Received: 7 September 2023 | Revised: 14 November 2023 | Accepted: 20 November 2023

DOI: 10.1002/elps.202300198

ELECTROPHORESIS

RESEARCH ARTICLE

A neural network model for rapid prediction of analyte focusing in isotachophoresis

Amit Jangra[1,2] | Shaurya Shriyam[1,3] | Juan G. Santiago[4] | Supreet Singh Bahga[1]

[1]Department of Mechanical Engineering, Indian Institute of Technology Delhi, New Delhi, India

[2]Government Polytechnic, Hisar, Haryana, India

[3]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, India

[4]Department of Mechanical Engineering, Stanford University, Stanford, California, USA

Correspondence
Supreet Singh Bahga, Department of Mechanical Engineering, Indian Institute of Technology Delhi, New Delhi, 110016, India.
Email: bahga@mech.iitd.ac.in

Funding information
Science and Engineering Research Board, Grant/Award Numbers: CRG/2022/002148, MTR/2022/000080

Abstract
We present the development and demonstration of a neural network (NN) model for fast and accurate prediction of whether or not a chosen analyte is focused by an isotachophoresis (ITP) buffer system. The NN model is useful in the rapid evaluation of possible ITP chemistries applicable to analytes of interest. We trained and tested the NN model for univalent species based on extensive data sets of over 10,000 anionic and 10,000 cationic ITP simulations. The NN model uses as inputs the mobilities and the acid dissociation constants of leading electrolyte ion, trailing electrolyte ion, counterion, and a single analyte as well as the leading-to-counterion concentration ratio of the leading zone. The output then indicates whether the chosen electrolyte system yields stable ITP focusing of the analyte. The prediction accuracy of the NN model is over 97.7%. We demonstrate the applicability of the NN by validating its predictions with reported experimental data for anionic and cationic ITP. We have packaged the NN model in a free, web-based application named IONN (isotachophoresis on neural network), which can be used to rapidly screen ITP electrolyte systems.

KEYWORDS
isotachophoresis, machine learning, neural network

1 | INTRODUCTION

Isotachophoresis (ITP) is a nonlinear electrophoresis technique that can be used to purify, separate, and/or preconcentrate ionic species in a sample mixture into distinct zones based on their electrophoretic mobilities [1–3]. In ITP, ionic analytes are introduced in a capillary or a microchannel between zones of a leading electrolyte (LE) and a trailing electrolyte (TE). The co-ions of LE and TE are chosen to have higher and lower effective mobility magnitudes than the analytes, respectively. Upon application of an external electric field through the capillary, the analytes focus and segregate into distinct zones in order of their mobilities. On the other hand, analytes whose effective mobilities are outside of the range targeted by the LE and TE buffers do not focus. In addition to the separation and identification of species, ITP has been applied to a wide range of applications, including preconcentration prior to other analytical techniques, DNA and RNA purification, single-cell analyses, and acceleration of chemical reactions [3].

ITP is a robust electrophoresis technique because the nonlinearity in the electromigration flux causes the zone boundaries to self-sharpen and counteract diffusion [4, 5]. This is unlike capillary zone electrophoresis (CZE), wherein the analyte zones diffuse continuously during separation, and the zone dispersion is typically irreversible. However, ITP's robustness comes at the cost

Abbreviations: ITP, isotachophoresis; LE, leading electrolyte; MLP, multilayer perceptron; NN, neural network; TE, trailing electrolyte; ZED, zone existence diagram.

of a relatively complicated choice of the discontinuous electrolyte system of ITP consisting of LE and TE (e.g., compared to CZE, which uses a single background electrolyte). The success of ITP depends on the proper choice of LE and TE ions and the common counterion in LE and TE to ensure that stable analyte zones form between the LE and TE zones with self-sharpening zone boundaries. We here refer to an ITP analyte zone as "stable" if a fixed, finite injection of the analyte results in a steady analyte concentration, and the analyte is focused between the TE and LE zones with self-sharpening zone boundaries. For the case of constant applied current in a uniform cross-section channel, the width of the zone and its boundaries also reach steady-state values [3]. The difficulty in choosing the proper electrolyte system for stable ITP focusing of a given set of analytes is likely the primary barrier to the adoption of ITP by beginners and students.

The choice of electrolyte systems in ITP often depends on empirical experience and electrolyte systems recommended in the literature [6–8]. Some formal strategies for choosing ITP electrolytes have also been reported in the literature based on the zone existence diagrams (ZED), which are used to visualize the dependence of and interrelation among effective mobilities of LE, TE, and analyte ions with the pH [9, 10]. However, due to the nonlinear dynamics of ITP, the pH and effective mobilities of the ionic species vary spatially and temporally during the ITP process. Consequently, the pH and the order of effective mobilities of various species during the ITP process are not known *a priori*, limiting the applicability of ZEDs in predicting whether the analyte mobilities will lie between those of LE and TE ions during the separation. The construction of ZEDs also requires significant knowledge of ITP principles.

