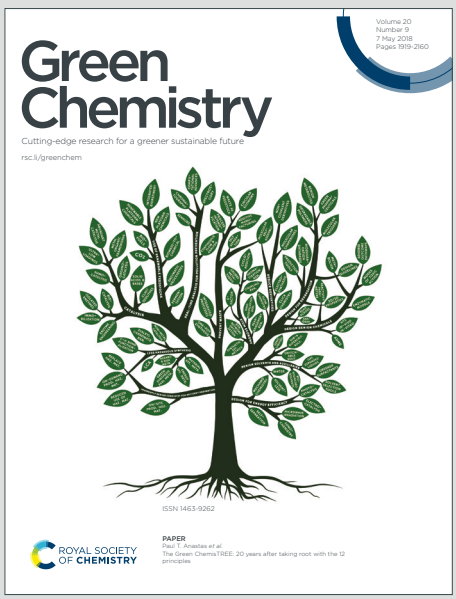
Green Chemistry

Cutting-edge research for a greener sustainable future

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: D. Usmanov, P. Yadav, G. M. Casanola-Martin, A. Mallya, S. Shirvanihosseini, A. Hubel and B. Rasulev, *Green Chem.*, 2026, DOI: 10.1039/D6GC01009A.



ROYAL SOCIETY OF CHEMISTRY

rsc.li/greenchem

View Article Online
DOI: 10.1039/D6GC01009A

Machine Learning Framework to Predict Glass Transition Temperature in Natural Deep Eutectic Solvents: A Step toward Green Functional Materials

Green Foundation

1. This work advances green chemistry by enabling data-driven design of natural deep eutectic solvents (NADES), biodegradable and low-toxicity alternatives to conventional organic solvents. We develop an interpretable machine-learning framework that predicts the glass transition temperature (Tg) directly from mixture-aware molecular descriptors, enabling virtual prescreening of formulations prior to synthesis.

2. By reducing trial-and-error experimentation and unnecessary differential scanning calorimetry measurements, the approach lowers material consumption, solvent waste, and energy input associated with formulation development. This strategy shifts solvent discovery toward predictive, sustainability-driven design.

3. Importantly, the framework provides chemically interpretable guidance that links hydrogenbond network topology and molecular flexibility to thermal behavior, thereby supporting the rational selection of sustainable compositions. Future work will expand compositional diversity and integrate multi-objective optimization (e.g., viscosity, toxicity, water tolerance) to further strengthen environmentally responsible solvent design.

View Article Online
DOI: 10.1039/D6GC01009A

ARTICLE

Machine Learning Framework to Predict Glass Transition Temperature in Natural Deep Eutectic Solvents: A Step toward Green Functional Materials

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Durbek Usmanov,[a, b] Priyanka Yadav,[c] Gerardo Casanola-Martin,[a] Akshat Mallya,[c] Seyedehelham Shirvanihosseini,[a] Allison Hubel,[c, d] Bakhtiyor Rasulev,*[a, b]

Natural Deep Eutectic Solvents (NADES) are a promising class of sustainable and environmentally-safe solvents with highly tunable physicochemical properties, including the glass transition temperature, which is critical for their functional performance, including ice control applications. Here, we present an interpretable machine learning (ML) framework to predict glass transition temperature (1) from the molecular structure of NADES combination, integrating descriptor-based feature engineering, unsupervised clustering, and ensemble regression. Combination of components and their mixing ratios for forming NADES were utilized to generate specific multi-component descriptors to describe NADES for ML modeling. A set of multicomponent descriptors was calculated based on individual descriptors from chemically diverse components of NADES. In result, a Random Forest (RF) model was developed to predict Tg values of NADES and the model achieved a very good performance with R^2 values in a range of 0.87–0.93, for both training and test sets. The analysis of contributing factors by Shapley Additive exPlanations (SHAP) analysis identified key features highlighting contributions of 3D geometry, atomic mass distribution and electronic effects. Finally, our results demonstrate that ML approaches combined with mixture descriptors approach and interpretable modeling, enables accurate and chemically meaningful prediction of Tg, facilitating the rational design of NADES for applications in green chemistry and sustainable materials science applications.

1. Introduction

Climate change drives the global scientific community to search for novel materials that can replace conventional, often toxic substances with biodegradable alternatives without compromising essential physical properties. In this context, Natural Deep Eutectic Solvents (NADES) have emerged as a promising class of materials due to their biodegradability (2), low toxicity (3), and tunable physicochemical properties, such as high thermal stability (4), conductivity (5), adjustable viscosity (6), polarity (6), low volatility (7), water solubility (7), stability in different concentrations, and a high solubilization capacity for a wide range of pharmacologically active compounds (7). Despite their relatively short history of research, it is well-known that NADES demonstrate a unique ability to form stable hydrogen-bonded networks, making them promising sustainable alternatives for various applications, including as pesticides (8), absorption materials for gases (9), coatings (10), cosmetics (11), drug delivery systems (12), biosensors (13), solar cells (14), supercapacitors (15), battery electrolytes (16), anti-freezing or anti-icing agents(17, 18), for cryopreservation(19) and in green chemistry processes (20).

One of the crucial thermophysical properties that significantly influences the materials, particularly in coatings, encapsulation, and flexible electronics, is the glass transition temperature (21-23). Tg defines the temperature range at which a material transforms from a rigid, glassy state to a soft, rubbery, phase (24). This transition is essential for determining thermal stability, mechanical integrity, and usability of materials under varying environmental conditions. For NADES-based systems, controlling the Tg is vital in ensuring functionality in temperature-sensitive applications; however, systematic studies on their Tg behavior remain limited due to the complex, multicomponent nature of their hydrogen-bonded networks (21-23). In cryopreservation applications, higher cryoprotectant Tg values have been linked to lower critical cooling and warming rates for vitrification (25), reduced likelihood of damaging ice crystal formation (26), and the reduction of thermal stresses in vitrified samples, leading to lower chances of catastrophic mechanical failure (27). While Tg is commonly measured using experimental techniques such as differential scanning calorimetry (DSC), thermomechanical analysis (TMA), and dynamic mechanical analysis (28), computational modelling and machine learning are increasingly being adopted to predict Tg from molecular descriptors (29, 30).

These data-driven approaches offer valuable insights into the relationship between molecular structure and thermophysical behavior, accelerating the design of NADES and other next-generation materials for advanced technological applications. However, despite the growing interest in NADES, comprehensive studies on their glass transition behavior remain limited, primarily because of the complexity of their multicomponent hydrogen-bonded networks and the variability in their physicochemical properties (8, 31). In recent years, accurately estimating the glass transition temperature of materials, particularly in confined

[a] Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND, USA.
[b] Materials and Nanotechnology (MNT) Program, North Dakota State University, Fargo, ND, USA.
[c] Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, USA. Address here.
[d] Department of Biomedical Engineering, University of Minnesota, Minneapolis, MN, USA.

View Article Online
DOI: 10.1039/D5GC01009A

geometries, has become increasingly important due to Tg's strong dependence on both composition and environmental factors (32). However, the experimental determination of Tg for newly developed material systems remains a significant challenge, as such measurements are often labor-intensive, time-consuming, and costly (23). Fortunately, the availability of experimental Tg datasets has opened the door to benchmarking computational approaches that offer a more efficient and scalable means of predicting Tg.

Beyond property prediction, machine learning is increasingly being used to optimize sustainable chemical processes, including biomass fractionation workflows (33, 34). These advances underscore the broader potential of data-driven approaches in green chemistry, where predictive modeling can reduce experimental burden, accelerate optimization, and improve decision-making in complex multivariable systems. In this context, machine learning (ML)-based quantitative structure-property relationship (QSPR) models as efficient tools for solving complex challenges in chemistry, biology and materials science (35-44). These models directly link molecular structures to their physicochemical properties, offering a faster and more cost-effective alternative to traditional trial-and-error experimentation. Scientists have successfully applied QSPR models to a wide range of chemical systems, including the prediction of Tg, significantly advancing the rational design of new materials (30, 45-51). Although many studies have predicted Tg values for various materials, researchers have yet to develop a dedicated dataset or model for Tg prediction in NADES. These advancements underscore the increasing effectiveness of ML-based approaches in capturing complex structure–property relationships within NADES systems. However, a review of the current literature reveals that most studies have predominantly focused on melting point(52, 53), density(54, 55), and viscosity of NADES (56, 57), but only a little attention given to Tg. This gap highlights the urgent need for development of dedicated computational and predictive models for Tg, specifically designed to address the unique structural and physicochemical complexities of NADES.

Considering these limitations and to the best of our knowledge, in this study, we developed a ML–based QSPR model for predicting the glass transition temperature of NADES. We integrated advanced combination of computational techniques to improve predictive accuracy and model interpretability. The Random Forest algorithm was selected as the primary modeling tool due to its robustness in capturing complex, nonlinear relationships across multivariate descriptor sets. The dataset, which comprises Tg values ranging from -122.87 to -52.54 °C, was analyzed using a combination of ML and dimensionality reduction methods. Specifically, Uniform Manifold Approximation and Projection (UMAP) was employed to explore the chemical space of NADES compounds, revealing meaningful structural clusters of NADES in the visualized chemical space. This unsupervised clustering guided the splitting of the dataset into training and test sets, thereby ensuring a structurally diverse and representative evaluation framework. To interpret the model predictions and assess the influence of individual molecular features on Tg, a SHAP approach was applied. This enabled a quantitative understanding of the most significant descriptors influencing the thermal behavior of NADES. This integrated approach not only enables accurate prediction of Tg but also provides valuable insights into the molecular determinants of glass transition behavior in eutectic solvent systems. Also, since today's materials science increasingly prioritizes environmentally friendly and sustainable alternatives, the developed model offers a powerful tool for the rational design of novel NADES formulations with tailored thermophysical properties for green and environmentally safe applications.

2. Methodology

2.1 Materials

L-Proline (Pro, ≥99%, Sigma Aldrich), Betaine (Bet, ≥98%, Sigma Aldrich), Guanidine Hydrochloride (Gua, ≥99.5%, Thermo Scientific Chemicals), Choline chloride (ChCl, 98%, Oakwood Chemical), Sarcosine (Sar, 98%, Sigma Aldrich), L-Lysine (Lys, ≥98%, Sigma Aldrich), Trimethylamine N-oxide (TMAO, ≥98%, Thermo Scientific Chemicals), L-Arginine (Arg, ≥98%, Sigma Aldrich), Urea (U, ≥99.5%, Fluka Chemie GmbH), D-(+)-Glucose (Glu, ≥99.5%, Sigma Aldrich), D-(-)-Fructose (Fru, ≥99%, Sigma Aldrich), Sucrose (Suc, puriss grade, Sigma Aldrich), D-(+)-Trehalose dihydrate (Tre, ≥99%, Sigma Aldrich), D-Sorbitol (Sor, ≥98%, MP Biomedicals), Xylitol (Xyl, ≥99%, Sigma Aldrich), Ethylene Glycol (EG, 99.8%, Sigma Aldrich), Glycerol (Gly, 99.5%, Humco) and Ultrapure-grade water was used from a Milli-Q filtration system to prepare the NADES samples.

