Journal of Molecular Liquids 418 (2025) 126707

Contents lists available at ScienceDirect

Journal of Molecular Liquids

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

# Machine learning-driven prediction of deep eutectic solvents' heat capacity for sustainable process design

Amit Kumar Halder [a,b], Reza Haghbakhsh [c], Elisabete S.C. Ferreira [a], Ana Rita C. Duarte [c], M. Natália D.S. Cordeiro [a,*]

[a] LAQV, REQUIMTE/Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
[b] Dr. B. C. Roy College of Pharmacy and Allied Health Sciences, Dr. Meghnad Saha Sarani, Bidhannagar, Durgapur 713212, West Bengal, India
[c] LAQV, REQUIMTE/Department of Chemistry, Faculty of Sciences and Technology, Nova University of Lisbon, 2829-516 Caparica, Portugal

ARTICLE INFO

*Keywords:*
Deep eutectic solvents
COSMO-RS
Machine learning
Heat capacity prediction
Thermodynamic modelling

ABSTRACT

Heat capacity, a crucial physical property for chemical processes, is often understudied in Deep Eutectic Solvents (DESs), which in turn are promising green alternatives to environmentally hazardous conventional solvents. This work addresses this gap by developing a machine learning model to predict DES heat capacity and identify key structural features influencing it. We employed a dataset of 530 DESs with corresponding experimental heat capacity values. Quantum-chemical COSMO-RS-based descriptors, capturing detailed information about DES structures, were calculated for each data point. Various machine learning algorithms, namely *k*-Nearest Neighbours (*k*NN), Random Forests (RF), Neural Network Multilayer Perceptron (MLP), and Support Vector Machines (SVM) were explored alongside a linear model (Multiple Linear Regression, MLR). Hyperparameter optimisation ensured all models were fine-tuned for optimal performance. The most successful model, based on the MLP technique, achieved remarkably low Average Absolute Relative Deviation (AARD) values of 0.500 % and 3.999 % for the training and test sets, respectively. This signifies a significant improvement in prediction accuracy compared to traditional methods. Furthermore, by applying a SHapley Additive exPlanations (SHAP) analysis, we identified the most crucial structural factors within DES components that govern their heat capacity. This comprehensive investigation offers valuable insights that can pave the way for an efficient design of novel DESs in the future.

## 1. Introduction

Heat capacity is a critical property in various chemical and industrial processes, and its optimal value varies depending on the specific application. In processes such as extraction and separation, it is essential that the heat capacity of the solvent be as low as possible to reduce the system's energy consumption. However, that is not the case in energy-related applications. Solar energy, for example, stands out as one of the most naturally available renewable sources, with electricity being produced through solar photovoltaic panels or heat via solar thermal systems. In the conventional solar thermal approach, toxic gases are emitted, contributing to global warming. In contrast, an alternative method harnesses solar radiation to generate vapour, using heat transfer or thermal fluids as a medium to store energy [1,2]. While typical thermal fluids have limited heat storage capacity and low thermal stability, molten salts suffer from high freezing points, low heat capacities, and corrosive characteristics. Addressing these challenges has been the focus of extensive research aimed at replacing environmentally harmful compounds with sustainable solvents.

Ionic liquids (ILs) initially emerged as a greener alternative due to their non-volatile, chemical and thermal stabilities, and tuneable features. However, the toxicity of some ILs, high costs, and viscosities have hindered their widespread industrial use [2–5]. In comparison, Deep Eutectic Solvents (DESs) have the advantages of being cheap, easy to prepare, and biodegradable, making them the next generation of sustainable solvents. More importantly, DESs possess less environmental toxicity than typical organic solvents and ILs. Therefore, designing industrially viable DES is crucial forward reducing the use of hazardous compounds in the near future [3].

Binary DESs are simply synthesized by combining at least one hydrogen bond donor (HBD) and one hydrogen bond acceptor (HBA) at a given molar ratio and temperature. The hydrogen bond network

* Corresponding author.
*E-mail address:* ncordeir@fc.up.pt (M.N.D.S. Cordeiro).

https://doi.org/10.1016/j.molliq.2024.126707
Received 25 July 2024; Received in revised form 6 November 2024; Accepted 7 December 2024
Available online 9 December 2024

established between the DES components gives rise to a mixture with a melting point significantly lower than either of its initial components [6,7]. Depending on the molar ratios of the components and their chemical nature, different DES properties can be attained, expanding the scope of their use [7].

The unique tunability of DESs leads to significant design opportunities, facilitating the materialisation of a wide array of this type of solvents with physical properties tailored to specific purposes [8–10]. So far, various DES usages have been reported, including gas absorption, solubilisation of chemicals and pharmaceuticals, biocatalysis, purification, electrochemical surface treatment, extraction, and chemical fuel processing [11–13]. Yet, given the multitude of possible combinations, experimental characterisation of each unique DES for a particular application is impractical. Consequently, the search for computational approaches capable of accurately predicting DES properties remains an active area of research.

Common computational approaches include, for example, empirical correlations based either solely on experimental data or incorporating also estimated data, with the latter typically calculated by means of group contribution methods like the Lydersen and Joback-Reid methods [14–19]. However, these empirical correlations may have limited applicability to specific DES compositions due to the complexity of DES structures, the scarcity of experimental data, and uncertainties in the estimated data. Molecular simulation methods instead, such as molecular dynamics (MD) simulations [7,20–22] can provide a detailed understanding of DES thermodynamic properties, particularly at a microscopic level. Nonetheless, these methods demand accurate force fields for the individual DES components and are computationally intensive.

More recently, Quantitative Structure-Property Relationships (QSPR) models have been employed to forecast various DES properties, such as density, viscosity, pH, heat capacity, surface tension, eutectic temperature, and $CO_2$ absorption [23–35], utilising molecular descriptors for the DES components. Particularly, quantum chemical molecular descriptors, based on the Conductor-like Screening Model for Realistic Solvation (COSMO-RS) approach, have proven effective in developing reliable models and understanding key structural factors influencing physicochemical values. For instance, it has been shown that DES non-polar regions enhance $CO_2$ absorption, while strong donor areas diminish their scavenger capacity [32]. Conversely, the same non-polar regions are accountable for a drop in values when predicting the DES surface tensions [29]. Similarly, COSMO-RS calculations have successfully modelled the heat capacity of ionic liquids [36]. This wealth of data facilitates the design of DESs with desired physicochemical characteristics.

Recognising heat capacity's crucial role as a thermodynamic property, this work seeks to develop a reliable QSPR model for predicting the heat capacities of deep eutectic solvents. Existing approaches that have been employed so far to predict DES heat capacities include empirical correlations [16], group contribution methods [17–20], and machine learning models with readily available data like molecular weight and critical pressure [26,27,35]. While these approaches have significantly contributed to building prediction models, they can be limited by pre-existing thermodynamic data availability, transferability, or reduced datasets, potentially leading to inaccuracies when predicting the heat capacities of diverse and novel DESs.

This work addresses these limitations by setting up QSPR models using the most updated experimental data available for binary DESs and employing molecular descriptors derived from the sigma profiles of their components, computed using COSMO-RS. Various machine learning and model-development techniques were systematically applied, and the model-predicted heat capacity values were compared with experimental data to identify the most predictive and statistically reliable approach. The procedure for this modelling work involves several essential steps: gathering the heat capacity dataset, deriving molecular descriptors, implementing linear and non-linear modelling, evaluating models, and interpreting the best model. Fig. 1 provides an overview of the methodology employed in this study. The outcomes of this study showcase its ability to offer valuable insights into the key characteristics impacting DES heat capacity, thereby making a substantial contribution to the realm of DES property predictions.

## 2. Materials and methods

**Dataset and Molecular Descriptors.** The dataset comprises 530 data points, covering thirty distinct binary DESs evaluated across various temperature ranges at constant atmospheric pressure. To compile this database, the majority of the data was gathered from the study carried out by Taherzadeh *et al.* [16], but it was subsequently updated by incorporating additional data points obtained from pertinent literature. Unlike properties such as density, viscosity, and surface tension, experimental heat capacity values for DESs are relatively scarce in the literature. Therefore, the dataset includes values for only eight different HBAs and thirteen HBDs. Despite this limitation, the presented dataset provides a broad spectrum of heat capacity values, spanning from 81 to 669 J $mol^{-1}$ $K^{-1}$. This diversity establishes a robust foundation for exploring the composition requirements of DESs to achieve desired heat capacities. Table 1 summarises the dataset employed for model development.

