Machine learning for determination of activity of water and activity coefficients of electrolytes in binary solutions

Guillaume Zante

Artificial Intelligence Chemistry 2 (2024) 100069

Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

Machine learning for determination of activity of water and activity coefficients of electrolytes in binary solutions

Guillaume Zante

Université Paris-Saclay, CEA, CNRS, NIMBE, LICSEN, Gif-sur-Yvette 91191, France

ARTICLE INFO

Keywords:
Activity coefficients
Electrolytes
Neural networks
Machine learning

ABSTRACT

Activity of water and electrolytes in aqueous solutions is of utmost importance for multiple industrial applications. However, experimental determination of such values is time-consuming, while calculation of activity coefficients using numerical methods is challenging. By training neural networks models on literature data, one could predict activity of water and electrolytes easily, without requiring any experiment. In this paper, multiple descriptors (or features) are compared to predict activity coefficients of electrolytes and activity of water in electrolyte solutions. A neural network based on the Levenberg-Marquardt algorithm (LM-NN) showed high accuracy to calculate values, despite the small size of the training datasets. Both activity coefficients of electrolytes and activity of water in electrolyte solutions can be predicted accurately even on unseen data, using simple descriptors such as electrolyte concentration, ion sizes and charges. However, some discrepancies were observed due to the lack of representativeness of the training dataset. This could be solved by selecting training data sets that are similar (*e.g.* same group of the periodic table) to the unknown values, or by including available experimental data for the salt considered. The ability of the LM-NN to solve non-linear least square curve fitting problems makes it a good candidate to fit experimental activity coefficient data, with the advantage of simplicity as compared to e-NRTL or UNIQUAC methods. This method paves the way for accurate and quick determination of thermodynamic data for electrolyte solutions (and beyond) using machine learning, without necessitating large training datasets.

1. Introduction

Activity coefficients measure the deviation from ideal behaviour of electrolyte solutions and are obviously highly dependent on the nature and concentration of electrolytes dissolved in the aqueous solution considered [1]. Activity of water ($a(H_2O)$), Eq. (1)) in such solutions is defined as the partial vapour pressure of water in equilibrium with the solution (P), divided by the partial vapour pressure of pure water at the same temperature (P_0).

$$a(H_2O) = \frac{P}{P_0} = \gamma_{H_2O} x_{H_2O} \quad (1)$$

where $\gamma_{H_2O}$is the activity coefficient of water and x_{H_2O} its molar fraction in the electrolyte solution. Similarly, activity of electrolyte (i) in the solution is expressed as the product of its activity coefficient ($\gamma_{electrolyte}$) with its molar fraction.

Deviation from ideal behaviour of aqueous solutions plays a major role in the food industry and beyond since essential biological processes such as micro-organism development are highly dependent on the activity of water [2]. For many other chemical processes such as desalination ([3]), crystallisation ([4]) or solvent extraction ([5]), accurate determination of water and salt activities is extremely important. Determining the activity of electrolytes in complex solutions containing multiple electrolytes is feasible using the mixing rules such as the Zdanovskii rule ([6]), which states that mixing multiple binary solutions (one salt and water) of electrolytes leads to a mixture with the same water activity than the one of the initial solutions, as long as there are no chemical interactions between the electrolytes. Application of such mixing rules require access to accurate and robust values for activity coefficients in binary solutions.

Different experimental techniques are available to measure $a(H_2O)$ and $\gamma_{electrolyte}$ in electrolyte solutions such as ion selective electrodes ([7]), hygrometric ([8]) and transpiration methods ([9]), freezing point and vapour pressure measurements ([10]), infrared spectroscopy coupled with microfluidics ([11,12]), etc. However, determining values experimentally is a cumbersome task especially for the purpose of

E-mail address: guillaume.zante@cea.fr.

https://doi.org/10.1016/j.aichem.2024.100069
Received 3 March 2024; Received in revised form 23 April 2024; Accepted 26 April 2024
Available online 27 April 2024

process modelling, which necessitates a large number of experimental data points to model the process appropriately.

Numerical and computational methods have been developed over the years to calculate the activity coefficients, reducing the experimental work required. The Debye-Hückel equation ([13], Eq. (2)), which solely considers interactions between ions in solution, is therefore only valid for dilute solutions (up to 0.1 mol/kg for modified versions):

$$\log(\gamma_i) = -A\frac{z_i\sqrt{I}}{1+Ba_i\sqrt{I}} \tag{2}$$

where γ_i is the activity coefficient of component (i) in the mixture, z_i its charge, I the ionic strength while a_i, A and B are temperature and mixture-dependent constants [14,15]. Extended versions of the Debye-Hückel equation, such as the one taking into account the variation of the dielectric constant of the electrolyte with its concentration [16,17], can expand the range of application of this equation to calculate activity coefficients in more concentrated electrolyte solutions.

The Pitzer equation [18] is widely used to calculate the mean activity coefficient of an electrolyte in water (Eq. (3)):

$$\ln(\gamma_{electrolyte}) = \left|z_i z_j\right| f^{\gamma} + \frac{2\nu_i\nu_j}{\nu_i+\nu_j} mB' + 2\frac{(\nu_i\nu_j)^{\frac{3}{2}}}{\nu_i+\nu_j} mC \tag{3}$$

where z_i and z_j are the charge of ions i and j, ν_i and ν_j being the stoichiometric coefficients of the ions constituting the electrolyte, f^{γ} is the long-range interaction term, m is the molal concentration of the electrolyte, while C and B' are ionic interaction parameters for which tabulated data can be found for numerous electrolytes. The expression of B' depends on the nature of the electrolyte and contains more parameters for cations with higher valencies. For 1:1 electrolytes, it can be expressed as follows (Eq. (4)):

$$B' = \beta^{(0)} + \frac{2\beta^{(1)}}{\alpha^2 I}\left\{1-(1+\alpha\sqrt{I})e^{-\alpha\sqrt{I}}\right\} \tag{4}$$

where α is a constant depending on the nature of the electrolyte. The Pitzer parameters $\beta^{(0)}$, $\beta^{(1)}$ and C can be determined for each electrolyte using unweighted least square fit of the experimental data [19], then used to calculate activity coefficients. Both the extended Debye-Hückel equation and the extended Pitzer equations depend on numerous parameters that are of empirical nature, while calculation of activity coefficients of highly concentrated electrolyte solutions (typically > 6 mol/kg) is challenging with both equations [20].

Widely adopted models such as electrolyte Non Random Two Liquid (eNRTL) and extended Universal Quasi Chemical (eUNIQUAC) have been extensively used for prediction of activity coefficient for processes such as solvent extraction [21,22]. The eNRTL model is based on the local composition theory and the two-liquid solution theory [23]. Within this model, activity coefficient can be expressed as a function of the molar fraction of components, the energy of interaction between them and a non-randomness factor. However, the physical meaning of the non-randomness factor is unclear, and this factor is often set a priori [24]. Interaction energies are obtained using a set of unique adjustable parameters (the salt/solvent parameter $\tau_{i,j}$ and the solvent salt parameter $\tau_{j,i}$) that are of empirical nature and are obtained by fitting experimental data. Those parameters are determined using experimentally measured activity coefficients ([25]), by minimising a least square objective function such as $F(\tau_{i,j}, \tau_{j,i})$ described in Eq. (5):

$$F\left(\tau_{i,j}, \tau_{j,i}\right) = \sum_i \left[\ln\left(\gamma^{\exp}\right)_i - \ln\left(\gamma^{calc}\right)\right]^2 \tag{5}$$

where $\gamma^{\exp}$ and γ^{calc} are the experimental and calculated activity coefficients, respectively. Deduction of adjustable parameters is not straightforward as local minima may exist, minimisation of the objective function could give highly initialization-dependent results.

The e-UNIQUAC model gives access to activity coefficients by summing its Debye-Hückel contribution (defined in Eq. (2)), combinatorial contribution and a residual contribution [26]. The last two contributions can be expressed as a function of the mole fraction (x), the volume (ϕ) and surface area (θ) fraction, the volume parameter (q) of components i and j, and the interaction parameters. Hence, surface area and volume parameters for each species composing the mixture must be known, while interaction parameters are required for each couple of species in the solution. Similarly to the eNRTL model, the interaction parameters are obtained by minimising an object function ([27]) such as F', described in Eq. (6):

$$F' = \sum_i \left[w_i\left(\gamma^{calc} - \gamma^{\exp}\right)_i\right]^2 + \sum_i \left[w_i \ln(SI)_i\right]^2 \tag{6}$$

where w_i are weighing factors which are set a priori, while SI is the solubility index of salt i (activity product of a salt divided by its solubility product), obtained from solid-liquid equilibrium data.

Conductor like screening model for real solvents (COSMO-RS) is based on quantum chemistry calculations and is used in combination with Pitzer/Debye-Hückel equations to determine activity coefficients of electrolytes solutions (COSMO-RS takes into account the short range interaction while the Pitzer/Debye-Hückel contribution takes into account the long range interactions, [28]. Hence, available models always necessitate access to numerous parameters, are often difficult to fit [29], and are sometimes based on specific softwares, which could require significant computing time and quantum chemistry calculations, such as COSMO-RS.

Machine learning (ML) approaches can streamline activity coefficient calculation since neural networks, support vector machines or other ML tools can learn from input data selected from literature and predict activity values with high accuracy [30]. So far, ML methods were mostly used for the determination of infinite dilution activity coefficients of various solutes in multiple solvents. Those methods include neural networks ([31]), matrix completion ([32]), natural language processing ([33]), etc. For most of these methods, a large amount of experimental data points need to be available to train the model (generating data points with COSMO-RS is sometimes required to increase the size of the database), making these methods computationally intensive. For electrolyte solutions, neural networks were used to calculate activity coefficients using parameters determined with the hard sphere equation of state (diameter and density of pure hard sphere of salts) as input data [34]. The activity coefficient of ions within ion exchange membranes were determined accurately by combining machine learning and molecular dynamics [35]. Evaluation of the performance of ML methods for the calculation of activity coefficients on a large panel of electrolytes is still lacking. Determination of the rules for selection of appropriate descriptors (or features) and ML models are needed, as well as a critical evaluation of the performances of the model and its generalisation ability.

In this paper, we investigate the possibility of determining activity of water and electrolytes in binary solutions using ML methods (neural networks) trained on small data sets. We describe a simple neural network trained on a few (<300) data points, which can take easily accessible data as input (ion size, charge…) to accurately predict activity coefficients (less than 5% deviation from experimental data in most cases).

2. Materials and methods

2.1. Neural networks used

Two different neural networks are used in the first section of this study. The first one (further referred to as NN), is coded in Python (version 3.11.5), using the Scikit-learn library and the keras module. Its architecture, similar to the one reported by Gbashi *et al.*, is made of five hidden layers composed of 256, 128, 64, 32 and 16 neurons [36]. A

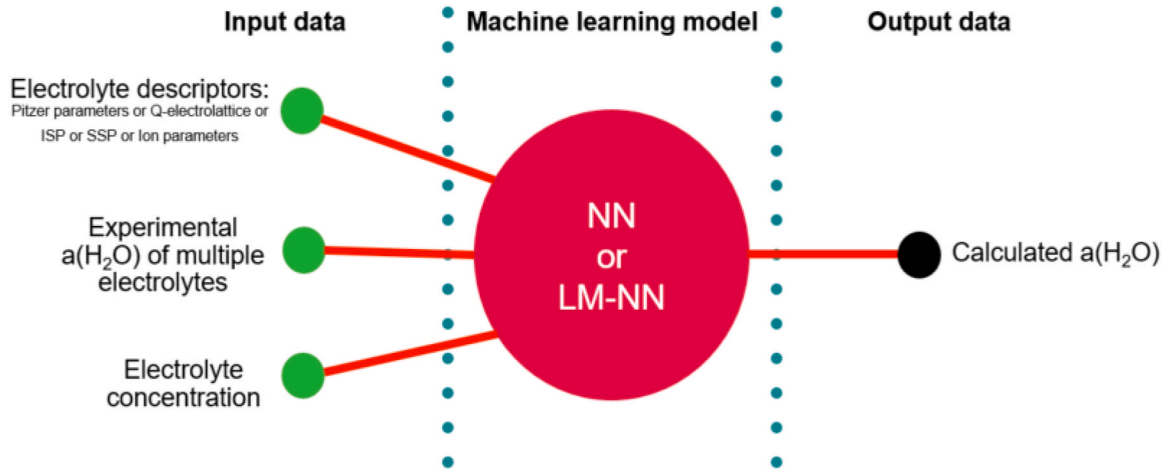

Fig. 1. Description of the ML-based approach used for the determination of $a(H_2O)$ in electrolyte solution.

dropout layer is added between the first and second layer and between the third and fourth layer of neurons to reduce risks of over-fitting. The learning rate, dropout and activation functions of hidden layers 1, 3 and output layers are the hyperparameters that are optimised (activation functions are selected amongst softmax, linear, sigmoid, Leakyrelu and relu, the last one being used for the other layers). The model includes a loss function (mean squared error), and aims to minimise it. The results presented for each input dataset selected are based on the architecture and hyperparameters that gave the most accurate results *i.e.* the architecture and hyperparameters that gave the lowest average absolute relative deviation (AARD) and root mean squared error (RMSE).

