Published in final edited form as:

Engineering (Beijing). 2024 August ; 39: 74–83. doi:10.1016/j.eng.2023.10.020.

Machine-Learning-Assisted Design of Deep Eutectic Solvents Based on Uncovered Hydrogen Bond Patterns

Usman L. Abbas[a], **Yuxuan Zhang**[b], **Joseph Tapia**[a], **Selim Md**[c], **Jin Chen**[c], **Jian Shi**[b], **Qing Shao**[a,*]

[a]Department of Chemical and Materials Engineering, University of Kentucky, Lexington, KY 40506, USA

[b]Department of Biosystems and Agricultural Engineering, University of Kentucky, Lexington, KY 40506, USA

[c]Institute for Biomedical Informatics, Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA

Abstract

Non-ionic deep eutectic solvents (DESs) are non-ionic designer solvents with various applications in catalysis, extraction, carbon capture, and pharmaceuticals. However, discovering new DES candidates is challenging due to a lack of efficient tools that accurately predict DES formation. The search for DES relies heavily on intuition or trial-and-error processes, leading to low success rates or missed opportunities. Recognizing that hydrogen bonds (HBs) play a central role in DES formation, we aim to identify HB features that distinguish DES from non-DES systems and use them to develop machine learning (ML) models to discover new DES systems. We first analyze the HB properties of 38 known DES and 111 known non-DES systems using their molecular dynamics (MD) simulation trajectories. The analysis reveals that DES systems have two unique features compared to non-DES systems: The DESs have ① more imbalance between the numbers of the two intra-component HBs and ② more and stronger inter-component HBs. Based on these results, we develop 30 ML models using ten algorithms and three types of HB-based descriptors. The model performance is first benchmarked using the average and minimal receiver operating characteristic (ROC)-area under the curve (AUC) values. We also analyze the importance of individual features in the models, and the results are consistent with the simulation-based statistical analysis. Finally, we validate the models using the experimental data of 34 systems. The extra trees forest model outperforms the other models in the validation, with an ROC-AUC of 0.88. Our work illustrates the importance of HBs in DES formation and shows the potential of ML in discovering new DESs.

[*]Corresponding author. qshao@uky.edu (Q. Shao).

Compliance with ethics guidelines
Usman L. Abbas, Yuxuan Zhang, Joseph Tapia, Selim Md, Jin Chen, Jian Shi, and Qing Shao declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2023.10.020.

Keywords

Machine learning; Deep eutectic solvents; Molecular dynamics simulations; Hydrogen bond; Molecular design

1. Introduction

Deep eutectic solvents (DESs) are liquid mixtures composed of hydrogen bond acceptors (HBAs) and donors (HBDs) that exhibit tunable properties [1–13]. DESs have gained attention as sustainable solvents in a number of applications, including carbon capture [2,14–16], pharmaceuticals [9,14,15,17–23], material synthesis [8,19,24], electrochemistry [9,14,25–37], decontamination [17,18], and extractions [6,8,24,38–41], due to their potential for recovery [42] and reuse [43]. Non-ionic DESs have several desirable properties—including biodegradability, high conductivity, low volatility, and low toxicity—as compared with conventional solvents [2,8,38,43,44]. Popularly classified as type V DESs [4,14–16,38], non-ionic DES can be made using natural compounds and exhibit low viscosity, making them particularly suitable for industrial applications such as liquid–liquid extraction and carbon nanomaterial production [12,38,45].

One of the main challenges in the field of DES is the discovery of a large collection of DES candidates, which would enable the community to have a vast pool to explore and to search for the ones with the desired properties. Numerous experimental and computational studies have shown the important role of hydrogen bonds (HBs) in the formation and properties of DESs [1,2,4,9,14,15,39,42,46]. Farias et al. [43] carried out an experimental study to understand the role of the HBDs of DESs in aqueous biphasic systems. They concluded that HBDs with high relative hydrophilicity mainly serve as adjuvants in biphasic systems, while HBDs with moderate hydrophilicity control the formation of biphasic systems, and HBDs with low hydrophilicity (high hydrophobicity) form aqueous biphasic systems, with the HBAs acting as adjuvants in such systems. Abranches et al. [1] investigated the suitability of betaine, a molecule with polarity imbalance, as a universal HBA in the formation of DESs. Their study used a combination of experiments and density functional theory calculations and concluded that betaine is a suitable choice for producing natural DESs due to its non-selective nature, low cost, and low toxicity. These fundamental studies highlight the important role of HBs in DES formation and properties, indicating that HB-based descriptors could serve as suitable inputs to discover new DESs.

Machine learning (ML) models are becoming increasingly popular for predicting the physicochemical and thermophysical properties of DESs [17,19,46–51]. A review by Hansen et al. [15] summarized studies that developed quantitative structure–property relationship models for predicting DES properties [6,15]. Halder et al. [51] used a cheminformatics approach to determine the structural attributes of DESs necessary for accurate predictions of densities in industrial applications. They utilized a consensus modeling approach and concluded that features such as the number of HBDs, lipophilicity, polarizability, and van der Waals surface area could be used to obtain highly accurate estimates of novel DES densities. Dietz et al. [6] used perturbed-chain statistical association fluid theory (PC-SAFT) modeling to predict the liquid–liquid equilibrium and solid–liquid equilibrium of mixtures of hydrophobic DES with water or hydroxymethyl furfural, demonstrating the efficacy of this approach for predicting the phase behavior of hydrophobic DES mixtures.

Other studies have employed ML algorithms to estimate the densities and viscosities of DESs. Abdollahzadeh et al. [19] compared seven ML algorithms and showed that least squares support vector regression had the highest accuracy in predicting the densities of 149 DESs, performing 74.5% better than the best results obtained via empirical correlations. Zamora et al. [16] compared the suitability of five ML algorithms, trained on experimental data, to predict the densities and viscosities of type V DESs. Their study concluded that support vector machines performed best at predicting densities, and Gaussian process regression models did best at predicting viscosities. Xu et al. [50] used gradient boosting models to predict DES viscosities; their model showed satisfactory results when trained and tested on experimental and simulation data. Overall, these studies demonstrate the potential of combining ML and molecular simulations to predict the properties of DESs.

In contrast to other studies that focus on predicting the properties of DESs, our work aims to use ML models to predict the formation of DES systems. Molecular dynamics (MD) simulation has emerged as a valuable technique for determining descriptors to be used as inputs for ML models [14,24,46]. We hypothesize that HB properties could serve as predictors for the formation of DESs. However, determining the relevant HB properties is not trivial. Our previous work [52] classified non-ionic DESs into three groups based on the ratio of intra- and inter-component HB numbers. These observations inspired us to explore the possibility of developing ML models using HB-based descriptors. To the best of our knowledge, our work is the first to use ML models to classify solvents as DES or non-DES.

In ML model training, data is a crucial element. To facilitate our research, we curated a library of 38 known DES and 111 non-DES systems from the literature. The construction of this library allows us to conduct statistical analysis on molecular simulation data, which can be used to develop training and testing datasets for model development. We curated a separate library of 34 systems to validate our model performance. Given the size of our database, this paper focuses on traditional ML algorithms. We utilized ten ML algorithms; however, we acknowledge that deep learning algorithms have emerged as a promising technique for designing materials. One obstacle to using deep learning algorithms for predicting the formation of DESs is the relative sparsity of experimentally verified DESs in the literature. Models such as the one we developed could help speed up the discovery of novel DESs by generating solvent mixtures likely to form DESs. The rest of this paper is structured as follows: Section 2 provides details on the computational methods, Section 3 presents the results and discussion, and Section 4 gives our conclusions.

2. Methodology

2.1. Library of DES and non-DES systems

Tables S1–S8 in Appendix A provide detailed information on the 183 systems that were simulated in this study. Of these systems, 38 are identified as known DES and 111 are known non-DES systems, as reported in the literature. These comprised our training and testing set. Additionally, 34 experimentally verified systems (17 DES and 17 non-DES) were reserved for validation. Classification of the DES and non-DES systems is based on the experimental results of van Osch et al. [53,54], with only the non-ionic DESs from their list being considered. DESs that lacked all three types of HBs (A–A, B–B, and A–B) were excluded from this study. The compounds used in the simulation are represented using three-letter abbreviations, such as “DEA” for “decanoic acid.” We follow the naming conventions from van Osch et al.’s work, in which component A in a system A–B is the expected HBD, and compound B is the expected HBA. The systems are denoted by the three-letter abbreviations of their compounds and the corresponding molar ratio; for example, DEA–MEN11 represents a 1:1 mixture of decanoic acid and menthol. Tables S9–S11 in Appendix A list the abbreviations used for chemical compounds in this study.

