membranes. Their study revealed that size exclusion

and hydrophobic membrane–solute interaction dominate the rejection. Sanches et al. [55] zoomed in on nanofiltration for the removal of micropollutants from drinking water sources and proved the importance of size and charge but also the geometry of the micropollutants in predicting their rejection. Barello et al. [56] focused on the desalination performance of the RO membrane under fouling conditions and were able to predict RO performance for different membranes in a wide range of feed salinity and operating pressure. More recently, other groups worked on the modeling of thin film nanocomposite (TFN) membranes for RO. For example, Yeo et al. [57] and Fetanat et al. [58] used ANN to analyze the literature data of a broad range of TFN membranes with a wide variety and concentration of nanoparticles to predict permeability, salt rejection and fouling behavior, and to steer the synthesis of optimized TFNs. Other researchers have also covered other pressure-driven membrane processes such as pervaporation [59], membrane distillation [60] and electrodialysis [61].

4.2. Data-Driven Modeling in OSN

4.2.1. Drivers for Data-Driven OSN

Next to the opportunities that data and data science offer in general, discussed above, we see two specific drivers for data-driven approaches in OSN: an operational one and a conceptual one. While in membrane filtration in water, the solvent is obviously fixed and the process is mainly governed by size exclusion, in OSN, this is supplemented by a whole set of important interactions between all three constituent components of the solute–solvent–membrane. This complexity currently makes detailed ab initio modeling very difficult. The lack of predictive power of the current models [5,6] has major implications on the development process of new operational OSN applications. For each given solute–solvent couple to be separated, membranes and operational parameters need to be screened nearly on a trial and error basis. This results in a repetitive, tedious and time-consuming development process in OSN. Consequently, data-driven models with good predictive power are sought simply because of their high operational value. At the same time, it can be questioned whether a single data-driven model can cover the whole domain and all needs in the field of OSN. This is connected to the need for explainability and more insight into the underlying processes, e.g., for delineating the application area of certain models and for the development and selection of new membrane properties. Also at this conceptual level, data-driven techniques have a role to play. They can help in uncovering key descriptors of the solute–solvent–membrane and correlate them with membrane performance. They can provide insight into the relation between these descriptors and the parameters of the mechanistic models and drive the understanding of the complex physics of OSN.

While most initiatives in data-driven OSN are recent, the field is developing fast, and we provide a review in the following subsection.

4.2.2. Data-Driven OSN: State of the Art

Santos et al. 2007 [62] were the first to use data-driven modeling to assess solvent fluxes for a series of first-generation polymeric OSN membranes with different chemistry (polyamide, polyimide, polyethersulphone, and PDMS). They combined their own experimental data with literature data for a series of solvent–solvent mixtures measured at variable pressures. They compared predictions of the SD model (mechanistic) and several ML models such as PLS, ANN, a combination of PCA and ANN, and all combined hybrid SD-ML models (data-driven models). The descriptors used include standard solvent properties (such as density, viscosity, dielectric constant, molar volume, and dipole moment), as well as the geometrical aspects of solvent molecules. To account for membrane–solvent interactions, the difference between membrane and solvent surface tension and Hansen solubility parameters (HSP) was taken into account. For the membrane itself, only the molecular weight cut-off (MWCO) was used. The results suggest that the solvent transport in OSN for the studied first-generation membranes is dominated by solvent density and viscosity and by the membrane MWCO. Furthermore, it can be concluded that membrane–

solvent affinity is best described by the HSP (and not surface tension) difference, and that the SD model particularly misses the incorporation of solvent polarity.

Goebel et al. 2020 [8] continued data-driven modeling to predict pure or mixed solvent fluxes for a first and a new generation of PDMS membranes (later supplied by GMT) and for different ceramic nanofiltration membranes (supplier Inopor). The peculiar ML approach, utilized on each membrane type separately, is a combination of nonlinear regression methods and genetic programming for automatic model development and optimization, leading to a compact (membrane specific) model equation with a strong data prediction capacity, able to outperform mechanistic models. Similar solvent descriptors were used, such as those by Santos, but Goebel also included the total and three partial HSP values (envisioned to play a role in membrane swelling), as well as the connectivity index, allowing to differentiate between linear and branched molecules. To catch the solvent–membrane affinity in the PDMS membranes, the HSP difference was again used, now in combination with the HSP interaction radius (not used for the ceramics). No further membrane–specific descriptor was added.

