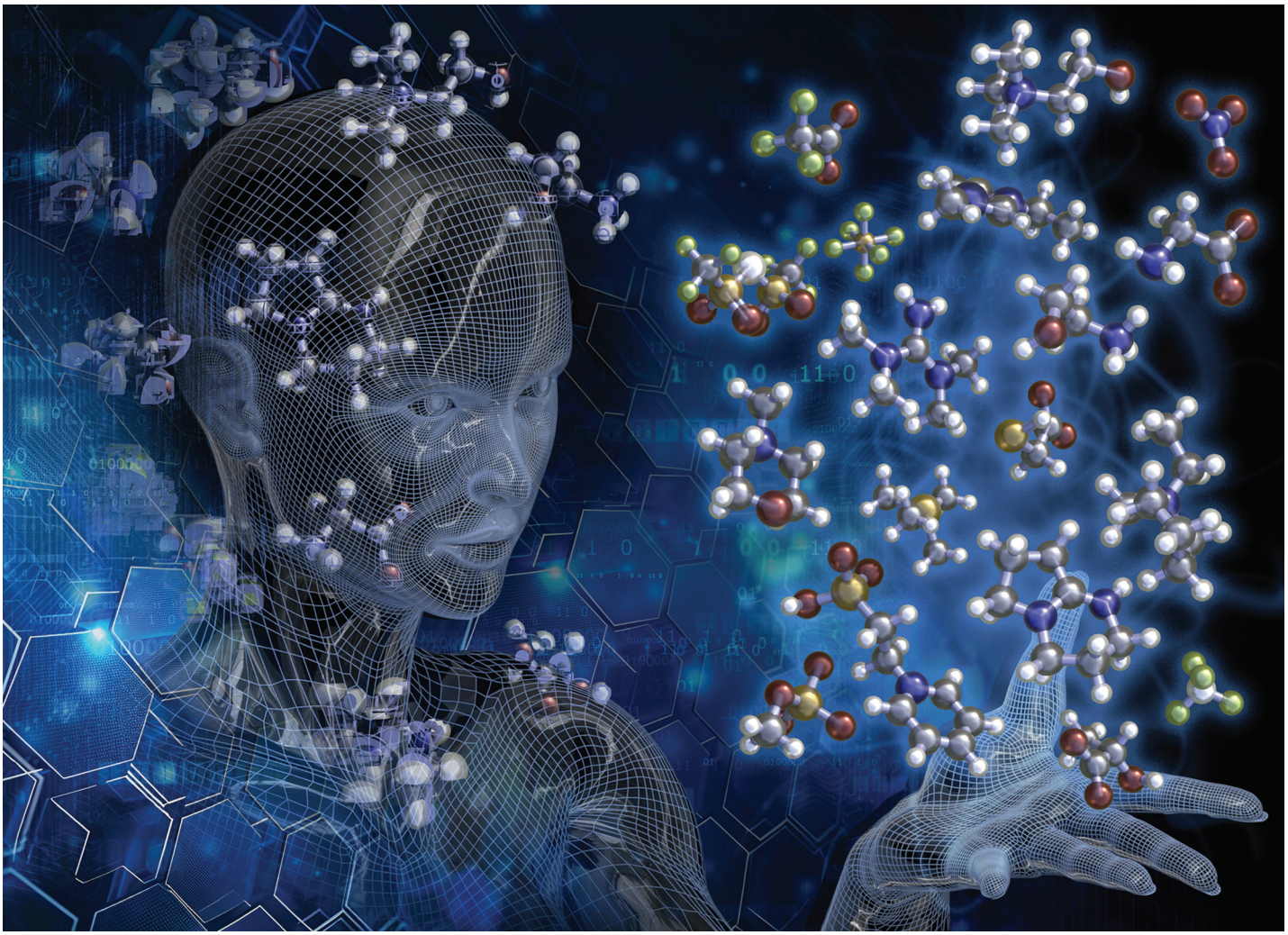


Showcasing research from the Biofuels Science Focus Area team, Oak Ridge National Laboratory, Tennessee, USA, and Prof. Jeremy C. Smith, Center for Molecular Biophysics, Oak Ridge National Laboratory, TN, USA.

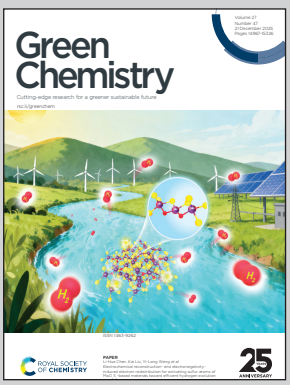
Molecular property prediction for very large databases with natural language processing: a case study in ionic liquid design

Machine learning with natural language processing–based molecular embeddings is used to represent ionic liquid structures, enabling accurate prediction of key physicochemical properties. This approach allows rapid screening of more than ten million ionic liquids to identify task-specific candidates for applications including biomass processing, CO_2 capture, and battery electrolytes, accelerating the design of efficient and sustainable solvents.

Cover artwork created by Andrew Sproles.

Image reproduced by permission of Andrew Sproles from *Green Chem.*, 2025, **27**, 15106.

As featured in:



See Mood Mohan, Jeremy C. Smith *et al.*, *Green Chem.*, 2025, **27**, 15106.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/greenchem

Registered charity number: 207890

Green Chemistry

PAPER

View Article Online
View Journal | View Issue

Cite this: *Green Chem.*, 2025, **27**, 15106

Molecular property prediction for very large databases with natural language processing: a case study in ionic liquid design

Mood Mohan, *[a] Michelle K. Kidder [c] and Jeremy C. Smith*[a,b]

Received 4th June 2025,
Accepted 27th October 2025

DOI: 10.1039/d5gc02803e

rsc.li/greenchem

The prospect of using artificial intelligence (AI) to accurately screen very large databases of compounds for multiple properties has yet to be realized. Here, we explore this possibility using ionic liquids (ILs) which offer unique physicochemical properties and excellent tunability, making them highly versatile solvents for various research applications. Screening millions of potential ILs for the best perfomance for use in specific tasks with experimental methods alone however, is impractical. Further, traditional' physics-based computational chemistry is hindered by high computational cost. To address this challenge, we leverage a natural language processing (NLP)-based molecular embedding technique with advanced machine learning (ML) models to predict seven key IL properties: viscosity, density, ionic conductivity, surface tension, melting temperature, toxicity, and water solubility. Comprehensive datasets for these properties are obtained, then NLP featurization with Mol2vec is compared with other featurization techniques such as 2D Morgan fingerprints, and 3D quantum chemistry-derived sigma profiles. NLP-based featurization exhibited the best predictive performance, achieving the highest R^2 and lowest RMSE values for all the studied IL properties. Further, we present case studies of how ILs might be screened using combined property criteria for practical cases – lignocellulosic biomass processing, CO_2 capture, and optimal electrolytes for batteries – screening a novel database of ~10.6 million generated feasible ILs. The results introduce NLP as a powerful tool for engineering many designer solvents with desirable properties for task specific applications.

Green foundation

1. Our work advances green chemistry by using AI and natural language processing to predict critical properties of ionic liquids, enabling rapid screening of over 10.6 million candidates. This avoids time- and resource-intensive experimental testing, supporting safer solvent design with minimal environmental impact. The approach identifies task-specific ILs for CO_2 capture, biomass processing, and reducing waste, energy use, and hazardous materials in solvent discovery.
2. We predicted seven key properties of ILs and identified optimal candidates using AI/ML. Our model achieved $R^2 > 0.9$ for most properties and revealed strong correlations between IL properties, enabling selection of safer solvents. This significantly reduces chemical waste, resource use, and trial-and-error experimentation, making solvent discovery faster, cleaner, and more sustainable.
3. The work could be greener by integrating metrics such as biodegradability, synthetic accessibility, and life cycle impacts into the screening pipeline. Future models could prioritize ILs from renewable feedstocks and include ecosystem toxicity data. Integrating this framework with experimental validation workflows would enable a closed-loop, AI-guided platform for sustainable solvent discovery, further aligning with green chemistry principles.

Introduction

Artificial intelligence (AI) raises the prospect of accurate fine-detailed exploration of large chemical spaces for specific desired combinations of key properties. Ionic liquids (ILs) provide an excellent trial system for such computational approaches. ILs are molten salts composed of organic cations and organic/inorganic anions with a melting temperature lower than 100 °C,[1,2] and exhibit several attractive properties, including high chemical and thermal stabilities, nonvolatility, a wide liquidus range, high

[a]*Biosciences Division and Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. E-mail: moodm@ornl.gov, smithjc@ornl.gov*
[b]*Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996, USA*
[c]*Manufacturing Science Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6201, USA*

ionic conductivity, and nonflammability.[3,4] ILs have been extensively investigated over the past few years as promising alternatives to conventional solvents and have been considered for various industrial applications (such as in electrochemistry, separation science, biomass processing, carbon capture, *etc.*).[5–9] Perhaps the most attractive feature of ILs is their tunability, which provides significant design flexibility, enabling the tailoring of ILs to meet specific applications.[5,10] However, prohibitive time and cost would be required for experimentally evaluating large numbers of different cation–anion combinations. Consequently, developing efficient and accurate tools for predicting IL properties is highly desirable.[11,12]

Over the past decades, researchers have developed a range of computational methods to predict the thermodynamics properties of ILs. These methods include equations of state (EoS) (PC-SAFT: perturbed-chain statistical associating fluid theory),[13] group contribution (GC) methods,[14] the conductor-like screening model for real solvents (COSMO-RS),[15] other quantum chemistry (QC) approaches,[16] and molecular dynamics (MD) simulations.[17–19] Among these, EoS methods possess a strong theoretical foundation for calculating thermodynamic properties, but their application is often limited by the complexity of estimating the required model parameters.[20] QC and MD calculations provide detailed atomistic and molecular insights into the behavior of ILs, but their high computational costs restrict their use in large-scale screening.[18,19,21] Further, COSMO-RS framework has emerged as a versatile quantum chemical method for calculating thermodynamic properties of ILs and deep eutectic solvents (DESs), but generating COSMO files remains computationally demanding and can sometimes yield only qualitative predictions rather than quantitative.[11,19,22]

To address the above issues, quantitative structure–property relationship (QSPR) models have emerged as a powerful tool for molecular property prediction, leveraging data-driven techniques to uncover complex patterns and relationships within the molecular datasets and driven by advances in AI and machine learning (ML).[23–26] The award of the 2024 Nobel Prize in Chemistry for the AlphaFold methods underscores the transformative potential of AI/ML across scientific disciplines, particularly in chemistry and materials science, by enabling predictive models that significantly reduce the time and resources required for complex research processes. Recent studies have demonstrated that COSMO-RS-derived sigma (σ) profiles can be effectively used to develop ML models for accurately predicting thermodynamic and physicochemical properties of ILs, DESs, and organic solvents.[11,22,25,27–30] However, the generation of σ-profiles for molecules is computationally intensive and impractical for large chemical libraries. As an alternative, other featurization techniques such as Morgan fingerprints, atomic counts, molecular descriptors, and functional group estimations are gaining much attention with less effort and computational time.[17,31,32] Although these approaches reduce computational cost, however, their predictive accuracy often remains limited compared to quantum-chemistry-based descriptors.[11,25,28–30]

In recent years, natural language processing (NLP) methods have introduced a new paradigm for molecular representation.[31,33] The molecular string notation such as SMILES (simplified molecular input line entry system),[34] encode chemical structures as sequences of characters, they can be treated analogously to text. Embedding techniques such as Mol2vec, inspired by Word2vec, are able to learn high-dimensional vector representations of molecular substructures that capture both local and global chemical environments. These embeddings can then be used as features in ML models, providing rich structural information without the computational burden of quantum calculations. Recent applications of Mol2vec have demonstrated promising predictive performance for small-molecule properties, but its potential for accurately modeling the diverse and complex chemical space of ILs has not been fully realized.

In this work, we leverage NLP-based Mol2vec embeddings combined with the CATBoost algorithm to predict seven critical IL properties-viscosity, surface tension, ionic conductivity, density, melting temperature, toxicity, and water solubility. The rationale for selecting these seven physicochemical properties were guided by three key considerations: (i) *fundamental relevance to IL performance across several industrial processes (e.g., biomass processing, CO_2 capture, battery electrolytes, chemical synthesis, etc.)*, (ii) *coverage of fundamental property classes (transport properties, thermodynamic and phase behavior, and environmental/biological interactions)*, and (iii) *availability of high-quality experimental datasets*. Other properties such as solvent polarity or cation–anion coulombic interaction energies are indeed important, but they are not as consistently available in large datasets, and their effects are largely reflected in the selected measurable properties. For example, polarity influences surface tension and water activity, whereas coulombic interactions affect viscosity and ionic conductivity. We compile large and diverse experimental datasets and compare Mol2vec with traditional featurization techniques, including Morgan fingerprints,[35] atom counts, and COSMO-RS-derived σ-profiles.[25] Finally, we demonstrate the practical utility of our approach through high-throughput screening of approximately 10.6 million systematically generated ILs, identifying potential candidates with desirable physicochemical properties for specific tasks of societal importance and potential enhanced performance. This work establishes NLP-based molecular embeddings as a powerful and computationally efficient technique for the accelerated discovery and design of task-specific ionic liquids.