The second neural network is based on the Levenberg-Marquadt algorithm (further reffered to as LM-NN), with an architecture similar to the one reported by Maamar et al. [30]. It was coded in Python (version 3.11.5) using the pyrenn module. To reduce risks of over-fitting, it is composed of a single hidden layer containing either 5, 7, 9, 16, 32, 64 or 128 neurons (this parameter being optimised), while the hyperbolic tangent (tanh) is used as an activation function for all layers. The LM-NN takes advantage of the Levenberg-Marquardt algorithm which locates the minimum of a multivariate function that is expressed as the sum of squares of non-linear real-valued functions ([37,38]). This kind of neural network combines the advantages of gradient descent and Gauss-Newton methods, converges rapidly, making it particularly indicated for small datasets [39]. For both models, data are scaled between −1 and 1 before training, the input layer contains the same number of neurons as the number of descriptors and the output layer contains a single neuron. Both models are trained on 100 epochs with 70% of the values included in the training dataset, the remaining 30% being used for testing. Tables are provided to describe the databases used for each calculation (which are also described in the supplementary information file). The database includes activity of water and activity of electrolytes at 298.15 K and 1 atm. The method reported here is therefore limited to activities of electrolyte solutions at ambient temperature and pressure, but the method reported here can be adapted to other conditions (as long as sufficient experimental data exist), by adapting the database and including temperature and pressure as features. Characteristics of the data sets used are given in Table S1 and Table S2, while the optimised parameters retained for the regular neural network and the LM-NN are displayed in Table S3 and Table S4, respectively. Several descriptors have been used as input data to describe different electrolytes (Pitzer parameters, cation and anion size and charge, etc), and their effect on the accuracy of the ML model compared.

2.2. Accuracy estimation

The accuracy of the two NN investigated is estimated with the AARD, defined as (Eq. (7)):

$$AARD(\%) = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_i^{\exp} - y_i^{calc}}{y_i^{\exp}} \right| \quad (7)$$

where N is the number of data points, $y_i^{\exp}$ and y_i^{calc} are the experimental and calculated values of either the activity of water or activity of the electrolyte. The RMSE is calculated with Eq. (8):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\exp} - y_i^{calc}\right)^2} \quad (8)$$

Finally, the plot of the calculated values as a function of the experimentally determined ones gives access to the correlation coefficient (R^2) which allows estimating the correlation between those values in the testing data set. The R^2 is an indicator of the performance of the model, indicating if the calculated and actual values match. AARD and RMSE are determined to characterise further the deviation of the model.

3. Results and discussion

3.1. Determination of activity of water in electrolyte solutions

3.1.1. Descriptors and ML model selection

The general approach adopted in this section is presented in Fig. 1. In order to train the two neural networks, experimental values of $a(H_2O)$ in electrolyte solutions were regrouped from multiple literature sources. The database used contained $a(H_2O)$ values at various salt molalities for different salts regrouped from the literature, most of them being alkaline and alkaline-earth cations associated with chloride ([40]), sulfate ([19]), or nitrate anions ([8], detailed presentation of the input datasets and literature sources used can be found in Table S1 and Table S2). The nature and concentration ranges of electrolytes used in the database for Fig. 2 and Fig. 3 are displayed in Table 1.

Various models are available in the literature to determine the activity of water. Within these models, descriptors having a direct or indirect effect on the activity of water can be found and have been collected for each salt. In the Pitzer model (briefly described in the introduction section, Eq. (3) and Eq. **(4)**), the Pitzer parameters $\beta^{(0)}$, $\beta^{(1)}$ and C can be used as descriptors to train the ML model. The Q-electrolattice model can be used to calculate activity of water in electrolyte solutions with adjustable parameters [41]. From this model, selected descriptors include interaction energies between the solvent and the cation ($\frac{u_0^{solvent-cation}}{R}$), and interaction energies between the solvent and the anion ($\frac{u_0^{solvent-anion}}{R}$), [42]. Similarly, different values for the solvent-cation and solvent-anion or solvent-salt interaction energies can be found in

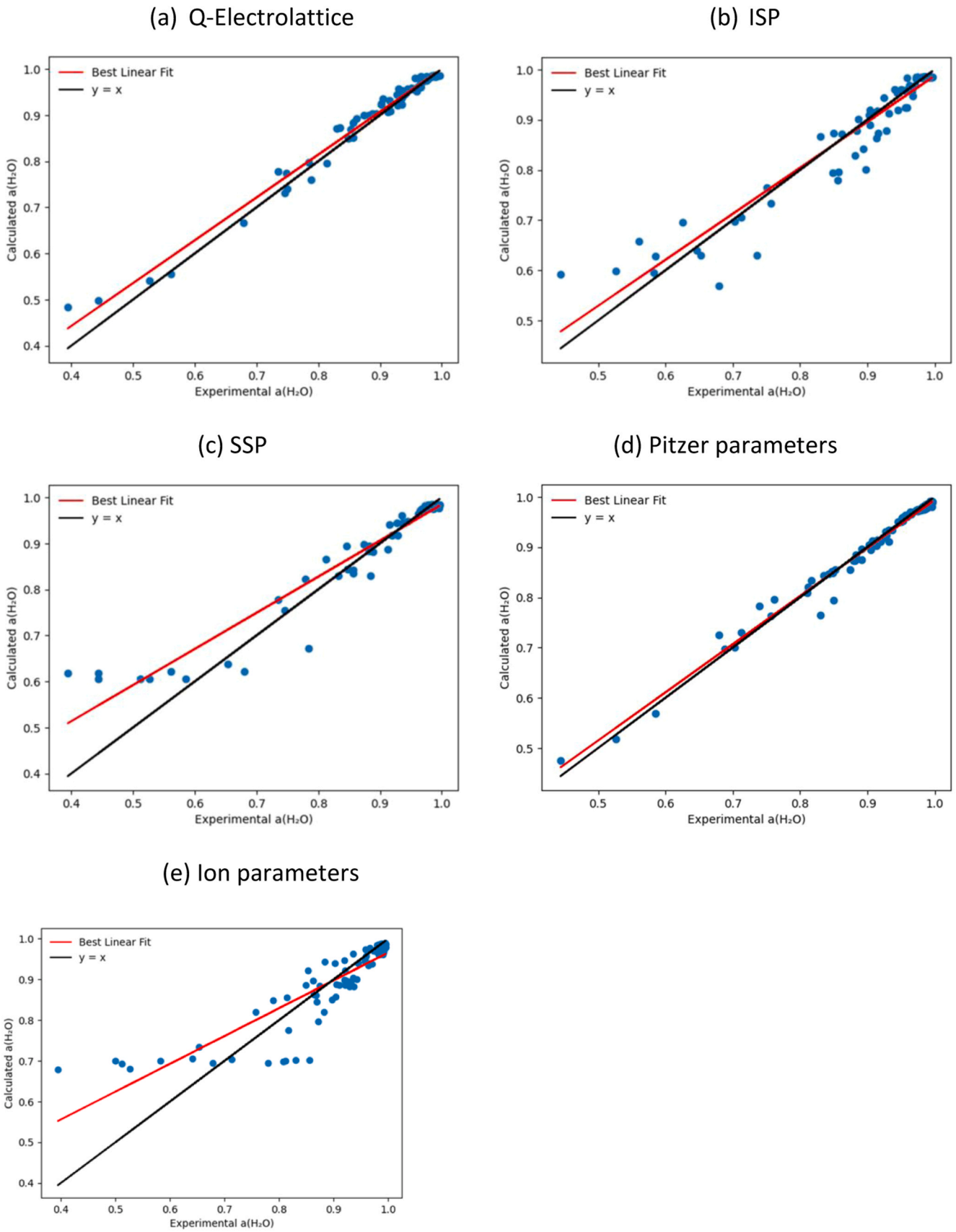

Fig. 2. Calculated and experimental activity coefficients of water in electrolyte solutions (at room temperature and pressure) using a neural network with different input datasets. (a) Q-electrolattice; (b) Ion specific parameters; (c) Salt specific parameters; (d) Pitzer parameters; (e) Ionic diameter and ion geometric parameters.

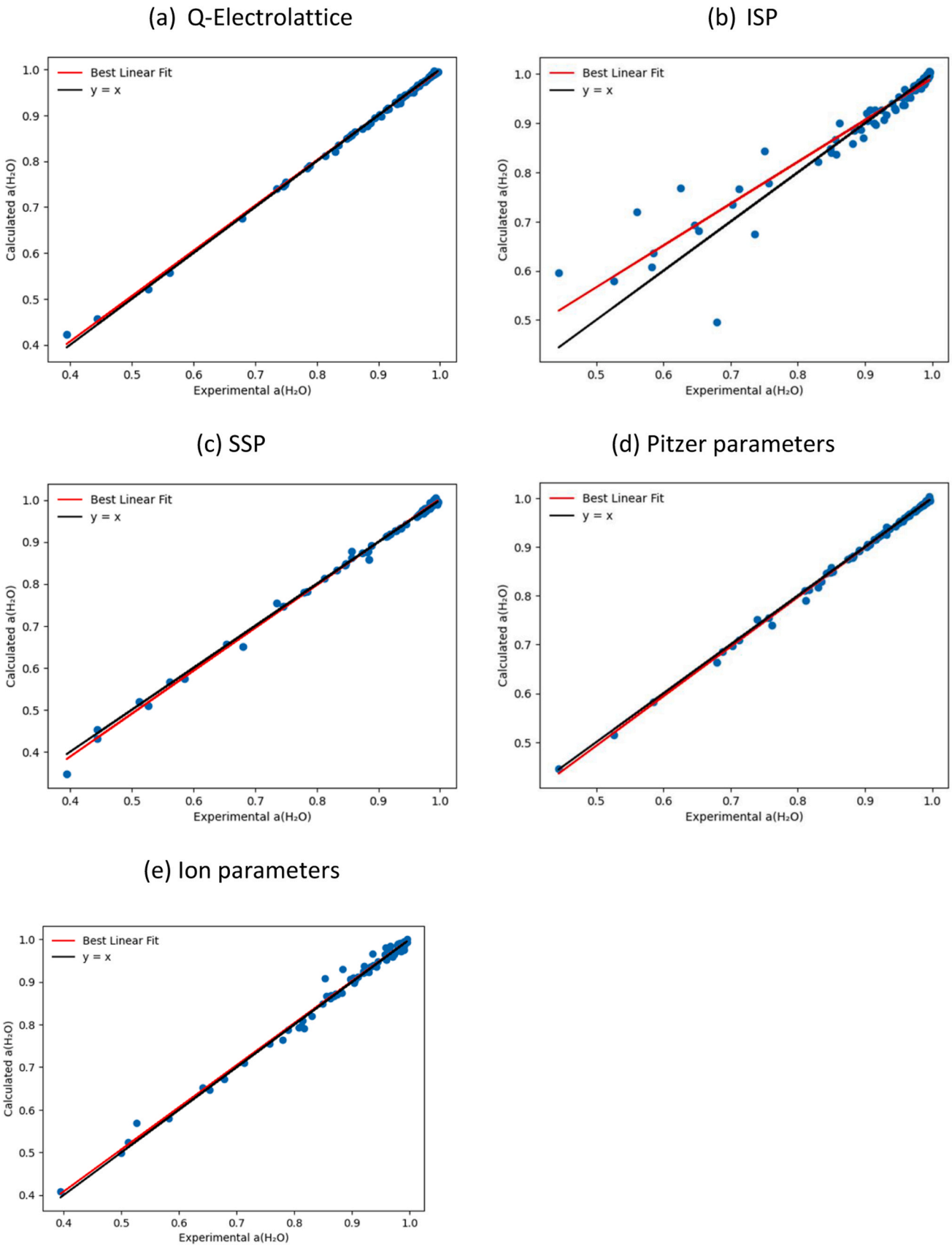

Fig. 3. Calculated and experimental activity coefficients of water in electrolyte solutions (at room temperature and pressure) using a LM neural network with different input datasets. (a) Q-electrolattice; (b) Ion specific parameters; (c) Salt specific parameters; (d) Pitzer parameters; (e) Ionic diameter and ion geometric parameters.