Shortly after the first paper, Goebel et al. 2020 [9] extended the same data-driven modeling approach to predict the rejection of a variety of solute–solvent couples for PDMS-based Evonik membranes (Puramem). In their introduction, the authors underlined that the strength of data-driven modeling for rejection predictions is even larger than that for flux predictions, due to the observed importance and complexity of all mutual interactions (solute–solvent–membrane), not well described in the mechanistic models. One of the signs of this complex interplay of interactions is the regular observation of negative rejections in OSN. The obtained membrane-specific models have very good prediction capacity and confirm the importance of solute and solvent polarity next to the solute size, consistent with the experimental observations. The developed models allow to draw practical triangular rejection maps.

All previously mentioned papers used relatively small datasets. Hu et al. 2021 [6] were the first to create large-scale data-driven models based on 38,430 data points from 67 sources in the literature (mainly from polymeric membranes but including some data for ceramic membranes as well). The dataset covers results from 35 commercial membranes, 11 solvents, and different solutes, mainly in high dilution. They made three different kinds of ML models, namely ANN, SVM, and random forests (RFs), which could predict flux and rejection with an accuracy (R^2) up to 98% and 91%, respectively, proving the strength of data-driven modeling in OSN. The only clear trend uncovered in their exploratory data analysis is an increase in the rejection as solute MW increases. The analysis also hints to a linear correlation between flux and the characteristic solvent parameter (given by $\delta_p/(\eta d_m^2)$, where δ_p is the polar HSP and d_m is the molar diameter of the solvent [63]). However, due to large data deviations (even for the same membrane), they concluded that the performance of OSN membranes is multi-dimensional and a complex function of many variables. To make sense of this, they used PCA to find the key descriptors affecting this multi-dimensional system and reduce the data dimension. Specifically, six descriptors out of the total 18 they considered were found to be important. Among these six were always the membrane MWCO, the solute MW, the concentration of the solute, the characteristic solvent parameter, the temperature and pressure, and the configuration (i.e., a categorical feature which is either dead-end or cross-flow). Their choice of solute concentration as a descriptor was made because of its impact on performance via the effects of osmotic pressure. Out of these descriptors, solute MW, solute concentration, and the solvent factor were deemed the most important for describing membrane permeability and rejection. The data, pre-processed with PCA, were used to build the successful ML models.

Kim et al. 2021 [45], among the collaborators of the Hu paper discussed in the last paragraph, conducted their own study on ML approaches for OSN modeling, finding the optimal operation specifications and process design. To achieve this, they collected and curated 884 OSN datasets, which were found to be highly structured. They then used SVM and hyperparameter-optimized SVM (HO-SVM) models that predict membrane perfor-

mance (flux and rejection) from the key descriptors influencing performance, which were assessed by a PCA. They used the same descriptors as Hu et al., namely membrane MWCO, solute MW, solute concentration, characteristic solvent parameter, pressure, temperature, and process configuration (dead-end or cross-flow). The ML models reached accuracies (R^2) for flux and rejection of, respectively, 91.4% and 90.6% for the SVM models, and 92.3% and 96.0% for the HO-SVM models, again showing the strength of ML models to predict OSN performance.

Ignacz et al 2022. [10] developed new data-driven models specifically for rejection prediction, which were trained on a large and chemically diverse dataset. The dataset was specifically created for this goal using the same medium-throughput system, three types of polyimide Duramem membranes (previously commercially available from Evonik), a wide variety of different solutes, two pressures, a fixed temperature, and only one solvent, methanol (see also Ignacz 2023 [40]). Their data are freely available on the OSN database website (www.osndatabase.com). Whereas in previous work, the solute was only described by its molecular weight, Ignacz et al. were the first to also take the solute structure specifically into account. The importance of structural solute properties is clearly underlined by the failure of building a nice MWCO curve for each membrane using the raw data. To be able to use structural features in the modeling, they included the molecular structures as SMILES (simplified molecular-input line-entry system, giving a direct string representation of a molecule) in the created open access OSN database. These SMILES, the open-access Mordred [64], and RDKit Python packages, were used to produce relevant molecular descriptors. This allows for the creation of a wide variety of QSAR descriptors (1241 in this case) that also contain information on the 3D electro-topological structure and available functional groups, not captured in the earlier hand-crafted descriptors (like molecular weight). Because the amount of descriptors was higher than the amount of rejection points (416 coming from as many different solutes), partial least square (PLS) regression was used in combination with variable importance in projection (VIP) for the removal of low-value descriptors, and a genetic algorithm (GA) for model optimization. This approach was compared with a graph-encoding or graph-convolutional deep neural network (DL) approach, working directly on the SMILES and the graphs derived from it. Both model approaches showed quite good prediction of solute rejection in methanol, with R^2 scores between 84 and 90%. From the PLS results, it can be concluded that descriptors catching charge, electronic and topological features of the solute molecules influence rejection the most, while molecular weight does not appear important at all (as was already clear from the raw data). We remark that this conclusion is only valid for rejections of Duramem membranes in methanol. From the DL results, conclusions are drawn using a visualization method highlighting which molecular features (functional groups and type of bonds) increase or decrease the rejection.

