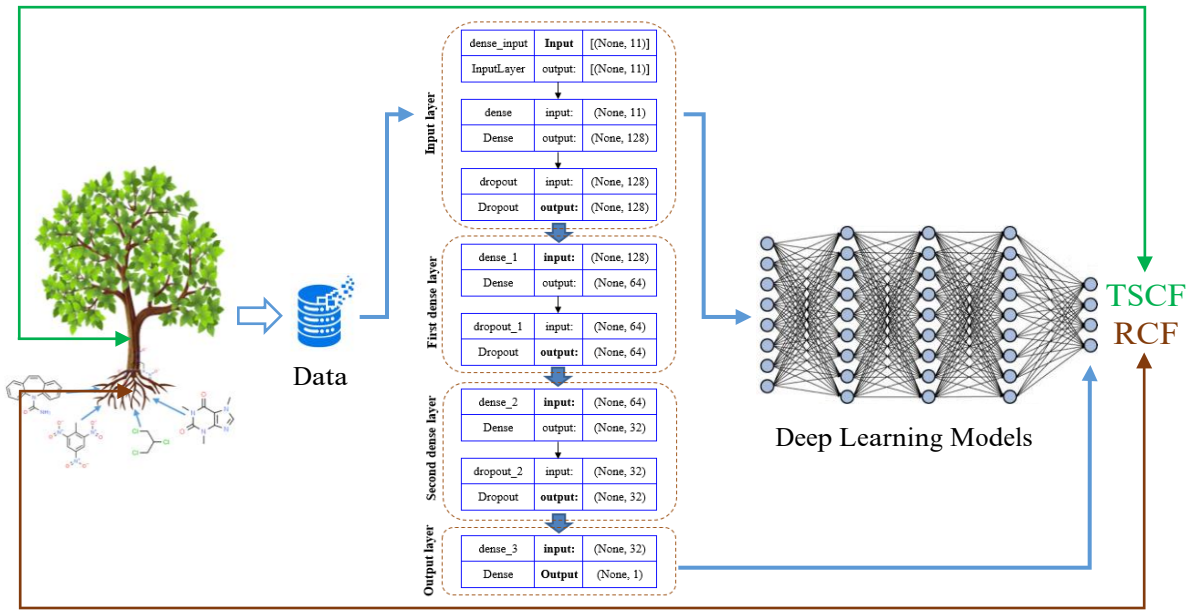
Journal of Hazardous Materials

Deep learning models for predicting plant uptake of emerging contaminants by including the role of plant macromolecular compositions

--Manuscript Draft--

Manuscript Number:	HAZMAT-D-24-12549R1
Article Type:	Research Paper
Keywords:	Emerging contaminants; TSCF; RCF; machine learning; deep learning
Corresponding Author:	Majid Bagheri Savannah State University Savannah, UNITED STATES
First Author:	Majid Bagheri
Order of Authors:	Majid Bagheri
	Shai McKenney
	Julie Gabriella Ware
	Nakisa Farshforoush
Suggested Reviewers:	Amin Shams, PhD amin.shams@semnan.ac.ir
	Lorenzo Rossi, PhD l.rossi@ufl.edu

Graphical Abstract



Deep learning models for predicting plant uptake of emerging contaminants by including the role of plant macromolecular compositions

Majid Bagheri[1, *], Shai McKenney[1], Julie Gabriella Ware[1], Nakisa Farshforoush[2]

[1] Department of Engineering Technology, Savannah State University, Savannah, GA 31404.
[2] Department of Electrical and Computer Engineering, Tabriz University, Tabriz, Iran.

*Corresponding author: Majid Bagheri. Department of Engineering Technology, Savannah State University, Savannah, GA 31404, USA.
Email: bagherim@savannahstate.edu, Phone: 912-358-3262

Abstract

Deep learning models can predict uptake of emerging contaminants in plants with improved accuracy because they leverage advanced data-driven approaches to capture non-linear relationships that traditional models struggle to address. Traditional models suffer from low accuracy in predicting transpiration stream concentration factor (TSCF) and root concentration factor (RCF). This study applied deep neural networks (DNN), recurrent neural networks (RNN), and long short-term memory (LSTM) to enhance the accuracy of predictive models for TSCF and RCF. The three models used nine chemical properties and two plant root macromolecular compositions for predicting TSCF and RCF. The results indicated that deep learning models predict TSCF and RCF with improved accuracy compared to mechanistic models. The coefficient of determination (R^2) for the DNN, RNN, and LSTM models in predicting TSCF was 0.62, 0.67, and 0.56, respectively. The corresponding mean squared error (MSE) on the test set for the models was 0.055, 0.035, and 0.060, respectively. The R^2 for the DNN, RNN, and LSTM models in predicting RCF was 0.90, 0.91, and 0.84, respectively. The corresponding MSE for the models was 0.124, 0.071, and 0.126, respectively. The results of feature extraction using extreme gradient boosting underlined the importance of lipophilicity and root lipid fraction.

Keywords: Emerging contaminants, TSCF, RCF, machine learning, deep learning.

1. Introduction

Predictive models aid in risk assessments, regulatory frameworks, and the formulation of sustainable strategies for mitigating environmental and human health concerns associated with emerging contaminants (Liu et al., 2024; Villeneuve et al., 2019). Predicting the uptake and translocation of emerging contaminants in plants is a critical task, especially in the context of assessing potential risks and impacts on both ecosystems and human health (Shi et al., 2022). This research area investigates the mechanisms by which plants take up and distribute emerging contaminants, such as pharmaceuticals and industrial chemicals, from soil or water into various plant tissues (Bagheri et al., 2023). Understanding these processes is essential for evaluating the bioaccumulation potential and potential transfer of contaminants along the food chain (Chormare and Kumar, 2022; Rossi et al., 2019). The modeling of transpiration stream concentration factor (TSCF) and root concentration factor (RCF) plays a vital role in unraveling the intricate dynamics of plant-contaminant interactions (Trapp, 2000). The TSCF represents the ratio of the concentrations of emerging contaminants in the plant's transpiration stream to those in the surrounding soil (exposure media). This factor sheds light on the contaminant's mobility within the plant and its potential transfer to other tissues (Bagheri et al., 2021). The RCF characterizes the accumulation of contaminants in the plant roots relative to the concentration in the exposure media (Li et al., 2022).

Uptake, translocation, and accumulation of contaminants in plants are generally encapsulated in mathematical models that integrate factors such as plant physiology, soil characteristics, and the physicochemical properties of chemicals (Dourado Junior et al., 2017). These mechanistic models provide valuable insights into the fate and transport of emerging contaminants (Brunetti et al., 2021; Trapp, 2004). Since 1974, a number of modeling studies have offered relationships between

the physicochemical properties of contaminants and their uptake by plants, mainly using the octanol/water partition coefficient (Briggs et al., 1982). These single-parameter relationships suffered from low accuracy and limited applicability for different plant species and chemical compounds. Compartmental models, which take into account more physicochemical and environmental properties and incorporate the complexity of uptake and translocation processes, did not offer high predictive accuracy (Collins and Finnegan, 2010). The accuracy of predictions for TSCF was improved by considering more physicochemical properties in a numerical modeling process (Limmer and Burken, 2014).

