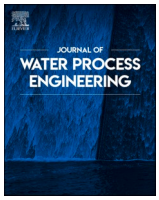
Journal of Water Process Engineering 62 (2024) 105379



Contents lists available at ScienceDirect

Journal of Water Process Engineering

journal homepage: www.elsevier.com/locate/jwpe



Machine learning screening tools for the prediction of extraction yields of pharmaceutical compounds from wastewaters

Ana Casas [a], Diego Rodríguez-Llorente [b,*], Guillermo Rodríguez-Llorente [c], Juan García [b], Marcos Larriba [b]

[a] School of Water, Energy and Environment, Cranfield University, Cranfield MK43 0AL, UK
[b] Catalysis and Separation Processes Research Group (CyPS), Department of Chemical Engineering and Materials, Complutense University of Madrid, Avda. Complutense s/n, 28040 Madrid, Spain
[c] Department of Artificial Intelligence, HI Iberia, C. de Juan Hurtado de Mendoza, 14, 28036 Madrid, Spain

ARTICLE INFO

Editor: Ludovic F. Dumée

Keywords:
Machine learning
COSMO-RS
Liquid-liquid extraction
Pharmaceuticals
Wastewater

ABSTRACT

Pharmaceutical compounds have become an increasingly important source of pollutants in wastewaters being conventional treatments ineffective in removing them, so they are commonly discharged into the environment. Pharmaceuticals can be successfully removed using liquid-liquid extraction, and COSMO-RS can be used to predict interactions and identify the most promising solvents. However, COSMO*therm* models cannot account for key process parameters, which reduces the accuracy of these computational models. Therefore, there is a need for alternative computational approaches to accurately predict the extraction yields of pharmaceuticals which can incorporate both processing and interaction variables. This work used machine learning to predict the extraction yield of eleven pharmaceuticals using eight solvents. Six regression models and two classification models were explored. The best performance was obtained with ANN regressor (test MAE: 4.510, test R^2: 0.884) and RF classifier (test accuracy: 0.938, test recall: 0.974). The RF regression analysis and classification also showed key extraction yield features: solvent-to-feed ratio, n–octanol–water partition coefficient, hydrogen bond and Van der Waals contributions to excess enthalpy, and pH distance to nearest pKa. Machine learning showed as an excellent tool for screening and selecting the most promising solvents and process conditions to remove pharmaceuticals from wastewater.

1. Introduction

Pharmaceuticals are a significant source of micropollutants in wastewater, with hospital wastewater being a major contributor [1,2]. Long-term exposure to pharmaceuticals in trace amounts can harm aquatic ecosystems and potentially impact human health. These substances enter sewage treatment plants but are not effectively removed by conventional treatment processes, leading to their discharge into the natural environment [3]. The presence of pharmaceuticals in water environments raises concerns as antimicrobial resistance and endocrine disruption effects [4,5]. Various techniques are being explored to eliminate pharmaceuticals from wastewater, including biological processes [6,7], oxidation methods [8–12], and separation processes [13–15]. Liquid-liquid extraction is a well-established method, but using hazardous conventional solvents is a concern. To address this, alternative and more sustainable solvents like natural eutectic solvents [16–21], and terpenoids [22,23] are being explored.

Eutectic solvents and terpenoids have been shown in literature as promising solvents in the extraction of pharmaceutical compounds from wastewater [24]. However, there is an unlimited number of combinations of cations, anions and hydrogen bond donors and acceptors that can be used to constitute eutectic solvents, and the literature on the use of terpenoids in this context is scarce. For those reasons and because laboratory testing is expensive and time consuming, the use of computational tools is becoming essential to identify the most promising solvents for the removal pharmaceuticals from wastewater [25].

The conductor-like screening model for realistic solvents (COSMO-RS) is an a priori molecular modelling method developed by Eckert and Klamt [26]. It uses quantum chemistry and statistical mechanics to predict thermodynamic properties of fluids based on their geometries and charge densities. [25,27]. COSMO-RS has been used in the literature for the screening of solvents such as conventional solvents, ionic liquids,

* Corresponding author.
E-mail address: diegor@ucm.es (D. Rodríguez-Llorente).

https://doi.org/10.1016/j.jwpe.2024.105379
Received 20 November 2023; Received in revised form 5 April 2024; Accepted 23 April 2024
Available online 30 April 2024
2214-7144/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

terpenoids and eutectic solvents for extraction applications [24,25,28–32].

COSMO-RS models are recognized for their efficacy in forecasting diverse chemical system behaviors, encompassing fluid phase properties, solvation effects on reactions, and liquid-liquid extraction phenomena [33–36]. Despite their utility, the precision of COSMO-RS predictions is subject to variations influenced by factors such as quantum chemical calculations and methodological choices [37]. While COSMO-RS is highly regarded for its accuracy in predicting solvation energies, it faces limitations, notably in forecasting CO_2 solubility in specific solvents [38] or with brominated compound predictions [39].

In pharmaceutical contexts, COSMO-RS exhibits superior performance relative to the perturbed-chain statistical associating fluid theory (PC-SAFT) equation of state, even in the absence of specific experimental data [40]. It surpasses PC-SAFT both qualitatively and quantitatively in predicting the solubility of active pharmaceutical ingredients (APIs) in pure solvents. However, both models may occasionally misjudge deviations from ideal behavior [40].

In our previous work, COSMO-RS was successfully employed for the prediction of interactions between pharmaceutical compounds and three types of solvents (i.e. conventional solvents, terpenoids and eutectic solvents). The selection of the drugs was made considering their higher presence in the wastewaters and that they belonged to different families of compounds. The most promising solvents were selected based on their excess enthalpies, and the extraction yields of the drugs using those solvents were predicted with COSMO*therm* and validated in the laboratory. However, as COSMO*therm* considers the pharmaceutical species as neutral, the COSMO*therm* model led to accurate predictions when those species where indeed neutral, but the absolute errors increased up to 15 % when the species were charged [24]. Therefore, there is a need for computational approaches to accurately predict the extraction yields of pharmaceutical products from wastewaters which can account for key processing parameters such as deviations from the neutral state in the pharmaceutical compounds.

Machine learning is a field of study that allows computers to learn from data. Supervised learning is a type of machine learning where a set of dependent and independent variables are provided to the algorithm and the relationship drawn between these variables is used to make predictions [41]. Machine learning algorithms can model complex non-linear relationships [42] and these methodologies have been widely used in literature in the context of the wastewater industry for a variety of applications, such as in the study of removal of pharmaceuticals from wastewaters by adsorption [43–46], in real-time prediction of wastewater biochemical oxygen demand [47], in anomaly classification in wastewater treatment plants [48], in microalgal wastewater treatment [49], to aid with the design of biological wastewater treatments [50], in water quality evaluation [51], in wastewater treatment plants [52,53], and in nutrient recovery [54]. Although machine learning is already being applied with eutectic solvents, among others, it is used for predicting physicochemical properties rather than for predicting extraction yields of certain solutes [55,56]. In this sense, the only recently published work addressing this problem applies a methodology using COSMO-RS and artificial neural networks to predict boron extraction yields from aqueous solutions with hydrophobic eutectic solvents [57]. Thus, there is a need to expand this field by exploring other solutes, systems, and models.

This work used machine learning methodologies to predict the extraction yield of eleven pharmaceutical compounds from wastewaters using eight different solvents. A range of processing parameters and thermodynamic properties provided by COSMO-RS were used to make the predictions. Two machine learning methodologies were explored: regression analysis and classification analysis. Regression analysis is the process of learning relationships between dependent variables and independent variables from a dataset and using that knowledge to predict outputs for novel inputs [58]. In contrast, classification analysis is used to classify instances based on various inputs [59,60].

Six regression models were used for the prediction of pharmaceuticals extraction yields, i.e. k-Nearest Neighbours (KNN), Support Vector Regressor (SVR), Decision Trees (DT), Random Forest (RF), Gradient Boosting (GB) and Artificial Neural Networks (ANN). Random Forest and ANN were also used for classification analysis to predict instances where the extraction yield was equal to or >90 %. These algorithms are commonly used in wastewater applications in literature [42,54,61,62]. These traditional machine learning models have been chosen for this study because of their simplicity, interpretability, simpler data requirements and lower computational costs [63].

This work aims to build computational screening tools capable of identifying the most promising solvents for extracting pharmaceutical compounds from hospital wastewater, overcoming the limitations of previously reported methodologies. These tools would also allow a better understanding of the impact of the thermodynamics and process features on the pharmaceuticals' extraction yield. To the best of the authors' knowledge, this is the first attempt made in the literature to predict the liquid-liquid extraction of pharmaceutical compounds from wastewater using machine learning methodologies and employing both processing and thermodynamic variables.

