Artificial neural network modeling on the polymer-electrolyte aqueous two-phase systems involving biomolecules

Chen, Yuqiu; Liang, Xiaodong; Kontogeorgis, Georgios M.

Published in:
Separation and Purification Technology

Link to article, DOI:
10.1016/j.seppur.2022.122624

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Artificial neural network modeling on the polymer-electrolyte aqueous two-phase systems involving biomolecules

Yuqiu Chen, Xiaodong Liang, Georgios M. Kontogeorgis*

Department of Chemical and Biochemical Engineering, Technical University of Denmark, DK-2800 Lyngby, Denmark

ARTICLE INFO

Keywords:
Biomolecule separation and purification
Aqueous two-phase systems (ATPS)
Modeling
Electrolytes
Artificial neural network

ABSTRACT

Polymer-electrolyte aqueous two-phase systems (ATPS) have demonstrated their superior performance in the separation and purification of high-value biomolecules. However, these powerful platforms are still a major academic curiosity, without their acceptance and implementation by industry. One of the major obstacles is the absence of models to predict the partition of biomolecules in ATPS in an easy and predictive way. To address this limitation, modelling studies on the binodal curve behavior of polymer-electrolyte ATPS and the partitioning of biomolecules in these aqueous electrolyte solutions are carried out in this work. First, a comprehensive database targeting the studied systems is established. In total, 11,998 experimental binodal data points covering 276 polymer-electrolyte ATPS at different temperatures (273.15 K-399.15 K) and 626 experimental partition data points involving 22 biomolecules in 42 polymer-electrolyte ATPS at different temperatures (283.15 K-333.15 K) are included. Then, a novel modeling strategy that combines a well-known machine learning algorithm, i.e., artificial neural network (ANN) and group contribution (GC) method is proposed. Based on this modeling strategy, an ANN-GC model (ANN-GC model1) is built to describe the binodal curve behavior of polymer-electrolyte ATPS, while another ANN-GC model (ANN-GC model2) is developed to predict the partition of biomolecules in these biphasic systems. ANN-GC model1 gives a mean absolute error (MAE) of 0.0132 and squared correlation coefficient ($R^2$) of 0.9878 for the 9,598 training data points, and for the 1,200 validation data points they are 0.0141 and 0.9858, respectively. Meanwhile, it also gives a MAE of 0.0143 and $R^2$ of 0.9846 for the 1,200 test data points. On the other hand, ANN-GC model2 gives root-mean-square deviation (RMSD) of 0.0577 for 501 training data points, and for the 62 validation data points and 63 test data points their RMSD are 0.0849 and 0.0885, respectively. Furthermore, the obtained results also indicate that the tie-line length of polymer-electrolyte ATPS calculated from ANN-GC model1 can be directly used in ANN-GC model2 for predicting the partition performance coefficient of biomolecules in these ATPS. The developed models offer the possibility to predict the partition of biomolecules in ATPS without any requirement of experimental data. Based on the developed ANN-GC models, some high-performance ATPS are identified to partition four well-known biomolecules.

1. Introduction

Aqueous two-phase systems (ATPS), also known as aqueous biphasic systems (ABS), are formed when two or more water-soluble components, such as polymers, salts, ionic liquids, alkaline and alcohols, are mixed at appropriate concentrations and temperatures in water. To date, various combinations of phase forming agents (e.g. polymer -salt/alkaline, polymer–polymer, ionic liquid-salt, alcohol-salt) have been proposed for the creation of ATPS [1,2]. Many ATPS and the combination of these two-phase systems with other techniques, such as microfluidic apparatus [3–5], have exhibited great technical and economic advantages in biotechnological applications [6,7]. Due to the presence of high-water content in both phases, ATPS can provide a biocompatible and non-denaturing environment for cells, proteins, and other biomolecules. Meanwhile, ATPS generally present less damage to the extracted biomolecules as they allow rapid phase separation and compounds partition, leading to much lower interfacial stress than that of organic-water solvent systems. In addition, ATPS can offer high recovery percentages and high purity of biomolecules in a one-step process. Besides these, ATPS show characteristics of high tailored space and they are also easily...
to be scaled-up [8–10].

In the past, ATPS have been widely studied as clean alternatives for conventional organic-water solvent extraction systems in different biotechnological applications [11], and many types of them have been used in the downstream recovering of biomolecules [12,13] (e.g. protein [14,15], enzyme [16], lactose [17], alkaldoids [18], nucleic acids [19], phenolic compounds [20], polyphenols [21], amino acids [22], hydroxycitric acid [23]) with high recovery percentages while preserving their activity. Despite the definite advantages of ATPS, the application of ATPS is still considered mostly of a major academic interest, without wide acceptance and implementation by industry [24]. This is mainly due to the difficulty in understanding the mechanism governing the phase formation of ATPS and the complexity in predicting the partition of biomolecules in ATPS [6]. In addition, the difficulty of finding low toxicity, low cost and high-performance ATPS for biomolecules with different properties is also a reason claimed to be the limitation for applying ATPS in industry. Besides these, how to efficiently reuse the phase-forming components of ATPS is another issue that challenges their industrial application [25]. A good mechanism knowledge that can guide the phase formation of ATPS is obviously of great importance for enhancing the opportunity of ATPS adoption in industry. Meanwhile, the ability of providing reliable predictions on the partition of biomolecules in ATPS is also essential given the fact that it would largely reduce the time and cost to find high-performance ATPS for biomolecules. Therefore, a systematic modeling study on the phase equilibria behavior of ATPS and the partition of biomolecules in ATPS is highly desirable for the transition of ATPS separation technique from pure academic focus to industrial implementation.