The applications of artificial intelligence (AI) and machine learning (ML) models to predict TSCF and RCF offered several advantages over traditional modeling approaches (Zhong et al., 2021). Multi-layer perceptron neural networks significantly improved the accuracy of predictions for both TSCF and RCF compared to the previous approaches. With solely relying on six physicochemical properties, the multi-layer perceptron neural networks outperformed traditional models and complemented the findings of previous studies in some aspects (Bagheri et al., 2020). The fuzzy logic technique also indicated that molecular weight is a significant factor in explaining the uptake efficiency of moderately hydrophobic and hydrophilic compounds (Bagheri et al., 2019). In a more recent study (Gao et al., 2022), the applications of several classical ML and ensemble learning algorithms resulted in improved prediction accuracy for RCFs. These ML models achieved high accuracy by learning nonlinear relationships between RCFs and the properties of contaminants, soils, and plants. Among ensemble learning models, gradient-boosted regression trees showed higher predictive performance for the root uptake of per- and polyfluoroalkyl substances, with accuracies up to 0.85 (Xiang et al., 2023).

The current research is an effort to improve the accuracy of predicting the TSCF and RCF by applying three deep learning models. To the best of our knowledge, this is the first study that employs deep neural networks (DNN), recurrent neural networks (RNN), and long short-term memory (LSTM) models to predict both TSCF and RCF for emerging contaminants. The predictions are based on nine chemical properties and two plant root macromolecular compositions. The feature importance analysis is performed for the input variables of the models using extreme gradient boosting (XGBoost). The relationship between fractions of macromolecules in the plant roots and concentration factors (TSCF and RCF) for the emerging contaminants is missing. The role of root macromolecular fractions in the uptake of emerging contaminants is examined through feature importance analysis.

2. Materials and Methods

2.1. Data sets

Comprehensive data sets were compiled from published studies for the modeling of both TSCF and RCF, see Supplementary Material. The selected TSCF values included 288 records of 151 compounds measured in 33 plant genera under various experimental approaches from 42 studies. The RCF data set included 342 values for 96 compounds in 44 plant genera measured under various experimental approaches from 19 studies. The inclusion of various chemicals and plant species in the data sets makes it possible to develop models that are not compound- or plant-specific. The data sets did not include TSCFs and RCFs from studies when there was no evidence of reaching a steady state, roots were damaged, depletion of dosing solution was higher than 50%, other modes of exposure were included, or calculations were not reliable (Limmer and Burken,

2014). The data sets also did not include TSCFs and RCFs when the metabolism of the parent compound in plants was observed or measurements included metabolites.

The data sets included nine physicochemical properties and two plant root properties. The physicochemical properties were octanol/water partition coefficient (log K_{OW}), molecular weight (MW), hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), rotatable bonds (RB), polar surface area (PSA), vapor pressure (VP), half-life (HL), and water solubility (WS). The experimental values of the properties were considered in the analyses when both predicted and experimental values were available. The chemical properties were obtained from chemical structure databases, including the US EPA Chemistry Dashboard and ChemSpider. The two plant root properties were fractions of the lipids and proteins in the roots. The macromolecular composition of the plant roots was obtained from published studies. Table 1 represents a summary of the chemical and plant root properties used for developing deep learning models for both TSCF and RCF.

2.2. Pattern recognition analysis

t-distributed stochastic neighbor embedding (t-SNE) was used for exploratory data analysis and to detect important patterns in the data sets. t-SNE involves a dimensionality reduction method to reveal and visualize patterns within complex data sets (Zhu et al., 2019). It focuses on preserving local relationships and capturing the intrinsic structure of the data. t-SNE is particularly effective for identifying clusters that might be challenging to discern in the high-dimensional spaces of plant uptake problems (Kim et al., 2023). This technique facilitates the visualization of intricate patterns and clusters within the TSCF and RCF data sets by mapping data points to a lower-dimensional space. However, t-SNE preserves pairwise similarities between the high-dimensional spaces.

2.3. Feature importance analysis

Feature importance analysis is an ML approach for examining the contribution of each input parameter to the predictive models (Zien et al., 2009). It particularly helps determine which parameters have the most influence on the prediction of TSCF and RCF. In this study, XGBoost was used to analyze the importance of the nine physiochemical and two plant root properties for the modeling processes. Feature extraction in XGBoost often refers to the importance scores assigned to each feature during the training process. XGBoost assigns weights to features based on their contribution to predictive accuracy (Wade and Glynn, 2020). These importance scores quantify the influence of each input parameter of the models in making predictions for the TSCF and RCF. The XGBoost models were developed using scikit-learn, which is an ML library in Python (Hackeling, 2017). Each data set was divided into 80% for training and 20% for testing the models. Both data sets were rescaled using StandardScaler to have a standard normal distribution with a mean of 0 and a standard deviation of 1. The hyperparameters of the XGBoost models were optimized through randomized search cross-validation (Bergstra and Bengio, 2012). The XGBoost models showed the highest performance, with a maximum depth of 4 and a learning rate of 0.15.

2.4. Deep learning model training

Three deep learning models, including DNN, RNN, and LSTM, were applied to predict TSCF and RCF. The models were designed and developed in Python using TensorFlow, which is an open-source machine learning library (Ramsundar and Zadeh, 2018). For the three deep learning models, 80% of the data was used for training and 20% for testing. The data was rescaled using StandardScaler to have a standard normal distribution before training. Optimized hyperparameters for the three deep learning models were achieved through randomized search cross-validation.

The DNN models are neural networks with multiple layers, typically including an input layer, one or more hidden layers, and an output layer. The DNN for predicting TSCF and RCF had two hidden layers. The architecture of the DNN model with eleven input variables for predicting TSCF and RCF is shown in Fig. 1. The input layer, first hidden layer, and second layer had 128, 64, and 32 computational neurons, respectively. A dropout of 0.2 was implemented after input and hidden layers to make the predictions reliable. During the training steps, the network adjusts its weights and biases through a backpropagation, minimizing the difference between predicted and measured outputs (Zhu et al., 2018). Adaptive moment (Adam), which is a replacement optimization algorithm for stochastic gradient descent, was used as the optimizer of the DNN models. Each layer of the DNN performs computations on the input data and transforms it into abstract representations. Activation functions apply non-linearity and enable the network to capture relationships within the data (Ding et al., 2018). The activation function of the input and hidden layers of the DNN models was a rectified linear unit (ReLU). Through multiple iterations of training on the measured data, a well-trained DNN can generalize its learned features to make accurate predictions on new test data (Larochelle et al., 2009). The depth and complexity of DNN models allow them to automatically extract important features, which in turn makes them highly effective in predicting plant uptake.

Deep RNN models are neural networks that handle sequential data by incorporating memory mechanisms. It consists of multiple layers of interconnected computational neurons, each processing information over time (Kanagachidambaresan et al., 2021). In this study, the RNN models had one input layer and two hidden layers with 150 computational neurons. Unlike traditional feedforward neural networks, deep RNN models have connections that form directed cycles. This allows them to maintain a memory of previous inputs and leverage temporal

information for prediction. For training the RNN models, the network is iteratively exposed to sequential input data. The network adjusts its internal parameters to minimize the discrepancy between predicted and measured outputs (Sutskever, 2013). The weights and biases of the RNN models were adjusted using adaptive moment estimation. Similar to the DNNs, the activation function of the input and hidden layers of the RNN models was a rectified linear unit.

