ACS Partner Journal
CHEM & BIO ENGINEERING

pubs.acs.org/ChemBioEng
Review

Solvent Screening for Separation Processes Using Machine Learning and High-Throughput Technologies

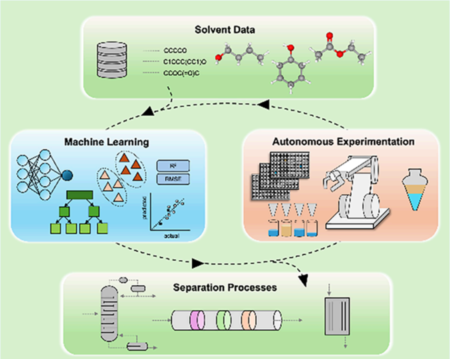
Justin P. Edaugal, Difan Zhang,* Dupeng Liu,* Vassiliki-Alexandra Glezakou, and Ning Sun*

Cite This: *Chem Bio Eng.* 2025, 2, 210−228

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: As the chemical industry shifts toward sustainable practices, there is a growing initiative to replace conventional fossil-derived solvents with environmentally friendly alternatives such as ionic liquids (ILs) and deep eutectic solvents (DESs). Artificial intelligence (AI) plays a key role in the discovery and design of novel solvents and the development of green processes. This review explores the latest advancements in AI-assisted solvent screening with a specific focus on machine learning (ML) models for physicochemical property prediction and separation process design. Additionally, this paper highlights recent progress in the development of automated high-throughput (HT) platforms for solvent screening. Finally, this paper discusses the challenges and prospects of ML-driven HT strategies for green solvent design and optimization. To this end, this review provides key insights to advance solvent screening strategies for future chemical and separation processes.



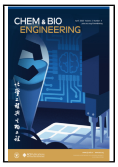
KEYWORDS: *Ionic liquids, Deep eutectic solvents, Artificial intelligence, Machine learning, Solvent extraction, High-throughput screening*

1. INTRODUCTION

Separation processes play an integral role in the production of a diverse range of products across the oil and gas, food, pharmaceutical, and chemical industries. Traditionally, these processes utilize large-scale physical and thermal treatments, such as centrifugation, filtration, and distillation, for the removal of various impurities and byproducts. However, such processes are energetically intensive and costly, accounting for approximately 10−15% of global energy consumption.[1] Solvent extraction, otherwise known as liquid−liquid extraction (LLE), is a well-established operation for the effective separation of a wide range of synthetic and bio-based compounds from liquid mixtures. By leveraging the differential affinity of target compounds across immiscible liquid phases, solvent extraction can achieve separation with a high extraction efficiency and selectivity. Additionally, solvent extraction offers a simpler process setup and reduced energy costs, positioning it as an attractive strategy for industrial-scale application.

Despite these advantages, the efficiency of an extraction process significantly hinges on the selection of the appropriate solvent. The optimal extractant solvent should exhibit a high distribution coefficient, low mutual solubility, low toxicity, high chemical stability, and low cost.[2] Most industrial extraction processes rely on conventional organic solvents, owing to their aqueous immiscibility and relatively low bulk costs. However, the environmental concerns regarding their nonrenewable, fossil-fuel origins, as well as their potentially high toxicity and volatility, have prompted the search for greener and safer alternatives.[3] In recent years, ionic liquids (ILs) and deep eutectic solvents (DESs) have emerged as potential "green" solvent alternatives, owing to their tunable properties and proposed environmental benefits.[4,5] Specifically, ILs are a class of salts composed of cations and anions with melting points below 100 °C, typically characterized by low volatility, minimal flammability, and high thermal stability. DESs are a neoteric class of solvents derived from eutectic mixtures of hydrogen bond donors (HBDs) and hydrogen bond acceptors (HBAs), which are reported to offer several key advantages including simpler synthesis, lower cost, reduced toxicity, and enhanced biodegradability.[6−8] The structures and properties of both ILs and DESs are highly customizable due to the vast number of possible cation/anions and HBA/HBD combinations.[9] While this flexibility enables their tailored use in a wide range of applications, it also introduces variability in the understanding of their structure−property relationships, often requiring case-by-case experimentation and validation, especially in more complex processes such as chemical and solvent-based separations.[10] As a result, their effectiveness in industrial-scale processes and their long-term environmental sustainability remain subjects of ongoing debate.[11,12] Given the broad

Received: November 7, 2024
Revised: February 13, 2025
Accepted: February 16, 2025
Published: March 5, 2025



https://doi.org/10.1021/cbe.4c00170

range of possible IL and DES structures, along with the time-consuming and costly nature of conventional experimentation, there is a growing demand for faster, more efficient strategies for green solvent screening and design.