Over the years, numerical simulations have become a preferred approach to select electrolyte systems for suitable ITP focusing of analytes. The most useful ITP simulators are based, at least in part, on numerical solutions of some form of the coupled set of equations for species transport, current continuity, electroneutrality, and acid–base equilibria in an electrolyte [11–14]. Included among these are useful and simplified simulations that neglect the details of the diffused interfaces of plateau-shaped zones in ITP and apply integral conservation laws across ITP zone boundaries. The latter integral approaches can be used to predict steady-state ITP zone conditions and verify ITP focusing conditions [1, 2, 10, 15–18]. On the other hand, modern electrophoresis simulation tools such as SIMUL [19], SPRESSO [20], SPYCE [21], and CAFES [22] allow simulation of complex time-dependent ITP dynamics. While existing ITP simulation techniques can simulate analyte focusing in ITP in a matter of minutes, the electrolyte selection process is still time-consuming, given the numerous combinations of LE ions, TE ions, and counterions available for screening. Moreover, performing and post-processing the simulations require a basic understanding of numerical methods, the simulation parameters such as time steps, number of grid points, grid-refinement parameters, and setting up the initial conditions.

Often, ITP practitioners are interested in quickly screening electrolyte systems from a large number of combinations of LE ions, TE ions, and counterions, which can then be analyzed in detail using simulations or experiments. For such preliminary "triage" of candidate electrolyte systems for ITP, there is a need for a computational tool that can quickly predict whether a given electrolyte system will lead to stable ITP focusing of a particular analyte without performing detailed simulations. Advances in machine learning (ML) algorithms have enabled the development of models trained over a large number of simulations and/or experimental data sets, which can have the potential to efficiently explore a wide range of process parameters to find high-performing designs [23, 24]. ML models trained on simulation data can also supplement existing simulation tools rather than substituting the first-principles simulation techniques. Despite the availability of various ITP simulators, currently, no computational tool exists that leverages ML models to quickly select electrolyte systems for stable ITP.

This paper presents a simple-to-use web-based application that quickly and accurately predicts whether a particular combination of LE, TE, and analyte results in stable ITP focusing. The application is based on a neural network (NN) model trained on extensive simulation databases of anionic and cationic ITP simulations with varying mobilities and acid dissociation constants (pK_a) of LE, TE, and analytes, and the relative concentration of the LE ion and the background counterion. The web application is named IONN (isotachophoresis on neural network) and can be accessed at https://web.iitd.ac.in/~bahga/IONN.html. IONN is available for free use through web browsers, including those on mobile devices. The application includes a database of over 200 anionic and cationic LE and TE co-ions and counterions, in addition to the ability to define a custom user-defined species. The only inputs to IONN are the fully ionized (limiting) mobilities and pK_a of LE, TE, counterion, and the analyte, and the ratio of background counterion and LE ion concentrations. The trained NN gives a nearly instantaneous prediction of whether the analyte will focus with the chosen electrolytes. Currently, the NN model employed by IONN is limited to analyzing buffered electrolyte systems with univalent species, and, in the future, its capability will be extended to multivalent species. In this paper, we report the development of the NN model for anionic and cationic ITP based on extensive numerical simulations and its validation with published experimental data.

2 | MATERIAL AND METHODS

In this work, we use an NN model trained on extensive databases of anionic and cationic ITP simulations to make fast and accurate predictions of ITP focusing of analytes. NN models are a class of ML algorithms inspired by the biological functioning of neurons that can recognize patterns from a large data set [23]. NNs have found widespread applications in clustering, classification, and regression using experimental and simulation data. In the current work, an NN is trained on the simulation data for anionic and cationic ITP for varying values of input parameters, which include limiting mobilities and pK_a of LE and TE co-ions, counterion, and the analyte, and the ratio of concentrations of background counterion and LE ion. We formulate the problem of prediction of stable ITP focusing as a supervised learning problem for binary classification. In this approach, the output of high-fidelity simulations corresponding to each set of input parameters is labeled into two classes based on whether or not stable ITP zones form. The data set is then randomly split into two subsets: the training and testing of data sets. The NN model is then trained by minimizing the error in predicting stable and unstable ITP zones compared with the training data set. Finally, we measure the accuracy of the trained NN by comparing its predictions with the testing data sets for anionic and cationic ITP.

The benefit of an NN over a full-scale ITP simulation is that, once the NN is well-trained, the prediction of whether or not stable ITP zones form is extremely fast and computationally efficient. This makes it possible to deploy the NN model as a web-based application. Moreover, due to its computational efficiency, the NN model can be used to quickly explore numerous combinations of LE ion, TE ion, and counterion for stable ITP focusing of given analytes. In this section, we describe the various steps involved in the training and testing of the NN model. We also describe the preparation of the simulation databases, the NN architecture, and the methods for training and validation.

2.1 | Database preparation

In any supervised ML problem, building the predictive model, NN in our case, requires a comprehensive data set for training over various possible scenarios. To make the NN model produce accurate predictions, we must have high-fidelity data for training and testing of the model. We generated data sets for anionic and cationic ITP focusing of a single analyte using numerical simulations for well-buffered electrolyte systems. For such electrolyte systems, hydronium and hydroxyl ion concentrations are sufficiently small such that they have a negligible effect on the electromigration flux of other species and contribute negligibly to the total current [3]. In particular, we considered the most widely used approach for pH buffering of ITP zones, wherein the LE counterion (which migrates from the LE zone to the TE zone) serves as the buffering ion, and where the LE, TE, and analyte ions are the titrants. For all our simulations, we considered a typical composition of electrolytes with initial concentrations of LE ion, TE ion, and analyte to be 10, 5, and 1 mM, respectively. We varied the concentration of the background counterion from 15 to 30 mM to incorporate the effect of varying LE composition in the NN model.