Table 1
Characteristics of the input datasets used in Fig. 1 and Fig. 2.

Nature of electrolytes (concentration range in mol/kg)	Reference descriptors	Reference activity values	Figure
Chlorides: H, Li, Na, K, Cs, Mg, Ca, Ba (0.2–6) Sulfates: Li, Na, K, Mg (0.1–3) Nitrates: Li, Na, K, Mg, Ca, Ba (0.1–6)	[41]	[8,19,40]	Fig. 2(a) and 3 (a)
Chlorides: H, Li, Na, K, NH_4, Cs, Mg, Ca, Ba (0.2–6) Sulfates: Li, Na, K, NH_4, Mg, Mn, Ni, Cu, Zn (0.1–5) Nitrates: Li, Na, K, NH_4, Mg, Ca, Ba (0.1–6)	[8,19,40]	[8,19,40]	Fig. 2(d), and 3 (d), 5, and 6
Chlorides: Li, Na, K, NH_4, Cs, Mg, Ca, Ba (0.2–6) Sulfates: Li, NH_4 (0.1–5) Nitrates: Li, Na, K, NH_4, Mg, Ca, Ba (0.1–6)	[41]	[8,19,40,41]	Fig. 2(c) and 3 (c)
Chlorides: Li, Na, K, NH_4, Cs, Mg, Ca, Ba (0.2–6) Sulfates: Li, Na, K, NH_4, Mg, Mn, Ni, Cu (0.1–5) Nitrates: Li, Na, Mg, Ca (0.1–6)	[41]	[8,19,40]	Fig. 2(e), 3 (e)

Table 2
AARD, R^2 and RMSE (during testing) obtained with the NN and the LM-NN trained on various input datasets.

Metric	Q-electrolattice	Pitzer parameters	SSP	Ion parameters	ISP
Optimised NN					
AARD	1.79	1.01	4.53	5.42	2.81
R^2	0.976	0.983	0.897	0.763	0.922
RMSE	0.0196	0.0136	0.0510	0.0616	0.0355
Optimised LM-NN					
AARD	0.356	0.286	0.867	0.882	2.70
R^2	0.999	0.998	0.996	0.991	0.905
RMSE	0.00456	0.00460	0.0102	0.0121	0.0392

the literature, denominated as ion specific parameters (ISP, [41,43] or salt specific parameters (SSP, [41,43]. The Q-electrolattice equations of state also allow to estimate the ion parameters (ion diameter and geometric diameter, [41,44] for cations and anions composing the salts, which were also used as descriptors. One should note that the ML models used are not "physics-informed". The descriptors are used as features in the input database and may have an effect on the accuracy of the calculated values, but they are not included in the ML model.

These different input data were used to predict values for water activities in various electrolyte solutions using the NN with five hidden layers and the LM-NN with a single hidden layer (described in Section 2.1). The plot of the predicted activities of water as a function of the experimentally determined values are given in Fig. 2 for the NN and in Fig. 3 for the LM-NN. Each graph corresponds to a different input dataset used to train the model. The best linear fit is plotted on each graph and can be seen as a red line while the y=x plot added as a black line. For the NN, major discrepancies are observed between the calculated and experimental values, in particular when ISP, SSP or ion parameters are used as descriptors. The calculated values using the Q-electrolattice or Pitzer parameters as descriptors seems to be in accordance with the experimental data.

For the LM-NN, large deviations between calculated and experimental values are observed when ISP are used as descriptors. The highest deviation between experimental and calculated values can be observed on the lowest $a(H_2O)$ values, *i.e.* when the deviation of the electrolyte solution from ideal behaviour is the highest. This discrepancy possibly arises from the lack of training data at high deviation from ideal behaviour (encountered at high electrolyte concentration). Overall, performance of the LM-NN seems to be better than the NN, with better accordance between the calculated and experimental values whatever the number of data points, nature and number of descriptors in the training and testing dataset.

These observations are confirmed on Table 2, which shows accuracy of predictions obtained with the NN and LM-NN. AARD, RMSE and R^2 are displayed for the NN and for the LM-NN using the various descriptors tested. It can clearly be seen that the performances of the LM-NN are better than the NN; AARD is lower, R^2 values are closer to one and RMSE values are 4–5 times lower (except when ISP are used as descriptors). Using the LM-NN, AARD values as low as 0.29% can be obtained, while AARD values obtained with the NN are Higher than 1 when using the NN. For the LM-NN, the accuracy of predictions depends on the descriptors selected and increases in the order: ISP < Ion parameters < SSP < Q-electrolattice < Pitzer parameters.

With the NN, Q-electrolattice and Pitzer parameters input datasets also allow to obtain the best performances, although the AARD and RMSE values are higher and the correlation between calculated and experimental values are lower (lower R^2 values). The performance obtained does not seem to depend on the size of the dataset or number of descriptors available. The Q-electrolattice model allows obtaining accurate predictions despite the fact that it contains less data points than the ISP model (241 against 294, see Table S1 and Table S2). The correlation between the input and output data seems to be relevant but accurate predictions can be obtained without a direct correlation between the descriptors and the output data. The representative nature of the initial dataset seems however important since the lack of data at high deviation from ideal behaviour (low $a(H_2O)$) could lead to poorer accuracy of the predictions. Overall, the quickly converging LM-NN performs well on the small datasets used (<300 data points) and allows to obtain more accurate values than the NN. Therefore, further calculations will be performed with the LM-NN, while the Pitzer parameters are used

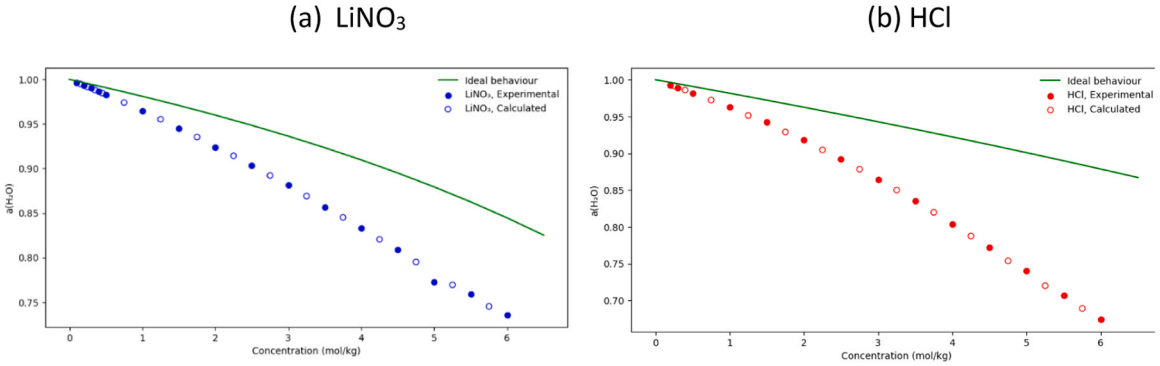

Fig. 4. Experimental and calculated activity of water (at room temperature and pressure) at various concentrations of (a) lithium nitrate and (b) hydrochloric acid.

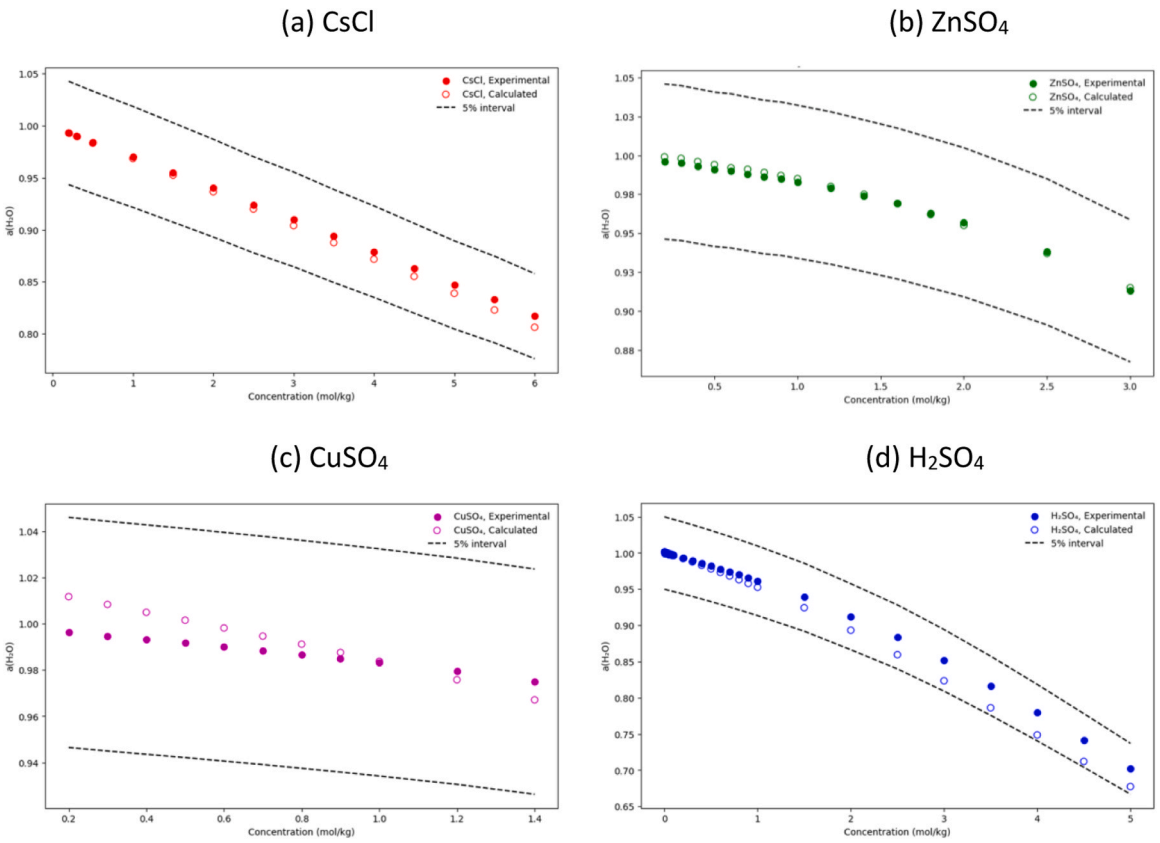

Fig. 5. Experimental and extrapolated activity of water (at room temperature and pressure) at various concentrations of (a) caesium chloride, (b) zinc sulfate, (c) copper sulfate and (d) sulfuric acid.

as input descriptors.

3.1.2. *Predictions and generalisation ability*

The LM-NN was therefore used to predict $a(H_2O)$ using Pitzer parameters of each electrolyte as descriptors. First, predictions were performed on electrolytes that were included in the dataset, *i.e.* $a(H_2O)$ values were predicted at electrolyte concentrations that were not included in the initial datasets (the approach is explained in Fig. S1). Experimental results (filled symbols) and calculated values (empty symbols) are shown in Fig. 4, which represents the activity of water as a function of electrolyte concentrations for lithium nitrate (Fig. 4(a)) and hydrochloric acid (Fig. 4(b)). As a matter of comparison, the ideal behaviour ($a(H_2O) = x(H_2O)$) is plotted as a green line for both electrolytes. The database used is the same as the one used in Fig. 2(d) and Fig. 3(d), described in Table 1.

It can be observed from this figure that very accurate predictions can be obtained for both electrolytes, with predicted values that can't be distinguished from the experimental ones. The predicted values reproduce the experimental behaviour observed, which translates the increased deviation from ideal behaviour when increasing electrolyte concentration. Hence, within the electrolyte concentration range for which experimental data are available, the LM-NN model can be used to "complete" a curve by predicting missing data points with very high accuracy, probing the ability of the LM-NN to interpolate data. The ability of the LM algorithm to solve least square fitting problems is therefore very useful when experimental data are available. It can compete with the current methods used for the same purpose (eNRTL, UNIQUAC), which require minimising a least square function for calculation of activity of water or activity coefficients of electrolytes. The LM-NN greatly simplifies such calculation, since the pyrenn module allows coding the NN in a few lines, while predicted results are obtained very quickly (100 epochs are used), minimising the computation time to

Table 3
Characteristics of the input datasets used in Fig. 5 and Fig. 7.