2.2. Molecular simulations

2.2.1. Molecular models—The all-atom optimized potentials for liquid simulations (OPLS-AA/M) force field [55] was used to describe the molecules in this study. The nonbonded and bonded parameters in the systems were determined based on the OPLS-AA/M force field, due to its proven ability to accurately model the behavior of organic molecules. The force field parameters were generated using the LigParGen [56] web server.

2.2.2. Simulation detail—The simulation systems were created by randomly inserting specific numbers (based on the molar ratio) of the chosen organic molecules in a cubic box. Fig. 1 shows a snapshot of the THY–MEN11 system generated using visual molecular dynamics (VMD) [57].

For each system, the simulation process comprised three steps: ① an energy minimization to remove any atomic overlaps; ② a 50 ns isobaric–isothermal (*NPT*, where *N* is the number of particles, *P* is pressure (*P* = 1 atm, 1 atm = 101.325 kPa), and *T* is the temperature (*T* = 295 K)) ensemble MD simulation to enable the system to reach thermodynamic equilibrium; and ③ a 10 ns canonical (*NVT*, where *V* is the system’s volume and *T* = 295 K) ensemble MD simulation to collect the data at a frequency of 10 ps. In step ②, the MD simulation used the Berendsen et al.’s method [58] to control the system pressure, while the velocity rescaling [59] method was used to control the system temperature.

The short- and long-range nonbonded interactions in the OPLS-AA/M force field were calculated using the Lennard–Jones 12–6 and Coulomb potential (*E*), respectively, using Eq. (1).

$$E = \sum_{i} \sum_{j < i} \left\{ \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j e^2}{r_{ij}} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \tag{1}$$

where r_{ij} is the distance between atoms i and j; q_i and q_j are the partial charges of atoms i and j, respectively; ε_0 is the free space permittivity; and ε_{ij} and σ_{ij} are energetic and geometric parameters, respectively. The particle mesh Ewald (PME) [60] sum was used to calculate long-range potentials, and the linear constraint solver (LINCS) algorithm [61] was used to constrain bonds involving hydrogen atoms. All energy minimization and MD simulations were conducted using GROMACS 2021.2 [62].

2.3. Hydrogen bond analysis

We characterized the HBs using the criteria developed by Luzar and Chandler [63]: ① the distance between the O(donor) and O(acceptor) is ≤0.35 nm; and ② the O(acceptor)–H(donor)–O(donor) angle is ≤30°. We calculated the HB lifetime in two steps:

(1) Calculate the correlation function $C(t)$, as shown in Eq. (2):

$$C(t) = \frac{\langle N_{\mathrm{HB}}(t)\rangle}{\langle N_{\mathrm{HB}}(0)\rangle} \tag{2}$$

where $\langle N_{\mathrm{HB}}(0)\rangle$ is the ensemble average of the number of HBs at the initial status, and $\langle N_{\mathrm{HB}}(t)\rangle$ is the ensemble average of the number of HBs still existing at time t. The HBs are counted even if they break intermittently, based on Rappaport’s definition [64].

(2) Calculate the lifetime $\boldsymbol{\tau}$ by numerically integrating the $C(t)$ curves.

2.4. Machine learning models

The literature-based library introduced in Section 2.1 contains more non-DES than DES systems, a data imbalance that may cause bias in model training. To attenuate this potential source of artificial effect, we curated a database containing 38 DES and 38 non-DES systems dedicated for each round of training during the ML model development. The 38 non-DES systems were selected randomly from the original 111 non-DES systems in the library. We further split this database into a training set consisting of 30 DES and 30 non-DES systems and a testing set consisting of eight DES and eight non-DES systems. We used fixed seeds when sampling from the DES and non-DES sets to ensure that all models were evaluated on the same dataset slices. All models were further validated with experimentally verified DES and non-DES systems, as described in Section 3.

We trained ten distinct ML algorithms utilizing algorithm implementations provided by the scikit-learn [65,66] and XGBoost [67] packages. These algorithms were: ① logistic regression, ② decision tree, ③ gradient boost, ④ AdaBoost, ⑤ random forest, ⑥ extra trees forest, ⑦ support vector machine, ⑧ k-nearest neighbors, ⑨ XGBoost, and ⑩ XGBoost-random forest. Hyperparameter optimization was performed using scikit-learn’s grid search method. Each model’s performance was measured via repeated k-fold cross-validation with six folds and ten repeats, using the receiver operating characteristic (ROC)-area under the curve (AUC) metric. The model with the highest ROC-AUC value during optimization was considered to be the best model. For each ML algorithm, subsequent training and testing were only conducted on the best-trained model.

HBs play a determining role in the formation of DESs. To obtain a full picture of the HB environment, it is imperative to know how many of the molecules in a system interact to form HBs (i.e., the HB number) and how long these HBs last (i.e., the HB lifetime). All ML algorithms consider three types of input features: ① HB numbers alone, ② HB lifetimes alone, and ③ HB numbers combined with lifetimes. The input features, generated from MD simulations, are shown in Tables S1–S8. A total of 30 models were trained in this study; the model hyperparameters are detailed in Tables S12–S14 in Appendix A.

The following Python packages were used to conduct the work presented in this study: Python (version 3.10.8), scikit-learn [66] (version 1.2.0), pandas [68] (version 1.5.2), NumPy [69] (version 1.22.3), matplotlib [70] (version 3.6.2), SciPy [71] (version 1.7.3), and XGBoost [67] (version 1.7.3). All ML work was executed on an 8th Gen Intel core i7–8750H processor.

2.5. Experiment

To validate the trained models, we used a list of solvent formulas from our previous study [72] to determine whether the formulas could form DESs or not. To prepare the systems, two components were mixed at a specific molar ratio, with heating and constant stirring to ensure complete mixing. More specifically, the required mass of each component was first calculated based on the molar ratio and sequentially weighed out into a glass bottle on an analytic balance (VWR-224AC, VWR International, USA). The compounds were pre-mixed using a glass rod, and a magnetic stir bar was subsequently added to the bottle. The bottle was then sealed and placed in an oil bath for heating. The temperature was generally maintained at 80 °C, with constant stirring at 500 r·min^{-1} for 1 h on a magnetic stirrer hotplate (Hei-Tec, Heidolph Instruments, Germany). For combinations that did not form a homogeneous liquid at this temperature, higher temperatures of 100 and 120 °C were further applied to check whether these combinations could transform into the liquid state at elevated temperatures. After the heating process, the mixture was air-cooled to room temperature and kept in a desiccator for 24 h. Samples that remained in a liquid form with no crystals within the 24 h period were considered to be DES candidates. However, we observed that some systems initially exhibited DES-like behavior but eventually formed a solid phase after several days. These systems were excluded from this study. Finally, 17 DES and 17 non-DES systems were selected.

3. Results and discussion

3.1. Statistical analysis of hydrogen bond features

3.1.1. Hydrogen bond number features—We first analyzed the probability density distribution of actual Inter- and intra-component HB numbers for DES and non-DES systems. Fig. 2 shows the distribution of the 38 DES and 111 non-DES systems based on their average inter- and intra-component HB numbers. The pattern distributions in Fig. 2 do not show distinct differences. As depicted in Fig. 2(a), the intra-component HB numbers (A–A and B–B) for the DESs skew to the left, indicating that most of the DESs in our dataset have average HB numbers that are less than 20. The B–B HB numbers are concentrated on the lower end of the spectrum compared with the A–A HBs. The inter-component HB numbers skew to the right, suggesting that most of the DESs have higher inter-component HB numbers compared with intra-component HB numbers. In Fig.

2(b), the intra-component HB numbers for the non-DESs also skew to the left. In addition, most of the inter-component HB numbers are skewed to the right. Thus, based on an analysis of Fig. 2, the actual number of intra- and inter-component HBs may not be a suitable HB feature to differentiate between DES and non-DES systems.

Distinct patterns emerge when plotting the average inter- and intra-component HB numbers as boxplots for DES and non-DES systems. As shown in Figs. 3(a) and (b), the DES systems present a large difference between the median values for the two intra-component HB numbers (A–A vs B–B) compared with the non-DES systems. In addition, the inter-component HBs in the DES systems exhibit a median value of 56.07, 6%–83% higher than the median values of A–A and B–B. For the non-DES systems, the A–B HBs only present a median value of 48.36, 55%–59% higher than those of A–A and B–B, respectively. In both the DES and non-DES systems, even when the intra-component HB numbers (A–A and B–B) are summed up, the inter-component HB number (A–B) is still greater. Such differences in the median imply that the ratio of the two intra-component and the inter-/intra-component HBs may serve as important features for DES and non-DES system classification.