Wang et al. 2023 [65] developed ML algorithms to predict performance for specifically thin film nanocomposite (TFN) membranes in OSN. The data they used were collected from 20 papers on 119 different TFN membranes in OSN, obtained through their tables and figures, which constitute 9252 data points. TFN membranes are fairly flexible in the sense that their properties can easily be tuned, properties which are determined by the membrane fabrication conditions, such as support type, nanoparticle size, type and loading, amine monomer and concentration, and chlorine monomer and concentration. Their aim here was to use ML models to really aid the development of membranes themselves by finding the optimal fabrication conditions. They did this by using these fabrication variables, and some common solute, solvent and membrane characteristics as descriptors, in total about 19. They used four different data-driven models, namely SVM, boosted tree (BT), ANN, and linear regression. The BT model turned out to be the best model with R^2 values of 92% for permeability and 85% for rejection. They pinpointed the most important of descriptors affecting permeance and rejection by parameter contribution analysis. From this, they concluded that nanoparticle loading, amine concentration and chloride concentration are the most important fabrication parameters for both permeability

and rejection. On top of this, the water contact angle, solvent viscosity and molar volume strongly influence permeability, while the rejection is affected more by the solute molecular weight. Additionally, they used partial dependence plots to evaluate the influence that the fabrication descriptors have on the prediction. From this analysis, it can be derived that nanoparticle loading is extremely critical, and is preferentially kept below 5 wt%. The nanoparticle type is quite unimportant. Similar conclusions can be drawn for the other TFN synthesis parameters.

Ignacz et al. 2023 [40] investigated specifically the effect of solvent descriptors on the solute rejection using graph neural networks (GNN). This study was triggered by the experimental observations that the solute structure plays a role in solute rejection (observed for one solvent, methanol, by Ignacz 2022 [41]) but that these characteristics seem to influence the performance mainly when the membrane–solvent affinity is low [66]. To train their model, they used data from their own experiments consisting of 5004 measurements with over 400 chemically diverse solutes in 11 green solvents using one particular polyimide membrane (Duramem 300 previously supplied by Evonik), created in a medium-throughput cross-flow OSN system already used in their previous study. The new data have, again, been added to their open database. The raw data show that the average rejection (over the wide range of solutes) is strongly related to the average solvent permeance and to the solvent polarity: lower polarity and permeance lead to lower average rejection. This observation is confirmed by the literature data in the database. As in their previous ML work, instead of using handmade descriptors (like solvent viscosity or dipole moment), Ignacz et al. again used more fundamental molecular descriptors obtained from the atom and bond scale, containing also a lot of structural information (using Mordred software). Their models were the first to use descriptors based on the chemical structure for both solutes and solvents. The study revealed that the rejection of solutes heavily depends on the solvent flux, as well as the solvent's electronic properties (e.g., polarity and LogP) and topology, which agrees with the literature. As the values of such properties might be hard to derive analytically, ML methods as described here can be very valuable for this. Furthermore, to demonstrate the robustness and generalization power of their model (R^2 = 86.4%), it was tested against literature data for the same membrane, showing good results (R^2 = 71.4%). However, it revealed some limitation of generalization to new solvents, as the model performed worse on some solvents that were not part of the original training dataset. They attributed this lack of generalization to using no features based on intermolecular (solute–solvent) interactions so that the model has to make predictions just based on similarity between solvents in the training set. Considering that they chose solvents with minimal similarity, their model was still able to make relatively good extrapolations to new solvents. Additionally, they showed that models not including solvent structural features performed significantly worse in predicting rejection, indicating the importance of considering molecular solvent descriptors, and thus the significant effect of the solvent type on rejection in OSN. The influence of swelling on this solvent effect in solute rejection was not assessed.