2.2 NADES Preparation

NADES-forming components were added to capped glass vials (WHEATON) in the molar ratios specified in this work until a total mass of 15 g was achieved. The glass vials containing the mixtures were then placed in a water bath at 60 °C and stirred at 200 rpm for 24 h. After mixing, the samples were allowed to rest at room temperature for 72 h before performing further analyses. Polarized optical microscopy was performed for all NADES mixtures to confirm the absence of crystals (58). The NADES formulations included in the study were prepared using a mixture of sugars (glucose, sucrose, etc.), sugar alcohols (glycerol, sorbitol, etc.), amino acids (proline, lysine, etc.) and other metabolites like urea and trimethylamine N-oxide. Two, three and four-component mixtures were prepared in a wide range integer molar ratios. Water was added to the prepared NADES to create 20% w/w to 80% w/w solutions of diluted NADES. Details about the composition of the NADES formulations are included in Table S2.

2.3 Thermophysical Characterization

Differential scanning calorimetry was carried out using a TA Instruments Q1000 calorimeter. For each measurement, 15-20 mg of a NADES sample was placed into an aluminum pan (Tzero, TA Instruments) and sealed with a hermetic lid (Tzero, TA Instruments). The samples were first cooled from 20 °C to -150 °C at a constant rate of 10 °C/min, followed by a 3 min isothermal equilibration at -150 °C to ensure thermal stability. Subsequently, the samples were heated from -150 °C back to 20 °C at 10 °C/min. Thermophysical properties of the NADES samples were evaluated using TA Universal Analysis software (version 4.5A). The glass transition temperature (1) was identified as the first local minimum observed in the plot of the

derivative of heat flow with respect to temperature during the heating step.

2.4 Computational Data Analysis and Structural Descriptors Calculation

The structural characterization of NADES' components began with the construction of individual molecular structures using ChemSketch software (59). The optimized components were subsequently processed by applying AlvaDesc code to generate a comprehensive array of molecular descriptors (60). For each NADES system, mixture-type descriptors were generated by combining the descriptors of individual components, multiplying on their respective molar ratios (61, 62).

This approach produced an initial set of 5,666 descriptors per each NADES system, spanning a broad spectrum of physicochemical, topological, geometrical, and electronic properties. These descriptors encompassed multiple-dimensional categories, including 0D, 1D, 2D, and 3D, and were drawn from various descriptor classes, such as constitutional indices, topological and connectivity metrics, 2D autocorrelations, geometrical parameters, atom-centered fragments, and quantum-chemical features.

To minimize redundancy and eliminate non-informative variables, descriptors exhibiting zero variance or near-constant values across the dataset were removed. Following this filtering process, a final set of 4,316 descriptors was retained for further machine learning modeling. Importantly, this descriptor set was not used directly in its entirety for interpretation or final decision-making. Instead, feature relevance was further refined in a data-driven manner using Random Forest–based learning and SHAP analysis, which effectively reduced the dimensionality to a compact subset of influential descriptors governing Tg prediction. A detailed summary of the initial descriptor set is provided in the SI (Table S2).

These curated descriptors formed the input features for all subsequent predictive modeling, providing a high-dimensional and chemically meaningful representation of the NADES space for accurate prediction of the glass transition temperature. Besides, Fig. 1 illustrates the general workflow.

2.5 Uniform Manifold Approximation and Projection (UMAP)

PUMAP is a nonlinear dimensionality reduction technique that constructs fuzzy topological representations of data by modeling the high-dimensional input space as a weighted graph. In this method, the probability of a connection between two data points in the high-dimensional manifold is expressed as:

$$p_{ij} = e^{-d(x_i,x_j)/\sigma_i}$$

Where $d(x_i, x_j)$ is the distances between points i and j, and σ_i is a normalization factor that controls the local connectivity. In the corresponding low-dimensional space, UMAP models the connection probability using:

$$q_{ij} = (1 + a(y_i - y_j)^2)^{-b}$$

Where y_i and y_j are the projected low-dimensional representations and a and b are hyperparameters that are the curve used in low-dimensional space.

Unlike to other methods, UMAP minimizes a cross-entropy loss function defined as:

$$CE(X,Y) = \sum_i \sum_j \left[p_{ij}(X) log\left(\frac{p_{ij}(X)}{q_{ij}(Y)}\right) + \left(1 - p_{ij}(X)\right) log\left(\frac{1 - p_{ij}(X)}{1 - q_{ij}(Y)}\right) \right]$$

This formulation allows UMAP to preserve both local and global data structures, enabling it to capture meaningful relationships in complex high-dimensional datasets.

The two principal hyperparameters in UMAP are the number of neighbors and the minimum distances. The number of neighbors defines the size of the local neighborhood used for manifold approximation – lower values enhance the preservation of local structures, whereas higher values emphasize global relationships. The minimum distance controls how tightly data points are packed in the low-dimensional space: smaller values promote tight clustering, while larger values preserve more of the global data topology.

To assess the clustering structure in the UMAP–reduced space, k-means clustering was applied. The optimal number of clusters was determined using internal validation metrics, namely the Davies-Bouldin Index (DBI) and the Silhouette Score (SS).

The SS $s(i)$ for each sample is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance. The overall score, averaged across all samples, ranged between -1 and 1, with values closer to 1 indicating well-defined clusters.

The DBI is defined as:

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

where Si is the average distance within cluster i and Mij is the centroid distance between clusters i and j. Lower DBI values signify better separation between clusters.

After selecting k=6 as the optimal number of clusters, final cluster assignments were computed using k-means, and centroid coordinates were extracted from the UMAP embedding space. Prior to dimensionality reduction, all numerical descriptors were standardized using StandardScaler (zero mean, unit variance). UMAP was then applied using the umap-learn implementation with n_neighbors = 20, min_dist = 0.1, n_components = 2, metric = 'euclidean', and random_state = 42 to ensure reproducibility. UMAP projections and cluster validation were implemented using the UMAP-learn and Scikit-learn libraries in Python. These results confirm that the UMAP transformation not only preserves meaningful chemical relationships in the descriptor space but also facilitates robust unsupervised clustering of NADES formulations.

2.6 Machine Learning Model and Evaluation Metrics

In this study, in the best model the Random Forest Regression (63) was employed to predict the Tg of the NADES systems. RF is a widely used ensemble learning method that builds a collection of decision trees and aggregates their outputs to produce robust predictions. The algorithm operates by constructing multiple regression trees from random subsets of the training data and features, and then averaging their outputs to mitigate variance and

overfitting. This approach is convenient in handling high-dimensional descriptor spaces and complex nonlinear relationships.

The RF model was implemented using the Scikit-learn library in Python. Hyperparameters – including the number of estimators (trees), maximum tree depth, and minimum samples per leaf – were optimized using grid search with 5-fold cross-validation to ensure generalization and minimize overfitting. The RF regression model was implemented in Python using the Scikit-learn library (RandomForestRegressor). Hyperparameter optimization was performed using grid search with 5-fold cross-validation (GridSearchCV, scoring = R^2) and a fixed random seed (random_state = 42) to ensure reproducibility. The hyperparameter search space included the number of trees (n_estimators = 100–400), maximum tree depth (max_depth = 4–10), minimum number of samples required to split an internal node (min_samples_split = 2–10), and minimum number of samples per leaf (min_samples_leaf = 1–4). Cross-validation was conducted exclusively within the training set, following the UMAP-based structure-aware train–test splitting strategy. The held-out test set was not used during model selection and was reserved solely for final performance evaluation, thereby minimizing information leakage and overfitting.

To evaluate the performance of the model, the following metrics were calculated on the training, test, and external validation datasets:
Mean Absolute Error (64):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - A_i|$$

Coefficient of Determination (R^2):

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(A_i - P_i)^2}{\sum_{i=1}^{n}(A_i - \mathring{A})^2}$$

Root Mean Squared Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - P_j)^2}$$

Where *Ai* and *Pi* denote the actual and predicted Tg values, respectively, for the i^{th} sample, *Å* is the mean of the exact values, and *n* is the total number of samples.

These statistical indicators were selected to ensure a comprehensive evaluation of the model's accuracy, error distribution, and generalization capacity. A high R^2, alongside low MAE and RMSE values, indicates the reliable predictive performance and robustness of the trained RF model..

2.7 Feature Selection and Model Interpretability

To enhance model transparency and ensure the reliability of Tg predictions, we employed a dual strategy that integrates both data-driven feature selection and post hoc model interpretability analysis. To interpret feature contributions and quantify their influence on the model output, we applied SHAP, a model-agnostic interpretability method based on cooperative game theory. SHAP values were computed using the *shap.TreeExplainer* module (SHAP v0.44.1), which is optimized for ensemble tree models such as Random Forest.

Prior to SHAP analysis, standard descriptor preprocessing steps were applied, including the removal of constant descriptors and highly correlated variables, to ensure numerical stability and reduce redundancy in the descriptor space. After initial descriptor preprocessing, SHAP-based feature importance analysis was used to identify the most informative variables. The final predictive models were constructed using the top 35 descriptors ranked by SHAP importance, thereby substantially reducing the dimensionality of the descriptor space relative to the number of samples. This dimensionality reduction step mitigates the risk of overfitting and improves model interpretability.

It should be noted that SHAP provides an additive decomposition of model predictions and does not explicitly reconstruct the full nonlinear functional form learned by the underlying machine learning model. Consequently, SHAP-based interpretations should not be viewed as mechanistic or causal descriptions of molecular interactions.

Nevertheless, SHAP remains a powerful interpretability tool for identifying dominant structure–property trends, directional effects, and threshold-like behaviours embedded within nonlinear models. In this study, SHAP analysis is used to reveal consistent, chemically plausible relationships between molecular descriptors and Tg, rather than to infer explicit nonlinear interaction mechanisms. The observed SHAP dependence patterns (e.g., saturation and sigmoidal responses) reflect the nonlinear decision boundaries learned by the model, providing physically meaningful insight into the factors governing vitrification in NADES.