Methodology

Datasets

Seven critical properties of ILs were investigated: surface tension, viscosity, density, ionic conductivity, melting temperatures, toxicity, and activity of water in ILs. The dataset employed was taken from our previous studies,[11,30] Song *et al.* (2024),[36] and ILThermo v2.0,[37,38] the last of which is maintained by the National Institute of Standards and Technology (NIST).[37] Detailed information about the number of data points and sources of properties are presented in Table 1. A detailed discussion of the IL dataset is provided in the SI (Section S1).

Table 1 Analysis of the ILs properties dataset along with their source

IL property	Number of ILs	Number of cations	Number of anions	Data points	Exp. T (K)	Exp. P (kPa)	Property range	Source of data
Surface tension, σ (mN m^{-1})	370	121	70	2663	263.08–533.2	101.325	16.9–76.2	Mohan *et al.*[29,30] and NIST[38]
Viscosity, ln(η) (mPa s)	967	419	172	11 721	278.15–353.15	60–950 000	−0.03–13.85	Mohan *et al.*[11]
Ionic conductivity, κ (S m^{-1})	414	180	85	5700	208.15–528.55	100–101.325	1.7 × $^{-7}$–14.54	Song *et al.*[36]
Density, ρ (kg m^{-3})	1687	906	341	52 278	90–573	81.5–300 000	780–2150	NIST[38]
Melting temperature, T_m (K)	3076	1763	205	3076	—	101.325	177.15–292.15	Feng *et al.*[42]
Activity of water, γ^{w}_{IL}	168	108	45	3578	288.05–433.15	80–101.33	0.028–9.36	NIST[38]
Toxicity, $\log_{10}$ EC$_{50}$	312	141	49	312	—	—	−0.24 to 4.9	Zhong *et al.*[61]

For all ILs in the dataset, SMILES strings were generated using the OPSIN method,[39] and those SMILES unavailable in OPSIN were retrieved from PubChem[40] or ChemDraw 3D.[41] SMILES is a widely used notation for representing molecular structures, in which compound has a specific canonical SMILES. For example, the canonical SMILES for 1,3-dimethylimidazolium acetate is CN1C=C[N +](=C1)C·CC(=O)[O–], where the dot (·) separates the cation and anions. The notation [N+] indicates a positively charged cation, while [O–] represents a negatively charged anion. Several studies in the literature,[36,42] have developed ML models using erroneous SMILES representations, such as charge imbalances in IL SMILES. In our work, we corrected these charge imbalances to ensure accurate IL featurization and ML model development. A detailed discussion on the erroneous IL SMILES can be found in the SI (Section S1).

Feature engineering

Efficient and chemically meaningful molecular representations are critical for developing predictive ML models of ionic liquids. Here we explore three different featurization techniques for ILs-Mol2vec molecular embeddings, Morgan FPs, and atom count descriptors-to predict IL properties. For comparison, COSMO-RS-derived σ-profiles from our previous studies were used as high-fidelity quantum-chemistry descriptors.[11,29,30] This multi-featurization approach systematically evaluates the balance between information richness and computational cost in predicting IL properties. Along with IL feature engineering, experimental parameters such as temperature (K) and pressure (kPa) were included as inputs in the ML. We employed the categorical boosting algorithm with well-optimized hyperparameters for each IL property prediction.

Natural language processing method: Mol2vec

Mol2vec is an unsupervised ML approach developed to learn vector representations of molecular substructures, inspired by the widely used Word2vec algorithm in NLP techniques.[31,43,44] In this framework, Mol2vec treats molecular substructures—identified using the Morgan algorithm—as "words" (*e.g.*, atoms, bonds, and fragments) and entire molecules as "sentences" (https://github.com/samoturk/mol2vec).[31] By applying the Word2vec algorithm to a large corpus of molecular structures, Mol2vec generates high-dimensional vector embeddings in which chemically similar substructures cluster together in the same part of the vector space. Here, this NLP molecular embedding approach is adapted for predicting the properties of ILs. We utilized Mol2vec pre-trained model with the Genism implementation of Word2vec, leveraging a dataset of 19.9 million molecules from the ZINC and ChEMBL (Chemical Database of Bioactive Molecules with Drug-Like Properties) databases. Consistent with steps in the Mol2vec paper,[31] SMILES in our datasets were identified by the extended-connectivity fingerprints (ECFP)-based tokenization process at radii 0 and 1. Each IL was then represented as a 300-dimensional embedding obtained by summing the vectors of all substructures. This representation captures both local and global chemical environments.[31] The overview of Mol2vec and feature generation steps are provided in Fig. 1a. A depiction of identifiers (radius 0 followed by radius 1 each) obtained with the Morgan algorithm on the structure of 1,3-dimethylimidazolium acetate forming a molecular sentence is shown in Fig. 1b. In the present study, Mol2vec was used with default hyperparameters to featurize ILs, and no hyperparameter optimization was performed.

Morgan fingerprints

Morgan fingerprints (Morgan FPs), also known as Extended-Connectivity Fingerprints (ECFPs), are molecular descriptors widely used in cheminformatics and ML to represent molecular structures.[45] Morgan FPs encode the presence of specific functional groups and structural motifs within a fixed-length binary vector. In this work, fingerprints were generated from canonical SMILES strings using RDKit (AllChem. GetMorganFingerprintAsBitVect) with a radius of 2 and a vector size of 2048 bits.[46] This configuration ensures that structural patterns extending up to two bonds are captured and hashed into a compact representation suitable for tree-based ML algorithms.

Atom-count features

Elemental composition of IL was calculated directly from canonical SMILES strings to provide simple and interpretable features reflecting molecular size and stoichiometry. First, we converted the SMILES strings into molecular formulas for each compound. Then, the number of each element type (C, H, N, O, S, P, F, Cl, Br, I, Si, Fe, and Al) was determined from the molecular for-

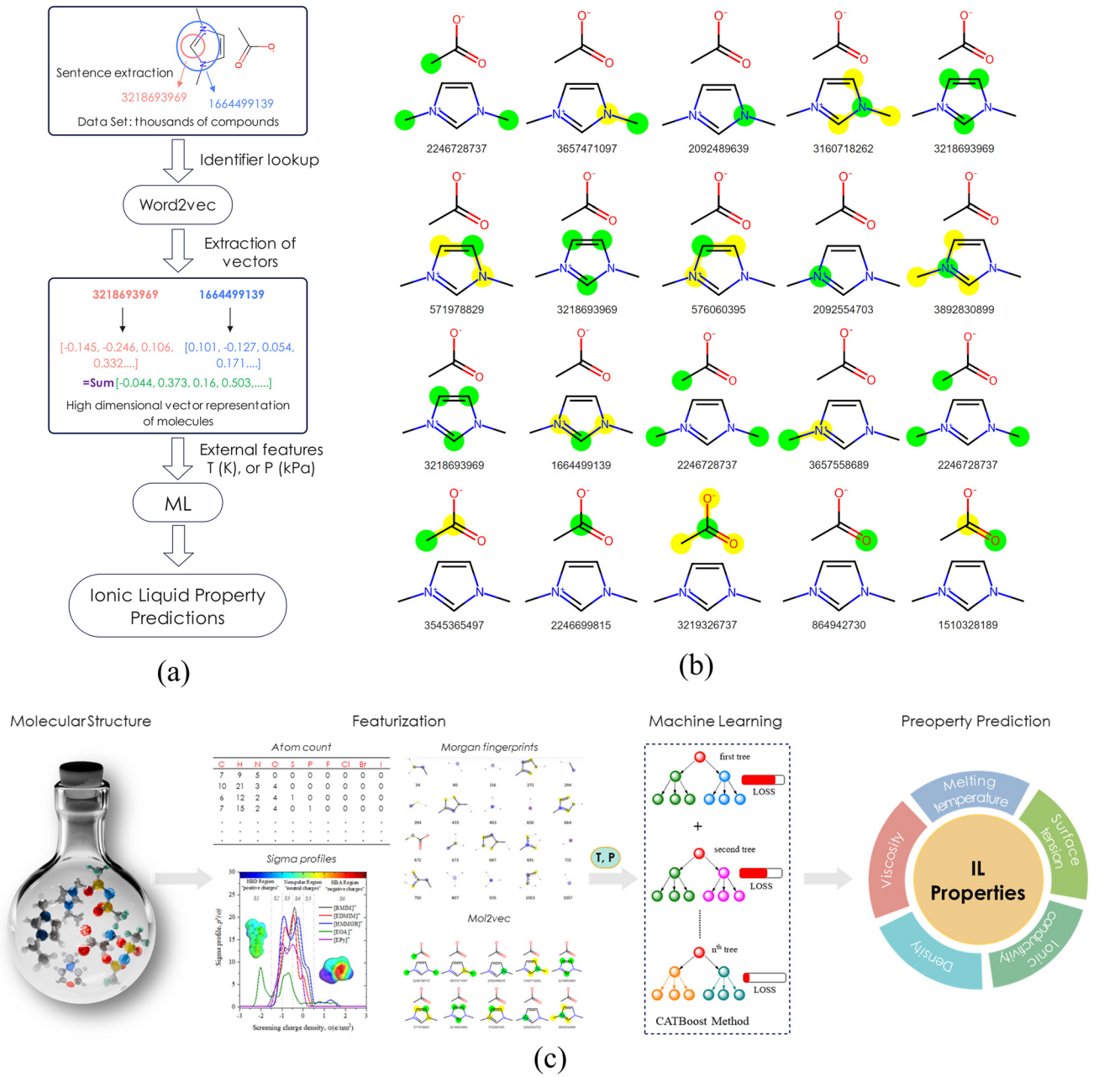


Fig. 1 (a) Overview of the generation and usage steps of Mol2vec. Mol2vec embeddings (*i.e.*, vector representations of substructures) are generated *via* an unsupervised pretraining. (b) Depiction of identifiers obtained with the Morgan algorithm on the structure of 1,3-dimethylimidazolium acetate forming a molecular sentence. If an atom has more than one identifier, the first identifier for that atom is the one for radius 0, followed by radius 1, *etc*. (c) Schematic overview of machine learning framework for predicting multiphysical properties based on different input featurization techniques.

mulae using the 'chemparse' library in Python.[32] These atom counts were combined with experimental temperature (K) and pressure (kPa) to incorporate experimental conditions. Table S1 reports the number of input features used in the ML model for each featurization technique.

ML model development: categorical boosting method

Our previous studies have demonstrated that advanced decision tree-based algorithms, particularly categorical boosting (CATBoost, developed by Yandex Corporation[47]), exhibit excellent predictive capabilities for IL viscosity and surface tension.[11,25] Thus, we chose CATBoost to develop ML models for the IL property predictions here. The choice of CATBoost was also driven by its ability to handle both categorical and numerical input features using ordered target statistics.[48] Unlike competing algorithms, such as XGBoost[49] and Light Gradient Boosted Machine (LightGBM),[50] CATBoost employs a unique tree-growing mechanism. CATBoost is available as

open-source software (https://catboost.ai/, accessed on 25 October 2023).[47] Adding random permutation in the ordered boosting prevents overfitting effectively, while the other two algorithms above grow asymmetrical trees quickly leading to potential overfitting. The ML models were developed and trained in Python 3.9.13 using the scikit-learn package.[51,52] The hyperparameters of CATBoost method was optimized using the Bayesian Optimization (BO) technique with the optuna python library.[53] The details of the hyperparameters search were provided in Table S2. Early stopping was implemented, and the model performance was monitored on the validation set, and training was stopped when no further improvements were observed in validation loss.

In addition, to further improve robustness and quantify predictive uncertainty, we trained multiple CATBoost models with different random seeds and applied the Virtual Ensembles (VE) method.[54] Each model generates multiple VE predictions, and the aggregated mean and variance provide both accurate point estimates and calibrated prediction intervals for the IL properties. This ensemble–VE framework enhances the reliability of IL property predictions without the computational cost of training a large number of independent models. This method prevents the overfitting of ML models. A schematic overview of the ML framework for predicting IL properties using various input featurization techniques is shown in Fig. 1c.

Model performance was evaluated on the independent test set using root-mean-squared error (RMSE) and mean absolute error (MAE) to evaluate error distribution, along with the coefficient of determination (R^2) to measure agreement between predicted and experimental IL property values. Lower RMSE and MAE with higher R^2 value indicate greater prediction accuracy and better model reliability. The final RMSE, MAE, and R^2 are reported based on the testing dataset.