The plot of A–A/B–B and A–B/(A–A + B–B) in Fig. 3(c) further confirms our hypothesis. The ratio of inter- to intra-component HB numbers is well above 1.5 for some DESs. On average, the inter-component HB numbers are 35% greater than the total intra-component HB numbers for DESs. Finally, we looked at the ratio of the intra-component bonds to obtain more insight into the magnitudes of their differences. On average, the ratio of A–A to B–B HBs is 8.01 for DESs. For non-DESs, the ratio of A–A to B–B HBs falls to 3.44. The average intra-component HB numbers (A–A and B–B) for the non-DESs are roughly the same (24.31 and 24.08, respectively). The median HB numbers for A–A and B–B are also similar in non-DESs, at 20.01 and 21.41, respectively. This finding suggests that there is no dominant intra-component HB in non-DESs. This is also shown in Fig. 3(c), in which most of the intra-component HB number ratios are clustered around 1.0 for non-DESs, with few outliers. For non-DESs, the average inter-component (A–B) HB numbers are close to twice (1.93–1.95 times) those of the average intra-component (A–A and B–B) HBs, respectively. Relative to DESs, the ratios of intra-component HBs in non-DESs are also smaller. For example, in the 25th, 50th, and 75th percentiles, the DESs display A–A/B–B of 1.13, 1.91, and 15.51 compared with 0.15, 0.49, and 0.99 for the non-DESs. Tables S15 and S16 in Appendix A provide more details.

3.1.2. Hydrogen bond lifetimes—

We also analyzed the probability density distribution of the inter- and intra-component HB lifetimes of the 38 DES and 111 non-DES systems. Across the bins, we observed two distinct scenarios for DESs: ① dominant inter-component (A–B) HBs; and ② dominant intra-component HBs (A–A or B–B). This finding agrees with our previous work, in which we classified several known DESs into inter- or intra-dominant groups. In Fig. 4(a), one of the intra-component HB lifetime bonds (A–A) is concentrated at 2.0–4.0 ns, while the B–B lifetime is concentrated at 0.25–2.50 ns for DESs. The inter-component HB lifetimes (A–B) appear to skew to the right and to last longer than the intra-component HB lifetimes.

Fig. 4(b) shows that one of the intra-component HB lifetimes for non-DESs dominates in different bins, but there is no clear trend; for example, B–B dominates at lifetimes less than 1.25 ns, but A–A dominates at lifetimes greater than 3.00 ns. In each bin, the A–B lifetimes appear to be more dominant than one of the intra-component HB lifetimes, while being similar to the other intra-component HB lifetimes. The lack of a clear pattern means that actual intra- and inter-component HB lifetime features alone might not be enough to differentiate between DES and non-DES systems.

Some differences emerge when we plot the intra- and inter-component HB lifetime distributions as boxplots. As shown in Fig. 5(a), the DESs present a small difference between the median values for the inter-component (A–B) and one of the intra-component (A–A) lifetimes; the difference is wider between the median values of A–B and the other intra-component (B–B) lifetimes. The A–B lifetimes have a median of 2.67, 14% and 39% greater than those of the A–A and B–B lifetimes, respectively. As shown in Fig. 5(b), the non-DESs present a smaller difference between the median values for the inter- and intra-component HB lifetime values. The A–B lifetimes have a median of 2.72, which is only 3.6% and 14.0% greater than those of the A–A and B–B lifetimes, respectively. These differences indicate that the ratios of inter- to intra-component HB lifetimes could be more useful as features than the actual lifetimes.

The plot of A–A/B–B versus A–B/(A–A + B–B) in Fig. 5(c) confirms this hypothesis. The A–A median lifetimes last about 7% longer than the B–B lifetimes in DESs, compared with 13% for non-DESs. Even though there are more inter-component HBs than intra-component HBs, the intra-component HBs last longer. The median value of A–B/(A–A + B–B) lifetimes is 0.63 for DESs and 0.53 for non-DESs. The ratio of inter- to intra-component HB lifetimes in DESs varies from 0.5 to 2.0 while most of the non-DESs have ratios of inter- to intra-component HB lifetimes clustered around 0.5. Similar to the HB numbers, ratios of the HB lifetimes might be more useful as features than the actual lifetime values.

3.2. Model development

We trained 30 models with ten algorithms (logistic regression, decision tree, gradient boost, AdaBoost, random forest, extra trees forest, support vector machine, *k*-nearest neighbors, XGBoost, and XGBoost-random forest) and three types of input features (HB number, HB lifetime, and a combination of HB number and lifetime features) to predict whether a system could be a DES. For each type of input feature, we used the five variables mentioned in Section 3.1 (A–A, B–B, A–B, A–A/B–B, and A–B/(A–A + B–B)). We trained each model for 100 rounds and calculated the average ROC-AUC values from each of the 100 rounds. The ROC is a probability curve, and the AUC represents the degree or measure of separability. The ROC-AUC shows how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting DES classes to be DESs and non-DES classes to be non-DESs. For each round, we randomly sampled 38 (30 for training, eight for testing) entries each from the DES and non-DES datasets. In each round, a six-fold grid search cross-validation was used for hyperparameter optimization, with the ROC-AUC as a metric. To ensure a fair comparison, each model was trained and tested with the same samples from the DES and non-DES datasets. Figs. S1–S11 in Appendix A show the variation in the ROC-AUC for each model during training.

We ranked the models using two criteria: ① average ROC-AUC score (Table 1), and ② minimum ROC-AUC score (Table 2).

With an average ROC-AUC score of 0.70, the AdaBoost and extra trees forest classifiers were tied for the best performing models when trained with HB lifetime features. When trained with HB number features, XGBoost-random forest, random forest, and XGBoost were the top-performing models, with an average ROC-AUC of 0.82, 0.81, and 0.81, respectively. When HB number and lifetime features were combined, the top-performing models were the random forest and XGBoost-random forest classifiers, with both having an average ROC-AUC of 0.79. Overall, the top-performing models were the XGBoost-random forest and extra trees forest, based on the average and minimum ROC-AUC values, respectively.

The minimum ROC-AUC score in 100 training rounds could also be used to evaluate the performance of a model. Table 2 lists the minimum ROC-AUC scores for the 30 models. The minimum ROC-AUC scores trained with HB lifetime features alone range from 0.15 to 0.30, lower than those trained with HB number or the number and lifetime. Such observations indicate that the HB lifetime alone might not be sufficient to develop an ML model for classifying DES systems. Across all categories, the extra trees forest classifier had the highest minimum ROC-AUC score of 0.70 when trained with HB number.

Some algorithms were among the top performers regardless of the criteria used for model selection. For models trained with HB numbers, the top-performing model was the extra trees forest, based on the minimum ROC-AUC, and this model was only slightly behind the XGBoost-random forest when judged by the average ROC-AUC score. For models trained with HB lifetime features, the extra trees forest and the AdaBoost were the top performers using either the average ROC-AUC scores or the highest minimum ROC-AUC scores. Among the models trained with the combined HB number and lifetime features, the extra trees forest classifier was the top performer using the highest minimum ROC-AUC score or the average ROC-AUC score. However, it should be noted that the stellar performance observed during training does not necessarily translate into excellence in the validation stage, as will be seen in the next section.

3.3. Model validation with experimental results

We validated the 30 trained models using 34 experimental results (17 DES and 17 non-DES systems). The results are presented in Table 3.

For models trained with the HB lifetime features, the XGBoost-random forest, logistic regression, and extra trees forest were the top performers, with ROC-AUC values of 0.68, 0.65, and 0.65, respectively, during validation. Support vector machine, extra trees forest, and gradient boost were the top performers, with ROC-AUC values of 0.80, 0.79, and 0.77, respectively, when the models were trained with the HB number features. Extra trees forest, logistic regression, and gradient boost were the top-performing models, with ROC-AUC values of 0.88, 0.84, and 0.81, respectively, among the models trained with both HB number and lifetime. In general, the ensemble algorithms (bagging and boosting) were observed to perform well. Bagging algorithms such as random forest, extra trees forest, and decision tree build and train independent estimators, and then average the independent predictions of these estimators to make a final prediction. This can help reduce variance in predictions and increase accuracy. The boosting algorithms (XGBoost, XGBoost-random forest, AdaBoost, and gradient boost), on the other hand, train several estimators sequentially. Each estimator focuses on reducing the errors of the previous estimator, and this typically reduces bias.