Recently, Xu et al. 2023 [11] looked at the optimization of the synthesis of polymer OSN membranes with intrinsic microporosity (PIM) using data-driven ML in combination with molecular simulation. They concentrated on finding a PIM with optimal solvent permeability. A literature-derived dataset of 152 solvent permeabilities originating from 35 different membranes (not only PIM) and 16 solvents drove the modeling. The raw data already show the strong effect of the membrane polymer type, solvent molecular size and viscosity on the permeability. Three different ML methods were used, namely kernel ridge regression (KRR), gradient boosting regression (GBR) and the least absolute shrinkage and selection operator (LASSO). As first descriptors, single solvent (diameter, viscosity, and HSP) and membrane (thickness, water contact angle, and nitrogen sorption capacity) properties were used in linear and logarithmic forms, as well as all possible combinations of one solvent plus one membrane descriptor (144 in total). This analysis led to a meaningful phenomenological equation for permeability containing the key descriptors:

solvent viscosity, membrane thickness and water contact angle. In a second approach, relatively more fundamental molecular representations (by numerically fingerprinting chemical structures) of solvents and membranes were used as descriptors. From this, the top 10 influencing fragments of solvents and membranes were derived. ML models using these descriptors were able to predict PIM membrane permeabilities well. In a next step, molecular dynamics simulations were used to quantify swelling in methanol for three types of PIMs, pinpointing the responsible structural features, and confirming its correlation with membrane–solvent affinity. Molecular dynamics simulations were also utilized to assess methanol permeability (using a pressure difference), which was shown to be correlated well to the ML results and the membrane-swelling degree. All modeling suggested an improved type of PIM, and the choice was confirmed by the performance of the synthesized membrane, outperforming existing membranes in the literature (highest methanol permeability, combined with high rejection). The study proved the synergy of ML and molecular dynamics simulations in obtaining microscopic insights in membrane transport and swelling.

Gallo-Molina et al. 2023 [12] explored the potential of hybrid data-driven modeling to better elucidate the link to the physical phenomena driving OSN. They focused on data from ceramic OSN membranes, both native and chemically modified (partly from the literature, partly created). This class of membranes was particularly chosen due to their non-swelling behavior, possibly leading to less-complex models. XGBoost data-driven modeling was used in combination with the mechanistic SD model, both in a parallel and a serial way. Non-idealities in the mixtures were taken into account via activity coefficients estimated via UNIFAC. This approach proved to lead to better flux and rejection predictions compared to the mechanistic model alone. A set of 22 available or easy-to-calculate physico-chemical properties for membranes, solutes and solvents (including HSPs for all) were used as descriptors. Shape indices were included as solute descriptors. The parallel hybrid modeling architecture proved to be the strongest with very high R^2 values of 99% for permeability, and 96% for rejection. From their results, it can be concluded that solvent transport is strongly influenced by solvent polarity and solvent–membrane affinity (captured by the difference in surface energy and in HSP values). Solute transport appears more complex: next to the size exclusion parameters, all mutual affinities between membrane, solute and solvent seem important (captured by the HSP differences, and ratios thereof). Solvent viscosity does not pop up as an important parameter in solvent transport but it does in solute transport, most likely pointing to a partly coupled solute–solvent transport, not included in the SD model. It was expected that hybrid modeling allows better extrapolations to unknown areas in the descriptor space. This was challenged by performance prediction for particular systems where strong membrane–solute affinities are expected to play a role. The relatively low rejections in these cases were quite well confirmed.

5. Conclusions

The field of OSN is following the trend of the widespread increasing interest in data-driven approaches. Recent advances in the field showcase data-driven models able to predict membrane performance to an unprecedented degree of accuracy, outperforming mechanistic models. Such advances may help to speed up process development in OSN, which is generally slow due to time-consuming membrane screening, and thus fast-tracking the industrial implementation of OSN. Moreover, not only can data-driven modeling help predict membrane performance but it can also assess the key descriptors and main transport mechanisms influencing the separation process, thus providing physical insight into the process. Recent work on data-driven modeling investigated both classical hand-crafted descriptors (like viscosity or molar volume) as well as a wide variety of more structural molecular descriptors. The latter group contains more fundamental information, like 2D geometry of the molecules or 3D electronic structure, which is not described by classical

descriptors. Such features, both for solute and solvent, appear important for a good description of OSN.