In our previous studies, we employed simple 0D–3D molecular descriptors to elucidate the main contributing factors that account for the density, surface tension, viscosity, and $CO_2$ absorption capability of binary DESs [24,28,31]. However, in this study, quantum chemical COSMO-RS-based molecular descriptors were utilized. COSMO-RS, initially proposed by Klamt [41] and later refined [42], is a theoretical modelling approach used to predict the thermodynamic behaviour of liquid systems, based on the screening charge density of their species ($\sigma$). More detailed information about the theory behind COSMO-RS and its applications can be found elsewhere [43]. In particular, COSMO-RS has gained prominence in recent years for investigating the physicochemical properties of DESs, providing critical structural insights into eutectic mixtures [23,25,29,30,32,44].

In this work, the three-dimensional (3D) structures of species were first drawn using the MOLDEN package [45]. Subsequently, geometry optimisations were carried out with the Gaussian09 software package [46], using the Becke-Perdew BVP86 density functional and the triple-zeta valence polarised (TZVP) basis set with DGA1 dispersion. The resulting COSMO files were then loaded into the COSMOThermX software (version 2023) to compute the sigma profiles ($\sigma$-profiles: 61 points ranging within ± 0.030 e/Å$^2$) of each cation, anion, and molecule. The $\sigma$-profiles were then divided into eight sections, and each section integrated with the help of the *trapz* function of *numPy* in Python to spawn several descriptors for the involved species ($S_i^{\text{species}}$ of section $i$, $i = 1, 2, \ldots, 8$).

Considering that the DES are in fact mixtures, the final descriptors for their corresponding combinations were calculated as follows:

$$S_i^{\text{DES}} = (x_{\text{HBA}})\left(S_i^{\text{anion}}\right) + \left(S_i^{\text{cation}}\right)\Big) + (x_{\text{HBD}})\left(S_i^{\text{HBD}}\right) \quad (1)$$

where $x_{\text{HBA}}$ and $x_{\text{HBD}}$ are the molar fractions of the HBA and HBD components of the DESs. $S_i^{\text{anion}}$ and $S_i^{\text{cation}}$ stand for the descriptors of the anions and cations, respectively, whereas $S_i^{\text{HBD}}$ represents the descriptors of HBD molecules. Along with these eight $S_i^{\text{DES}}$ descriptors, the temperature ($T$ in K) was also used as an independent variable for establishing the QSPR models. The final $S_i^{\text{DES}}$ descriptors are hereafter simply designated as $S_1$–$S_8$.

**Model Set-up and Validation.** To begin with, the dataset was divided into training and test sets by employing the 'mixture-out' methodology [47]. This approach involves placing all data points of a specific type of DES either in the training set or the test set (i.e.: every mixture with different ratios is present in either the training or the test

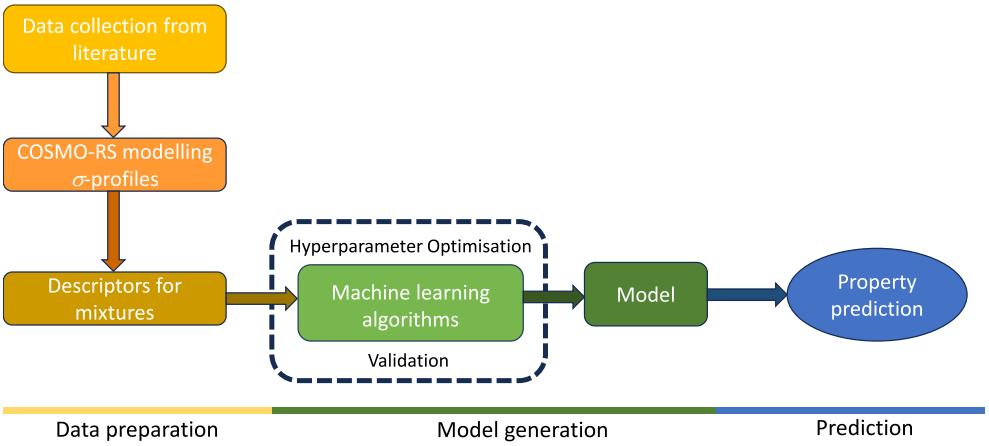


**Fig. 1.** Overview of the methodology workflow followed in the present study.

**Table 1**
List of DESs investigated in this study, along with their composition, temperatures, heat capacity ($C_p$) values, number of data points (Ndp), dataset distribution, and corresponding references.

| DES | HBA | HBD | Molar ratio HBA:HBD | Ndp | Temperatures (K) | $C_p$ (J mol$^{-1}$ K$^{-1}$) | Dataset | Ref. |
|---|---|---|---|---|---|---|---|---|
| D01 | Sodium acetate | Glycerol | 0.33 | 7 | 360–300 | 92.3–81.1 | Train | [37] |
| D02 | Betaine | Ethylene glycol | 0.17 | 11 | 363.15–313.15 | 173.2–165.6 | Test | [38] |
| D03 | Betaine | Ethylene glycol | 0.25 | 11 | 363.15–313.15 | 177.5–161.7 | Test | [38] |
| D04 | Betaine | Glycerol | 0.25 | 1 | 298.15–298.15 | 143.1–143.1 | Train | [16] |
| D05 | Betaine | Propylene glycol | 0.22 | 1 | 298.15–298.15 | 139.9–139.9 | Test | [16] |
| D06 | Choline chloride | Ethylene glycol | 0.33 | 11 | 353.15–303.15 | 205.6–190.8 | Train | [16] |
| D07 | Choline chloride | Glycerol | 0.33 | 28 | 353.15–278.15 | 254.3–211.5 | Test | [16] |
| D08 | Choline chloride | Resorcinol | 0.33 | 23 | 338.15–283.15 | 249.7–222.0 | Train | [39] |
| D09 | Choline chloride | Triethylene glycol | 0.33 | 23 | 353.15–298.15 | 314.6–299.0 | Train | [16] |
| D10 | Choline chloride | Citric acid | 0.33 | 23 | 353.15–298.15 | 449.4–422.0 | Train | [16] |
| D11 | Choline chloride | Fructose | 0.67 | 23 | 353.15–298.15 | 340.1–311.5 | Train | [16] |
| D12 | Choline chloride | Glucose | 0.67 | 23 | 353.15–298.15 | 356.3–327.5 | Test | [16] |
| D13 | Choline chloride | Malonic acid | 0.50 | 23 | 353.15–298.15 | 240.8–226.9 | Train | [16] |
| D14 | Choline chloride | Oxalic acid | 0.33 | 23 | 353.15–298.15 | 282.9–270.8 | Train | [16] |
| D15 | Choline chloride | Phenol | 0.25 | 23 | 353.15–298.15 | 236.8–219.3 | Train | [16] |
| D16 | Choline chloride | Urea | 0.33 | 21 | 338.15–288.15 | 197.7–180.8 | Train | [40] |
| D17 | DL-Menthol | Oleic Acid | 0.48 | 20 | 423–328 | 669.6–551.0 | Train | [2] |
| D18 | L-Carnitine | Ethylene glycol | 0.17 | 11 | 363.15–313.15 | 188.2–180.05 | Train | [38] |
| D19 | L-Carnitine | Ethylene glycol | 0.20 | 11 | 363.15–313.15 | 193.1–170.6 | Train | [38] |
| D20 | L-Carnitine | Ethylene glycol | 0.25 | 11 | 363.15–313.15 | 199.3–177.6 | Train | [38] |
| D21 | Methyl triphenyl phosphonium bromide | Ethylene glycol | 0.20 | 23 | 353.15–298.15 | 256.7–237.6 | Test | [16] |
| D22 | Methyl triphenyl phosphonium bromide | Glycerol | 0.25 | 23 | 353.15–298.15 | 350.2–328.5 | Train | [16] |
| D23 | Methyl triphenyl phosphonium bromide | Malonic acid | 0.40 | 23 | 353.15–298.15 | 354.2–336.9 | Train | [16] |
| D24 | N,N-Diethylethanolammonium chloride | Ethylene glycol | 0.33 | 11 | 353.15–303.15 | 219.8–204.1 | Train | [16] |
| D25 | N,N-Diethylethanolammonium chloride | Glycerol | 0.33 | 11 | 353.15–303.15 | 269.3–250.4 | Test | [16] |
| D26 | Tetrabutylammonium chloride | Ethylene glycol | 0.25 | 23 | 353.15–298.15 | 312.6–288.3 | Train | [16] |
| D27 | Tetrabutylammonium chloride | Glycerol | 0.17 | 23 | 353.15–298.15 | 310.9–281.2 | Test | [16] |
| D28 | Tetrabutylammonium chloride | Triethylene glycol | 0.50 | 23 | 353.15–298.15 | 479.7–445.0 | Test | [16] |
| D29 | Tetrabutylammonium chloride | Malonic acid | 0.25 | 23 | 353.15–298.15 | 342.8–299.8 | Train | [16] |
| D30 | Tetrabutylammonium chloride | Urea | 0.80 | 19 | 353.15–308.15 | 605.9–590.1 | Train | [16] |

set, but never in both sets). The choice of dataset division method plays a crucial role in QSPR modelling of mixtures, and two other methods, namely 'points-out' and 'compounds-out', have been previously suggested. In the 'points-out' method, the test set comprises randomly selected data points from various mixtures, while in the 'compounds-out' method, specific components are identified, and all mixtures containing these components are assigned to either the training or test set. As highlighted by Oprisiu *et al.* [47] and further supported by our prior studies on DESs [24,28,31], the 'mixture-out' validation method is deemed more reliable for establishing models for mixtures compared to the 'points-out' validation method. In contrast, the 'compounds-out' validation method is considered overly stringent and may not be well-suited for small datasets like ours, with a limited number of DESs.