Results and discussion

IL properties

To prevent ML model overfitting and enable generalizability, we performed t-distributed stochastic neighbor embedding (t-SNE)[55] feature dimensionality reduction with different sampling techniques using the instant Similarity (iSIM) method.[56] The iSIM method was used to split the dataset into training, validation, and testing sets based on the t-SNE clustering and chemical space analysis. The iSIM method uses molecular fingerprints (Morgan fingerprints) to calculate pairwise similarity scores and enables the selection of representative samples to ensure balanced representation of molecular diversity. Using this iSIM method, various approaches such as medoid sampling (similar molecules), outlier sampling (dissimilar molecules), and stratified sampling (cover both medoid and outlier sampling) approaches were explored.[56] Fig. S1 shows the graphical representation of medoid, outlier, and stratified sampling methods.

Medoid sampling. This sampling method selects the most representative ILs with low complementary similarity within the chemical space. These are structurally similar, minimizing the average distance to all other ILs in the group.[56]

Outlier sampling. This method selects structurally dissimilar ILs with high complementary similarity (ILs from low density regions) in the chemical space. This method is important for exploring extreme regions of the ILs landscape.[57]

Stratified sampling. In the stratified sampling approach, ILs are first divided into b equally sized groups, and these groups are called strata. From each stratum, ILs with the lowest complementary similarity (*i.e.*, medoid) are picked based on the defined percentage. The stratified sampling method covers the well-balanced medoid sampling and outlier sampling in their chemical space.[57,58]

In this study, we did not use manually or random data splitting. Instead, we employed iSIM-based stratified, medoid, and outlier sampling techniques for data splitting, which clusters ILs using Morgan fingerprints and Tanimoto similarity and then draws representative points from each cluster. This split acts as a scaffold-like split, reducing the chance that highly similar ILs (or duplicates in chemical space) dominate a single split and thereby mitigating leakage relative to random splitting. Further, the availability of dataset is small for many IL properties (example, surface tension- 370 ILs; toxicity-312 ILs, activity of water-168, viscosity-967, and ionic conductivity-414) and the dataset is strongly imbalanced with IL classes. Imidazolium-based ILs constitute a major fraction (~60%) and remaining dataset belonging to pyridinium, pyrrolidinium, ammonium, phosphonium, and sulfonium families. To ensure that the model trained across IL classes, we utilized iSIM-based sampling technique to maintain structural diversity across the training, validation, and testing sets. This sampling technique prevents the model from being biased toward dominant classes and improves its generalizability across IL families. The dataset was split into 70% for training, 10% for validation, and 20% for testing the model. The training set was used for model development and the validation set used for fine-tuning and hyperparameter optimization, and the test set was used for testing the model on unseen data. Fig. 2–6 show comparisons of the different featurization techniques for each of the properties examined. The metrics are tabulated in Table 2. The main conclusion from these plots is that CATBoost with the Mol2vec NLP featurization method outperforms the other approaches for each of the properties studied and with very high predictive power.

Fig. 2(a–d) illustrate the correlation between experimental and predicted surface tension of ILs in the training, validation, and testing sets for different featurization techniques using the CATBoost model. The parity plots in Fig. 2(a–d) show that the ML model performs relatively poorly when using atom-count, sigma profiles, and Morgan FP features, yielding relatively low R^2 (0.880–0.951) and high RMSE (1.668–2.644 mN m^{-1}) on the test sets. However, CATBoost performs well on the training sets with these features. CATBoost with Mol2vec featurization demonstrates excellent predictive performance on

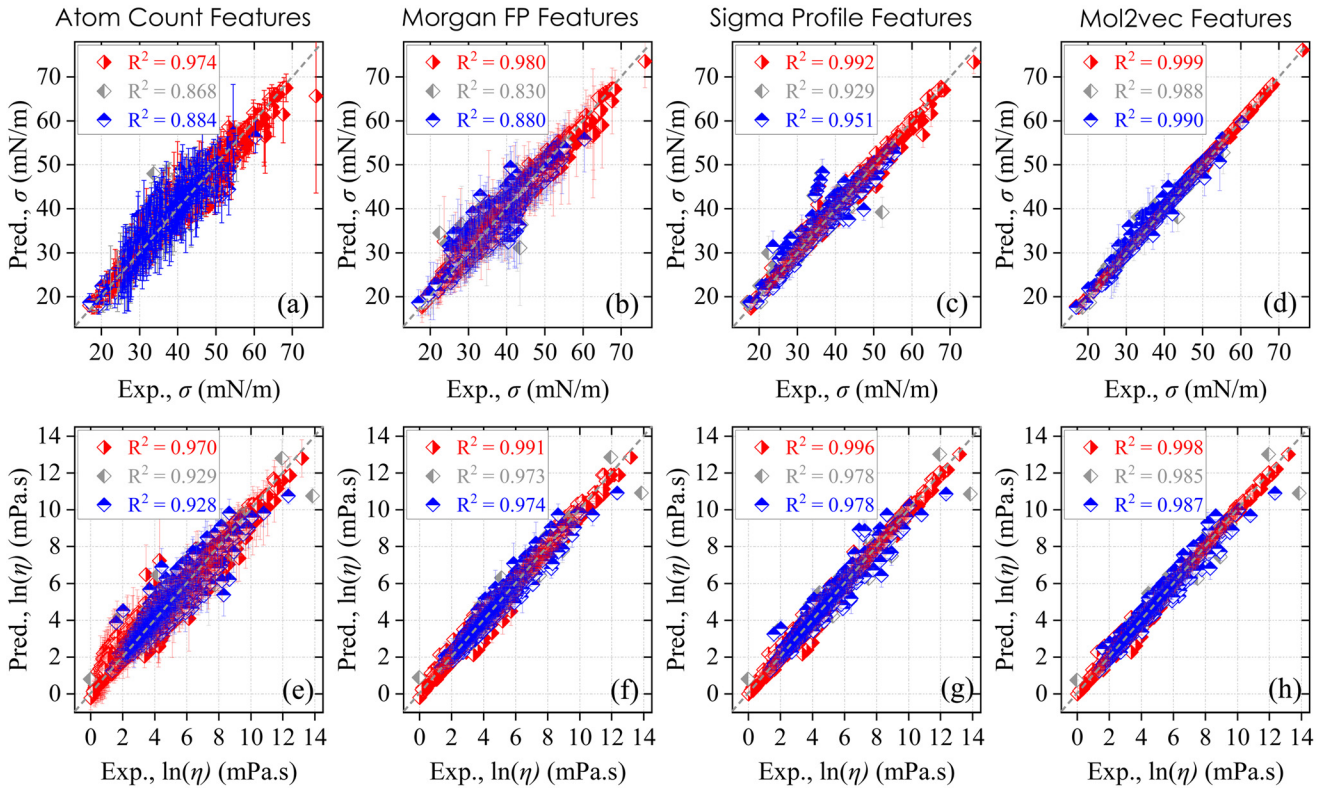


Fig. 2 Experimental and ML-predicted surface tension of ILs using various approaches: (a) atom count, (b) Morgan FPs, (c) σ-profiles, and (d) Mol2vec with the CATBoost method. Similarly, the experimental and ML-predicted viscosities of ILs is shown for (e) atom count, (f) Morgan FPs, (g) σ-profiles, and (h) Mol2vec featurization with the CATBoost method. Symbol colors: red – training data; grey – validation, and blue – testing sets. The uncertainties are calculated using the virtual ensemble (VE) module implemented in CATBoost.

training and test sets, achieving a high R^2 value of 0.990 and a low RMSE of 0.755 mN m^{-1} on the test set (see Table 2). To further assess model reliability, Virtual Ensemble (VE)-based uncertainty quantification was applied, with 95% prediction intervals (PIs) shown as vertical error bars in the plots. Across all featurization techniques, the Mol2vec-based model exhibits the narrowest and most consistent uncertainty bounds, indicating high predictive capability and lower epistemic uncertainty and also achieved 95% VE–PI coverage after applying calibration scaling factor. Lemaoui *et al.* (2024)[59] developed ML and deep learning (DL) model to predict the surface tension of ILs using σ-profiles, and the performance of their DL model is comparable to our models based on the Morgan FP and atom-count featurization technique. In summary, based on the ML performance the ranking of these featurization techniques is as follows: Morgan FP < atom count < σ-profile < Mol2vec. To further illustrate model performance, Fig. S2(a–d) presents the residual deviations between the experimental and ML-predicted surface tensions. Mol2vec exhibits the lowest residual deviations, confirming that this model provides the most accurate predictions with the smallest errors.

Fig. 2(e–h) illustrates the correlation between experimental and CATBoost predicted IL viscosity in the training, validation, and testing sets and employing the different featurization methods. Apart from the atom-count (Fig. 2e), all the investigated featurization techniques demonstrated excellent predictive accuracy. Among these, Mol2vec outperformed Morgan FPs, achieving higher R^2 values and lower RMSE and MAE, and this model also improved IL viscosity prediction relative to recently published models.[11,59] For example, Lemaoui *et al.* (2024)[59] developed a DL model using COSMO-RS-derived σ-profiles to predict IL viscosity, but its performance (R^2 = 0.907 and RMSE = 0.477 mPa s) was somewhat weaker than that of the present study. The calculated experimental standard deviation (SD) of IL viscosities is closer to the SD values predicted using Mol2vec. To further analyze prediction errors, residuals (the difference between experimental and predicted values) were plotted against experimental viscosity (Fig. S2(e–h)). The residual plot indicates that again Mol2vec yields lower residual deviations than the other featurization approaches.

We have also developed predictive CATBoost models for the ionic conductivity of ionic liquids using the three different featurization techniques: atom count, Morgan FPs, and Mol2vec (Fig. 3(a–c)). For all featurization techniques the correlation between experimental and ML predicted ionic conductivities is excellent. However, again the Mol2vec-based model demonstrated marginally better performance. Table 2 presents the performance metrics for the investigated models, with the models are ranked based on their predictive accuracy. The

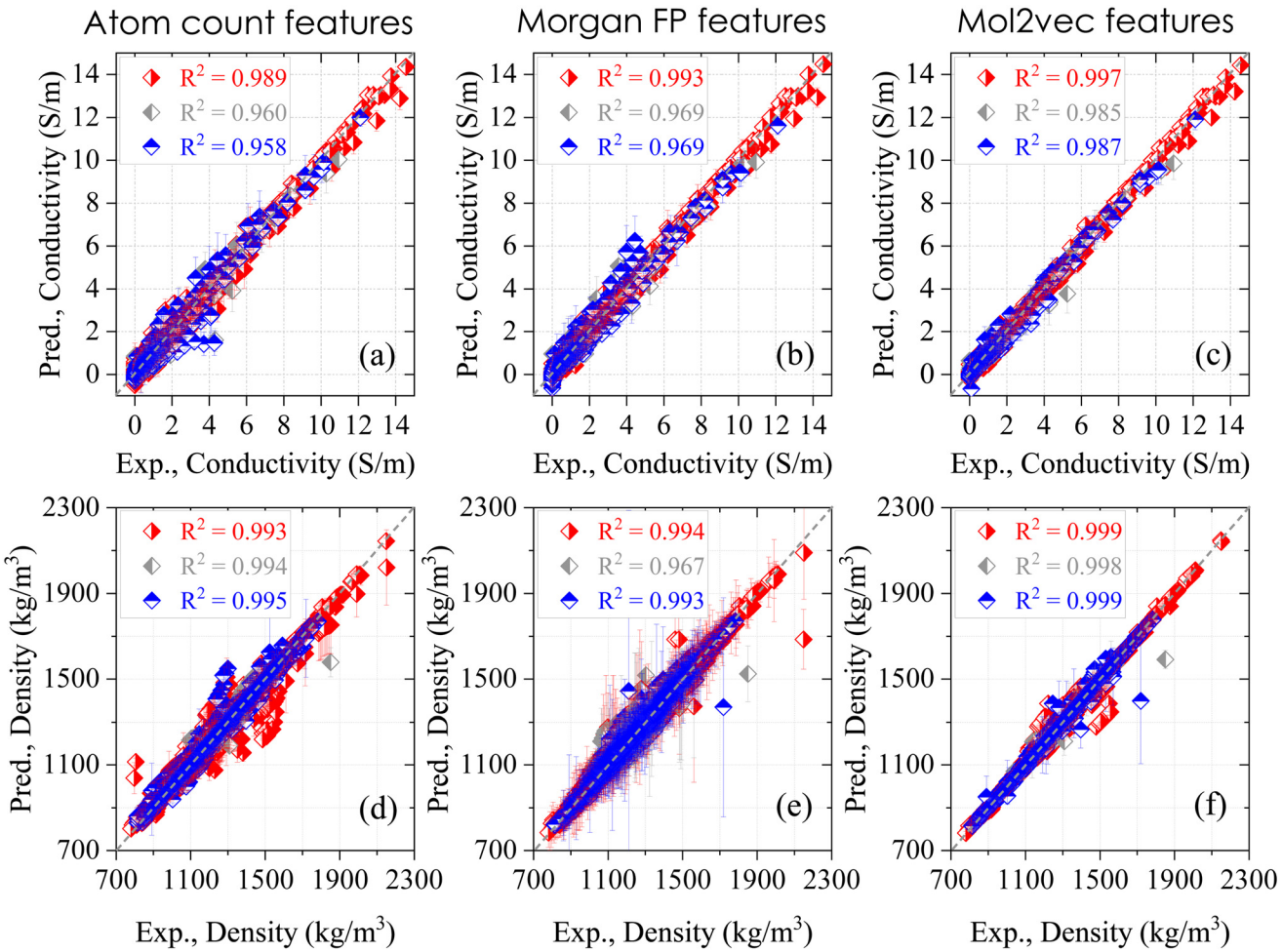


Fig. 3 Experimental and ML-predicted ionic conductivity of ILs using various approaches: (a) atom count, (b) Morgan FPs, and (c) Mol2vec with the CATBoost method. Similarly, the experimental and ML-predicted densities of ILs is shown for (d) atom count, (e) Morgan FPs, and (f) Mol2vec featurization with the CATBoost method. Symbol colors: red – training data; grey – validation, and blue – testing sets. The uncertainties are calculated using the virtual ensemble (VE) module implemented in CATBoost.