To accelerate the solvent screening process, researchers have turned to the use of *in silico* methods for property estimation. Several useful methods have been broadly utilized in the past decades, including molecular-scale modeling, macroscale modeling, and empirical/semiempirical methods.[13] Molecular-scale modeling, such as Density Functional Theory (DFT) and classical molecular dynamics, describes atomic bonding and elucidates the interactions of particles (atoms or electrons) at molecular levels. These methods have offered valuable insights into the fundamental molecular behavior, aiding researchers in the synthesis of new materials and the understanding of solvent−solute interactions.[13−15] In the context of solvent-based separation processes, a notable example is the study by Sun et al., which employed DFT calculations to analyze the extraction mechanism of metal ions in a tributyl phosphate (TBP)−$FeCl_3$/brine system for lithium recovery.[16] Their DFT results revealed that the electrostatic interactions between the solvent and central metal ions were the driving force behind the extraction and co-extraction processes. Another recent example is the work by Kaim-Sevalneva et al., which explored the solvent extraction of scandium(III) from rare earth elements using quaternary ammonium-based ILs.[17] By applying DFT methods, they were able to validate the neutral exchange mechanism behind the selective extraction of scandium(III). Similarly, other studies have combined experimental work and DFT or classical molecular dynamics simulations to elucidate the molecular mechanisms underlying various extraction processes.[18−21] Macroscale modeling, also known as thermodynamic modeling, utilizes mathematical relations based on thermodynamic principles to describe bulk system properties, such as phase transitions and heat transfer. One well-established model is COSMO-RS (Conductor-like Screening Model for Realistic Solvation), widely recognized for its effectiveness in estimating a variety of bulk solvent properties, such as solubilities, partition coefficients, and vapor−liquid equilibria.[22] A notable example is the study by Filly et al., which utilized COSMO-RS to estimate the σ-profiles of aroma compounds within eight organic solvents, revealing ethyl acetate and dimethyl carbonate as more efficient alternatives to *n*-hexane for aroma extraction from caraway seeds.[23] Similarly, Wojeicchowski et al. developed a range of hydrophobic DESs specifically for extracting carnosic acid and carnosols from rosemary by utilizing COSMO-RS to determine the optimal pairings of hydrogen bond acceptors and donors.[24] The readers can refer to several review articles[13,14,25] regarding more applications of COSMO-RS methods in solvent screening. Moreover, empirical/semiempirical methods blend theoretical and empirical data to estimate molecular properties. GC-based methodologies effectively bridge molecular-scale and macroscale contexts by employing a semiempirical, additive framework. These approaches decompose molecules into functional groups, assign quantitative parameters to each, and sum these values to estimate the system properties. The UNIFAC (Universal Quasi-Chemical Functional Group Activity Coefficient) and UNIQUAC (Universal Quasi-Chemical Activity Coefficient) methods have been widely employed in estimating the properties of a variety of solvents. Birajdar et al., for instance, applied an optimized UNIFAC model to predict the distribution coefficients of bio-based 2,3-butanediol in various organic solvents from aqueous fermentation broth. Their model successfully identified 1-butanol and 2-secondary butyl phenol as the optimal extractant solvents for 2,3-butanediol recovery.[26] In addition, the hybrid method, such as the integration of COSMO-RS and GC into the GC-COSMO method,[27,28] shows higher efficiency and reliability. Researchers have also leveraged the combination of GC-based methods with other traditional property estimation methods for solvent screening. For instance, Mu and Gmehling first proposed a combined GC-COSMO model, integrating GC methods with the COSMO-RS and COSMO-SAC models, to enhance their predictive accuracy.[29,30] A study by Dong et al. employed a combined COSMO-SAC-UNIFAC model to predict phase equilibrium data in IL systems.[31] Their results indicated that the combined model provided relatively accurate predictions, even in the absence of certain UNIFAC group binary parameters, suggesting its potential utility in data sets with missing parameters. In a subsequent study, Zhu et al. incorporated 648 new vacant parameter pairs for 51 main functional groups into the COSMO-SAC-UNIFAC model to expand the model's applicability to a broader range of IL systems.[32] In doing so, their model produced more accurate predictions of vapor−liquid equilibrium and liquid−liquid equilibrium data for 16 binary and 2 ternary systems, outperforming both the individual COSMO-SAC and UNIFAC models. Additionally, Liu et al. developed a GC-COSMO model to suggest new reaction kinetics for reaction solvents in organic synthesis and validated their model in the selected Menschutkin reaction.[33] Peng et al. built an extended GC-COSMO model to estimate the σ-profile and cavity volume of ionic liquid for separation processes and tested their method in the extraction of benzene from cyclohexane and postcombustion carbon capture.[34] Despite their widespread application, these traditional methods for property estimation still face several inherent challenges, including high computational costs and slow processing times.[13] Furthermore, these methods heavily rely on theoretical frameworks, which may fall short of accurately capturing real-world complexities and nuances observed in actual data. Thus, more advanced, real-world, data-driven models are highly desirable.