In this work, the dataset division was executed using the open-access Python-based QSAR-Mx software (https://github.com/ncordeirfcup/QSAR-Mx) [24,31], specifically utilising Module-1 with a seed value of 1 and an interval of 4. This module automatically generated the training and test sets along with the $S_i^{DES}$ descriptors (Eq. (1)). Moreover, within the 'mixture-out' method, each binary DES (defined by a specific combination of its two components) is initially sorted in descending order. Then, starting from the second DES, every fourth DES is selected and placed in the test set. This resulted in assigning twenty-two mixtures for training and eight mixtures for testing. To enhance data distribution, the DES labelled as D28 was included in the test set. Consequently, nine DESs with a total of 154 data points constituted the test set, while the remaining 375 data points from twenty-one DESs formed the training set. Importantly, as such, we guaranteed that at least some components were present in both the training and test sets. In this specific case, only glucose and propylene glycol (HBDs) are present exclusively in the test set. Details about the training and test sets distribution are outlined in Table 1.

Regarding the modelling framework, we opted to develop both linear and non-linear models using multiple linear regression (MLR) and a variety of machine learning (ML) algorithms, respectively. To mitigate unbalanced skewness, the response variable was converted to natural logarithmic scaling ($-\ln C_p$), and all models were developed based on the nine descriptors (i.e.: the eight $S_1$–$S_8$ plus $T$). Module-1 of QSAR-Mx was employed to establish the best MLR model. Concerning the non-linear models, these were developed by resourcing to four different machine learning regression algorithms, namely *k*-Nearest Neighbours (*k*NN) [48], Random Forests (RF) [49], Neural Network Multilayer Perceptron (MLP) [50], and Support Vector Machines (SVM) [51]. The non-linear models were generated using the Python-based Scikit-learn libraries, and the descriptors were standardised using the *StandardScaler* function implemented in Scikit-learn [52]. Additionally, a 5-fold cross-validation scheme was utilised to optimise the hyperparameters associated with each machine learning algorithm, employing a systematic grid search for the best model hyperparameters through the *GridSearchCV* function in Scikit-learn.

The parameters optimised during model development are shown in Table 2. However, in the case of MLP, the hidden layer sizes, the number of neurons in each layer, and the alpha values (essentially the learning rate) were initially varied, while keeping fixed the activation function and solver. Once the optimal hidden layers and alpha values were determined, the other parameters were further fine-tuned, such as the activation function and solver, using *GridSearchCV*. Where applicable, a random state of 42 was consistently set for both model development and validation.

To assess the performance of the models, the average absolute relative deviation (AARD) was calculated. AARD, a well-known metric often used to judge the models' ability to predict physicochemical characteristics, is expressed as follows:

$$\text{AARD}\,(\%) = \frac{100}{k} \times \sum_{i}^{k} \frac{\left|\left(Y_{\text{Pred},i} - Y_{\text{Exp},i}\right)\right|}{Y_{\text{Exp},i}} \tag{2}$$

Here, $Y_{\text{Pred},i}$, $Y_{\text{Exp},i}$, and $k$ refer to the predicted heat capacity by the model, the experimental heat capacity values, and the number of data points, respectively. Given that multiple models were established in the present study, the AARD values were used as a guide for selecting the most predictive ones. Additional statistical metrics considered included the coefficient of determination ($R^2$), root mean square error (RMSE), mean absolute residual (MAR), and standard deviation (SD), calculated as follows:

**Table 2**
List of parameters considered for hyperparameter optimisation of the machine learning algorithms exploited in this study.

| Machine learning tool | Parameters |
|---|---|
| *k*NN | n-neighbours: 1–50<br>weights: uniform, distance<br>algorithm: auto, ball_tree, kd_tree, brute |
| SVM | C: 0.1, 1.0, 10, 100, 1000<br>gamma: 1, 0.1, 0.01, 0.001<br>kernel: rbf, linear, poly, sigmoid |
| RF | criterion: MSE, MAE<br>max_features: auto, sqrt, log2<br>max_depth: 10, 30, 50, 70, 90, 100, 200<br>n_estimators: 50, 100, 200<br>min_samples_leaf: 1, 2, 4<br>min_samples_split: 2, 5, 10 |
| MLP | hidden_layer_sizes: varied in a stepwise manner<br>activation: relu, tanh, logistic, identity<br>solver: lbfgs, adam, sgd<br>alpha: 0.1, 0.01, 0.001, 0.0001 |

$$R^2 = 1 - \frac{\sum_i^k \left(Y_{\text{Exp},i} - Y_{\text{Pred},i}\right)^2}{\sum_i^k \left(Y_{\text{Exp},i} - \overline{Y}\right)} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_i^k \left(Y_{\text{Exp},i} - Y_{\text{Pred},i}\right)^2}{k}} \tag{4}$$

$$\text{MAR} = \frac{1}{k} \times \sum_{i}^{k} \left|\left(Y_{\text{Exp},i} - Y_{\text{Pred},i}\right)\right| \tag{5}$$

$$\text{SD} = \sqrt{\frac{\sum_i^k \left(Y_{\text{Pred},i} - \overline{Y}\right)^2}{k}} \tag{6}$$

in which, $\overline{Y}$ refers to the mean value of the experimental heat capacity.

Furthermore, internal $R^2$ cross-validation measures based on the leave-one-out ($Q^2_{\text{LOO}}$) or leave-many-out ($Q^2_{\text{5-fold}}$) approaches were taken into account [53]. For external validation, the model's predictive ability ($R^2_{\text{Pred}}$) was calculated [54]. The equations for these parameters are similar to $R^2$, but their application is different. For instance, in leave-one-out (LOO), the LOO predicted values ($Y_{\text{LOO}}$) are used in place of $Y_{\text{Pred}}$ to calculate $Q^2_{\text{LOO}}$ applying Eq. (3). In contrast, the *k*-fold cross-validation is employed in the leave-many-out approach to assess the internal predictivity of the ML-based non-linear models. To do so, the dataset is partitioned into five equal subsets for the determination of $Q^2_{\text{5-fold}}$, with 80 % of the data (four subsets) utilised as the training set and the remaining 20 % used as the prediction set. This procedure is repeated five times to collect all the predicted results for all five test sets, and the resulting values are used as $Y_{\text{Pred}}$ in Eq. (3). As previously mentioned, the 5-fold cross-validation results were exploited not only for assessing the model's internal predictivity, but also for hyperparameter optimisation during the set-up of ML models. In the latter case, the dependent variable for each DES in the test set is predicted ($Y_{\text{Pred}}$) using the specific derived model, and then, these values are employed in the $R^2_{\text{Pred}}$ calculation. It is worth mentioning that while calculating these statistical parameters, the logarithmic response variable values were converted back to their original values (*i.e.*, the heat capacity in J mol$^{-1}$ K$^{-1}$).

**Applicability domain of the models.** To ensure the reliability of the QSPR model predictions, it is crucial to define their applicability domain, which outlines the chemical space boundaries within which the models are valid. In this study, two complementary methodologies were adopted to determine the applicability domain of our models. In the first method, popularly known as the 'Leverage approach', a Williams plot (standardised residuals vs. leverage values) is built to define structural outliers (in terms of leverage values) and response outliers (in terms of standardised residuals). Specifically, data-points whose leverage values exceed the threshold $h^*$ ($h^* = 3p/N$, where $p$ is the number of descriptors in the model plus one, and $N$ is total number of data-points in the training set) were spotted as structural outliers [55]. The second method that has been proposed by Roy et al. [56], known as the 'Standardization approach', involves standardising the model descriptors and applying simple rules to identify structural outliers. Data-points flagged as structural outliers by either of these methods were treated as outliers of the model.

**Model Interpretation.** The linear models may simply be interpreted from the standardised coefficients associated with the descriptors. On the other hand, tools such as SHAP (SHapley Additive exPlanations) approach help understand the contributions of different independent parameters of ML non-linear models [57].