2. Materials and methods

2.1. Prediction of thermodynamic properties using COSMO-RS

BIOVIA TmoleX 2023 version 23.0.0 with COSMO continuum solvation method was used to optimise the geometries of the compounds with a BP86/TZVP computational level using the solvent effect and the obtained COSMO files were then introduced in BIOVIA COSMO*therm* 2023 software version 23.0.0 [64]. Therefore, computational results were obtained by using Dassault Systèmes BIOVIA software programs. Equimolar binary mixtures containing the pharmaceutical compounds and the different solvents were considered and the excess enthalpies and the n-octanol-water partition coefficients for the solute were calculated at 298.2 K as reference. For eutectic solvents, composed of a binary mixture of compounds, computation was conducted as a mixture at the molar composition of the eutectic point of hydrogen bond acceptor and donor compounds as in previous work. The excess enthalpy was further detailed in its three contributions (hydrogen bond, Van der Waals forces, and electrostatic and misfit interactions) [24].

2.2. The dataset

The dataset used for the machine learning analysis consisted of 2156 datapoints, where each of these represents an extraction experiment for a unique combination of factors. The dependent variable was the extraction yield or extraction class for the regression and the classification analysis, respectively, and the independent variables are summarised in Table 1. These independent variables were selected based on their potential to influence the extraction process. Two of the independent variables were categorical: the water matrix and the solvent type.

In this study, the impacts of the pH distance to the closest pKa and the n-octanol-water partition coefficient on the pharmaceuticals' extraction yield have been explored for the first time in literature using machine learning.

11 pharmaceutical compounds were included in this analysis, 5 analgesics (acetaminophen, diclofenac, ibuprofen, naproxen, and phenazone), 1 antidepressant (carbamazepine), 4 antibiotics (sulfamethoxazole, tetracycline, ciprofloxacin, and trimethoprim), and 1 blood pressure regulator (atenolol). The solvents used in the analysis were three terpenoids (eucalyptol, thymol and carvacrol), three eutectic solvents (thymol + octanoic acid, thymol + decanoic acid and thymol + dodecanoic acid), and two conventional solvents (ethyl acetate and diisopropyl ether).

The two types of water matrix considered in this analysis were ultrapure water obtained from a Purelab flex Elga Veolia water

Table 1
Independent variables.

Variable	Additional information
Solvent-to-feed ratio (S/F)	N/A
Type of water matrix	Hospital wastewater matrix or laboratory ultrapure water
Initial concentration of the pharmaceutical product	Measured in mg/L
Excess enthalpy (HE)	Measured in kcal/mol; source: BIOVIA COSMO*therm* 2023
Individual contributions to the excess enthalpy: Hydrogen bond (H(HB)), Van der Waals (H(VDW)), Misfit or electrostatic interactions (H(MF))	Measured in kcal/mol; source: BIOVIA COSMO*therm* 2023
n-Octanol–water partition coefficient (log10(P))	source: BIOVIA COSMO*therm* 2023
Solvent type	Conventional, eutectic or terpenoid.
pH distance to closest pKa	Absolute difference of the pH of aqueous solution to the closest pKa. pKa values found in previous work [24]
Charge	Zero, positive or negative
Charge of the next species to the closest pKa	Zero, positive or negative

purification system and a hospital wastewater effluent collected from a hospital in the southern region of Madrid (Spain).

The extraction yield was obtained from our previous study [24]. The temperature chosen for the tests was 40 °C, which was the minimum to ensure the liquid state of the solvents, and close to room temperature for the convenience of operation and energy saving. More details of the materials used to generate the data in this study, the preparation of the eutectic solvents and the methodology used for the removal of the pharmaceuticals from wastewaters are described in our previous work [24].

Descriptive statistics were calculated for the numerical variables. The correlation matrix was generated to understand the strength and direction of the relationship of the dependent and the independent variables, and to identify any potential multicollinearity issues.

2.3. Prediction of extraction yields using regression and classification analysis

2.3.1. Data pre-processing

Python was the programming language used in this study, and the analysis was run on the Jupyter Notebook platform using the open-source libraries scikit learn [65], Keras [66] and TensorFlow [67]. The categorical variables were transformed into dummy variables prior to the analysis, and one of the dummies for each variable was left out of the analysis to avoid multicollinearity issues. Multicollinearity occurs when independent variables are highly correlated, which might affect the stability of the regression model [68]. The dataset was randomly split into a test and a train dataset using a test size of 20 % and a training size of 80 %, which is a standard split used in this type of analysis that ensures a reasonable amount of data is used for both model training and testing purposes [69]. As some of the models used in the analysis were sensitive to scale (i.e. KNN, SVR and ANN), and the independent variables in this study had different scales, the data was scaled for those models using a StandardScaler class to improve the accuracy of these models [70]. Standardisation scaling is a commonly used technique to transform features to have zero mean and unit variance [71]. A grid search with 5-fold cross-validation was performed to select the most promising model and reduce overfitting (except for the ANN model). In k-fold validation the train set is randomly split in k-folds of approximately equal size and each fold is considered as a test set and the remaining k-1 folds are considered as train set; the model assessment metrics are then averaged or combined [42]. For the search of best hyperparameters for the ANN model, Bayesian Optimization [72] was performed.

2.3.2. Regression analysis

Regression analysis was used in this study to predict the extraction yield of eleven pharmaceuticals in eight solvents. The response was the extraction yield of the pharmaceutical compound, measured as percentage. Six regression models were used for the analysis, k-Nearest Neighbours, Decision Trees, Support Vector Regressor, Random Forest, Gradient Boosting and Artificial Neural Network. k-Nearest Neighbours and Decision Trees are simple and interpretable models, but more prone to overfitting, whereas Support Vector Regressor, Random Forest, Gradient Boosting and Artificial Neural Network are suited for complex datasets and less prone to overfitting, but they are less interpretable [42,61,62,73–79].

The performance of those algorithms was measured using the mean absolute error (MAE), the root mean squared error (RMSE) and the coefficient of determination or R squared (R^2), which are common performance metrics in regression analysis [42]. The selection of the features to include in the model was performed using a stepwise approach, using either the individual contributions to the enthalpy or the global enthalpy. The regression algorithms are briefly described in this section.

The k-Nearest Neighbours (KNN) algorithm is one of the simplest and most intuitive machine learning algorithms. It is a non-parametric algorithm which can be used for both classification and regression tasks. In regression, it predicts the value of a new instance as the mean of the nearest k neighbours. The Euclidean distance is used to determine the similarity between two instances [61,73].

Support Vector Regressors (SVR) are an extension of the Support Vector Machine classification algorithm. The SVR algorithm finds a hyperplane in the feature space that best fits the train set and minimises the error. SVRs are capable of handling nonlinear relationships and therefore they are well suited to model complex systems while avoiding overfitting. [61,74].

The Decision Tree algorithm (DT) is a binary recursive algorithm which is widely used in literature due to its relative simplicity and interpretability. It can be used for both regression and classification tasks [62]. The DT algorithm splits the features into different nodes and the nodes with the most homogeneous sub-nodes are selected based on an error metric or a splitting requirement [42,62,75]. Due to its simplicity, DTs can suffer from low accuracy and stability issues [62].

The Random Forest (RF) is an ensemble algorithm widely used for classification and regression tasks due to its simplicity, flexibility, and ease of implementation. In RF regression, a group of decision trees are trained on random subsets of the original dataset with bootstrapping (i.e. resampling with replacement) and the predictions of all the trees are averaged to obtain a more accurate prediction while reducing the variance. In addition to using random subsets of the dataset, the RF also uses random feature selection to build the decision trees [42,59,75,76].

Gradient Boosting (GB) is a popular ensemble algorithm used for both regression and classification tasks. The GB algorithm uses a gradient descent algorithm to minimise the target residuals. It builds weak learners, usually decision trees, in an iterative manner, where each tree is built to correct the errors made by previous trees. This makes the model highly robust, but it also requires more trees to achieve high accuracy [42,61].

Artificial Neural Networks are versatile and powerful algorithms which have the potential to approximate any output given a related input [77]. The ANN is based in non-linear combinations of inputs in multiple layers which are fed via a series of weights. These weights are learned during training through backpropagation [78], which computes the specific contribution of each weight to the global error defined in the loss function. The ANN used in this study for regression and classification tasks is Feedforward Neural Network (FNN) (also known as Multilayer-Perceptron, MLP) [79]. To prevent overspecification of neurons and overfitting, regularisation techniques were used during the training, i.e. Batch Normalisation [80] and Dropout [81].

2.3.3. Classification analysis

Classification analysis, and specifically binary classification, was used to predict whether a solvent leads to an extraction yield of the pharmaceutical compounds equal to or >90 %. This value of 90 % was considered by the authors as an appropriate threshold to identify the most promising solvents. For this task, instances where the extraction yield was equal to or >90 %, were labelled as yield class 1 (i.e. positive), whereas instances where the extraction yield was lower than that value were labelled as 0 (i.e. negative).