SHAP analysis allowed for a detailed decomposition of Tg predictions into additive feature attributions, revealing both global feature importance and local instance-specific effects. Visualization tools, including beeswarm and summary plots, were used to identify the most influential descriptors. Among the top contributors were 3D-MoRSE and GETAWAY descriptors (e.g., *Mor08p*, *Mor26p*, *R6m+*) and topological autocorrelations (e.g., *MATS4i*), highlighting the importance of molecular geometry, polarizability, and electronic distribution in modulating the thermal behavior of NADES systems.

This framework not only improved the interpretability of our model but also provided mechanistic insights into structure–property relationships within the NADES chemical space, thereby supporting rational design and formulation strategies.

2.8 Visualization

All graphical analyses and visualizations were performed using Python (version 3.10.11) with the Matplotlib (version 3.7.1) and Seaborn (version 0.12.2) libraries. These tools were used to generate model performance plots (e.g., R^2 comparisons, SHAP summary plots), feature importance rankings, UMAP projections, clustering diagrams, and correlation heatmaps. Custom plotting scripts ensure consistent visual styling, interpretability, and high-resolution rendering [≥300 dpi), making the figures suitable for publication in both print and digital formats. All visualization codes are fully reproducible and provided in the Supplementary Information (SI).

3. Results

This work contributes to green chemistry by enabling data-driven pre-screening of NADES formulations, reducing experimental waste, solvent consumption, and energy-intensive trial-and-error synthesis. In this study, we used an in-house dataset comprising 263 NADES systems with experimentally measured Tg values to develop several machine learning methods for Tg prediction. This is unique experimental dataset, collected by the group, since all the other public data available are scarce and made with different protocols.

View Article Online
DOI: 10.1039/D6GC01009A

Therefore, the collection of 263 NADES with measured Tg under the same experimental protocol is a unique set for a proper ML study. To investigate Tg of NADES in detail, we systematically evaluated different algorithms using a curated dataset of NADES formed from various components' combinations. The results presented in this work include model development steps, performance metrics, feature importance analysis based on SHAP analysis, and insights into molecular descriptors that most significantly influence Tg values.

After careful data curation, the dataset of NADES was used to develop predictive models, which were trained and evaluated using a "mixture-based split" strategy, ensuring that structurally related compounds were not included in both the training and test sets simultaneously. This approach provides a more realistic assessment of the ML model's generalizability, especially for prospective applications involving prescreening of novel NADES formulations. The overall workflow of the study is shown in Fig. 1.

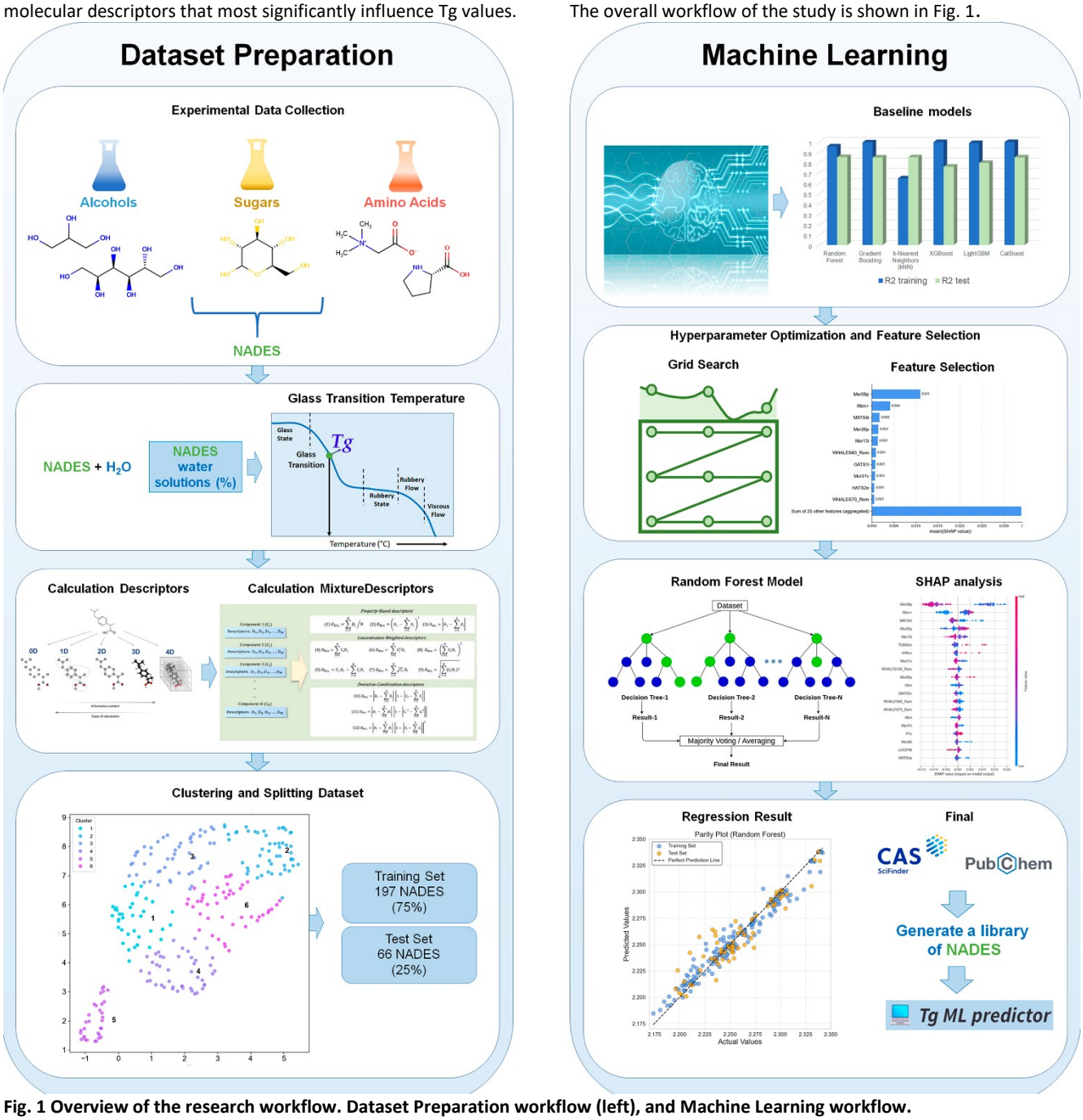


Fig. 1 Overview of the research workflow. Dataset Preparation workflow (left), and Machine Learning workflow.

3.1 Experimental Analysis

The glass transition temperature was determined from the DSC thermogram by analyzing the first derivative of heat flow with respect to temperature, where the peak of the first local maximum during the heating step was taken as Tg. Upon cooling, pure NADES (without dilution) samples exhibited vitrification without undergoing crystallization, thereby forming a metastable supercooled liquid state, eventually undergoing a glass transition into a non-equilibrium glassy phase. This was observed in their thermograms by noting the presence of only a glass transition without a melting peak, consistent

with eutectic behavior (5, 65, 66). The Tg values of pure NADES were found to lie between -114.54 °C (Fruc: EG 1:66) to -53.64 °C (Gly:Glu:Sorb:Water 1:1:1:3), lower than those observed for the water-diluted samples. These findings agree with previous reports showing that NADES remain in the liquid state across a broad low-temperature range (66, 67).

Water addition was found to disrupt the supramolecular organization of NADES, increase molecular mobility, and lower Tg (66, 68). Tg decreased with increasing dilution up to 50% but remained nearly constant at higher dilutions (18). The reduction in Tg upon hydration reflected the well-known plasticizing role of water (5). Consistent with the Flory-Fox equation, Tg was observed to be directly proportional to molecular weight. NADES formulated with low-molecular-weight ethylene glycol reported the lowest Tg, whereas those with high-molecular-weight sorbitol exhibited higher Tg values (69).

3.2 Dataset Preparation for ML-QSPR Modeling

A crucial first step before performing any ML modeling is a proper splitting of the dataset to avoid data leakage and achieve a balanced structural diversity distribution of the NADES dataset through unbiased train-test data splitting. For this, we applied the UMAP unsupervised classification approach (70) to perform dimensionality reduction on the dataset and facilitate visualization. UMAP effectively projects the high-dimensional molecular features (descriptors) space into two dimensions while preserving both local and global topological structures (71, 72). To quantitatively determine the optimal number of clusters, we evaluated the Silhouette Score (SS) (73) and Davies–Bouldin index (DBI) (74) across multiple cluster configurations (from 2 to 17 clusters). As shown in Fig. 2a, the SS reached its maximum (SS=0.476) and the DBI its minimum (DBI=0.695) at k = 6, indicating the optimal balance between cluster cohesion and separation. As shown in Fig. 2b, the UMAP revealed the presence of six well-defined clusters, each representing distinct regions of chemical space, thus confirming the molecular diversity within the dataset. These six clusters served as the foundation for splitting the dataset into training and test sets, ensuring that the model would be trained and evaluated on structurally diverse and representative subsets. Furthermore, the UMAP-based clustering approach provides insights into the underlying structural patterns of the NADES compounds, as seen in the distribution of the clusters. Each cluster exhibits distinct chemical characteristics, which may correlate with differences in its glass transition behavior. By using this structure-aware data partitioning strategy, we minimized potential data leakage and enhanced the predictive model's generalization ability.

