Article

Prediction of pH Value of Aqueous Acidic and Basic Deep Eutectic Solvent Using COSMO-RS σ Profiles’ Molecular Descriptors

Manuela Panić 1, Mia Radović 1, Marina Cvjetko Bubalo 1, Kristina Radošević 1, Marko Rogošić 2, João A. P. Coutinho 3, Ivana Radojčić Redovniković 1,* and Ana Jurinjak Tušek 1

1 Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva ulica 6, 10000 Zagreb, Croatia; mpanic@pbf.hr (M.P.); mradovic@pbf.hr (M.R.); mcvjetko@pbf.hr (M.C.B.); krado@pbf.hr (K.R.); ana.tusek.jurinjak@pbf.unizg.hr (A.J.T.)
2 Faculty of Chemical Engineering and Technology, University of Zagreb, Marulićev trg 19, 10000 Zagreb, Croatia; mrogosic@fkit.hr
3 CICECO—Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal; jcoutinho@ua.pt
* Correspondence: irredovnikovic@pbf.hr

Abstract: The aim of this work was to develop a simple and easy-to-apply model to predict the pH values of deep eutectic solvents (DESs) over a wide range of pH values that can be used in daily work. For this purpose, the pH values of 38 different DESs were measured (ranging from 0.36 to 9.31) and mathematically interpreted. To develop mathematical models, DESs were first numerically described using σ profiles generated with the COSMOtherm software. After the DESs’ description, the following models were used: (i) multiple linear regression (MLR), (ii) piecewise linear regression (PLR), and (iii) artificial neural networks (ANNs) to link the experimental values with the descriptors. Both PLR and ANN were found to be applicable to predict the pH values of DESs with a very high goodness of fit ($R^2_{\text{independent validation}} > 0.8600$). Due to the good mathematical correlation of the experimental and predicted values, the σ profile generated with COSMOTHERM could be used as a DES molecular descriptor for the prediction of their pH values.

Keywords: artificial neural networks; COSMO-RS; deep eutectic solvents; multiple linear regression; piecewise linear regression

1. Introduction

Green chemistry presents a way of creating and applying chemical products and processes that reduce or eliminate the use or production of substances that are hazardous to human health and the environment [1]. A growing area of research in green technology development is devoted to the design of new, more environmentally friendly solvents whose use would meet technological and economic requirements. Requirements for alternative solvents include a reasonable price, non-toxicity to humans and the environment, non-flammability, biodegradability, and possibility of regeneration or recovery [2,3]. Currently, known green solvents are water, carbon dioxide, bio-solvents, ionic liquids, and deep eutectic solvents. In the last decade, deep eutectic solvents (DESs) have received enormous attention in the academic community and the number of articles published has increased exponentially.
DEs were first described by Abbott et al. in 2003 as a mixture of a hydrogen bond

donor (HBD) with a hydrogen bond acceptor (HBA), which exhibited much lower melting

definitions due to the formation of hydrogen bonds between constituting compounds [4–6]. Lately, DEs have shown great potential for industrial application

thanks to their acceptable costs, the versatility of their physicochemical properties,
and simple preparation. They also often present low cytotoxicity and good biodegrada-
ility. The properties that have gained them the environmentally friendly label are low
volatility (reduced air pollution), nonflammability (process safety), and stability (potential
for recycling and reuse). The number of structural combinations encompassed by DEs
is tremendous; thus, it is possible to design DEs with unique physicochemical properties
for a particular purpose. The physicochemical properties, such as the viscosity, density,
and pH value, of DEs are crucial for industrial application of these solvents in terms of

equipment materials, mass transfer, filtration, or pumping [7].

The pH values of aqueous solutions affect the enzyme activity, extraction efficiency,
and stability of biologically active molecules. As such, the pH value is an important property
of a solvent and, especially for DE design, one of the critical parameters. Though
several papers have analyzed the pH behavior of DEs, there are still gaps in the under-
standing of how DE-forming compounds influence its pH value [8,9]. Despite this, some
general conclusions can be outlined. For example, DEs containing organic acids (i.e.,
malic acid or oxalic acid) are, as expected, more acidic than those containing polyalcohols
or sugars. The role of the water content in DEs regarding the pH behavior is still not









entirely clear; however, it was observed that an increase in pH values with an increasing
water content was reported for DEs with extremely low pH values while the pH values
of DEs with pH in the higher range of values (lower acidity region) decreased with an
increasing water content [7].

So far, the search for an ideal DE for a particular system has been guided by an
empirical trial-and-error approach, with no systematic research into the structure–activity
of DEs. Therefore, the rational design of these solvents for specific purposes is still in its
infancy. Data collection on the application properties of DEs and the development of
mathematical methods as a tool for the design of novel solvents are imperative for the
industrial application of these solvents. The Conductor-like Screening Model for Real Sol-
vents (COSMO-RS) is an ab initio computational method that may be used for the gener-

dation of the σ profile of a molecule. The σ profile shows the probability of finding surface
segments with σ polarity on the surface of the molecule and contains the most relevant
chemical information needed to predict the compound’s electrostatic, hydrogen bonding,
and dispersion interactions [10]. The distribution of the charge, the width, and the height
of the peaks in the σ profile vary with the nature of the molecules. Therefore, any change
in the molecular structure can be quantified. By coupling the σ profile of DE-forming
compounds with experimental data using model-generating methods such as multiple
linear regression (MLR), piecewise linear regression (PLR), or artificial neural networks
(ANNs), models for the description of DEs’ physicochemical properties can be devel-
oped [11–14]. In most studies, good model fitting of the literature viscosity, density, and
pH values of the DEs was obtained [12,13]. The results showed that simple linear models
such as MLR and more complex ones such as ANN could be used efficiently to predict the
physical properties of specific DE groups (e.g., amine or sugar-based DEs), whereas it
was difficult to create a single model covering the whole range of possible DE systems
[11]. Commonly, simple mathematical models such as MLR were good enough for viscose-
tivity and density prediction while in the case of the pH value, more complex ANN models
had to be used [11,13,15].
In this work, we report a model for the prediction of the pH values of acidic and basic DESs. For this purpose, the experimental pH values of 38 different DESs were evaluated, described, and mathematically interpreted. For the development of mathematical models, DESs were firstly numerically described using σ profiles estimated by the COSMOtherm software. After the description of DESs, the following models were used: (i) MLR, (ii) PLR, and (iii) ANN to link the experimental values with the descriptors. In the end, the prepared models were statistically verified.