A Random Forest classifier was used for the classification task, due to its simplicity and flexibility while providing high accuracy and low variance. The RF classifier follows a similar approach as the equivalent regressor, but in classification tasks, it uses the majority voting method to combine the results from multiple decision trees and therefore selecting the most frequent class for each instance [48,82,83].

An ANN classifier was also developed for the classification task. This task was performed by modifying the activation function in the last layer, which has one neuron and a sigmoid activation function [84]. The sigmoid function gives output values between 0 and 1, which allows binary classification. For the loss function, binary cross-entropy [85] was selected, which measures the difference between predicted and actual probability distributions in binary outcomes.

The classifiers were evaluated using common performance metrics, i.e. accuracy, confusion matrix, precision, recall and F-1 score. Accuracy refers to the proportion of classes identified correctly. A confusion matrix is a matrix showing the number of true positives, true negatives, false positives, and false negatives obtained with the model. Precision is defined as the proportion of predicted positives that are actual positives. Recall or sensitivity refers to the proportion of positives which are correctly classified as such. The F-1 score is calculated as the harmonic mean between precision and recall [86].

3. Results and discussion

3.1. Exploratory data analysis

Table 2 shows the descriptive statistics of the numerical features in the dataset. As seen in Table 2. Descriptive statistics of the dataset 2, the solvent-to-feed ratio varied from 0.05 to 3 in the dataset and the initial concentration ranged from 5 to 50 mg/L. In terms of the interactions between the drugs and the solvents, the three contributions to the excess enthalpy ranged from exothermic to endothermic values and the n-octanol–water partition coefficient had values between −0.18 and 3.20. The pH distance to the nearest pKa was between 0.03 and 6.75 units. The charge of the species was either −1, 0 or 1, and the charge of the next specie to the closest pKa was either −1, 0, 1 or 2. With respect to the response, i.e. the extraction yield, the dataset had values between 0 and 100 %, with most of the datapoints above 70 %.

Table S1 of Supplementary material shows the correlation matrix of the numerical variables of the dataset. As seen in Table S1, the extraction yield showed a mild positive correlation with the solvent-to-feed ratio (0.327) and with the n-octanol–water partition coefficient (0.196). These results agree with our previous findings, which showed an increase in yield with the solvent-to-feed ratio [24]. In terms of the octanol-water partition coefficient, this is a measure of the lipophilicity of the pharmaceutical compound, so high values of this coefficient are associated with more tendency to migrate from the water to the solvent [32].

In contrast, the extraction yield showed a weak negative correlation with the pH distance to the closest pKa (−0.075), and a mild negative correlation with the excess enthalpy (−0.366) and with the hydrogen bond (−0.344) and misfit (−0.307) contributions to the excess enthalpy. This suggests that exothermic excess enthalpies, hydrogen bond and misfit contributions and low pH distances to the nearest pKa improve the

Table 2
Descriptive statistics of the dataset – a) full dataset, b) training dataset, c) test dataset.

a)

Full dataset	S/F	Initial Conc. (mg/L)	H(HB)	H(MF)	H (VDW)	log10 (P)	pH distance to closest pKa	Charge	Charge of the next specie to the closest pKa	Yield (%)
Mean	0.986	13.036	−1.119	−0.146	0.018	1.505	2.348	−0.133	0.107	79.7
Min.	0.050	5.000	−3.543	−0.847	−0.048	−0.177	0.030	−1	−1	0
25 %	0.100	5.000	−2.101	−0.242	0.002	0.949	0.990	−1	0	70.9
50 %	0.500	5.000	−0.734	−0.122	0.014	1.166	2.200	0	0	96.5
75 %	2.000	5.000	−0.389	−0.059	0.032	2.571	3.000	0	1	99.6
Max.	3.000	50.000	0.093	0.183	0.115	3.201	6.750	1	2	100.0

b)

Train set	S/F	Initial Conc. (mg/L)	H(HB)	H (VDW)	log10 (P)	pH distance to closest pKa	Charge	Charge of the next specie to the closest pKa	Yield (%)
Mean	0.988	13.066	−1.123	0.019	1.505	2.334	−0.136	0.106	79.6
Min.	0.050	5.000	−3.543	−0.048	−0.177	0.030	−1.000	−1.000	0
25 %	0.100	5.000	−2.101	0.002	0.949	0.990	−1.000	0.000	70.7
50 %	0.500	5.000	−0.725	0.014	1.166	2.200	0.000	0.000	96.4
75 %	2.000	5.000	−0.389	0.032	2.571	3.000	0.000	1.000	99.6
Max.	3.000	50.000	0.093	0.115	3.201	6.750	1.000	2.000	100

c)

Test set	S/F	Initial Conc. (mg/L)	H(HB)	H (VDW)	log10 (P)	pH distance to closest pKa	Charge	Charge of the next specie to the closest pKa	Yield (%)
Mean	0.976	12.917	−1.105	0.018	1.506	2.405	−0.123	0.113	80.3
Min.	0.050	5.000	−3.543	−0.048	−0.177	0.030	−1.000	−1.000	0.0
25 %	0.100	5.000	−2.101	0.001	0.949	0.980	−1.000	−1.000	72.1
50 %	0.500	5.000	−0.743	0.014	1.166	2.200	0.000	0.000	97.1
75 %	2.000	5.000	−0.344	0.031	2.571	3.300	0.000	1.000	99.7
Max.	3.000	50.000	0.093	0.089	3.201	6.750	1.000	2.000	100.0

extraction yields. These results agree with our previous findings, which showed that exothermic excess enthalpies imply more affinity of the pharmaceutical compounds for the solvent, whereas endothermic excess enthalpies worsen the extraction yields [24]. The correlation matrix also shows that the hydrogen bond contribution to the excess enthalpy is the individual contribution with the strongest correlation with the yield, which also agrees with our previous findings [24]. As for the distance to pH, it is more complex. When the next molecule to the nearest pKa is neutral, i.e. at the current pH the molecule is charged, an increase of the pH distance to the pKa would imply a reduction of species in the neutral state. As seen experimentally, an increase in charged species implies a higher affinity for water than for hydrophobic solvents, which would reduce extraction yields [24]. However, when the main species is neutral, a higher pH to pKa distance would imply a reduction of the charged states and, therefore an increase in the extraction yield.

The correlation matrix also showed a strong correlation between the hydrogen bond and the misfit contributions to the excess enthalpy (0.597), so the latest was not included in the regression and classification models to avoid multicollinearity issues in the analysis.

3.2. Prediction of extraction yields using regression analysis

Various combinations of the features were used in the regression models to predict the extraction yield of pharmaceutical compounds using conventional, eutectic and terpenoid solvents. The best performance was obtained using the individual contributions to the enthalpy (HB and VDW) instead of the global HE, and including the S/F, initial concentration, log10(P), water matrix, solvent type, pH distance to closest pKa, charge and charge of the next species to the closest pKa in the model.

Various hyperparameters were tuned for the models. The aim was to reach a balance that minimised overfitting, while maintaining a good performance in the model, illustrated by the test set performance measures.

Table 3 shows the values of the MAE, RMSE and R^2 for the six regression models obtained for that combination of features.

As seen in Table 3, the ANN regressor led to the best performance with a MAE of 4.510, an RMSE of 10.186 and an R^2 of 0.884. Adam optimization algorithm [87] was used for Gradient Descent learning, with learning rate 10^{-3}, MAE as loss function and decay rate of 0.98. The architecture of the model consists of 5 dense layers with 512 neurons in the first and last layer and 1024 neurons in the deep layers. Dropout has been applied in all the layers, and batch normalisation in the last two layers to prevent overfitting during training. The activation function applied in the layers is LeakyReLU, which prevents gradient vanishment and fast computation [88], except for the last layer, which uses a sigmoid function as the yield is bounded between 0 and 1.

The Random Forest regressor led to a similar performance, with a MAE of 5.732, an RMSE of 10.264 and an R^2 of 0.881. The hyperparameters tuned were the maximum number of trees in the model (i.e. number of estimators), the maximum depth of the tree, the minimum number of samples required to be at a leaf node, the minimum number of samples required to split an internal node and the maximum number of features to consider when looking for the best split [65]. The best results were obtained with 3 features, 100 estimators, a maximum depth of 80, a minimum of 2 samples required to split and a minimum of 1 sample in the leaf.

The SVR also led to promising results, with a MAE of 6.481, an RMSE of 12.364, and a R^2 of 0.827. The regularisation parameter C and the hyperparameter epsilon were tuned for SVR. Large values of C decrease the regularisation and therefore the model has more tendency to overfitting. Epsilon specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value. This means that if the deviation is less than epsilon, no penalty will be imposed in the model, and when the deviation is greater than epsilon, the corresponding penalty will be imposed. [89]. The best results were obtained with values of 400 for C and 0.4 for epsilon, and a Radial Basis Function (RBF) kernel, well suited for non-linear relationships [90].