Nature of electrolytes (concentration range in mol/kg)	Reference descriptors	Reference activity values	Figure
Chlorides: H, Li, Na, K, NH_4, Mg, Ca, Ba (0.2–6) Sulfates: Li, Na, K, NH_4, Mg, Mn, Ni, Cu, Zn (0.1–5) Nitrates: Li, Na, K, NH_4, Mg, Ca, Ba (0.1–6)	[8,19,40]	[8,19,40]	Fig. 5 (a)
Chlorides: H, Li, Na, K, NH_4, Cs, Mg, Ca, Ba (0.2–6) Sulfates: Li, Na, K, NH_4, Mg, Mn, Ni, Cu (0.1–5) Nitrates: Li, Na, K, NH_4, Mg, Ca, Ba (0.1–6)	[8,19,40]	[8,19,40]	Fig. 5 (b)
Chlorides: H, Li, Na, K, NH_4, Cs, Mg, Ca, Ba (0.2–6) Sulfates: Li, Na, K, NH_4, Mg, Mn, Ni, Zn (0.1–5) Nitrates: Li, Na, K, NH_4, Mg, Ca, Ba (0.1–6)	[8,19,40]	[8,19,40]	Fig. 5 (c)
Chlorides: H, Li, Na, K, NH_4, Cs, Mg, Ca, Ba (0.2–6) Sulfates: Li, Na, K, NH_4, Mg, Mn, Ni, Cu, Zn (0.1–5) Nitrates: Li, Na, K, NH_4, Mg, Ca, Ba (0.1–6)	[8,19,40]	[8,19,40]	Fig. 5 (d)
Osmotic coefficient of: UO_2Cl_2 (0.1–5.5), $Th(NO_3)_4$ (0.1–5) Lanthanide nitrates: La, Nd, Eu (0.088–3.247)	[47]	[46,48]	Fig. 7

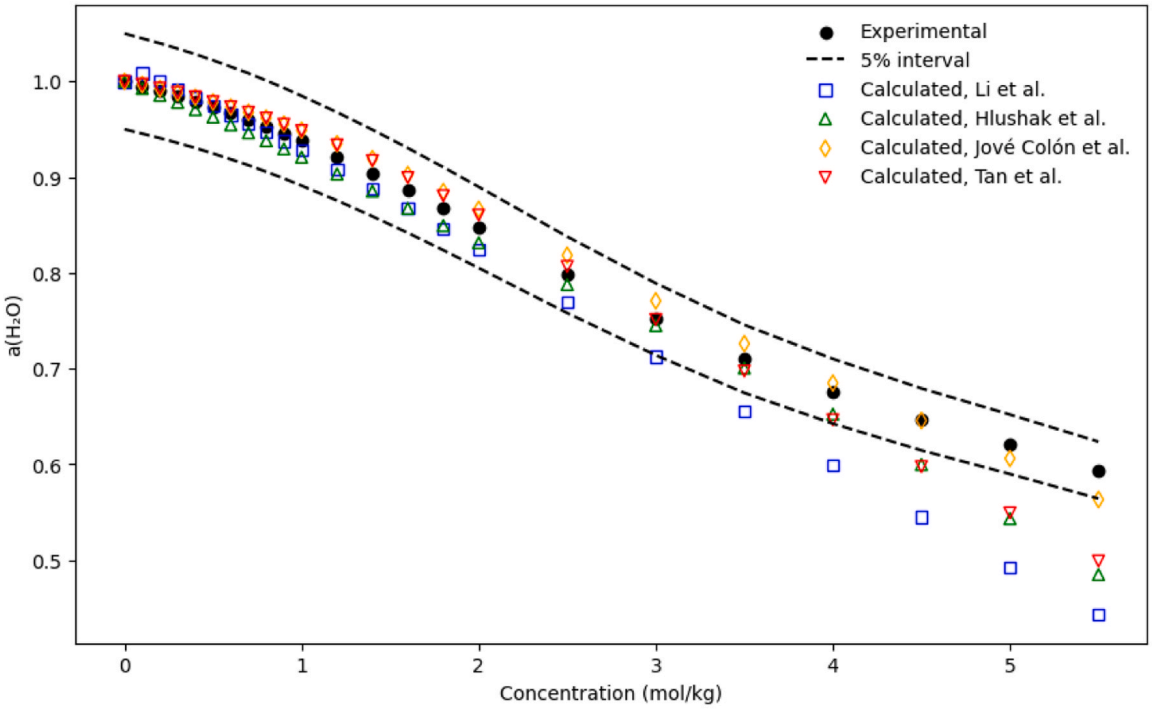

Fig. 6. Experimental and extrapolated activity of water (at room temperature and pressure) at various concentrations of uranyl nitrate using different Pitzer parameters reported by Li et al., Hlushak et al., Jové Colón et al. and Tan et al. The 5% interval for experimental data is shown as a dotted black line.

less than 30 seconds in most cases.

However, such an interpolation tool would be inherently limited to electrolytes for which experimental data are available. But the LM-NN is not limited to "missing" data points to complete an already available experimental curve. It also allows calculation of activity coefficients without including experimentally determined values in the input dataset, which will be demonstrated in the following sections. In the following sections, the term "extrapolated" is used in the figure captions when the target value is determined for an electrolyte not included in the training and testing dataset. When the target electrolyte is included in the training and testing dataset, the term "calculated" is used in the caption.

In order to determine the generalisation ability of the model, the LM-NN is used to predict $a(H_2O)$ values for electrolytes that were not included in the input dataset (neither in the training nor in the testing data set, whatever the concentration considered). This generalisation ability is important to ensure that the model performs well even if the target electrolyte is not included in the data set. It helps verify the absence of over-fitting and possibly allows calculating activity coefficients in the absence of experimental values. The results of these predictions can be seen in Fig. 5 (predicted values shown as empty symbols, experimental ones shown as filled symbols), where predicted values are reported at the same concentration as the experimental ones (approach used for each electrolyte is detailed in Fig. S2, while the databse used is described in Table 3). It can be seen from this figure that accurate predictions can be obtained for all four electrolytes reported (caesium chloride, zinc sulfate, copper sulfate and sulfuric acid, [19,40,45]) even if the calculated values are not included in the initial data set, probing the good generalisation ability of the model developed and the pertinence of the selected descriptors. Such performances are very satisfying knowing the small size of the dataset. It once again confirms that the LM-NN is able to learn very quickly on a limited number of training data points. Some discrepancies between experimental and calculated values can however be observed, particularly for copper sulfate and sulfuric acid. For copper sulfate, calculated values are slightly higher than 1 at diluted concentrations, which is not possible. Therefore, the ML model can easily be adapted to include constraints in order to improve accuracy of calculation, *e.g.* by limiting calculated $a(H_2O)$ values between 0 and 1. Predictions remain however within a 5% range of the experimental values (this interval being shown as dotted black lines on all graphs). It should also be noted that 10% deviation between calculated and experimental activity coefficients is often encountered using Pitzer equations ([46]), modification of this equation being often needed at high electrolyte concentrations.

As evidenced in Fig. 5, minor variation from ideal behaviour are appropriately anticipated, with accurate values obtained even if water activity remains higher than 0.9 in the concentration range considered. However, water activity is decreasing linearly with electrolyte concentrations in most cases and remains included between 0 and 1, which makes it relatively easy to model. Osmotic coefficients are another way to describe deviation from ideal behaviour of aqueous electrolyte concentrations. The method described in this work can possibly be applied to calculation of osmotic coefficients, which would necessitate to build a database with osmotic coefficients values.

To further determine the generalisation ability of the model, water activities were calculated for uranyl nitrate solutions [49]. Knowing the activity of water in such solutions is very important for modelling the solvent extraction of uranyl nitrate in the Purex process, an important step of spent nuclear fuel reprocessing [21]. Tabulated values of Pitzer parameters for many electrolytes (>200, [47]) are available, which makes the use of these descriptors very attractive. However, experimental determination by different authors has led to the co-existence of multiple values of Pitzer parameters. Predictions for uranyl nitrate would also allow comparison of prediction accuracy obtained with different values of the same Pitzer parameters for uranyl nitrate, such as the ones reported by Li *et al.* ([50]), Hlushak *et al.* ([22]), Jové Colón *et al.* ([51]), and Tan *et al.* [52]. Predicted values (empty symbols) can be found in Fig. 6, along with the experimental values (filled symbol). The database used is the same as the one used in Fig. 2(d) and Fig. 3(d), described in Table 1.

It can be seen that predicted values are within a 5% range from the experimental ones, whatever the Pitzer parameters used, as long as the

Table 4
RMSE (expressed in percentage) obtained when calculating activity of water in uranyl nitrate solutions using either Pitzer parameters, e-NRTL, e-UNIQUAC or the LM-NN.

Method	RMSE (%)	Reference
Pitzer	0.351	[21]
e-NRTL	1.13	[21]
e-UNIQUAC	0.623	[21]
LM-NN	1.33	This work

electrolyte concentration does not exceed 3 mol/kg. This demonstrates that the proposed method works even for electrolytes like actinides which are not included in the training or testing dataset. However, significant deviations can be seen at high electrolyte concentrations, apart from the Pitzer parameters of Jové Colón *et al.*, which give calculated values within a 5% range of the experimental even for electrolyte concentrations up to 5.5 mol/kg. This emphasises the need for proper selection of the training values, which should ideally be taken from a single reference. This is an inherent limitation of the method. It would ideally be used for the calculation of "unknown" activity values; it is therefore advantageous to build ML models that perform well on small databases. But since the database is small, the ML model shows a limited number of examples. Therefore, conflicts (different values for the same electrolyte obtained from different references) in the input database will be reflected in the final extrapolated result. This indicate the need to select recommended values (when available), values from various references have been critically reviewed.

Discrepancies at higher concentration could be related to the fact that Pitzer equations are generally not valid for calculating activities at high electrolyte concentrations. It is more likely related to a lack of training data at high electrolyte concentrations, making generalisation more challenging. Hence, a proper selection of the input data based on the targeted prediction could probably overcome these issues. Hence, the proposed LM-NN combined with training on small data sets containing Pitzer parameters as descriptors is appropriate for calculating water activities accurately, with good generalisation capabilities. Accuracy in the calculation is favoured by the small variation of the activity of water for most electrolytes, up to relatively high concentration ranges. The determination of activity of water in uranyl nitrate solutions allows comparing the performances of the ML-based method described herein with the methods reported by Balasubramonian *et al.* for the same electrolyte [21]. In this reference, activity of water in uranyl nitrate

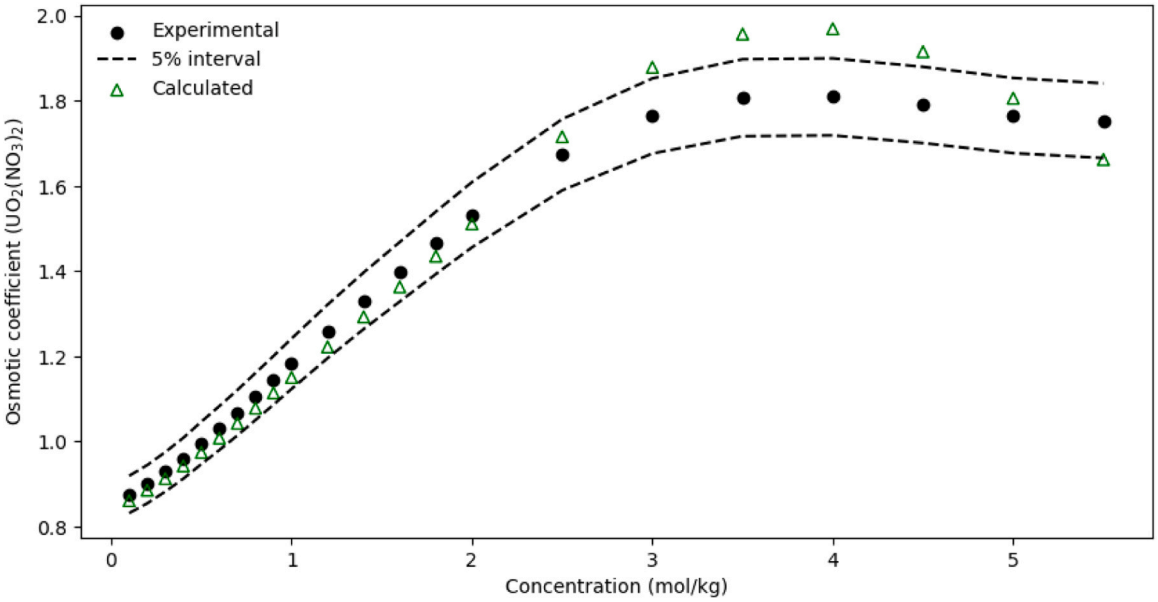

Fig. 7. Experimental ([53]) and extrapolated osmotic coefficients values (at room temperature and pressure) of uranyl nitrate using Pitzer parameters as descriptors.