2. Results and Discussion

2.1. DES Characteristics: Experimental pH Values and σ Profiles

This work aimed to develop a simple and robust mathematical model for predicting the pH values of DESs based on \( S_{\text{mix}} \) descriptors. To develop a user-friendly model to predict pH values in the wide range, we selected both acidic and basic DESs from our database. We chose 38 DESs by carefully selecting and varying different HBA, HBD, and water shares (Table 1). Selected HBAs and HBDs can be roughly classified as quaternary ammonium salts (choline chloride, betaine), amino acids (proline), organic acids (citric and malic acid), and sugars (fructose, glucose, sucrose, xylose). In comparison to HBA, there are more HBD candidates from previously mentioned classes and it has been shown that they have an immediate effect on pH values (Table 1). Overall, all synthesized DESs cover a wide range of pH values from 0.36 for Ch:CA containing 30% water \((w/w)\) to 9.31 for Ch:U containing 10% water \((w/w)\). Monitoring the pH values of the same HBA/HBD pair while varying the DES water content shows that water influences the measured pH value. However, this influence is a distinctive characteristic of an individual DES and cannot be extended to all DESs studied in this work.

### Table 1. Experimentally measured pH values.

| DES                | Abbreviation | Molar Ratio | \( w \cdot \text{H}_{2}\text{O} \) [%] | pH (20 °C) ± st.dev. |
|--------------------|--------------|-------------|---------------------------------|---------------------|
| Betaine:citric acid| B:CA         | 1:1         | 30                              | 2.46 ± 0.04         |
|                    |              |             | 50                              | 2.46 ± 0.02         |
| Betaine:ethylene glycol | B:EG     | 1:2         | 30                              | 6.86 ± 0.00         |
| Betaine:glucose    | B:Glc        | 1:1         | 30                              | 6.64 ± 0.35         |
|                    |              |             | 10                              | 6.44 ± 0.35         |
| Betaine:glycerol   | B:Gly        | 1:2         | 30                              | 6.77 ± 0.04         |
|                    |              |             | 50                              | 6.38 ± 0.07         |
| Betaine:malic acid | B:Ma         | 1:1         | 30                              | 2.98 ± 0.01         |
|                    |              |             | 50                              | 2.92 ± 0.01         |
| Betaine:sucrose    | B:Suc        | 4:1         | 30                              | 7.85 ± 0.11         |
| Choline chloride:citric acid | Ch:CA         | 2:1          | 30                              | 0.34 ± 0.04         |
|                    |              |             | 50                              | 0.71 ± 0.00         |
| Choline chloride:ethylene glycol | ChCl:EG  | 1:2         | 30                              | 6.60 ± 0.57         |
|                    |              |             | 50                              | 4.58 ± 0.14         |
|                    |              |             | 80                              | 4.41 ± 0.00         |
| Choline chloride:fructose | ChCl:Fru  | 1:1         | 30                              | 3.51 ± 0.05         |
|                    |              |             | 50                              | 3.35 ± 0.03         |
| Choline chloride:glucose | ChCl:Glc  | 1:1         | 30                              | 4.83 ± 0.06         |
|                    |              |             | 50                              | 3.56 ± 0.01         |
| Choline chloride:glycerol | ChCl:Gly  | 1:2         | 30                              | 3.71 ± 0.06         |
|                    |              |             | 50                              | 2.67 ± 0.11         |
|                    |              |             | 80                              | 3.06 ± 0.01         |
| Molecules 2022, 27, 4489 | 4 of 15 |
|--------------------------|----------------------------------|
| **Choline chloride:malic acid** | **ChCl:MA** 1:1 |
| 30 | 0.63 ± 0.01 |
| 50 | 1.03 ± 0.00 |
| **Choline chloride:proline:malic acid** | **ChCl:Pro:MA** 1:1:1 |
| 10 | 3.23 ± 0.00 |
| 30 | 2.82 ± 0.01 |
| 50 | 2.63 ± 0.03 |
| **Choline chloride:sorbitol** | **ChCl:Sol** 1:1 |
| 50 | 4.92 ± 0.04 |
| 80 | 3.80 ± 0.08 |
| **Choline chloride:urea** | **ChCl:U** 1:2 |
| 30 | 8.85 ± 0.06 |
| 50 | 8.23 ± 0.04 |
| **Choline chloride:urea:ethylene glycol** | **ChCl:U:EG** 1:2:2 |
| 10 | 8.29 ± 0.07 |
| **Choline chloride:urea:glycerol** | **ChCl:U:Gly** 1:2:2 |
| 10 | 8.72 ± 0.05 |
| 30 | 2.86 ± 0.04 |
| **Choline chloride:xylose** | **ChCl:Xyl** 1:1 |
| 50 | 3.32 ± 0.03 |
| 80 | 3.93 ± 0.01 |
| **Choline chloride:xyritol** | **ChCl:Xyol** 5:2 |
| 30 | 6.90 ± 0.06 |
| 50 | 6.50 ± 0.01 |
| **Choline chloride:fructose** | **ChCl:Fru** 1:1 |
| 30 | 3.51 ± 0.05 |
| 50 | 3.35 ± 0.03 |
| **Citric acid:glucose** | **CA:Glc** 1:1 |
| 30 | 0.53 ± 0.04 |
| **Citric acid:sucrose** | **CA:Suc** 1:1 |
| 30 | 0.83 ± 0.00 |
| **Fructose:ethylene glycol** | **Fru:EG** 1:2 |
| 30 | 5.31 ± 0.09 |
| **Fructose:glucose:ethylene glycol** | **Fru:Glc:EG** 1:1:2 |
| 50 | 3.67 ± 0.06 |
| **Fructose:glucose:sucrose** | **Fru:Glc:Suc** 1:1:1 |
| 50 | 2.63 ± 0.03 |
| 80 | 2.99 ± 0.01 |
| **Fructose:glucose:urea** | **Fru:Glc:U** 1:1 |
| 30 | 8.22 ± 0.06 |
| **Glucose:ethylene glycol** | **Glc:EG** 1:2 |
| 50 | 4.03 ± 0.02 |
| **Glucose:glycerol** | **Glc:Gly** 1:2 |
| 50 | 4.33 ± 0.04 |
| **Malic acid:fructose** | **MA:Fru** 1:1 |
| 30 | 0.77 ± 0.01 |
| **Malic acid:fructose:glycerol** | **MA:Fru:Gly** 1:1 |
| 30 | 2.77 ± 0.01 |
| **Malic acid:glucose** | **MA:Glc** 1:1 |
| 30 | 0.83 ± 0.01 |
| **Malic acid:glucose:sucrose** | **MA:Glc:Suc** 1:1:1 |
| 10 | 0.92 ± 0.00 |
| **Malic acid:sucrose** | **MA:Suc** 2:1 |
| 30 | 0.66 ± 0.01 |
| 10 | 2.63 ± 0.01 |
| **Proline:malic acid** | **Pro:MA** 1:1 |
| 30 | 2.78 ± 0.02 |
| 50 | 2.73 ± 0.03 |
| **Sucrose:ethylene glycol** | **Suc:EG** 1:2 |
| 30 | 6.05 ± 0.06 |
| **Sucrose:glucose:urea** | **Suc:Glc:U** 1:1 |
| 30 | 8.14 ± 0.25 |
| **Xylose:ethylene glycol** | **Xyl:EG** 1:2 |
| 30 | 4.57 ± 0.06 |
Furthermore, DESs were mathematically described using the $\sigma$ profile defined with the COSMOtherm software. The HBA and HBD molecules were optimized in TmoleX, both from an energy and geometry point of view. The generated COSMO files contain all information necessary for the calculation of the $\sigma$ profile function and thus for the calculation of the $\sigma$ profile descriptors. For the preparation of the descriptor set, the DESs were modeled as a molar mixture of HBA and HBD according to Table 1. The $\sigma$ profile curves for each HBA and HBD were divided into 10 regions, the area under each region was calculated, and their numerical values were correlated with the experimental pH values using mathematical models.