The remaining three algorithms, KNN, DT and GB, led to less promising results than SVR, RF and ANN. Thus, the R^2 values were lower for these algorithms than for SVR, RF and ANN, with an R^2 value of 0.812 for KNN, 0.738 for DT and 0.771 for GB. In addition, the values of MAE and RMSE were greater for KNN, DT and GB than for SVR, RF and ANN, with MAE ranging between 6.862 and 7.722, and RMSE ranging between 12.887 and 15.209 for KNN, DT and GB. The differences observed between the train and test sets for the R^2, MAE and RMSE values also suggest that the KNN, DT and GB models tended to overfit the data. For these reasons, we considered that the most promising models in this study were SVR, RF and ANN.

Fig. 1 shows a scatter plot of the predicted yield versus the observed yield for the ANN (a) SVR (b) and the RF (c) models. As seen in Fig. 1, all models provided an adequate fit, as the data is close to the diagonal; the dispersion of the data in the RF and ANN model was lower than in the SVR, as expected due to the higher R^2 value obtained for those models (see Table 3).

The random forest model allowed us to determine the feature importances and identify the key features driving the drug extraction process (see Table 4). The results showed that the five most important features were the solvent-to-feed ratio (0.198), followed by the hydrogen bond contribution to the excess enthalpy (0.189), the n-octanol–water partition coefficient (0.144), the Van der Waals contribution to the excess enthalpy (0.117) and the pH distance to the nearest pKa (0.086). These results reflect how both the processing and thermodynamic features are important in the accurate prediction of extraction yields and are consistent with previous findings, which highlighted the importance of excess enthalpies, solvent-to-feed ratio and pH in the process [24]. These findings also highlight for the first time the importance of the n-octanol–water partition coefficient in the prediction of pharmaceuticals extraction yields. In the work conducted by [57] on the prediction of extraction yields for boron extraction, the critical parameters were also pH and S/F ratio. This information will allow researchers to select the most favourable solvents and conditions to extract pharmaceutical compounds from wastewaters and will aid in the process design at large scale. Moreover, not only as a tool for predicting the most suitable solvents, but also as a tool for predicting extraction yields for the selected S/F using real effluents.

Table 3
MAE, RMSE and R^2 of the regression models (features: S/F, initial concentration, log10(P), H(HB), H(VDW), water matrix, solvent type, pH distance to closest pKa, charge and charge of the next species to the closest pKa; dependent variable: extraction yield (%)).

	Train set			Test set		
Model	MAE	RMSE	R^2	MAE	RMSE	R^2
ANN	3.280	5.780	0.965	4.510	10.186	0.884
RF	2.040	3.630	0.986	5.732	10.264	0.881
SVR	4.593	9.969	0.891	6.481	12.364	0.827
KNN	0	0	1	6.862	12.887	0.812
DT	3.422	7.765	0.934	7.545	15.209	0.738
GB	4.741	7.752	0.934	7.722	14.212	0.771

3.3. Prediction of extraction yields using classification analysis

For applications where it is not necessary to calculate the actual yield value, having a screening tool capable of accurately classify solvents as excellent or less promising might be an appropriate alternative. With those applications in mind, we performed a classification analysis with a random forest classifier and ANN, using a yield label as a response – 1 in cases where the extraction yield was equal to or >90 %, and 0 in cases where the extraction yield was lower than that value.

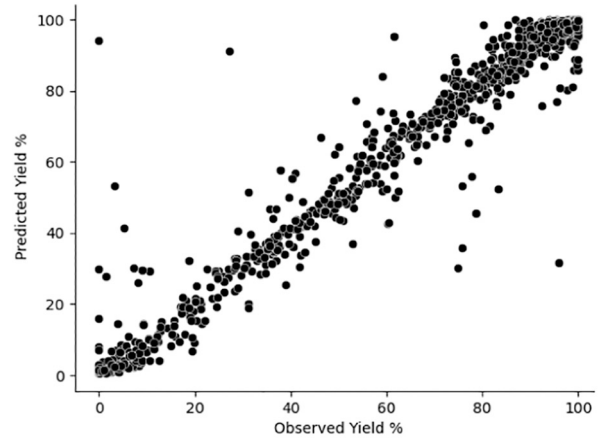


Fig. 1. Predicted yield vs observed yield using the ANN (a), SVR (b) and RF regressor (c) (features: S/F, initial concentration, log10(P), H(HB), H(VDW), water matrix, solvent type, pH distance to closest pKa, charge and charge of the next species to the closest pKa; dependent variable: extraction yield (%)).

Table 5 shows the error metrics of the RF classifier for the train and test set after hyperparameter tuning using grid search. The features used in this classification model were the same as in the random forest regressor. The best results were obtained with 3 features, 25 estimators, a maximum depth of 70, a minimum of 3 samples required to split and a minimum of 1 sample in the leaf.

Table 4
Feature importances in the random forest regression model (features: S/F, initial concentration, log10(P), H(HB), H(VDW), water matrix, solvent type, pH distance to closest pKa, charge and charge of the next species to the closest pKa; dependent variable: extraction yield (%)).

Feature	Importance
S/F	0.198
H(HB)	0.189
Log10(P)	0.144
H(VDW)	0.117
pH distance to pKa	0.086
Conventional solvent (dummy variable)	0.063
Eutectic solvent (dummy variable)	0.057
Charge of the next species to the closest pKa	0.042
Charge	0.039
Water matrix (lab/hospital)	0.038
Initial concentration	0.026

Table 5
Performance metrics of the random forest and ANN classifiers (features: S/F, initial concentration, log10(P), H(HB), H(VDW), water matrix, solvent type, pH distance to closest pKa, charge and charge of the next species to the closest pKa; dependent variable: yield class – 1: yield $\geq$ 90 %, 0: yield $<$ 90 %).

	Random Forest		ANN	
Metric	Train set	Test set	Train set	Test set
Accuracy	0.994	0.938	0.984	0.924
F-1 score	0.995	0.951	0.988	0.939
Precision	0.993	0.929	0.986	0.941
Recall	0.997	0.974	0.989	0.937

Table 5 also shows the error metrics of the ANN classifier for the train and test set after hyperparameter tuning using Bayesian Optimization. The features used in this classification model were the same as in the ANN regressor. After 500 trials, the best results were obtained with 4 layers, with 192, 544, 64 and 1 neurons, with batch normalisation only in the last deep layer and dropout and LeakyReLU activation functions in all non-output layers. For gradient descent, an Adam optimizer, with 0.93 decay rate and initial learning rate of 10^{-3} was used.

As seen in Table 5, the best performance was obtained with the RF model, which led to an accuracy of 0.938 on the test set and very high recall (0.974), precision (0.929) and F-1 score (0.951). This suggests that 97 % of the times that a solvent led to an extraction yield equal to or >90 %, the model was able to correctly classify it. Table 6 shows the confusion matrix, where only 7 of the positive instances in the test set were wrongly classified as 0 by the RF model. Therefore, the random forest classifier was shown as an excellent tool for the accurate screening of the most promising solvents for pharmaceuticals extraction.

The feature importances were also calculated for the random forest classifier (see Table 7). The results showed that the five most important features were the solvent-to-feed ratio (0.182), followed by the n-octanol–water partition coefficient (0.179), the hydrogen bond contribution to the excess enthalpy (0.137), the Van der Waals contribution to the excess enthalpy (0.129) and the pH distance to the nearest pKa (0.117). These results are very similar to the feature importances obtained with

Table 6
Confusion matrix for the random forest and ANN classifiers (features: S/F, initial concentration, log10(P), H(HB), H(VDW), water matrix, solvent type, pH distance to closest pKa, charge and charge of the next species to the closest pKa; dependent variable: yield class – 1: yield $\geq$ 90 %, 0: yield $<$ 90 %).

		Random Forest Predicted		ANN Predicted	
		Negative	Positive	Negative	Positive
Actuals	Negative	142	20	145	16
	Positive	7	263	17	254

Table 7
Feature importances in the random forest classifier (features: S/F, initial concentration, log10(P), H(HB), H(VDW), water matrix, solvent type, pH distance to closest pKa, charge and charge of the next species to the closest pKa; dependent variable: yield class – 1: yield ≥ 90 %, 0: yield < 90 %).

Feature	Importance
S/F	0.182
Log10(P)	0.179
H(HB)	0.137
H(VDW)	0.129
pH distance to pKa	0.117
Eutectic solvent (dummy variable)	0.074
Conventional solvent (dummy variable)	0.050
Charge of the next species to the closest pKa	0.041
Water matrix (lab/hospital)	0.038
Charge	0.033
Initial concentration	0.021

the random forest regressor, with both models agreeing on the top 5 features, but in the classifier the hydrogen bond contribution was the third most important feature, whereas in the regressor it was the second.