Needless to say, the strength of data-driven models depends on the structure of the data space on which the model is built. In particular, for an optimal performance, the data space should be densely and diversely filled. Recently, dedicated efforts were aimed at the creation of data to cover the whole chemical space spanned by OSN, using, for example, medium-throughput filtration systems. However, achieving this goal will still prove to be an intricate objective in the future, certainly considering the near infinite amount of possible solute–solvent couples in OSN. Furthermore, much work is still needed to comply with the FAIR (findability, accessibility, interoperability, and reusability of data) principles. In particular, we note a lack of standardization of the measurement process, lack of a standardized membrane characterization test (like MWCO measurements in aqueous filtrations), and insufficient accessible datasets. Also in this regard, recent advances have been taking steps in the right direction with, for example, the establishment of an open-source database, which forms a platform on which data can be freely shared and collected, and further calls for standardization. We believe that a next, ambitious step could be an orchestrated approach to lift interests to the level of the entire discipline, finding ways to overcome hurdles of confidentiality and avoid focusing only on specific projects or short-term goals. Such a realization would obviously benefit not only data-driven models but certainly also mechanistic models and crucially, in our view, a combination of both. Where the balance between the two will lie in a data-intensive discipline, the future will tell.

Mechanistic models, successful in describing membrane separations in water, show only limited applicability in organic solvents. This is due to fluxes and rejections being strongly solvent dependent, even for a fixed set of solute and membrane, which results from the solute–solvent–membrane interactions that complicate the separation process. However, the rise of data-driven models does not at all mean the end of mechanistic models. In fact, fitting the membrane performance to mechanistic models can still reveal information on the relative importance of different transport mechanisms, like diffusion and viscous flow, or reveal the effects of non-idealities. Moreover, in combination with data-driven modeling, so-called hybrid models have a great potential to provide extra insights, and possible relationships of model parameters with specific solute–solvent–membrane properties can be revealed. These approaches using mechanistic models, alongside using data-driven models to investigate descriptor importance, can provide physical insight into the separation process and can ultimately help unravel the transport processes in OSN.

Author Contributions: Writing—original draft preparation, P.-J.P.; writing—review and editing, P.-J.P., P.B., B.C., J.H. and A.B.; visualization, P.-J.P.; supervision, J.H. and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable to this review article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sholl, D.S.; Lively, R.P. Seven chemical separations to change the world. *Nature* **2016**, *532*, 435–437. [CrossRef] [PubMed]
2. Rundquist, E.M.; Pink, C.J.; Livingston, A.G. Organic solvent nanofiltration: A potential alternative to distillation for solvent recovery from crystallisation mother liquors. *Green Chem.* **2012**, *14*, 2197–2205. [CrossRef]
3. *Nanofiltration—An Overview of Technology Development, Status and Trends, D16E*; Frost & Sullivan: San Antonio, TX, USA, 2008.
4. Marchetti, P.; Jimenez Solomon, M.F.; Szekely, G.; Livingston, A. Molecular separation with organic solvent nanofiltration: A critical review. *Chem. Rev.* **2014**, *114*, 10735–10806. [CrossRef] [PubMed]
5. Galizia, M.; Bye, K.P. Advances in organic solvent nanofiltration rely on physical chemistry and polymer chemistry. *Front. Chem.* **2018**, *6*, 511. [CrossRef]
6. Hu, J.; Kim, C.; Halasz, P.; Kim, J.F.; Kim, J.; Szekely, G. Artificial intelligence for performance prediction of organic solvent nanofiltration membranes. *J. Membr. Sci.* **2021**, *619*, 118513. [CrossRef]