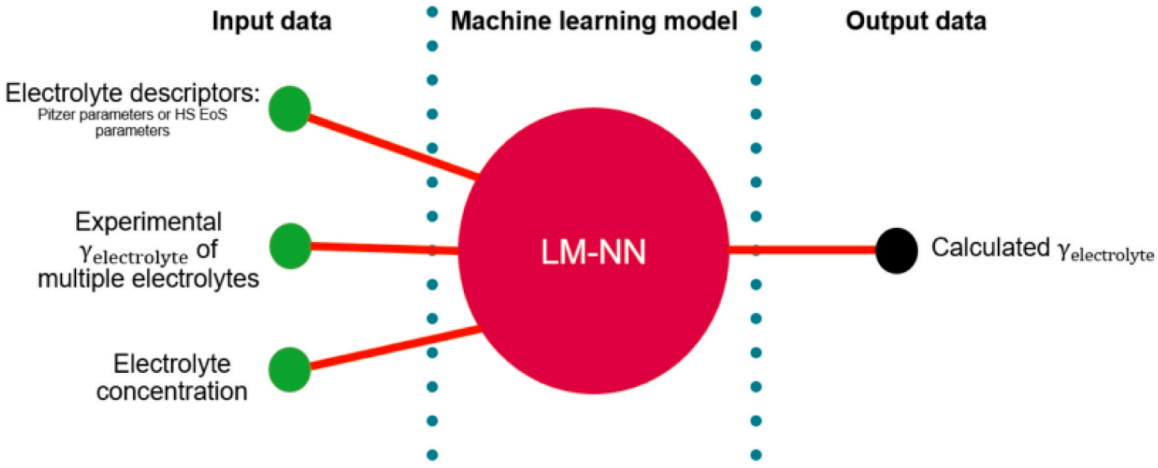

Fig. 8. Description of the ML approach used for the determination of γ(electrolyte) in electrolyte solution.

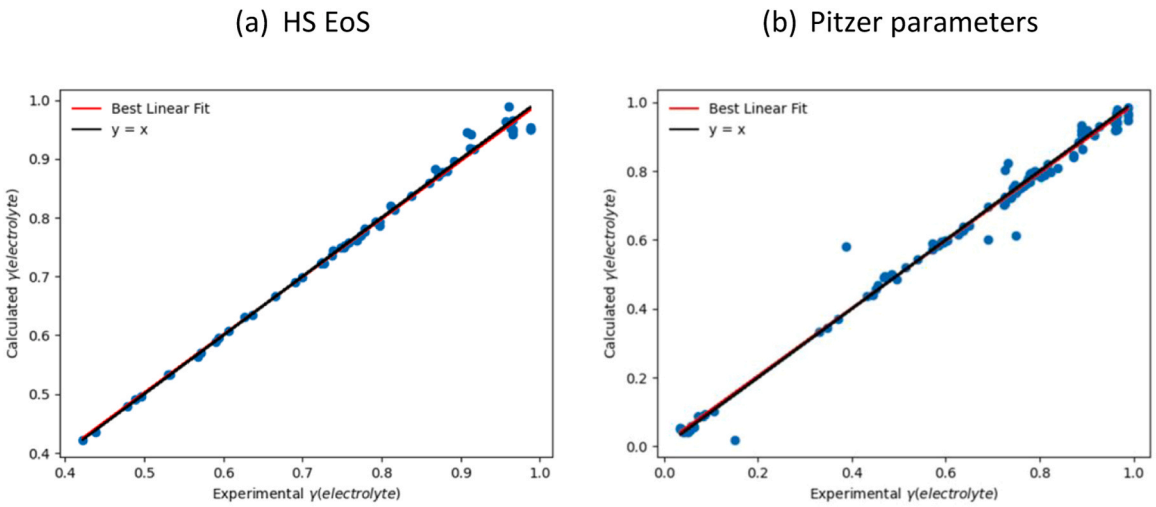

Fig. 9. Calculated and experimental activity coefficients of electrolytes (at room temperature and pressure) using a LM neural network using (a) HS EoS and (b) Pitzer parameters as descriptors.

Table 5
Characteristics of the input datasets used in Fig. 9.

Nature of electrolytes (concentration range in mol/kg)	Reference descriptors	Reference activity values	Figure
Chlorides: H, K, Cs (10^{-4}-11) Bromides and iodides: H, Li, Na (10^{-4}-11) Nitrates: Li, Na (10^{-4}-10.5)	[34]	[54]	Fig. 9 (a)
Chlorides: H, Li, Na, K, Cs, Mg, Sr, Zn (10^{-4}-11) Bromides and iodides: H, Li, Na (10^{-4}-11) Nitrates: H, Li, Na, Ca, Cu (10^{-4}-10.5) Sulfates: Mg, Mn, Ni, Zn (2×10^{-4}-5)	[8,19,40,47]	[54]	Fig. 9 (b)

solutions was calculated using either the empirically determined Pitzer parameters, e-Uniquac or e-NRTL methods. The RMSE (expressed in percentage) obtained with each of those methods is shown in Table 4. It indicates that the lowest RMSE is obtained with the Pitzer parameters (0.351%), followed by the e-UNIQUAC and e-NRTL methods (0.623% and 1.13%). The LM-NN shows the highest deviation between the experimental and calculated values (1.33%), which can be improved by including known experimental values in the training and testing dataset and/or training the model with values of highly concentrated electrolyte concentrations.

Another important aspect is the variation and the range of the input data. The activity of water is relatively easy to model since it often decreases linearly with the electrolyte concentration and it remains included between 0 and 1. Therefore, ML is not necessary for such a simple case, but could rather be used for determination of osmotic coefficients, which are an estimation of the deviation of water from ideal behaviour. From an ML point of view, osmotic coefficients values are more dispersed, which makes them more challenging to extrapolate. However, the ML-based method proposed is still valid to extrapolate osmotic coefficient values as evidenced in Fig. 7, which shows the osmotic coefficient of uranyl nitrate. Extrapolated values are obtained by training the neural network on osmotic coefficients of uranium chloride as well as actinide and lanthanide nitrates (Th, La, Nd, Eu), using Pitzer coefficients as descriptors (see Table 3). Since the values in the training database are close to the extrapolated ones, the calculated values are relatively accurate, most of them being within a 5% range of the experimental ones, all of them being within a 10% range.

Table 6
AARD, R^2 and RMSE (during testing) obtained with the LM-NN trained on various input datasets.

Metric	Pitzer parameters	HS-EoS
AARD	6.30	1.03
R^2	0.988	0.992
RMSE	0.0324	0.0139

3.2. Determination of the mean activity coefficient of electrolytes

3.2.1. Descriptors comparison

Performances of the LM-NN for the calculation of the mean activity coefficient of electrolytes was determined. The activity coefficient values of various electrolytes differ greatly and are not limited between 0 and 1, contrarily to water activities. First, descriptor selection was performed using the approach described in Fig. 8. Pitzer parameters were selected since they allowed obtaining satisfying results for the calculation of $a(H_2O)$ and can easily be found in the literature. Performances obtained using diameter and density of pure hard spheres of electrolytes as descriptors (obtained using the Hard sphere equation of state, HS EoS) were compared to performances obtained using Pitzer parameters as input data (Fig. 9). The databases used are in both cases are described in Table 5. HS EoS parameters were retained since their use as descriptors for calculation of activity coefficients of electrolytes was described in the literature [34].

Good correlation between experimental and predicted values are obtained using both the HS EoS descriptors (R^2=0.992) and Pitzer parameters (R^2=0.988), as displayed in Table 6. However, the HS EoS seems to perform better than the Pitzer parameters and showed lower RMSE (0.014 against 0.032) and most of all, a much lower AARD (1.03 against 6.30). Training data are obtained from mean activity coefficients values for various electrolytes, which are lower than one [54]. The lower performances of the Pitzer parameters are possibly due again to a higher dispersion of the input data, the HS EoS being restrained to fewer data points (188 vs 332) and being limited to a few alkali and alkaline-earth chlorides, nitrates and bromides.

3.2.2. Predictions and generalisation ability

Ability of the LM-NN to predict unseen data was again checked on multiple electrolytes using both kinds of descriptors. The results (Fig. 10, while the corresponding databases are described in Table 5) show that predictions remain accurate for some electrolytes such as RbCl, with predicted values within a 5% range, whatever the descriptors used (HS

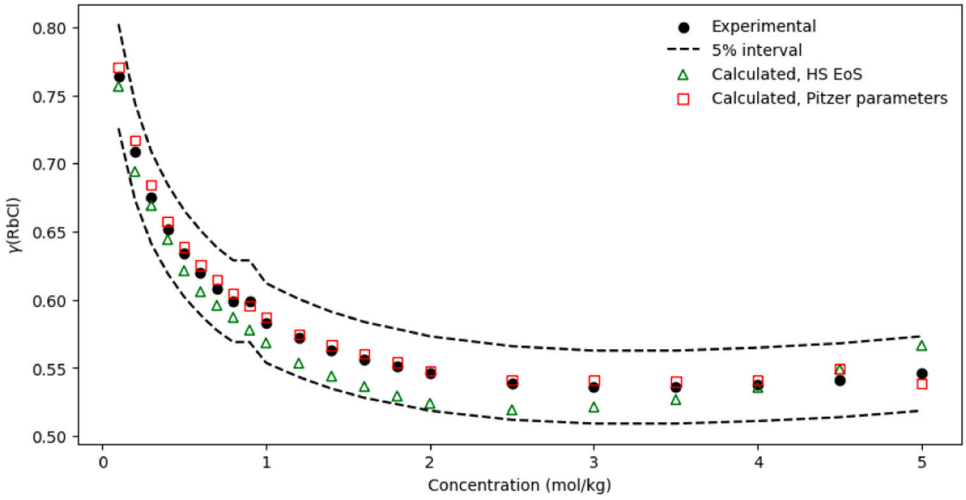

Fig. 10. Experimental and extrapolated activity coefficients of rubidium chloride (at room temperature and pressure) using either Pitzer parameters or HS EoS.

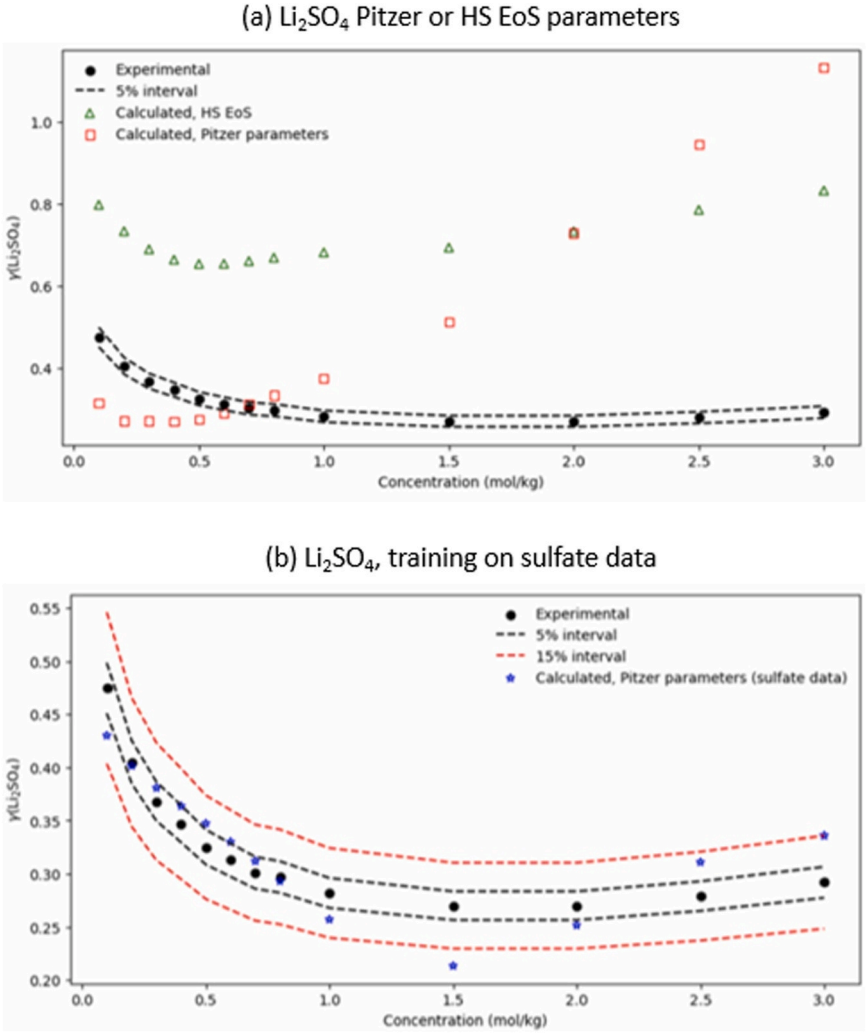

Fig. 11. Experimental and extrapolated activity (at room temperature and pressure) of lithium sulfate using either (a) Pitzer and HS EoS descriptors and (b) Pitzer descriptors, training on sulfate data (see Table 7).