The LSTM models are RNNs that capture long-term dependencies in sequential data and address vanishing gradient problems (Sherstinsky, 2020). During training, LSTMs utilize backpropagation through time to compute gradients and adjust the weights. The activation functions play a crucial role in information flow and memory cell modulation. LSTMs introduce memory cells with self-regulating mechanisms, including input, forget, and output gates. The input gate determines which information is stored, the forget gate regulates what information is discarded, and the output gate decides what information is passed to the next time step (Manaswi and Manaswi, 2018). This architecture enables LSTM models to selectively retain or forget information over long sequences and allows them to capture and remember relevant patterns. The LSTM models for predicting TSCF and RCF had an input layer and two hidden layers with 40 computational neurons in the optimal conditions. The activation function of the input and hidden layers was a rectified linear unit. The adaptive moment estimation outperformed other methods in adjusting the weights and biases.

2.5. Performance evaluation

Evaluating the performance of deep learning models is a critical step in assessing the effectiveness and reliability of their predictions. Mean squared error (MSE), which is a common loss function in regression problems, was used to measure the average squared difference between

predicted values and actual values. The lower values of MSE indicate better performance. The coefficient of determination (R-squared or R^2) was used to evaluate the goodness of fit of regression models. The MSE and R^2 are calculated as follow:

$$\mathrm{MSE} = 1/\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_{pi}} - \mathrm{y_{ti}})^2 \quad (1)$$

$$\mathrm{R}^2 = 1 - \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_{ti}} - \mathrm{y_{pi}})^2 / \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_{ti}} - \bar{\mathrm{y}})^2 \quad (2)$$

where $\bar{\mathrm{y}}$ is the average of y over the n data, y_t is the actual value, and y_p is the predicted value.

The performance of XGBoost for the feature importance analysis was examined based on the F1 score. The F1 score combines precision and recall into a single metric, considering both false positives and false negatives of the predictions. The F1 is calculated as follow:

$$\text{F1 score} = \mathrm{TP}/(\mathrm{TP} + 1/2(\mathrm{FP} + \mathrm{FN})) \quad (3)$$

where TP is true positive, FP is false positive, and FN is false negative.

2.6. Plant macromolecular compositions

Chemicals either accumulate in the roots or cross the plant root membranes and transport to the upper tissues through the vascular pathways. Plant roots and shoots are composed of water, wax, lignin, cellulose, lipids, phenolics, and non-structural carbohydrates (Gupta and Singh, 1981). While lipophilicity is an important factor, predicting the uptake of emerging contaminants solely based on lipids is a simplified approach. The fraction of root macromolecules such as protein is not negligible since these materials (protein, lignin, and cellulose) were shown to be important in other biological systems (Endo et al., 2012; Jonker, 2008; Stoklosa et al., 2013). It is assumed that the partitioning of emerging contaminants into plant roots and other tissues is equal to the partitioning of compounds into macromolecules. In this study, the changes in concentration factors, including TSCF and RCF, will be examined over a wide range of root macromolecular

fractions. The results of feature extraction will be used to analyze the uptake and translocation of emerging contaminants based on the fractions of lipids and proteins in the roots.

3. Results and Discussion

3.1. Hidden patterns in the data

The results of pattern recognition using t-SNE to visualize the possible clusters in the TSCF and RCF data sets are shown in Fig. 2. The first dimension is derived in such a way that similar data points in the original space are also close to each other in this new dimension. Similarly, the second dimension is also obtained by preserving pairwise similarities between data points. The results for both TSCF and RCF showed different clusters in the data sets. For the RCF, the clusters for the compounds with the higher root concentration factors formed separate clusters. The perfluoroalkyl family of chemicals was found to have higher RCFs, as shown in the separate clusters in Fig. 2. The observations of another study also indicated that the clusters in the RCF data reflect the similarities across different combinations of chemicals, plants, and soils (Gao et al., 2022). For the TSCF, the clusters are less distinguishable compared to the RCF data. However, the plant species and compounds with higher uptake efficiency were clustered closer. This is in line with the results of a previous study, which stated that tomato is a species with higher uptake potential (Bagheri et al., 2019).

3.2. Significant features for predicting TSCF and RCF

The feature importance analysis using XGBoost based on nine physicochemical properties and two plant root properties yielded insightful findings (Fig. 3). The octanol/water partition coefficient (log K_{OW}) and molecular weight (MW) were two paramount predictors for both TSCF

and RCF, as reflected by their high F1 scores. Particularly, the log K_{OW}, with F1 scores of 209 and 215 for TSCF and RCF, respectively, emphasized its critical role in predicting these concentration factors. The F1 scores of the MW for predicting TSCF and RCF were 167 and 131. Previous studies using statistical methods such as stepwise regression also demonstrate that log K_{OW} and MW are significant variables in the prediction of both TSCF and RCF (Bagheri et al., 2020; Bagheri et al., 2019). The feature importance analysis using XGBoost also showed that lipid and protein fractions are other significant predictors, with high F1 scores for both variables. The lipid fraction with high F1 scores of 189 and 187 for the TSCF and RCF models was the second significant predictor. The feature importance analyses using neural networks and regression trees also indicated the lipid fraction as a major predictor for the RCF (Gao et al., 2022). The protein fraction of the plant roots, with high F1 scores of 126 and 100 for the TSCF and RCF, was found to be another significant predictive feature. Despite their importance, the role of proteins and macromolecules other than lipids in the uptake and translocation of emerging contaminants in plants has never been deeply studied. These results underscore the importance of understanding the interactions between chemical and root properties in plant uptake modeling and provide valuable insights for risk assessments and ecological management strategies.

3.3. Predictive models for TSCF

The three deep learning models showed improved accuracy on the test data sets for the prediction of TSCF. The results indicated that deep RNN models have the highest performance based on R-squared and MSE values (Fig. 4). The deep RNN achieved the highest accuracy with an R^2 of 0.67 and an MSE of 0.35. The training history based on train and test data sets indicated that the RNN models are reliable without any overfitting or underfitting. The close values for the

train and test losses confirm the reliability of the predictions by the RNN models. The performance of the RNN models based on the tests was positive compared to the traditional models and simple neural networks, with MSEs of 0.25 and 0.037, respectively (Bagheri et al., 2019; Doucette et al., 2018; Schriever and Lamshoeft, 2020). Despite the high accuracy of the simple neural networks, their predictive performance was not consistent for different compounds since the models did not consider important chemical and plant properties (Bagheri et al., 2019). The RNN models were followed by DNNs with an R^2 of 0.62 and an MSE of 0.55. The training history for the DNN models showed that train and test losses are close over 100 epochs. The LSTM model predicted the TSCF with lower accuracy compared to the RNN and DNN models. The best values of R^2 and MSE on the test data set for the LSTM models were 0.56 and 0.06, respectively.

The results of this study demonstrated the importance and need for considering plant properties in the modeling of plant uptake. The macromolecular compositions were significant and relevant parameters for the prediction of concentration factors. One of the main drawbacks of the single-parameter models based on lipophilicity was their applicability to specific chemicals or plant species (Limmer and Burken, 2014). The models considered fractions of lipids and proteins in the plant roots, which in turn make the models applicable for different plant species. The model also considered vapor pressure and biodegradation half-lives for the chemicals as two new input parameters. These chemical properties improved the reliability of the predictive models, particularly for volatile and degradable compounds. The feature importance analysis showed that these properties are important since they had relatively high F1 scores. The vapor pressure had F1 scores of 105 and 114, and the biodegradation half-lives had F1 scores of 100 and 51 for the TSCF and RCF.