To better understand the strengths and weaknesses of the classification model, the percentage of correct predictions made in the test set by solvent type (Table 8) and by pharmaceutical compounds (Table 9) were examined. As shown in Table 8, the highest accuracy was obtained for the eutectic solvents (95 %), followed by the conventional solvents (93 %) and the terpenoids (91 %). The fact that the lowest accuracy is obtained for the terpenoids is likely due to the tendency of COSMO-RS to overestimate the properties of terpenoids. This has been seen in different experimental trials in previous work on the extraction of organic molecules from aqueous effluents such as BPA [91], phenols [92], or azole fungicides [93], in which the extraction yields of terpenes were lower than those experienced in comparison with eutectic solvents when using the excess enthalpies and infinite dilution activity coefficients as a solvent screening tool.

In terms of the predictions by pharmaceutical compound (Table 9), the percentage of correct responses in the test set was above 90 % for all the pharmaceuticals except for carbamazepine (87 %). The best performance was obtained for the drugs tetracycline and ibuprofen, with 98 and 97 % of instances correctly predicted by the model in the test set, respectively. Considering these results, the drugs with the lowest predictability are phenazone and carbamazepine. These drugs coincide with the compounds, in which the pH that were operated, were always in the neutral state and the distance between pH and pKas are the largest with respect to the rest of the compounds. This lower accuracy might be due to the fact that we are using a single model and generalising for various types of pharmaceuticals, so it is possible that the inclusion of the distance between pH and pKas as a parameter in the model is improving the accuracy of the prediction for compounds affected by the pH, but reducing it for compounds not affected by it, as is the case for phenazone and carbamazepine.

4. Conclusions

In this work, we have used machine learning methodologies to predict the extraction yield of eleven pharmaceutical compounds from wastewater using eight solvents. The data was analysed using regression and classification analysis.

Table 8
Percentage of correct predictions obtained with the RF classifier by solvent type.

Solvent type	Percentage of correct predictions
Eutectic	95 %
Conventional	93 %
Terpenoid	91 %

Table 9
Percentage of correct predictions obtained with the RF classifier by pharmaceutical compound.

Compound	Percentage of correct predictions
Tetracycline	98 %
Ibuprofen	97 %
Sulfamethoxazole	96 %
Acetaminophen	95 %
Diclofenac	95 %
Ciprofloxacin	95 %
Atenolol	94 %
Trimethoprim	93 %
Naproxen	92 %
Phenazone	91 %
Carbamazepine	87 %

The results showed that the ANN and RF regression and classification models are excellent screening tools for the prediction of extraction yields of pharmaceuticals from wastewaters. These models also address shortcomings of previous prediction methodologies, as they incorporate key processing parameters such as the pH distance to the closest pKa. The RF and ANN classification models were proven as excellent candidates when a very high accuracy, precision and recall is needed but less detail on the response is required, whereas the RF and ANN regression models were more promising in applications needing a more detailed response (i.e. the actual yield value) but a less accurate prediction. RF models, compared to ANN, improve the interpretability of the results as they allow for the identification of the key model parameters and their importances. Thus, both the RF classification and regression models showed that the key features affecting the extraction yield were the solvent-to-feed ratio, the n-octanol–water partition coefficient, the hydrogen bond and Van der Waals contributions to the excess enthalpy and the pH distance to the nearest pKa. Therefore, the learnings obtained from the data modelling have allowed a better understanding of the impact of the processing and interaction parameters in the drugs removal and identification of the key factors affecting the process. This provides valuable insights into the variables to modify to improve the process during experimental or process design stages, as well as for more effective process control. These models have also been shown as excellent screening tools to identify the most suitable solvents for drugs removal from wastewater, contributing to the reduction of experimental time and costs.

Further work would be needed to generate additional data and improve the performance and generalisation of the regression models. Among them, the inclusion of additional data regarding antidepressant and blood pressure-regulating compounds, which were comparatively underrepresented in this study. Novel ANN architectures could be implemented, such as Physics-informed Neural Networks (PiNN) [94] which uses the physical laws as a regularisation technique during learning to improve generalisation and compensate the lack of data. Even more novel models such as Neural Operators (NOs) [95], which approximates hidden operators between function spaces, could be implemented, to obtain a discretisation-invariant function of yield depending on variables, for example the solute concentration. In addition, data from larger scale experiments should also be generated to understand the impact of scaling up on the process; for that stage, data-centric approaches would need to be considered for the optimisation of the machine learning models [96].

In conclusion, applying machine learning to the extraction of solutes from aqueous solutions allows a fast and accurate approach of predicting extraction yields for given conditions. The inputs for this system are obtained from molecular simulation. This tool would not only allow screening solvents that have not yet been applied or solutes that have not yet been studied for extraction, but it could also be incorporated into process simulation, using the predicted extraction yields in larger scale or techno-economic simulations.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jwpe.2024.105379.

CRediT authorship contribution statement

Ana Casas: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Diego Rodríguez-Llorente:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Conceptualization. **Guillermo Rodríguez-Llorente:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Juan García:** Writing – review & editing, Project administration, Funding acquisition. **Marcos Larriba:** Writing – review & editing, Writing – original draft, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

Data and code will be made available upon reasonable request.

Acknowledgments

This work was supported by Comunidad Autónoma de Madrid [project numbers P2018/EMT-4341 and PR65/19-22441]. Diego Rodríguez-Llorente thanks Ministerio de Ciencia, Innovación y Universidades for awarding an FPU grant (FPU18/01536).