In the current work, we limited the training of the NN to univalent species only. Therefore, each data point in the data set was characterized by the limiting (fully ionized) ionic mobilities (μ) and acid-dissociation constants (pK_a) of LE and TE ions, the analyte co-ion, and counterion, and the ratio of concentrations of background counterion and LE ion (c_{BG}/c_{LE}). That is, there were nine input parameters (or features) based on which we predicted whether stable ITP zones formed or not. For all the simulations, we also neglected the effects of the ionic strength on mobility and pK_a of species as the ionic strength rarely changes the relative order of zones, particularly for univalent valences [25].

To generate the simulation data sets for anionic and cationic ITP, we performed 30,000 simulations each for varying values of the nine input parameters. All the simulations were performed using an in-house diffusion-free solver written in Python 3 programming language. The solver was validated with transient, one-dimensional numerical simulations using the SPYCE simulator [21, 26] prior to the generation of the data set. The values of mobilities and pK_a of the species were sampled randomly from the corresponding uniform distributions with the ranges mentioned in Table 1. The ratio of concentrations of the background counterion and the LE ion (c_{BG}/c_{LE}) was sampled from a uniform distribution between 1.5 and 3. Because the desired output of the simulations was whether the analyte focuses or not using the chosen ITP electrolytes, we used a relatively fast diffusion-free simulation approach [1, 18], for predicting the zone concentrations and stability. The diffusion-free approach neglects molecular diffusion and applies species conservation and current continuity across the sharp zone boundaries of ITP while accounting for local acid–base equilibria. The simulations yield the concentrations, pH, and effective mobilities of the species in the LE, analyte, and TE zones (upon adjusting to a new concentration).

Finally, to verify the stability of ITP zones, we check whether the ITP stability conditions are satisfied [3]. The first set of conditions is that the analyte (A) must have

TABLE 1 The range of physiochemical properties of species used for the simulations to generate the data sets. The mobilities and pK_a were sampled randomly from the corresponding uniform distributions over the ranges mentioned here.

Species	Anionic ITP		Cationic ITP	
	Mobility (μ) (10^{-9} m^2 V^{-1}s^{-1})	**pK_a**	**Mobility (μ) (10^{-9} m^2 V^{-1}s^{-1})**	**pK_a**
LE ion	[−85, −20]	[−2, 8]	[20, 85]	[6, 14]
TE ion	[−85, −20]	[−2, 10]	[20, 85]	[4, 14]
counterion	[20, 85]	[4, 10]	[−85, −20]	[4, 10]
Analyte	[−85, −20]	[−2, 10]	[20, 85]	[4, 14]

Abbreviations: ITP, isotachophoresis; LE, leading electrolyte; TE, trailing electrolyte.

a lower mobility magnitude than the LE ion (L) in the analyte and LE zones,

$$|\bar{\mu}_A^L| < |\bar{\mu}_L^L| \text{ and } |\bar{\mu}_A^A| < |\bar{\mu}_L^A|. \tag{1}$$

Here, $\bar{\mu}$ denotes the effective mobility, and the subscript denotes the identity of the ion. The superscripts A and L denote the analyte and LE zones (locations), respectively. This condition ensures that the interface separating the LE and analyte zones is self-sharpening. Next, to ensure the stability of the interface separating the adjusted TE zone and the analyte zone, we must have,

$$|\bar{\mu}_A^T| > |\bar{\mu}_T^T| \text{ and } |\bar{\mu}_A^A| > |\bar{\mu}_T^A|. \tag{2}$$

Consistently, the subscript T here denotes the TE ion, and the superscript T denotes the adjusted TE zone [3]. Lastly, for stable ITP, the TE ion mobility magnitude must be lower than that of the LE ion in the LE and adjusted TE zones,

$$|\bar{\mu}_L^T| > |\bar{\mu}_T^T| \text{ and } |\bar{\mu}_L^L| > |\bar{\mu}_T^L|. \tag{3}$$

If all these stability conditions were met by the zones computed by the diffusion-free model, we assigned the combination of electrolytes and the analyte to a class with label 1, corresponding to stable ITP focusing. Otherwise, we assigned the choice of electrolytes and analytes to class 0, corresponding to a violation of ITP focusing conditions. Therefore, each data point in the databases for anionic and cationic ITP consisted of the nine input features (μ and pK_a of species and c_{BG}/c_{LE}) and the corresponding simulated class label (0 or 1).

Because stable ITP focusing occurs for a restrictive choice of input parameters, the databases for anionic and cationic ITP had a large imbalance in the number of cases for stable and unstable ITP. Such imbalance in the data set can lead to an NN model becoming biased towards the majority class [24]. Therefore, we performed undersampling by randomly removing the simulated data points for cases with no ITP focusing to ensure an equal number of both classes (0 and 1) in the data sets. After undersampling, the databases for anionic and cationic ITP had 10 950 and 11 302 data points, respectively, with an equal number of data points for each class.