Machine learning (ML) offers a promising approach for solvent screening, leveraging mathematical relations to analyze patterns and unveil connections within large and complex data sets. In the context of solvent screening, ML models can be employed to rapidly analyze and predict solvent−solute interactions, optimize recovery yields, and enhance selectivity for a wide range of chemical and separation processes.[35−38] Additionally, ML models can "learn" from new experimental and/or computational data, drive self-optimization, and as a result potentially surpass traditional property estimation methods in both accuracy and efficiency, for effective solvent screening.[39,40] This review explores the latest advancements in AI-assisted solvent screening with a specific focus on ML models for solvent physicochemical property prediction and separation process parameter design. Additionally, this review highlights recent progress in the development of automated high-throughput (HT) platforms for solvent screening and investigation. Furthermore, the review addresses the challenges and future opportunities for integrating ML-driven HT approaches in the screening, design, and optimization of green solvents (Figure 1).

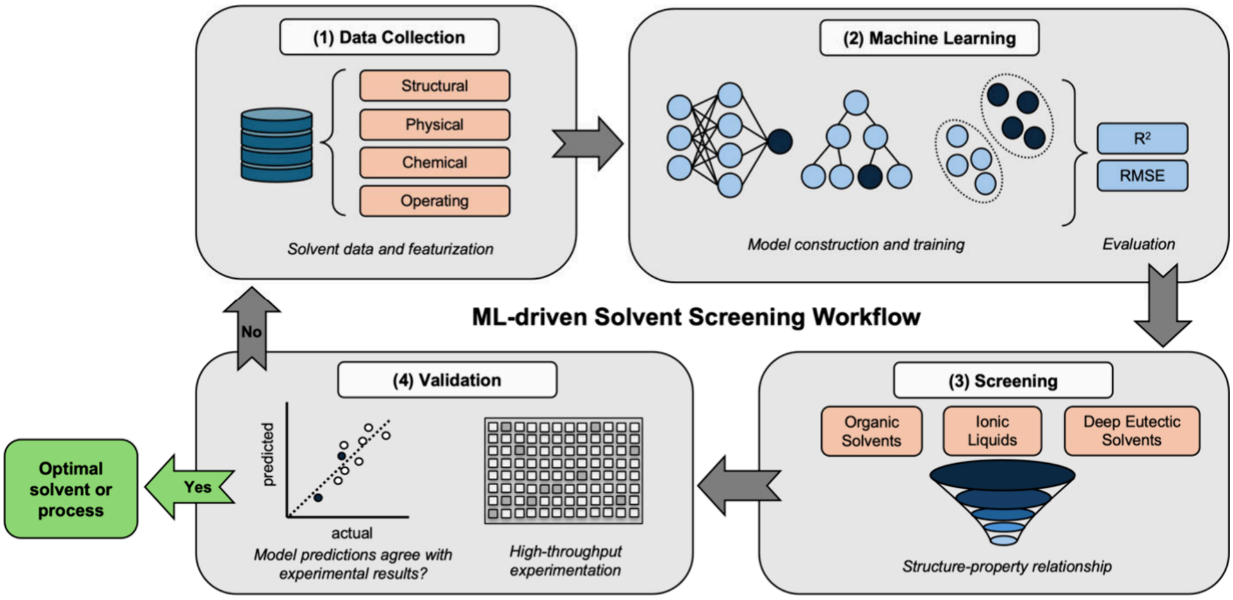

(1) Data Collection
Structural
Physical
Chemical
Operating
Solvent data and featurization
(2) Machine Learning
R²
RMSE
Model construction and training
Evaluation
ML-driven Solvent Screening Workflow
No
(4) Validation
predicted
actual
Model predictions agree with experimental results?
High-throughput experimentation
Yes
Optimal solvent or process
(3) Screening
Organic Solvents
Ionic Liquids
Deep Eutectic Solvents
Structure-property relationship

Figure 1. Machine-learning-driven framework for high-throughput solvent screening.

2. MACHINE LEARNING MODELING

The integration of ML in chemical engineering endeavors is not a novel concept. As far back as early 1980s, researchers had already crafted an expert system designed to forecast the physical properties of intricate fluid mixtures.[41] Subsequently, the swift adoption of neural network methodologies revealed their intrinsic ability to extract insights from vast data sets, consequently enhancing the precision of modeling for complex systems. Nonetheless, the initial application of ML encountered significant hurdles, including insufficient data, limited data accessibility and inadequate computational resources.[41] These limitations hindered ML from realizing its anticipated revolutionary impact. However, over the past few decades, the introduction of crucial technologies, including more complex deep learning architectures such as convolutional neural nets (CNNs) and recurrent neural networks (RNN), advanced parametrization methods such as reinforcement learning and transfer learning, as well as hardware advances such as GPU computing, has the capability to address more complicated "big data" domain problems.[42,43] Notably, the recent advent of Generative Pretrained Transformer (GPT) models, which develop specialized large-scale language models (LLMs), has offered an accessible platform that enhances text mining and complex data analysis.[44] This advancement significantly speeds up the traditional research methodology, marking a pivotal shift in how data is utilized and understood in various fields.[45] In addition, a deep generative model such as stable diffusion and transformer has also emerged as a new powerful family of AI with record-breaking performance in graph-based applications, language models, and beyond.[46−49]

2.1. ML Solvent Screening Workflow. In the context of solvent screening, the ML process requires a balanced integration of chemistry and chemical engineering knowledge with computer science methodologies. This integrated ML process operates through a systematic workflow, resulting in the development of algorithms designed to identify patterns from data and generalize those patterns to make predictions or decisions based on unseen data. This can be distilled into three key stages: data collection and preprocessing, model development, and model deployment.