Mol2vec-based model, with the highest performance, has R^2 = 0.987, MAE = 0.087 S m^{-1}, and RMSE = 0.142 S m^{-1} on the test set. In comparison, Chen *et al.* (2024)[60] developed two ML models using σ-profile featurization for predicting IL ionic conductivity, but their models exhibited relatively weak performance with an R^2 of 0.77 on the test set. Recently, Song *et al.* (2024)[36] developed four ML models to predict ionic conductivity using graph neural networks (GNN) featurization and reported excellent predictive accuracy with a high R^2 values of 0.97–0.99 and low RMSEs on the test set. Our model utilizing the NLP-based Mol2vec featurization approach performs better as compared previous attempts to predict ionic conductivity, demonstrating its effectiveness in capturing structure–property relationships.

The predictive performance of the CATBoost model using three different featurization techniques for IL density prediction is depicted in Fig. 3(d–f). NLP-based featurization technique exhibit strong correlation with experiments. In contrast, the Morgan FP and atom count technique showed relatively weaker predictive performance, yielding R^2 and RMSE values of 0.99 and 12.20–13.81 kg m^{-3}, respectively, on the test set. Among the three approaches, the Mol2vec featurization demonstrated the highest accuracy, achieving R^2, MAE, and RMSE values of 0.999, 2.25 kg m^{-3}, and 5.74 kg m^{-3}, respectively. Fig. S3(d–f) evaluates the relative deviation distribution for the Mol2vec model, revealing that most data points fall within a ±200 kg m^{-3} range, predominantly concentrated between ±100 kg m^{-3}. This distribution suggests minimal prediction bias compared to the other two featurization techniques. Finally, the toxicity ($\log_{10}$ EC$_{50}$) of ionic liquids (leukemia rat cell line IPC-81) was also predicted. For IL toxicity, ML model uses only structural information encoded through Mol2vec embeddings generated from canonical SMILES. These embeddings, which capture both local and global substructural chemical features, are used as input to train ML models that predict experimental EC$_{50}$ values, and the results presented in Fig. 4. Toxicity data for 312 ILs were sourced from Zhong *et al.* (2024)[61] in which the $\log_{10}$ EC$_{50}$ values range from

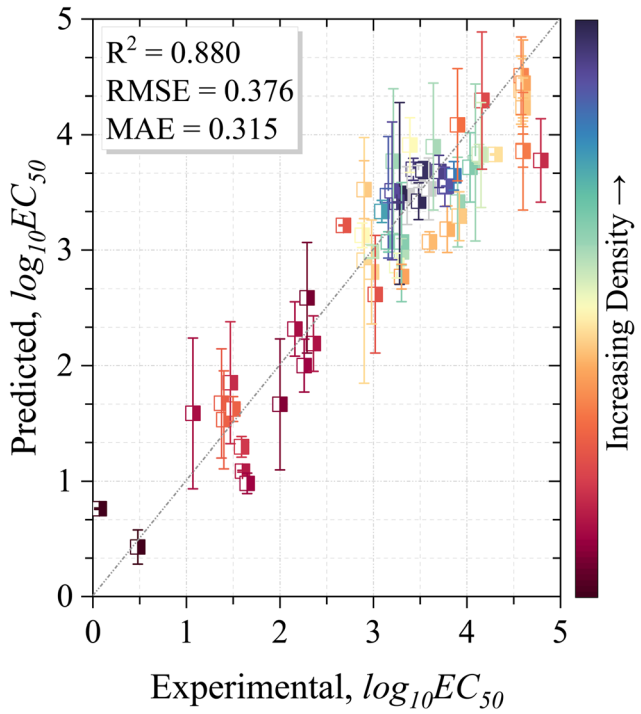


Fig. 4 Correlation between experimental and ML predicted ionic liquids' toxicity using the CATBoost method with Mol2vec features. The uncertainties are calculated using the virtual ensemble (VE) module implemented in CATBoost.

−0.24 to 4.9. The ML model again demonstrated strong predictive accuracy, achieving an excellent (if not close to perfect) correlation with experimental data (high R^2 = 0.880 and low RMSE = 0.376). This model exhibits slightly improved predictive performance than the model reported by Zhong *et al.* (2024)[61] (R^2 = 0.859). The vertical error bars in Fig. 4 represent 95% prediction intervals, capturing the spread of model predictions. Notably, the PI widths are relatively large for certain ILs, indicating higher uncertainty in toxicity predictions compared to other properties.

As discussed, Mol2vec features are generated by summing the vectors of molecular substructures identified from SMILES strings, which provide a chemically meaningful yet highly abstract representation. Unlike traditional featurization techniques (*e.g.*, σ-profiles, Morgan FPs, and atom count), understanding the importance of functional groups or local chemical environments in Mol2vec-based models is challenging.

Melting temperature

In addition to predicting IL surface tension, viscosity, ionic conductivity, density, and toxicity of ILs, we developed ML models for predicting melting temperatures, T_m. For this we used the top performing Mol2vec featurization benchmarked against existing literature models. Fig. 5a shows the correlation between the predicted and experimental T_m values. The model achieves quite good predictive accuracy on the test set, with an RMSE of 39.87 K, MAE of 30.43 K, and an R^2 value of 0.720. The NLP-based featurization technique, combined with a larger dataset than Venkatraman *et al.*,[62] significantly enhances predictive performance compared to previously reported models.[62,63] For instance, Makarov *et al.* (2022)[63] trained convolutional neural networks (CNNs) on SMILES strings to predict ILs T_m using 3073 data points, achieving an R^2 of 0.66 and an RMSE of 45 K. Similarly, Venkatraman *et al.* (2018)[62] explored six different ML models for predicting the T_m of 2212 ILs using the quantum chemistry-derived molecular descriptors. With the gradient boosting method (GBM), they obtained an R^2 value of 0.67 and RMSE of 45 K on the test set, which is again less accurate than the present model's performance.

We have also performed an error analysis to identify the structural classes responsible for the largest deviations in melting temperature predictions. We observed that ILs containing sulfonate (O═S(═O)([O–])) and sulfone (O═S═O) anions exhibited higher residuals when combined with piperidinium, pyrrolidinium, and imidazolium-based cations. Furthermore, we retrained the CATBoost model by augmenting the original Mol2vec embeddings with two classes of physically meaningful descriptors: (i) atom-count features (11 elemental descriptors) and (ii) RDKit-derived molecular descriptors capturing topology, polarity, and hydrogen-bonding capacity, *etc.* This hybrid input featurization results in slightly improved performance, increasing R^2 from 0.713 to 0.758 and reducing the MAE and RMSE from 30.43 K and 39.87 K to 27.64 K and 36.85 K, respectively (see Fig. S4).

However, despite these refinements, the accuracy of the regression model leaves significant room for improvement, and therefore, we explored a classification-based approach, using the Mol2vec features and the CATBoost method. The ILs were categorized into two classes based on their T_m values: those with T_m below 300 K were classified as "liquid", and those above 300 K were classified as "solid". As shown in Fig. 5(b–d), we conducted a comprehensive evaluation of the classification model's performance using metrics such as accuracy, the confusion matrix, and the ROC/AUC curve. The model achieves an accuracy of 0.844 with an ROC curve yielding an AUC of 0.88 (Fig. 5b), indicating that the model's accuracy on the test set is notably high, and the confusion matrix (Fig. 5b) indicates a robust classification performance. Additionally, precision, recall, and F1-score metrics were evaluated (Fig. 5d), demonstrating balanced performance across different classes and underscoring the reliability of the predictions. Relative to the regression models, the classification model performance is better with higher accuracy and precision. The better performance of classification model can be attributed to the model's ability to simplify the prediction task, thus effectively capturing the structure–property relationships in ILs. The classification approach offers more accurate and precise metrics for predicting the phase behavior of ILs at given temperatures, providing valuable insights for applications, where knowing the liquid or solid phase of ILs is critical. Furthermore, we also developed a classification model using hybrid featurization (Mol2vec, atom counts, and RDKit molecular descriptors); however, its performance was weaker than the model using only Mol2vec features.

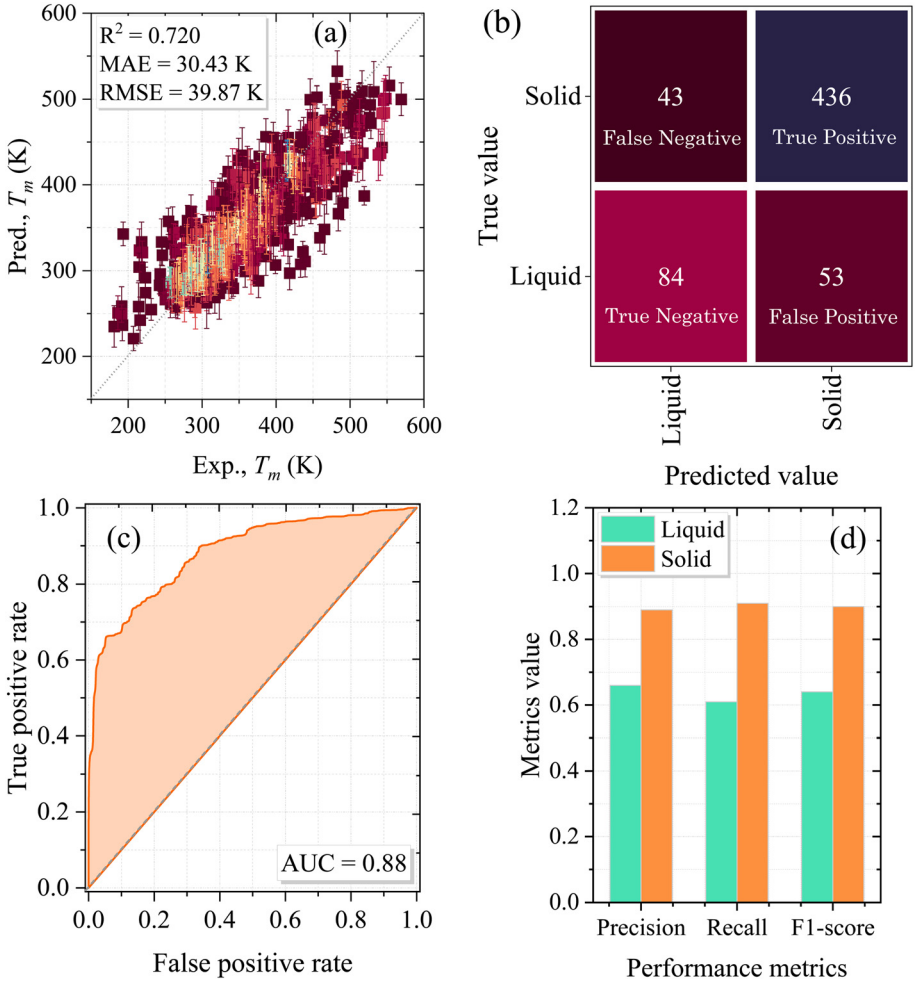


Fig. 5 (a) Correlation between experimental and ML predicted ionic liquid melting temperatures using the CATBoost method with Mol2vec NLP featurization technique. (b) Confusion matrix (c) AUC of ROC on the independent test set for the ILs melting temperature using Mol2vec feature and the CATBoost model. ILs melting temperature data classification liquid: $T_m \leq 300$ K, solid: $T_m > 300$ K. (d) Performance metrics of ILs melting temperature for classification method. The accuracy of classification model is 0.844. The uncertainties are calculated using the virtual ensemble (VE) module implemented in CATBoost.