7. Claessens, B.; Hitsov, I.; Verliefde, A.; Nopens, I. Analyzing transport in ceramic membranes for organic solvent nanofiltration using Maxwell-Stefan theory. *Chem. Eng. Sci.* **2022**, *264*, 118133. [CrossRef]
8. Goebel, R.; Skiborowski, M. Machine-based learning of predictive models in organic solvent nanofiltration: Pure and mixed solvent flux. *Sep. Purif. Technol.* **2020**, *237*, 116363. [CrossRef]
9. Goebel, R.; Glaser, T.; Skiborowski, M. Machine-based learning of predictive models in organic solvent nanofiltration: Solute rejection in pure and mixed solvents. *Sep. Purif. Technol.* **2020**, *248*, 117046. [CrossRef]
10. Ignacz, G.; Szekely, G. Deep learning meets quantitative structure–activity relationship (QSAR) for leveraging structure-based prediction of solute rejection in organic solvent nanofiltration. *J. Membr. Sci.* **2022**, *646*, 120268. [CrossRef]
11. Xu, Q.; Gao, J.; Feng, F.; Chung, T.S.; Jiang, J. Synergizing machine learning, molecular simulation and experiment to develop polymer membranes for solvent recovery. *J. Membr. Sci.* **2023**, *678*, 121678. [CrossRef]
12. Gallo-Molina, J.P.; Claessens, B.; Buekenhoudt, A.; Verliefde, A.; Nopens, I. Capturing unmodelled phenomena: A hybrid approach for the prediction of the transport through ceramic membranes in organic solvent nanofiltration. *J. Membr. Sci.* **2023**, *686*, 122024. [CrossRef]
13. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
14. Vandezande, P.; Gevers, L.; Vankelecom, I. Solvent resistant nanofiltration: Separating on a molecular level. *Chem. Soc. Rev.* **2008**, *37*, 365–405. [CrossRef] [PubMed]
15. Mason, E.; Lonsdale, H. Statistical-mechanical theory of membrane transport. *J. Membr. Sci.* **1990**, *51*, 1–81. [CrossRef]
16. Bird, R.; Stewart, W.; Lightfoot, E. *Transport Phenomena, Revised 2nd Edition*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
17. Taylor, R.; Krishna, R. *Multicomponent Mass Transfer*; John Wiley & Sons: Hoboken, NJ, USA, 1993.
18. Cussler, E. *Diffusion: Mass Transfer in Fluid Systems*; Cambridge University Press: Cambridge, UK, 2009.
19. Noordman, T.; Wesselingh, J. Transport of large molecules through membranes with narrow pores: The Maxwell-Stefan description combined with hydrodynamic theory. *J. Membr. Sci.* **2002**, *210*, 227–243. . [CrossRef]
20. Callen, H.B. *Thermodynamics and an Introduction to Thermostatistics*; John Wiley & Sons: Hoboken, NJ, USA, 1991.
21. Mehta, G.; Morse, T.; Mason, E.; Daneshpajooh, M. Generalized Nernst–Planck and Stefan–Maxwell equations for membrane transport. *J. Chem. Phys.* **1976**, *64*, 3917–3923. [CrossRef]
22. Bowen, W.R.; Welfoot, J.S. Modelling the performance of membrane nanofiltration—Critical assessment and model development. *Chem. Eng. Sci.* **2002**, *57*, 1121–1137. [CrossRef]
23. Bye, K.P.; Galizia, M. Fundamental origin of flux non-linearity in organic solvent nanofiltration: Formulation of a thermodynamic/diffusion framework. *J. Membr. Sci.* **2020**, *603*, 118020. [CrossRef]
24. Shi, B.; Peshev, D.; Marchetti, P.; Zhang, S.; Livingston, A.G. Multi-scale modelling of OSN batch concentration with spiral-wound membrane modules using OSN Designer. *Chem. Eng. Res. Des.* **2016**, *109*, 385–396. [CrossRef]
25. Wright, P. Remarks on the Stefan-Maxwell equations for diffusion in a dusty gas. *J. Chem. Soc. Faraday Trans. 2* **1972**, *68*, 1951–1954. [CrossRef]
26. Mason, E.; Viehland, L.A. Statistical–mechanical theory of membrane transport for multicomponent systems: Passive transport through open membranes. *J. Chem. Phys.* **1978**, *68*, 3562–3573. [CrossRef]
27. Wijmans, J.G.; Baker, R.W. The solution-diffusion model: A review. *J. Membr. Sci.* **1995**, *107*, 1–21. [CrossRef]
28. Paul, D.R. Reformulation of the solution-diffusion theory of reverse osmosis. *J. Membr. Sci.* **2004**, *241*, 371–386. [CrossRef]
29. Kedem, O.; Katchalsky, A. Thermodynamic analysis of the permeability of biological membranes to non-electrolytes. *Biochim. Biophys. Acta* **1958**, *27*, 229–246. [CrossRef]
30. Spiegler, K.; Kedem, O. Thermodynamics of hyperfiltration (reverse osmosis): Criteria for efficient membranes. *Desalination* **1966**, *1*, 311–326. [CrossRef]
31. Mulder, M. *Basic Principles of Membrane Technology*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1996.
32. Bowen, W.R.; Mukhtar, H. Characterisation and prediction of separation performance of nanofiltration membranes. *J. Membr. Sci.* **1996**, *112*, 263–274. [CrossRef]
33. Matsuura, T.; Sourirajan, S. Reverse osmosis transport through capillary pores under the influence of surface forces. *Ind. Eng. Chem. Process. Des. Dev.* **1981**, *20*, 273–282. [CrossRef]
34. Niemi, H.; Palosaari, S. Flowsheet simulation of ultrafiltration and reverse osmosis processes. *J. Membr. Sci.* **1994**, *91*, 111–124. [CrossRef]
35. Lonsdale, H.; Merten, U.; Riley, R. Transport properties of cellulose acetate osmotic membranes. *J. Appl. Polym. Sci.* **1965**, *9*, 1341–1362. [CrossRef]
36. Bhanushali, D.; Kloos, S.; Bhattacharyya, D. Solute transport in solvent-resistant nanofiltration membranes for non-aqueous systems: Experimental results and the role of solute–solvent coupling. *J. Membr. Sci.* **2002**, *208*, 343–359. [CrossRef]
37. Stafie, N.; Stamatialis, D.; Wessling, M. Insight into the transport of hexane–solute systems through tailor-made composite membranes. *J. Membr. Sci.* **2004**, *228*, 103–116. [CrossRef]
38. Chollet, F. *Deep Learning with Python*; Simon and Schuster: New York, NY, USA, 2021.