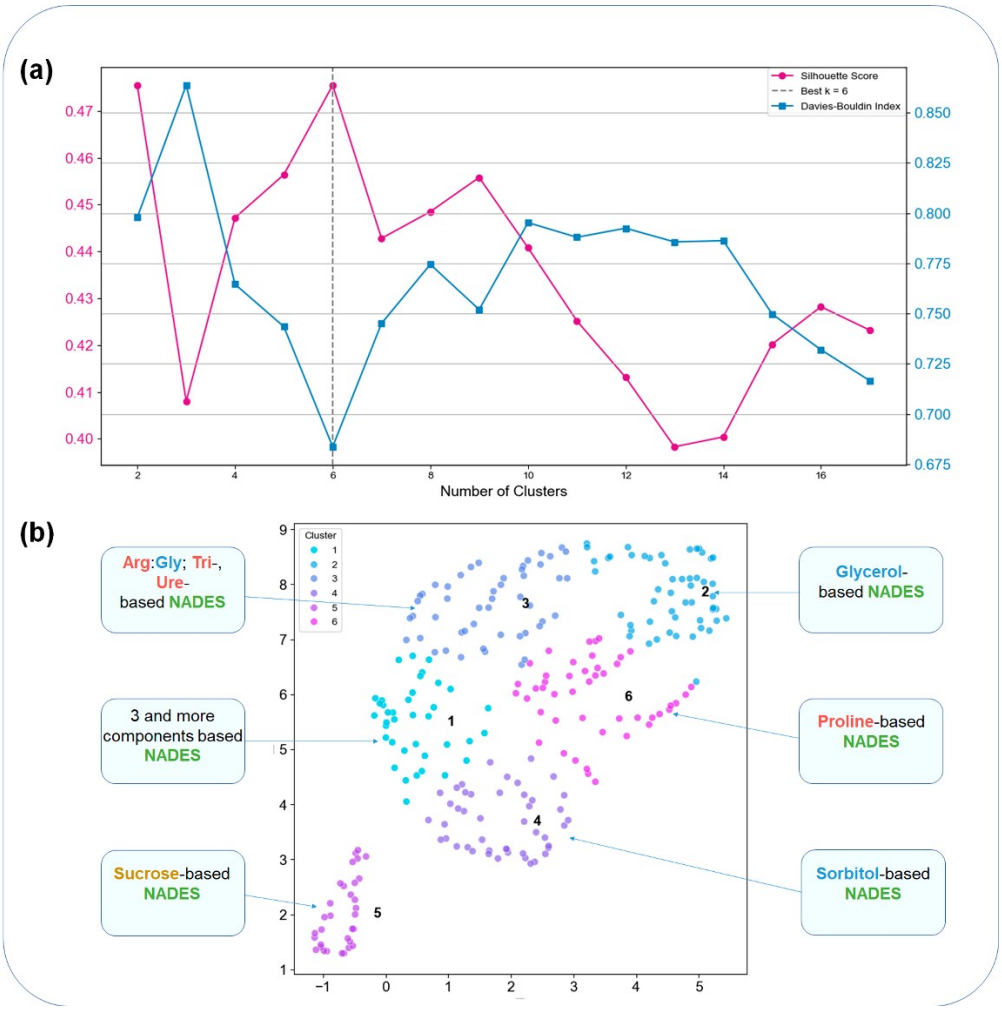


Fig. 2 Clustering Analysis and Structural Distribution of NADES Systems; **a**) An evaluation of clustering performance using the Silhouette score and the Davies-Bouldin index; **b**) UMAP-Driven Clustering of molecular descriptors using K-Means

View Article Online
DOI: 10.1039/D6GC01009A

It is worth noting that water content can significantly influence the Tg value. To explore this, the importance of water concentration on the Tg within structurally similar NADES systems was investigated as well, and a comparative violin plot was generated (Fig. 3), highlighting various combinations of the same core components with differing water content. As can be noticed in the plot, the systems Bet:Ure:Wat**,** Bet:Sor:Wat**,** Bet:Sor:Ure:Wat**,** and Bet:Lys:Wat exhibit minimal shifts and smooth transitions in Tg values as water concentration increased. Water concentration was explicitly reported for each formulation (Table 2) and treated as part of the mixture composition during descriptor generation for the machine learning model. The consistently narrow violin shapes across these systems indicate that the incorporation of water had only a moderate impact on their thermal behavior. This stability likely reflects robust intermolecular interactions among the primary components (e.g., betaine, sorbitol, urea, lysine), which are only marginally affected by hydration. Conversely, other NADES systems demonstrated broader or more irregular Tg distributions, with pronounced changes in response to variations in water. This suggests that their physicochemical properties are more sensitive to hydration, possibly due to weaker or more dynamic hydrogen bonding networks. These findings emphasize that specific NADES formulations preserve structural and thermal coherence even under changing moisture conditions, making them attractive candidates for applications requiring thermophysical stability. This analysis underscores the chemical and functional diversity within the NADES formulations, highlighting the need to employ advanced machine learning techniques to capture the complex, nonlinear relationships between molecular structure and physicochemical properties. The applied methodology enhances the interpretability and robustness of the predictive framework, enabling more accurate and generalizable predictions of the glass transition temperatures across a structurally diverse range of NADES systems.

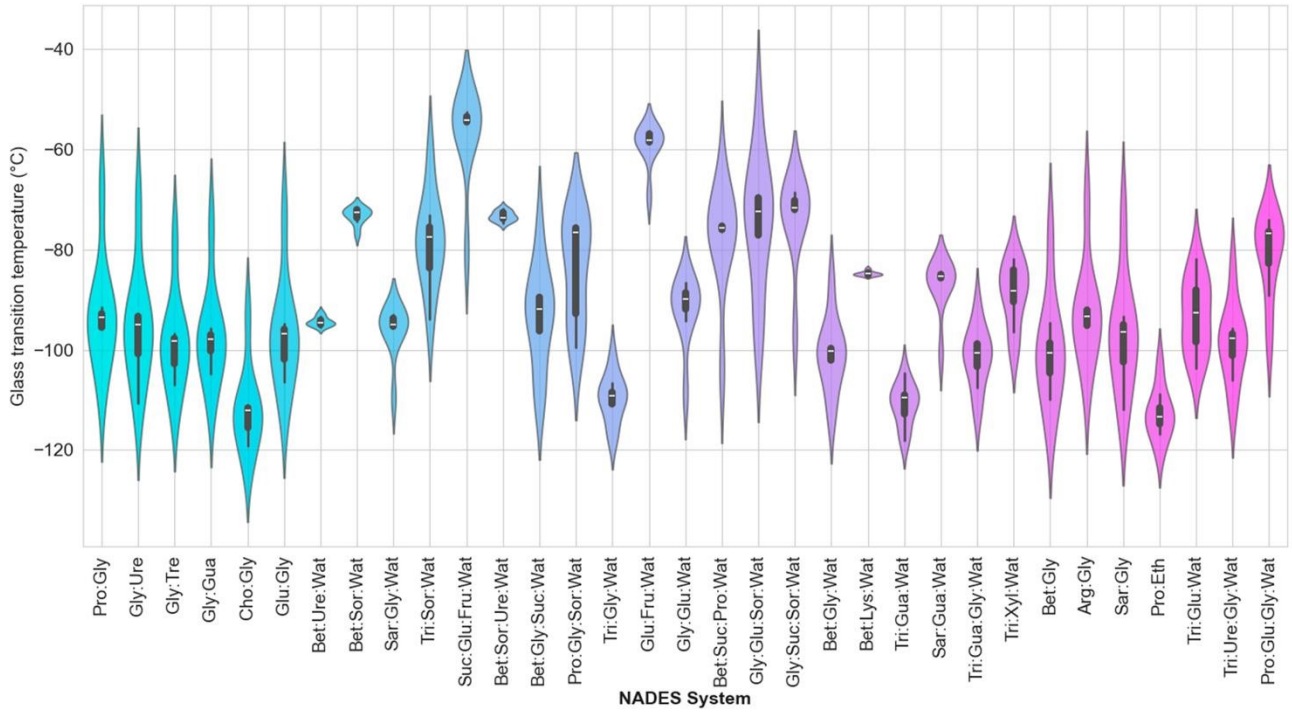


Fig. 3 Distribution of Tg across NADES systems with varying water concentration.

3.3 Development of ML-QSPR Model

To assess confidently the Tg of NADES, the capabilities of several machine learning algorithms were analyzed, including RF (75), CatBoost (76), XGBoost (77), Gradient Boosting (78), LightGBM (79), Support Vector Regression (SVR) (80), k-Nearest Neighbors (kNN) (81), Gaussian Process (82), Artificial Neural Networks (ANN) (83), and Lasso Regression (84). Each individual ML model was optimized through hyperparameter tuning, and the models' performances were evaluated using the coefficient of determination (R^2) for the training set, mean squared error (85), and R^2 values for the test set (Supplementary Information file, Table S1). To establish a performance benchmark, we initially evaluated a set of baseline ML models. This included both linear and non-linear regression algorithms. Among these, the RF model demonstrated the most promising performance, achieving R^2 of 0.87 for the test set and a low mean squared error (MSE = 0.0001), while maintaining a strong fit on the training data (R^2 = 0.96), indicating good generalization (Table S1). In contrast, some models displayed tendencies toward overfitting; for instance, XGBoost and CatBoost achieved nearly perfect R^2 values on the training set (0.99), their predictive

View Article Online
DOI: 10.1039/D6GC01008A

performance declined on the test set to 0.75 and 0.89, respectively (Table S1), indicating limited generalization ability. Similarly, Gradient Boosting and LightGBM models exhibited high training R^2 values but decreased test performance, suggesting that these models capture complex patterns in the training data, but they lack robustness when exposed to unseen data without further regularization. These results show that RF algorithm is highly effective in modeling the complex, nonlinear relationships between molecular descriptors and the Tg of NADES formulations. Its superior generalization performance underscores the robustness of ensemble tree-based approaches for capturing structure–property relationships in chemically diverse systems. Furthermore, the observed tendency of specific advanced boosting methods to overfit emphasizes the necessity of careful model selection and validation to ensure predictive reliability.

To further enhance the predictive performance of the RF model, we conducted an extensive hyperparameter optimization using a grid search approach. Key parameters, including the number of trees [n_estimators], maximum tree depth [max_depth], minimum number of samples required to split an internal node [min_samples_split], and minimum number of samples needed for a leaf node [min_samples_leaf], were systematically varied across a defined range. GridSearchCV with 5-fold cross-validation was employed to evaluate each parameter combination, with R^2 as the primary scoring metric. The optimal configuration identified through 5-fold cross-validation consisted of 180 trees [n_estimators=180], a maximum depth of 6, min_samples_split=2, and min_samples_leaf=1. The hyperparameter-tuned RF model preserved high predictive accuracy while improving computational efficiency and model stability, ensuring that predictions of Tg remain both interpretable and generalizable across a chemically diverse NADES space. Although RF provides built-in feature importance measures such as mean decrease in Gini impurity (86, 87) or permutation importance (88), these approaches are limited by their global scope, bias toward high-cardinality or continuous variables, and instability in correlated descriptor spaces, which are common in cheminformatics. To overcome these shortcomings, we employed SHAP approach (89), a framework that ensures mathematical consistency, delivers both global rankings and local, sample-specific contributions, and distributes importance more equitably across correlated descriptors. By directly linking descriptor effects to predicted Tg values at multiple levels, SHAP offers a more reliable and chemically meaningful interpretation of structure–property relationships. SHAP was therefore used to complement RF feature importance, providing unbiased insights that enhance mechanistic interpretability and strengthen confidence in the model's predictive performance across the chemically diverse NADES space.

3.4 Model Interpretability via SHAP Analysis

An additional layer of analysis was conducted using SHAP approach to deepen our understanding of the internal decision-making process of the optimized model. This approach was implemented not to assess robustness, but rather to provide a better understanding of the relative importance of the molecular descriptors selected in the ML model and to explain how they influence the model's predictions. By decomposing each prediction into additive contributions from individual features, SHAP values provide a transparent and quantitative interpretation of how specific chemical descriptors influence the predicted glass transition temperature. Representative outcomes from this interpretability analysis are illustrated in Fig.4.