To describe the liquid–liquid equilibria of ATPS, different empirical equations and thermodynamic models have been studied. The
et al. [31] introduced it to the ATPS containing water-miscible ionic liquids and water-structuring salts. This expression has proved to be able to describe the binodal curve of polymer-based ATPS, and then Gutowski et al. [32] proposed an equation to reproduce the binodal curves of different ATPS with a relatively high accuracy. On the other hand, activity-coefficient models (e.g. Wilson [37], NRTL [33], UNIFAC, UNIQUAC [34]) and equation of state (EoS) models such as the Perturbed Hard Sphere Chain (PHSC) electrolyte EoS [35] and the electrolyte Perturbed-Chain Statistical Associating Fluid Theory (ePC-SAFT) [36–38] have also been used to study the phase behavior of ATPS. Most recently, a hybrid model of Flory–Huggins modified Flory-Huggins [28] to correlate the partition coefficients of hydrolytic enzymes in ATPS composed of PEG and dextran [40]. A model derived from the modified Flory–Huggins equation was proposed by Lin et al. in order to predict the partitioning of five proteins in PEG-dextran ATPS covering a wide range of polymer molecular weights and polymer concentration [41]. Pazuki et al. used an artificial neural network model to predict the partition coefficients of different biomolecules (e.g. α-amylase, β-amylase, albumin) in PEG-dextran ATPS [42]. Partitioning study of biomolecules in electrolyte involved ATPS has also been paid attention in the past decade. Madeira [43] and Doozandeh [44], respectively, employed the modified Wilson and Electrolyte-SAFT method to study the protein partitioning in polymer-salt ATPS. Shahriari et al., [45] applied the UNIFAC-FV model to predict the partitioning behaviour of β-amylase and amyloglucosidase in ATPS composed of PEG and KOPO₄/Na₂SO₄. So far, most proposed correlations and models are limited to a single or some specific biomolecule-ATPS.

To the best of our knowledge, models that are capable to describe the partitioning behaviour of biomolecules in ATPS in a widespread way are not yet available. Due to the high complexity of these systems, empirical correlations and theory-driven models cannot simultaneously provide reliable predictions on the partition of different biomolecules in various ATPS. In this respect, machine learning (ML) algorithms may have potentials given the fact that algorithms such as ANN and support vector machine (SVM) have been successfully used to model thermodynamic and transport properties of complex systems including ionic liquid-based ATPS [46,47] andionic liquid–water mixtures [48,49]. Nonetheless, some issues should also be addressed when using ML algorithms for the partition prediction of biomolecules in polymer-electrolyte ATPS. For example, how to describe the structure of studied polymer-electrolyte ATPS. For the partitioning behaviour of biomolecules in ATPS, modeling studies have also been performed by some researchers using different empirical correlations and theory-driven models. Furuya applied the modified Flory-Huggins to correlate the partition coefficients of hydrolytic enzymes in ATPS composed of PEG and dextran [40]. A model derived from the modified Flory-Huggins equation was proposed by Lin et al. in order to predict the partitioning of five proteins in PEG-dextran ATPS covering a wide range of polymer molecular weights and polymer concentration [41]. Pazuki et al. used an artificial neural network model to predict the partition coefficients of different biomolecules (e.g. α-amylase, β-amylase, albumin) in PEG-dextran ATPS [42]. Partitioning study of biomolecules in electrolyte involved ATPS has also been paid attention in the past decade. Madeira [43] and Doozandeh [44], respectively, employed the modified Wilson and Electrolyte-SAFT method to study the protein partitioning in polymer-salt ATPS. Shahriari et al., [45] applied the UNIFAC-FV model to predict the partitioning behaviour of β-amylase and amyloglucosidase in ATPS composed of PEG and KOPO₄/Na₂SO₄. So far, most proposed correlations and models are limited to a single or some specific biomolecule-ATPS.

To the best of our knowledge, models that are capable to describe the partitioning behaviour of biomolecules in ATPS in a widespread way are not yet available. Due to the high complexity of these systems, empirical correlations and theory-driven models cannot simultaneously provide reliable predictions on the partition of different biomolecules in various ATPS. In this respect, machine learning (ML) algorithms may have potentials given the fact that algorithms such as ANN and support vector machine (SVM) have been successfully used to model thermodynamic and transport properties of complex systems including ionic liquid-based ATPS [46,47] andionic liquid–water mixtures [48,49]. Nonetheless, some issues should also be addressed when using ML algorithms for the partition prediction of biomolecules in polymer-electrolyte ATPS. For example, how to describe the structure of studied polymer-electrolyte ATPS. Due to the large number of involved electrolytes, it is not good to treat each electrolyte as a separate input element. On the other hand, the group contribution (GC) method is ideally suited as each electrolyte can be described as a combination of different cations and anions, which would largely reduce the number of the model’s input elements. In addition, GC-based models can also be easily integrated in the computer-aided design method [50], which makes it possible to design optimal polymer-electrolyte ATPS for specific biomolecule from a molecular level. Therefore, this work aims to investigate the possibility of combining ANN algorithm and GC method to model the partitioning behaviour of biomolecules in polymer-electrolyte ATPS. Besides this, some major issues that govern the formation of polymer-electrolyte ATPS and the partition of biomolecules in these aqueous electrolyte solutions are also discussed.