2.2. Multiple Linear Regression and Piecewise Linear Regression

The assessment of the MLR and PLR model applicability to predict the pH values of DESs was based on the correlation coefficient values, $R^2$, $R^2_{\text{adj}}$, and RMSE. The obtained model coefficient values and the basic statistical analysis are presented in Table 2 while a comparison between the experimental and model-estimated pH values is given in Figure 1.
Figure 1. Comparison between experimental data and (a) MLR model, (b) PLR model, and (c) ANN model. (○) data set for model development, (◆) data set for model validation.

Table 2. MLR and PLR regression coefficients. Statistically significant coefficients are marked in bold.

| Break point | MLR Regression Coeff. ± st. Error | p-Value | PLR Regression Coeff. ± st. Error | p-Value |
|-------------|-----------------------------------|---------|-----------------------------------|---------|
|             | Regression Coeff. ± st. Error | p-Value | Regression Coeff. ± st. Error | p-Value |
| b₀          | -13.4623 ± 4.9782                | 0.0078  | -1.9449 ± 0.1556 ± 80.4560 ± 10.6436 | 0.0001  |
| b₁ (S₁mix) | 16.4623 ± 5.1388                 | 0.0022  | 14.8847 ± 2.1908 ± 23.1982 ± 1.8558 | 0.0001  |
| b₂ (S₂mix) | 9.1349 ± 2.4418                  | 0.0003  | 10.2415 ± 2.3918 ± 27.8095 ± 2.2247 | 0.0001  |
| b₃ (S₃mix) | 9.7560 ± 2.5748                  | 0.0002  | 9.1933 ± 1.7354 ± 35.1992 ± 2.8159 | <0.0001 |
| b₄ (S₄mix) | 4.2440 ± 1.1602                  | 0.0004  | 4.8581 ± 1.1221 ± 11.2879 ± 1.1902 | <0.0001 |
| b₅ (S₅mix) | 2.2980 ± 0.6482                  | 0.0006  | 2.5621 ± 0.1188 ± 10.1747 ± 1.3976 | <0.0001 |
| b₆ (S₆mix) | -0.9176 ± 1.0696                 | 0.3927  | -2.4281 ± 0.8779 ± 14.7126 ± 1.1770 | 0.2666  |
| b₇ (S₇mix) | -4.5381 ± 1.1435                 | 0.0020  | -4.1497 ± 0.6632 ± 9.6777 ± 0.7742 | <0.0001 |
| b₈ (S₈mix) | -8.9573 ± 1.9634                 | <0.0001 | -9.2237 ± 1.6373 ± 25.6581 ± 2.0526 | <0.0001 |
| b₉ (S₉mix) | -10.0312 ± 2.8589                | 0.0006  | -11.4736 ± 3.6473 ± 32.0013 ± 2.5601 | 0.0001  |
| b₁₀ (S₁₀mix) | -12.9604 ± 3.6943               | 0.0006  | -13.9250 ± 4.4560 ± 42.7492 ± 3.4199 | 0.0001  |

| R²          | 0.7758                          | 0.9654 |
| R²adj       | 0.7564                          | 0.9624 |
| RMSE        | 1.1865                          | 0.6588 |
| F value     | 39.8120                         | 39.8120 |
| p-value     | <0.0001                         | <0.0001 |
As described in the literature, linear regression calculates an equation that minimizes the distance between the fitted line and all data points. In general, a model fits the data well if the discrepancies between the observed and predicted value are minimal and unbiased. According to Cheng et al. (2014) [16], the coefficient of determination and adjusted coefficient of determination can be considered as summary measures for the goodness of fit of any linear regression model. Moreover, Le Mann et al. (2010) stated that the model can be regarded as appropriate if the coefficient of determination is above 0.75 [17]. Based on this, it can be concluded that both the MLR \((R^2 = 0.7758)\) and PLR \((R^2 = 0.9654)\) models developed in this work are applicable for the description of DESs’ pH values based on \(S_{mix}\) descriptors but not with the same accuracy. When analyzing RMSE errors, it is evident that the PLR model (Figure 1b) ensures significantly smaller data dispersion \((RMSE = 0.6558)\) in comparison to the MLR model \((RMSE = 1.1865)\) (Figure 1a). As previously described, a high-accuracy model is strongly desired. However, the increase in the accuracy is usually accomplished by the increase in the complexity of the models by increasing the number of model parameters. For practical application, a model with fewer parameters is easier to interpret and, therefore, more suitable for the application.