Water activity in ionic liquids

To assess the hydrophilicity and hydrophobicity of ILs, a ML model was developed to predict the activity coefficient of water in ILs (γ_{IL}^{W}). The activity coefficient is often used as a quantitative descriptor for the dissolution power of a solvent.[15,19] According to solid–liquid equilibria (SLE) assumptions,[64] the solubility of water in an IL is characterized by the reciprocal of the activity coefficient: a lower activity coefficient value ($\gamma_{IL}^{W} < 1$) indicates higher water solubility, classifying the IL as hydrophilic. Conversely, $\gamma_{IL}^{W} > 1$ suggests lower water solubility, indicating hydrophobicity. To train the model, an extensive experimental dataset for γ_{IL}^{W} was collected from the ILThermo v2.0 database[37,38] comprising 3578 data points for γ_{IL}^{W} at concentrations of water varying from 0 (infinite dilution of water) to 1 (infinite dilution of IL), temperatures ranging from 288.05 K to 433.15 K, and pressures ranging from 80 kPa to 101.33 kPa. The dataset comprised 168 distinct ILs with combinations of 108 cations and 45 unique anions. Fig. 6a shows the relationship between the number of carbon atoms and water activity as a function of water concentration. The majority of the data points are clustered around carbon numbers below 20, with higher experimental water activities occurring within this range. In contrast, only a few scattered data points appear at higher carbon numbers, indicating that as the carbon number increases the water activity data becomes more dispersed and less frequent. Notably, no clear linear correlation is observed between the number of carbon atoms in ILs and the water activity, indicating that a non-linear model is required to accurately predict γ_{IL}^{W}.

We employed the Mol2vec featurization method to predict γ_{IL}^{W}. Fig. 6b illustrates the correlation between ML predicted and experimental γ_{IL}^{W} values, demonstrating excellent predictive performance on the test set, with low RMSE and MAE values of 0.063 and 0.017, respectively, and a high R^2 value of 0.99. In comparison, Paduszyński (2016)[65] developed various tra-

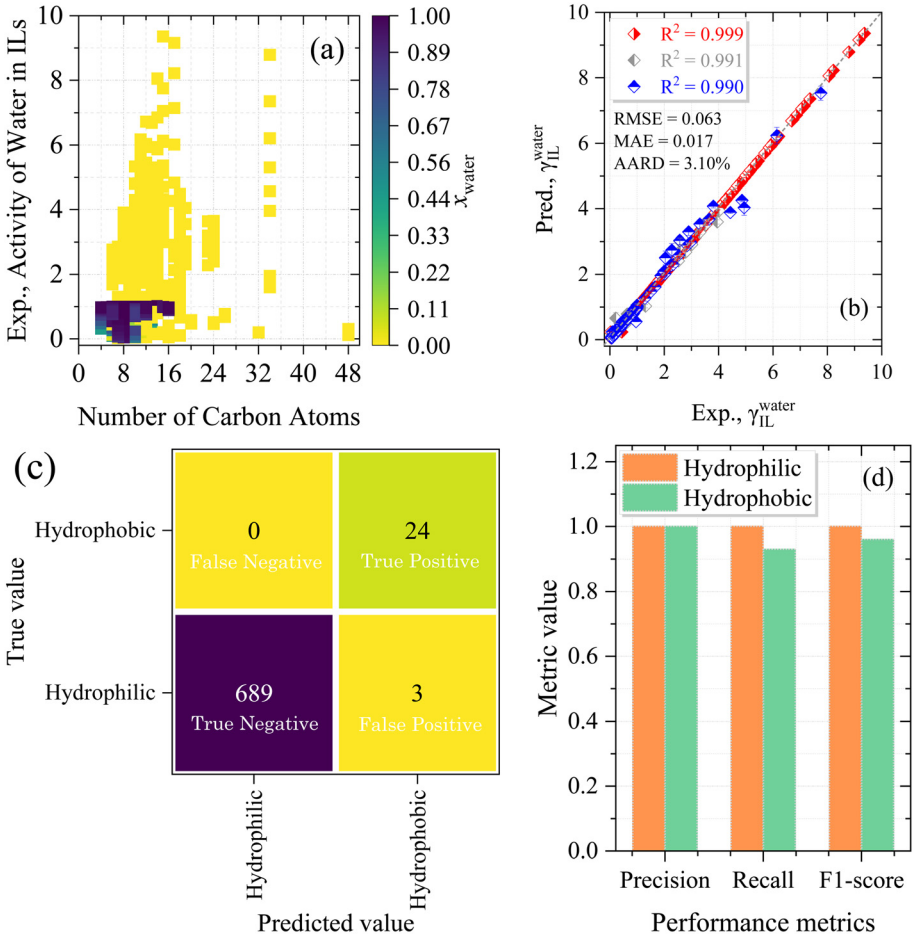


Fig. 6 (a) Relationship between experimental activity coefficient of water in ionic liquids and number of carbon atoms in ILs as a function of water mole fraction. The color scale on the figure indicates the density of data points, with darker colors (purple) representing lower density with high water concentration and brighter colors (yellow) showing higher densities with low water concentration. (b) Correlation between experimental and ML predicted activity coefficient of water in IL using the CATBoost method with Mol2vec NLP featurization technique for training (red), validation (grey), and testing (blue) sets. The performance metrics reported for testing set. (c) Confusion matrix, and (d) performance metrics of water activity coefficient in ionic liquid for classification method. Classification of water activity coefficient in ionic liquid: $\gamma \leq 1$ then hydrophilic, and $\gamma > 1$ then hydrophobic. The accuracy of classification model is 0.997. The AUC of ROC curve is 1.0. The uncertainties are calculated using the virtual ensemble (VE) module implemented in CATBoost.

ditional ML models to predict the activity of molecular solvents in ILs, achieving an RMSE of 0.205 with a feed-forward neural network (FFNN), which was less accurate than our NLP-based approach. Further, we also explored a classification approach using the same Mol2vec featurization in combination with the CATBoost method to categorize the ILs. For this, the ILs were divided into two classes based on their γ_{IL}^{W} values: those with γ_{IL}^{W} below 1 were classified as "hydrophilic", and those above 1 were classified as "hydrophobic". Fig. 6(c and d) shows a comprehensive evaluation of the classification model's performance using metrics such as accuracy, ROC/AUC curves, and the confusion matrix. On the test set the model achieves an accuracy of 0.997 with an ROC curve yielding an AUC value of 1.0, signifying excellent performance. The confusion matrix (Fig. 6c) indicates a robust classification performance with accurate predictions. Moreover, the precision, recall, and F1-score metrics show balanced performance across both classes, reinforcing the reliability of the predictions (Fig. 6d).

Comparison of NLP predictions with literature reported models

A comparison is made of the present NLP-based ML results with recently reported literature models for all the investigated IL properties in Table 3, using RMSE and R^2 values as evaluation metrics. For IL surface tension Mohan *et al.* (2023)[30] and Lemaoui *et al.* (2024)[59] developed ML and DL models utilizing the COSMO-RS-derived σ-profiles as inputs. Lemaoui *et al.* (2024)[59] reported a weaker predictive performance for their deep learning model (R^2 = 0.931, RMSE = 2.251 mN m^{-1}) than the present NLP-based model. In our earlier studies,[29,30] we developed eight ML models for surface tension prediction, with the XGBoost method demonstrating the best performance among them; however, relative to the present work lower R^2

Table 2 Performance metrics of CATBoost model on the ionic liquids' properties for different input feature sets

IL property	Feature	Data set	R^2	MAE	RMSE
σ, mN m^{-1}	Atom count	Training	0.974	0.921	1.520
		Testing	0.884	1.865	2.601
σ, mN m^{-1}	Morgan FP	Training	0.980	0.950	1.355
		Testing	0.880	1.986	2.644
σ, mN m^{-1}	Sigma profiles	Training	0.992	0.526	0.863
		Testing	0.951	1.057	1.668
σ, mN m^{-1}	Mol2vec	Training	0.999	0.105	0.170
		Testing	0.990	0.407	0.755
ln(η), mPa s	Atom count	Training	0.970	0.193	0.301
		Testing	0.928	0.235	0.363
ln(η), mPa s	Morgan FP	Training	0.991	0.114	0.167
		Testing	0.974	0.148	0.218
ln(η), mPa s	Sigma profiles	Training	0.996	0.069	0.111
		Testing	0.978	0.115	0.201
ln(η), mPa s	Mol2vec	Training	0.998	0.049	0.080
		Testing	0.987	0.084	0.151
κ, S m^{-1}	Atom count	Training	0.989	0.121	0.185
		Testing	0.958	0.149	0.255
κ, S m^{-1}	Morgan FP	Training	0.993	0.100	0.149
		Testing	0.969	0.140	0.220
κ, S m^{-1}	Mol2vec	Training	0.997	0.059	0.095
		Testing	0.987	0.087	0.142
ρ, kg m^{-3}	Atom count	Training	0.993	6.689	14.267
		Testing	0.995	6.119	12.201
ρ, kg m^{-3}	Morgan FP	Training	0.994	6.320	12.928
		Testing	0.993	6.997	13.815
ρ, kg m^{-3}	Mol2vec	Training	0.999	2.067	5.952
		Testing	0.999	2.246	5.742
$\log_{10}EC_{50}$	Mol2vec	Training	0.935	0.197	0.261
		Testing	0.880	0.315	0.376
T_m, (K)	Mol2vec	Training	0.946	14.39	18.57
		Testing	0.720	30.43	39.87
γ_{IL}^{W}	Mol2vec	Training	0.999	0.0034	0.008
		Testing	0.990	0.017	0.063

(0.963) and higher RMSE (1.716 mN m^{-1}) values were obtained in our earlier studies.

Liu *et al.* (2023),[66] Mohan *et al.* (2024),[11] and Lemaoui *et al.* (2024)[59] developed ML models to predict the IL viscosities at different temperatures and pressures. Liu *et al.* (2023)[66] and Lemaoui *et al.* (2024)[59] has developed ML models based on the norm indexes and σ-profile featurization techniques and reported a R^2 and RMSE values of 0.91 and 0.477 mPa s, respectively, *i.e.*, lower predictive performance compared to our study. It is also important to mention that Lemaoui *et al.* (2024)[59] compiled a dataset for surface tension and viscosity that is 2–2.6 times larger than the present study. This increase in data size is because the authors did not remove duplicate and triplicate IL data points. The inclusion of these redundant data points, which had larger experimental deviations, together with an inability to generate stable IL conformers, resulted in the developed model underperforming in predicting surface tension and viscosity.[59] Chen *et al.* (2023)[67] developed an IL transfer learning of representations model (ILTransR) based on a language model to predict IL viscosities. However, this also exhibited weaker predictive performance (MAE = 0.17 mPa s) than the present study (MAE = 0.09 mPa s). Furthermore, norm index-based models, which reduce ILs to global dimensionless parameters such as size (NS), polarity (NP), and symmetry (NQ), have also been used to predict temperature-dependent IL properties such as density,[66] viscosity,[66] and surface tension.[68] These indices provide simple and computationally inexpensive descriptors; they compress complex structural information into a few coarse values. However, this coarse-grained representation overlooks the influence of local substructures, branching, charge localization, and specific cation–anion interactions that strongly affect dynamic properties like ionic conductivity.

Song *et al.* (2024)[36] employed a deep learning-based GNN featurization approach to predict the ionic conductivity of 414 ILs across various temperature and pressure conditions, achieving high R^2 and low RMSE values (Table 3), indicating excellent predictive performance. Similarly, Chen *et al.* (2024)[60] developed several ML models to predict IL ionic conductivity using the σ-profiles as input features; however, this approach yielded poorer performance with an R^2 value of 0.773. In contrast, our NLP-based featurization technique resulted in excellent predictive capability of IL ionic conductivity, as evidenced by high R^2 and lower RMSE values (Table 3). In our previous work, we demonstrated that COSMO-RS-derived σ-profiles can indeed yield accurate ML models for IL properties such as viscosity, surface tension, and speed of sound.[11,29,30] We have performed the conformer analysis of ILs and generated their most stable conformer for input featurization. In contrast, Chen *et al.*[60] reported weaker performance for ionic conductivity when using σ-profiles, partly because their approach did not explicitly include conformer analysis, a key limitation of σ-profile featurization. Sigma profiles and other quantum chemistry-derived features are highly sensitive to molecular conformer, even a small conformational change can lead to large deviations in electronic descriptors, and thus in ML predictions.

Our Mol2vec-based CATBoost models improve IL property predictions because they avoid dependence on conformers and also capture molecular details differently. Mol2vec embeddings represent molecules through their local substructures and chemical context (such as headgroup type, chain branching, and anion functionalities), which are important for ion mobility and other IL properties. Unlike physics-derived descriptors, Mol2vec encodes cation and anion in the same vector space, allowing the ML model to capture patterns related to their combined structural features without the need for manual feature engineering.