2.2 | Neural network model

We used the data sets generated by running ITP simulations to train and test an NN model for anionic and cationic ITP. All NN models have an artificial neuron or node as their building block. Many of these nodes are interconnected, and these connections are associated with specific weight parameters. When a node receives input signals from the upstream nodes, the inputs get modified by the weights corresponding to the interconnections. These modified inputs are then summed up, and a bias value is added to the sum to obtain the output. The output of every node (termed activation) is typically modified using a nonlinear function called the activation function, which mimics the firing mechanism of a biological neuron. In this work, we use a multilayer perceptron (MLP) NN, as shown in Figure 1, which is a widely used NN architecture for

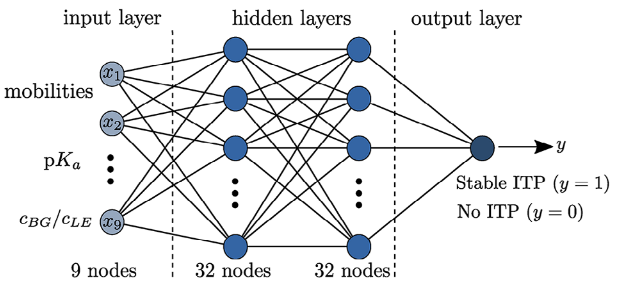

FIGURE 1 Schematic of the neural network (NN) based on MLP architecture for predicting ITP focusing. The NN consists of an input layer with nine nodes, an output layer with a single node, and two hidden layers with 32 nodes each. The nine input features of the NN are the mobilities and pK_a of the univalent LE ion, TE ion, counterion, and the analyte, and c_{BG}/c_{LE}. The NN outputs 1 or 0 corresponding to the prediction of whether the analyte focuses between LE and TE zones or not, respectively.

representing a nonlinear mapping between several inputs and a single output. The MLP consists of nodes arranged in three types of layers: (i) the input layer that receives the model inputs, (ii) the output layer that gives the output of the model (0 or 1 for binary classification), and (iii) hidden layers whose output is not accessible outside the network.

2.2.1 | Neural network architecture

Figure 1 shows a schematic of the MLP architecture used in the current work. The MLP consists of two hidden layers, each having an equal number of nodes, $N = 32$, in addition to the input and output layers. Therefore, the input and output layers are the first and fourth layers of the MLP, while the second and third layers are the hidden layers. Each node x_i of the input layer represents the input features such as mobilities, acid dissociation constants, and the counterion-to-LE concentration ratio. The N nodes of the second layer (first hidden layer) transform the input features to

$$z_i^{(2)} = \sum_{j=1}^{9} w_{ij}^{(2)} x_j + b_i^{(2)}, \quad i = 1, \dots, N. \tag{4}$$

Here $w_{ij}^{(2)}$ denote the weights for the second layer acting on the input features x_j, and $b_i^{(2)}$ are the bias values. To model the nonlinear dependence of the output on the inputs, the linear combinations $z_i^{(2)}$ are transformed using a nonlinear activation function, $g^{(2)}(z)$, to get the activations (outputs) $a_i^{(2)}$ of the nodes of the second layer. We use the rectified linear unit (RELU) activation function, which yields the following activations:

$$a_i^{(2)} = g^{(2)}(z_i^{(2)}) = \max(0, z_i^{(2)}), \qquad i = 1, \dots, N. \tag{5}$$

Following the same procedure, the activations of the second layer nodes are used to obtain the activations of the third layer nodes. In the current work, we use the same nonlinear activation function (RELU) and the same number of nodes N for all the hidden layers (layers 2 and 3). Therefore, the activation $a_i^{(3)}$ of ith node of the third layer is given by

$$a_i^{(3)} = g^{(3)}\left(\sum_{j=1}^{N} w_{ij}^{(3)} a_j^{(2)} + b_i^{(3)}\right), \qquad i = 1, \dots, N. \tag{6}$$

Similarly, we obtain the output y of the single node in the output layer using the activations of the third layer nodes. However, we use a logistic sigmoid function as the activation function for the output node. If the resulting value is greater than a threshold (0.5), the final output of the NN is assigned to the class with label 1 corresponding to stable ITP; otherwise, the output is set to $y = 0$. That is, the network output y for binary classification is given by

$$y = g^{(4)}\left(z^{(4)}\right), \qquad z^{(4)} = \sum_{j=1}^{N} w_{1j}^{(4)} a_j^{(3)} + b^{(4)}. \tag{7}$$

Here, the activation function $g^{(4)}$ is defined as

$$g^{(4)}(x) = H(\sigma(x) - 0.5), \qquad \sigma(x) = \frac{1}{1 + e^{-x}}, \tag{8}$$

where H is the Heaviside function and σ is the logistic sigmoid function. Therefore, the output y takes on values of 1 or 0 depending on whether the input features correspond to stable or unstable ITP, respectively. On the other hand, the output of the sigmoid function $\sigma(z^{(4)})$ can be interpreted as a probability-like score that provides confidence in the predictions. That is, the values of $\sigma(z^{(4)})$ close to 1 and 0 correspond to a higher likelihood of stable and unstable ITP, respectively.