A high \(R^2\) value alone does not guarantee that the model fits the data well, so the model’s goodness of fit was further confirmed by residual analysis. The residuals from a fitted model are the differences between the responses observed and the corresponding prediction of the response computed using the regression function. If the model’s fit to the data was correct, the residuals would approximate the random errors that make the relationship between the explanatory variables and the response variable a statistical relationship. Therefore, if the residuals appear to behave randomly, it would suggest that the model fits the data well [18]. Analyzing the results presented in Figure 2, the residuals for the MPLR and PLR models were found to be normally distributed (Figure 2a,b). Furthermore, because the residual plots were gathered roughly along a straight line, the normality condition was met. The bell-shaped histograms that display the measurement distribution also verified the normal distribution of the residuals (Figure 2a,b). The residual vs. predicted value plots (Figure 2a,b) reveal that the residuals have no pattern, implying that the models match the experimental data well. Additionally, the residuals were found to range around the central value (Figure 2a,b) without obvious outliers, which means that the level of randomization was appropriate and that the sequence of testing had no effect on the findings [19].
Figure 2. Analysis of the residuals for the MLR model (a–d), PLR model (e–h), and ANN mode (i–l).
Analysis of the MLR and PLR model coefficients showed that all coefficients, except $b_6$ (coefficient multiplying $S_{6\text{mix}}$), were statistically significant. It can also be noticed that for both models, the coefficients from $b_1$ to $b_5$ have a positive influence on the output variable while the coefficients from $b_6$ to $b_{10}$ have a negative influence on the analyzed model output. The results are easily interpreted in terms of $b_1$ to $b_5$, which are associated with the negative potential region and thus with hydrogen bond accepting and basicity properties on the one hand, and $b_7$ to $b_{10}$, which are associated with the positive potential region and thus with hydrogen bond donating and acidity properties on the other hand. $b_6$ turns out to be related to the neutral potential region insignificantly contributing to the pH value. As for the other $b$ coefficient values, the more distant the potential region is from the zero (neutral value), the stronger its influence (whether positive or negative) on the pH value. Thus, the model seems to have a clear and rather simple physical significance. Although statistical analysis showed that the coefficient $b_6$ was not significant, the variable $S_{6\text{mix}}$ was not excluded from the modeling. This result indicates that there is no correlation with the dependent variable at the population level, but this could be changed if a different data set was used.

The ANOVA revealed that the created MLR and PLR models were statistically significant, with $p$ values $< 0.001$. Moreover, higher $F$-test results ($F$ value $= 39.8120$) and lower $p$ values, according to Greenland et al. (2016) [20], show the relative relevance of the created models. Based on the presented results it can be concluded that the collected findings demonstrate the dependability of the created models throughout the spectrum of variables evaluated.

2.3. Artificial Neural Network Modelling

The applicability of the artificial neural network models for predicting the DES pH values based on the $\sigma$ profiles was also studied. The best neural network was chosen based on the following criteria: $R^2$ and $RMSE$ for training, test, and validation sets taking into account the number of neurons in the hidden layer. The properties of the created networks that were chosen are shown in Table 3. Based on the goodness of fit and validation error and considering the number of neurons in the hidden layer, the MLP model 10-5-1 was selected as optimal. Fewer neurons in the hidden layer make the ANN architecture simpler. The selected ANN was characterized by 10 neurons in the input layer, 5 neurons in the hidden layer, and 1 neuron in the output layer. The hidden activation function for the selected ANN was Tanh while the output activation function was Logistic. The described ANN provides a good agreement between the experimental data and the data predicted by the model ($R^2_{\text{validation}} = 0.9797$, $RMSE_{\text{validation}} = 0.0012$). As presented in Figure 1c, it can be observed that the data are distributed around the fitted function and that there are no evident outliers. As for the MLP and PLR models, the residual analysis was also performed for the ANN model (Figure 2c) and confirmed the ANN model’s goodness of fit through a normal probability plot of the residuals (Figure 2c), residuals versus the predicted values plot (Figure 2c), histogram of the residuals (Figure 2c), and residuals versus the order of the data plot (Figure 2c).
Table 3. Architecture of the developed ANN (selected network is marked in bold). The numbers in the network name denote the number of neurons in the input, hidden, and output layers, respectively.

| Network Name | Training Perf./Training Error | Test Perf./Test Error | Validation Perf./Validation Error | Hidden Activation | Output Activation |
|--------------|-------------------------------|-----------------------|----------------------------------|-------------------|------------------|
| MLP 10-13-1  | 0.9734, 0.0021                | 0.9751, 0.0031        | 0.9578, 0.0042                   | Logistic          | Logistic         |
| MLP 10-11-1  | 0.9812, 0.0013                | 0.9802, 0.0018        | 0.9794, 0.0018                   | Tanh              | Exponential      |
| MLP 10-10-1  | 0.9803, 0.0013                | 0.9827, 0.0016        | 0.9788, 0.0019                   | Tanh              | Tanh             |
| MLP 10-10-1  | 0.9808, 0.0017                | 0.9806, 0.0021        | 0.9716, 0.0019                   | Tanh              | Logistic         |
| MLP 10-5-1   | 0.9868, 0.0011                | 0.9799, 0.0012        | 0.9797, 0.0012                   | Tanh              | Logistic         |

Based on the presented results, it can be concluded that the σ profiles are good molecular descriptors of DESs since the mathematical correlation of the experimental and predicted values is high. Moreover, based on the obtained $R^2$ values and the residual analysis, it can be concluded that both the PLR and ANN model can be efficiently applied for the prediction of the DES pH values based on the σ profiles. Due to the simplicity of the PLR model, this model is proposed for the prediction of physicochemical properties.