Lemaoui *et al.* (2024)[59] developed nine ML models for predicating IL density using the σ-profile as inputs, achieving a relatively poor performance with an RMSE value of 14.36 kg m^{-3} compared to our approach. In recent years, various IL melting temperature prediction models have been developed, employing diverse methodologies with varying degrees of success. Feng *et al.* (2024)[42] and Makarov *et al.* (2022)[63] utilized

Table 3 Comparison of IL property prediction models

IL property	Model	Featurization	No. of ILs	Data points	R^2	RMSE	Reference
σ, mN m^{-1}	XGBoost	σ-Profiles	360	2524	0.963	1.716	Mohan *et al.*[30]
σ, mN m^{-1}	DL	σ-Profiles	579	6599	0.931	2.251	Lemaoui *et al.*[59]
σ, mN m^{-1}	CATBoost	Mol2vec	370	2663	0.990	0.755	Present study
ln(η), mPa s	QSPR	Norm indexes	832	9238	0.910	—	Liu *et al.*[66]
ln(η), mPa s	CATBoost	σ-Profiles	967	11 721	0.984	0.200	Mohan *et al.*[11]
ln(η), mPa s	DL	σ-Profiles	2026	25 243	0.907	0.477	Lemaoui *et al.*[59]
ln(η), mPa s	CATBoost	Mol2vec	967	11 721	0.987	0.151	Present study
κ, S m^{-1}	XGBoost	GNN	414	5700	0.988	0.180	Song *et al.*[36]
κ, S m^{-1}	XGBoost	σ-Profiles	242	2168	0.870	—	Chen *et al.*[60]
κ, S m^{-1}	CATBoost	Mol2vec	414	5700	0.987	0.142	Present study
ρ, kg m^{-3}	DL	σ-Profiles	2100	40 860	0.993	14.360	Lemaoui *et al.*[59]
ρ, kg m^{-3}	CATBoost	Mol2vec	1687	52 278	0.999	5.742	Present study
$\log_{10}EC_{50}$	CATBoost	Mol2vec	332	332	0.880	0.376	Present study
$\log_{10}EC_{50}$	CATBoost	C-MF	332	332	0.859	0.338	Zhong *et al.*[61]
$\log_{10}EC_{50}$	SVM	2D descriptors	355	355	0.927	0.288	Wang *et al.*[82]
$\log_{10}EC_{50}$	MLR	2D descriptors	304	304	0.77	0.51	Sosnowska *et al.*[83]
T_m, (K)	GNN	GNN	3080	3080	0.760	37.06	Feng *et al.*[42]
T_m, (K)	Transformer CNN	SMILES	3073	3073	0.66	45.0	Makarov *et al.*[63]
T_m, (K)	DL	σ-profiles	1145	1145	0.875	17.45	Lemaoui *et al.*[59]
T_m, (K)	CATBoost	Mol2vec	3080	3080	0.720	39.87	Present study
γ_{IL}^{W}	CATBoost	Mol2vec	168	3578	0.990	0.063	Present study
γ_{IL}^{W}	FFANN	GC	53	399	0.921	0.205	Paduszyński[65]
γ_{IL}^{W}	LSSVM	Critical features	53	318	0.999	0.018	Benimam *et al.*[84]

deep learning techniques based on the GNN and CNNs to predict T_m. Feng *et al.* (2024)[42] obtained an R^2 value of 0.76 and RMSE of 37.06 °C, whereas Makarov *et al.* (2022)[63] achieved an R^2 value of 0.66 and RMSE of 45 K. Additionally, Lemaoui *et al.* (2024)[59] developed a deep learning model for 1145 ILs T_m with σ-profile as input features, and reported the R^2 value of 0.875 and RMSE of 17.45 K. In comparison, our study covers more T_m data with a wider structural diversity of ILs (3080 unique ILs), encompassing 3080 unique ILs with large structural chemical diversity. This extensive dataset allows our model to capture a wider range of structural variations and also predict the melting temperatures with greater accuracy and generalizability.

High-throughput screening of novel ionic liquids for biomass processing, CO_2 capture and electrolyte

The above results show that using ML with NLP featurization seven physicochemical properties of ILs can be predicted accurately, thus enabling the design of task-specific ILs at ambient conditions. We have predicted the viscosity of ILs that are not included in the training, validation, or testing datasets, and the results are shown in Fig. S5. The CATBoost model using Mol2vec features demonstrated good agreement with experimental data, indicating its potential for reliable predictions in high-throughput ILs properties. Further, we evaluated the performance of our ML model for predicting the density and melting temperatures of synthesized ILs reported by Qiu *et al.* (2024),[69] and the predictions show excellent agreement with experimental observations (Table S5). Fig. 7 demonstrates the systematic generation of novel cation and anion molecules, incorporating a wide array of alkyl chain lengths and functional groups at various positions. By systematically varying the alkyl chains and functional groups attached to core ion categories, we were able to explore a broad spectrum of ILs. These variations in ion SMILES is key to fine-tuning optimized physical and chemical properties for a range of research applications, including biomass processing, carbon capture, battery electrolytes, phase separation processes, and for improving the solubility of pharmaceutical compounds (drug delivery). We utilized the RDKit cheminformatics tool for generating novel cations and anions.[70] RDKit ensures the chemical validity of generated molecules by enforcing proper valence states for all atoms and automatically identifies and eliminates configurations that exceed the normal valence limits for each element. For example, oxygen (O), which typically has a valence of 2, will trigger an error in RDKit if a configuration attempts to exceed this valence limit. This rigorous validation process guarantees that the output consists only of chemically plausible molecular structures. Using this method 7200 cations and 1474 anions were combined to yield ~10.61 million unique ILs, and their physicochemical properties were subsequently predicted using our top performed ML models. In addition to IL properties, we have also calculated synthesizability scores (SA score) of ILs using the method developed by Ertl and Schuffenhauer.[71] The SA score ranges from 1 to 10, where lower values indicate easier synthesis and higher values corres-

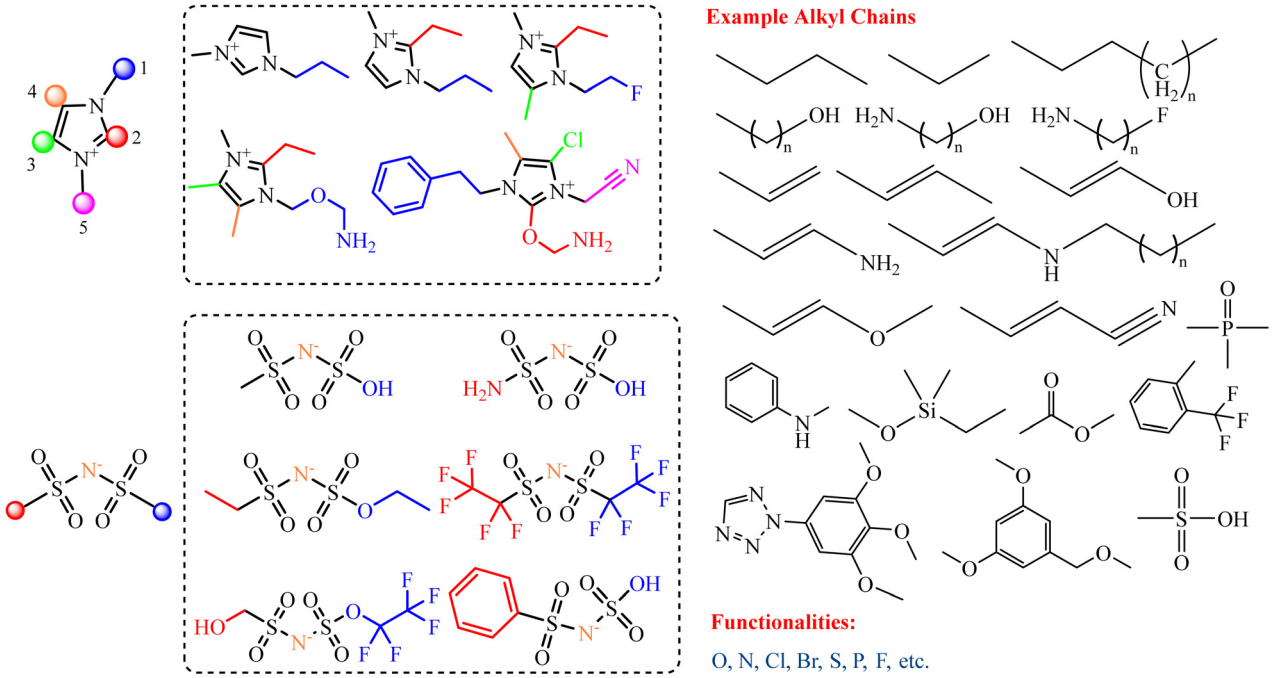


Fig. 7 Generation of novel cation and anion molecules with diverse combinations of alkyl chain lengths and functional groups at various positions in both cation and anion categories. For example, the top-left panel illustrates the generation of imidazolium-based cations with varying alkyl chain lengths and functional groups, and the bottom-left panel shows the generation of $[Tf_2N]^-$-derived anions with different combinations of alkyl chain lengths and functionalities. Example of various alkyl chains and functionalizations are shown on the right.

pond to increased synthetic complexity. In our earlier work, we computed COSMO-RS viscosities for several ionic liquids; however, we found that COSMO-RS often exhibits large deviations for IL viscosity.[11] In contrast, literature studies by Palomar *et al.* (2007)[72] and Alcantara *et al.* (2025)[73] have reported that COSMO-RS model provide lower deviations for density. Therefore, to externally validate one of our IL properties, we calculated IL densities using COSMO-RS and observed a linear correlation with the ML-predicted values (Fig. S6). These results confirm the robustness and transferability of our ML model.

To analyze relationships among IL properties, we calculated a correlation matrix between the properties, and the results are shown in Fig. 8. It is interesting to note that the density is positively correlated with the surface tension and toxicity, with Pearson correlation coefficients of 0.62–0.63. As proposed by MacLeod,[74] surface tension is directly proportional to the density, reflecting stronger cohesive interactions among tightly packed ions. As the temperature increases, molecules move farther apart, weakening intermolecular forces. According to van der Waals, these forces decrease with the fourth power of intermolecular distance.[74] As a result, surface tension of a liquid is a function of the distance between the molecules and thus depends on density. A negative correlation was observed between the viscosity and ionic conductivity, in line with the Stokes–Einstein equations, which demonstrates a strong relationship between these two properties.[75] Further, the molar volume of ILs shows a negative correlation with surface

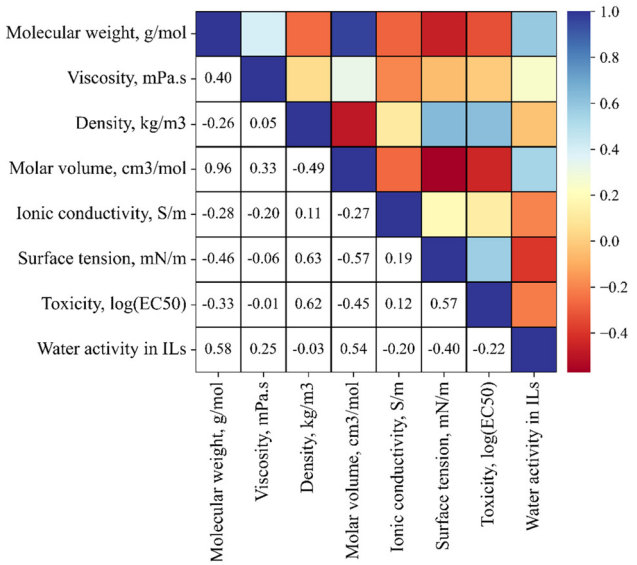


Fig. 8 Pearson correlation matrix of ionic liquid properties.

tension and this observation is in line with experiments, as the alkyl chain length of IL increases the surface tension decreases.[29] Higher alkyl chain length corresponds to higher molar volume. Gardas and Coutinho proposed an inverse relationship between surface tension and molar volume in ILs.[76] In addition, the positive correlation is observed between the surface tension and the toxicity, with a 0.57 Pearson R cor-

relation. Higher surface tension indicates strongly interacting ions that can disrupt biological membranes and increase toxicity.