In the initial stage, researchers compile relevant empirical and/or computational data, typically obtained from previous literature and/or databases such as QM9,[50] GDB-17,[51] PubChem,[52] ChEMBL,[53] PDBbind,[54] and ChemSpider,[55] to serve as the input data set for a given ML model. Additionally, databases such as NIST Chemistry WebBook,[56] AqSolDB,[57] Dortmund Data Bank,[58] and Cambridge Structural Database[59] are widely used repositories that provide key molecular data that are directly relevant to solvent screening and solvent-based separations, including solvent properties, aqueous solubilities, phase equilibria, molecular bonding interactions and structural details. The obtained data set must be machine-readable in order to be incorporated into ML models. Researchers can employ various methods to convert and display solvent data for computational analyses. One common approach involves the use of string-based representations, such as International Chemical Identifier (InChI), MDL Molfiles, and Chemical Markup Language (CML), which utilize sequences of letters, numbers, and/or symbols to depict molecular structures in a text-based format suitable for computational processing. Notably, the Simplified Molecular Input Line Entry System (SMILES) notation and its derivatives, such as SMILES Arbitrary Target Specification (SMARTS), are among the most widely adopted string-based methods for the representation of a variety of organic and inorganic compounds. It has enabled the processing of several complex solvent structures using ML models, including regression and deep learning algorithms.[60−62] Alternatively, graph-based representations and molecular-fingerprint-based representations have also been employed to represent the molecular structures of numerous solvents and compounds for ML models. Each of these notational methods offers unique advantages in terms of detail, specificity, and compatibility with computational tools and, therefore, should be guided by the specific requirements of the data set in question.[63] The readers can refer to previous works for detailed discussions on different fingerprints or molecular features.[64,65]

In addition, solvent data may be represented through property-based features. These features can serve as supplemental input data for ML models to further enhance the accuracy of the solvent predictions. For example, researchers can generate specific structural features, known as structural descriptors, from solvent data, such as bond length, bond angle, and molecular weight, typically through numerical, vector, or matrix configurations.[66,67] Notably, a study by Aghaie and Zendehboudi demonstrated the prediction of CO_2 solubility in ILs using four distinct ML models.[68] These ML models were fed with two separate CO_2 solubility data sets, which were characterized by either thermodynamic properties or structural descriptors. Interestingly, the ML models trained using the structural descriptor data set generally exhibited higher predictive accuracies when compared to those based on the thermodynamic properties, suggesting a strong relationship between anionic and cationic structural features for dictating CO_2 solubility.

Solvent data may be further characterized by their distinct chemical and physical characteristics, which are otherwise known as physicochemical descriptors. These descriptors, which encompass electron charge, electron distribution, and solvation energies, are frequently computed by using thermodynamic and quantum chemistry based property estimation methods. A noteworthy example is the study by Amar et al., which utilized a hybridized ML solvent screening approach to identify optimal solvents for asymmetric catalysis.[69] In particular, Amar et al. utilized a library of 459 solvents, calculating 12 conventional molecular descriptors, two reaction-specific descriptors, and additional descriptors based on screening charge density. Cross-validation analyses revealed that the combination of both physicochemical descriptors and screening charge density descriptors resulted in enhanced accuracy for predicting the reactive conversion in solvents. Though, it is essential to note that most raw solvent data, along with structural and/or physicochemical descriptors, may still present irregularities such as missing values, outliers, or inconsistent units, which can significantly hinder the overall performance of a constructed ML model. Hence, researchers typically subject raw data sets to preprocessing, which includes data cleaning, scaling, and/or normalization.

Given the inherent complexity of molecular properties, no single molecular descriptor or fingerprint is capable of fully characterizing all aspects of these properties. Thus, the selection of molecular representation is often problem dependent. The scalar descriptors (also referred to as zero-dimensional descriptors) contain very general information on molecules, such as molecular weight and number of hydrogens, but lack structural information in detail. Since they are usually easy and quick to compute, they are often considered together with other descriptors in practice. The one-dimensional descriptors, such as fingerprints, contain a string-like vector with substructural information on a particular pattern (e.g., whether a chemical moiety exists). They are also often easily computed and may be sufficient when the targeting property is highly correlated to a specific substructure (e.g., acidity from acid groups). The two-dimensional descriptors offer more detailed structural information, such as connectivity, symmetry, and shape, and are mostly computationally feasible within modern cheminformatics. Therefore, they have been widely used with machine learning.[70,71] The three-dimensional descriptors are typically derived from the spatial coordinates of atoms within a molecule, providing comprehensive geometric information about its structure. However, it is usually time-consuming to calculate these descriptors due to their inherent complexity. Nonetheless, the three-dimensional representation must be employed in certain cases, such as atomic force evaluation. The reader can refer to other articles that comprehensively reviewed molecular representations.[72−75]