2. Methodology

In this work, a comprehensive database targeting polymer-electrolyte ATPS and biomolecules in these aqueous electrolyte systems is established. In total, 11,998 experimental binodal data points...
covering 276 polymer-electrolyte ATPS at a temperature range of 273.15 K-399.15 K (summarized in Table 1) and 626 experimental partition data points involving 22 biomolecules in 42 polymer-electrolyte ATPS (89 different systems in total) at a temperature range of 283.15 K-333.15 K (summarized in Table 3) are included. All studied biomolecules and polymer-electrolyte ATPS together with their experimental binodal and partition data are given in the Supporting Information. The binodal curve and partition data are, respectively, organized and stored in Database1 (Table S2) and Database2 (Table S3) for the further model development.

Based on the developed Database1 and Database2, a novel modeling strategy for the partitioning behavior of biomolecules in polymer-electrolyte ATPS is proposed, as shown in Fig. 1. Testing is an essential process in the model development and here both experimental binodal and partition data are divided into a training set and a test set. About 80% and 10% of the data points are, respectively, randomly selected as training sets (Training dataset1 and Training dataset2) and validation sets (Validation dataset1 and Validation dataset2) for the model development and the rest of data points are used as test sets (Test dataset1 and Test dataset2) to evaluate the predictive performance of the developed models (ANN-GC model1 and ANN-GC model2). The training set, validation set and test set of Database1 are provided in Tables S4-S6, and for the Database2 they are given in Tables S7-S9, respectively. It’s worth mentioning that a classification process on Database2 will be performed before splitting it into training and test sets. In this work, a neural network is applied to model the binodal curve of polymer-electrolyte ATPS and the partitioning behavior of biomolecules in these aqueous electrolyte systems, while the GC method is introduced to describe the structure of involved electrolytes. The partition performance coefficient of biomolecules $P_{bin}(\text{see Section 4})$ in polymer-electrolyte ATPS can be predicted from ANN-GC model2 with the input including the system’s temperature, the molecular weight of biomolecule, the tie-line length (TLL) and the structure of these aqueous electrolyte systems. The TLL of polymer-electrolyte ATPS with specific feed compositions can be obtained from ANN-GC model1 with the input including the system’s temperature, the type and molecular weight of polymer, the GC descriptor and the weight fraction of electrolyte.

### 3. Ann-GC model1

In this section, artificial neural network modeling will be performed on the binodal curve prediction of polymer-electrolyte ATPS. The ANN concept that inspired by biological neural systems has been widely applied in many technical fields due to its simplicity, flexibility, and ability in different modeling systems [161]. It is characterized by layered architectures and feed-forward connections between neurons, or back connections. Weights are assigned to these connections between the neurons of one layer and the next. The benefits of ANN modeling include: (1) ANN can learn organically, and this means that their outputs aren’t limited entirely by inputs, and they can generalize their inputs; (2) nonlinear systems are able to find shortcuts to reach computationally expensive solutions; (3) ANN has the potential for high fault tolerance; (4) ANN is capable to do more than routing around parts of the network that no longer operate [162]. In this work, a popular ANN architecture that comprises of a three-layer feed forward network is employed, as presented in Fig. 2. In this ANN, $P$ represents the input vector containing the system’s information of polymer-electrolyte ATPS, while $W_1$, $W_2$ are the weight matrices and $b_1$, $b_2$ are the bias vectors of the transfer functions in the hidden and output layers, respectively.

The input layer reads the inputs (50 elements) that include the system’s temperature, the weight fraction of electrolyte and the structure information of polymer-electrolyte ATPS. To describe the structure of these aqueous electrolyte systems, the polymers/copolymers are described as a combination of their type and molecular weight, while the electrolytes are described by their cations and anions. Table 2 gives all the types of studied polymers/copolymers and the ions decomposed from all involved electrolytes. This input is transferred in the hidden layer and then delivered to the output layer, in which the weight fractions of polymer/copolymers are predicted. The summation of the errors between the experimental and model-predicted weight fractions of polymer/copolymer are then quantified. For a given input vector $P$, the output of the hidden layer $A_1$ and the output layer $Y$ are, respectively, calculated by Eq. 1 and Eq. 2.

$$A_1 = f_1(W_1 \times P + b_1)$$  \hspace{1cm} (2)

$$Y = f_2(W_2 \times A_1 + b_2)$$  \hspace{1cm} (2)

In this work, the combination of tansig transfer function (see Eq. 3) in the hidden layer and purelin transfer function (see Eq. 4) in the output layer (in MATLAB) is applied due to it generally can provide good performance for the three-layer neural network modeling [163].

$$f_1(x) = \frac{2}{1 + e^{-2x}} - 1$$  \hspace{1cm} (3)

$$f_2(x) = x$$  \hspace{1cm} (4)

To avoid either underfitting or overfitting, the training and validation datasets are used to optimize the number of neurons in the hidden layer as well as the weight matrices ($W_1$, $W_2$) and the bias vectors ($b_1$, $b_2$).
Comparison between ANN-GC model2 (this work) and published models (from literature) on the partition prediction of biomolecules in polymer-electrolyte ATPS.