3.4. Predictive models for RCF

The deep learning algorithms predicted the RCF with higher performance compared to the TSCF. The three deep learning models outperformed previous traditional and data-driven models for predicting the RCF. The three deep learning models predicted the logarithm of the RCF for better visualization representation. Similar to the results for the TSCF, the deep RNN models indicated the highest performance for the prediction of the RCF (Fig. 5). The deep RNN achieved the highest accuracy with an R^2 of 0.91 and an MSE of 0.071. This model was followed by the DNN with an R^2 of 0.9 and an MSE of 0.124. The training of the RNN and DNN models was successful, as shown by the decreasing losses of the train and test data sets. The close and decreasing losses for train and test sets proved the lack of overfitting or underfitting in both deep learning models. The RNN and DNN models showed significant improvement over the traditional single-parameter relationships with low accuracy and specificity for limited compounds (Briggs et al., 1982; Chen et al., 1989). These deep learning models indicated a higher accuracy for the prediction of the RCF than simple neural networks, with an R-squared of 0.82 (Bagheri et al., 2020). The prediction of the log RCF using fully connected neural networks and by considering different chemical and plant properties achieved an accuracy of 0.79 and a mean average error of 0.22 (Gao et al., 2022). In this study, even the LSTM models with an R^2 of 0.84 and an MSE of 0.126 showed higher performance compared to these neural network models (Fig. 5). The results of this study indicated that the three deep learning models outperformed traditional and classical machine learning models in predicting the TSCF and RCF (Table 2).

3.5. Plant uptake and root macromolecular fractions

The feature importance analyses and previous studies have emphasized the role of macromolecules in plant uptake of emerging contaminants. Correlation analysis was performed to examine the roles of fractions of root macromolecules in the accumulation and distribution of emerging contaminants in plants. Fig. 6 demonstrates the correlation of fractions of the lipids and proteins in the plant roots with the log RCF. The results of this study indicated that RCFs for the emerging contaminants correlate negatively with the root lipids ($P < 0.05$) and positively with the root proteins ($P < 0.05$). The result for the lipids is supported by the generally accepted understanding that compounds with higher lipophilicity have higher TSCFs and lower RCFs (Burken and Schnoor, 1998; Dettenmaier et al., 2009). The results of a study on the uptake and accumulation of perfluorooctane sulfonate and perfluorooctanoate emphasized the importance of both lipids and proteins (Wen et al., 2016). The results indicated that the perfluorooctane sulfonate and perfluorooctanoate accumulations in roots correlate positively with root protein contents and negatively with root lipid contents.

4. Conclusions

This study employed deep neural networks (DNN), recurrent neural networks (RNN), and long short-term memory (LSTM) models to enhance the predictive accuracy of TSCF and RCF. The findings demonstrated significant improvements in the predictive accuracy of these deep learning models compared to the traditional models. DNN showed the highest accuracy in predicting the TSCF and FCF with coefficients of determination equal to 0.67 and 0.91, respectively. The mean squared error for TSCF and FCF was 0.035 and 0.071, respectively. The findings of this study underscore the potential of deep learning techniques to improve predictive

models for plant uptake and translocation of emerging contaminants. This study also indicated the importance of physicochemical properties and fractions of macromolecules for reliable prediction of the TSCF and RCF. Including important physicochemical properties such as degradation and fractions of macromolecules such as lipids and proteins in the modeling process enhanced the reliability of predictions.

Declaration of Competing Interest

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

Data Statement

The data for this study is available in Excel format as Supplementary Material.

Acknowledgments

The authors declare that they have no competing interests. This work was supported by National Science Foundation under Award Number 2300369 and National Science Foundation under Award Number 2348805.

References

Bagheri, M., Al-Jabery, K., Wunsch, D., Burken, J.G., 2020. Examining plant uptake and translocation of emerging contaminants using machine learning: Implications to food security. Sci. Total Environ. 698, 133999. https://doi.org/10.1016/j.scitotenv.2019.133999

Bagheri, M., Al-Jabery, K., Wunsch, D.C., Burken, J.G., 2019. A deeper look at plant uptake of environmental contaminants using intelligent approaches. Sci. Total Environ. 651, 561-569. https://doi.org/10.1016/j.scitotenv.2018.09.048

Bagheri, M., He, X., Al-Lami, M.K., Oustriere, N., Liu, W., Limmer, M.A., Shi, H., Burken, J.G., 2023. Assessing plant uptake of organic contaminants by food crops tomato, wheat, and corn through sap

concentration factor. Int. J. Phytoremediat. 25(9), 1215-1224. https://doi.org/10.1080/15226514.2022.2144797

Bagheri, M., He, X., Oustriere, N., Liu, W., Shi, H., Limmer, M.A., Burken, J.G., 2021. Investigating plant uptake of organic contaminants through transpiration stream concentration factor and neural network models. Sci. Total Environ. 751, 141418. https://doi.org/10.1016/j.scitotenv.2020.141418

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13(2), 281-305.

Briggs, G.G., Bromilow, R.H., Evans, A.A., 1982. Relationships between lipophilicity and root uptake and translocation of non-ionised chemicals by barley. Pestic. Sci. 13(5), 495-504. https://doi.org/10.1002/ps.2780130506

Brunetti, G., Kodeš ová , R., S̆ vecová , H., Fér, M., Nikodem, A., Klement, A., Grabic, R., S̆ imů nek, J.í., 2021. On the use of mechanistic soil−plant uptake models: a comprehensive experimental and numerical analysis on the translocation of carbamazepine in green pea plants. Environ. Sci. Technol. 55(5), 2991-3000. https://dx.doi.org/10.1021/acs.est.0c07420?ref=pdf

Burken, J.G., Schnoor, J.L., 1998. Predictive relationships for uptake of organic contaminants by hybrid poplar trees. Environ. Sci. Technol. 32(21), 3379-3385.

Chen, Q., Briggs, G., Evans, A., 1989. Relationships between lipophilicity and root uptake and translocation of non-ionised chemicals by rice. Acta Agric. Nucleatae Sinica 3, 1-3.

Chormare, R., Kumar, M.A., 2022. Environmental health and risk assessment metrics with special mention to biotransfer, bioaccumulation and biomagnification of environmental pollutants. Chemosphere 302, 134836. https://doi.org/10.1016/j.chemosphere.2022.134836

Collins, C.D., Finnegan, E., 2010. Modeling the plant uptake of organic chemicals, including the soil− air− plant pathway. Environ. Sci. Technol. 44(3), 998-1003.

Dettenmaier, E.M., Doucette, W.J., Bugbee, B., 2009. Chemical hydrophobicity and uptake by plant roots. Environ. Sci. Technol. 43(2), 324-329.