References

[1] K. Kümmerer, The presence of pharmaceuticals in the environment due to human use - present knowledge and future challenges, J. Environ. Manage. 90 (2009) 2354–2366, https://doi.org/10.1016/j.jenvman.2009.01.023.
[2] L. Castillo Meza, P. Piotrowski, J. Farnan, T.L. Tasker, B. Xiong, B. Weggler, K. Murrell, F.L. Dorman, J.P. Vanden Heuvel, W.D. Burgos, Detection and removal of biologically active organic micropollutants from hospital wastewater, Sci. Total Environ. 700 (2020) 134469, https://doi.org/10.1016/j.scitotenv.2019.134469.
[3] A.S. Adeleye, J. Xue, Y. Zhao, A.A. Taylor, J.E. Zenobio, Y. Sun, Z. Han, O. A. Salawu, Y. Zhu, Abundance, fate, and effects of pharmaceuticals and personal care products in aquatic environments, J. Hazard. Mater. 424 (2022), https://doi.org/10.1016/j.jhazmat.2021.127284.
[4] A. Majumder, A.K. Gupta, P.S. Ghosal, M. Varma, A review on hospital wastewater treatment: a special emphasis on occurrence and removal of pharmaceutically active compounds, resistant microorganisms, and SARS-CoV-2, J. Environ. Chem. Eng. 9 (2021) 104812, https://doi.org/10.1016/j.jece.2020.104812.
[5] D. Papagiannaki, M.H. Belay, N.P.F. Gonçalves, E. Robotti, A. Bianco-Prevot, R. Binetti, P. Calza, From monitoring to treatment, how to improve water quality: the pharmaceuticals case, Chemical Engineering Journal Advances 10 (2022) 100245, https://doi.org/10.1016/j.ceja.2022.100245.
[6] A.S. Oberoi, Y. Jia, H. Zhang, S.K. Khanal, H. Lu, Insights into the fate and removal of antibiotics in engineered biological treatment systems: a critical review, Environ. Sci. Technol. 53 (2019) 7234–7264, https://doi.org/10.1021/acs.est.9b01131.
[7] R. Tormo-Budowski, J.C. Cambronero-Heinrichs, J.E. Durán, M. Masís-Mora, D. Ramírez-Morales, J.P. Quirós-Fournier, C.E. Rodríguez-Rodríguez, Removal of pharmaceuticals and ecotoxicological changes in wastewater using Trametes versicolor: a comparison of fungal stirred tank and trickle-bed bioreactors, Chem. Eng. J. 410 (2021) 128210, https://doi.org/10.1016/j.cej.2020.128210.
[8] A.C. Álamo, C. González, M.I. Pariente, R. Molina, F. Martínez, Fenton-like catalyst based on a reticulated porous perovskite material: activity and stability for the on-site removal of pharmaceutical micropollutans in a hospital wastewater, Chem. Eng. J. 401 (2020) 126113, https://doi.org/10.1016/j.cej.2020.126113.
[9] G.A. Kifle, Y. Huang, M. Xiang, W. Wang, C. Wang, C. Li, H. Li, Heterogeneous activation of peroxygens by iron-based bimetallic nanostructures for the efficient remediation of contaminated water. A review, Chemical Engineering Journal 442 (2022) 136187, https://doi.org/10.1016/j.cej.2022.136187.
[10] Y. Feng, Y. Tao, Q. Meng, J. Qu, S. Ma, S. Han, Y. Zhang, Microwave-combined advanced oxidation for organic pollutants in the environmental remediation: an overview of influence, mechanism, and prospective, Chem. Eng. J. 441 (2022) 135924, https://doi.org/10.1016/j.cej.2022.135924.
[11] K. Fedorov, K. Dinesh, X. Sun, R. Darvishi Cheshmeh Soltani, Z. Wang, S. Sonawane, G. Boczkaj, Synergistic effects of hybrid advanced oxidation processes (AOPs) based on hydrodynamic cavitation phenomenon – a review, Chem. Eng. J. 432 (2022), https://doi.org/10.1016/j.cej.2021.134191.
[12] M.F. Dapaah, Q. Niu, Y.Y. Yu, T. You, B. Liu, L. Cheng, Efficient persistent organic pollutant removal in water using MIL-metal–organic framework driven Fenton-like reactions: a critical review, Chem. Eng. J. 431 (2022), https://doi.org/10.1016/j.cej.2021.134182.
[13] J.R. de Andrade, M.F. Oliveira, M.G.C. da Silva, M.G.A. Vieira, Adsorption of pharmaceuticals from water and wastewater using nonconventional low-cost materials: a review, Ind. Eng. Chem. Res. 57 (2018) 3103–3127, https://doi.org/10.1021/acs.iecr.7b05137.
[14] P. Krasucka, B. Pan, Y. Sik Ok, D. Mohan, B. Sarkar, P. Oleszczuk, Engineered biochar – a sustainable solution for the removal of antibiotics from water, Chem. Eng. J. 405 (2021) 126926, https://doi.org/10.1016/j.cej.2020.126926.
[15] G. Alam, I. Ihsanullah, M. Naushad, M. Sillanpää, Applications of artificial intelligence in water treatment for optimization and automation of adsorption processes: recent advances and prospects, Chem. Eng. J. 427 (2022), https://doi.org/10.1016/j.cej.2021.130011.
[16] C. Florindo, F. Lima, L.C. Branco, I.M. Marrucho, Hydrophobic deep eutectic solvents: a circular approach to purify water contaminated with ciprofloxacin, ACS Sustain Chem Eng. 7 (2019) 14739–14746, https://doi.org/10.1021/acssuschemeng.9b02658.
[17] H. Sereshti, G. Abdolhosseini, S. Soltani, F. Jamshidi, N. Nouri, Natural thymol-based ternary deep eutectic solvents: application in air-bubble assisted-dispersive liquid-liquid microextraction for the analysis of tetracyclines in water, J. Sep. Sci. 44 (2021) 3626–3635, https://doi.org/10.1002/jssc.202100495.
[18] F. Bergua, M. Castro, J. Muñoz-Embid, C. Lafuente, M. Artal, Hydrophobic eutectic solvents: thermophysical study and application in removal of pharmaceutical products from water, Chem. Eng. J. 411 (2021) 128472, https://doi.org/10.1016/j.cej.2021.128472.
[19] A.G. Pekel, E. Kurtulbaş, İ. Toprakçı, S. Şahin, Menthol-based deep eutectic solvent for the separation of carbamazepine: reactive liquid-liquid extraction, Biomass Convers. Biorefinery 12 (2022) 1249–1256, https://doi.org/10.1007/s13399-020-00707-z.
[20] F. Bergua, M. Castro, J. Muñoz-Embid, C. Lafuente, M. Artal, L-menthol-based eutectic solvents: characterization and application in the removal of drugs from water, J. Mol. Liq. 352 (2022) 118754, https://doi.org/10.1016/j.molliq.2022.118754.
[21] P. Gutiérrez-Sánchez, D. Rodríguez-Llorente, P. Navarro, V.I. Águeda, S. Álvarez-Torrellas, J. García, M. Larriba, Extraction of antibiotics identified in the EU Watch List 2020 from hospital wastewater using hydrophobic eutectic solvents and terpenoids, Sep. Purif. Technol. 282 (2022) 120117, https://doi.org/10.1016/j.seppur.2021.120117.
[22] D. Rodríguez-Llorente, Y. Liu, J.C. de Mello, J. García, M. Larriba, Flow-based recovery of acetic acid from aqueous solutions using bio-derived terpenes as extracting solvents, Sep. Purif. Technol. 324 (2023), https://doi.org/10.1016/j.seppur.2023.124524.
[23] L. Liu, H. Fang, Q. Wei, X. Ren, Extraction performance evaluation of amide-based deep eutectic solvents for carboxylic acid: molecular dynamics simulations and a mini-pilot study, Sep. Purif. Technol. 304 (2023) 122360, https://doi.org/10.1016/j.seppur.2022.122360.
[24] D. Rodríguez-Llorente, E. Hernández, P. Gutiérrez-Sánchez, P. Navarro, V. Ismael Águeda, S. Álvarez-Torrellas, J. García, M. Larriba, Extraction of pharmaceuticals from hospital wastewater with eutectic solvents and terpenoids: computational, experimental, and simulation studies, Chem. Eng. J. 451 (2023) 138544, https://doi.org/10.1016/J.CEJ.2022.138544.
[25] A.S. Darwish, T. Lemaoui, J. AlYammahi, H. Taher, Y. Benguerba, F. Banat, I. M. AlNashef, Molecular insights into potential hydrophobic deep eutectic solvents for furfural extraction guided by COSMO-RS and machine learning, J. Mol. Liq. 379 (2023) 121631, https://doi.org/10.1016/J.MOLLIQ.2023.121631.
[26] A. Klamt, COSMO-RS: from quantum chemistry to fluid phase thermodynamics, Computer Aided Chemical Engineering 43 (2018) 9, https://doi.org/10.1016/B978-0-444-64235-6.50003-6.
[27] D. Wang, M. Zhang, C.L. Law, L. Zhang, Natural deep eutectic solvents for the extraction of lentinan from shiitake mushroom: COSMO-RS screening and ANN-GA optimizing conditions, Food Chem. 430 (2024) 136990, https://doi.org/10.1016/J.FOODCHEM.2023.136990.
[28] M. Lazović, I. Cvijetić, M. Jankov, D. Milojković-Opsenica, J. Trifković, P. Ristivojević, COSMO-RS in prescreening of Natural Eutectic Solvents for phenolic extraction from Teucrium chamaedrys, J. Mol. Liq. 387 (2023) 122649, https://doi.org/10.1016/J.MOLLIQ.2023.122649.
[29] D.S. Raut, V.A. Joshi, S. Khan, D. Kundu, A-Priori screening of deep eutectic solvent for enhanced oil recovery application using COSMO-RS framework, J. Mol. Liq. 377 (2023) 121482, https://doi.org/10.1016/J.MOLLIQ.2023.121482.
[30] C. Chen, Y. Cao, A. Ali, S. Toufouki, S. Yao, How to apply terpenoid-based deep eutectic solvents for removal of antibiotics and dyes from water: theoretical prediction, experimental validation and quantum chemical evaluation, Environ. Res. 231 (2023) 116180, https://doi.org/10.1016/J.ENVRES.2023.116180.
[31] R. Santiago, I. Díaz, M. González-Miquel, P. Navarro, J. Palomar, Assessment of bio-ionic liquids as promising solvents in industrial separation processes: computational screening using COSMO-RS method, Fluid Phase Equilib. 560 (2022) 113495, https://doi.org/10.1016/J.FLUID.2022.113495.
[32] E. Hernández, R. Santiago, A. Belinchón, G. Maria Vaquerizo, C. Moya, P. Navarro, J. Palomar, Universal and low energy-demanding platform to produce propylene