<table>
<thead>
<tr>
<th>Biomolecule</th>
<th>ATPS</th>
<th>RMSD</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laccase</td>
<td>UCON 50-HB-5100 + KHCOO/-K2CO3, NaHCOO/-K2CO3 [159]</td>
<td>0.0284</td>
<td>NA</td>
</tr>
<tr>
<td>α-amylase</td>
<td>PEG1500 + Na2CO3 [83]</td>
<td>0.1466</td>
<td>NA</td>
</tr>
<tr>
<td>Amyloglucosidase</td>
<td>PEG6000 + KH2PO4 [45], PEG4000/PEG10000 + Na2SO4</td>
<td>0.1274</td>
<td>0.1813</td>
</tr>
<tr>
<td>Bovine serum albumin</td>
<td>PEG2000 + KH2PO4 [92]</td>
<td>0.0159</td>
<td>NA</td>
</tr>
<tr>
<td>Caffeine</td>
<td>PEG400 + Na2SO4 [165]</td>
<td>0.0405</td>
<td>NA</td>
</tr>
<tr>
<td>Cefalexin</td>
<td>Tween 20 + MgSO4/Na2SO4/Na2C4H4O4 [156]</td>
<td>0.0732</td>
<td>NA</td>
</tr>
<tr>
<td>Cefazolin</td>
<td>PEG6000/PEG800 + Na2SO4/Na2C4H4O4 [124], Tween 20 + MgSO4/Na2C4H4O4</td>
<td>0.0465</td>
<td>NA</td>
</tr>
<tr>
<td>Cephalexin</td>
<td>PEG1500/6000 + Na2C4H4O4 [167]</td>
<td>0.1205</td>
<td>0.2547</td>
</tr>
<tr>
<td>DNP-alanine</td>
<td>PEG4000/PEG6000/PEG800 + K2CO3, Na2C4H4O4/K2CO3/Na2C4H4O4 [168,169]</td>
<td>0.0324</td>
<td>NA</td>
</tr>
<tr>
<td>DNP-glycine</td>
<td>PEG4000/PEG6000/PEG800 + K2CO3, Na2C4H4O4/K2CO3/Na2C4H4O4 [168,169]</td>
<td>0.0277</td>
<td>NA</td>
</tr>
<tr>
<td>DNP-lysine</td>
<td>PEG4000/PEG6000/PEG800 + K2CO3, Na2C4H4O4/K2CO3/Na2C4H4O4 [168,169]</td>
<td>0.0194</td>
<td>NA</td>
</tr>
<tr>
<td>DNP-Valine</td>
<td>PEG4000/PEG6000/PEG800 + K2CO3, Na2C4H4O4/K2CO3/Na2C4H4O4 [168,169]</td>
<td>0.0189</td>
<td>NA</td>
</tr>
<tr>
<td>Human insulin</td>
<td>L62/L64/F68 + KH2PO4 [170]</td>
<td>0.0465</td>
<td>NA</td>
</tr>
<tr>
<td>l-lysine HCl</td>
<td>PEG4000/PEG10000 + KH2PO4/Na2C4H4O4 [171]</td>
<td>0.0829</td>
<td>0.2042</td>
</tr>
<tr>
<td>l-methionine</td>
<td>PEG4000 + NaH2PO4/Na2HPO4/Na2SO4 [172]</td>
<td>0.0637</td>
<td>0.0807</td>
</tr>
<tr>
<td>l-phenylalanine</td>
<td>PEG6000 + (NH4)2SO4/MgSO4/Na2SO4 [173]</td>
<td>0.0933</td>
<td>0.1166</td>
</tr>
<tr>
<td>l-tryptophan</td>
<td>PEG6000 + (NH4)2SO4/MgSO4/Na2SO4 [173]</td>
<td>0.0950</td>
<td>0.0822</td>
</tr>
<tr>
<td>l-tyrosine</td>
<td>PEG6000 + (NH4)2SO4/MgSO4/Na2SO4 [173]</td>
<td>0.0689</td>
<td>0.1612</td>
</tr>
<tr>
<td>Oligonucleotide</td>
<td>PEG6000/PEG10000/PEG1500/PEG4000 + Na2SO4 [174]</td>
<td>0.0088</td>
<td>0.0754</td>
</tr>
<tr>
<td>Penicillin G</td>
<td>PEG2000/PEG3500 + KH2PO4/Na2C4H4O4 [175]</td>
<td>0.0595</td>
<td>0.0647</td>
</tr>
<tr>
<td>β-amylase</td>
<td>PEG4000/PEG6000/PEG10000 + Na2SO4 [45], PEG6000/PEG10000 + KH2PO4 [176]</td>
<td>0.0302</td>
<td>0.0359</td>
</tr>
</tbody>
</table>

Squared correlation coefficient ($R^2$) and mean absolute error (MAE) versus the number of neurons in the hidden layer of ANN-GC model1.