2.4. MLR, PLR, and ANN Models’ Independent Validation

Validation of the MLR, PLR, and ANN models developed for the prediction of the DES pH values based on the σ profiles was performed on the independent set of data. The validation set included the σ profiles of 16 DESs. Comparisons between the experimental data and model-predicted data are shown in Figure 2. The validation performance of the developed models was estimated based on $R^2$ and RMSE and the obtained values were as follows: (i) for MLR $R^2 = 0.7097$, RMSE = 1.1140; (ii) for PLR $R^2 = 0.8605$, RMSE = 0.7652; and (iii) for ANN $R^2 = 0.8885$, RMSE = 0.82926.

It can be noticed that all three proposed models predict the pH value with high accuracy. As expected, the highest $R^2$ between the experiment and model-predicted data was obtained for ANN prediction of the analyzed DES pH values while the lowest $R^2$ between the experiment and model-predicted data was obtained for the MLR model. These findings demonstrate that σ profile ANN modeling is a useful and reliable method for predicting DES pH values based on the σ profiles. Nevertheless, considering RMSE, it can be noticed that the PLR model can efficiently be used for the prediction of pH values based on the σ profiles. As described, the $R^2$ values are scaled between 0 and 1, whereas the RMSE is not scaled to a specific value and, therefore, provides explicit information about how much the prediction deviates.

As stated before, it was relatively easy to link the parameters of the MLR and PLR models to their physical significance. On the other hand, ANNs, by definition, belong to a class of agnostic models and, thus, it is difficult, if not impossible, to reveal their physical meaning. At the same time, this is the reason why they behave much better in interpolation than in extrapolation. The independent validation presented here may be considered as interpolation since the DES members of the independent validation dataset belong to the same DES classes as those used for constructing the model. However, given the rather simple and rather clear relation between the σ profile and pH as revealed by MLR, there is no true reason to believe that the models would behave poorly in extrapolation, even for ANN, i.e., for DES classes not involved in the development of the models. However, this is yet to be checked, e.g., for DESs based on metal chlorides or DESs containing ionic liquids, etc.
The current literature data refer to the prediction of other physicochemical properties (such as viscosity and density) and only a narrow range of values characteristic for limited groups of structurally related DESs [11–14]. Based on our current knowledge, only one study has investigated the development of a mathematical model for DES pH value prediction [13]. In that study, the pH literature data of 41 DESs were processed in a similar way using the COSMO-RS and mathematical models, MLR and ANN, also covering a variety of cations, anions, and functional groups. The literature study [12] used literature data and included different temperatures (with temperature as an input parameter) while our study used our data obtained at a single temperature. The literature study also showed the potential of MLR and ANN modeling for the prediction of the pH value, however, with more complex models (models with more coefficients) than those developed in this work. Taking into consideration the specific future application of the developed models, it is recommended that they are as simple as possible and as robust as possible. Summing up the presented results, it can be concluded that the PLR model developed in this research can efficiently be used for the prediction of a wide range of DES pH values based on the σ profiles.

3. Materials and Methods

3.1. Materials

Betaine, choline chloride, glucose, L-(-)-proline, oxalic acid, sucrose, sorbitol, and xylitol were all purchased from Acros Organics, USA. Citric acid, D-fructose, D-(+)-xylose, D,L-malic acid, ethylene glycol, glycerol, and urea were all purchased from Sigma-Aldrich, USA. BIOVIA TmoleX19 version 2021 software (Dassault Systèmes, Vélizy-Villacoublay, France) was used for geometry and energy optimization of the HBAs and HBDs used in this study. BIOVIA COSMOTHERM 2020 version 20.0.0 software (Dassault Systèmes) was used for the σ profile calculations of the defined DESs.

3.2. Methods

3.2.1. DES Preparation

DESs were prepared by mixing defined molar ratios of HBA to HBD. The two or more components were weighed in a specific ratio in a round-bottomed glass flask, adding 10–50% (w/w) of water. Then, the flasks were sealed, and the mixtures stirred and heated to 50 °C for 2 h until homogeneous transparent colorless liquids formed. The DES abbreviations and corresponding molar ratios are given in Table 1.

3.2.2. pH Value Measurement

The pH values for each DES were determined with a pH/ion meter S220 using an InLab Viscous Pro-ISM pH-electrode (Mettler Toledo, Greifensee, Switzerland), all within the pH measuring range 0.36–9.31 at room temperature. The instrument was calibrated using standard pH buffer solutions. Additionally, the pH values were checked with litmus paper (range 1–14). All measurements were carried out in duplicates and the results were expressed as an average value ± standard deviation.

3.2.3. Calculation of DES Constituents’ σ Profiles and Descriptors

All molecules used for DES preparation: HBA, HBD, and water, were geometrically and energetically optimized in the BIOVIA TmoleX19 version 2021 (Dassault Systèmes) software. Quantum chemical calculations were performed by adopting DFT (density functional theory) with the BP86 functional level of theory and def-TZVP basis set [10]. To create a simplified and user-friendly database, for each molecule, the single most abundant non-ionized conformer with the lowest energy was chosen and used for further calculations. Molecules consisting of two or more ions (e.g., choline chloride) were treated as ion pairs and their structures were optimized according to Abranches et al. (2019) [21]. Finally, the software-generated COSMO file for each optimized molecule contained its σ
profile curve that provided a quantitative representation of the molecules’ polar surface screen charge on the polarity scale. HBAs are characterized by peaks in the negative potential region, HBDs by peaks in the positive potential region, and nonpolar molecules by peaks in the potential region around zero.

To define the molecular descriptors for all DES constituents, the $\sigma$ profile curve for each HBA, HBD, and water was divided into 10 regions. The width of each region was $0.005 \text{ e/Å}^2$, covering the range from $-0.025$ to $+0.025 \text{ e/Å}^2$. The areas under the curve were integrated separately for each defined region. This was achieved by simple summation of the tabulated $\sigma$ profile data point ordinate values as presented by the BIOVIA COSMOtherm 2020 software. The ordinate values lying on the boundaries of the regions were split into halves and each half was attributed to one of the neighboring regions. Thus, 10 $S$ descriptors ($S^1$–$S^{10}$) of the $\sigma$ profiles were calculated exactly as the numerical values of these 10 areas (Table A1).