The SHAP approach, proposed by Lundberg and Lee [57], is based on the well-known Shapley value concept, initially introduced by the mathematician Lloyd Shapley [58]. SHAP is a tool for model interpretability, aiming to make machine learning models more transparent and

understandable. It leverages concepts from cooperative game theory to fairly allocate the contribution of each feature to the model's predictions. SHAP achieves this by distributing the overall reward among different "players" (descriptors) based on their relative contributions to the model's outcome. It provides information about each descriptor's positive (increasing values) and negative (decreasing values) contributions to the response variable under consideration, ranking descriptors based on their overall influence on the model. This approach thus offers a nuanced understanding of the impact of individual descriptors within non-linear machine learning models [59].

## 3. Results and discussion

**Decoding the Physical Significance of $\sigma$-Profiles.** Essentially, COSMO-RS $\sigma$-profiles characterise the surface polarity of species and serve as a means of providing insights into the nature of intermolecular interactions among the components of a mixture, including both polar and nonpolar interactions. Fig. 2 showcases illustrative examples of the probability distribution of the computed $\sigma$-profiles for selected ions and HBD components, alongside representations of their 3D structures. In this representation, red indicates negative (electron-rich) regions, blue represents (electron-deficient) regions, and green highlights the nonpolar, neutral regions of the species. Additionally, three main areas can be identified in this figure depending on the screening charge density: the HBD area, the nonpolar (hydrophobic) area, and the HBA area. The two dotted perpendicular lines, positioned at −0.007 and +0.007 e/Å$^2$, demarcate the boundaries between the nonpolar area and the two polar areas (HBD and HBA).

As depicted in Fig. 2**a**, it is evident that anions and cations with a single functional group exhibit only one thin peak either in the HBA area ($Cl^-$ and $Br^-$) or in the HBD area ($Na^+$) according to their respective charges. Other cations, such as methyltriphenylphosphonium ($MTPP^+$), choline ($Ch^+$), and tetrabutylammonium ($TBA^+$) cations display prominent peaks mainly in the nonpolar area due to the presence of nonpolar alkyl groups (like methyl and butyl). Upon comparing $Ch^+$ and $TBA^+$, it is noticeable that the former has a $\sigma$-profile curve more shifted towards the left (more positive), whereas the latter displays a single broad peak in the nonpolar region. This distinction arises from the structural differences between the two cations. $Ch^+$ contains a positively charged nitrogen ($N^+$) and a hydroxyl group (–OH) with a positively charged hydrogen, with its fewer carbon atoms unable to balance such a positive charge, resulting in a more positive overall profile. In contrast, $TBA^+$ features a symmetrical structure with four identical butyl groups attached to the central nitrogen. As a result, the positive charge on the nitrogen is effectively neutralised by the surrounding butyl groups, leading to a high and broad peak in the nonpolar area. It is also seen that aromatic-containing cations (C=C), like $MTPP^+$, feature two peaks in the nonpolar area instead of just one peak.

From Fig. 2b, it can be observed that all the HBD components have multiple broad peaks spanning both the HBA and HBD areas of the $\sigma$-profile, indicating that these molecules can act as both HBAs and HBDs. Moreover, the height of these peaks is much lower than the ones of the ions, save for the case of triethylene glycol (TG) that has a dominant high peak in the nonpolar area. This can be explained by the presence of a longer hydrocarbon chain with two ether linkages in TG that contributes to a more prominent nonpolar hydrophobic region in its $\sigma$-profile compared to the other diols, like propylene glycol (PG) and ethylene glycol (EG).

Overall, the analysis of Fig. 2 reinforces the concept of $\sigma$-profiles reflecting the charge distribution and polarity of species. It demonstrates how these profiles can be utilised to compare and understand the properties of different DES components and highlights the usefulness of descriptors derived from the integration of target $\sigma$-profile sections for QSPR modelling studies.

**QSPR Models and Evaluation.** Following the strategy formerly outlined, we began by seeking the best linear model relating the heat capacity ($-\ln C_p$) and the eight $S_i^{DES}$-based descriptors ($S_1$-$S_8$) along with the measured temperature ($T$) for the training set. However, it became evident that $S_2$ held less significance in the model, as indicated by a probability value ($p$-value) exceeding 0.05. Consequently, the final model was established using the remaining descriptors. The MLR equation and its detailed statistical results are presented in Table 3. It can be observed that the MLR model achieved an AARD of 8.516 % through leave-one-out (LOO) cross-validation and an AARD of 7.845 % during its evaluation using the test set comprising 154 data points.

The maximum intercorrelation (Pearson $R$) between any two descriptors of this model was found to be 0.73, which is satisfactory. Furthermore, although the developed MLR model demonstrated good

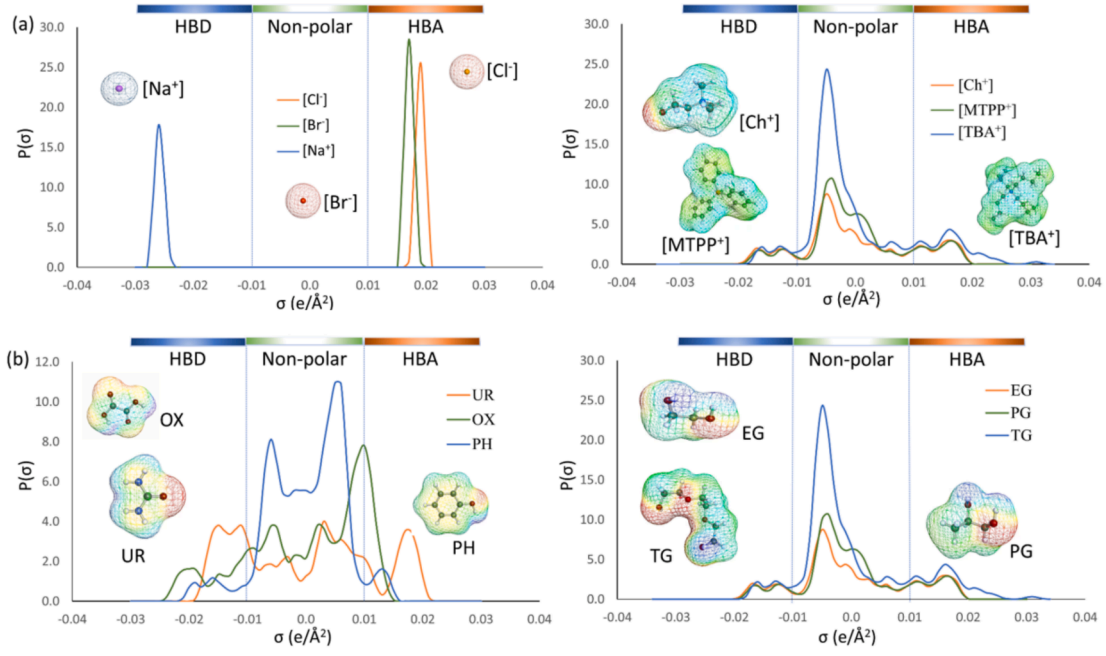


**Fig. 2.** Probability distribution of the derived $\sigma$-profiles for six selected **(a)** ions and **(b)** HBD components.

**Table 3**
Statistical performance of the developed MLR model.

| Equation | Statistical results |
|---|---|
| $-\ln C_p = -65.766(\pm3.426)\ S_1 - 5.659(\pm1.850)\ S_3 + 4.429(\pm0.526)\ S_4 + 9.527(\pm0.962)\ S_5 + 24.073(\pm1.284)\ S_6 + 23.783(\pm2.942)\ S_7 - 34.373(\pm7.848)\ S_8 + 0.001(\pm0.000)\ T + 4.058(\pm0.102)$ | $N_{training} = 376$, $R^2 = 0.942$, $Q^2_{LOO} = 0.940$, $AARD_{LOO} = 8.620$ %, $MAR_{LOO} = 25.170$, $RMSE_{LOO} = 30.648$, $N_{test} = 154$, $R^2_{Pred} = 0.889$, $AARD_{test} = 8.257$ %, $MAR_{test} = 25.140$, $RMSE_{test} = 30.992$ |