In the second stage, researchers construct and train an ML model using the preprocessed input data. Generally, ML models can be classified into three main categories: supervised learning, unsupervised learning, and reinforcement learning.[42,76] Supervised learning involves training an ML algorithm on a labeled input data set with the goal of predicting a labeled output or target variable. This includes algorithms such as Linear Regressions (LR), Logistic Regression (LogReg), Artificial Neural Networks (ANNs), Support Vector Machines (SVMs)/Support Vector Regressions (SVRs), Decision Trees (DTs), and Random Forest (RF). The parametrization of these models typically aims to map the correlation between the solvent descriptors as input and the targeting property (e.g., solubility) as output, enabling a predictive evaluation of unseen solvents. The readers can find more comprehensive explanations of these algorithms in previous publications.[76−79] In contrast, unsupervised learning involves training an ML algorithm on data without labeled responses to infer the inherent structure or pattern present within a set of data points. Unsupervised learning algorithms, such as K-Means Clustering, Hierarchical Clustering, Principal Component Analysis (PCA), and Association Rule Learning (ARL)/Association Rule Mining (ARM), aim to explore large data sets and freely identify inherent patterns typically through clustering analyses. These models reveal the underlying patterns of a group of solvents and allow for efficient evaluation of new solvents that exhibit desired properties. Several previous articles have provided more explanations regarding these unsupervised learning methods.[76,80−83] The third category of ML models is reinforcement learning, in which an agent learns to form decisions through feedback within a given system or environment. Examples include Q-Learning, Deep Q-Networks, and Monte Carlo Tree Search, which aim to uncover the optimal behavior or patterns within a data set to obtain the maximum reward or output.

It is important to note that each of these ML algorithms possess unique strengths and limitations, particularly in the context of solvent screening and separation process design.[84,85] For instance, ANNs are among the most popular ML models used in these applications.[86] ANNs are composed of multiple layers of processing units, or "nodes", which receive data inputs, compute weighted values based on these inputs, and generate an output value.[87] These values are then passed as inputs to the subsequent layer of nodes, progressing through all layers in the network until reaching the final output layer, resulting in the final prediction. ANNs are particularly valuable for identifying complex, nonlinear relationships within intricate data sets such as molecular structural data and real-time separation process data.[86] However, this advantage comes with its trade-offs, as ANNs often require large training data sets, which can lead to increased training time and computational resource demands.[88] Additionally, ANN models can be susceptible to data overfitting if they are overtrained or have too many layers.[89] SVMs are another widely used ML model in solvent screening. Specifically, these models represent data points as vectors in space and aim to identify the optimal boundary or hyperplane that maximizes the margin between distinct groups of vectors. By utilizing quadratic functions, SVMs can effectively model both linear and nonlinear interactions, only requiring small training data sets, which can enable quicker analyses and enhance generalization.[90] This performance of SVMs, however, can be compromised when applied to imbalanced data sets, as these models tend to bias toward the majority class of data points that are present.[91] Besides, LR is relatively simple and interpretable, but it is also sensitive to multicollinearity. PCA is more data-driven with less user-dependent parameters to tune but may show a loss of interpretability with new variables created. Q-learning is adaptable to learn optimal results through trial and error, but it is sensitive to hyperparameters and may also show convergence issues. Readers can refer to earlier articles for further details on ML model types.[92] To this end, researchers must select the appropriate model (or combination of models) and descriptors based on the specific objectives of their tasks as well as the size and complexity of the respective data set.

The final stage of the ML process involves deployment of the trained ML model and evaluation of its predictive performance. The performance evaluation of a ML model is crucial and is conducted using various metrics and statistical techniques tailored to the specific task at hand, such as regression, classification, or clustering. In the context of solvent screening, where the goal is often to quantify complex structure−property relationships and optimize process parameters, regression models are predominantly used due to their capability to provide quantitative evaluation of a given solvent property. To gauge the performance of such models, researchers rely on key statistical metrics, such as the coefficient of determination (R^2) and the root-mean-square error (RMSE), which serve as comparative benchmarks regarding accuracy and precision. More specifically, R^2 is a statistical measure that represents the proportion of variation in a dependent variable that can be predicted by an independent variable. R^2 is mathematically expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2} \tag{1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\overline{y}_i$ is the mean of the actual values. The RMSE measures the average difference between a model's predicted and actual values. The RMSE is mathematically expressed as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)} \tag{2}$$

where n is the number of data points, y_i is the actual value, and $\hat{y}_i$ is the predicted value. When assessing the performance of different ML models for solvent screening applications, an R^2 score close to 1 alongside a low RMSE is greatly desired, indicating a high degree of predictive accuracy and precision. It is also worthy to note that the predictive capability of a trained ML model should be evaluated based on not only its own testing or validation data sets but also unseen external data from other resources. Readers can refer to educational articles for more discussions on developing ML models.[93]