Comparison between the experimental and ANN-GC model1 calculated weight fractions of polymer/copolymer ($w_{ex}$) in ATPS.

Error of experimental versus ANN-GC model1 calculated weight fractions of polymer/copolymer ($w_{ex}$) in ATPS.

In this work, the weight and bias parameters are optimized using the Levenberg-Marquardt algorithm (in MATLAB), in which the minimization of the summation of absolute errors between the experimental and model-calculated weight fractions of polymer/copolymer in the training dataset is applied as the optimization target. Results show that the network with 14 neurons in the hidden layer provides the best model performance, as illustrated in Fig. 3. The optimized values of the weight and bias parameters for the ANN-GC model1 with 14 neurons in the hidden layer are provided in Table S9 (see Supporting Information). Evaluating the model accuracy is an essential step in the process of developing machine learning models to describe how good the models are performing in their predictions. In this study, evaluation metrics including squared correlation coefficient ($R^2$) and mean absolute error (MAE) are used, as given by Eqs.5 and 6.

$$R^2 = \frac{\sum_{i=1}^{N} (w_{exp,i} - w_{pred,i})^2}{\sum_{i=1}^{N} (w_{exp,i} - \bar{w}_{exp})^2}$$  \hspace{1cm} (5)
The partition coefficient of biomolecules ($K_{bio}$) in ATPS is an important metric for evaluating the separation performance of ATPS and it is defined by Eq. 7. In this section, artificial neural network modeling will be performed on the partition prediction of biomolecules in polymer-electrolyte ATPS, and a three-layer feed forward network will also be used. Fig. 7 gives the schematic diagram of network modeling (ANN-GC model2) that targets for the partition prediction of biomolecules in polymer-electrolyte ATPS. In this ANN-GC model, the inputs contain the system’s temperature, the molecular weight of biomolecule, the structure (i.e. the molecular weight and the type of polymer/copolymer, the ion combination of electrolyte) and the TLL of polymer-electrolyte ATPS. Similar to the development of ANN-GC model1, the combination of tansig transfer function in the hidden layer and purelin transfer function in the output layer is also applied. Similarly, the number of neurons in the hidden layer, and the weight and bias parameters of ANN-GC model2 are optimized from the training and validation datasets. In this model, the minimization of the summation of absolute errors between the experimental and model calculated partition coefficients of biomolecules in polymer-electrolyte ATPS in the training dataset is set as the objective function.

$$K_{bio} = \frac{w_{bio}^T}{w_{bio}^B}$$

(7)

$$\frac{1}{K_{bio}} = \frac{w_{bio}^B}{w_{bio}^T}$$

(8)

Where $w_{bio}^T$ and $w_{bio}^B$ are the weight fraction of biomolecules in the top and bottom phases, respectively.

After numerous network modeling tries on the partition of 22 biomolecules in 42 polymer-electrolyte ATPS, the model consistently performs poorly with different number of neurons in the hidden layer. We find that this is mainly due to the large range of the $K_{bio}$ values of the studied biomolecule partitioning systems. For this reason, Database2 is classified based on the $K_{bio}$ values less or greater than 1 for the further dataset split. Then, $K_{bio}$ values that greater than 1 are replaced by $1/K_{bio}$ (see Eq.8). It means that the output of ANN-GC model2 is the element ($P_{bio}$) from set C, as described by Eq.9, which is a union of sets A and B. Obviously, the closer to zero an $P_{bio}$ is, the greater is the ability of polymer-electrolyte ATPS to partition biomolecule either in the top phase or the bottom phase. Therefore, we define $P_{bio}$ as the partition performance coefficient of biomolecules in ATPS. Using $P_{bio}$ as the output, ANN-GC model2 with 32 neurons in the hidden layer have the best model performance and it gives a root-mean-square deviation (RMSD) of 0.0577 for 501 training data points, and for the 62 validation data points and 63 test data points their RMSD are 0.0849 and 0.0885, respectively. The optimized values of the weight and bias parameters for ANN-GC model2 are provided in Table S10. Fig. 4 presents the comparison between the experimental and model calculated partition coefficients of all 22 biomolecules in 42 polymer-electrolyte ATPS for all training, validation, and test sets. Clearly, ANN-GC model2, to some extent, is able to predict the partition coefficient of biomolecules in polymer-electrolyte ATPS in a general way. The applicability of this ANN-GC model is further validated through the comparison between the ANN-GC model2 and other published models on the predictions of partition coefficient of biomolecules in polymer-electrolyte ATPS, as given in Table 3. An example of using ANN-GC model2 to predict the partition coefficient of biomolecules in polymer-electrolyte ATPS is provided in the Appendix. It should be noted that the output of ANN-GC model2 is $P_{bio}$, and therefore the output result cannot determine whether the biomolecules are enriched in the top phase or the bottom phase. Nonetheless, this ANN-GC model is able to predict the system’s partition performance as the closer to zero an output is, the greater is the ability of polymer-electrolyte ATPS to partition biomolecule.
\[ C = A \cup B \]  
\[ P_{\text{bio}} \in C, K_{\text{bio}} \in A, \frac{1}{K_{\text{bio}}} \in B \]  
\[ \text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (P_{\text{bio}}^{\text{exp}}, i - P_{\text{bio}}^{\text{cal}}, i)^2}{N}} \]  

In Eq.10, \( P_{\text{bio}}^{\text{exp}} \) and \( P_{\text{bio}}^{\text{cal}} \) are the experimental and model-calculated partition performance coefficient of biomolecules in ATPS for each data point, respectively. \( N \) represents the total number of data points in the training dataset or the test dataset.