statistical predictivity in terms of $Q^2_{LOO}$ (= 0.940) and $R^2_{Pred}$ (= 0.889), the $MAR_{LOO}$, $RMSE_{LOO}$, $MAR_{test}$ and $RMSE_{test}$ values were considered slightly high. Given the potential limitations of the MLR model, we explored whether different machine learning tools could yield more statistically significant results using all the nine descriptors (*i.e.*: $S_1$–$S_8$ and $T$). In search of potentially better non-linear models, four machine learning (ML) algorithms (*k*NN, RF, SVM, and MLP) were explored, utilising hyperparameter optimisation for fine-tuning. As mentioned, a slightly different hyperparameter optimisation technique was employed for MLP compared to other ML tools. For *k*NN, RF, and SVM, all parameters listed in Table 2 were simultaneously varied. However, for setting up the final MLP model, we took a step-by-step approach. Initially, we adjusted two crucial parameters – the number of hidden layers and alpha values – while keeping default values for all other parameters, except for the solver and maximum number of iterations. Preliminary trials indicated that the 'lbgfs' (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) solver, coupled with a higher number of iterations, consistently produced superior model quality. Therefore, the solver was initially set as lbgfs and the maximum number of iterations was fixed at 7000. Subsequently, we optimised the number of hidden layers, number of neurons in the hidden layers and alpha value. We started with a single hidden layer, varying the number of neurons from 5 to 30 in increments of 5. We then explored a two-layer architecture, with the number of neurons in the second layer being less than the first. Additionally, for each neuron configuration, we tested four different alpha values: 0.1, 0.01, 0.001, and 0.0001. The predictive accuracy of the training set model was assessed by 5-fold cross-validated AARD values ($AARD_{5\text{-fold}}$). The $AARD_{5\text{-fold}}$ of the training set improved (> 10 %) in the presence of two layers compared to one layer. As depicted in Fig. 3, the lowest $AARD_{5\text{-fold}}$ was achieved when 30 and 25 neurons were added in two layers while maintaining the alpha value at 0.01. No significant improvement (*i.e.*, 5 %) in $AARD_{5\text{-fold}}$ was achieved when a third layer was introduced.

After fixing the number of hidden layers, neurons and alpha values, the activity function and solver parameters were varied, but with no improvement on the model behaviour.

The AARD values of the final fine-tuned machine learning models thus established, along with the optimised parameters, are presented in Table 4.

Among these ML algorithms, *k*NN and RF showed high overfitting, evidenced by significantly lower AARD values in 5-fold cross-validation on the training set (*i.e.*, $AARD_{5\text{-fold}}$) compared to the test set (i.e., $AARD_{test}$). Conversely, SVM produced a balanced model with AARD values of 7.097 % (5-fold) and 10.308 % (test). Yet, MLP yielded the most predictive non-linear model, with an AARD of 0.500 % (5-fold) and superior performance on the test set. The detailed statistical results of this model are presented in Table 5.

As shown in Table 5, the final MLP model exhibited satisfactory statistical predictive performance for both the training and test sets. Fig. 4 displays a comparison between the experimental heat capacities and the predicted values (Fig. 4A), as well as the residual values (Fig. 4B), alongside with the correlation matrix for the descriptors in the MLP model. By observing Fig. 4A and Fig. 4B, it becomes apparent that this model accurately predicts DESs with both low and high heat capacities, while its performance is less optimal for those with intermediate values. Additionally, the correlation matrix presented in Fig. 4C indicates that the model's descriptors do not demonstrate strong inter-collinearity (> 0.80).

**Applicability domain of the models**. As per the Williams plot, the results revealed that all data-points belonging to D01 (Table 1) of the training set were projected as structural outliers. Interestingly, these same data-points were depicted as outliers according to the 'Standardization approach'. In addition, the data-points in D17 (Table 1) were further flagged as outliers by the 'Standardization approach'. Nevertheless, none of the test set data-points were classified as outliers by either method, indicating that the model's predictions for the test set are likely reliable within the established applicability domain.

**Table 4**
Comparison of performance metrics for four machine learning models.

| Model | $AARD_{5\text{-fold}}$ (%) | $AARD_{test}$ (%) | Selected Parameters |
|---|---|---|---|
| *k*NN | 0.452 | 14.669 | n_neighbours = 2, weights = 'distance' |
| RF | 0.445 | 12.833 | max_depth = 10, max_features = 'sqrt', random_state = 42 |
| SVM | 7.097 | 10.308 | gamma = 0.1 |
| MLP | 0.500 | 3.999 | alpha = 0.01, hidden_layer_sizes = (30, 25,), max_iter = 7000, random_state = 42, solver = 'lbfgs' |

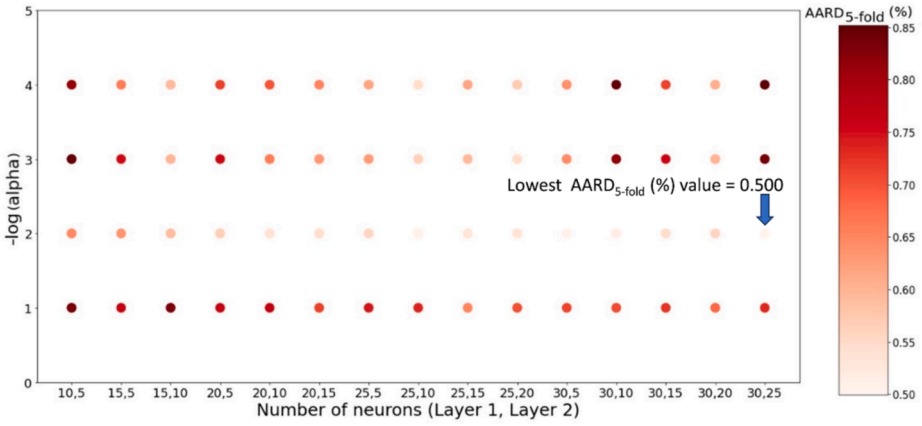


**Fig. 3.** Variation of the predictivity of the MLP model developed in the first step with two hidden layers in the presence of different numbers of neurons and alpha values (represented as −log(alpha) for clarity; alpha = 0.0001, 0.001, 0.01, 0.1).

**Table 5**
Detailed statistical results for the final MLP model.

| Training set parameters | | Test set parameters | |
|---|---|---|---|
| $R^2$ | 0.999 | – | – |
| $Q^2_{5\text{-fold}}$ | 0.999 | $R^2_{\text{Pred}}$ | 0.976 |
| $\text{AARD}_{5\text{-fold}}$ (%) | 0.500 | $\text{AARD}_{\text{test}}$ (%) | 3.999 |
| $\text{MAR}_{5\text{-fold}}$ (J mol$^{-1}$ K$^{-1}$) | 1.358 | $\text{MAR}_{\text{test}}$ (J mol$^{-1}$ K$^{-1}$) | 11.316 |
| $\text{RMSE}_{5\text{-fold}}$ (J mol$^{-1}$ K$^{-1}$) | 2.860 | $\text{RMSE}_{\text{test}}$ (J mol$^{-1}$ K$^{-1}$) | 14.282 |

**Feature importance.** To understand the explainability behind the final MLP model's predictions for DES critical properties ($C_p$), we employed the SHAP approach to analyse the relative importance and contributions of the model's features (descriptors). Particularly, we focused on the significance of its $S_i$ descriptors in capturing the DES's ability to act as a hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), and its non-polar character. As previously mentioned, we divided the $\sigma$-profile curves into eight sections. By calculating the integral area under each section, we obtained eight molecular descriptors ($S_1$-$S_8$) that quantify these specific regions of the $\sigma$-profile (see Table 6).

Fig. 5 shows the SHAP results of the feature analysis of the MLP model. In the subplots of this figure, the features are sorted from the most significant one to the least significant. Higher feature values (highlighted in red in Fig. 5A) of the descriptor indicate a higher heat capacity if they are associated with positive SHAP values. Conversely, if a higher feature value of any descriptor is linked to a negative SHAP value (highlighted in blue in Fig. 5A), it suggests that a higher value of this descriptor decreases the heat capacity of DES.

The SHAP analysis applied to the MLP model pinpointed the non-polar descriptor $S_4$ as the primary contributor to higher heat capacity values. Elevated $S_4$ values are attributed to the presence of non-polar fragments like –CH, $–CH_2$, and $–CH_3$. Consequently, DESs containing components such as tetrabutylammonium chloride (D28, D30), DL-menthol (D17), and oleic acid (D17) exhibit higher heat capacities as they possess these non-polar fragments in their structures. Similarly, belonging to a non-polar region, $S_5$ was also found to contribute to increased heat capacities, albeit to a lesser extent than $S_4$.

Additionally, $S_8$ emerged as the second most relevant descriptor of the model, indicating that fragments comprising strong HBAs exhibit low DES' heat capacity. For example, HBAs such as sodium acetate, betaine, and carnitine that have high $S_8$ values consistently exhibit lower heat capacities. Interestingly, triethylene glycol (TG) is the only HBD in the dataset that has a positive value for this descriptor.

The favourable effect of $S_7$ can be seen in D30, which has high heat capacity values (590–606 J mol$^{-1}$ K$^{-1}$) owing to the presence of urea. Moreover, $S_7$ values are bolstered by the existence of HBAs containing anions such as $Cl^-$ and $Br^-$. The weak acceptor region ($S_6$) also contributes to increase the heat capacity values, as observed in D10, where the HBD, citric acid, presents a high $S_6$ value, thereby influencing the rise of DES's heat capacity.