Ding, B., Qian, H., Zhou, J., 2018. Activation functions and their characteristics in deep neural networks, 2018 Chin. Control Decis. Conf. IEEE, pp. 1836-1841. https://doi.org/10.1109/CCDC.2018.8407425

Doucette, W.J., Shunthirasingham, C., Dettenmaier, E.M., Zaleski, R.T., Fantke, P., Arnot, J.A., 2018. A review of measured bioaccumulation data on terrestrial plants for organic chemicals: metrics, variability, and the need for standardized measurement protocols. Environ. Toxicol. Chem. 37(1), 21-33. https://doi.org/10.1002/etc.3992

Dourado Junior, S., Nunes, E., Marques, R., Rossino, L., Quites, F., Siqueira, J., Moreto, J., 2017. Controlled release behavior of sulfentrazone herbicide encapsulated in Ca-ALG microparticles: preparation, characterization, mathematical modeling and release tests in field trial weed control. J. Mater. Sci. 52, 9491-9507. https://doi.org/10.1007/s10853-017-1103-9

Endo, S., Bauerfeind, J., Goss, K.-U., 2012. Partitioning of neutral organic compounds to structural proteins. Environ. Sci. Technol. 46(22), 12697-12703. https://doi.org/10.1021/es303379y

Gao, F., Shen, Y., Sallach, J.B., Li, H., Zhang, W., Li, Y., Liu, C., 2022. Predicting crop root concentration factors of organic contaminants with machine learning models. J. Hazard. Mater. 424, 127437. https://doi.org/10.1016/j.jhazmat.2021.127437

Gupta, S., Singh, J., 1981. The effect of plant species, weather variables and chemical composition of plant material on decomposition in a tropical grassland. Plant and Soil 59(1), 99-117. https://doi.org/10.1007/BF02183596

Hackeling, G., 2017. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd. Birmingham, UK.

Jonker, M.T., 2008. Absorption of polycyclic aromatic hydrocarbons to cellulose. Chemosphere 70(5), 778-782. https://doi.org/10.1016/j.chemosphere.2007.07.020

Kanagachidambaresan, G., Ruwali, A., Banerjee, D., Prakash, K.B., 2021. Recurrent neural network, in: Prakash, K.B., Kanagachidambaresan, G. R. (Eds.), Programming with TensorFlow: Solution for Edge Computing Applications, Springer Cham, Ghent, Belgium, pp. 53-61. https://doi.org/10.1007/978-3-030-57077-4_7

Kim, J., Lee, C., Park, J., Kim, N., Kim, S.-L., Baek, J., Chung, Y.-S., Kim, K., 2023. Comparison of Various Drought Resistance Traits in Soybean (Glycine max L.) Based on Image Analysis for Precision Agriculture. Plants 12(12), 2331. https://doi.org/10.3390/plants12122331

Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P., 2009. Exploring strategies for training deep neural networks. J. Mach. Learn. Res. 10(1).

Li, Y., Sallach, J.B., Zhang, W., Boyd, S.A., Li, H., 2022. Characterization of plant accumulation of pharmaceuticals from soils with their concentration in soil pore water. Environ. Sci. Technol. 56(13), 9346-9355. https://doi.org/10.1021/acs.est.2c00303

Limmer, M.A., Burken, J.G., 2014. Plant translocation of organic compounds: molecular and physicochemical predictors. Environ. Sci. Technol. Lett. 1(2), 156-161. https://doi.org/10.1021/ez400214q

Liu, S., Qiu, Y., He, Z., Shi, C., Xing, B., Wu, F., 2024. Microplastic-derived dissolved organic matter and its biogeochemical behaviors in aquatic environments: A review. Crit. Rev. Environ. Sci. Technol. 54(11), 865-882. https://doi.org/10.1080/10643389.2024.2303294

Manaswi, N.K., 2018. RNN and LSTM, in: Manaswi, N.K.(Eds.), Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras. Springer, Berkeley, CA, pp. 115-126. https://doi.org/10.1007/978-1-4842-3516-4_9

Ramsundar, B., Zadeh, R.B., 2018. TensorFlow for deep learning: from linear regression to reinforcement learning. " O'Reilly Media, Inc.", Sebastopol, CA.

Rossi, L., Bagheri, M., Zhang, W., Chen, Z., Burken, J.G., Ma, X., 2019. Using artificial neural network to investigate physiological changes and cerium oxide nanoparticles and cadmium uptake by Brassica napus plants. Environ. Pollut. 246, 381-389. https://doi.org/10.1016/j.envpol.2018.12.029

Schriever, C., Lamshoeft, M., 2020. Lipophilicity matters–a new look at experimental plant uptake data from literature. Sci. Total Environ. 713, 136667. https://doi.org/10.1016/j.scitotenv.2020.136667

Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenom. 404, 132306. https://doi.org/10.1016/j.physd.2019.132306

Shi, Q., Xiong, Y., Kaur, P., Sy, N.D., Gan, J., 2022. Contaminants of emerging concerns in recycled water: Fate and risks in agroecosystems. Sci. Total Environ. 814, 152527. https://doi.org/10.1016/j.scitotenv.2021.152527

Stoklosa, R.J., Velez, J., Kelkar, S., Saffron, C.M., Thies, M.C., Hodge, D.B., 2013. Correlating lignin structural features to phase partitioning behavior in a novel aqueous fractionation of softwood Kraft black liquor. Green Chem. 15(10), 2904-2912. https://doi.org/10.1039/C3GC41182F

Sutskever, I., 2013. Training recurrent neural networks. University of Toronto, Toronto, ON, Canada.

Trapp, S., 2000. Modelling uptake into roots and subsequent translocation of neutral and ionisable organic compounds. Pest Manag. Sci. 56(9), 767-778. https://doi.org/10.1002/1526-4998(200009)56:9%3C767::AID-PS198%3E3.0.CO;2-Q

Trapp, S., 2004. Plant uptake and transport models for neutral and ionic chemicals. Environ. Sci. Pollut. Res. 11, 33-39. https://doi.org/10.1065/espr2003.08.169

Villeneuve, D.L., Coady, K., Escher, B.I., Mihaich, E., Murphy, C.A., Schlekat, T., Garcia-Reyero, N., 2019. High-throughput screening and environmental risk assessment: State of the science and emerging applications. Environ. Toxicol. Chem. 38(1), 12-26. https://doi.org/10.1002/etc.4315

Wade, C., Glynn, K., 2020. Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd., Birmingham, UK.

Wen, B., Wu, Y., Zhang, H., Liu, Y., Hu, X., Huang, H., Zhang, S., 2016. The roles of protein and lipid in the accumulation and distribution of perfluorooctane sulfonate (PFOS) and perfluorooctanoate (PFOA) in plants grown in biosolids-amended soils. Environ.l Pollut. 216, 682-688. https://doi.org/10.1016/j.envpol.2016.06.032

Xiang, L., Qiu, J., Chen, Q.-Q., Yu, P.-F., Liu, B.-L., Zhao, H.-M., Li, Y.-W., Feng, N.-X., Cai, Q.-Y., Mo, C.-H., 2023. Development, Evaluation, and Application of Machine Learning Models for Accurate Prediction of Root Uptake of Per-and Polyfluoroalkyl Substances. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.2c09788

Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., 2021. Machine learning: new ideas and tools in environmental science and engineering. Environ. Sci. Technol. 55(19), 12741-12754. https://doi.org/10.1021/acs.est.1c01339

Zhu, H., Akrout, M., Zheng, B., Pelegris, A., Jayarajan, A., Phanishayee, A., Schroeder, B., Pekhimenko, G., 2018. Benchmarking and analyzing deep neural network training, 2018 IEEE Int. Symp. Workload Charact. IEEE, pp. 88-100. https://doi.org/10.1109/IISWC.2018.8573476

Zhu, W., Webb, Z.T., Mao, K., Romagnoli, J., 2019. A deep learning approach for process data visualization using t-distributed stochastic neighbor embedding. Ind. Eng. Chem. Res. 58(22), 9564-9575. https://doi.org/10.1021/acs.iecr.9b00975

Zien, A., Krämer, N., Sonnenburg, S., Rätsch, G., 2009. The feature importance ranking measure, Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20. Springer, pp. 694-709. https://doi.org/10.1007/978-3-642-04174-7_45

Lists of figure captions:

Fig. 1. Architecture of the deep neural networks with eleven inputs for predicting TSCF and RCF.