Before engaging in a more in-depth interpretation of SHAP-based feature attributions, it is essential to recognize that several factors inherently bound the interpretability of any ML model. These include the model's architectural constraints (e.g., ensemble tree-based structure in RF), the nature of the input representation (in this case, applied molecular mixture-descriptors), and the specific chemical space spanned by the training dataset. Therefore, the SHAP values discussed herein should be understood as explanations tied to the Tg predictions within the scope of the NADES systems represented in this study. Importantly, the limited contribution or apparent absence of specific physicochemical descriptors traditionally considered relevant for glass transition phenomena does not imply that these features lack significance across broader or alternative molecular contexts. Instead, their marginal influence in our model may reflect dataset-specific patterns or descriptor redundancy within the confined chemical space.

With this framework in mind, a SHAP beeswarm plot (Fig. 4a) was generated to visualize the relative impact of the top 20 selected molecular descriptors on the predicted glass transition temperature of NADES. Among these, molecular topological descriptors such as *Mor08p, R6m+, MATS4i*, and *Mor26p* emerged as the most influential features, each displaying a distinctive effect on model output. In this context, descriptors with higher positive SHAP values are associated with an increased predicted Tg value, whereas those with negative values tend to lower the predicted Tg value. The feature importance ranking, Fig. 4c, corroborates these findings, indicating that a small subset of descriptors accounts for the majority of the model's predictive ability. Further, the individualized SHAP analyses (Fig. 4b, d) provide a detailed understanding of how specific molecular descriptors contribute to Tg predictions in individual NADES systems. For instance, descriptor *Mor08p* is a 3D-MoRSE descriptor weighted by polarizability, exhibits a strong positive contribution to Tg, indicating that spatially extended and polarizable molecular structures tend to elevate glass transition temperatures. At the same time, *R6m+* is a GETAWAY descriptor capturing mass distribution at topological lag 6, contributes negatively, suggesting that mass clustering at intermediate distances may reduce Tg. These trends are substantiated by SHAP dependence and force plots, which reveal both the directionality and magnitude of feature effects. A representative SHAP force plot, Fig. 4d, demonstrates the local interpretability of a single prediction: *Mor08p* contributed a significant positive shift (+0.572) toward the predicted Tg value, whereas *R6m+* had a minor negative effect (–0.025). Together, these feature contributions reduced the prediction from the model's base value (2.26) to the final output (2.23). This visualization reinforces the utility of SHAP in uncovering feature-specific impacts within the Random Forest framework, offering mechanistic insights into the structural underpinnings of Tg behavior in NADES. Additionally, the presented figure illustrates a SHAP decision plot, which visualizes the contribution of individual molecular descriptors to the predicted values of the glass transition temperature expressed as Tg. The x-axis represents the model's output (predicted values), while the y-axis

View Article Online
DOI: 10.1039/D6GC01000A

lists the most influential descriptors, ranked by their average SHAP importance. Each line on the plot corresponds to a single observation in the dataset, tracing how cumulative SHAP values (feature attributions) evolve from the model's base value to its final output for that instance. The color gradient encodes the normalized value of the respective descriptor: blue indicates lower values, whereas blue to pink denotes higher values.

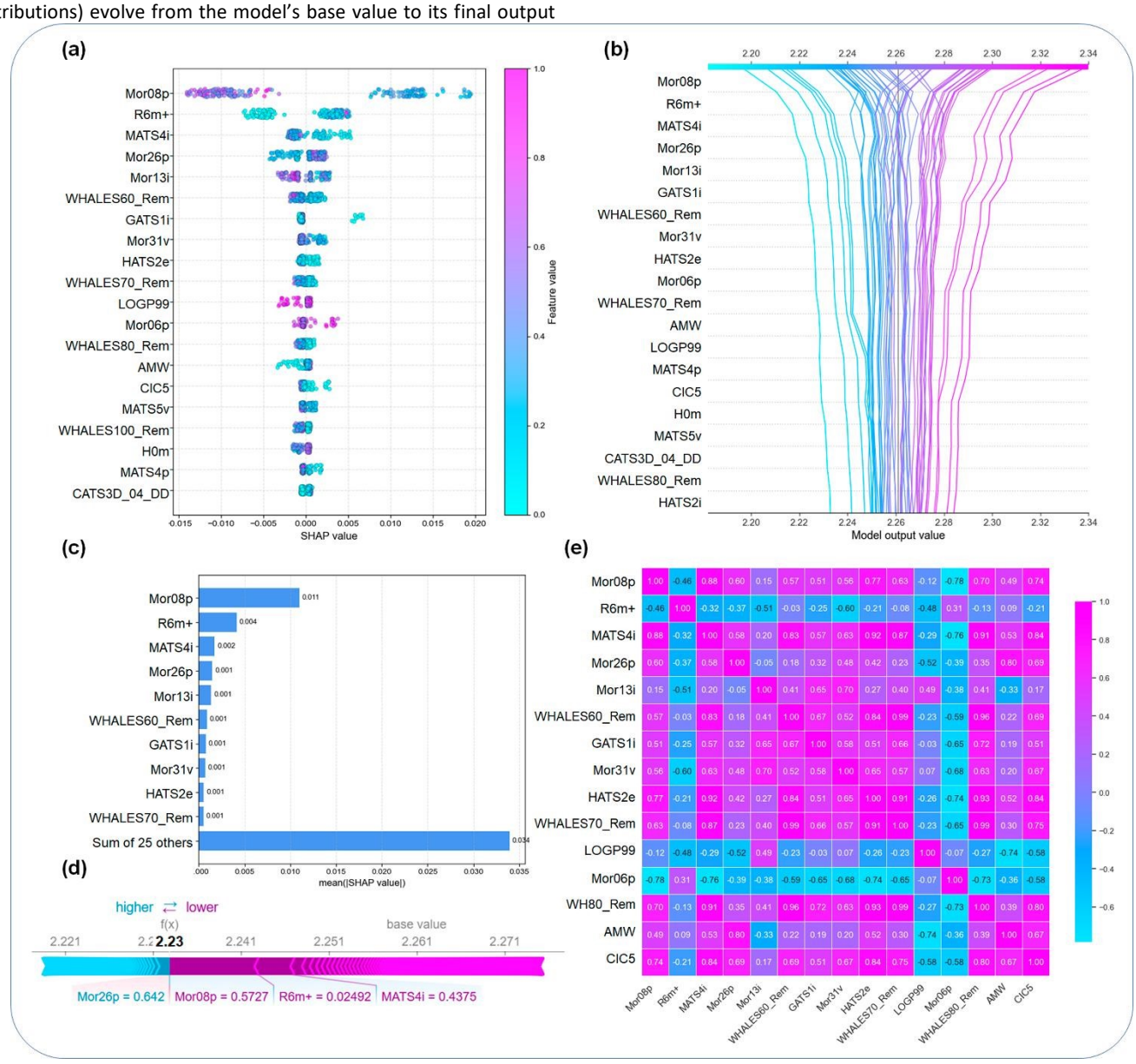


Fig. 4. SHAP-based interpretation and validation of the Random Forest model predicting the glass transition temperature of NADES. **a**) Feature contribution from SHAP values; **b**) Descriptor effects on model output; **c**) Global importance ranking highlighting *Mor08p, R6m+, MATS4i*, and *Mor26p*; **d**) Single-sample force plot; **e**) Pearson correlation matrix of top descriptors

Although the RF algorithm is inherently robust to multicollinearity and does not require strict linear independence among input features, where evaluated pairwise correlations remain essential for interpreting model's behavior. To this end, a Pearson correlation matrix was constructed for the top 15 descriptors ranked by SHAP importance, Fig. 4e. This visualization enables an assessment of feature redundancy and helps to identify potential duplication of encoded physicochemical information. For instance, descriptor pairs such as *MATS4i-HATS2e* and *WHALES60_Rem-WHALES80_Rem* exhibited strong cross-correlation (|r| > 0.85), indicating that they encode similar structural characteristics. In contrast, features like *Mor08p* and *R6m+* displayed weak correlation (r ≈ –0.46), underscoring low similarity and high structural diversity among the most informative variables. Overall, the analysis confirms the presence of both partially redundant and chemically unique descriptors, ensuring broad and diverse coverage of the NADES chemical space. This reinforces the interpretability of the model and highlights the value of integrating SHAP-based feature importance with correlation-based filtering in molecular modeling workflows.

View Article Online
DOI: 10.1039/D6GC01009A

These results highlight the model's strong generalization ability and stable predictive accuracy across diverse datasets (Fig.5). Moreover, the combination of rigorous regularization through optimized tree depth and node splitting, along with descriptor-level interpretability via SHAP, not only mitigates overfitting but also provides mechanistic insights, ensuring that complex nonlinear relationships within the descriptor space are captured in a transparent and chemically meaningful manner. To further validate the model's robustness and feature interpretability, we examined the distribution of the most influential descriptors across the training and test datasets. As shown in Fig. 3a-b, descriptors such as *Mor08p*, *R6m+*, *MATS4i*, and *Mor26p* exhibit consistent distributions between training and test sets, indicating good representativeness and minimal sampling bias. Specifically, *Mor08p* follows a near-normal distribution centered around 0.4 in both datasets, while *R6m+* displays a positively skewed distribution, reflecting its sparsity in the chemical space. Similar trends are observed for *MATS4i* and *Mor26p*, where the descriptors' values are well-aligned across the splits**.**

Although descriptors such as Mor08p and R6m+ do not explicitly encode hydrogen bonding, they capture structural features that indirectly influence it. Mor08p reflects the three-dimensional distribution of polarizability and is consistent with differences in molecular packing and supramolecular rigidity, whereas R6m+ describes medium-range mass distribution and can be associated with structural compactness and segmental mobility. Together, these effects are relevant to the stability of hydrogen-bond networks and the vitrification behavior of NADES.

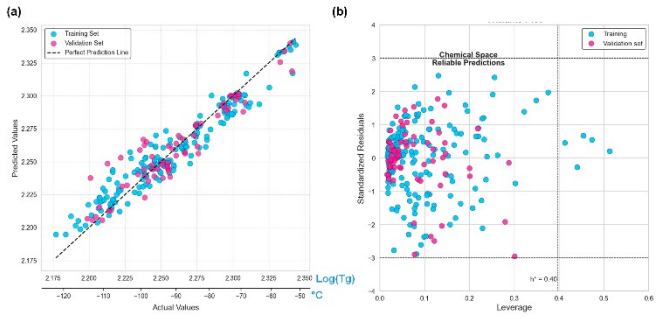


Fig. 5 Model performance and applicability domain for Tg prediction in NADES. **a**) Predicted vs. Experimental Log(1) values showing strong agreement; **b**) William's plot confirming applicability domain and reliability of predictions.