39. Le Phuong, H.A.; Blanford, C.F.; Szekely, G. Reporting the unreported: The reliability and comparability of the literature on organic solvent nanofiltration. *Green Chem.* **2020**, *22*, 3397–3409. [CrossRef]
40. Ignacz, G.; Alqadhi, N.; Szekely, G. Explainable machine learning for unraveling solvent effects in polyimide organic solvent nanofiltration membranes. *Adv. Membr.* **2023**, *3*, 100061. [CrossRef]
41. Ignacz, G.; Yang, C.; Szekely, G. Diversity matters: Widening the chemical space in organic solvent nanofiltration. *J. Membr. Sci.* **2022**, *641*, 119929. [CrossRef]
42. Song, K.; Li, G.; Zu, X.; Du, Z.; Liu, L.; Hu, Z. The fabrication and application mechanism of microfluidic systems for high throughput biomedical screening: A review. *Micromachines* **2020**, *11*, 297. [CrossRef] [PubMed]
43. Vandezande, P.; Gevers, L.E.; Weyens, N.; Vankelecom, I.F. Compositional optimization of polyimide-based SEPPI membranes using a genetic algorithm and high-throughput techniques. *J. Comb. Chem.* **2009**, *11*, 243–251. [CrossRef]
44. Cano-Odena, A.; Spilliers, M.; Dedroog, T.; De Grave, K.; Ramon, J.; Vankelecom, I. Optimization of cellulose acetate nanofiltration membranes for micropollutant removal via genetic algorithms and high throughput experimentation. *J. Membr. Sci.* **2011**, *366*, 25–32. [CrossRef]
45. Kim, C.; You, C.; Ngan, D.T.; Park, M.; Jang, D.; Lee, S.; Kim, J. Machine learning-based approach to identify the optimal design and operation condition of organic solvent nanofiltration (OSN). *Comput. Aided Chem. Eng.* **2021**. *50*, 933–938. [CrossRef]
46. Marchetti, P.; Livingston, A.G. Predictive membrane transport models for Organic Solvent Nanofiltration: How complex do we need to be? *J. Membr. Sci.* **2015**, *476*, 530–553. [CrossRef]
47. Ignacz, G.; Beke, A.K.; Szekely, G. Data-driven future for nanofiltration: Escaping linearity. *J. Membr. Sci. Lett.* **2023**, *3*, 100040. [CrossRef]
48. Yangali-Quintanilla, V.; Verliefde, A.; Kim, T.U.; Sadmani, A.; Kennedy, M.; Amy, G. Artificial neural network models based on QSAR for predicting rejection of neutral organic compounds by polyamide nanofiltration and reverse osmosis membranes. *J. Membr. Sci.* **2009**, *342*, 251–262. [CrossRef]
49. Zhang, Z.; Luo, Y.; Peng, H.; Chen, Y.; Liao, R.Z.; Zhao, Q. Deep spatial representation learning of polyamide nanofiltration membranes. *J. Membr. Sci.* **2021**, *620*, 118910. [CrossRef]
50. Galinha, C.F.; Crespo, J.G. From black box to machine learning: A journey through membrane process modelling. *Membranes* **2021**, *11*, 574. [CrossRef] [PubMed]
51. Galinha, C.F.; Guglielmi, G.; Carvalho, G.; Portugal, C.A.; Crespo, J.G.; Reis, M.A. Development of a hybrid model strategy for monitoring membrane bioreactors. *J. Biotechnol.* **2013**, *164*, 386–395. [CrossRef] [PubMed]
52. Sá, M.; Monte, J.; Brazinha, C.; Galinha, C.F.; Crespo, J.G. Fluorescence coupled with chemometrics for simultaneous monitoring of cell concentration, cell viability and medium nitrate during production of carotenoid-rich Dunaliella salina. *Algal Res.* **2019**, *44*, 101720. [CrossRef]