3.2.4. Calculation of DES Descriptors

Any change in the DES composition can be described by a change in its $\sigma$ profile and the associated numerical value of its descriptors. To obtain a unique descriptor set for each particular DES, the $\sigma$ profiles of its constituents were processed in the following manner. The descriptors of the studied DESs ($S_{\text{mix}}$) were calculated from the HBA and HBD component (and in some cases water) descriptors according to Equation (1) proposed by Benguerba et al. (2019) [11]:

$$S_{\text{mix}}^i = \sum_{j=1}^{NC} X_j S_{\sigma-\text{profile},j}^i$$

where $i$ denotes the descriptor number (1–10), $j$ stands for the DES constituent number, $X_j$ is the molar fraction of HBA or HBD or some other constituent such as water if present in the mixture, $S_{\sigma-\text{profile},j}^i$ is the $j$-th constituent $i$-th descriptor, and $NC$ is the total number of constituents from which DES is prepared. All the experiments were performed at 20 °C.

3.2.5. Modeling of Correlation between pH and Descriptors

In further calculations, it was assumed that the measured DES pH value can be described as a function of the $\sigma$ profile of the mixture, expressed by a set of $S_{\text{mix}}$ descriptors in Equation (2):

$$\text{pH} = f\left(S_{\text{mix}}^1, S_{\text{mix}}^2, S_{\text{mix}}^3, S_{\text{mix}}^4, S_{\text{mix}}^5, S_{\text{mix}}^6, S_{\text{mix}}^7, S_{\text{mix}}^8, S_{\text{mix}}^9, S_{\text{mix}}^{10}\right)$$

Multiple linear regression (MLR) with Equation (3), piecewise linear regression (PLR) with Equation (4), and artificial neural network (ANN) models were attempted to describe the relationship between the input and output variables. The dataset included 142 data points (that included replicates), of which 126 were used for model development and 16 (randomly selected) for independent model validation:

$$\text{pH} = b_0 + b_1 \cdot S_{\text{mix}}^1 + b_2 \cdot S_{\text{mix}}^2 + b_3 \cdot S_{\text{mix}}^3 + b_4 \cdot S_{\text{mix}}^4 + b_5 \cdot S_{\text{mix}}^5 + b_6 \cdot S_{\text{mix}}^6 + b_7 \cdot S_{\text{mix}}^7 + b_8 \cdot S_{\text{mix}}^8 + b_9 \cdot S_{\text{mix}}^9 + b_{10} \cdot S_{\text{mix}}^{10}$$

$$\text{pH} = \begin{cases} b_{01} + \sum_{i=1}^{10} b_{1i} \cdot S_{\text{mix}}^i & \text{if } (\text{pH} \leq b_n) \\ b_{02} + \sum_{i=1}^{10} b_{2i} \cdot S_{\text{mix}}^i & \text{if } (\text{pH} > b_n) \end{cases}$$
The PLR technique is based on estimating the parameters of two linear regression equations: one for dependent variable values \( y \) less than or equal to the breakpoint \( (bn) \) and the other for dependent variable values \( y \) higher than the breakpoint. The MLR parameters in Equation (3) were estimated using least square regression while the PLR parameters in Equation (4) were estimated using the Levenberg–Marquardt algorithm implemented in the software Statistica 13.0 (Tibco Software Inc, Palo Alto, Santa Clara, CA, USA). The algorithm searches for optimal solutions in the function parameter space using the least squares method. The calculations were performed in 50 repetitions with a convergence parameter of \( 10^{-6} \) and a confidence interval of 95% [22].

In addition, multilayer perceptron (MLP) ANNs were used for the prediction of DES pH values based on the \( S_{mix} \) descriptors. The ANN models included an input layer, hidden layer, and output layer. The input layer included 10 neurons representing the \( S_{mix} \) descriptors, the output layer had only one neuron, and the number of neurons in the hidden layer varied between 4 and 13 and was randomly selected by the algorithm. The hidden activation function and output activation function were selected randomly from the following set: Identity, Logistic, Hyperbolic tangent, and Exponential. The dimension of the data set for ANN modeling was 126 × 11 and was randomly divided into 70% for network training, 15% for network testing, and 15% for model validation. Model training was carried out using a back error propagation algorithm and the error function was a sum of squares implemented in Statistica v.13.0 Automated Neural Networks. The developed model’s performance was estimated by calculating the \( R^2 \) and root mean squared error (\( RMSE \)) values for the training, test, and validation sets.

Validation of the developed MLR, PLR, and ANN models was performed on an independent data set, including the \( S_{mix} \) descriptors for 16 randomly selected DESs. The validation performance of the developed models was estimated based on the \( R^2 \) and root mean squared error (\( RMSE \)).

4. Conclusions

The applicability of MLR, PLR, and ANN to predict the pH values of DESs was evaluated. The results indicate that although simple linear regression can be used for the description and prediction, its effectiveness and applicability are limited. On the other hand, PLR and ANN are applicable to predict the pH values of DESs with a very high goodness of fit \((R^2 > 0.8600)\). The contribution of this work lies in the development of a user-friendly model to predict pH values in a wide range (from 0.525 to 9.25), indicating that the developed models are good for the prediction of the pH value of newly synthesized DESs. However, due to the simplicity of the developed PLR model, it could be suggested as a model of choice for use in daily work and screening purposes.

Nevertheless, this approach can also be extended to other physicochemical properties since this study confirmed previous findings that showed how the \( \sigma \) profile generated in COSMOtherm is a valuable DES molecular descriptor. It could be a good basis for the evaluation of various mathematical models to develop a simple and applicable prediction model for everyday laboratory or industrial applications.