The data-fitted values for the weights and biases for all the layers are computed using an optimization approach by minimizing the error between the model's predicted values and the ground-truth values for the training data set. In particular, the training of an NN model begins by assigning random values to the weights and biases. The outputs predicted by Equation (7) are compared with the corresponding labels in the training data set, and the resulting error is calculated. Based on the error, the weights and the biases are updated using the backpropagation algorithm [23, 27]. This procedure is iteratively repeated to update the weights and biases until the error is contained within an acceptable tolerance limit.

2.2.2 | Model implementation, training, and validation

We implemented and trained the NN model in Python 3 programming language using the MLPClassifier function and Adams optimizer available in the Scikit-learn library [28]. In particular, we optimized a separate set of weights and biases for anionic and cationic ITP using the respective data sets. We used the GridSearchCV function for cross-validation purposes and for tuning the hyperparameters of NN. In Figure S1 of the Supporting Information, we show the model training and validation workflow. We began by preprocessing the data by scaling the input features between 0 and 1. Next, we split each simulation data set into training and testing sets, with 80% and 20% data from the original data set. The training set was used to optimize the parameters of the NN using the cross-entropy loss function, and the test set was used for the final evaluation of the model. We used fivefold cross-validation on the training set to ensure that training leads to a generalized NN model. That is, the training set was partitioned into five subsets, and multiple training rounds were done by rotating between four subsets for training and the remaining subset for validation. We also used the grid-search-based hyperparameter tuning method of the Scikit-learn library to tune the hyperparameters (parameters whose values remain constant during the training), such as the number of hidden layers, the number of nodes in each hidden layer, and the regularization parameter. Hyperparameter tuning suggested the MLP architecture with two hidden layers having 32 nodes each.

The trained NN model yields a rapid prediction of ITP zone stability. A single prediction using the NN model takes, on average, 0.15 ms on a personal computer (AMD Ryzen 7 6800H, 16 GB RAM), which is over 400 times faster than a single calculation based on the diffusion-free model. By comparison, the computational time for a typical calculation based on the diffusion-free model is approximately 60 ms. Hence, the NN model has the potential for exploring and comparing among numerous combinations of LE ion, TE ion, and counterions. For example, consider the evaluation of ITP chemistries for a single analyte. Given 10 choices each of LE ion, TE ion, and counterion at specified (fixed) concentrations, the NN model can make 1000 predictions in less than a second on a personal computer.

The final performance of the trained NN model was evaluated using the unseen test data set. The evaluation was based on the usual performance metrics, including accuracy, precision, recall, and F_1-score [24]. These metrics are based on the number of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*). The accuracy (*ACC*) is defined as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{9}$$

Here, positive and negative cases correspond to stable ITP and no ITP, respectively. The accuracy (*ACC*) represents the fraction of cases for which a correct prediction of ITP (whether stable or unstable) is made. While accuracy is a good metric for balanced data with similar data points for both classes in a binary classification problem, we also calculated the precision and recall metrics. Precision or the positive predictive value (*PPV*) is defined as

$$PPV = \frac{TP}{TP + FP}, \tag{10}$$

and recall or the true positive rate (*TPR*) metrics is defined as

$$TPR = \frac{TP}{TP + FN}. \tag{11}$$

PPV measures how many predictions of a particular class (stable or no ITP) by the NN model actually belong to the same class. Additionally, *TPR* measures the fraction of correct predictions of a particular class by the NN model out of all the class cases in the testing data set. For a perfect predictive model, *PPV* and *TPR* should be unity. The *PPV* and *TPR* scores are usually combined in the form of a geometric mean to get the F_1 score,

$$F_1 = 2 \times \frac{PPV \times TPR}{PPV + TPR}. \tag{12}$$

2.2.3 | Web application

We packaged the final NN model into a web application using the Flask framework in Python 3. Figure S2 of the Supporting Information shows the graphical user interface (GUI) of the application, which we named IONN. The GUI allows the users to input the mobilities and pK_a of LE and TE co-ions, the counterion and the analyte, and the counterion to LE concentration ratio (c_{BG}/c_{LE}). The GUI then calls the NN model to predict whether or not stable ITP focusing of the analyte will occur, along with a probability-like score. The IONN application allows the users to input the mobilities and pK_a from a database of commonly used species, in addition to custom user-defined species. The NN model should preferably be used for input values within their respective ranges used for the model training. That is, the mobility and pK_a values should preferably lie within the bounds given in Table 1 and c_{BG}/c_{LE} should be chosen between 1.5 and 3. Therefore, the GUI suggests these ranges to the user during input, although the user can override these suggestions.

3 | RESULT AND DISCUSSION

This section presents the results of testing the NN model with the test data sets and experimental observations of anionic and cationic ITP.

3.1 | Model testing

First, we tested the trained NN model using the testing data sets for anionic and cationic ITP with 2190 and 2261 test cases, respectively. In Figure 2, we present the results

TABLE 2 The results of testing the NN model for anionic and cationic ITP cases in terms of precision (*PPV*), recall (*TPR*), and F_1 scores. The classes 0 and 1 correspond to no ITP and stable ITP, respectively.