Fig. 6 presents confusion matrices for the top-performing models under each of the three input feature categories during validation. The confusion matrices present true positives, true negatives, false positives, and false negatives for each model's predictions. In this study, DESs are positives, while non-DESs are negatives. The sensitivity measures how many DESs were correctly predicted to be DESs, while the specificity measures how many non-DESs were correctly predicted to be non-DESs by a model. Some models were better at predicting DESs (high sensitivity), while some were better at predicting non-DESs (high specificity).

XGBoost-random forest was the top-performing algorithm among the models trained with HB lifetime features. It performed best at predicting which systems were DESs, as shown by its high sensitivity of 0.82 (Fig. 6(a)), but it was not good at predicting which systems were non-DESs (with a low specificity of 0.47, Fig. 6(a)). Among the models trained with HB number features, the support vector machine was the top-performing algorithm. It had a specificity of 0.88 (Fig. 6(c)), which means that it performed best at predicting which systems were non-DESs. Its low sensitivity of 0.35 (Fig. 6(c)) means that it was not good at predicting DESs. When models were trained with combined HB lifetime and number features as inputs, the extra trees forest model performed best. It had a sensitivity of 0.76 (Fig. 6(e)), indicating it was among the top performers at predicting which systems were DESs. It had a specificity of 0.94 (Fig. 6(e)), indicating it was the top performer at predicting which systems were non-DESs. Relative to the top-performing models in other input feature categories, the extra trees forest algorithm was the best overall at predicting DESs and non-DESs. The confusion matrices for all models are shown in Figs. S18–S20 in Appendix A.

3.4. Prediction probabilities

'Prediction probabilities are useful indicators of how well each model separates DESs and non-DESs. A model with good separation capability would have all its non-DES predictions with the probability of being DES < 0.5 and as close to 0 as possible, and its DES predictions with the probability of being DES > 0.5 and as close to 1 as possible. Fig. 7 presents the distribution of prediction probabilities for the best models during validation. It can be seen from the probabilities in Fig. 7(a) that the predictions of the XGBoost-random forest are closely distributed around 0.49 to 0.51, suggesting that there is not much separation for models trained with HB lifetime features. Notably, all of the XGBoost-random forest's 14 DES predictions made with confidence > 0.5 were correct. The separation improves in Fig. 7(b), with the probabilities distributed around 0.46 to

0.54, suggesting that HB number features helped the models detect non-DESs relatively better than HB lifetimes alone. This is backed up by the observation that all 15 non-DES predictions made by the support vector machine model with the probability of being DES < 0.5 were correct. The probabilities are distributed between 0.30 and 0.70 in Fig. 7(c), indicating better confidence in the extra trees forest model's predictions when HB number and lifetime features were combined as inputs. The extra trees forest model shows better separation in its classifications and is relatively more confident in its non-DES predictions, and this is backed up by its specificity of 0.94 (Fig. 6(e)). It got only one non-DES prediction wrong. The prediction probabilities for all the other models are shown in Figs. S21–S23 in Appendix A.

It is useful to have some insight into which input features carry the most weight when the ML models make predictions. Fig. 8 shows how the models ranked the importance of input features. Models that were trained with HB lifetime features alone over-whelmingly ranked the ratio of inter- to intra-component HB lifetime as the most important feature for predictions, followed by the inter-component HB lifetime. When trained with HB numbers features alone, the models ranked the inter-component HB numbers as the most important feature; however, it should be noted that the ratio of inter- to intra-component HB numbers was not far behind in second place. When number and lifetime were combined, the trained models ranked the inter-component HB numbers as the most important feature, closely followed by the ratio of inter- to intra-component HB lifetimes.

4. Conclusions

We analyzed the HB features of 38 known DES and 111 known non-DES systems using MD simulation trajectories. The statistical analysis of inter- and intra-component HB numbers and lifetimes revealed two unique features for DES systems in comparison with non-DES systems: The DESs exhibited an imbalance between the two intra-component HB numbers, and more and stronger inter-component HBs. We then developed 30 ML models by training ten algorithms on three types of input features and validated the models using 17 DES and 17 non-DES systems that had been experimentally verified. Using the two criteria of highest average and highest minimum ROC-AUC scores, we found the logistic regression, gradient boost, support vector machine, and extra trees forest models among the top performers when trained using the HB lifetime, number, and combined lifetime and number features. When testing against the experimental validation, the extra trees forest classifier was the top-performing model overall, with an ROC-AUC of 0.88 with the HB number and lifetime combined as inputs. Intuitively, it makes sense that models would perform better when fed information about the population of HB numbers and how long those HBs last. All models ranked the inter-component and the ratio of inter- to intra-component HB number and lifetime as the most important features for classifying a system as a DES or not.

DESs are promising solvents that hold huge potential. Due to the sheer size of the candidate pool, it is important to have models that can accurately predict which compounds will or will not form DESs when mixed. The purpose of the ML models developed in this work was to determine whether a binary system could be a DES based on MD simulation data. These ML models could assist in DES research by accelerating the discovery of new DES candidates.

Our work sheds light on which compounds are likely to form DESs but does not suggest what their physicochemical properties are likely to be. In the future, more work needs to be done to predict which compounds will form DESs with application-specific properties.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by Ignite Research Collaborations (IRC), Startup funds, and the UK Artificial Intelligence (AI) in Medicine Research Alliance Pilot (NCATS UL1TR001998 and NCI P30 CA177558), University of Kentucky Center for Computational Sciences and Information Technology Services Research Computing for the use of the Lipscomb Compute Cluster of the University of Kentucky.

References

[1]. Abranches DO, Silva LP, Martins MAR, Pinho SP, Coutinho JAP. Understanding the formation of deep eutectic solvents: betaine as a universal hydrogen bond acceptor. ChemSusChem 2020;13(18):4916–21. [PubMed: 32672893]

[2]. Stephens NM, Smith EA. Structure of deep eutectic solvents (DESs): what we know, what we want to know, and why we need to know it. Langmuir 2022;38 (46):14017–24. [PubMed: 36346803]

[3]. Celebi AT, Dawass N, Moultos OA, Vlugt TJH. How sensitive are physical properties of choline chloride–urea mixtures to composition changes: molecular dynamics simulations and Kirkwood–Buff theory. J Chem Phys 2021;154(18):184502. [PubMed: 34241035]

[4]. Abranches DO, Coutinho JAP. Type V deep eutectic solvents: design and applications. Curr Opin Green Sustain Chem 2022;35:100612.

[5]. Alcalde R, Gutiérrez A, Atilhan M, Aparicio S. An experimental and theoretical investigation of the physicochemical properties on choline chloride—lactic acid based natural deep eutectic solvent (NADES). J Mol Liq 2019;290:110916.

[6]. Dietz CHJT, Erve A, Kroon MC, van Sint AM, Gallucci F, Held C. Thermodynamic properties of hydrophobic deep eutectic solvents and solubility of water and HMF in them: measurements and PC-SAFT modeling. Fluid Phase Equilib 2019;489:75–82.

[7]. Florindo C, Branco LC, Marrucho IM. Development of hydrophobic deep eutectic solvents for extraction of pesticides from aqueous environments. Fluid Phase Equilib 2017;448:135–42.

[8]. Kivelä H, Salomäki M, Vainikka P, Mäkilä E, Poletti F, Ruggeri S, et al. Effect of water on a hydrophobic deep eutectic solvent. J Phys Chem B 2022;126 (2):513–27. [PubMed: 35001628]

[9]. Kovács A, Neyts EC, Cornet I, Wijnants M, Billen P. Modeling the physicochemical properties of natural deep eutectic solvents. ChemSusChem 2020;13(15):3789–804. [PubMed: 32378359]

[10]. Křížek T, Bursová M, Horsley R, Kuchař M, Tůma P,Čabala R, et al. Menthol-based hydrophobic deep eutectic solvents: towards greener and efficient extraction of phytocannabinoids. J Clean Prod 2018;193:391–6.

[11]. Li K, Jin Y, Jung D, Park K, Kim H, Lee J. *In situ* formation of thymol-based hydrophobic deep eutectic solvents: application to antibiotics analysis in surface water based on liquid–liquid microextraction followed by liquid chromatography. J Chromatogr A 2020;1614:460730. [PubMed: 31812273]

[12]. Lukaczynska-Anderson M, Mamme MH, Ceglia A, Van den Bergh K, De Strycker J, De Proft F, et al. The role of hydrogen bond donor and water content on the electrochemical reduction of Ni^{2+} from solvents—an experimental and modelling study. Phys Chem Chem Phys 2020;22(28):16125–35. [PubMed: 32638784]

[13]. Martins MAR, Silva LP, Schaeffer N, Abranches DO, Maximo GJ, Pinho SP, et al. Greener terpene–terpene eutectic mixtures as hydrophobic solvents. ACS Sustain Chem Eng 2019;7(20):17414–23.