EoS or Pitzer parameters). However, very important deviations are observed for electrolytes such as lithium sulfate or barium chloride, for which the experimental values remain in a small range.

For lithium sulfate (Fig. 11, database shown in Table 7), the values calculated with HS EoS descriptors reproduce the shape of the curve but are far from the experimental values. The ones calculated with Pitzer

Table 7
Characteristics of the input datasets used in Fig. 11.

Nature of electrolytes (concentration range in mol/kg)	Reference descriptors	Reference activity values	Figure
Chlorides: H, K, Cs (10^{-4}-11) Bromides and iodides: H, Li, Na (10^{-4}-11) Nitrates: Li, Na (10^{-4}-10.5) Sulfates: Mg, Mn, Ni, Zn (2×10^{-4}-5)	[34]	[54]	Fig. 11 (a), HS EoS
Chlorides: H, Li, Na, K, Cs, Mg, Sr, Zn (10^{-4}-11) Bromides and iodides: H, Li, Na (10^{-4}-11) Nitrates: H, Li, Na, Ca, Cu (10^{-4}-10.5) Sulfates: Mg, Mn, Ni, Zn (2×10^{-4}-5)	[8,19,40,47]	[54]	Fig. 11 (a), Pitzer parameters
Sulfate salts: Na, K, NH_4, Mg, Mn, Ni, Cu, Zn (0.1–5)	[19]	[19]	Fig. 11 (b)

parameters have a different shape from the one expected, but are closer to the experimental values, at least at low electrolyte concentrations. Prediction accuracy was improved by changing the input dataset. When using only experimental values of alkali and alkaline-earth sulfate salts ([19], lithium sulfate remains excluded from the training and testing data set), calculated values (shown as blue stars in Fig. 11 (b), database shown in Table 7) are much closer to the experimental values, most of them being included in a 5% range, the further staying in a 15% range.

A similar strategy was applied to barium chloride (Fig. 12 (a), along with the corresponding database shown in Table 8) which showed very poor agreement between calculated and experimental values when using Pitzer parameters as descriptors. Prediction accuracy is improved when using a training and testing data set that only includes alkaline-earth iodides and bromides ([55], barium chloride is still excluded from the training and testing dataset, see Fig. 12 (a)), with predicted values within a 15% range of the experimentally determined ones. The predicted values are obviously much better when the barium chloride data are included in the training and testing data set, with high-accuracy calculations (shown in Fig. 12 (b)) obtained for electrolyte concentrations that were not included in the initial data set.

Prediction accuracy also depends on the nature of the Pitzer parameters used. In Fig. 13, calculated and experimental activity coefficients were obtained for neodymium nitrate (training and testing data sets were composed of lanthanide nitrate salts ([46]), neodymium nitrate excluded) using either four Pitzer parameters ($\beta^{(0)}$, $\beta^{(1)}$ and C) or five Pitzer parameters ($\beta^{(0)}$, $\beta^{(1)}$, $\beta^{(2)}$, α and C, which were found from the modified Pitzer equation expressed by Guignot *et al.*, [56]). The results obtained with the five parameters version of the Pitzer equation are more accurate than the ones obtained with the three parameters one, which reflects the fact that consistency on the selected input data is important. Ideally, experimental activity coefficients and Pitzer parameters determined in the same reference should be used.

Calculation of the mean activity coefficient of uranyl nitrate was attempted using the different values of Pitzer coefficients reported in the literature (uranyl nitrate values were not included in the initial data set). The results (available in Fig. S3 (a)) indicate that accurate predictions can be obtained up to uranyl nitrate concentrations of around 1 mol/kg. Large deviations between the experimental and calculated values occur at higher concentrations, whatever the Pitzer parameters used. This can be again related to the lack of training data available at high concentrations. The deviation observed is much higher than for water activities calculations, which could be due to the highest variation of the uranyl nitrate activity coefficients as compared to water activities (calculated values are no longer restrained between 0 and 1). In order to improve prediction accuracy, the training dataset was augmented to include activity coefficients at high concentrations (the initial dataset contained 24 values at electrolyte concentration higher than 3 mol/kg, the augmented one contained 137). In order to keep consistency in the Pitzer coefficients used as descriptors, the values from the same reference were used ([47], uranyl nitrate data are still excluded from the training or testing data set). Calculated values for uranyl nitrate concentrations $\geq$ 2 mol/kg are visible in Fig. S3 (b). The quality of the prediction is improved, although relatively large deviations remain (most of the calculated values are within a 15% interval of the

Table 8
Characteristics of the input datasets used in Figs. 12, 13, and 14.

Nature of electrolytes (concentration range in mol/kg)	Reference descriptors	Reference activity values	Figure
Chlorides: H, K, Cs (10^{-4}-11) Bromides and iodides: H, Li, Na (10^{-4}-11) Nitrates: Li, Na (10^{-4}-10.5) Sulfates: Mg, Mn, Ni, Zn (2×10^{-4}-5)	[34]	[54]	Fig. 12 (a) Pitzer parameters
Alkaline earth bromides and iodides: Mg, Ca, Sr (0.01–2.5)	[55]	[55]	Fig. 12 (a) Alkaline earth data
Alkaline earth bromides and iodides: Mg, Ca, Sr, Ba (0.01–2.5) Chlorides: Ba (0.2–2)	[55]	[55]	Fig. 12 (b)
Lanthanide nitrates: La, Ce, Pr, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu (0.028–3.446)	[56]	[46]	Fig. 13
Alkali chlorides, nitrates, sulfates: Na, K	[47]	[8,19,40]	Fig. 14

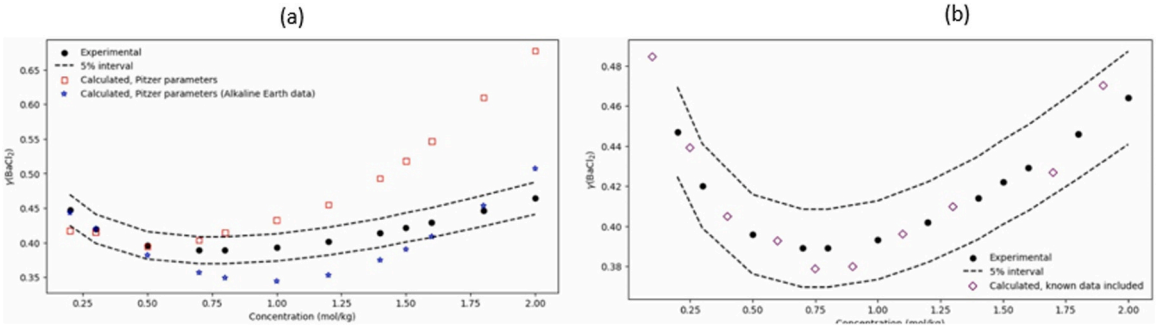

Fig. 12. (a) Experimental and extrapolated values of activity coefficients of barium chloride (at room temperature and pressure) using either a large database (red squares) or a small database containing experimental values of alkaline-earth bromides and iodide (blue stars), (b) experimental and calculated values with a small database, experimental values of alkaline-earths bromides and iodides, known values of barium chloride included.

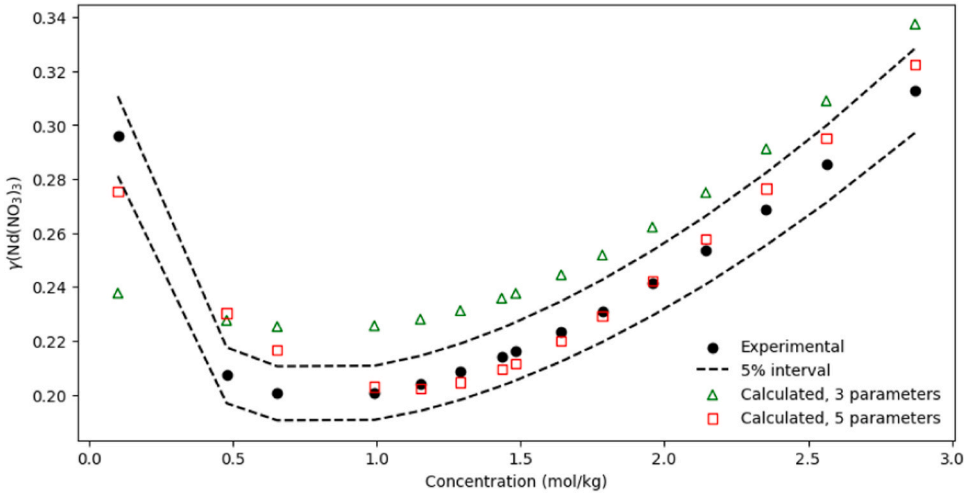

Fig. 13. Experimental and extrapolated activities values (at room temperature and pressure) of neodymium nitrate using using either 3 or 5 Pitzer parameters as descriptors.

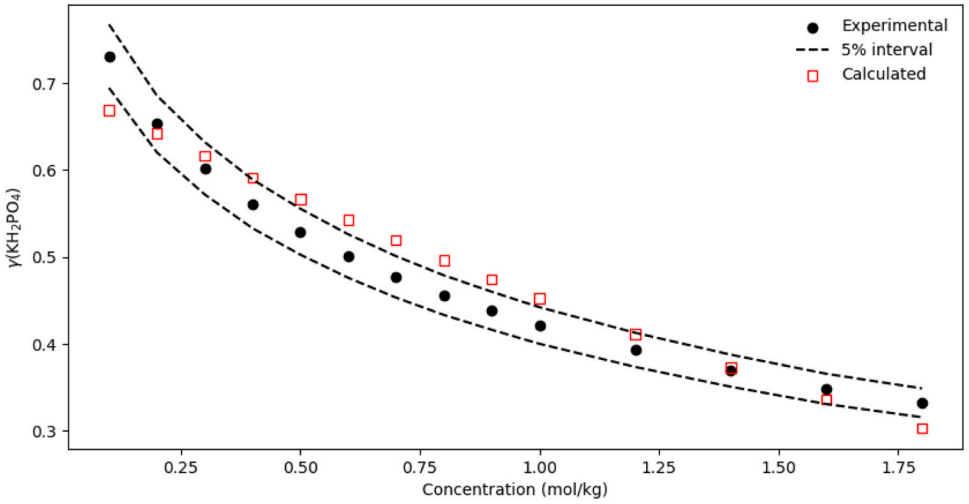

Fig. 14. Experimental ([57]) and extrapolated activities values (at room temperature and pressure) of potassium hydrogen phosphate using Pitzer parameters as descriptors.

experimental data). The generalisation ability of the LM-NN has therefore limits, but can be improved by selecting carefully the input data, ideally by making sure experimental activity coefficients and Pitzer parameters are obtained from the same reference. Prediction accuracy can easily be improved by including the uranyl nitrate data in the input dataset, which however limits predictions to "unseen" concentrations. This strategy was applied by including known data for uranyl nitrate and calculating activity coefficients at concentrations that weren't included in the input dataset. Training and testing data included salts close to uranyl nitrate in the periodic table (throrium nitrate [48], and lanthanide nitrates, [46]). The results (available in Fig. S3 (c)) show a clear improvement with most of the calculated values within a 5% range of the available experimental ones, whatever the concentration range considered.

Hence, the nature of the descriptors plays a role in the accuracy of the calculation. As evidenced in Fig. 9, different descriptors used as input data give different results using the same ML model. However, similarity between input and output values seems to be the most important parameter to ensure that accurate calculation are obtained. Values included in the training database should be as close as possible to the extrapolated ones, which can be achieved by selecting training electrolytes from the same group in the periodic table than the extrapolated ones, or training and target electrolytes having a common ion. This possibility is demonstrated in Fig. 14, with extrapolated activity of potassium dihydrogen phosphate trained on sodium and potassium chlorides, sulfates and nitrates (Table 8). The extrapolated activity values are within a 10% range of the experimental ones, despite the fact that the training database does not contain any value related to a phosphate salt.