Classes	Anionic ITP			Cationic ITP		
	Precision (*PPV*)	Recall (*TPR*)	F_1	Precision (*PPV*)	Recall (*TPR*)	F_1
0	0.99	0.97	0.98	0.99	0.97	0.98
1	0.97	0.99	0.98	0.97	0.99	0.98

Abbreviations: ITP, isotachophoresis; PPV, positive predictive value; TPV, true predictive value.

of testing the NN model for binary classification in the form of confusion matrices. The diagonal terms of a confusion matrix represent the number of correctly classified cases (true negatives and true positives). In contrast, the off-diagonal terms correspond to the number of misclassified cases (false negatives and false positives). The data presented in Figure 2 yields an accuracy of 97.7% and 98.0% for anionic and cationic ITP, respectively. The precision, recall, and F_1 scores, presented in Table 2, suggest that the trained NN model accurately predicts the electrolyte system and analyte combination that results in stable ITP zones.

We also analyzed the small number of misclassified cases and identified two primary types of incorrect (false) predictions by the NN model. These types of false predictions can be described in terms of the thermophysical parameters governing the physics of the problem. The majority of misclassified cases involved pK_a of one or more co-ions (LE, TE, and analyte ions) close to the pH of one or more ITP zones, with the difference between pK_a and pH less than 0.5 pH units. The effective mobilities of such co-ions are most sensitive to the pH, which led to incorrect prediction of ITP focusing conditions by the NN model. The second type of misclassification was associated with extreme values of either one or more of the mobility and pK_a of the species. This type of misclassification is associated with values of these parameters that were close to the limits of these physical parameters under which the NN model was trained. Irrespective of the type of misclassification, the probability-like scores given by the NN model indicated a degree of uncertainty in the prediction for most misclassified cases. A detailed analysis of the misclassified cases and their relation to the physical parameters of the problem is presented Tables S2 and S3 of the Supporting Information.

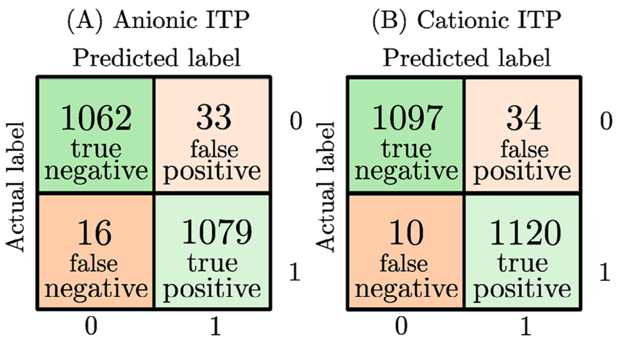

FIGURE 2 Results of testing the NN model for anionic and cationic ITP presented in the form of confusion matrices. Label 0 corresponds to the violation of ITP focusing conditions, and label 1 corresponds to the formation of stable ITP zones. We tested the NN model with 2190 and 2261 test cases for anionic and cationic ITP, respectively. The confusion matrices show that the trained NN model accurately predicts ITP focusing.

3.2 | Validation with experimental data

In addition to testing the performance of the NN using the simulated data, we also validated the NN model using published experimental data for anionic and cationic ITP. The mobility and pK_a values for all the species used for validation are provided in Table S1 of the Supporting Information. For anionic ITP, we considered the experiments of Chambers et al. [29], where an anionic nonfocusing fluorescent tracer, Alexa Fluor (AF488), was mixed with the TE to visualize the various zones in ITP. In the experiment, the LE ion was 100 mM MES, the TE ion was 100 mM tricine, and the counterion was 200 mM bis-tris ($c_{BG}/c_{LE} = 2$). The analytes were MOPS and HEPES, which formed stable zones between the LE and adjusted TE zones. In contrast, AF488 did not focus between the LE and TE zones. In Table 3, we compare the predictions of the NN model with experimental observations. The NN model correctly predicts stable ITP focusing of MOPS and HEPES. Moreover, the model correctly predicts that AF488 will not focus between LE and TE zones.

Next, we compared the predictions of the NN model with the anionic ITP experiments of Everaerts et al. [6], wherein the LE ion was chloride, and the TE ion was MES. The buffering counterion was histidine, and the pH of LE was 6.02, corresponding to $c_{BG}/c_{LE} = 2$. Note that histidine has ionization states of −1, +1, and +2 with pK_a values of 9.33, 6.04, and 2.01, respectively. However, these multiple pK_a values are sufficiently spaced apart (in pH units) such that histidine behaves as an univalent weak base at the pH at which the experiment was performed. Therefore, even though our NN model was trained using only univalent species, the model can be applied to this electrolyte system. We predicted ITP focusing of five anionic analyte

TABLE 3 Comparison of the predictions of the NN model with published experimental data for anionic and cationic ITP.