It is interesting to comment on the influence of the addition of water to a DES. In our previous article [7], based on a limited set of data, it was noticed that the addition of water to extremely acidic DESs increases their pH values, and the addition of water to highly basic DESs decreases their pH values. Thus, it seemed that the addition of water somehow mellowed the pH environments. On the other hand, on a larger set of data, as presented here, this conclusion does not hold any more: there are difficult-to-predict exemptions to the rule. On the other hand, the COSMO-RS calculation results in combination with the non-presumptive numerical models, such as MLR, PLR, and ANN, are perfectly suitable to tackle those difficult-to-predict systems.
Author Contributions: Conceptualization, I.R.R., M.P., and A.J.T.; methodology, M.P. and A.J.T.; software, M.P., M.R. (Mia Radović), M.R. (Marko Rogošić), and J.A.P.C.; validation, M.R. (Mia Radović), M.C.B., K.R., and J.A.P.C.; formal analysis, M.P., M.R. (Mia Radović), and M.C.B.; investigation, M.P., M.C.B., K.R., and M.R. (Mia Radović); resources, I.R.R.; data curation, M.P., M.R. (Mia Radović), M.R. (Marko Rogošić), and A.J.T.; writing—original draft preparation, M.P. and A.J.T.; writing—review and editing, M.P., M.R. (Mia Radović), M.C.B., K.R., M.R. (Marko Rogošić), I.R.R., A.J.T., and J.A.P.C.; visualization, M.P., M.R. (Marko Rogošić), and A.T.; supervision, I.R.R.; project administration, I.R.R.; funding acquisition, I.R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly developed within the scope of the project CICECO-Aveiro Institute of Materials, UIDB/50011/2020 & UIDP/50011/2020, financed by national funds through the Portuguese Foundation for Science and Technology/MCTES. This work was also financed by the Croatian science foundation (grant No.7712).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Samples of the compounds are available from the authors.

Appendix A

Table A1. S descriptors (S1–S10) of the σ profiles from compounds from which DESs were prepared.

| Intervals | B Betaine | ChCl Choline Chloride | Pro LDproline | CA Citric Acid | MA Malic Acid | OxA Oxalic Acid | U Urea | H2O |
|-----------|-----------|-----------------------|---------------|----------------|---------------|-----------------|-------|-----|
| [-0.025; 0.02] | 1 | 0 | 0 | 0.506 | 4.861 | 3.5955 | 0 | 0 | 0 |
| [-0.025; 0.015] | 2 | 0 | 0 | 0 | 5.186 | 14.9695 | 10.5215 | 7.5105 | 6.35 | 6.35 |
| [-0.015; 0.01] | 3 | 11.869 | 16.1615 | 6.9485 | 13.5665 | 9.368 | 20.482 | 10.027 | 10.027 |
| [-0.01; 0.005] | 4 | 59.1185 | 66.196 | 17.199 | 29.212 | 28.535 | 9.0145 | 3.5195 | 3.5195 |
| [-0.005; 0.0] | 5 | 36.625 | 34.4875 | 60.605 | 29.3465 | 23.3925 | 7.9265 | 2.1635 | 2.1635 |
| [0.0; 0.005] | 6 | 4.5285 | 5.6435 | 21.7815 | 23.467 | 18.1455 | 13.051 | 2.8725 | 2.8725 |
| [0.005; 0.01] | 7 | 3.2405 | 6.6525 | 10.6 | 37.877 | 25.726 | 7.606 | 4.055 | 4.055 |
| [0.01; 0.015] | 8 | 7.719 | 18.3 | 17.614 | 38.933 | 30.6435 | 11.679 | 5.2285 | 5.2285 |
| [0.015; 0.02] | 9 | 22.3525 | 30.0465 | 5.2065 | 1.0135 | 2.3845 | 13.8265 | 8.2765 | 8.2765 |
| [0.02; 0.025] | 10 | 8.202 | 0.0525 | 1.3475 | 0 | 0 | 0 | 0.172 | 0.5775 |

| Intervals | EG ethylene glycol | Sol sorbitol | Gly glycol | Xyl xylitol | Fru Dfructose | Glc Dglucose | Suc sucrose | Xyl Dxylose |
|-----------|-------------------|-------------|----------|-----------|-------------|---------|-----------|-----------|
| [-0.025; 0.02] | 1 | 0.1725 | 0 | 0.013 | 0.037 | 0.1655 | 0.213 | 0.108 | 0.037 |
| [-0.02; 0.015] | 2 | 3.8055 | 15.884 | 7.828 | 9.216 | 11.8325 | 23.022 | 11.1905 | 8.7015 |
| [-0.015; 0.01] | 3 | 7.638 | 20.8955 | 11.4065 | 15.941 | 16.9895 | 28.444 | 14.5935 | 12.9035 |
| [-0.01; 0.005] | 4 | 20.2675 | 41.5705 | 19.6085 | 43.0415 | 36.2095 | 61.232 | 35.406 | 34.6755 |
| [-0.005; 0.0] | 5 | 28.038 | 30.5525 | 34.6465 | 35.965 | 39.319 | 54.567 | 28.5165 | 45.0735 |
| [0.0; 0.005] | 6 | 9.973 | 18.7645 | 15.0725 | 19.083 | 19.565 | 29.1605 | 15.066 | 17.7475 |
| [0.005; 0.01] | 7 | 7.9725 | 21.5775 | 10.103 | 20.848 | 19.4555 | 26.6145 | 20.4555 | 17.2715 |
| [0.01; 0.015] | 8 | 10.5605 | 28.283 | 17.194 | 28.977 | 30.9465 | 47.3685 | 29.0795 | 23.517 |
| [0.015; 0.02] | 9 | 9.6155 | 20.752 | 10.7865 | 10.9745 | 11.901 | 26.0425 | 8.4485 | 12.688 |
| [0.02; 0.025] | 1 | 0.0035 | 0.15 | 0.0115 | 0 | 0 | 1.082 | 0 | 0.005 |
References

1. Anastas, P.T.; Beach, E.S. Green Chemistry: The Emergence of a Transformative Framework. Green Chem. Lett. Rev. 2008, 1, 9–24. https://doi.org/10.1080/175182520701882441.

2. Cvjetko Bubalo, M.; Vidović, S.; Radojičić Redovniković, I.; Jokić, S. Green Solvents for Green Technologies. J. Chem. Technol. Biotechnol. 2015, 90, 1631–1639. https://doi.org/10.1002/jctb.4668.

3. Lanza, V.; Vecchio, G. New Conjugates of Superoxide Dismutase/Catalase Mimetics with Cyclodextrins. J. Inorg. Biochem. 2009, 103, 381–388. https://doi.org/10.1016/j.inorgbio.2008.11.017.

4. Abbott, A.P.; Capper, G.; Davies, D.L.; Rasheed, R.K.; Tambryjahaj, V. Novel Solvent Properties of Choline Chloride/Urea Mixtures. Chem. Commun. 2003, 10, 70–71. https://doi.org/10.1039/b210714g.