[14]. Tolmachev D, Lukasheva N, Ramazanov R, Nazarychev V, Borzdun N, Volgin I, et al. Computer simulations of deep eutectic solvents: challenges, solutions, and perspectives. Int J Mol Sci 2022;23(2):645. [PubMed: 35054840]

[15]. Hansen BB, Spittle S, Chen B, Poe D, Zhang Y, Klein JM, et al. Deep eutectic solvents: a review of fundamentals and applications. Chem Rev 2021;121 (3):1232–85. [PubMed: 33315380]

[16]. Zamora L, Benito C, Gutiérrez A, Alcalde R, Alomari N, Bodour AA, et al. Nanostructuring and macroscopic behavior of type V deep eutectic solvents based on monoterpenoids. Phys Chem Chem Phys 2021;24(1):512–31. [PubMed: 34904590]

[17]. Bergua F, Castro M, Lafuente C, Artal M. Thymol + *L*-menthol eutectic mixtures: thermophysical properties and possible applications as decontaminants. J Mol Liq 2022;368(Pt B):120789.

[18]. Bergua F, Castro M, Muñoz-Embid J, Lafuente C, Artal M. *L*-Menthol-based eutectic solvents: characterization and application in the removal of drugs from water. J Mol Liq 2022;352:118754.

[19]. Abdollahzadeh M, Khosravi M, Hajipour Khire Masjidi B, Samimi Behbahan A, Bagherzadeh A, Shahkar A, et al. Estimating the density of deep eutectic solvents applying supervised machine learning techniques. Sci Rep 2022;12 (1):4954. [PubMed: 35322084]

[20]. Dai Y, Witkamp GJ, Verpoorte R, Choi YH. Tailoring properties of natural deep eutectic solvents with water to facilitate their applications. Food Chem 2015;187:14–9. [PubMed: 25976992]

[21]. Gutiérrez A, Aparicio S, Atilhan M. Design of arginine-based therapeutic deep eutectic solvents as drug solubilization vehicles for active pharmaceutical ingredients. Phys Chem Chem Phys 2019;21(20):10621–34. [PubMed: 31080981]

[22]. Gutiérrez A, Atilhan M, Aparicio S. A theoretical study on lidocaine solubility in deep eutectic solvents. Phys Chem Chem Phys 2018;20(43):27464–73. [PubMed: 30357182]

[23]. Zainal-Abidin MH, Hayyan M, Ngoh GC, Wong WF, Looi CY. Emerging frontiers of deep eutectic solvents in drug discovery and drug delivery systems. J Control Release 2019;316:168–95. [PubMed: 31669211]

[24]. Zhong X, Velez C, Acevedo O. Partial charges optimized by genetic algorithms for deep eutectic solvent simulations. J Chem Theory Comput 2021;17 (5):3078–87. [PubMed: 33885293]

[25]. Chaabene N, Ngo K, Turmine M, Vivier V. New hydrophobic deep eutectic solvent for electrochemical applications. J Mol Liq 2020;319:114198.

[26]. Hanada T, Goto M. Synergistic deep eutectic solvents for lithium extraction. ACS Sustain Chem Eng 2021;9(5):2152–60.

[27]. Yurramendi L, Hidalgo J, Siriwardana A. A sustainable process for the recovery of valuable metals from spent lithium ion batteries by deep eutectic solvents leaching. Mater Proc 2021;5(1):100.

[28]. Du K, Ang EH, Wu X, Liu Y. Progresses in sustainable recycling technology of spent lithium-ion batteries. Energy Environ Mater 2022;5(4):1012–36.

[29]. Neumann J, Petranikova M, Meeus M, Gamarra JD, Younesi R, Winter M, et al. Recycling of lithium-ion batteries—current state of the art, circular economy, and next generation recycling. Adv Energy Mater 2022;12(17):2102917.

[30]. Tang S, Zhang M, Guo M. A novel deep-eutectic solvent with strong coordination ability and low viscosity for efficient extraction of valuable metals from spent lithium-ion batteries. ACS Sustain Chem Eng 2022;10 (2):975–85.

[31]. Zhang J, Wenzel M, Steup J, Schaper G, Hennersdorf F, Du H, et al. 4-Phosphoryl pyrazolones for highly selective lithium separation from alkali metal ions. Chemistry 2022;28(1):e202103640. [PubMed: 34652866]

[32]. Chen Y, Wang Y, Bai Y, Duan Y, Zhang B, Liu C, et al. Significant improvement in dissolving lithium-ion battery cathodes using novel deep eutectic solvents at low temperature. ACS Sustain Chem Eng 2021;9(38):12940–8.

[33]. Wang K, Hu T, Shi P, Min Y, Wu J, Xu Q. Efficient recovery of value metals from spent lithium-ion batteries by combining deep eutectic solvents and coextraction. ACS Sustain Chem Eng 2022;10(3):1149–59.

[34]. Zante G, Boltoeva M. Review on hydrometallurgical recovery of metals with deep eutectic solvents. Sustain Chem 2020;1(3):238–55.

[35]. Chen L, Chao Y, Li X, Zhou G, Lu Q, Hua M, et al. Engineering a tandem leaching system for the highly selective recycling of valuable metals from spent Li-ion batteries. Green Chem 2021;23(5):2177–84.

[36]. Tran MK, Rodrigues MTF, Kato K, Babu G, Ajayan PM. Deep eutectic solvents for cathode recycling of Li-ion batteries. Nat Energy 2019;4(4):339–45.

[37]. Wang S, Zhang Z, Lu Z, Xu Z. A novel method for screening deep eutectic solvent to recycle the cathode of Li-ion batteries. Green Chem 2020;22 (14):4473–82.

[38]. Aguilar N, Barros R, Antonio Tamayo-Ramos J, Martel S, Bol A, Atilhan M, et al. Carbon nanomaterials with thymol + menthol type V natural deep eutectic solvent: from surface properties to nano-Venturi effect through nanopores. J Mol Liq 2022;368:120637.

[39]. Tiecco M, Cappellini F, Nicoletti F, Del Giacco T, Germani R, Di Profio P. Role of the hydrogen bond donor component for a proper development of novel hydrophobic deep eutectic solvents. J Mol Liq 2019;281:423–30.

[40]. Zainal-Abidin MH, Hayyan M, Wong WF. Hydrophobic deep eutectic solvents: current progress and future directions. J Ind Eng Chem 2021;97:142–62.

[41]. Paul R, Mitra A, Paul S. Phase separation property of a hydrophobic deep eutectic solvent–water binary mixture: a molecular dynamics simulation study. J Chem Phys 2021;154(24):244504. [PubMed: 34241334]

[42]. Makoś P, Słupek E, Gębicki J. Extractive detoxification of feedstocks for the production of biofuels using new hydrophobic deep eutectic solvents—experimental and theoretical studies. J Mol Liq 2020;308:113101.

[43]. Farias FO, Pereira JFB, Coutinho JAP, Igarashi-Mafra L, Mafra MR. Understanding the role of the hydrogen bond donor of the deep eutectic solvents in the formation of the aqueous biphasic systems. Fluid Phase Equilib 2020;503:112319.

[44]. Vainikka P, Thallmair S, Souza PCT, Marrink SJ. Martini 3 coarse-grained model for type III deep eutectic solvents: thermodynamic, structural, and extraction properties. ACS Sustain Chem Eng 2021;9(51):17338–50.

[45]. Atilhan M, Aparicio S. Molecular dynamics simulations of mixed deep eutectic solvents and their interaction with nanomaterials. J Mol Liq 2019;283:147–54.

[46]. Alkhatib III, Bahamon D, Llovell F, Abu-Zahra MRM, Vega LF. Perspectives and guidelines on thermodynamic modelling of deep eutectic solvents. J Mol Liq 2020;298:112183.

[47]. Adeyemi I, Abu-Zahra MRM, AlNashef IM. Physicochemical properties of alkanolamine-choline chloride deep eutectic solvents: measurements, group contribution and artificial intelligence prediction techniques. J Mol Liq 2018;256:581–90.