Descriptors associated with the hydrogen bond donor region ($S_1$-$S_3$) tend to have higher heat capacities, as their higher values (red points in Fig. 5A) are situated in the positive region of the SHAP analysis.

**Table 6**
Partitioning within the $\sigma$ range into eight descriptors.

| Descriptors | $\sigma$-range (e/Å$^2$) | Significance |
|---|---|---|
| **Hydrogen bond donor area** | | Strong HBD |
| $S_1$ | $-0.030 < \sigma < -0.023$ | (e.g.: $Na^+$) |
| $S_2$ | $-0.023 < \sigma < -0.015$ | HBD (e.g.: $H^+$) |
| $S_3$ | $-0.014 < \sigma < -0.007$ | Weak HBD (e.g.: $N^+$, $P^+$) |
| **Non-polar area** | | Non-polar |
| $S_4$ | $-0.007 < \sigma < 0.000$ | (e.g.: C, H) |
| $S_5$ | $0.000 < \sigma < +0.007$ | Non-polar |
| **Hydrogen bond acceptor area** $S_6$ | $+0.007 < \sigma < +0.015$ | Weak HBA (e.g.: $C{=}O^{(\delta-)}$) |
| $S_7$ | $+0.015 < \sigma < +0.023$ | HBA (e.g.: $Cl^-$, $Br^-$) |
| $S_8$ | $+0.023 < \sigma < +0.030$ | Strong HBA |

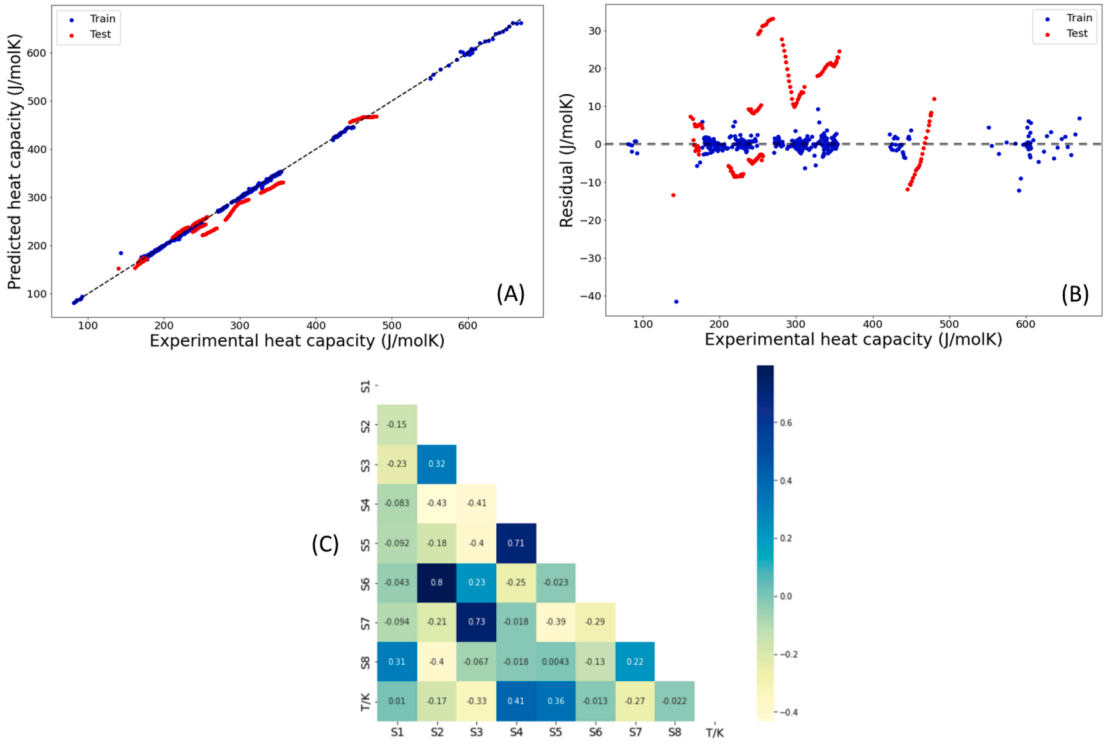


**Fig. 4.** Plots of (A) observed vs. predicted heat capacity values and (B) residual vs. observed heat capacity values as *per* the MLP model. (C) Heatmap showing the correlation matrix for the descriptors of the MLP model.

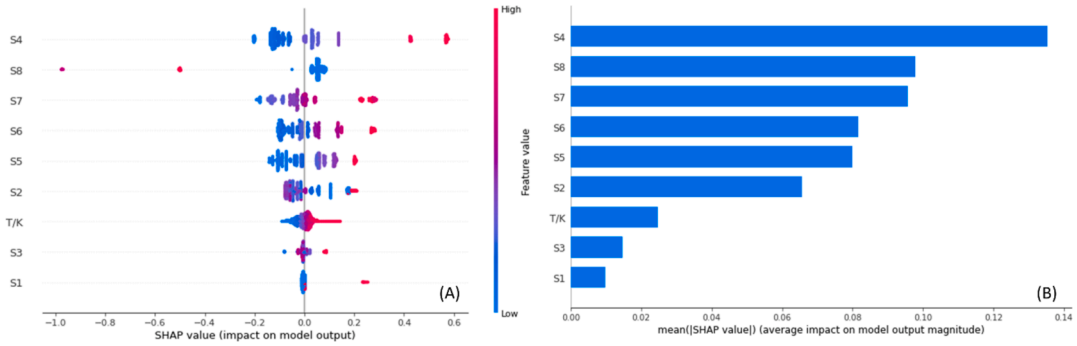


**Fig. 5.** SHAP Feature Analysis of the MLP Model. (A) Summary plot showing feature contributions (colour) and importance (dot size) for model predictions (Positive SHAP values indicate favourable contributions, while negative values indicate unfavourable contributions.). (B) Bar plot depicting the mean absolute SHAP value for each feature, highlighting their relative importance to the model.

However, the significance of these descriptors is relatively less as compared to those of $S_4$-$S_8$. The measuring temperature is positively correlated since, in most cases, an increase in temperature leads to a steady increase in heat capacity values.

To evaluate the performance of our model in anticipating these occurrences, we plotted the predicted and experimental heat capacity values as a function of temperature for the DESs test set (Fig. 6). It is noteworthy that the majority of the test set DESs reflects changes in heat capacity with increasing temperature. It is noticeable also that, save for the DES formed by choline chloride and glucose, all other DES are accurately predicted by the model.

Finally, we sought to determine whether our proposed model could effectively predict the heat capacity for entirely distinct data, such as ternary DESs. To accomplish this, we gathered data from 44 ternary DESs consisting of choline chloride, urea, and L-arginine, as documented in Table 7. Here, it is worth noting that all these ternary DESs were found to lie within the model's applicability domain. The model exhibited an AARD of 3.132 %, indicating its potential applicability for predicting the heat capacities of ternary DESs as well. This also underscores the model robustness and versatility across diverse compositions, further validating its utility in predictive modelling.

**Comparison with previous models.** Our literature review revealed three previous attempts to develop QSPR models for the heat capacity of DESs. These models, however, exclusively relied on thermodynamic critical features of the DESs. In contrast, our work pioneers the use of quantum chemical COSMO-RS descriptors to characterise and predict the heat capacity of DESs, offering deeper insights into the underlying mechanisms governing their heat capacities. Taherzadeh et al. [16] proposed an equation based on critical properties and molecular weight to calculate the isobaric heat capacity of eutectic solvents. Using a dataset comprising 28 DESs with 505 data points, their model achieved AARD values of 4.3 % and 5.5 % for the training and test sets, respectively. Our model, which incorporates a larger dataset and exhibits improved statistical performance, therefore significantly surpasses these previous results. Later, Bagherzadeh et al. [27] employed fifteen machine learning techniques to predict the heat capacity of DESs, utilising a combination of critical properties with experimental temperature. Despite achieving an excellent overall AARD of 0.27 %, their model relied on a random split of data into training and test sets (i.e., 'points-out' validation; *Cf.* Section 2), limiting its robustness. Very recently, Darwish et al. [35] developed a group-contribution model using a dataset of 2,696 data-points. However, all these data-points were not reported as DESs but simply as IL mixtures in the original literature, potentially neglecting the specific hydrogen bonding interactions characteristic of DESs. By focussing specifically on DESs and leveraging quantum chemical descriptors, our proposed model provides a more accurate and insightful approach to predicting their heat capacity.