5. Martins, M.A.R.; Pinho, S.P.; Coutinho, J.A.P. Insights into the Nature of Eutectic and Deep Eutectic Mixtures. J. Solut. Chem. 2019, 48, 962–982. https://doi.org/10.1007/s10953-018-0793-1.

6. Paiva, A.; Matias, A.A.; Duarte, A.R.C. How Do We Drive Deep Eutectic Systems towards an Industrial Reality? Curr. Opin. Green Chem. 2018, 11, 81–85. https://doi.org/10.1016/j.coq.2018.05.010.

7. Mitar, A.; Panić, M.; Prlić Kardum, J.; Halambek, J.; Zagajski Kučan, K.; Radojičić Redovniković, I.; Radšević, K. Physicochemical Properties, Cytotoxicity, and Antioxidative Activity of Natural Deep Eutectic Solvents Containing Organic Acid. Chem. Biochem. Eng. Q. 2019, 33, 1–18. https://doi.org/10.15255/CABEQ.2018.1454.

8. Abbott, A.P.; Alabudullah, S.S.M.; Al-Mursheidi, A.Y.M.; Ryder, K.S. Brønsted Acidity in Deep Eutectic Solvents and Ionic Liquids. Faraday Discuss. 2017, 206, 365–377. https://doi.org/10.1039/C7FD00153C.

9. Farias, F.O.; Passos, H.; Coutinho, J.A.P.; Mafra, M.R. PH Effect on the Formation of Deep-Eutectic-Solvent-Based Aqueous Two-Phase Systems. Ind. Eng. Chem. Res. 2018, 57, 16917–16924. https://doi.org/10.1021/acs.iecr.8b04256.

10. Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J.C.W. Refinement and Parametrization of COSMO-RS. J. Phys. Chem. A 1998, 102, 5074–5085. https://doi.org/10.1021/jp980017s.

11. Benguerba, Y.; Alnashef, I.M.; Erto, A.; Balsamo, M.; Ernst, B. A Quantitative Prediction of the Viscosity of Amine Based DESs Using So-Profile Molecular Descriptors. J. Mol. Struct. 2019, 1184, 357–363. https://doi.org/10.1016/j.molstruc.2019.02.052.

12. Lemaoui, T.; Hammoudi, N.E.H.; Alnashef, I.M.; Balsamo, M.; Erto, A.; Ernst, B.; Benguerba, Y. Quantitative Structure-Property Relationships for Deep Eutectic Solvents Using So-Profile as Molecular Descriptors. J. Mol. Liq. 2020, 309, 113165. https://doi.org/10.1016/j.molliq.2020.113165.

13. Lemaoui, T.; Abu Hatab, F.; Darwish, A.S.; Attoui, A.; Hammoudi, N.E.H.; Almustafa, G.; Benaicha, M.; Benguerba, Y.; Alnashef, I.M. Molecular-Based Guide to Predict the PH of Eutectic Solvents: Promoting an Efficient Design Approach for New Green Solvents. ACS Sustain. Chem. Eng. 2019, 7, 5783–5808. https://doi.org/10.1021/acssuschemeng.9c07367.

14. Silva, I.P.; Fernandez, L.; Conceicao, J.H.F.; Martins, M.A.R.; Sosa, A.; Ortega, J.; Pinho, S.P.; Coutinho, J.A.P. Design and Characterization of Sugar-Based Deep Eutectic Solvents Using Conductor-like Screening Model for Real Solvents. ACS Sustain. Chem. Eng. 2018, 6, 10724–10734. https://doi.org/10.1021/acssuschemeng.8b02042.

15. Hayyan, A.; Mjallí, F.S.; Alnashef, I.M.; Al-Wahaibi, T.; Al-Wahaibi, Y.M.; Hashim, M.A. Fruit Sugar-Based Deep Eutectic Solvents and Their Physical Properties. Thermochim. Acta 2012, 541, 70–75. https://doi.org/10.1016/j.tca.2012.04.030.

16. Cheng, C.L.; Salalbh, G. Coefficient of Determination for Multiple Measurement Error Models. J. Multivar. Anal. 2014, 126, 137–152. https://doi.org/10.1016/j.jmva.2014.01.006.

17. Le Man, H.; Behera, S.K.; Park, H.S. Optimization of Operational Parameters for Ethanol Production from Korean Food Waste Leachate. Int. J. Environ. Sci. Technol. 2009, 7, 157–164. https://doi.org/10.1007/s12641-012-0306-0.

18. Cheng, C.; Feng, C.; Li, L.; Sadeghpour, A. A Comparison of Residual Diagnosis Tools for Diagnosing Regression Models for Count Data. BMC Med. Res. Methodol. 2020, 20, 175. https://doi.org/10.1186/s12874-020-01055-2/TABLES/7.

19. Matešić, N.; Jurina, T.; Benković, M.; Panić, M.; Valinger, D.; Gajdoš Kljusurić, J.; Jurinjak Tušek, A. Microwave-Assisted Extraction of Phenolic Compounds from Cannabis Sativa L.: Optimization and Kinetics Study. Sep. Sci. Technol. 2020, 56, 2047–2060. https://doi.org/10.1080/01496395.2020.1804938.

20. Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations. Eur. J. Epidemiol. 2016, 31, 337–350. https://doi.org/10.1007/s10654-016-0143-9.

21. Abranches, D.O.; Larriba, M.; Silva, L.P.; Melle-Franco, M.; Palomar, J.F.; Pinho, S.P.; Coutinho, J.A.P. Using COSMO-RS to Design Choline Chloride Pharmaceutical Eutectic Solvents. Fluid Phase Equilibria 2019, 497, 71–78. https://doi.org/10.1016/j.fluid.2019.06.005.

22. Jurinjak Tušek, A.; Jurina, T.; Benković, M.; Valinger, D.; Belščak-Cvitanović, A.; Kljusurić, J.G. Application of Multivariate Regression and Artificial Neural Network Modelling for Prediction of Physical and Chemical Properties of Medicinal Plants Aqueous Extracts. J. Appl. Res. Med. Aromat. Plants 2020, 16, 100229. https://doi.org/10.1016/J.JARMAP.2019.100229.