[48]. Shahbaz K, Bagh FSG, Mjalli FS, AlNashef IM, Hashim MA. Prediction of refractive index and density of deep eutectic solvents using atomic contributions. Fluid Phase Equilib 2013;354:304–11.

[49]. Bagh FSG, Shahbaz K, Mjalli FS, AlNashef IM, Hashim MA. Electrical conductivity of ammonium and phosphonium based deep eutectic solvents: measurements and artificial intelligence-based prediction. Fluid Phase Equilib 2013;356:30–7.

[50]. Xu X, Range J, Gygli G, Pleiss J. Analysis of thermophysical properties of deep eutectic solvents by data integration. J Chem Eng Data 2020;65(3):1172–9.

[51]. Halder AK, Haghbakhsh R, Voroshylova IV, Duarte ARC, Cordeiro MNDS. Density of deep eutectic solvents: the path forward cheminformatics-driven reliable predictions for mixtures. Molecules 2021;26(19):5779. [PubMed: 34641322]

[52]. Abbas UL, Qiao Q, Nguyen MT, Shi J, Shao Q. Molecular dynamics simulations of heterogeneous hydrogen bond environment in hydrophobic deep eutectic solvents. AIChE J 2022;68:e17382.

[53]. van Osch DJGP, Dietz CHJT, Warrag SEE, Kroon MC. The curious case of hydrophobic deep eutectic solvents: a story on the discovery, design, and applications. ACS Sustain Chem Eng 2020;8(29):10591–612.

[54]. van Osch DJGP, Dietz CHJT, van Spronsen J, Kroon MC, Gallucci F, van Sint AM, et al. A search for natural hydrophobic deep eutectic solvents based on natural components. ACS Sustain Chem Eng 2019;7(3):2933–42.

[55]. Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved peptide and protein torsional energetics with the OPLS-AA force field. J Chem Theory Comput 2015;11(7):3499–509. [PubMed: 26190950]

[56]. Dodda LS, Cabeza de Vaca I, Tirado-Rives J, Jorgensen WL. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. Nucleic Acids Res 2017;45(W1):W331–6. [PubMed: 28444340]

[57]. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14(1):33–8. [PubMed: 8744570]

[58]. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81(8):3684–90.

[59]. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys 2007;126(1):014101. [PubMed: 17212484]

[60]. Darden T, York D, Pedersen L. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J Chem Phys 1993;98(12):10089–92.

[61]. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comput Chem 1997;18(12):1463–72.

[62]. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1–2:19–25.

[63]. Luzar A, Chandler D. Hydrogen-bond kinetics in liquid water. Nature 1996;379 (6560):55–7.

[64]. Luzar A Resolving the hydrogen bond dynamics conundrum. J Chem Phys 2000;113(23):10663–75.

[65]. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project [presentation]. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases; 2013 Sep 23–27; Prague, Czech Republic; 2013.

[66]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12: 2825–30.

[67]. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 13–17; San Francisco, CA, USA. New York City: Association for Computing Machinery; 2016. p. 785–94.

[68]. McKinney W Data structures for statistical computing in Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010 Jun 28–Jul 3; Austin, TX, USA; 2010. p. 56–61.

[69]. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357–62. [PubMed: 32939066]

[70]. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9 (3):90–5.

[71]. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261–72. [PubMed: 32015543]

[72]. Zhang Y, Qiao Q, Abbas UL, Liu J, Zheng Y, Jones C, et al. Lignin derived hydrophobic deep eutectic solvents as sustainable extractants. J Clean Prod 2023;388:135808.

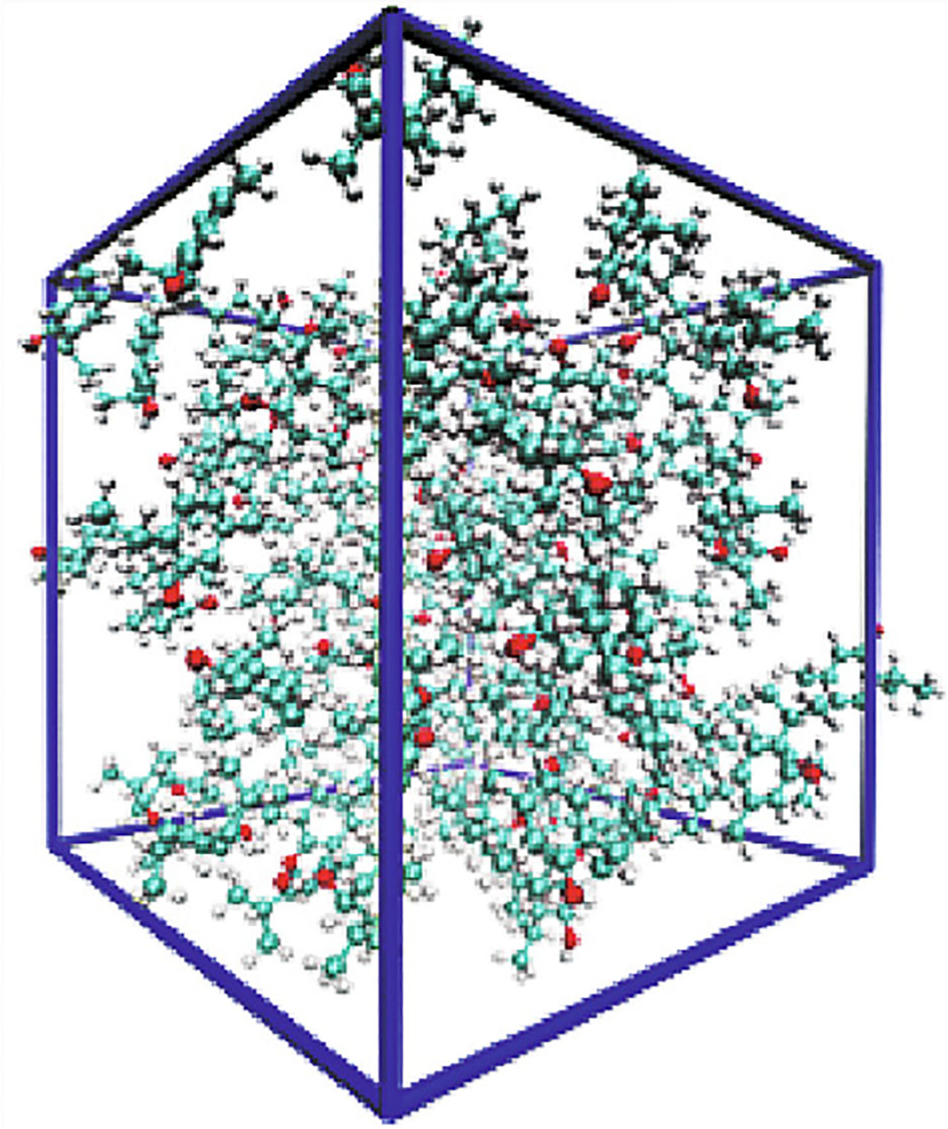

Fig. 1.

A snapshot of the equilibrated THY–MEN11 system. The molecules are displayed in Corey–Pauling–Koltun (CPK) model (color scheme of atoms: C, cyan; O, red; H, white).

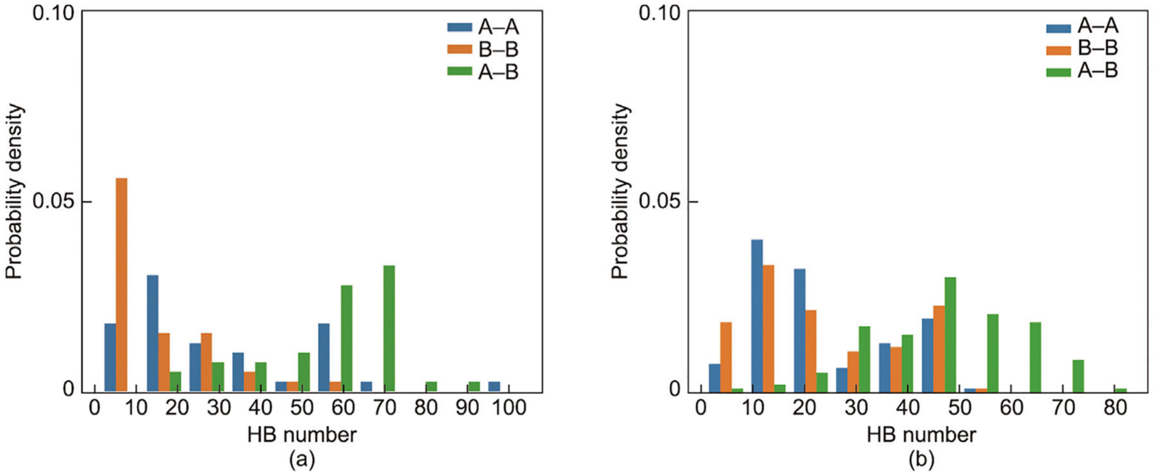

Fig. 2.