Fig. 2. Visualization of the hidden patterns in the TSCF and RCF data sets using t-SNE.

Fig. 3. Significance of different predictive variables for the TSCF and RCF models based on feature importance analysis using XGBoost.

Fig. 4. Performance of the deep learning models for the TSCF based on test data. Regression plots and training history for the deep neural networks (A1 and A2), recurrent neural networks (B1 and B2), and long short-term memory (C1 and C2).

Fig. 5. Performance of the deep learning models for the RCF based on test data. Regression plots and training history for the deep neural networks (A1 and A2), recurrent neural networks (B1 and B2), and long short-term memory (C1 and C2).

Fig. 6. Relationships between lipid and protein contents in plant roots and RCFs of emerging contaminants.

Table 1 Characteristics of the parameters used in the deep learning modeling processes.

Input parameter	Minimum–Maximum		Output parameter	Min-Max
	TSCF	RCF		
log Kow	-2.19–6.75	-0.88–8.15	TSCF	0.001–1.16
MW (Da)	32–616.4	52–765	RCF	0.009–497.15
HBD	0–6	0–5	log RCF	-2.046–2.696
HBA	0–16	0–11		
RB	0–36	0–13		
PSA (A^2)	0–196.2	0–161		
VP (mm Hg)	4.6e-18–538	0–167		
HL (day)	0.5–832	3–832		
WS (mg/L)	2.0e-4–1.0e-6	2.0e-4–1.0e-6		
Root lipids (%)	0.1–7.2	0.16–9		
Root proteins (%)	1.35–23	1.35–28		

Table 2 Predictive performance of the three deep learning models compared to the reported studies.

Model	Input	Output	R^2	Error	Reference
RNN	Nine chemical and two plant root properties	log RCF	0.91	MSE = 0.071	This study
DNN			0.90	MSE = 0.124	
LSTM			0.84	MSE = 0.126	
RNN	Nine chemical and two plant root properties	TSCF	0.67	MSE = 0.035	This study
DNN			0.62	MSE = 0.055	
LSTM			0.56	MSE = 0.060	
ANN	Six physicochemical properties	TSCF	0.54	MSE = 0.037	Bagheri et al. (2019)
ANN	Six physicochemical properties	RCF	0.80	MSE = 922.2	Bagheri et al. (2020)
GBRT	Molecular, soil, and root properties	log RCF	0.76	MAE = 0.23	Gao et al. (2022)
RF			0.71	MAE = 0.25	
FCNN			0.79	MAE = 0.22	
SVR			0.68	MAE = 0.26	
BPNN	Three physicochemical and soil properties	RCF	0.80	-	Wang et al. (2021)

RNN: Recurrent neural networks, DNN: Deep neural networks, LSTM: Long short-term memory, ANN: Artificial neural networks, GBRT: Gradient boosted regression trees, RF: Random forest, FCNN: Fully connected neural networks, SVR: Support vector regression, BPNN: Backpropagation neural networks.