carbonate from CO2 using hydrophilic ionic liquids, Sep. Purif. Technol. 295 (2022) 121273, https://doi.org/10.1016/J.SEPPUR.2022.121273.
[33] J. Reinisch, M. Diedenhofen, R. Wilcken, A. Udvarhelyi, A. Glöß, Benchmarking different QM levels for usage with COSMO-RS, J. Chem. Inf. Model. 59 (2019) 4806–4813, https://doi.org/10.1021/acs.jcim.9b00659.
[34] K. Paduszyński, U. Domańska, COSMO-RS screening for ionic liquid to be applied in extraction of 2-phenylethanol from aqueous solutions, J. Mol. Liq. 271 (2018) 305–312, https://doi.org/10.1016/j.molliq.2018.07.111.
[35] A. Klamt, The COSMO and COSMO-RS solvation models, WIREs Comput. Mol. Sci. 8 (2018), https://doi.org/10.1002/wcms.1338.
[36] H. Chang, F. Zhong, R. Cao, Y. Liu, M. Wang, X. Pu, Prediction of the solubilities of water in hydrocarbons with COSMO-RS and interpretation of the solubility characteristics, J. Solution Chem. 49 (2020) 365–382, https://doi.org/10.1007/s10953-020-00966-4.
[37] R. Franke, B. Hannebauer, On the influence of basis sets and quantum chemical methods on the prediction accuracy of COSMO-RS, Phys. Chem. Chem. Phys. 13 (2011) 21344, https://doi.org/10.1039/c1cp22317h.
[38] Y. Liu, Z. Dai, Z. Zhang, S. Zeng, F. Li, X. Zhang, Y. Nie, L. Zhang, S. Zhang, X. Ji, Ionic liquids/deep eutectic solvents for CO2 capture: reviewing and evaluating, Green Energy & Environment 6 (2021) 314–328, https://doi.org/10.1016/j.gee.2020.11.024.
[39] E. Villarroel, F. Olea, C. Araya-López, G. Merlet, R. Cabezas, J. Romero, E. Quijada-Maldonado, COSMO-RS evaluation as a tool for prediction of solvents in dispersive liquid-phase microextraction: evaluation of conventional solvents and ionic liquids as extractants, J. Mol. Liq. 354 (2022) 118861, https://doi.org/10.1016/j.molliq.2022.118861.
[40] M. Klajmon, Purely predicting the pharmaceutical solubility: what to expect from PC-SAFT and COSMO-RS? Mol. Pharm. 19 (2022) 4212–4232, https://doi.org/10.1021/acs.molpharmaceut.2c00573.
[41] P.P. Mondal, A. Galodha, V.K. Verma, V. Singh, P.L. Show, M.K. Awasthi, B. Lall, S. Anees, K. Pollmann, R. Jain, Review on machine learning-based bioprocess optimization, monitoring, and control systems, Bioresour. Technol. 370 (2023) 128523, https://doi.org/10.1016/J.BIORTECH.2022.128523.
[42] A.S. Qambar, M.M.M. Al Khalidy, Optimizing dissolved oxygen requirement and energy consumption in wastewater treatment plant aeration tanks using machine learning, J. Water Proc. Eng. 50 (2022) 103237, https://doi.org/10.1016/j.jwpe.2022.103237.
[43] J.A. Okolie, S. Savage, C.C. Ogbaga, B. Gunes, Assessing the potential of machine learning methods to study the removal of pharmaceuticals from wastewater using biochar or activated carbon, Total Environment Research Themes 1–2 (2022) 100001, https://doi.org/10.1016/J.TOTERT.2022.100001.
[44] S.K. Bhagat, K.E. Pilario, O.E. Babalola, T. Tiyasha, M. Yaqub, C.E. Onu, K. Pyrgaki, M.W. Falah, A.H. Jawad, D.A. Yaseen, N. Barka, Z.M. Yaseen, Comprehensive review on machine learning methodologies for modeling dye removal processes in wastewater, J. Clean. Prod. 385 (2023) 135522, https://doi.org/10.1016/J.JCLEPRO.2022.135522.
[45] T.K. Chakraborty, S. Ghosh, M.S. Islam, M.S. Nice, K.R. Islam, B.N. Netema, M. S. Rahman, A. Habib, S. Zaman, G. Chandra Ghosh, M.R. Hossain, K. Tul-Coubra, K. Adhikary, A. Munna, M.M. Haque, H. Bosu, M. Halder, Removal of hazardous textile dye from simulated wastewater by municipal organic solid waste charcoal using machine learning approaches: kinetics, isotherm, and thermodynamics, Heliyon 9 (2023) e18856, https://doi.org/10.1016/J.HELIYON.2023.E18856.
[46] V. Godvin Sharmila, V. Kumar Tyagi, S. Varjani, J. Rajesh Banu, A review on the lignocellulosic derived biochar-based catalyst in wastewater remediation: advanced treatment technologies and machine learning tools, Bioresour. Technol. 387 (2023) 129587, https://doi.org/10.1016/J.BIORTECH.2023.129587.
[47] A.S. Qambar, M.M.M. Al Khalidy, Development of local and global wastewater biochemical oxygen demand real-time prediction models using supervised machine learning algorithms, Eng. Appl. Artif. Intel. 118 (2023) 105709, https://doi.org/10.1016/J.ENGAPPAI.2022.105709.
[48] F. Bellamoli, M. Di Iorio, M. Vian, F. Melgani, Machine learning methods for anomaly classification in wastewater treatment plants, J. Environ. Manage. 344 (2023) 118594, https://doi.org/10.1016/J.JENVMAN.2023.118594.
[49] R.K. Oruganti, A.P. Biji, T. Lanuyanger, P.L. Show, M. Sriariyanun, V.K. K. Upadhyayula, V. Gadhamshetty, D. Bhattacharyya, Artificial intelligence and machine learning tools for high-performance microalgal wastewater treatment and algal biorefinery: a critical review, Sci. Total Environ. 876 (2023) 162797, https://doi.org/10.1016/J.SCITOTENV.2023.162797.
[50] N.K. Singh, M. Yadav, V. Singh, H. Padhiyar, V. Kumar, S.K. Bhatia, P.L. Show, Artificial intelligence and machine learning-based monitoring and design of biological wastewater treatment systems, Bioresour. Technol. 369 (2023) 128486, https://doi.org/10.1016/J.BIORTECH.2022.128486.
[51] M. Zhu, J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu, L. Ye, A review of the application of machine learning in water quality evaluation, Eco-Environment & Health 1 (2022) 107–116, https://doi.org/10.1016/J.EEHL.2022.06.001.
[52] M. Alvi, D. Batstone, C.K. Mbamba, P. Keymer, T. French, A. Ward, J. Dwyer, R. Cardell-Oliver, Deep learning in wastewater treatment: a critical review, Water Res. (2023) 120518, https://doi.org/10.1016/J.WATRES.2023.120518.
[53] S. Safeer, R.P. Pandey, B. Rehman, T. Safdar, I. Ahmad, S.W. Hasan, A. Ullah, A review of artificial intelligence in water purification and wastewater treatment: recent advancements, Journal of Water Process Engineering. 49 (2022) 102974, https://doi.org/10.1016/J.JWPE.2022.102974.
[54] A. Soo, L. Wang, C. Wang, H.K. Shon, MachIne learning for nutrient recovery in the smart city circular economy – a review, Process Saf. Environ. Prot. 173 (2023) 529–557, https://doi.org/10.1016/J.PSEP.2023.02.065.
[55] A. Roosta, R. Haghbakhsh, A.R.C. Duarte, S. Raeissi, Machine learning coupled with group contribution for predicting the density of deep eutectic solvents, Fluid Phase Equilib. (2022) 113672, https://doi.org/10.1016/j.fluid.2022.113672.
[56] J. Li, J.L. Anderson, E.A. Smith, Determination of infinite dilution activity coefficients of molecular solutes in ionic liquids and deep eutectic solvents by factorization-machine-based neural networks, ACS Sustain. Chem. Eng. 10 (2022) 13927–13935, https://doi.org/10.1021/acssuschemeng.2c02600.
[57] N.E. Awaja, G. Almustafa, A.S. Darwish, T. Lemaoui, Y. Benguerba, F. Banat, H. A. Arafat, I. AlNashef, Molecular-based artificial neural networks for selecting deep eutectic solvents for the removal of contaminants from aqueous media, Chem. Eng. J. 476 (2023), https://doi.org/10.1016/j.cej.2023.146429.
[58] F. Stulp, O. Sigaud, Many regression algorithms, one unified model: a review, Neural Netw. 69 (2015) 60–79, https://doi.org/10.1016/J.NEUNET.2015.05.005.
[59] L. Díaz-González, O. Alejandro Uscanga-Junco, M. Rosales-Rivera, Development and comparison of machine learning models for water multidimensional classification, J Hydrol (Amst) 598 (2021) 126234, https://doi.org/10.1016/J.JHYDROL.2021.126234.
[60] A. Elsayed, S. Rixon, J. Levison, A. Binns, P. Goel, Application of classification machine learning algorithms for characterizing nutrient transport in a clay plain agricultural watershed, J. Environ. Manage. 345 (2023) 118924, https://doi.org/10.1016/J.JENVMAN.2023.118924.
[61] A. Malloum, K.A. Adegoke, J.O. Ighalo, J. Conradie, C.R. Ohoro, J.F. Amaku, K. O. Oyedotun, N.W. Maxakato, K.G. Akpomie, E.S. Okeke, C. Olisah, Computational methods for adsorption study in wastewater treatment, J. Mol. Liq. (2023) 123008, https://doi.org/10.1016/J.MOLLIQ.2023.123008.
[62] M. Bahramian, R.K. Dereli, W. Zhao, M. Giberti, E. Casey, Data to intelligence: the role of data-driven models in wastewater treatment, Expert Syst. Appl. 217 (2023) 119453, https://doi.org/10.1016/J.ESWA.2022.119453.
[63] R. Gatchalian, P. Tsvetkov, Data to intelligence: traditional machine learning methods in predicting the physics of subcritical systems in source-equilibrium, Ann. Nucl. Energy 201 (2024) 110413, https://doi.org/10.1016/j.anucene.2024.110413.
[64] F. Eckert, A. Klamt, Fast solvent screening via quantum chemistry: COSMO-RS approach, AICHE J. 48 (2002) 369–385, https://doi.org/10.1002/aic.690480220.
[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.
[66] A. Gulli, S. Pal, Deep Learning With Keras, 2017, p. 318. https://www.oreilly.com/library/view/deep-learning-with/9781787128422/ (accessed October 28, 2023).
[67] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, G. Research, TensorFlow: large-scale machine learning on heterogeneous distributed systems. https://arxiv.org/abs/1603.04467v2, 2016 (accessed October 28, 2023).
[68] K.I. Sundus, B.H. Hammo, M.B. Al-Zoubi, A. Al-Omari, Solving the multicollinearity problem to improve the stability of machine learning algorithms applied to a fully annotated breast cancer dataset, Inform Med Unlocked. 33 (2022) 101088, https://doi.org/10.1016/J.IMU.2022.101088.
[69] X.C. Nguyen, T.T.H. Nguyen, Q.V. Le, P.C. Le, A.L. Srivastav, Q.B. Pham, P. M. Nguyen, D.D. La, E.R. Rene, H.H. Ngo, S.W. Chang, D.D. Nguyen, Developing a new approach for design support of subsurface constructed wetland using machine learning algorithms, J. Environ. Manage. 301 (2022) 113868, https://doi.org/10.1016/J.JENVMAN.2021.113868.
[70] N. Istiqomah, D. Jung, J. Khim, Tree-based ensemble machine learning model for nitrate reduction by zero-valent iron, Journal of Water Process Engineering 56 (2023) 104303, https://doi.org/10.1016/j.jwpe.2023.104303.
[71] I. Ibrar, S. Yadav, A. Altaee, A. Braytee, A.K. Samal, S.M. Javaid, A.H. Hawari, A machine learning approach for prediction of reverse solute flux in forward osmosis, Journal of Water Process Engineering 54 (2023) 103956, https://doi.org/10.1016/J.JWPE.2023.103956.
[72] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, Adv Neural Inf Process Syst. 4 (2012) 2951–2959. https://arxiv.org/abs/1206.2944v2 (accessed October 28, 2023).
[73] R. Wang, Y. Yu, Y. Chen, Z. Pan, X. Li, Z. Tan, J. Zhang, Model construction and application for effluent prediction in wastewater treatment plant: data processing method optimization and process parameters integration, J. Environ. Manage. 302 (2022) 114020, https://doi.org/10.1016/J.JENVMAN.2021.114020.
[74] S.M.Z. Hossain, N. Sultana, M.S. Jassim, G. Coskuner, L.M. Hazin, S.A. Razzak, M. M. Hossain, Soft-computing modeling and multiresponse optimization for nutrient removal process from municipal wastewater using microalgae, Journal of Water Process Engineering 45 (2022) 102490, https://doi.org/10.1016/J.JWPE.2021.102490.
[75] L. Vaughan, M. Zhang, H. Gu, J.B. Rose, C.C. Naughton, G. Medema, V. Allan, A. Roiko, L. Blackall, A. Zamyadi, An exploration of challenges associated with machine learning for time series forecasting of COVID-19 community spread using wastewater-based epidemiological data, Sci. Total Environ. 858 (2023) 159748, https://doi.org/10.1016/J.SCITOTENV.2022.159748.
[76] A.P. De Miranda Ramos Soares, F. De Oliveira Carvalho, C.E. De Farias Silva, A.H. Da Silva Gonçalves, A.K. De Souza Abud, Random Forest as a promising application