2.2. Prediction of Solvent Physicochemical Properties via ML Models.

A thorough understanding of a solvent's physicochemical properties is crucial for its effective use in chemical and separation processes. As highlighted earlier, the efficacy of a chemical or separation process is contingent on several factors, such as the selection of a solvent system and the resulting interactions with targeted compounds or reagents. While previous experimental studies have documented numerous solvents and their physicochemical properties, the vast array of untested solvent structures and solvent mixtures, compounded with the immense number of potential applications, renders the trial-and-error methodology of conventional experimentation impractical. To solve this constraint, researchers have utilized ML algorithms to rapidly predict the physicochemical properties of organic and ionic solvents, as summarized in Table 1.

In the context of organic solvent screening, a prominent study by Boobier et al. successfully demonstrated the application of ML for the prediction of organic solvent and water solubilities.[94] By leveraging a curated set of 14 molecular descriptors, the ExtraTrees (ET) model achieved high predictive accuracy, significantly outperforming conventional *in silico* prediction tools including AquaSol, EPI Suite, and COSMO*therm*. Another notable example is the study by Saini and Kumar, which involved the development of an ANN model with four quantum chemical descriptors, one topological descriptor, and one categorical descriptor to predict the polarity values of various organic solvents.[95] Despite the high predictive accuracy of their model, the reliance on proprietary software for calculating the quantum chemical descriptors could restrict its broader applicability. To address this limitation, Saini and Singh streamlined their approach in a later study by implementing feature importance and reduction techniques.[96] Their refined ANN model eliminated the need for categorical descriptors while maintaining high predictive performance, making it a simpler, quicker, and more effective tool for ML-based organic solvent screening.

In addition to polarity, predicting the Gibbs free energy of solvation (ΔG_{sol}) for organic solvents, an important but challenging task, has also been achieved via the ML approach. For example, Ferraz-Caetano et al. used experimentally measured ΔG_{sol} along with structural and electronic descriptors, such as chi connectivity indices, topological polar surface area, volume surface area state, molecular weight, molar refractivity, and electrotopological state atom, to parametrize several regression models including RF, gradient boosting (GB), SVM, and ANN.[67] They identified RF and GB as the top two models that can reach improved performances for an RMSE of 0.42 kcal/mol. Similarly, Lim and Jung developed an ANN model for ΔG_{sol} of organic solvents using structural descriptors encoded into feature vectors, and they benchmarked the model's performance against other theories.[97] They achieved a MAE of 0.24 kcal/mol for nonaqueous systems, and their prediction displays comparable results to several quantum-mechanism-based solvation models. Later, they further improved their modeling based on the group-contribution method by considering pairwise atomic interactions in the atomic feature vectors.[98] This led to a ML model exhibiting the potential to offer more physicochemical insights on the solvation process. There are also several other studies reporting the successful application of ML in predicting ΔG_{sol} of organic solvents, but due to the length limit of this article, we only demonstrated them in Table 1 and omitted their detailed discussion in our text. Overall, the prediction of ΔG_{sol} has been successfully achieved by various ML methods with high accuracy. However, compared to the aqueous phase, the available data of ΔG_{sol} for organic solvent are relatively much

Table 1. Summary of Studies Using ML Models for the Prediction of Solvent Physicochemical Properties