Probability density distribution of average inter-component (A–B) and intra-component (A–A and B–B) HB numbers: (a) DES systems and (b) non-DES systems.

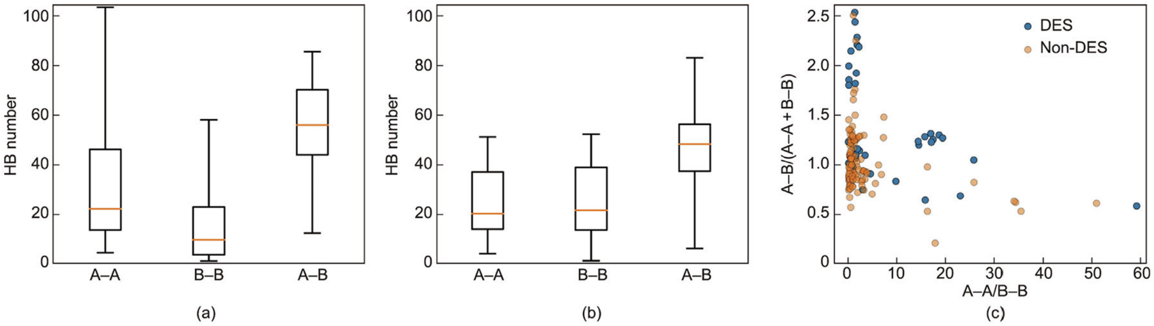

Fig. 3.

HB number features for DESs and non-DESs. (a, b) Distributions of average HB numbers: (a) DES systems and (b) non-DES systems. Starting from the bottom whisker, the boxplots show the minimum value; the 25th, 50th, and 75th percentile; and the maximum value. (c) Ratio of inter- to intra-component HB numbers vs ratio of intra-component HB numbers.

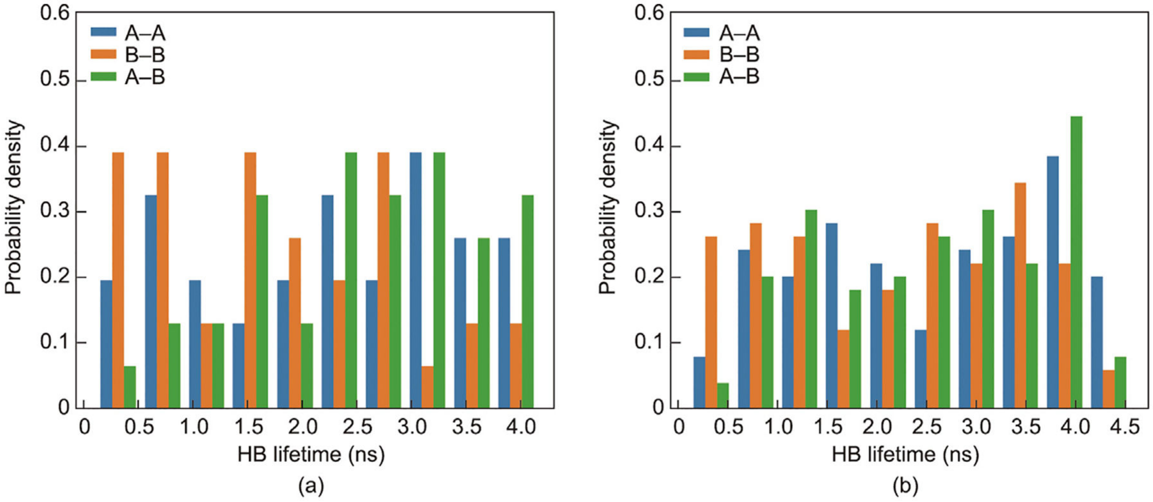

Fig. 4.

Probability density distribution of average inter-component (A–B) and intra-component (A–A and B–B) HB lifetimes: (a) DES systems and (b) non-DES systems.

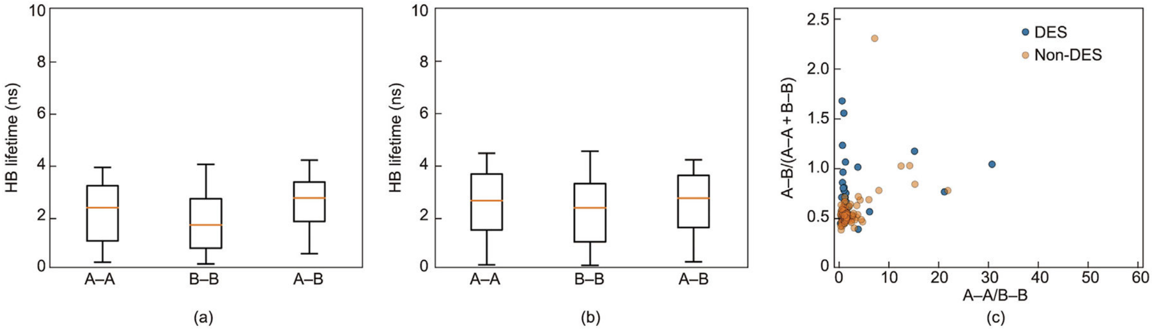

Fig. 5.

HB lifetime features for DESs and non-DESs. (a, b) Distributions of average HB lifetimes: (a) DES systems and (b) non-DES systems. Starting from the bottom whisker, the boxplots show the minimum value; the 25th, 50th, and 75th percentiles; and maximum value. (c) Ratio of inter- to intra-component HB lifetimes vs ratio of intra-component HB lifetimes.

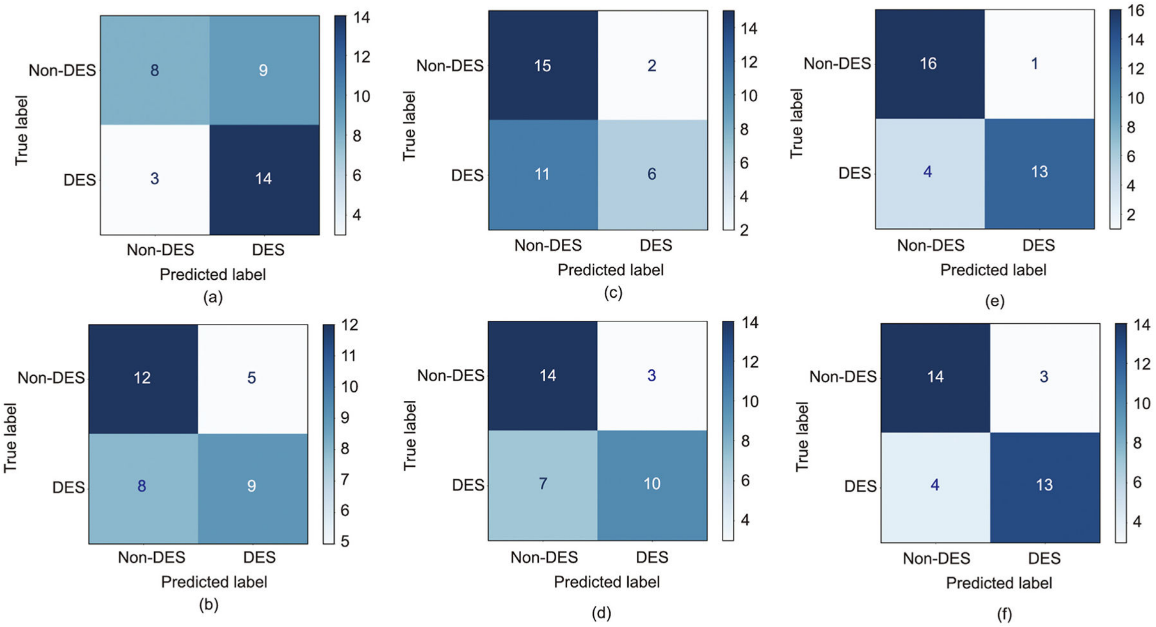

Fig. 6.

Confusion matrices for the top-performing models during validation. (a, b) HB lifetime features: (a) XGBoost-random forest and (b) logistic regression. (c, d) HB number features: (c) support vector machine and (d) extra trees forest. (e, f) Combined HB number and lifetime features: (e) extra trees forest and (f) logistic regression. The color bar on the right indicates the performance of the models, with white to blue indicating the performance from bad to good.