In ANN-GC model2, the tie-line length of polymer-electrolyte ATPS is an important input, and it is also the only input that needs to be identified from the phase compositions. However, it would be time consuming and expensive to obtain this parameter from experimental works given the fact that the number of potential polymer-electrolyte ATPS is very large. Therefore, an effective way to obtain TLL of different polymer-electrolyte ATPS is quite important for maximizing the application potential of ANN-GC model2 in the partition prediction of biomolecules in these aqueous electrolyte solutions. Here we will investigate the possibility of using the ANN-GC model1 calculated TLL to replace the experimental TLL as an input to predict the partition coefficient of biomolecules in polymer-electrolyte ATPS. Experimental data (from literature) containing the information of the top and bottom phase compositions are selected (see Table S11) for such a purpose. Fig. 9 illustrates the comparison between the ANN-GC model2 calculated partition performance coefficients \( P_{\text{bio}} \) of biomolecules in polymer-electrolyte ATPS from the experimental and ANN-GC model1 calculated TLL. Results show that the TLL of polymer-electrolyte ATPS calculated from ANN-GC model1 can be used as acceptable alternatives when their experimental counterparts are not available.

5. Results and discussions

Modeling results indicate that ANN-GC model1 is capable to describe
well the binodal curve of polymer-electrolyte ATPS, as shown in Figs. 4-6. This model gives a mean absolute error (MAE) of 0.0132 and squared correlation coefficient ($R^2$) of 0.9878 for the 9,598 training data points, and for the 1,200 validation data points they are 0.0141 and 0.9858, respectively. Meanwhile, it also gives a MAE of 0.0143 and $R^2$ of 0.9846 for the 1,200 test data points. Most of the prediction errors from this model are close to zero and only a very limited number of predictions have absolute errors higher than 0.05. On the other hand, ANN-GC model2, to some extent, is able to predict the partition coefficient of biomolecules in polymer-electrolyte ATPS, as shown in Fig. 8. This model gives a RMSD of 0.0577 for the 501 training data points and a RMSD of 0.0849 for 62 the test data points, and for the 63 test data points the RMSD is 0.0885. Except the calculations on the system involving l-Tryptophan, all the rest of the calculations from ANN-GC model2 are better than those obtained by other methods from literature. Moreover, ANN-GC model2 can predict the partition of different biomolecules in various polymer-electrolyte ATPS in a general way, while the correlations and models from literature are limited to a single or some specific biomolecule-ATPS. In addition, the TLL of polymer-electrolyte ATPS calculated from ANN-GC model1 has proved to be an acceptable alternative for their experimental counterparts when using ANN-GC model2 to predict the partition coefficient of biomolecules in polymer-electrolyte ATPS.

In addition to the development of predictive models to describe the binodal curve of polymer-electrolyte ATPS and the partitioning behaviour of biomolecules in these aqueous biphasic systems, the effects of polymer, electrolyte, and system temperature on the phase formation of polymer-electrolyte ATPS and the partitioning of biomolecules in these ATPS are discussed using experimental data and modeling results. Fig. 10 shows an example of the influence from temperature on the phase formation of polymer-electrolyte ATPS. It is well known that the closer the binodal curve is to the axis origin, the greater the phase split capability of an ATPS. Obviously, higher temperatures are favourable to the phase formation of polymer-electrolyte ATPS, which is quite

![Fig. 12. Binodal curves of ATPS composed of Na$_2$C$_6$H$_5$O$_7$ and polymer PEG with different molecular weights at room temperature. Symbols are experimental data [88,183] and lines are calculated values from ANN-GC model1 developed in this work.](image1)

![Fig. 13. Binodal curves of ATPS composed of PEG4000 and different sodium salts at room temperature. Symbols are experimental data [69,88,177,184-186] and lines are calculated values from ANN-GC model1 developed in this work.](image2)

![Fig. 14. Binodal curves of ATPS composed of PEG4000 and sulfate salts at room temperature. Symbols are experimental data [187-192] and lines are calculated values from ANN-GC model1 developed in this work.](image3)

![Fig. 15. Binodal curves of ATPS composed of PEG4000 and sulfate salts at room temperature. Symbols are experimental data [187-192] and lines are calculated values from ANN-GC model1 developed in this work.](image4)

![Fig. 16. Binodal curves of ATPS composed of Na$_2$C$_6$H$_5$O$_7$ and polymer PEG with different molecular weights at room temperature. Symbols are experimental data [88,183] and lines are calculated values from ANN-GC model1 developed in this work.](image5)

![Fig. 17. Binodal curves of ATPS composed of PEG4000 and different sodium salts at room temperature. Symbols are experimental data [69,88,177,184-186] and lines are calculated values from ANN-GC model1 developed in this work.](image6)

![Fig. 18. Binodal curves of ATPS composed of PEG4000 and sulfate salts at room temperature. Symbols are experimental data [187-192] and lines are calculated values from ANN-GC model1 developed in this work.](image7)

![Fig. 19. Effect of temperature on biomolecular partitioning in polymer-electrolyte ATPS. Symbols are experimental data [45,166,170,175] and lines are calculated values from ANN-GC model2 developed in this work.](image8)
may not have a significant influence on the phase formation of some polymer-electrolyte ATPS (e.g. PEG6000 + Na2C4H3O7 + H2O).