to predict basic-dye biosorption process using orange waste, J. Environ. Chem. Eng. 8 (2020) 103952, https://doi.org/10.1016/J.JECE.2020.103952.
[77] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366, https://doi.org/10.1016/0893-6080(89)90020-8.
[78] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536, https://doi.org/10.1038/323533a0.
[79] H. Taud, J.F. Mas, Multilayer Perceptron (MLP), 2018, pp. 451–455, https://doi.org/10.1007/978-3-319-60801-3_27.
[80] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: 32nd International Conference on Machine Learning, ICML 2015 vol. 1, 2015, pp. 448–456. https://arxiv.org/abs/1502.03167v3 (accessed October 28, 2023).
[81] N. Srivastava, G. Hinton, A. Krizhevsky, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.
[82] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, X. Liang, An improved random forest based on the classification accuracy and correlation measurement of decision trees, Expert Syst Appl. (2023) 121549, https://doi.org/10.1016/J.ESWA.2023.121549.
[83] X. Lin, M. Wu, X. Shao, G. Li, Y. Hong, Water turbidity dynamics using random forest in the Yangtze River Delta Region, China, Science of The Total Environment. 903 (2023) 166511, https://doi.org/10.1016/J.SCITOTENV.2023.166511.
[84] S.R. Dubey, S.K. Singh, B.B. Chaudhuri, Activation functions in deep learning: a comprehensive survey and benchmark, Neurocomputing 503 (2021) 92–108, https://doi.org/10.1016/j.neucom.2022.06.111.
[85] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, J. Gonzalez-Rodriguez, Deconstructing cross-entropy for probabilistic binary classifiers, Entropy 20 (2018) 208, https://doi.org/10.3390/E20030208.
[86] M. Billah, A.K.M.S. Islam, W. Bin Mamoon, M.R. Rahman, Random forest classifications for landuse mapping to assess rapid flood damage using Sentinel-1 and Sentinel-2 data, Remote Sens Appl. 30 (2023) 100947, https://doi.org/10.1016/J.RSASE.2023.100947.
[87] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2014. https://arxiv.org/abs/1412.6980v9 (accessed October 28, 2023).
[88] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier Nonlinearities Improve Neural Network Acoustic Models, 2013.
[89] B. Xiao, H. Wu, X. Zhang, R. Wu, Y. Liu, A novel approach for the open-circuit voltage estimation of lithium-ion batteries by epsilon SVR, Int. J. Electrochem. Sci. 17 (2022) 22059, https://doi.org/10.20964/2022.05.14.
[90] G.O. Anyanwu, C.I. Nwakanma, J.M. Lee, D.S. Kim, RBF-SVM kernel-based model for detecting DDoS attacks in SDN integrated vehicular network, Ad Hoc Netw. 140 (2023) 103026, https://doi.org/10.1016/J.ADHOC.2022.103026.
[91] D. Rodríguez-Llorente, P. Navarro, R. Santiago, V.I. Águeda, S. Álvarez-Torrellas, J. García, M. Larriba, Extractive removal and recovery of bisphenol A from aqueous solutions using terpenoids and hydrophobic eutectic solvents, J. Environ. Chem. Eng. 9 (2021) 106128, https://doi.org/10.1016/j.jece.2021.106128.
[92] D. Rodríguez-Llorente, D. Martín-Gutiérrez, P. Suárez-Rodríguez, P. Navarro, S. Álvarez-Torrellas, J. García, M. Larriba, Sustainable recovery of phenolic antioxidants from real olive vegetation water with natural hydrophobic eutectic solvents and terpenoids, Environ. Res. 220 (2023) 115207, https://doi.org/10.1016/j.envres.2022.115207.
[93] D. Rodríguez-Llorente, E. García-Fernández de la Puente, J. García, M. Larriba, Novel approach for azole fungicides extraction from aqueous environments using terpenes and eutectic solvents, J. Environ. Chem. Eng. 11 (2023), https://doi.org/10.1016/j.jece.2023.110562.
[94] S.A. Faroughi, N. Pawar, C. Fernandes, M. Raissi, S. Das, N.K. Kalantari, S. K. Mahjour, Physics-guided, physics-informed, and physics-encoded neural networks in scientific computing. http://arxiv.org/abs/2211.07377, 2022 (accessed May 28, 2023).
[95] N. Kovachki, Z. Li, Z. Edu, C.B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, A. Anandkumar, A. Edu, L. Rosasco, Neural operator: learning maps between function spaces, Journal of Machine Learning Research 23 (2021) 1–97. https://arxiv.org/abs/2108.08481v5 (accessed June 5, 2023).
[96] G. Liu, Z.D. Savic, G. Fu, Short-term water demand forecasting using data-centric machine learning approaches, J. Hydroinf. 25 (2023) 895–911, https://doi.org/10.2166/hydro.2023.163 (accessed April 4, 2024).



https://dspace.org/

Machine learning screening tools for the prediction of extraction yields of pharmaceutical compounds from wastewaters

Casas, Ana; Rodríguez-Llorente, Diego; Rodríguez-Llorente, Guillermo; García, Juan; Larriba, Marcos

2024-05-01
Elsevier
https://doi.org/10.1016/j.jwpe.2024.105379
Casas A, Rodríguez-Llorente D, Rodríguez-Llorente G, et al., (2024) Machine learning screening tools for the prediction of extraction yields of pharmaceutical compounds from wastewaters. Journal of Water Process Engineering, Volume 62, May 2024, Article number 105379

Downloaded from a DSpace repository. https://dspace.org/