ITP	LE	TE	$\frac{c_{BG}}{c_{LE}}$	Analyte	Experiment	NN prediction	Data source
Anionic	MES + Bis-tris	Tricine + bis-tris	2	MOPS	Focused	Focused	Chambers and Santiago [29]
				HEPES	Focused	Focused	
				AF488	Not focused	Not focused	
	Chloride + Histidine	MES + histidine	2	Perchloric acid	Focused	Focused	Everaerts et al. [6]
				Formic acid	Focused	Focused	
				Acetic acid	Focused	Focused	
				Lactic acid	Focused	Focused	
				Caproic acid	Focused	Focused	
	MOPS + imidazole	Taurine + imidazole	2.8	HEPES	Focused	Focused	Bahga and Santiago [30]
				Tricine	Focused	Focused	
Cationic	Ethanolamine + tricine	Tris + tricine	2	Lysine	Focused	Focused	Garcia-Schwarz et al. [8]
				Arginine	Focused	Focused	
				R6G	Not focused	Not focused	
	Sodium + HEPES	Pyridine + HEPES	2	Bis-tris	Focused	Focused	Bahga et al. [18]

Abbreviations: ITP, isotachophoresis; LE, leading electrolyte; TE, trailing electrolyte; NN, neural network.

ions: perchlorate, formate, acetate, lactate, and caproate ions. As shown in Table 3, the NN model for anionic ITP correctly predicts stable analyte zones for these ions, as observed in the experiment of Everaerts et al. [6].

We also tested the NN model with the data of anionic ITP experiments of Bahga and Santiago [30] for $c_{BG}/c_{LE} = 2.8$. In this experiment, the LE and TE ions were MOPS and taurine, respectively, and imidazole was the background counterion. In the pH range of this experiment, taurine can be modeled as a univalent acid despite having ionization states of −1 and +1. As shown in Table 3, the NN model correctly predicts the focusing of two analytes, HEPES and tricine.

To validate the NN model's capability to predict cationic ITP focusing, we considered the ITP experiments of Garcia-Schwarz et al. [8] for separating two amino acids (lysine and arginine). In the latter experiment, the LE ion was 100 mM ethanolamine, the TE ion was 20 mM tris, and the buffering counterion was tricine with concentrations of 200 and 40 mM in LE and TE, respectively. The ITP zones were visualized using Rhodamine 6G, which was a nonfocusing tracer. Even though arginine and lysine are multivalent amino acids, under the pH conditions of the experiment, both behaved as univalent weak bases. Hence, we can apply our NN model to this case. The comparison of model predictions and experimental observations in Table 3 shows that the NN model correctly predicts that arginine and lysine will focus between LE and TE zones, and Rhodamine 6G will not focus for the chosen electrolyte system.

Lastly, we compared the predictions of the NN model with the data for the cationic ITP experiment of Bahga et al. [18]. In the latter experiment, LE was 10 mM sodium hydroxide and 20 mM HEPES, and TE was 10 mM pyridine and 20 mM HEPES. One analyte, bis-tris, was focused between LE and TE zones. The NN model also predicts stable ITP focusing of bis-tris with these LE and TE.

4 | CONCLUSION

We demonstrated an NN model for fast and accurate prediction of stable and unstable zones in ITP. We separately trained the NN weights and biases for anionic and cationic ITP using extensive data sets of ITP simulations. In particular, the NN model uses the mobilities and acid dissociation constants of the species and the LE solution composition to predict whether the chosen electrolyte chemistry yields stable analyte focusing on ITP. We have presented the benchmarking of the NN model with simulated test data and validation with published experimental data. The NN model rapidly identifies whether or not a given electrolyte system results in stable ITP focusing of a particular analyte with an accuracy of over 97%.

We have packaged the NN model in a free, web-based application named IONN. The IONN application enables fast and computationally efficient prediction of ITP focusing, with the choice of electrolytes and analytes as the only user-defined inputs. Therefore, IONN can be used by experimenters to quickly screen various ITP electrolyte systems without prior experience in performing electrophoresis simulations. We note that ML-based models, such as NN, are not substitutes for high-fidelity ITP simulations but offer a method of rapid calculation for design and optimization applications. A small number of predicted system designs offered by the NN can then be validated using more accurate, high-fidelity simulations.

Currently, our NN model is trained for handling univalent species. This limitation is primarily because the mobilities and pK_as of higher ionization states of multivalent species lead to additional input features of the NN model. Training an NN model with additional input features will require more layers and nodes and, correspondingly, much more training simulations. However, in many ITP applications, the pH ranges of interest and pK_a values of interest are such that the multivalent species behave as univalent acids or bases. Our NN model accurately predicts ITP focusing with such multivalent species, as demonstrated by model validation based on ITP experiments involving multivalent amino acids. In the future, we will work towards extending the capability of the NN model to handle multivalent species. The fast prediction capability of the NN model can also be leveraged to develop computational tools to automatically suggest the electrolytes for ITP focusing of given analytes by rapidly screening numerous possible combinations of the LE and TE co-ions and the counterion.

ACKNOWLEDGMENTS

We acknowledge the financial support received from the Science and Engineering Research Board (SERB), Government of India, under the Core Research Grant scheme (Grant No. CRG/2022/002148) and the Mathematical Research Impact Centric Support (MATRICS) scheme (Grant No. MTR/2022/000080).