Apart from the influence of temperature, the phase formation ability of polymer-electrolyte ATPS is also highly associated with the involved polymer and electrolyte. After comparing polymer-electrolyte ATPS that composed of different polymers, we find that the type of polymers has very limited impact on the phase formation of polymer-electrolyte ATPS. Fig. 11 gives three examples of the influence of the polymer type on the phase formation of polymer-electrolyte ATPS involving polymers PEG, PEGDME and PEO. On the other hand, we also find that polymers with higher molecular weights tend to have greater phase splitting ability for polymer-electrolyte ATPS. This can be explained by that the hydrophobicity of a polymer generally increases with the increase of its molecular weight, resulting in a higher tendency toward salting out from water solution. However, the polymer molecular weight increase may have no obvious impact on the formation of these aqueous biphasic systems when the polymer is within a certain molecular weight range. An example of this trend is illustrated by ATPS composed of Na2C4H3O7 and polymer PEG with different molecular weights (i.e. 2000, 3000, 4000, 6000) at system’s temperature of 298.15 K, as shown in Fig. 12.

Similar to many other types of ATPS, both the cation and anion of the phase-forming agent also have a great influence on the phase formation of polymer-electrolyte ATPS. The amount of electrolyte required to produce ATPS significantly depends on the salting-out strength of the electrolyte anion. An example of this trend is illustrated by ATPS composed of Na2C4H3O7 and NaH2PO4 with different electrolyte anions (i.e. K+, Na+, Mg2+ and Ca2+) at temperature of 298.15 K, as shown in Fig. 13. As mentioned above, the closer to zero a Pbio is, the greater is the ability of polymer-electrolyte ATPS to partition biomolecule. The partition performance of L64 + KH2PO4 + H2O for insulin increases significantly with increasing temperature, while the effect of temperature on the partitioning of cefazolin in Tween + Na2C4H3O7 + H2O is small. The ability of PEG + Na2SO4 + H2O to partition β-amylase increases obviously first and then decreases gradually with the increase of temperature. On the contrary, the partition performance of PEG + KH2PO4 + H2O for penicillin decreases greatly first and then increases slightly with increasing temperature. On the other hand, increasing the tie-line length (TLL) of ATPS promotes the partitioning of β-amylase and penicillin in PEG + Na2SO4/ KH2PO4 + H2O. This increase has, however,
much smaller impact on the partitioning of cefazolin in Tween + Na3C6H5O7 + H2O and insulin in L64 + K3PO4 + H2O, as presented in Fig. 16. Overall, greater TLLs favour the partitioning of biomolecules in polymer-electrolyte ATPS.

In addition to temperature and the TLL of ATPS, the molecular weight of both polymers and biomolecules also lead to changes on the partitioning of biomolecules in polymer-electrolyte ATPS. As shown in Fig. 17, modeling results indicate that when the polymer molecular weight increases, the partitioning of biomolecules β-amylose (in PEG + Na2SO4 + H2O), penicillin (in PEG + KH2PO4 + H2O) and insulin (in L64 + K3PO4 + H2O) increases. Meanwhile, only the partitioning of cefazolin in Tween + Na3C6H5O7 + H2O decreases with increasing the molecular weight of polymers. On the other hand, the ability of Tween + Na3C6H5O7 + H2O to partition biomolecules decreases first and then increase with the increase of the molecular weight of biomolecules, as illustrated in Fig. 18. For PEG + Na2SO4 + H2O, the partitioning ability of ATPS decreases when the molecular weight of biomolecules increases. However, biomolecules with higher molecular weights are favourable to increase in the molecular weight of biomolecules. As discussed in Fig. 14, when the polymer molecular weight increases, the partitioning of biomolecules increases. This may be attributed to the fact that the hydrophobicity of biomolecules generally increases with the increase of their molecular weight, and therefore promoting their partition in the polymer-rich phase.

The ANN-GC model1 is able to describe the binodal curve of polymer-electrolyte ATPS and the ANN-GC model2 can predict the partition of different biomolecules in various polymer-electrolyte ATPS in a general way. It should be noted that unlike the conventional models such as NRTL and UNIFAC, the proposed ANN-GC models are not derived from thermodynamic mechanisms/principles and therefore it is not possible to use them to explain the phase formation mechanisms of ATPS or the biomolecule partition mechanisms in ATPS. Furthermore, being data-driven models, ANN-GC models have limited predictive scalability, and their predictive performance is highly dependent on the quality and quantity of experimental data included. In this work, the polymers of most studied ATPS are PEGs and the molecular weights of most biomolecules are less than 1000 (g/mol). This means that the ANN-GC model2 may present poor predictions for the partitioning of biomolecules with molecular weights above 1000 in ATPS formed by non-PEG polymers. For this reason, experimental data on the partitioning of large biomolecules in non-PEG-based ATPS are highly desirable for improving the predictive reliability of ANN-GC model2 for these systems. Nonetheless, the ANN-GC model2 enable us to predict the partitioning of biomolecules in polymer-electrolyte ATPS under different conditions. Moreover, this machine learning model allows us to understand the trends of the partitioning properties of specific ATPS with the changes of temperature, tie-line length, and molecular weight of polymers and biomolecules. This can be used to guide the experimentalists to obtain high-performance ATPS for different biomolecules (as presented in Figs. 15-18). In addition, GC models can be easily integrated in the computer-aided design method, which makes it possible to apply this cost and time-effective technique to design optimal polymer-electrolyte ATPS for the separation and purification of specific biomolecules. Some examples of using this design method are given in Table 4. We hope that the promising performance of these proposed ATPS will invite further experimental validations.