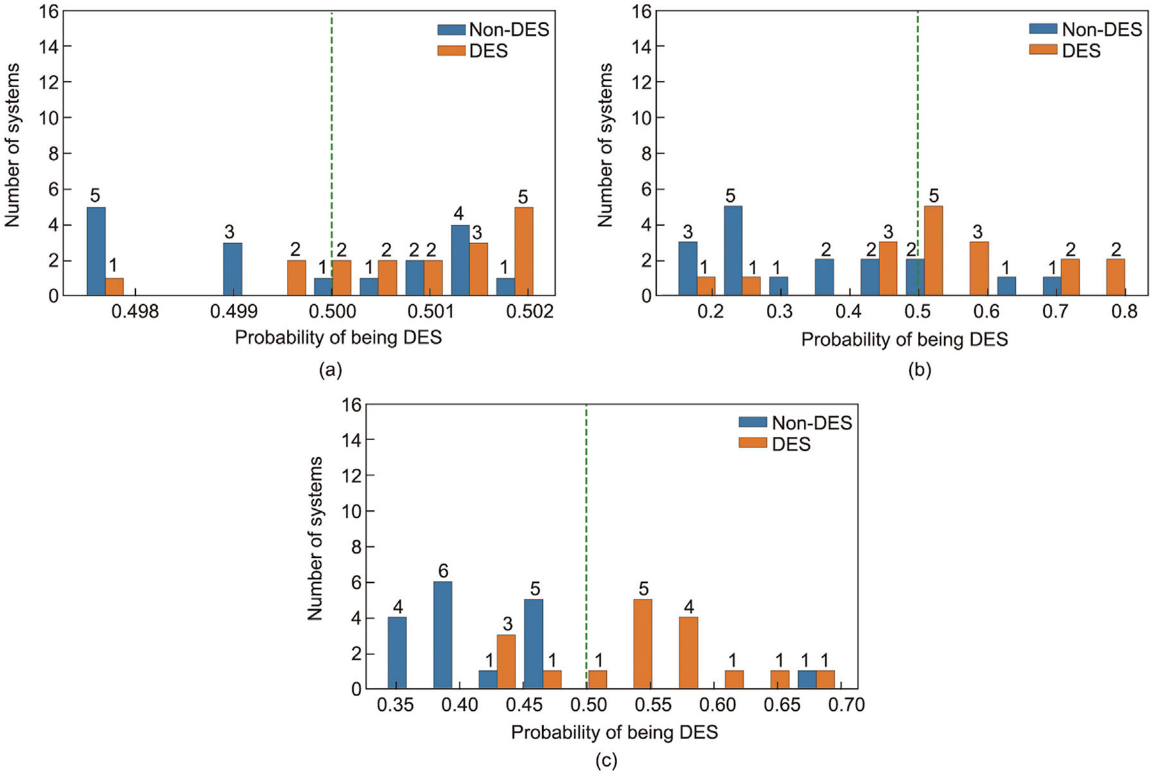

Fig. 7.

Distribution of prediction probabilities for the top-performing models during validation: (a) XGBoost-random forest model verified with HB lifetime features; (b) support vector machine model verified with HB number features; and (c) extra trees forest model verified with combined HB number and lifetime features. The number of systems within each bin is indicated on the bars. The vertical dashed line indicates the classification threshold. Perfect models will have all non-DESs on the left and DESs on the right of the vertical dashed line.

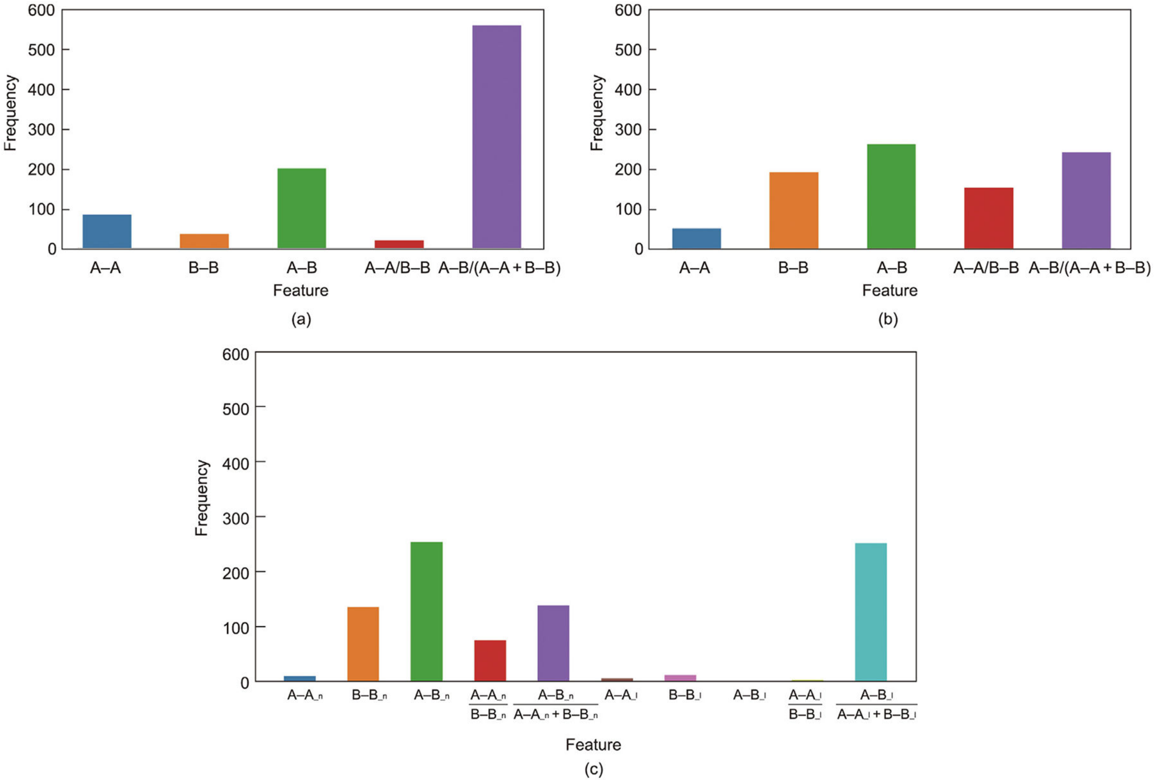

Fig. 8.

Important features during training iterations for all models: (a) models trained with HB lifetimes; (b) models trained with HB numbers; and (c) models trained with both HB numbers and lifetimes. In part (c), the "_n" and "_l" subscripts denote HB number and lifetime features, respectively.

Table 1

Average ROC-AUC values of the 30 models (best values are in bold).

Algorithm	Lifetime	Number	Number + lifetime
Logistic regression	0.68	0.78	0.77
Decision tree	0.63	0.74	0.68
Gradient boost	0.66	0.78	0.76
AdaBoost	**0.70**	0.78	0.75
Random forest	0.69	0.81	**0.79**
Extra trees forest	**0.70**	0.80	0.78
Support vector machine	0.64	0.77	0.77
k-nearest neighbors	0.63	0.77	0.77
XGBoost	0.67	0.81	0.77
XGBoost-random forest	0.62	**0.82**	**0.79**

Table 2

Minimum ROC-AUC scores for the 30 models (best values are in bold).

Algorithm	Lifetime	Number	Number + lifetime
Logistic regression	0.25	0.55	0.50
Decision tree	**0.30**	0.50	0.45
Gradient boost	0.25	0.50	0.40
AdaBoost	**0.30**	0.40	0.38
Random forest	**0.30**	0.45	0.45
Extra trees forest	**0.30**	**0.70**	**0.55**
Support vector machine	0.15	0.10	0.10
k-nearest neighbors	**0.30**	0.45	0.45
XGBoost	0.20	0.50	0.45
XGBoost-random forest	0.20	0.55	0.45

Table 3

ROC-AUC values of the trained models when tested with validation data (top-performing model under each feature type has its ROC-AUC value in bold).

Algorithm	Lifetime	Number	Number + lifetime
Logistic regression	0.65	0.66	0.84
Decision tree	0.52	0.69	0.65
Gradient boost	0.57	0.77	0.81
AdaBoost	0.61	0.74	0.66
Random forest	0.54	0.76	0.79
Extra trees forest	0.65	0.79	**0.88**
Support vector machine	0.56	**0.80**	0.80
k-nearest neighbors	0.47	0.53	0.57
XGBoost	0.61	0.65	0.74
XGBoost-random forest	**0.68**	0.69	0.79