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author on reasonable request.

ORCID

Supreet Singh Bahga https://orcid.org/0000-0001-7277-9015

REFERENCES

1. Everaerts FM, Beckers JL, Verheggen TP. Isotachophoresis: Theory, instrumentation and applications. Amsterdam: Elsevier; 1976.
2. Boček P, Deml M, Gebauer P, Dolnik V. Analytical isotachophoresis. New York: VCH Publishers; 1988.
3. Ramachandran A, Santiago JG. Isotachophoresis: theory and microfluidic applications. Chem Rev. 2022;122:12904–76.
4. Babskii VG, Zhukov MY, Yudovich V. Mathematical theory of electrophoresis. Berlin: Springer Science & Business Media; 1989.
5. Bahga SS, Moza R, Khichar M. Theory of multi-species electrophoresis in the presence of surface conduction. Proc R Soc A Math Phys. 2016;472:20150661.
6. Everaerts FM, Mikkers FEP, Verheggen TPEM. Isotachophoresis. Separ Purif Method. 1977;6:287–351.
7. Hirokawa T, Nishino M, Aoki N, Kiso Y, Sawamoto Y, Yagi T, et al. Table of isotachophoretic indices: I. Simulated qualitative and quantitative indices of 287 anionic substances in the range pH 3–10. J Chromatogr A. 1983;271:D1–D106.
8. Garcia-Schwarz G, Rogacs A, Bahga SS, Santiago JG. On-chip isotachophoresis for separation of ions and purification of nucleic acids. JoVE-J Vis Exp. 2012;61:e3890.
9. Křivánková L, Foret F, Gebauer P, Boček P. Selection of electrolyte systems in isotachophoresis. J Chromatogr A. 1987;390:3–16.
10. Boček P, Gebauer P. Some problems encountered in the selection of electrolyte systems in isotachophoresis. Electrophoresis. 1984;5:338–42.
11. Bier M, Palusinski O, Mosher R, Saville D. Electrophoresis: mathematical modeling and computer simulation. Science. 1983;219:1281–87.
12. Saville D, Palusinski O. Theory of electrophoretic separations. Part I: Formulation of a mathematical model. AIChE J. 1986;32:207–14.
13. Mosher R, Thormann W, Bier M. Computer aided analysis of electric field gradients within isotachophoretic boundaries between weak electrolytes. J Chromatogr A. 1985;320:23–32.
14. Bahga SS, Bercovici M, Santiago JG. Robust and high-resolution simulations of nonlinear electrokinetic processes in variable cross-section channels. Electrophoresis. 2012;33:3036–51.
15. Beckers J, Everaerts F. Isotachophoresis. The qualitative separation of cation mixtures. J Chromatogr A. 1972;68:207–30.
16. Beckers J, Everaerts F. Isotachophoresis: the qualitative separation of anions. J Chromatogr A. 1972;69:165–79.
17. Gebauer P, Boček P. Zone order in isotachophoresis: the concept of the zone existence diagram and its use in cationic systems. J Chromatogr A. 1983;267:49–65.
18. Bahga SS, Kaigala GV, Bercovici M, Santiago JG. High-sensitivity detection using isotachophoresis with variable cross-section geometry. Electrophoresis. 2011;32:563–72.
19. Gaš B, Bravenec P. Simul 6: A fast dynamic simulator of electromigration. Electrophoresis. 2021;42:1291–99.
20. Bercovici M, Lele SK, Santiago JG. Open source simulation tool for electrophoretic stacking, focusing, and separation. J Chromatogr A. 2009;1216:1008–18.
21. Bahga SS, Gupta P. Electrophoresis simulations using Chebyshev pseudo-spectral method on a moving mesh. Electrophoresis. 2022;43:688–95.
22. Avaro AS, Sun Y, Jiang K, Bahga SS, Santiago JG. Web-based open-source tool for isotachophoresis. Anal Chem. 2021;93:15768–74.
23. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. Information Science and Statistics, vol. 4. Berlin: Springer; 2006.

24. Müller AC, Guido S. Introduction to machine learning with Python: A guide for data scientists. Sebastopol, CA: O'Reilly; 2016.
25. Bahga SS, Bercovici M, Santiago JG. Ionic strength effects on electrophoretic focusing and separations. Electrophoresis. 2010;31:910–19.
26. Gupta P, Bahga SS. High-resolution numerical simulations of electrophoresis using the Fourier pseudo-spectral method. Electrophoresis. 2021;42:890–98.
27. Hinton GE. Connectionist learning procedures. In: Machine learning. Amsterdam: Elsevier; 1990. p. 555–610.
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
29. Chambers RD, Santiago JG. Imaging and quantification of isotachophoresis zones using nonfocusing fluorescent tracers. Anal Chem. 2009;81:3022–28.
30. Bahga SS, Santiago JG. Concentration cascade of leading electrolyte using bidirectional isotachophoresis. Electrophoresis. 2012;33:1048–59.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jangra A, Shriyam S, Santiago JG, Bahga SS. A neural network model for rapid prediction of analyte focusing in isotachophoresis. Electrophoresis. 2023;1–10. https://doi.org/10.1002/elps.202300198