Thus, the discussed visualizations together link structural diversity to thermal behavior, thereby enhancing the mechanistic interpretability of the model outputs. These distributional consistencies bolster the generalization ability of the optimized RF model and ensure that SHAP-derived feature contributions are not merely the result of sampling bias. Moreover, the consistency between feature importance and descriptor availability across datasets enhances the interpretability and chemical relevance of the model, supporting its use in virtual screening and rational NADES design.

3.5 Detailed Descriptor Interpretation

The glass transition behavior of NADES exhibits clear compositional dependence across chemically distinct formulation regimes defined by the nature of the constituent components and their relative ratios**.** Fig.6 illustrates how the glass transition temperature of NADES varies across chemically distinct compositional regimes, as identified through UMAP-based clustering of the descriptor space and visualized using representative component systems. Rather than reflecting isolated descriptor effects, the observed trends primarily arise from differences in NADES composition, hydrogen-bond donor/acceptor balance, and water content, which collectively govern network rigidity and molecular mobility.

Clusters associated with higher Tg values are predominantly composed of sugar-rich NADES systems, such as sucrose–glucose–fructose–water (1:1:1:11) and glucose–fructose–water (1:1:8). These systems feature dense, multidirectional hydrogen-bonding networks formed by polyhydroxylated carbohydrates, leading to enhanced structural rigidity and reduced segmental mobility. The presence of multiple hydroxyl groups per molecule promotes extensive intermolecular interactions, which stabilizes the amorphous phase and shifts Tg toward higher values. In these compositions, water acts as a plasticizer at low concentrations but becomes structurally integrated into the hydrogen-bonded network at higher ratios, contributing to network stabilization rather than disruption.

In contrast, clusters characterized by lower Tg values are enriched in amino acid– and choline-based NADES, such as proline–ethylene glycol (1:3.3) and choline chloride–glycerol (1:2). These systems exhibit more flexible hydrogen-bonding motifs, dominated by fewer donor–acceptor sites and increased conformational freedom of aliphatic chains. Ethylene glycol and glycerol introduce mobility through rotational flexibility, while zwitterionic or ionic components such as proline and choline chloride generate localized interactions that are less effective in producing rigid, three-dimensional networks. As a result, these formulations display lower resistance to molecular rearrangement and reduced glass transition temperatures.

Intermediate clusters, including mixed systems such as glycerol–glucose–sorbitol–water (1:1:1:3), exhibit Tg values between the two extremes. These compositions combine rigid carbohydrate backbones with flexible polyol or amino-alcohol components, yielding partially constrained hydrogen-bonded networks. The coexistence of rigid and flexible molecular motifs produces heterogeneous interaction landscapes, reflected in broader Tg distributions and cluster overlap. This behavior highlights the non-additive nature of Tg in multicomponent eutectic systems, where collective interactions dominate over individual component properties.

Importantly, the cluster-resolved analysis demonstrates that Tg in NADES is governed primarily by component chemistry and formulation strategy, rather than by isolated molecular descriptors. The machine-learning descriptors serve as an effective abstraction of these chemical features—capturing size, polarity, and interaction density—but the underlying physical origin of Tg variation remains rooted in hydrogen-bond topology, molecular flexibility, and water-mediated plasticization effects.

View Article Online
DOI: 10.1039/D6GC01009A

From a green chemistry perspective, these findings provide chemically intuitive design guidelines for tailoring Tg in NADES. Carbohydrate-rich, highly hydrogen-bonded systems are suitable for applications requiring elevated Tg and enhanced thermal stability, such as cryopreservation or solid-state encapsulation. Conversely, polyol- and amino acid–based NADES offer lower Tg and greater flexibility, making them attractive for low-temperature processing, coatings, and solvent applications. By linking machine-learning insights to concrete compositional motifs, this study advances the rational, sustainability-driven design of NADES with application-specific thermophysical properties.

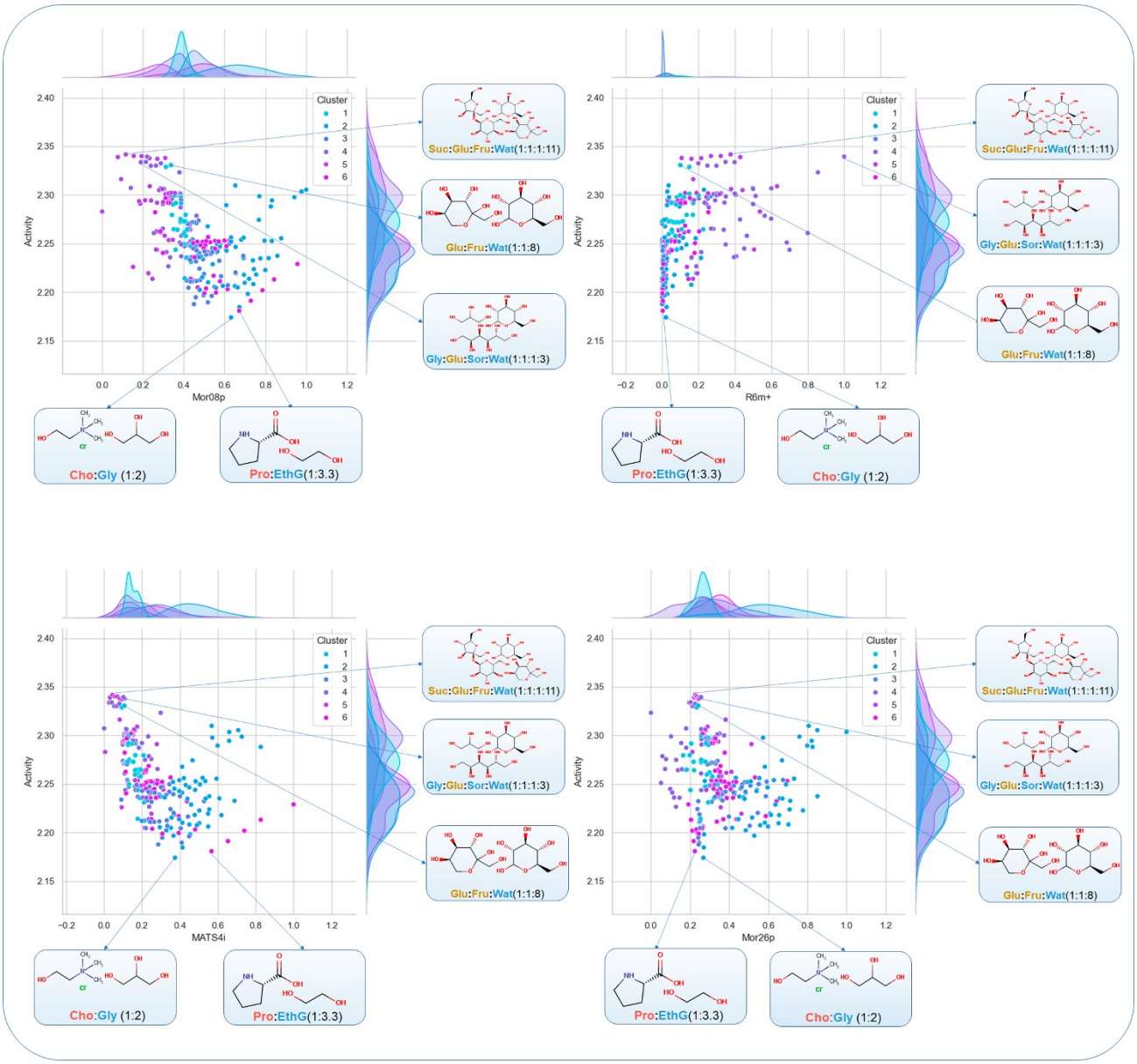


Fig. 6 Cluster-resolved distribution of key molecular descriptors and their influence on glass transition temperature. Bivariate density plots showing the relationship between Tg and four highly influential descriptors — *Mor08p, R6m⁺, MATS4i*, and *Mor26p,* colored by KMeans clusters derived from UMAP projections of descriptor space

3.6 Implications for Process and Formulation Design

Accurate prediction of Tg enables the rational selection of NADES formulations for temperature-sensitive processes, such as cryopreservation, anti-icing, and sub-zero storage. By identifying compositions with Tg values safely below operating temperatures, the proposed ML framework allows prescreening of mechanically stable, amorphous systems without extensive experimental trial-and-error.

In practical formulation design, Tg represents only one of several parameters governing the performance of NADES systems. Other properties, such as viscosity, toxicity, conductivity, cost, and solubilization capacity, may also strongly influence the suitability of a formulation for specific applications. Although the present study

View Article Online
DOI: 10.1039/D5GC01009A

focuses on Tg prediction, the proposed descriptor-based machine learning framework is readily extendable to multi-property modelling. By combining Tg prediction with existing ML models for the viscosity, density, or surface tension of deep eutectic solvents, multi-objective screening strategies could be implemented to support rational NADES design. Such approaches would enable the identification of formulations that simultaneously satisfy multiple performance constraints, thereby accelerating the development of sustainable functional materials.

On the other hand, Tg must often be balanced against viscosity and water tolerance. Descriptor-level interpretation further indicates that molecular geometry, polarizability, and mass distribution simultaneously influence Tg and molecular mobility, providing guidance for navigating trade-offs between thermal stability, flow behavior, and hydration sensitivity. These insights support the development of multi-objective formulation strategies rather than single-property optimization.

From a process perspective, virtual Tg screening significantly reduces reliance on low-throughput DSC measurements by narrowing experimental validation to high-confidence candidates. As a result, the framework accelerates NADES formulation workflows while reducing material consumption and experimental cost.

3.6.1 Practical workflow

As a practical illustration, the model can be used to guide the design of NADES formulations targeting a Tg around −80 °C for cryopreservation-related applications. A researcher may first examine experimentally characterized systems in the dataset and identify formulations whose Tg values fall within or below the desired temperature range. The distribution of Tg values across related NADES families, including hydrated systems (Fig. 3), provides an initial empirical guide for selecting promising compositional motifs and hydration regimes.

If the required components are experimentally accessible, these formulations may be directly prioritized for preparation and validation. Otherwise, structurally related components, such as alternative sugars, polyols, amino acids, or quaternary ammonium derivatives, can be used to construct a virtual library of candidate systems across feasible molar ratios and water contents. For each candidate formulation, mixture descriptors are calculated using the same stoichiometry-weighted protocol applied in this study, and the trained Random Forest model is then used to predict Tg.