6. Conclusion

A comprehensive database including 11,998 experimental binodal data in 276 polymer-electrolyte ATPS and 626 experimental partition data in 89 biomolecule-ATPS at different temperatures is established. Two machine learning models that combine ANN algorithm and GC method have been developed, respectively, in order to predict the phase composition of polymer-electrolyte ATPS and the partition performance coefficient of biomolecules in these aqueous electrolyte solutions. The modeling results show that ANN-GC model1 can give reliable predictions on the binodal curve behavior of polymer-electrolyte ATPS and ANN-GC model2, to some extent, is capable to predict the partition coefficient of biomolecules in these biphasic systems in a widespread way. In addition, the obtained results also indicate that the tie-line length of polymer-electrolyte ATPS calculated from ANN-GC model1 can be directly used in ANN-GC model2 for predicting the partition coefficient of biomolecules in these ATPS.

Besides the development of the ANN-GC models, some major issues affecting the partitioning of biomolecules in polymer-electrolyte ATPS are also discussed from the modeling studies. It is found that greater TLLs favor the partitioning of biomolecules in these aqueous electrolyte solutions. However, there are no general rules to describe the influence of temperature and the molecular weight of polymers and biomolecules. This is due to these factors generally presenting very different effects on the partitioning of biomolecules in different polymer-electrolyte ATPS. Nonetheless, the ANN-GC models enable us to have the knowledge on the partitioning trends of specific ATPS with the changes of these factors. This could serve as a guidance in the screening of high-performance ATPS to partition biomolecules. Moreover, when compared with the experimental method, these ANN-GC models largely reduce the time and cost to find high-performance ATPS for the separation and purification of specific biomolecules. Some high-performance ATPS that were identified from ANN-GC model calculations are proposed to partition four well-known biomolecules. In addition, these machine learning models can also provide reliable process design results, which would speed up the industrial implementation of ATPS. Besides these, although this work mainly focuses on the polymer-electrolyte ATPS, the proposed modeling strategy can also be used in the partitioning study of biomolecules in other types of ATPS such as the ionic liquid-salt ATPS and the alcohol-salt ATPS.

CRediT authorship contribution statement

Yuqiu Chen: Conceptualization, Methodology, Software, Investigation, Writing – original draft. Xiaodong Liang: Project administration, Writing – review & editing. Georgios M. Kontogeorgis: Supervision, Writing – review & editing, Resources.

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.

Data availability

Data will be made available on request.

Acknowledgements

The authors thank the European Research Council (ERC) for funding this research under the European Union’s Horizon 2020 research and innovation program (grant agreement no.832460), ERC Advanced Grant project “New Paradigm in Electrolyte Thermodynamics.”
\[ \text{PEG}6000 + \text{Na}_2\text{SO}_3 + \text{H}_2\text{O} \text{ at } 298.15 \text{ K}, \text{PEG}4000 + \text{Na}_2\text{MoO}_4 + \text{H}_2\text{O} \text{ at } 308.15 \text{ K}. \]

Another example of using ANN-GC model2 to calculate the partition performance coefficient of three biomolecules (i.e. oligonucleotide, bovine serum albumin, laccase) in polymer-electrolyte ATPS is also presented. The weight fraction of polymers (from ANN-GC model1) or the partition performance coefficient of biomolecules (from ANN-GC model2) can be obtained as an output:

\[
\text{Output} = W_2 \left( \frac{2}{1 + e^{-W_1P + b_1}} - 1 \right) - b_2
\]

The vector \( p \) and the mode parameters \((W_1, b_1, W_2, b_2)\) of ANN-GC model1 and ANN-GC model2 are, respectively, given in Table S9 and Table S10 (Supporting Information). Fig. A1 illustrates the comparison between the experimental and ANN-GC model1 calculated binodal curves of the exemplified polymer-electrolyte ATPS. Meanwhile, the experimental and ANN-GC model2 calculated partition performance coefficient of the exemplified biomolecules are given in Fig. A2.

**Fig. A1.** Example of using ANN-GC model1 to calculate the binodal curve of polymer-electrolyte ATPS. Symbols are experimental data [177,185,193] and lines are ANN-GC model1 calculated binodal curves.

**Fig. A2.** Example of using ANN-GC model2 to calculate the partition performance coefficient of biomolecules in polymer-electrolyte ATPS [92,159,174]. (Olig: oligonucleotide, BSA: bovine serum albumin, Lac: laccase).

References


Y. Chen et al.