Figure 1

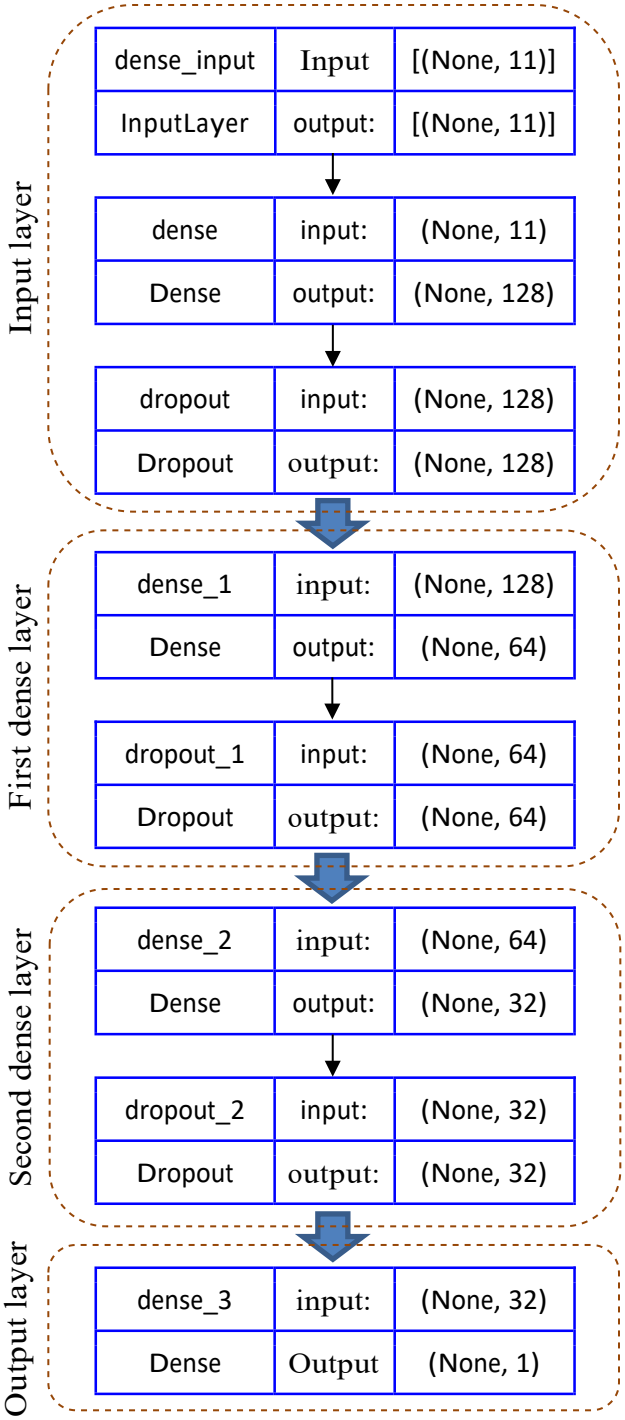


Figure 2

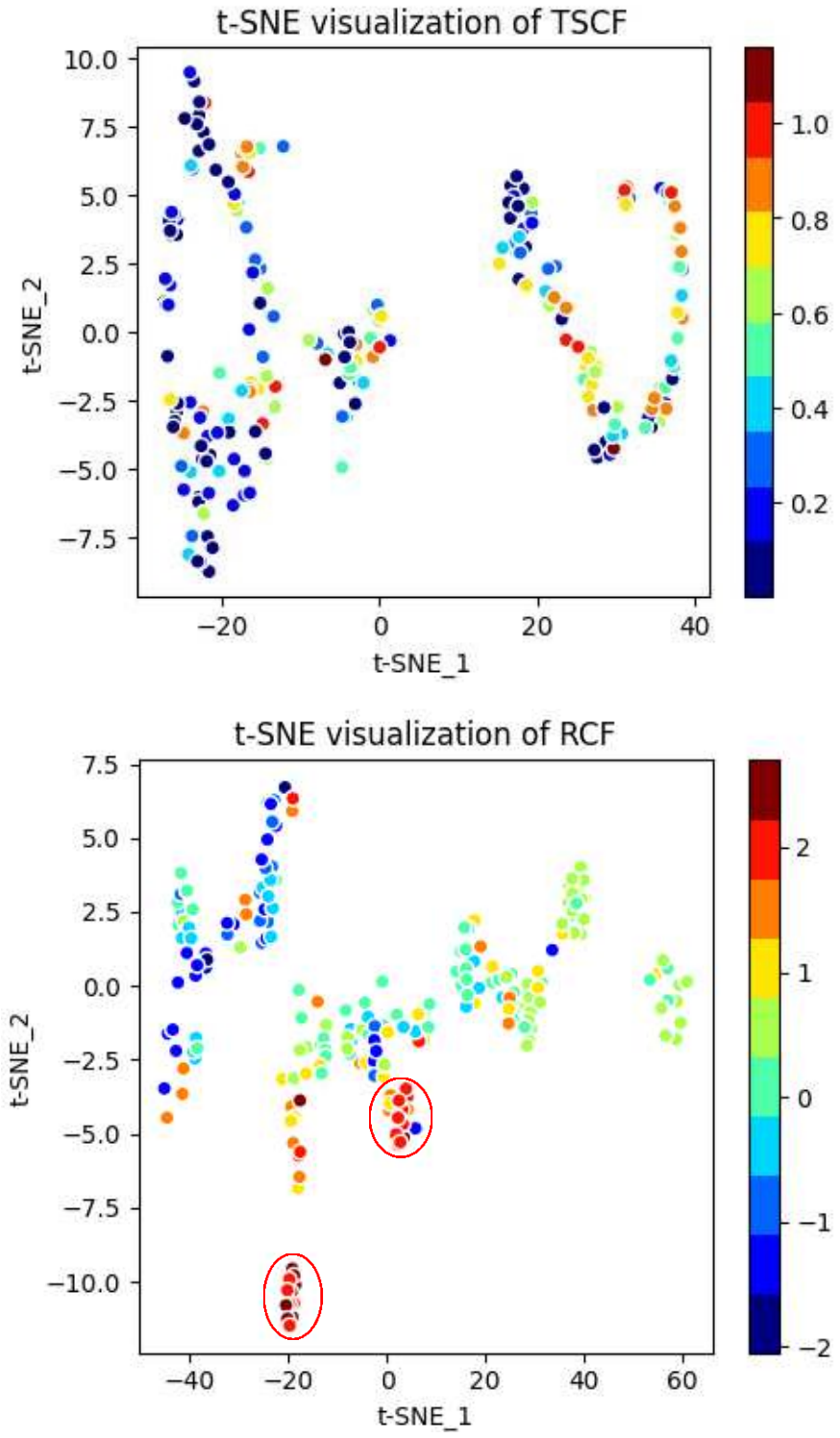


Figure 3

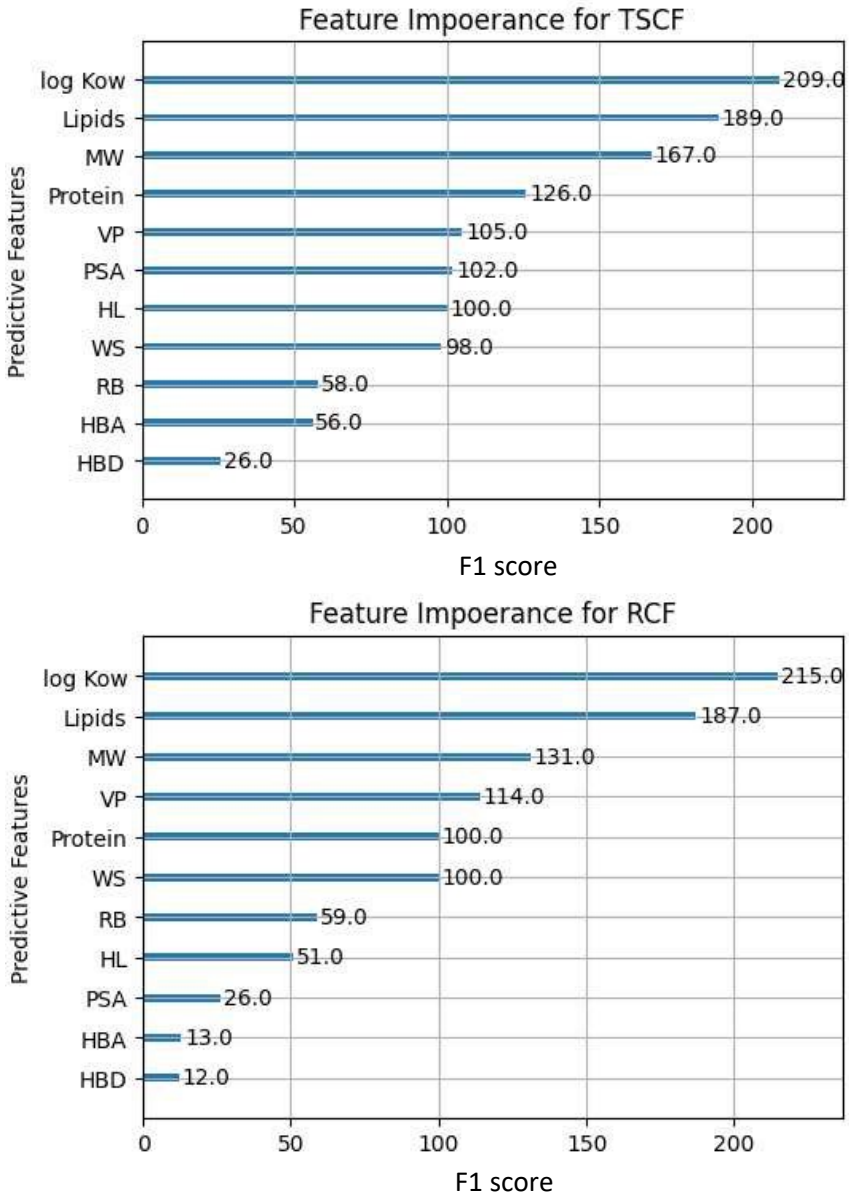