For case studies, ILs were screened for their potential applications in lignocellulosic biomass processing, carbon capture, and electrolytes for lithium and sodium-ion batteries. Table S3 summarizes the desirable IL property ranges required for effective biomass pretreatment, high CO_2 capture, and act as an electrolyte in battery research. For lignocellulosic biomass pretreatment, the key characteristics of ILs include low viscosity, moderate surface tension and ionic conductivity, being liquid at room temperature, and low toxicity. Therefore, based on Table S3, the optimal design criteria for ILs in biomass applications are as follows: $\ln(\eta) < 5$ mPa s, σ: 30–45 mN m^{-1}, $\kappa < 0.8$ S m^{-1}; $\log_{10} EC_{50} > 2.1$; hydrophilic IL, synthesizability scores (SA score) < 6, and $T_m < 298.15$ K (liquid). Based on these criteria, 1937 ionic liquids were initially identified, and 206 ILs of these were selected based on the polarity for further

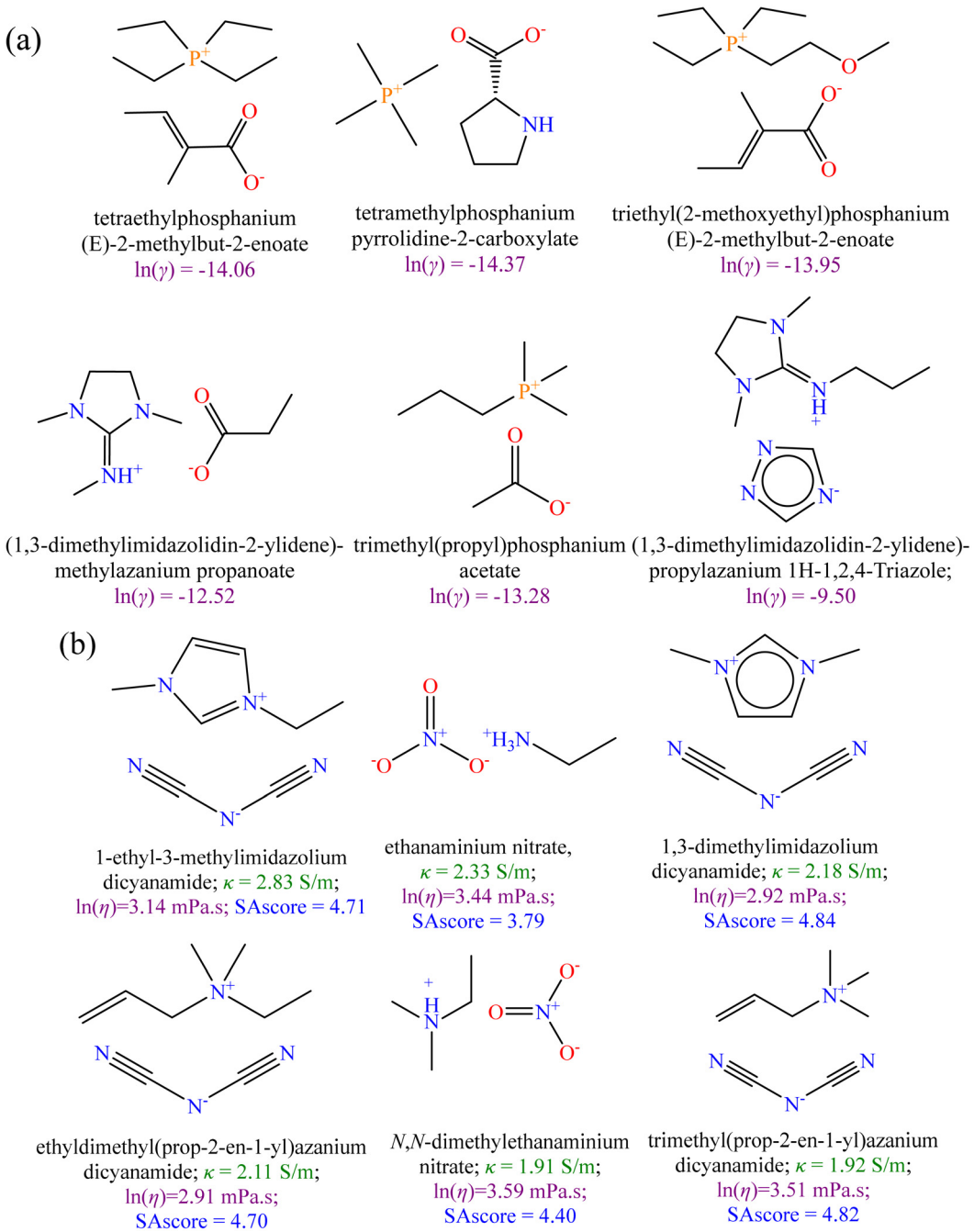


Fig. 9 (a) COSMO-RS predicted logarithmic activity coefficient of lignin in ML developed ionic liquids. (b) Ionic conductivity of newly engineered ionic liquids for battery research applications. All these ILs are predicted to be easy to synthesize with a SA score below 6.

evaluation using COSMO-RS calculations. For this, COMSO-RS was employed to calculate the logarithmic activity coefficient ($\ln(\gamma)$) and excess enthalpy (H^E, kJ mol^{-1}) of lignin, cellulose, and hemicellulose in ILs at 298.15 K (Fig. S7–S9); details of COSMO-RS methodology used for calculating $\ln(\gamma)$ and H^E can be found in our previous publications.[4,15,77,78] These two properties are critical, as low $\ln(\gamma)$ and H^E indicate strong interactions between the IL and lignin, enhancing lignin solubility and facilitating effective biomass fractionation. First, the $\ln(\gamma)$ and H^E were calculated for well-studied ILs from the literature, including 1-ethyl-3-methylimidazolium acetate ([EMIM][OAc]), 1-ethyl-3-methylimidazolium chloride ([EMIM]Cl), 1-allyl-3-methylimidazolium chloride ([AMIM]Cl), and choline lysinate ([Ch][Lys]).[4,79–81] The corresponding COSMO-RS values are provided in Table S4. We primarily focused on lignin solubility because complete lignin removal remains the main challenge for effective fractionation, and solvents capable of dissolving lignin are urgently needed. Based on these COSMO-RS results, an additional selection criterion for ILs targeting lignin solvation was established: $\ln(\gamma)_{\text{lignin}} < -8$ and $H^E < -5.0$ kcal mol^{-1}. From the set of 206 ILs, 57 ILs lower $\ln(\gamma)$ and H^E for lignin and the other stated desirable IL properties were thus retained (Fig. S7). Notably, phosphonium-based ILs dominate this pool of promising candidates. The top-predicted ILs, exhibiting lower $\ln(\gamma)$ and H^E than the literature reported ILs, include: tetraethylphosphonium (*E*)-2-methylbut-2-enoate, tetramethylphosphonium pyrrolidine-2-carboxylate, triethyl(2-methoxyethyl)phosphonium (*E*)-2-methylbut-2-enoate, (1,3-dimethylimidazolidin-2-ylidene)-methylazanium propanoate, and trimethyl(propyl)phosphonium acetate (Fig. 9a).

For carbon capture, reference properties of ILs were obtained from the literature and are listed in Table S3. The key selection criteria are: $\ln(\eta) < 4.5$ mPa s, $\sigma < 45$ mN m^{-1}, κ: 0.1–0.5 S m^{-1}; hydrophilic IL, SA score < 6, and liquid at ambient condition. Based on these criteria, 3986 ILs were retained. A number of interesting ILs were found based on the (methylsulfonyl)acetonitrile ([MSA]$^-$) anion. Recent work by Qiu *et al.* (2024)[69] explored the potential of [MSA]-based ILs for CO_2 capture, demonstrating a cascade insertion mechanism of two CO_2 molecules *via* consecutive C–C and O–C bond formation with [MSA]$^-$. In our screening, [MSA]$^-$ formed ILs with various cations exhibiting desirable properties for CO_2 capture. Furthermore, cyano (–C#N), carboxylate, borate, and TF_2N-derived ILs are also potential candidates for CO_2 capture. The chemical structures of a few selected ILs are illustrated in Fig. S10.

Finally, ILs were also assessed for their potential suitability in lithium- and sodium-ion batteries. The conventional Li-ion battery electrolyte, LP30, consists of 1 M $LiPF_6$ in a 1 : 1 mixture of ethyl carbonate (EC) and dimethylcarbonate (DMC), with an ionic conductivity of 1.26 at 298.15 K. To serve as a viable electrolyte additive or replacement for LP30, ILs must exhibit an ionic conductivity greater than 1.5 S m^{-1}. The criteria for selecting ILs as electrolytes thus comprised: $\ln(\eta) <$ 5 mPa s, $\kappa > 1.5$ S m^{-1}; $\log_{10} EC_{50} > 2.0$; SA score < 6; and T_m = liquid. Based on these parameters, 117 IL candidates were identified. The top predicted ILs share dicyanamide anions paired with imidazolium, triazolium, pyridinium, ammonium, and sulfonium cations, and the chemical structures of the top six are depicted in Fig. 9b.

In summary, our ML model with NLP-featurization offers a reliable and efficient tool for predicting desirable IL properties and enables large-scale high-throughput screening, paving the way for precision carbon capture and energy storage applications. In our future studies, we will expand the ML framework to include additional physicochemical properties, facilitating the design and generation of optimal ILs for diverse research applications, and experimentally validating the ML predicted results.

Conclusions

In this work, we demonstrate the potential of NLP-based molecular embeddings (Mol2vec) combined with the CATBoost ML method to accurately predict multiple key physicochemical properties of ILs. A comprehensive dataset of IL properties was compiled from literature sources and the NIST ILThermo database and was used for model development. In addition to Mol2vec featurization, we examined Morgan fingerprints, atom count, and COSMO-RS-derived sigma profiles as input features to predict IL properties. Among these features, Mol2vec consistently outperformed the sigma profiles and Morgan fingerprints, yielding higher R^2 and lower RMSE values for all properties. Sigma profiles do provide valuable interpretability, aiding in understanding feature importance and potential molecular interactions. However, they are computationally expensive to calculate for large sets of IL chemical libraries. For melting point predictions, the regression model exhibited weaker performance with an R^2 of 0.720 and an RMSE of 39.87 K. To address this, we performed an alternative classification task, which significantly improved predictive accuracy to 0.844, enabling better data classification. Overall, the combination of comprehensive molecular representation, contextual embedding, a rich and informative feature set, and efficient learning makes Mol2vec a powerful featurization tool for IL property prediction, resulting in its superior performance compared to Morgan FP, sigma profiles, and GNN-based techniques.

Furthermore, we generated ~10.61 million novel ILs by systematically combining 7200 cations and 1474 anions and calculating from them the seven critical physicochemical properties using Mol2vec featurization with the pre-trained ML models. Notably, among the predicted properties, surface tension exhibited the strongest positive linear correlation with IL toxicity, with a Pearson correlation coefficient of 0.81. Further, the density shows a positive correlation with both surface tension and toxicity, with correlation coefficients ranging from 0.66 to 0.68, and this observation is in line with the theoretically proposed MacLeod equation, which establishes a direct proportionality between surface tension and density.[74] Finally, we demonstrated the capability of the

Mol2vec-based ML model in high-throughput screening of ILs for specific topical research applications identifying promising IL candidates with desirable properties for lignocellulosic biomass, CO_2 capture, and electrolytes. These findings highlight the advantages of data-driven NLP approaches and pave the way for their integration into experimental high-throughput screening pipelines for chemical and materials discovery.

Conflicts of interest

The authors declare no competing financial interest.

Notes

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://energy.gov/downloads/doe-public-access-plan).

Data availability

Code and data for this study is available and can be accessed *via* the link to the GitHub repository at: https://github.com/MohanMood/NLP_Ionic-Liquid_Properties. The generated ILs (10.6 million) and their properties can be found at https://doi.org/10.5281/zenodo.15580854. For the ML algorithms, the authors used the Python packages CATBoost (https://catboost.ai/), word2vec (https://www.tensorflow.org/text/tutorials/word2vec), and scikit-learn (https://scikit-learn.org/stable/) packages. All the codes were run with Python3.

Supplementary information (SI): Relative deviations of ML models for the IL properties, IL property datasets, numbers of input features in the featurization techniques, COSMO-RS predicted activity coefficients and excess enthalpies of biomass in ILs, correlations between experimental *vs.* ML IL properties, and screened ILs for CO_2 capture. See DOI: https://doi.org/10.1039/d5gc02803e.

Acknowledgements

This work was supported and provided by the U. S. Department of Energy (DOE), Office of Science, through the Genomic Science Program, Office of Biological and Environmental Research (contract no. FWP ERKP752), and U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences (CSGB), Award No. DE-SC0022214; FWP 3ERKCG25.