Therefore, best results are obtained with a training database containing values "close" to the target electrolyte. It seems necessary to include at least 5 electrolytes of the same group in the periodic Table than the target electrolyte in the training database. Preferentially, the training electrolytes have a common cation or anion with the target electrolyte.

3.3. Performance of the model with simple descriptors and on highly concentrated electrolytes

Overall, the previous results showed that high accuracy can be obtained using a LM-NN and the Pitzer parameters as descriptors. Those parameters are good candidates to be used as descriptors since tabulated values exist for a large panel of electrolytes. However, more simple descriptors can be used to describe the electrolytes. It was hypothesized

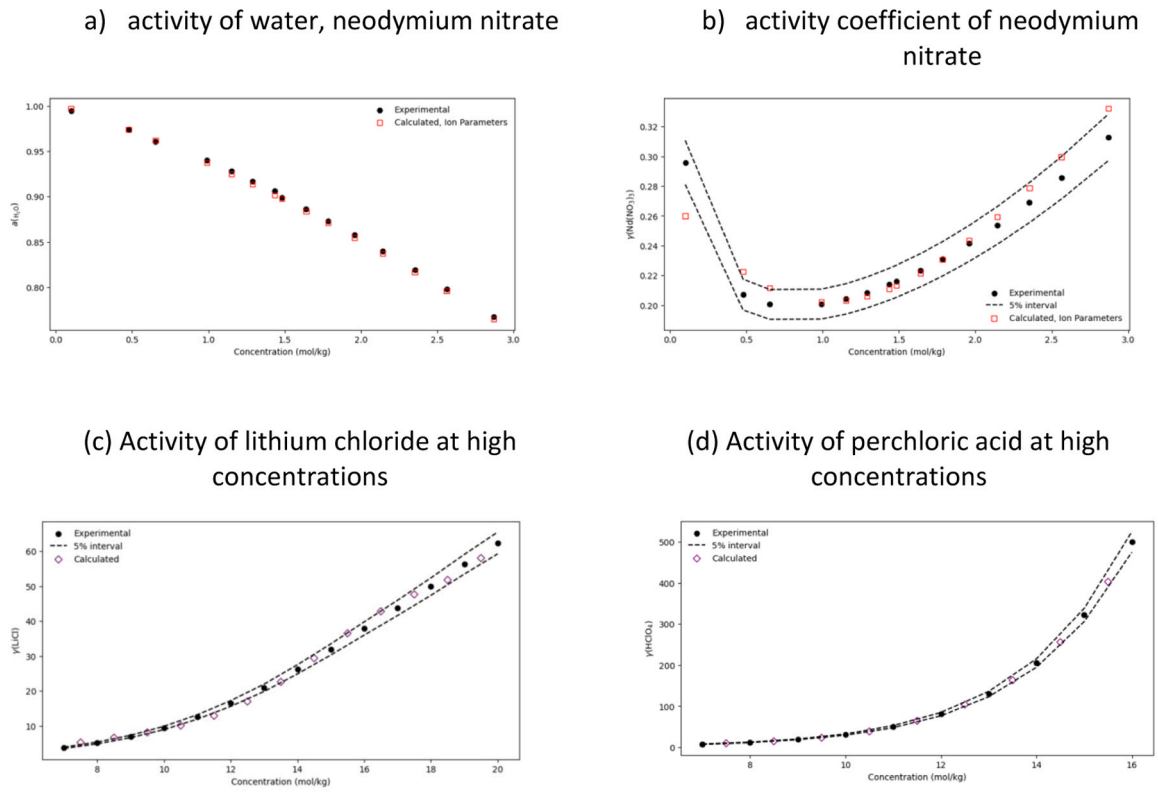

Fig. 15. Experimental and extrapolated activity of (a) water and (b) activity coefficient of electrolyte in neodymium nitrate solutions; experimental and calculated activity of (c) lithium chloride and (d) perchloric acid at high concentration. Ion parameters (ion size, charge and stoichiometric coefficient) are used as descriptor, values at room temperature and pressure.

Table 9
Characteristics of the input datasets used in Fig. 15.

Nature of electrolytes (concentration range in mol/kg)	Reference descriptors	Reference activity values	Figure
Lanthanide nitrates: La, Ce, Pr, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu (0.028–3.446)	-	[46]	Fig. 15 (a) and (b)
Chlorides: H, Li, Na, K, Cs (0.2 -6) Perchlorates: H, Na (0.1–16) Nitrates, bromides: Li (0.1–20)	-	[40,48]	Fig. 15 (c) and (d)

that accuracy of the prediction is related to the similarity between the input dataset and the unknown values, *i.e.* when trying to predict lithium sulfate values, training and testing using sulfate salts values is more appropriate than using chloride salts. Activity of water and activity coefficients of neodymium nitrate (which was outside of the training and testing dataset) were calculated using simple descriptors (cation and anion size and charge, stoichiometric coefficients), using experimental values of the other lanthanide nitrates. As evidenced in Fig. 15 ((a) for activity of water in neodymium nitrate solutions, (b) for activity coefficients of neodymium nitrate, database used is described in Table 9) the LM-NN is able to predict values very accurately even when using extremely simple and easily accessible descriptors as input data.

So far, most existing models for activity calculation struggle to calculate activities for highly concentrated electrolytes (>6 mol/kg). This is why activity coefficients of lithium chloride (Fig. 15 (c)) and perchloric acid (Fig. 15 (d)) were calculated with the LM-NN at high electrolyte concentration (> 6 mol/kg). Values in the training and testing data set were selected to include chlorides or perchlorate-containing electrolytes at high concentrations ([48], approach described in Fig. S4), available data for lithium chloride and perchloric acid being included in the training and testing data set. Stoichiometric coefficients, diameters and charges of cations and anions were used as descriptors. Calculated values are in very good accordance with the experimental ones. However, inclusion of available experimental data in the training and testing data sets seems mandatory in that case to ensure accurate predictions. The lack of data at high concentration leads to a poorer generalisation ability of the model, but it remains very accurate at calculating missing data as long as the experimental ones are included in the training and testing dataset. Therefore, the ML-based approach reported is highly dependent on the database used, but does not take into account the physics of the system considered. It makes the calculation of activities simpler, but also reveals an inherent limitation of the method, which could lead to aberrant results. This is why physics-informed ML-models could not only improve the accuracy but also eliminate the risks of outliers.

4. Conclusion

The experimental determination of water activity and activity coefficients of electrolytes requires a tremendous amount of experimental work, while calculation of these activity coefficients with models such as e-NRTL or UNIQUAC suffers from limitations (difficulty to fit the adjustable parameters, lower accuracy at high concentration…). Machine learning offers the opportunity to determine accurate water activities and activity coefficients in electrolyte solutions, while being practically very simple to use.

Two neural networks were developed to determine water activity and activity coefficients in electrolyte solutions using different descriptors relative to the electrolytes considered (Pitzer parameters, specific salt parameters…). The LM-NN used was proven to be more accurate and allowed predicting values accurately without needing large datasets. Representativeness of the training and testing data sets are probably an important factor, the lack of data at high electrolyte

concentration leading to poorer predictions depending on the concentration range considered. Generalisation ability of the LM-NN model was tested and was found to be correct as long as the training and testing data sets are representative of the values that need to be calculated. Therefore, a "differential" training was employed, where calculation of values related to nitrate salts were performed using other nitrate salts for training and testing. Prediction accuracy could be better with smaller and more representative datasets, which reduces the computing time required. In this work, using 100 epochs, all of the predictions were obtained in less than 30 s. In most cases, the predicted values are within a 5% range of the experimental ones. If the prediction is not satisfying, one could include the experimental values available and predict the activities at missing concentrations, which greatly improves the prediction accuracy. The ML-based method developed could therefore be used to predict very easily water activities and activity coefficients of electrolytes using simple and easily accessible descriptors such as the charge and diameters of ions composing the electrolyte. The method could be applied to a very large panel of electrolytes as long as sufficient experimental data are available. It could be used to generate accurate values in order to obtain activity coefficients for more complex solutions (*i.e.* containing multiple electrolytes), or to feed more complex machine learning models to predict activity coefficients in such solutions. Future work could include application of data augmentation techniques to improve accuracy of the model when experimental data are lacking, or the use of physics-informed ML, *e.g.* by modifying the loss function of the NN to include the Gibbs-Duhem equation [58]. Optimised models could then be used to calculate activity of water in electrolytes in more complex electrolyte solutions. The method presented could also be applied for the determination of activity coefficient of other species (water, electrolytes and extracting molecules in organic solvents within the frame of solvent extraction) and for the calculation of different thermodynamic constants, as long as accurate experimental data exist.

CRediT authorship contribution statement

Guillaume Zante: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgements

This work benefited from French government aid managed by the Agence Nationale de la Recherche under the France 2030 programme, project CYCLAMET under the reference "ANR-22-PERE-0002". The author is grateful to Dr. J-C. Gabriel and Dr. N. Charpentier for their helpful comments and fruitful discussions.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.aichem.2024.100069.