To ensure reliable interpretation, predicted candidates should also be evaluated within the applicability domain of the model, meaning that descriptor values remain within or close to the range represented in the training dataset. In this way, the model functions as a prescreening tool that reduces the number of experimental trials and accelerates the rational identification of NADES formulations with targeted low-temperature properties.

3.7 Limitations and transferability

Despite the strong predictive performance of the proposed machine-learning framework, several limitations should be acknowledged. First, the model is trained on a finite dataset of experimentally reported NADES systems and therefore reflects the chemical diversity, compositional ranges, and water contents represented in the available data. As a result, predictions for formulations that lie far outside the explored chemical space, such as systems involving unconventional hydrogen-bond donors or extreme stoichiometries, should be interpreted with caution. As noted in the supplementary data (Table S2), the Tg data used to train the ML models were based on single data points for each composition at each dilution. However, we have found that the Tg for NADES measured using the method described in this study have a low degree of variability (18, 68). This experience gives us confidence in the reproducibility of the Tg measurements.

Second, the descriptor-based representation provides an effective abstraction of component chemistry but does not explicitly encode specific intermolecular interactions or dynamic hydrogen-bond rearrangements. Consequently, the model captures statistically learned structure–property relationships rather than explicit mechanistic pathways governing glass transition behavior. While SHAP analysis improves interpretability, it does not establish causal links between individual descriptors and Tg.

Regarding transferability, the applicability of the model is expected to be highest for NADES composed of chemically related components and comparable formulation strategies, particularly polyol-, carbohydrate-, and amino acid–based systems. The applicability domain analysis further supports this view by identifying regions of descriptor space in which predictions are reliable. Extension of the model to fundamentally different eutectic systems or to non-natural DES formulations would require retraining or recalibration using representative experimental data.

Nevertheless, the overall modeling strategy, combining structure-aware data splitting, ensemble learning, and interpretability analysis, is transferable and can be readily adapted to other thermophysical properties or solvent classes. With the continued expansion of high-quality experimental datasets, the predictive scope and generalizability of such data-driven frameworks are expected to improve, supporting the rational and sustainable design of next-generation eutectic materials.

Although the present model was evaluated using a structure-aware held-out test set, prospective validation on newly prepared NADES formulations remains an important next step. Such validation would further establish the practical utility of the framework for formulation design and targeted screening.

Conclusions

In this work, we combined an in-house dataset with a data-driven framework to develop an ML model for predicting the Tg of NADES. A unique, internally consistent dataset of 263 NADES systems with Tg values ranging from -122.9 to -52.5 °C, measured under identical DSC protocols, was used to ensure reliable model training and validation. A stoichiometry-weighted combination of single-component descriptors generated mixture-specific molecular descriptors. UMAP-based clustering (k = 6) was used to guide structure-aware train-test splitting, preventing data leakage and ensuring evaluation on chemically distinct NADES families.

Among the machine-learning algorithms tested, the RF model achieved the best balance between accuracy and generalization, with R^2 = 0.93 (training) and R^2 = 0.87 (test), whereas advanced

boosting models showed higher overfitting. SHAP analysis identified a small subset of descriptors dominating Tg prediction, with the most influential being related to 3D molecular geometry, polarizability, atomic mass distribution, and electronic effects, directly linking Tg to molecular mobility and hydrogen-bond network rigidity.

Overall, this study demonstrates that Tg of NADES can be predicted with high accuracy using mixture-aware descriptors and interpretable machine learning. The proposed framework enables rapid prescreening and rational formulation of NADES with targeted thermal properties, reducing experimental effort and supporting scalable design of green solvents for cryopreservation, anti-icing, and materials applications.

Author contributions

D.U. - Writing – original draft; D.U., G.C-M, B.R. - Software; D.U., G.C-M., A.M., P.Y., S.S., A.H., B.R. - Methodology; D.U., G.C-M., A.H. B.R. - Conceptualization; D.U., G.C-M., A.M., P.Y., S.S., A.H., B.R. - Writing – review and editing; A.H., B.R. - Supervision; A.H., B.R. - Resources; A.H., B.R. - Funding acquisition.

Conflicts of interest

No competing interests to declare

Data availability

The experimental dataset and data used to develop the machine learning models shown in Supplementary Information (SI) file 1 and Supplementary Information (SI) file 2. The dataset is curated and listed in the SI. The authors declare that all data supporting the findings and those used for reproducing the figures in this paper are available within the paper and its Supplementary Information. Source data are provided with this paper.

Acknowledgements

Authors thank the Characterization Facility, University of Minnesota, where most of the experiments were carried out. Also, authors acknowledge the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University, resources of which were used. This was made possible in part by National Science Foundation (90), MRI Award number 2019077 (Lead PI – B. Rasulev). Supercomputing support provided by the CCAST HPC System at NDSU is gratefully acknowledged.

Funding

A.H and B.R. thank the U.S. DARPA for the support through the Award Number: HR001124C0315. A.H. receives partial support for this work from the NSF through the MRSEC (Award No. DMR-2011401) and the NNCI (Award No. ECCS-2025124) programs. This work was made possible in part by NSF MRI Award No. 2019077 (B.R. – supercomputing facility).

References

View Article Online
DOI: 10.1039/D6GC01009A

1. Yetgin A. Revolutionizing multi-omics analysis with artificial intelligence and data processing. Quantitative Biology. 2025;13(3).
2. Dai YT, van Spronsen J, Witkamp GJ, Verpoorte R, Choi YH. Natural deep eutectic solvents as new potential media for green technology. Anal Chim Acta. 2013;766:61-8.
3. Sharma A, Lee B-S. Toxicity test profile for deep eutectic solvents: A detailed review and future prospects. Chemosphere. 2024;350:141097.
4. 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.
5. Craveiro R, Aroso I, Flammia V, Carvalho T, Viciosa MT, Dionísio M, et al. Properties and thermal behavior of natural deep eutectic solvents. J Mol Liq. 2016;215:534-40.
6. Spaggiari C, Carbonell-Rozas L, Zuilhof H, Costantino G, Righetti L. Structural elucidation and long-term stability of synthesized NADES: A detailed physicochemical analysis. J Mol Liq. 2025;424:127105.
7. Aguilar N, Benito C, Martel-Martín S, Gutiérrez A, Rozas S, Marcos PA, et al. Insights into Carvone: Fatty Acid Hydrophobic NADES for Alkane Solubilization. Energ Fuel. 2024;38(24):23633-53.
8. Arunachalam A, Oosterhoff T, Breet I, Dijkstra P, Yunus RAM, Parisi D, et al. Harnessing the bio-adhesive power of natural deep eutectic solvents for trichome-inspired pest control. Commun Mater. 2025;6(1).
9. Wang BR, Lin L, Zhang WX, Zhu CF, Ren SH, Hou YC, et al. Specific heat capacities of deep eutectic solvents applicable for absorbing sulfur dioxide. Sep Purif Technol. 2025;370.
10. Boiteux J, Espino M, Azcarate S, Silva MF, Gomez FJV, Pizzuolo P, et al. NADES blend for bioactive coating design as a sustainable strategy for postharvest control. Food Chem. 2023;406.
11. Freitas DS, Rocha D, Santos J, Noro J, Tavares TD, Teixeira MO, et al. NADES-in-Oil Emulsions Enriched with Essential Oils for Cosmetic Application. Processes. 2025;13(2).
12. Li H, Yang KN, Yang YM, Ding LQ, Li XX. Natural deep eutectic solvents (NADES) in drug delivery systems: Characteristics, applications, and future perspectives. Int J Pharmaceut. 2025;675.
13. Aschemacher NA, Teglia CM, Siano Á S, Gutierrez FA, Goicoechea HC. Development of an electrochemical sensor using carbon nanotubes and hydrophobic natural deep eutectic solvents for the detection of α-glucosidase activity in extracts of autochthonous medicinal plants. Talanta. 2024;268(1):125313.
14. Boldrini CL, Manfredi N, Perna FM, Capriati V, Abbotto A. Eco-Friendly Sugar-Based Natural Deep Eutectic Solvents as Effective Electrolyte Solutions for Dye-Sensitized Solar Cells. ChemElectroChem. 2020;7(7):1707-12.
15. Julião D, Xavier M, Mascarenhas X. Deep eutectic solvents: viable sustainable electrolytes for supercapacitors. Materials Today Energy. 2024;42:101432.
16. Zhang C, Ding Y, Zhang L, Wang X, Zhao Y, Zhang X, et al. A Sustainable Redox-Flow Battery with an Aluminum-Based, Deep-Eutectic-Solvent Anolyte. Angewandte Chemie International Edition. 2017;56(26):7454-9.
17. Li X, Li J-Y, Manzoor MF, Lin Q-Y, Shen J-l, Liao L, et al. Natural deep eutectic solvent: A promising eco-friendly food bio-inspired antifreezing. Food Chem. 2024;437:137808.
18. Mallya AS, Yadav P, Zakhia S, Hubel A. Low-temperature Characterization of Sugar Alcohol-based Type V Deep Eutectic Solvents for Anti-icing and Cryopreservation Applications. 2025.