53. Teodosiu, C.; Pastravanu, O.; Macoveanu, M. Neural network models for ultrafiltration and backwashing. *Water Res.* **2000**, *34*, 4371–4380. [CrossRef]
54. Bowen, W.R.; Jones, M.G.; Welfoot, J.S.; Yousef, H.N. Predicting salt rejections at nanofiltration membranes using artificial neural networks. *Desalination* **2000**, *129*, 147–162. [CrossRef]
55. Sanches, S.; Galinha, C.; Crespo, M.B.; Pereira, V.; Crespo, J. Assessment of phenomena underlying the removal of micropollutants during water treatment by nanofiltration using multivariate statistical analysis. *Sep. Purif. Technol.* **2013**, *118*, 377–386. [CrossRef]
56. Barello, M.; Manca, D.; Patel, R.; Mujtaba, I.M. Neural network based correlation for estimating water permeability constant in RO desalination process under fouling. *Desalination* **2014**, *345*, 101–111. [CrossRef]
57. Yeo, C.S.H.; Xie, Q.; Wang, X.; Zhang, S. Understanding and optimization of thin film nanocomposite membranes for reverse osmosis with machine learning. *J. Membr. Sci.* **2020**, *606*, 118135. [CrossRef]
58. Fetanat, M.; Keshtiara, M.; Keyikoglu, R.; Khataee, A.; Daiyan, R.; Razmjou, A. Machine learning for design of thin-film nanocomposite membranes. *Sep. Purif. Technol.* **2021**, *270*, 118383. [CrossRef]
59. Tan, M.; He, G.; Li, X.; Liu, Y.; Dong, C.; Feng, J. Prediction of the effects of preparation conditions on pervaporation performances of polydimethylsiloxane (PDMS)/ceramic composite membranes by backpropagation neural network and genetic algorithm. *Sep. Purif. Technol.* **2012**, *89*, 142–146. [CrossRef]
60. Dudchenko, A.V.; Mauter, M.S. Neural networks for estimating physical parameters in membrane distillation. *J. Membr. Sci.* **2020**, *610*, 118285. [CrossRef]
61. Kadel, S.; Daigle, G.; Thibodeau, J.; Perreault, V.; Pellerin, G.; Lainé, C.; Bazinet, L. How physicochemical properties of filtration membranes impact peptide migration and selectivity during electrodialysis with filtration membranes: Development of predictive statistical models and understanding of mechanisms involved. *J. Membr. Sci.* **2021**, *619*, 118175. [CrossRef]
62. Santos, J.; Hidalgo, A.; Oliveira, R.; Velizarov, S.; Crespo, J. Analysis of solvent flux through nanofiltration membranes by mechanistic, chemometric and hybrid modelling. *J. Membr. Sci.* **2007**, *300*, 191–204. [CrossRef]
63. Karan, S.; Jiang, Z.; Livingston, A.G. Sub–10 nm polyamide nanofilms with ultrafast solvent transport for molecular separation. *Science* **2015**, *348*, 1347–1351. [CrossRef] [PubMed]
64. Moriwaki, H.; Tian, Y.S.; Kawashita, N.; Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4. [CrossRef]

65. Wang, C.; Wang, L.; Soo, A.; Pathak, N.B.; Shon, H.K. Machine learning based prediction and optimization of thin film nanocomposite membranes for organic solvent nanofiltration. *Sep. Purif. Technol.* **2023**, *304*, 122328. [CrossRef]
66. Thiermeyer, Y.; Blumenschein, S.; Skiborowski, M. Fundamental insights into the rejection behavior of polyimide-based OSN membranes. *Sep. Purif. Technol.* **2021**, *265*, 118492. [CrossRef]