Solvent	Property	Descriptors	No. of data points (training/test)	ML model	R^2	Ref.
Organic	Solubility	Structural descriptors, quantum chemical descriptors	Water = 805/95, Water = 445/115, Ethanol = 553/142, Benzene = 370/94, Acetone = 360/92	MLR	0.64	94
				PLS	0.64	
				ANN	0.67	
				SVM	0.71	
				GP	0.70	
	Polarity	Quantitative structure, DFT descriptors	378/53	RF	0.90	60
				ANN	0.97	
	Polarity	Structural, quantum chemical and thermodynamic descriptors	378/43	MLR	0.67	95
				PLS	0.66	
				KNN	0.81	
				SVM	0.86	
				ET	0.95	
				BR	0.90	
				RF	0.93	
				NN	0.96	
	Polarity	Quantitative structure, PaDEL descriptors, PaDEL fingerprints	378/53	NN	0.97	96
	Reactivity (conversion; diastereomeric excess)	Screening charge density, COSMO*therm* descriptors	25/9	GP	Accuracy = 0.94	69
	Gibbs free energy of solvation	RDKit structural descriptors	513/64 and 4477/559	RF	0.99 (RMSE = 0.5 kcal/mol)	67
				GB	0.99 (RMSE = 0.43 kcal/mol)	
	Gibbs free energy of solvation	Morgan structural descriptors encoded in fixed-size feature vectors	1996/249 and 514/64	ANN	0.96 (RMSE = 0.24 kcal/mol)	97
	Gibbs free energy of solvation	Molecular graph	5615/623	Message passing NN	0.98 (MAE = 0.16 kcal/mol)	164
	Gibbs free energy of solvation	Molecular graph	10145 (total)	Message passing NN	No R^2 reported (RMSE = 0.44 kcal/mol, MAE = 0.21 kcal/mol)	165
	Gibbs free energy of solvation	Molecular graph	Solute = 8366 (total), Solvation free energy = 20253 (total), Solvation enthalpy = 6322 (total)	Message passing NN	0.93 (RMSE = 1.32 kcal/mol, MAE = 0.91 kcal/mol)	166
	Gibbs free energy of solvation	Structural descriptors, thermodynamic descriptors	Self-solvation free energy = 211 (total)	SVR	(RMSE = 0.3 kcal/mol, MAE = 0.2 kcal/mol)	167
IL-organic mixtures	Surface tension and viscosity	Group contribution	Surface tension = 2430/511; Viscosity = 19872/4282	ANN	0.97; 0.96	103
				XGB	0.98; 0.86	
				LightGBM	0.92; 0.82	
	Heat capacity and density	Group contribution	Heat capacity = 1601/353; Density = 24282/5335	ANN	0.99; 0.99	104
				XGB	0.99; 0.98	
				LightGBM	0.99; 0.98	
IL	CO_2 solubility	Thermodynamic properties, structural descriptors	1655 (total)	DT	0.94; 0.96	68
				RF	0.96; 0.98	
				LSSVM	0.75; 0.76	
				MLR	0.55; 0.28	
	CO_2 solubility	Group contribution, temperature, pressure	8093/2023	ANN	0.98	105
				SVM	0.98	
	CO_2 solubility	Structural (experimental)	10116 (total)	ANN	>0.98 (RMSE < 0.025)	15
	Toxicity	Electrostatic potential surface area, charge distribution area	99/20	ELM	0.94	106

Table 1. continued

Solvent	Property	Descriptors	No. of data points (training/test)	ML model	R^2	Ref.
				MLR	0.87	
				SVM	0.89	
	Toxicity	Quantitative structure (SMILES), structural descriptors	284/71	ANN	0.89	99
				SVM	0.92	
	Toxicity	Quantitative structure (SMILES), group contribution, structural descriptors	284/71	MLR	0.61	100
				MLR-EN	0.76	
				ANN	0.65	
	Water solubility	COSMO-RS solubilities, PM3 descriptors	16137 (total)	ARM	N/A	107
				DT	Accuracy = 0.93	
				Multilayer NN (DL)	0.84; 0.99	
	Melting temperature	Quantitative structure (SMILES), structural and physicochemical descriptors	1253 (total)	RNN (DL)	0.90	101
	Melting temperature	Thermodynamic and structural descriptors	2212 (total)	RF	0.67 (RMSE = 44, MAE = 14)	168
	Viscosity	Group contribution, quantitative structure	15251 (total)	NLR	0.96	102
				SVM	0.99	
	Conductivity	Structural descriptors	IL-split = 324/40; Data-point split = 3406/425	ANN	0.96 (RMSE = 1.63)	169
	Conductivity	Structural descriptors	1190/132	SVR	0.99 (RMSE = 0.15, MAE = 0.06)	170
				ANN	0.99 (RMSE = 0.02, MAE = 0.09)	
	Conductivity	Structural descriptors	2582/286	MLR	0.853 (RMSE = 0.322, MAE = 0.204)	171
				RF	0.96 (RMSE = 0.16, MAE = 0.09)	
				XGB	0.99 (RMSE = 0.09, MAE = 0.05)	
DES	Surface tension	COSMO-RS descriptors ($S_{\sigma\text{-profiles}}$)	1084/487	ANN	0.98	109
	Viscosity	Basic properties (molar mass, ratio, temperature); content of H_2O in DES (W_{H_2O}); Morgan fingerprint of HBA/HBD	795/198	SVR	0.93	110
				RF	0.96	
				NN	0.98	
				XGB	0.99	
	Viscosity	COSMO-RS derived descriptors	3464/1484	CatBoost	0.99 (RMSE = 0.21 mPa·s)	172
	Density	Group contribution	1058/352	MLPANN	0.99	111
				LSSVM	0.99	
	Density	Group contribution, temperature, critical pressure and temperature, acentric factor	1053/186	LSSVR	0.99	117
				MLP	0.98	
				CFF	0.97	
				GR	0.95	
				RBF	0.89	
				RNN	0.89	
				ANFIS	0.97	
	pH	Quantitative structure (SMILES), COSMO-RS descriptors ($S_{\sigma\text{-profiles}}$)	473/175	MLR	0.99; 0.99	113
				ANN	0.99	
	Eutectic temperature and eutectic composition	Quantitative structure (SMILES), COSMO-RS descriptors	1318/329	GBR	0.68; 0.04	112
				RFR	0.43; −0.21	
				SVR	0.74; 0.34	
				KNN	0.56; 0.03	