19. Hornberger K, Li R, Duarte ARC, Hubel A. Natural deep eutectic systems for nature-inspired cryopreservation of cells. AIChE J. 2021;67(2).
20. Fu MT, Zhang HM, Bai JL, Cui MT, Liu ZY, Kong XJ, et al. Application of Deep Eutectic Solvents with Modern Extraction Techniques for the Recovery of Natural Products: A Review. Acs Food Sci Technol. 2025;5(2):444-61.
21. Gupta V, Thakur R, Das AB. Effect of natural deep eutectic solvents on thermal stability, syneresis, and viscoelastic properties of high amylose starch. Int J Biol Macromol. 2021;187:575-83.
22. Tian Y, Sun D-W, Xu L, Fan T-H, Zhang S-T, Zhu Z. Bioinspired Cryoprotectants Enabled by Binary Natural Deep Eutectic Solvents for Sustainable and Green Cryopreservation. Acs Sustain Chem Eng. 2022;10(23):7677-91.
23. Craveiro R, Castro VIB, Viciosa MT, Dionísio M, Reis RL, Duarte ARC, et al. Influence of natural deep eutectic systems in water thermal behavior and their applications in cryopreservation. J Mol Liq. 2021;329:115533.
24. Kauzmann W. The Nature of the Glassy State and the Behavior of Liquids at Low Temperatures. Chemical Reviews. 1948;43(2):219-56.
25. Bojic S, Murray A, Bentley BL, Spindler R, Pawlik P, Cordeiro JL, et al. Winter is coming: the future of cryopreservation. BMC Biology. 2021;19(1):56.
26. Sansinena M, Santos MV, Taminelli G, Zaritky N. Implications of storage and handling conditions on glass transition and potential devitrification of oocytes and embryos. Theriogenology. 2014;82(3):373-8.
27. Kavian S, Sellers R, Sanchez GA, Alvarez C, Aguilar G, Powell-Palm MJ. Higher glass transition temperatures reduce thermal stress cracking in aqueous solutions relevant to cryopreservation. Sci Rep-Uk. 2025;15(1):27903.
28. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. npj Digital Medicine. 2020;3(1).
29. Karuth A, Alesadi A, Xia W, Rasulev B. Predicting glass transition of amorphous polymers by application of cheminformatics and molecular dynamics simulations. Polymer. 2021;218:123495.
30. Casanola-Martin GM, Karuth A, Pham-The H, González-Díaz H, Webster DC, Rasulev B. Machine learning analysis of a large set of homopolymers to predict glass transition temperatures. Communications Chemistry. 2024;7(1):226.
31. Ayres LB, Gomez FJV, Silva MF, Linton JR, Garcia CD. Predicting the formation of NADES using a transformer-based model. Sci Rep-Uk. 2024;14(1).
32. Xiao X, Kong D, Qiu X, Zhang W, Zhang F, Liu L, et al. Shape-Memory Polymers with Adjustable High Glass Transition Temperatures. Macromolecules. 2015;48(11):3582-9.
33. Madadi M, Kargaran E, Hashemi SS, Sun C, Denayer JFM, Karimi K, et al. Scalable Lignin Monomer Production Via Machine Learning-Guided Reductive Catalytic Fractionation of Lignocellulose. Advanced Science. 2025;12(42):e10496.
34. Madadi M, Kargaran E, Al Azad S, Saleknezhad M, Zhang E, Sun F. Machine learning-driven optimization of biphasic pretreatment conditions for enhanced lignocellulosic biomass fractionation. Energy. 2025;326:136241.
35. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature. 2023;620(7972):47-60.
36. Kaspar C, Ravoo BJ, van der Wiel WG, Wegner SV, Pernice WHP. The rise of intelligent matter. Nature. 2021;594(7863):345-55.
37. Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari TP, Michalkova A, et al. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nat Nanotechnol. 2011;6(3):175-8.
38. Pourmousa M, Jain S, Barnaeva E, Jin W, Hochuli J, Itkin Z, et al. AI-driven discovery of synergistic drug combinations against pancreatic cancer. Nat Commun. 2025;16(1):4020.
39. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev. 2020;49(11):3525-64.
40. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Science Advances. 2018;4(7):eaap7885.
41. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model. 2010;50(7):1189-204.
42. Toropov AA, Rasulev BF, Leszczynski J. QSAR modeling of acute toxicity by balance of correlations. Bioorg Med Chem. 2008;16(11):5999-6008.
43. Erickson ME, Ngongang M, Rasulev B. A Refractive Index Study of a Diverse Set of Polymeric Materials by QSPR with Quantum-Chemical and Additive Descriptors. Molecules. 2020;25(17):3772.
44. Sizochenko N, Rasulev B, Gajewicz A, Mokshyna E, Kuz'min VE, Leszczynski J, et al. Causal inference methods to assist in mechanistic interpretation of classification nano-SAR models. Rsc Adv. 2015;5(95):77739-45.
45. Tran H, Gurnani R, Kim C, Pilania G, Kwon H-K, Lively RP, et al. Design of functional and sustainable polymers assisted by artificial intelligence. Nature Reviews Materials. 2024;9(12):866-86.
46. Gregoire JM, Zhou L, Haber JA. Combinatorial synthesis for AI-driven materials discovery. Nature Synthesis. 2023;2(6):493-504.
47. Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY-J. Explainable machine learning in materials science. npj Computational Materials. 2022;8(1):204.
48. Ahmed L, Rasulev B, Kar S, Krupa P, Mozolewska MA, Leszczynski J. Inhibitors or toxins? Large library target-specific screening of fullerene-based nanoparticles for drug design purpose. Nanoscale. 2017;9(29):10263-76.
49. Zhuravskyi Y, Iduoku K, Erickson ME, Karuth A, Usmanov D, Casanola-Martin G, et al. Quantitative Structure-Permittivity Relationship Study of a Series of Polymers. ACS Mater Au. 2024;4(2):195-203.
50. Daghighi A, Casanola-Martin GM, Iduoku K, Kusic H, González-Díaz H, Rasulev B. Multi-Endpoint Acute Toxicity Assessment of Organic Compounds Using Large-Scale Machine Learning Modeling. Environ Sci Technol. 2024;58(23):10116-27.
51. Karuth A, Szwiec S, Casanola-Martin GM, Khanam A, Safaripour M, Boucher D, et al. Integrated machine learning, computational, and experimental investigation of compatibility in oil-modified silicone elastomer coatings. Prog Org Coat. 2024;193:108526.
52. Lavrinenko AK, Chernyshov IY, Pidko EA. Machine Learning Approach for the Prediction of Eutectic Temperatures for Metal-Free Deep Eutectic Solvents. Acs Sustain Chem Eng. 2023;11(42):15492-502.
53. Khajeh A, Shakourian-Fard M, Parvaneh K. Quantitative structure-property relationship for melting and freezing points of deep eutectic solvents. J Mol Liq. 2021;321:114744.
54. 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-Uk. 2022;12(1):4954.
55. Ayres LB, Bandara M, McMillen CD, Pennington WT, Garcia CD. eutXG: A Machine-Learning Model to Understand and Predict the

Melting Point of Novel X-Bonded Deep Eutectic Solvents. Acs Sustain Chem Eng. 2024;12(30):11260-73.
56. Odegova V, Lavrinenko A, Rakhmanov T, Sysuev G, Dmitrenko A, Vinogradov V. DESignSolvents: an open platform for the search and prediction of the physicochemical properties of deep eutectic solvents. Green Chem. 2024;26(7):3958-67.
57. Mohan M, Jetti KD, Smith MD, Demerdash ON, Kidder MK, Smith JC. Accurate Machine Learning for Predicting the Viscosities of Deep Eutectic Solvents. Journal of Chemical Theory and Computation. 2024;20(9):3911-26.
58. Joules A, Burrows T, I. Dosa P, Hubel A. Characterization of eutectic mixtures of sugars and sugar-alcohols for cryopreservation. J Mol Liq. 2023;371:120937.
59. Hunter AD. ACD/ChemSketch 1.0 (freeware); ACD/ChemSketch 2.0 and its tautomers, dictionary, and 3D plug-ins; ACD/HNMR 2.0; ACD/CNMR 2.0. ACS Publications; 1997.
60. Mauri A. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs: Springer; 2020. p. 801-20.
61. Mahini RA, Casanola-Martin G, Ludwig SA, Rasulev B. MixtureMetrics: A comprehensive package to develop additive numerical features to describe complex materials for machine learning modeling. SoftwareX. 2024;28:101911.
62. Mahini RA, Casanola-Martin G, Szwiec S, Ludwig SA, Rasulev B. CombinatorixPy: Advancing mixture descriptors for computational chemistry. SoftwareX. 2025;29:102060.
63. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020;180(4):688-702.e13.
64. Tanoli Z, Fernández-Torras A, Özcan UO, Kushnir A, Nader KM, Gadiya Y, et al. Computational drug repurposing: approaches, evaluation of in silico resources and case studies. Nature Reviews Drug Discovery. 2025;24(7):521-42.
65. Castro VIB, Craveiro R, Silva JM, Reis RL, Paiva A, AR CD. Natural deep eutectic systems as alternative nontoxic cryoprotective agents. Cryobiology. 2018;83:15-26.
66. Dai Y, van Spronsen J, Witkamp GJ, Verpoorte R, Choi YH. Natural deep eutectic solvents as new potential media for green technology. Anal Chim Acta. 2013;766:61-8.
67. Savi LK, Dias MCGC, Carpine D, Waszczynskyj N, Ribani RH, Haminiuk CWI. Natural deep eutectic solvents (NADES) based on citric acid and sucrose as a potential green technology: a comprehensive study of water inclusion and its effect on thermal, physical and rheological properties. Int J Food Sci Tech. 2019;54(3):898-907.
68. Mallya AS, Yadav P, Zakhia S, Hubel A. Computational Design of Natural Deep Eutectic Systems Using COSMO-RS for Ice Control Applications. Acs Sustain Chem Eng. 2025;13(36):14683-92.
69. Joardder MUH, Bosunia MH, Hasan MM, Ananno AA, Karim A. Significance of Glass Transition Temperature of Food Material in Selecting Drying Condition: An In-Depth Analysis. Food Reviews International. 2024;40(3):952-73.
70. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. The Journal of Open Source Software. 2018;3(29):861.
71. Rugard M, Jaylet T, Taboureau O, Tromelin A, Audouze K. Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages. PLoS One. 2021;16(5):e0252486.
72. Moshkov N, Becker T, Yang K, Horvath P, Dancik V, Wagner BK, et al. Predicting compound activity from phenotypic profiles and chemical structures. Nat Commun. 2023;14(1):1967.
73. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65.
74. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979;PAMI-1(2):224-7.
75. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
76. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018:6639-49.
77. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
78. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics. 2013;Volume 7 - 2013.
79. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30.
80. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and computing. 2004;14(3):199-222.
81. Zhang Z. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine. 2016;4(11):218.
82. Williams C, Rasmussen C. Gaussian processes for regression. Advances in neural information processing systems. 1995;8.
83. Grossi E, Buscema M. Introduction to artificial neural networks. European journal of gastroenterology & hepatology. 2007;19(12):1046-54.
84. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996;58(1):267-88.
85. Zengin G, Ceylan R, Katanic J, Aktumsek A, Matic S, Boroja T, et al. Exploring the therapeutic potential and phenolic composition of two Turkish ethnomedicinal plants – Ajuga orientalis L. and Arnebia densiflora (Nordm.) Ledeb. Ind Crop Prod. 2018;116:240-8.
86. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics. 2009;10(1):213.
87. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711-8.
88. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340-7.
89. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
90. Li N, Lewin A, Ning S, Waito M, Zeller MP, Tinmouth A, et al. Privacy-preserving federated data access and federated learning: Improved data sharing and AI model development in transfusion medicine. Transfusion. 2025;65(1):22-8.

View Article Online
DOI: 10.1039/D6GC01009A

Data availability

The experimental dataset and data used to develop the machine learning models shown in Supplementary Information (SI) file 1 and Supplementary Information (SI) file 2. The dataset is curated and listed in the SI. The authors declare that all data supporting the findings and those used for reproducing the figures in this paper are available within the paper and its Supplementary Information. Source data are provided with this paper.