References

1 M. Mohan, C. Balaji, V. V. Goud and T. Banerjee, *J. Solution Chem.*, 2015, **44**, 538–557.
2 M. Mohan, T. Banerjee and V. V. Goud, *ChemistrySelect*, 2016, **1**, 4823–4832.
3 M. Mohan, T. Banerjee and V. V. Goud, *ACS Omega*, 2018, **3**, 7358–7370.
4 M. Mohan, H. Choudhary, A. George, B. A. Simmons, K. Sale and J. M. Gladden, *Green Chem.*, 2021, **23**, 6020–6035.
5 N. V. Plechkova and K. R. Seddon, *Chem. Soc. Rev.*, 2008, **37**, 123–150.
6 C. Dou, H. Choudhary, Z. Wang, N. R. Baral, M. Mohan, R. Aguilar, A. Holiday, D. Banatao, S. Singh, C. D. Scown, J. D. Keasling, B. Simmons and N. Sun, *One Earth*, 2023, **6**(11), 1576–1590.
7 M. Mohan, N. N. Deshavath, T. Banerjee, V. V. Goud and V. V. Dasu, *Ind. Eng. Chem. Res.*, 2018, **57**, 10105–10117.
8 T. Numpilai, L. K. H. Pham and T. Witoon, *Ind. Eng. Chem. Res.*, 2024, **63**, 19865–19915.
9 M. Galiński, A. Lewandowski and I. Stępniak, *Electrochim. Acta*, 2006, **51**, 5567–5580.
10 R. Hayes, G. G. Warr and R. Atkin, *Chem. Rev.*, 2015, **115**, 6357–6426.
11 M. Mohan, K. D. Jetti, S. Guggilam, M. D. Smith, M. K. Kidder and J. C. Smith, *ACS Sustainable Chem. Eng.*, 2024, **12**, 7040–7054.
12 S. Koutsoukos, F. Philippi, F. Malaret and T. Welton, *Chem. Sci.*, 2021, **12**, 6820–6843.
13 A. P. Carneiro, C. Held, O. Rodriguez, G. Sadowski and E. A. Macedo, *J. Phys. Chem. B*, 2013, **117**, 9980–9995.
14 L. Constantinou and R. Gani, *AIChE J.*, 1994, **40**, 1697–1710.
15 M. Mohan, J. D. Keasling, B. A. Simmons and S. Singh, *Green Chem.*, 2022, **24**, 4140–4152.
16 A. Lee, S. Sarker, J. E. Saal, L. Ward, C. Borg, A. Mehta and C. Wolverton, *Commun. Mater.*, 2022, **3**, 73.
17 R. L. Gardas and J. A. Coutinho, *Fluid Phase Equilib.*, 2008, **266**, 195–201.
18 M. Mohan, K. L. Sale, R. S. Kalb, B. A. Simmons, J. M. Gladden and S. Singh, *ACS Sustainable Chem. Eng.*, 2022, **10**, 11016–11029.
19 M. Mohan, B. A. Simmons, K. L. Sale and S. Singh, *Sci. Rep.*, 2023, **13**, 271.
20 H. S. Majdi, A.-B. F. Raheem, S. J. Abdullah, I. M. Mohammed, Y. Yasin, A. Yadav, S. K. Hadrawi and R. Shariyati, *Chem. Eng. Sci.*, 2023, **265**, 118246.
21 M. Mohan, P. Viswanath, T. Banerjee and V. V. Goud, *Mol. Phys.*, 2018, **116**, 2108–2128.

22 M. Mohan, O. Demerdash, B. A. Simmons, J. C. Smith, M. K. K. Kidder and S. Singh, *Green Chem.*, 2023, **25**, 3475–3492.
23 D. L. Shrestha and D. P. Solomatine, *Neural Networks*, 2006, **19**, 225–235.
24 Z.-M. Win, A. M. Cheong and W. S. Hopkins, *J. Chem. Inf. Model.*, 2023, **63**, 1906–1913.
25 M. Mohan, K. D. Jetti, M. D. Smith, O. N. Demerdash, M. K. Kidder and J. C. Smith, *J. Chem. Theory Comput.*, 2024, **20**, 3911–3926.
26 J. C. Smith, M. D. Smith, S.-H. Liu, S. J. Rukmani, M. Mohan, Y. Yu and M. Goswami, *Biophys. J.*, 2025, DOI: 10.1016/j.bpj.2025.09.006.
27 M. Mohan, O. N. Demerdash, B. A. Simmons, S. Singh, M. K. Kidder and J. C. Smith, *ACS Omega*, 2024, **9**(17), 19548–19559.
28 M. Mohan, N. Gugulothu, S. Guggilam, T. R. Rajeshwar, M. K. Kidder and J. C. Smith, *Green Chem. Eng.*, 2024, **6**(2), 275–287.
29 M. Mohan, M. D. Smith, O. Demerdash, M. K. Kidder and J. C. Smith, *J. Chem. Phys.*, 2023, **158**, 214502.
30 M. Mohan, M. D. Smith, O. N. Demerdash, B. A. Simmons, S. Singh, M. K. Kidder and J. C. Smith, *ACS Sustainable Chem. Eng.*, 2023, **11**, 7809–7821.
31 S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model.*, 2018, **58**, 27–35.
32 D. H. Kenney, R. C. Paffenroth, M. T. Timko and A. R. Teixeira, *J. Cheminf.*, 2023, **15**, 9.
33 S. Chithrananda, G. Grand and B. Ramsundar, arXiv, 2020, preprint, arXiv:2010.09885, DOI: 10.48550/arXiv.2010.09885.
34 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
35 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
36 C. Song, C. Wang, F. Fang, G. Zhou, Z. Dai and Z. Yang, *J. Chem. Eng. Data*, 2024, **69**(12), 4310–4319.
37 Q. Dong, C. D. Muzny, A. Kazakov, V. Diky, J. W. Magee, J. A. Widegren, R. D. Chirico, K. N. Marsh and M. Frenkel, *J. Chem. Eng. Data*, 2007, **52**, 1151–1159.
38 A. F. Kazakov, J. W. Magee, R. D. Chirico, V. Paulechka, V. Diky, C. D. Muzny, K. G. Kroenlein and M. D. Frenkel, *NIST Standard Reference Database 147: NIST Ionic Liquids Database - (ILThermo). Ionic Liquids Database-ILThermo (v2. 0)*, https://ilthermo.boulder.nist.gov/, (accessed October 5, 2023).
39 D. M. Lowe, P. T. Corbett, P. Murray-Rust and R. C. Glen, *J. Chem. Inf. Model.*, 2011, **51**(3), 739–753.
40 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen and B. Yu, *Nucleic Acids Res.*, 2019, **47**, D1102–D1109.
41 N. Mills, *J. Am. Chem. Soc.*, 2006, **128**(41), 13649–13650.
42 H. Feng, L. Qin, B. Zhang and J. Zhou, *ACS Omega*, 2024, **9**, 16016–16025.
43 T. P. Adewumi, F. Liwicki and M. Liwicki, *Open Comput. Sci.*, 2022, **12**, 134–141.
44 M. Grohe, PODS'20: Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, 2020, pp. 1–16, DOI: 10.1145/3375395.3387641.
45 S. Zhong and X. Guan, *Environ. Sci. Technol.*, 2023, **57**, 18193–18202.
46 D. Y. Shi, F. Y. Zhou, W. B. Mu, C. Ling, T. C. Mu, G. Q. Yu and R. Q. Li, *Phys. Chem. Chem. Phys.*, 2022, **24**, 26029–26036.
47 A. V. Dorogush, V. Ershov and A. Gulin, arXiv, 2018, preprint, arXiv:1810.11363, DOI: 10.48550/arXiv.1810.11363.
48 J. T. Hancock and T. M. Khoshgoftaar, *J. Big Data*, 2020, **7**, 94.
49 T. Chen and C. Guestrin, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, DOI: 10.1145/2939672.2939785.
50 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, 3149–3157.
51 O. Kramer, in *Machine Learning for Evolution Strategies*, Springer, 2016, vol. 20, pp. 45–53.
52 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn Res.*, 2011, **12**, 2825–2830.
53 R. D. A. Fernandes, F. R. Seppe, C. T. Júnior, I. G. Torné, C. D.S Filho and S. B. Rego, 2023 15th IEEE International Conference on Industry Applications (INDUSCON), 2023, pp. 842–849, DOI: 10.1109/INDUSCON58041.2023.10374661.
54 A. Joshi, P. Saggar, R. Jain, M. Sharma, D. Gupta and A. Khanna, *Adv. Data Sci. Adapt. Anal.*, 2021, **13**, 2141002.
55 M. C. Cieslak, A. M. Castelfranco, V. Roncalli, P. H. Lenz and D. K. Hartline, *Mar. Genomics*, 2020, **51**, 100723.
56 K. López-Pérez, T. D. Kim and R. A. Miranda-Quintana, *Digital Discovery*, 2024, **3**, 1160–1171.
57 K. Lopez-Perez, B. Zhao and R. A. M. Miranda-Quintana, *J. Chem. Inf. Model.*, 2025, **65**(13), 6797–6808.
58 G. Furman, Enhancing generalizability and feasibility in sample selection: a methodological study of cluster analysis for stratifying populations, 2023, DOI: 10.26153/tsw/51489.
59 T. Lemaoui, T. Eid, A. S. Darwish, H. A. Arafat, F. Banat and I. AlNashef, *Mater. Sci. Eng., R*, 2024, **159**, 100798.
60 Z. Chen, J. Chen, Y. Qiu, J. Cheng, L. Chen, Z. Qi and Z. Song, *ACS Sustainable Chem. Eng.*, 2024, **12**, 6648–6658.
61 S. Zhong, Y. Chen, J. Li, T. Igou, A. Xiong, J. Guan, Z. Dai, X. Cai, X. Qu and Y. Chen, *Environ. Sci. Technol. Lett.*, 2024, **11**, 1193–1199.
62 V. Venkatraman, S. Evjen, H. K. Knuutila, A. Fiksdahl and B. K. Alsberg, *J. Mol. Liq.*, 2018, **264**, 318–326.
63 D. M. Makarov, Y. A. Fadeeva, L. E. Shmukler and I. V. Tetko, *J. Mol. Liq.*, 2022, **366**, 120247.
64 M. Mohan, T. Banerjee and V. V. Goud, *J. Chem. Eng. Data*, 2016, **61**, 2923–2932.
65 K. Paduszyński, *J. Chem. Inf. Model.*, 2016, **56**, 1420–1437.

66 X. Liu, M. Yu, Q. Jia, F. Yan, Y.-N. Zhou and Q. Wang, *J. Mol. Liq.*, 2023, **388**, 122711.
67 G. Chen, Z. Song, Z. Qi and K. Sundmacher, *Digital Discovery*, 2023, **2**, 591–601.
68 X. Liu, Y. Gu, M. Yu, Q. Jia, Y.-N. Zhou, F. Yan and Q. Wang, *ACS Sustainable Chem. Eng.*, 2023, **11**, 13429–13440.
69 L. Qiu, B. Li, J. Hu, A. Ganesan, S. Pramanik, J. T. Damron, E. Li, D.-e. Jiang, S. M. Mahurin and I. Popovs, *J. Am. Chem. Soc.*, 2024, **146**, 29588–29598.
70 G. Landrum, *Greg Landrum*, 2013, **8**, 31.
71 P. Ertl and A. Schuffenhauer, *J. Cheminf.*, 2009, **1**, 1–11.
72 J. Palomar, V. R. Ferro, J. S. Torrecilla and F. Rodríguez, *Ind. Eng. Chem. Res.*, 2007, **46**, 6041–6048.
73 M. L. Alcantara, G. L. Bressan, P. V. Santos, M. F. Nobre, J. A. Coutinho, C. A. Nascimento and L. A. Follegatti-Romero, *J. Mol. Liq.*, 2025, **417**, 126616.
74 D. Macleod, *Trans. Faraday Soc.*, 1923, **19**, 38–41.
75 Z. K. Koi, W. Z. N. Yahya and K. A. Kurnia, *New J. Chem.*, 2021, **45**, 18584–18597.
76 R. L. Gardas and J. A. Coutinho, *Fluid Phase Equilib.*, 2008, **265**, 57–65.
77 E. C. Achinivu, M. Mohan, H. Choudhary, L. Das, K. Huang, H. D. Magurudeniya, V. R. Pidatala, A. George, B. A. Simmons and J. M. Gladden, *Green Chem.*, 2021, **23**, 7269–7289.
78 K. Huang, M. Mohan, A. George, B. A. Simmons, Y. Xu and J. M. Gladden, *Green Chem.*, 2021, **23**, 6036–6049.
79 M. Mohan, K. Huang, V. R. Pidatala, B. A. Simmons, S. Singh, K. L. Sale and J. M. Gladden, *Green Chem.*, 2022, **24**, 1165–1176.
80 E. K. Coronado-Aldana, C. L. Ferreira-Salazar, N. Y. Piñeros-Castro, R. Vázquez-Medina and F. A. Perdomo, *Chin. J. Chem. Eng.*, 2023, **60**, 143–154.
81 J.-F. Liu, Y. Cao, M.-H. Yang, X.-J. Wang, H.-Q. Li and J.-M. Xing, *Chin. Chem. Lett.*, 2014, **25**, 1485–1488.
82 Z. Wang, Z. Song and T. Zhou, *Processes*, 2020, **9**, 65.
83 A. Sosnowska, M. Grzonkowska and T. Puzyn, *J. Mol. Liq.*, 2017, **231**, 333–340.
84 H. Benimam, C. Si-Moussa, M. Laidi and S. Hanini, *Neural Comput. Appl.*, 2020, **32**, 8635–8653.