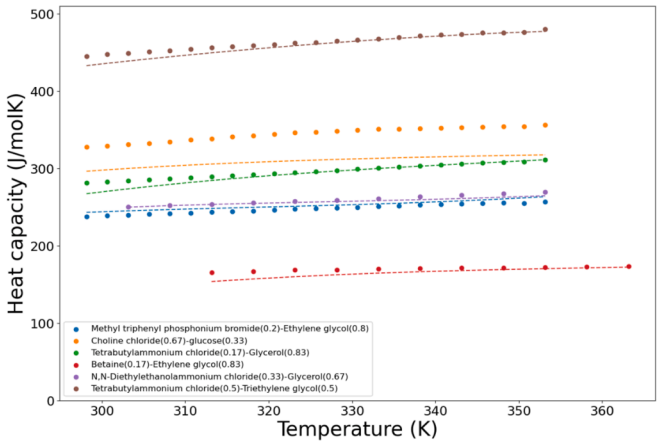


**Fig. 6.** Comparison of heat capacities calculated by the MLP model (lines) with experimental data (points) from the literature across varying measurement temperatures (in K).

## 4. Conclusions

This work highlights the importance of heat capacity, an often underreported yet crucial physicochemical property for deep eutectic solvents. Our objective was to bridge this gap by developing a robust predictive model for DES heat capacities. Through the compilation of an extensive dataset comprising over 500 data points, we explored various machine learning approaches. Our most statistically robust model emerged from the MLP technique, utilising descriptors derived from quantum-chemical COSMO-RS sigma profiles. This innovative approach represents a significant advancement in DES heat capacity prediction, as it leverages the detailed information captured within COSMO-RS $\sigma$-profiles.

In comparison to a prior model reliant on correlations with thermodynamic descriptors, which encompassed a broader array of DES types (24 binary and 4 ternary) [16], our research focused specifically on binary DESs (30 types). This focus, coupled with the power of COSMO-RS descriptors, yielded a remarkable improvement in accuracy. Our model achieved exceptionally low AARD values of 0.500 % and 3.999 % for the training and test sets, respectively, demonstrating a significant reduction in error compared to the previous model's 4.3 % and 5.5 % AARD [16].

Beyond establishing a highly accurate predictive tool, this research offers valuable insights for future DES design. By employing the SHAP approach, we were able to identify and understand the key chemical features within DES components that significantly contribute to their heat capacity values. This knowledge enables researchers to customise DES development for specific applications where heat capacity is crucial, such as thermal energy storage. In conclusion, this work paves

**Table 7**
List of ternary DESs investigated in this study, along with their composition, number of data points (Ndp), temperatures, heat capacity ($C_p$) values, and corresponding references.

| DES | HBA | HBD1, HBD2 | Molar Fractions HBA:HBD1:HBD2 | Ndp | Temperatures (K) | $C_p$ (J mol$^{-1}$ K$^{-1}$) | Ref. |
|---|---|---|---|---|---|---|---|
| T01 | Choline chloride | Urea, L-arginine | 0.328:0.656:0.016 | 11 | 303.15–353.15 | 182.21–193.74 | [16] |
| T02 | Choline chloride | Urea, L-arginine | 0.323:0.645:0.032 | 11 | 303.15–353.16 | 184.38–195.82 | [16] |
| T03 | Choline chloride | Urea, L-arginine | 0.317:0.635:0.048 | 11 | 303.15–353.17 | 186.54–197.89 | [16] |
| T04 | Choline chloride | Urea, L-arginine | 0.313:0.625:0.063 | 11 | 303.15–353.18 | 188.70–199.96 | [16] |

the way for a new era of accurate DES heat capacity prediction by combining MLP and COSMO-RS descriptors. It not only offers a reliable tool for future DES design but also equips researchers with essential knowledge to engineer DESs with tailored functionalities.

## CRediT authorship contribution statement

**Amit Kumar Halder:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Reza Haghbakhsh:** Writing – review & editing, Validation, Formal analysis, Data curation. **Elisabete S.C. Ferreira:** Writing – review & editing, Writing – original draft, Validation, Formal analysis, Data curation. **Ana Rita C. Duarte:** Validation, Resources. **M. Natália D.S. Cordeiro:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

## Funding

This work received financial support from FCT/MCTES (UIDB/50006/2020 DOI 10.54499/UIDB/50006/2020) through national funds.

## 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.

## Acknowledgements

This work received financial support from FCT/MCTES (LA/P/0008/2020 DOI 10.54499/LA/P/0008/2020, UIDP/50006/2020 DOI 10.54499/UIDP/50006/2020, and UIDB/50006/2020 DOI 10.54499/UIDB/50006/2020), through national funds. The authors also acknowledge FCT by funding CEEC projects 2020.01423.CEECIND/CP1596/CT0003 and 10.54499/2022.05803.CEECIND/CP1725/CT0003.

## Data availability

The data underlying the key findings of this study are available in the published article (*e.g.*, Tables). For inquiries about data beyond what is included in the article, please contact the corresponding author.

## References

[1] N.K. Das, S. Santra, P.K. Naik, M.S. Vasa, R. Raj, S. Bose, et al., Evaluation of thermophysical properties and thermal performance of amine-functionalized graphene oxide/deep eutectic solvent nanofluids as heat-transfer media for desalination systems, ACS Sustain. Chem. Eng. 11 (2023) 5376–5389.
[2] P. Dehury, U. Mahanta, R. Singh, T. Banerjee, Potential of deep eutectic solvent based nanofluids as a new generation heat transfer media, J. Mol. Liq. 379 (2023) 121700.
[3] A.K. Halder, M.N.D.S. Cordeiro, Probing the environmental toxicity of deep eutectic solvents and their components: an in silico modeling approach, ACS Sustain. Chem. Eng. 7 (2019) 10649–10660.
[4] K. Roy, R.N. Das, P.L.A. Popelier, Predictive QSAR modelling of algal toxicity of ionic liquids and its interspecies correlation with Daphnia toxicity, Environ. Sci. Pollut. Res. 22 (2014) 6634–6641.
[5] M. Deetlefs, K.R. Seddon, Assessing the greenness of some typical laboratory ionic liquid preparations, Green Chem. 12 (2010) 17–30.
[6] A.P. Abbott, G. Capper, D.L. Davies, R.K. Rasheed, V. Tambyrajah, Novel solvent properties of choline chloride/urea mixtures, Chem. Commun. 70–71 (2003).
[7] B.B. Hansen, S. Spittle, B. Chen, D. Poe, Y. Zhang, J.M. Klein, et al., Deep eutectic solvents: a review of fundamentals and applications, Chem. Rev. 121 (2020) 1232–1285.
[8] D.C. de Andrade, S.A. Monteiro, J. Merib, A review on recent applications of deep eutectic solvents in microextraction techniques for the analysis of biological matrices, Adv. Sample Prep. 1 (2022) 100007.
[9] A.E. Ünlü, A. Arıkaya, S. Takaç, Use of deep eutectic solvents as catalyst: a mini-review, Green Process. Synth. 8 (2019) 355–372.
[10] Q. Zhang, V.K. De Oliveira, S. Royer, F. Jérôme, Deep eutectic solvents: syntheses, properties and applications, Chem. Soc. Rev. 41 (2012) 7108–7146.
[11] J. Naser, F.S. Mjalli, Z.S. Gano, Molar heat capacity of selected type III deep eutectic solvents, J. Chem. Eng. Data 61 (2016) 1608–16015.
[12] A. Kityk, M. Hnatko, V. Pavlik, M. Boča, Electrochemical surface treatment of manganese stainless steel using several types of deep eutectic solvents, Mater. Res. Bull. 141 (2021) 111348.
[13] E.L. Smith, A.P. Abbott, K.S. Ryder, Deep eutectic solvents (DESs) and their applications, Chem. Rev. 114 (2014) 11060–11082.
[14] A. Bakhtyari, R. Haghbakhsh, A.R.C. Duarte, S. Raeissi, A simple model for the viscosities of deep eutectic solvents, Fluid Phase Equilibria 521 (2020) 112662.
[15] R. Haghbakhsh, M. Taherzadeh, A.R.C. Duarte, S. Raeissi, A general model for the surface tensions of deep eutectic solvents, J. Mol. Liq. 307 (2020) 112972.
[16] M. Taherzadeh, R. Haghbakhsh, A.R.C. Duarte, S. Raeissi, Estimation of the heat capacities of deep eutectic solvents, J. Mol. Liq. 307 (2020) 112940.
[17] R. Haghbakhsh, S. Sona Raeissi, A.R.C. Duarte, Group contribution and atomic contribution models for the prediction of various physical properties of deep eutectic solvents, Sci. Rep. 11 (2021) 6684.
[18] T. Di Pietro, L. Cesari, F. Mutelet, Group contribution models for densities and heat capacities of deep eutectic solvents, Fluid Phase Equilib. 572 (2023) 113854.
[19] A. Boublia, T. Lemaoui, G. Almustafa, A.S. Darwish, Y. Benguerba, F. Banat, et al., Critical properties of ternary deep eutectic solvents using group contribution with extended Lee–Kesler mixing rules, ACS Omega 8 (2023) 13177–13191.
[20] I.I.I. Alkhatib, D. Bahamon, F. Llovell, M.R.M. Abu-Zahra, L.F. Vega, Perspectives and guidelines on thermodynamic modelling of deep eutectic solvents, J. Mol. Liq. 298 (2020) 112183.
[21] E.S.C. Ferreira, I.V. Voroshylova, N.M. Figueiredo, M.N.D.S. Cordeiro, Molecular dynamic study of alcohol-based deep eutectic solvents, J. Chem. Phys. 155 (2021) 064506.
[22] K. Jeong, J.G. McDaniel, A. Yethiraj, Deep eutectic solvents: molecular simulations with a first-principles polarizable force field, J. Phys. Chem. B 125 (2021) 7177–7186.
[23] T. Lemaoui, A.S. Darwish, A. Attoui, F. Abu Hatab, N.E.H. Hammoudi, Y. Benguerba, et al., Predicting the density and viscosity of hydrophobic eutectic solvents: towards the development of sustainable solvents, Green Chem. 22 (2020) 8511–8530.
[24] A.K. Halder, R. Haghbakhsh, I.V. Voroshylova, A.R.C. Duarte, M.N.D.S. Cordeiro, Density of deep eutectic solvents: the path forward cheminformatics-driven reliable predictions for mixtures, Molecules 26 (2021) 5779.
[25] T. Lemaoui, F. Abu Hatab, A.S. Darwish, A. Attoui, N.E.H. Hammoudi, G. Almustafa, et al., Molecular-based guide to predict the pH of eutectic solvents: promoting an efficient design approach for new green solvents, ACS Sustain. Chem. Eng. 9 (2021) 5783–5808.
[26] R.M.A. Bunquin, A.R. Caparanga, Predicting the heat capacities of ammonium- and phosphonium-based deep eutectic solvents using artificial neural network, J. Phys. Conf. Ser. 1893 (2021) 012001.