Figure 4

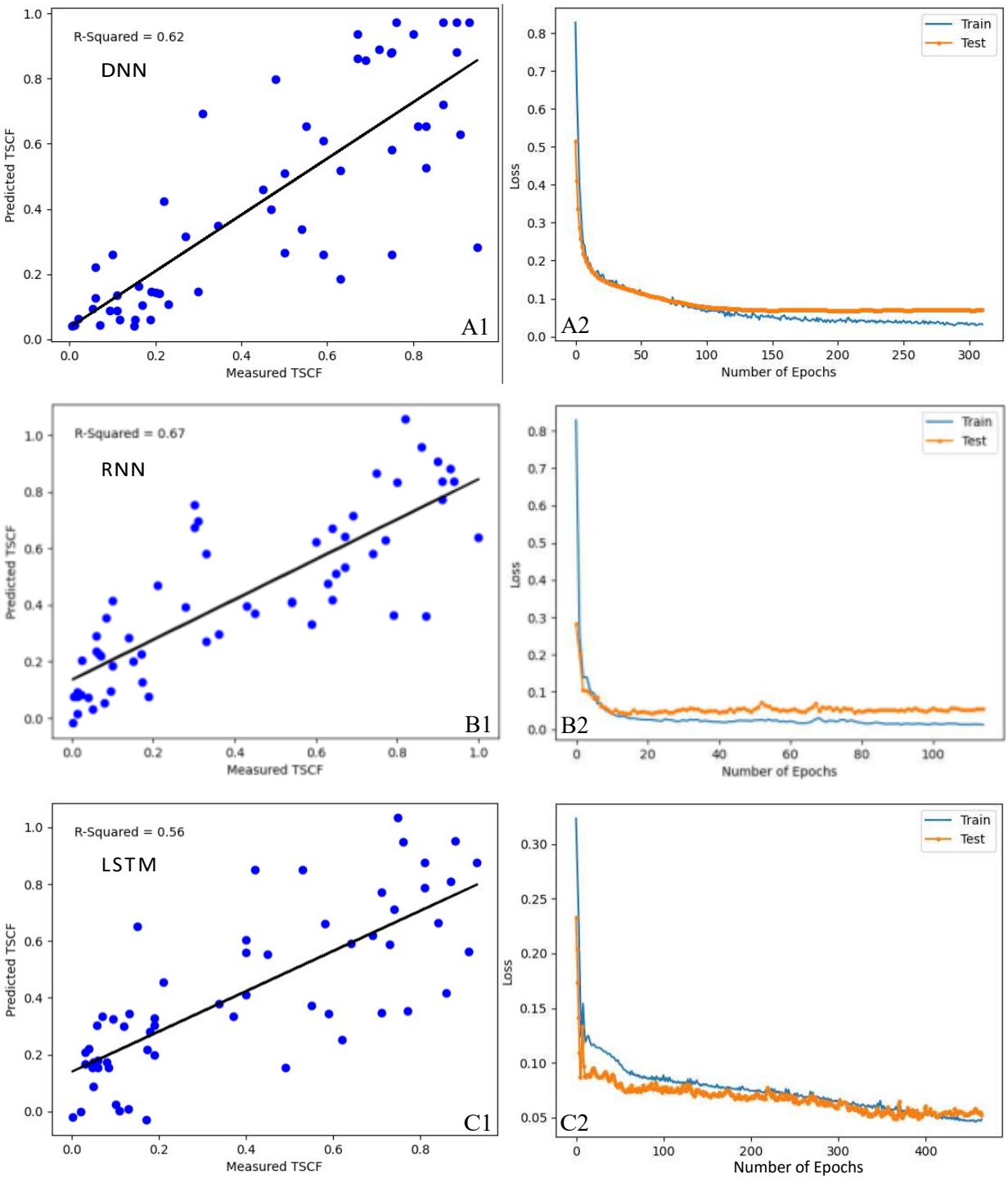


Figure 5

Click here to access/download;Figure;Figure 5.docx

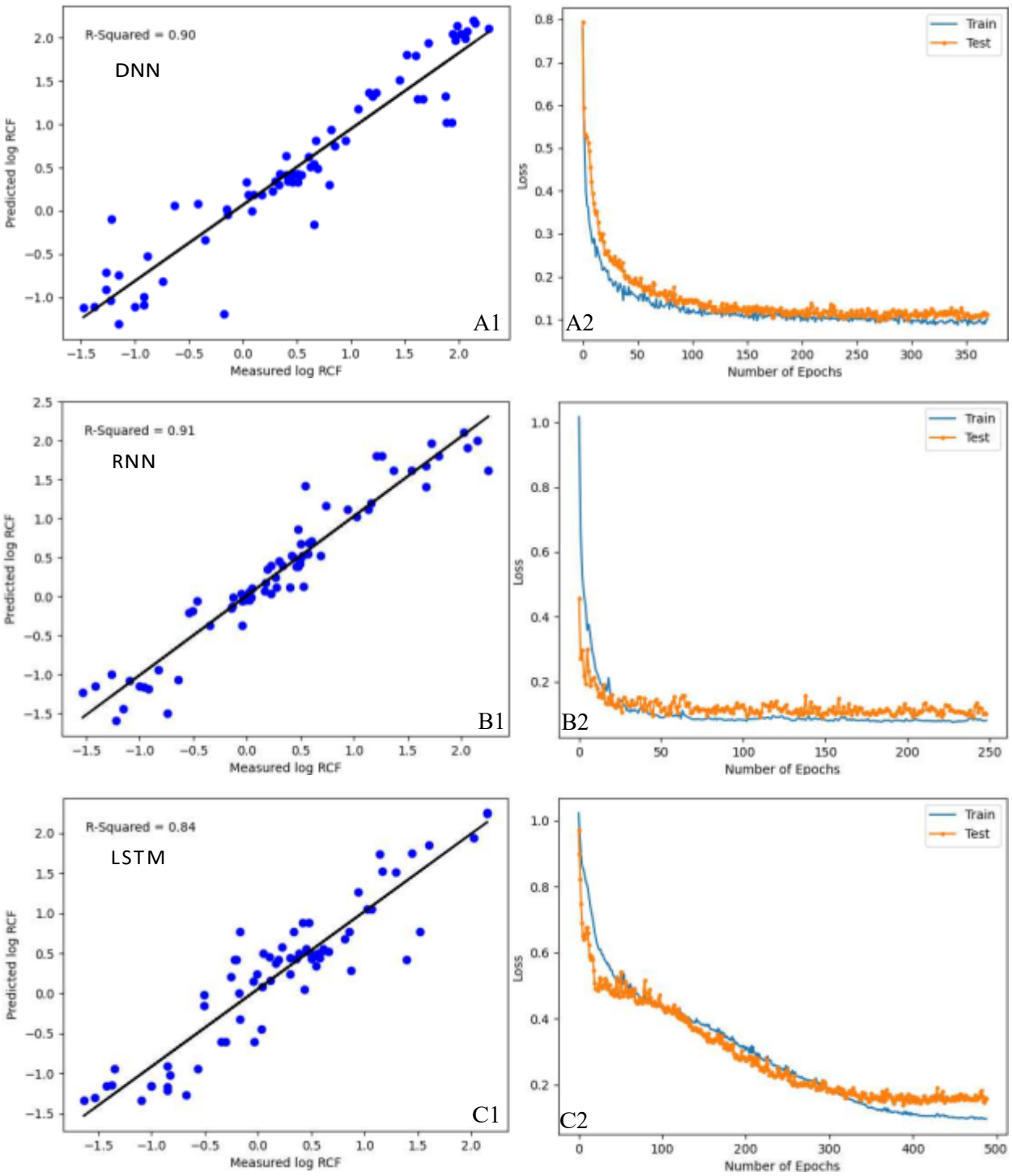


Figure 6