References

[1] K.S. Pitzer. Activity Coefficients in Electrolyte Solutions, 2nd ed, CRC Press, 2018, https://doi.org/10.1201/9781351069472.
[2] J. Chirife, A.J. Fontana, Introduction: Historical Highlights of Water Activity Research, in: G.V. Barbosa-Cánovas, A.J. Fontana, S.J. Schmidt, T.P. Labuza (Eds.), Water Activity in Foods, 1st ed, Wiley, 2020, pp. 1–11, https://doi.org/10.1002/9781118765982.ch1.
[3] K.H. Mistry, H.A. Hunter, J.H. Lienhard V, Effect of composition and nonideal solution behavior on desalination calculations for mixed electrolyte solutions with comparison to seawater, Desalination 318 (2013) 34–47, https://doi.org/10.1016/j.desal.2013.03.015.
[4] M. Steiger, Crystal growth in porous materials—I: the crystallization pressure of large crystals, J. Cryst. Growth 282 (2005) 455–469, https://doi.org/10.1016/j.jcrysgro.2005.05.007.
[5] Y. Marcus, Solvent extraction of inorganic species, Chem. Rev. 63 (1963) 139–170, https://doi.org/10.1021/cr60222a004.
[6] A.B. Zdanovskii, Fundamental Aspects of Variation of Properties of Mixed Solutions: works of Salt Laboratory, Trudy Solyanoï Laboratorii (Transactions of the Salt Laboratory) 6 (1936) 5–70.
[7] V. Taghikhani, J.H. Vera, H. Modarress, Measurement and correlation of the individual ionic activity coefficients of aqueous electrolyte solutions of KF, NaF and KBr, Can. J. Chem. Eng. 78 (2000) 175–181, https://doi.org/10.1002/cjce.5450780123.
[8] M.E. Guendouzi, A. Dinane, Determination of water activities, osmotic and activity coefficients in aqueous solutions using the hygrometric method, J. Chem. Thermodyn. 32 (2000) 297–310, https://doi.org/10.1006/jcht.1999.0574.
[9] W. Davis, H.J. De Bruin, New activity coefficients of 0–100 per cent aqueous nitric acid, J. Inorg. Nucl. Chem. 26 (1964) 1069–1083, https://doi.org/10.1016/0022-1902(64)80268-2.
[10] J.A. MacNeil, G.B. Ray, D.G. Leaist, Activity coefficients and free energies of nonionic mixed surfactant solutions from vapor-pressure and freezing-point osmometry, J. Phys. Chem. B 115 (2011) 5947–5957, https://doi.org/10.1021/jp201500y.
[11] C. Penisson, A. Wilk, J. Theisen, V. Kokoric, B. Mizaikoff, J.-C.P. Gabriel, Water activity measurement of NaCl/H2O mixtures via substrate-integrated hollow waveguide infrared spectroscopy with integrated microfluidics, Nanotech 2018 - 20th Annu. Nanotech Conf., Anaheim, U. S. (2018) 198–201. ⟨https://cea.hal.science/cea-03323266⟩.
[12] V. Kokoric, J. Theisen, A. Wilk, C. Penisson, G. Bernard, B. Mizaikoff, J.-C. P. Gabriel, Determining the partial pressure of volatile components via substrate-integrated hollow waveguide infrared spectroscopy with integrated microfluidics, Anal. Chem. 90 (2018) 4445–4451, https://doi.org/10.1021/acs.analchem.7b04425.
[13] P. Debye, E. Hückel, Zur Theorie der Elektrolyte, Phys. Z. 24 (1923) 185–206.
[14] G.G. Manov, R.G. Bates, W.J. Hamer, S.F. Acree, Values of the Constants in the Debye—Hückel Equation for Activity Coefficients, J. Am. Chem. Soc. 65 (1943) 1765–1767, https://doi.org/10.1021/ja01249a028.
[15] J. Baezabaeza, G. Ramisramos, A series expansion of the extended Debye-Hückel equation and application to linear prediction of stability constants, Talanta 43 (1996) 1579–1587, https://doi.org/10.1016/0039-9140(96)01942-X.
[16] I.Yu Shilov, A.K. Lyashchenko, The role of concentration dependent static permittivity of electrolyte solutions in the debye–hückel theory, J. Phys. Chem. B 119 (2015) 10087–10095, https://doi.org/10.1021/acs.jpcb.5b04555.
[17] I.Yu Shilov, A.K. Lyashchenko, Modeling activity coefficients in alkali iodide aqueous solutions using the extended Debye-Hückel theory, J. Mol. Liq. 240 (2017) 172–178, https://doi.org/10.1016/j.molliq.2017.05.010.
[18] K.S. Pitzer, Thermodynamics of electrolytes. I. Theoretical basis and general equations, J. Phys. Chem. 77 (1973) 268–277, https://doi.org/10.1021/j100621a026.
[19] M.E. Guendouzi, A. Mounir, A. Dinane, Water activity, osmotic and activity coefficients of aqueous solutions of Li2SO4, Na2SO4, K2SO4, (NH4)2SO4, MgSO4, MnSO4, NiSO4, CuSO4, and ZnSO4 at T=298.15K, J. Chem. Thermodyn. 35 (2003) 209–220, https://doi.org/10.1016/S0021-9614(02)00315-4.
[20] W. Voigt, Chemistry of salts in aqueous solutions: applications, experiments, and theory, Pure Appl. Chem. 83 (2011) 1015–1030, https://doi.org/10.1351/PAC-CON-11-01-07.
[21] S. Balasubramonian, N.K. Pandey, R.V. Subba Rao, Comparison of activity coefficient models for the estimation of uranyl nitrate and nitric acid distribution coefficients in phosphoric solvent, Prog. Nucl. Energy 128 (2020) 103472, https://doi.org/10.1016/j.pnucene.2020.103472.
[22] S.P. Hlushak, J.P. Simonin, B. Caniffi, P. Moisy, C. Sorel, O. Bernard, Description of partition equilibria for uranyl nitrate, nitric acid and water extracted by tributyl phosphate in dodecane, Hydrometallurgy 109 (2011) 97–105, https://doi.org/10.1016/j.hydromet.2011.05.014.
[23] S. Gebreyohannes, B.J. Neely, K.A.M. Gasem, Generalized interaction parameter for the modified nonrandom two-liquid (NRTL) activity coefficient model, Ind. Eng. Chem. Res. 53 (2014) 20247–20257, https://doi.org/10.1021/ie503135c.
[24] C. Zhang, Y. Xing, D. Tao, A two-parameter theoretical model for predicting the activity and osmotic coefficients of aqueous electrolyte solutions, J. Solut. Chem. 49 (2020) 659–694, https://doi.org/10.1007/s10953-020-00987-z.
[25] L.S. Belvèze, J.F. Brennecke, M.A. Stadtherr, Modeling of activity coefficients of aqueous solutions of quaternary ammonium salts with the electrolyte-NRTL equation, Ind. Eng. Chem. Res. 43 (2004) 815–825, https://doi.org/10.1021/ie0340701.
[26] S. Balasubramonian, N.K. Pandey, K. Shekhar, R.V. Subba Rao, Thermodynamic modeling of nitric acid speciation using eUNIQUAC activity coefficient model, J. Solut. Chem. 50 (2021) 1300–1314, https://doi.org/10.1007/s10953-021-01124-0.
[27] K. Thomsen, P. Rasmussen, R. Gani, Correlation and prediction of thermal properties and phase behaviour for a class of aqueous electrolyte systems, Chem. Eng. Sci. 51 (1996) 3675–3683, https://doi.org/10.1016/0009-2509(95)00418-1.
[28] O. Toure, F. Audonnet, A. Lebert, C.-G. Dussap, COSMO-RS-PDHS: a new predictive model for aqueous electrolytes solutions, Chem. Eng. Res. Des. 92 (2014) 2873–2883, https://doi.org/10.1016/j.cherd.2014.06.020.

[29] D.O. Abranches, E.J. Maginn, Y.J. Colón, Activity coefficient acquisition with thermodynamics-informed active learning for phase diagram construction, AIChE J. 69 (2023) e18141, https://doi.org/10.1002/aic.18141.
[30] H. Benimam, C. Si-Moussa, M. Laidi, S. Hanini, Modeling the activity coefficient at infinite dilution of water in ionic liquids using artificial neural networks and support vector machines, Neural Comput. Applic 32 (2020) 8635–8653, https://doi.org/10.1007/s00521-019-04356-w.
[31] E.I. Sanchez Medina, S. Linke, M. Stoll, K. Sundmacher, Graph neural networks for the prediction of infinite dilution activity coefficients, Digit. Discov. 1 (2022) 216–225, https://doi.org/10.1039/D1DD00037C.
[32] F. Jirasek, R.A.S. Alves, J. Damay, R.A. Vandermeulen, R. Bamler, M. Bortz, S. Mandt, M. Kloft, H. Hasse, Machine learning in thermodynamics: prediction of activity coefficients by matrix completion, J. Phys. Chem. Lett. 11 (2020) 981–985, https://doi.org/10.1021/acs.jpclett.9b03657.
[33] B. Winter, C. Winter, J. Schilling, A. Bardow, A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing, Digit. Discov. 1 (2022) 859–869, https://doi.org/10.1039/D2DD00058J.
[34] K. Golzar, S. Amjad-Iranagh, H. Modarress, Evaluation of compressibility factor and mean ionic activity coefficient for aqueous electrolyte solutions with hard sphere equations of state in the MSA model and artificial neural network method, J. Mol. Liq. 207 (2015) 50–59, https://doi.org/10.1016/j.molliq.2015.02.043.
[35] H.K. Gallage Dona, T. Olayiwola, L.A. Briceno-Mena, C.G. Arges, R. Kumar, J. A. Romagnoli, Determining ion activity coefficients in ion-exchange membranes with machine learning and molecular dynamics simulations, Ind. Eng. Chem. Res. 62 (2023) 9533–9548, https://doi.org/10.1021/acs.iecr.3c00636.
[36] S. Gbashi, T.L. Maselesele, P.B. Njobeh, T.B.J. Molelekoa, S.A. Oyeyinka, R. Makhuvele, O.A. Adebo, Application of a generative adversarial network for multi-featured fermentation data synthesis and artificial neural network (ANN) modeling of bitter gourd–grape beverage production, Sci. Rep. 13 (2023) 11755, https://doi.org/10.1038/s41598-023-38322-3.
[37] K. Levenberg, A method for the solution of certain non-linear problems in least squares, Quart. Appl. Math. 2 (1944) 164–168, https://doi.org/10.1090/qam/10666.
[38] D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, J. Soc. Ind. Appl. Math. 11 (1963) 431–441, https://doi.org/10.1137/0111030.
[39] B.M. Wilamowski, J.D. Irwin. The industrial electronics handbook Intelligent systems, 2nd ed, CRC Press, Boca Raton, FL, 2011.
[40] M.E. Guendouzi, A. Dinane, A. Mounir, Water activities, osmotic and activity coefficients in aqueous chloride solutions atT= 298.15 K by the hygrometric method, J. Chem. Thermodyn. 33 (2001) 1059–1072, https://doi.org/10.1006/jcht.2000.0815.
[41] N. Silva, J.M.O. Bacicheti, L. Campana, D. Rossoni, M. Castier, V.F. Cabral, Calculation of water activity in electrolytic solutions using the Electrolattice and Q-Electrolattice equations of state, Fluid Phase Equilibria 563 (2023) 113569, https://doi.org/10.1016/j.fluid.2022.113569.
[42] A. Zuber, R.F. Checoni, M. Castier, Thermodynamic properties of aqueous solutions of single and multiple salts using the Q-electrolattice equation of state, Fluid Phase Equilibria 362 (2014) 268–280, https://doi.org/10.1016/j.fluid.2013.10.021.
[43] A. Zuber, R.F. Checoni, R. Mathew, J.P.L. Santos, F.W. Tavares, M. Castier, Thermodynamic Properties of 1:1 salt aqueous solutions with the electrolattice equation of state, Oil Gas. Sci. Technol. – Rev. IFP Energ. Nouv. 68 (2013) 255–270, https://doi.org/10.2516/ogst/2012088.
[44] Y. Marcus, Ionic radii in aqueous solutions, Chem. Rev. 88 (1988) 1475–1498, https://doi.org/10.1021/cr00090a003.
[45] K.S. Pitzer, R.N. Roy, L.F. Silvester, Thermodynamics of electrolytes. 7. Sulfuric acid, J. Am. Chem. Soc. 99 (1977) 4930–4936, https://doi.org/10.1021/ja00457a008.
[46] S. Chatterjee, E.L. Campbell, D. Neiner, N.K. Pence, T.A. Robinson, T.G. Levitskaia, Aqueous binary lanthanide(III) nitrate $Ln(NO_3)_3$ electrolytes revisited: extended pitzer and bromley treatments, J. Chem. Eng. Data 60 (2015) 2974–2988, https://doi.org/10.1021/acs.jced.5b00392.
[47] P.M. May, D. Rowland, G. Hefter, E. Königsberger, A generic and updatable pitzer characterization of aqueous binary electrolyte solutions at 1 bar and 25 °C, J. Chem. Eng. Data 56 (2011) 5066–5077, https://doi.org/10.1021/je2009329.
[48] R.A. Robinson, R.H. Stokes, Tables of osmotic and activity coefficients of electrolytes in aqueous solution at 25° C, Trans. Faraday Soc. 45 (1949) 612–624, https://doi.org/10.1039/TF9494500612.
[49] A. Ochkin, D. Gladilov, S. Nekhaevskiy, A. Merkushkin, Activity coefficients of uranyl nitrate and nitric acid in aqueous mixtures, Procedia Chem. 21 (2016) 87–92, https://doi.org/10.1016/j.proche.2016.10.013.
[50] Z. Li, J. Chen, T. Bao, Y. Shang, Y. Li, Prediction of phase equilibria in tributyl phosphate extraction system using the unifac group contribution method, Thermochim. Acta 169 (1990) 287–300, https://doi.org/10.1016/0040-6031(90)80155-R.
[51] C.F. Jové Colón, H.K. Moffat, R.R. Rao, Modeling of liquid-liquid extraction (LLE) equilibria using gibbs energy minimization (GEM) for the system TBP–HNO_3–UO_2–H_2O–Diluent, Solvent Extr. Ion.-. Exch. 31 (2013) 634–651, https://doi.org/10.1080/00397911.2013.785882.
[52] B. Tan, C. Chang, D. Xu, Y. Wang, T. Qi, Modeling of the competition between uranyl nitrate and nitric acid upon extraction with Tri- *n* -butyl phosphate, ACS Omega 5 (2020) 12174–12183, https://doi.org/10.1021/acsomega.0c00583.
[53] R.A. Robinson, C.K. Lim, The osmotic and activity coefficients of uranyl nitrate, chloride, and perchlorate at 25°, J. Chem. Soc. 0 (1951) 1840–1843, https://doi.org/10.1039/JR9510001840.
[54] E. Moggia, B. Bianco, Mean Activity Coefficient of Electrolyte Solutions, J. Phys. Chem. B 111 (2007) 3183–3191, https://doi.org/10.1021/jp067133c.
[55] J.I. Partanen, Traceable activity and osmotic coefficients in pure aqueous solutions of alkaline earth metal bromides and iodides at 25 °C, J. Chem. Eng. Data 59 (2014) 2530–2540, https://doi.org/10.1021/je500298g.
[56] S. Guignot, A. Lassin, C. Christov, A. Lach, L. André, P. Henocq, Modeling the osmotic and activity coefficients of lanthanide nitrate aqueous solutions at 298.15 K from low molalities to supersaturation, J. Chem. Eng. Data 64 (2019) 345–359, https://doi.org/10.1021/acs.jced.8b00859.
[57] J.M. Stokes, The osmotic and activity coefficients of sodium and potassium dihydrogen phosphate at 25°, Trans. Faraday Soc. 41 (1945) 685–688, https://doi.org/10.1039/TF9454100685.
[58] J.G. Rittig, K.C. Felton, A.A. Lapkin, A. Mitsos, Gibbs–Duhem-informed neural networks for binary activity coefficient prediction, Digit. Discov. 2 (2023) 1752–1767, https://doi.org/10.1039/D3DD00103B.