[27] A. Bagherzadeh, N. Shahini, D. Saber, P. Yousefi, S.M.S. Alizadeh, S. Ahmadi, F. T. Shahdost, Developing a global approach for determining the molar heat capacity of deep eutectic solvents, Measurement 188 (2022) 110630.
[28] A.K. Halder, R. Haghbakhsh, I.V. Voroshylova, A.R.C. Duarte, M.N.D.S. Cordeiro, Predicting the surface tension of deep eutectic solvents: a step forward in the use of greener solvents, Molecules 27 (2022) 4896.
[29] T. Lemaoui, A. Boublia, A.S. Darwish, M. Alam, S. Park, B.-H. Jeon, et al., Predicting the surface tension of deep eutectic solvents using artificial neural networks, ACS Omega 7 (2022) 32194–32207.
[30] A.K. Lavrinenko, I.Y. Chernyshov, E.A. Pidko, Machine learning approach for the prediction of eutectic temperatures for metal-free deep eutectic solvents, ACS Sustain. Chem. Eng. 11 (2023) 15492–15502.
[31] A.K. Halder, P. Ambure, Y. Perez-Castillo, M.N.D.S. Cordeiro, Turning deep-eutectic solvents into value-added products for CO2 capture: a desirability-based virtual screening study, J. $CO_2$ Util. 58 (101926) (2022).
[32] T. Lemaoui, A. Boublia, S. Lemaoui, A.S. Darwish, B. Ernst, M. Alam, et al., Predicting the $CO_2$ capture capability of deep eutectic solvents and screening over 1000 of their combinations using machine learning, ACS Sustain. Chem. Eng. 11 (2023) 9564–9580.
[33] I. Salahshoori, A. Baghban, A. Yazdanbakhsh, Novel hybrid QSPR-GPR approach for modeling of carbon dioxide capture using deep eutectic solvents, RSC Adv. 13 (2023) 30071–30085.
[34] V. Odegova, A. Lavrinenko, T. Rakhmanov, G. Sysuev, A. Dmitrenko, V. Vinogradov, DESignSolvents: an open platform for the search and prediction of the physicochemical properties of deep eutectic solvents, Green Chem. 26 (2024) 3958–3967.
[35] A.S. Darwish, R. Abu Alwan, A. Boublia, T. Lemaoui, Y. Benguerba, I.M. AlNashef, et al., Machine learning approach for mapping the heat capacity of deep eutectic solvents for sustainable energy applications, Fuel 381 (2025).
[36] Z. Dai, Y. Chen, C. Liu, X. Lu, Y. Liu, X. Ji, Prediction and verification of heat capacities for pure ionic liquids, Chin. J. Chem. Eng. 31 (2021) 169–176.
[37] W. Sun, Q. Liu, J. Zhao, H. Muhammad Ali, Z. Said, C. Liu, Experimental study on sodium acetate trihydrate/glycerol deep eutectic solvent nanofluids for thermal energy storage, J. Mol. Liq. 372 (2023) 121164.
[38] Y. Hou, B. Zhang, M. Gao, S. Ren, W. Wu, Densities, viscosities and specific heat capacities of deep eutectic solvents composed of ethanediol + betaine and ethanediol + L-carnitine for absorbing $SO_2$, J. Chem. Thermodyn. 179 (2023) 106999.
[39] L. Lomba, F. Tucciarone, B. Giner, M. Artal, C. Lafuente, Thermophysical characterization of choline chloride: Resorcinol and its mixtures with water, Fluid Phase Equilib. 557 (2022) 113435.
[40] D. Lapeña, F. Bergua, L. Lomba, B. Giner, C. Lafuente, A comprehensive study of the thermophysical properties of reline and hydrated reline, J. Mol. Liq. 303 (2020) 112679.
[41] A. Klamt, Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena, J. Phys. Chem. 99 (1995) 2224–2235.
[42] A. Klamt, V. Jonas, T. Bürger, J.C.W. Lohrenz, Refinement and parametrization of COSMO-RS, J. Phys. Chem. A 102 (1998) 5074–5085.
[43] A. Klamt, The COSMO and COSMO-RS solvation models, WIREs Comput. Mol. Sci. 8 (2018) e1338.
[44] J.P. Wojeicchowski, A.M. Ferreira, D.O. Abranches, M.R. Mafra, J.A.P. Coutinho, Using COSMO-RS in the design of deep eutectic solvents for the extraction of antioxidants from rosemary, ACS Sustain. Chem. Eng. 8 (2020) 12132–12141.
[45] G. Schaftenaar, J.H. Noordik, Molden: a pre- and post-processing program for molecular and electronic structures, J. Comput. Aided Mol. Des. 14 (2000) 123–134.
[46] Gaussian 09, Revision A.02, Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H, Li X, Caricato M, Marenich A, Bloino J, Janesko BG, Gomperts R, Mennucci B, Hratchian HP, Ortiz JV, Izmaylov AF, Sonnenberg JL, Williams-Young D, Ding F, Lipparini F, Egidi F, Goings J, Peng B, Petrone A, Henderson T, Ranasinghe D, Zakrzewski VG, Gao J, Rega N, Zheng G, Liang W, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Throssell K, Montgomery, Jr. JA, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Keith T, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Millam JM, Klene M, Adamo C, Cammi R, Ochterski JW, Martin RL, Morokuma K, Farkas O, Foresman JB, Fox DJ, Gaussian, Inc., Wallingford CT, 2016.
[47] I. Oprisiu, S. Novotarskyi, I.V. Tetko, Modeling of non-additive mixture properties using the Online CHEmical database and Modeling environment (OCHEM), J. Cheminf. 5 (2013) 4.
[48] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1967) 21–27.
[49] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[50] G. Huang, H.A. Babri, Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions, IEEE Trans. Neural Netw. 9 (1998) 224–229.
[51] A.M. Deris, A.M. Zain, R. Sallehuddin, Overview of support vector machine in modeling machining performances, Procedia Eng. 24 (2011) 308–312.
[52] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, et al., Machine learning for neuroimaging with scikit-learn, Front. Neuroinf. 8 (2014) 14.
[53] M. Stone, Cross-validatory choice and assessment of statistical predictions (with discussion), J. Roy. Stat. Soc.: Ser. B (Methodol.) 38 (1976) 102.
[54] A. Golbraikh, A. Tropsha, Beware of q2!, J. Mol. Graph. Model. 20 (2002) 269–276.
[55] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.
[56] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemom. Intel. Lab. Syst. 145 (2015) 22–29.
[57] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
[58] L.S. Shapley, A value for N-person games, in: H.W. Kuhn, A.W. Tucker (Eds.), Contributions to the Theory of Games, Annals of Mathematical Studies; Princeton University Press, 1953, pp. 307–317.
[59] R. Rodríguez-Pérez, J. Bajorath, Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values, J. Med. Chem. 63 (2020) 8761–8777.