Table 1. continued

Solvent	Property	Descriptors	No. of data points (training/test)	ML model	R^2	Ref.
	CO_2, CO, CH_4, H_2 and N_2 gas solubility	Thermodynamic descriptors	Not specified	Hybrid SVM/LSTM	No R^2 reported (RMSE ranging from 1.6×10^{-4} to 9.8×10^{-4})	173
	CO_2 solubility	Thermodynamic and structural descriptors	711/300	RF	0.95 (RMSE = 0.22)	174
	CO_2 solubility	COSMO-RS descriptors	2055/272	ANN	0.99 (RMSE = 7.08 g/kg)	175
	CO_2 solubility	COSMO-RS descriptors	1084/889	ANN	0.99 (RMSE = 0.13, MAE = 0.09)	176
	CO_2 absorption capability	Structural descriptors	DES = 1971 (total); IL = 9615 (total)	RF	0.98 (RMSE = 0.10, MAE = 0.06)	177
				ANN	0.97 (RMSE = 0.10, MAE = 0.08)	
	SO_2 adsorption capability	Thermodynamic descriptors	384/96	ANN	0.98 (RMSE = 0.001)	178
	Cobalt solubility	Structural and thermodynamic descriptors, experimental condition	Cathode solubility = 592 (total); Metal oxide solubility = 199 (total)	XGB	0.91 (MSE = 0.004)	179

smaller and there is still a lack of data for many organic solute molecules such as 2,3-butanediol in organic solvents. The extrapolation capability of the currently developed ML models requires further examinations.

Several research groups have extended the application of ML models to predict the physicochemical properties of green solvents, such as ILs.[99−104] For instance, Song et al. constructed two hybridized GC-based models (ANN and SVM) to predict the CO_2 solubility of ILs at various temperature and pressure conditions.[105] Their findings indicated the superior performance of their ANN-GC model compared to its SVM counterpart in CO_2 solubility prediction. Similarly, Cao et al. developed three quantitative structure−property predictive models (Multilinear Regression (MLR), SVM, and ELM) using a data set comprising 119 ILs to predict the toxicity values of ILs.[106] Based on their results, their ELM model exhibited the highest predictive accuracy compared to their MLR and SVM models, revealing a strong relationship between increasing alkyl chain length (dictated by the chosen cation) and increasing toxicity. In a separate study, Can et al. employed a combination of three ML models (ARM, DT, multilayer ANN) with molecular descriptors derived from COSMO-RS and semiempirical estimation methods to predict the water capacities of ILs.[107] Their findings demonstrated high predictive performances of all three models in determining the descriptor effects, heuristic cation−anion pairing rules, and predicted water capacity of ILs. Additional examples of IL are highlighted in Table 1. The readers can also refer to previous articles for a wider range of ML applications in the field of ILs.[86,108] Considering the vast number of possible ILs that can be formed, the size of the current IL database is still relatively small. How to efficiently generate unbiased IL data with comprehensive property characterization in batches remains an interesting question. In this context, data bias may significantly contribute to the wide variation in the accuracy of machine learning models reported in the literature.

ML models have also been applied to predict the physicochemical properties of newer classes of solvents, such as DESs.[109−112] For instance, Lemaoui et al. conducted a study in which they developed and compared two novel quantitative structure−property models based on MLR and ANN algorithms to predict the pH of DESs.[113] They found that their ANN model exhibited greater predictive performance, while their MLR model provided simpler interpretability. Another notable example is the study by Abdollahzadeh et al., which involved the construction and comparison of seven different ML algorithms to predict the density of DESs.[117] Their least-squares SVR (LSSVR) exhibited the highest predictive accuracy and evaluated the effects of temperature variation, HBA selection, and HBD selection on the predicted densities of DESs. Other examples of ML applications to the studies of viscosity and gas capture capability in DES are provided in Table 1. Earlier computational data-driven works of DESs are also reviewed in several previous articles.[114−117] Similar to ILs, the limited amount of existing DES data relative to the large number of possible DES compositions remains a bottleneck in generating new, unbiased results. Therefore, a deeper fundamental understanding of the DES structures is still needed.

The application of ML models for green solvent screening introduces distinct challenges when compared to conventional solvent screening, primarily due to the relatively low availability of reported data and a limited understanding of their specific structure−property relationships. Thus, researchers often adopt a combination of feature selection techniques to refine the accuracy of their ML models, especially for more complex solvent systems such as ILs/DESs.[118] As an example, the study conducted by Cao et al. employed stepwise regression as a feature selection technique to optimize the predictive accuracy of their IL toxicity models. By utilizing a stepping criterion and combination of forward selection and backward elimination procedures, Cao et al. successfully refined their MLR model with an eight-coefficient descriptor equation.[106] In the study by Aghaie and Zendehboudi, Genetic Algorithm Multilinear Regression (GA-MLR) was employed as a feature selection technique to identify the most important descriptors for the construction of linear ML models for IL CO_2 solubility prediction. Their GA-MLR results highlighted specific geometrical descriptors, including Chi_G/D 3D, HOMO−LUMO fraction for anions, Disps, and SpMax_RG for cations, as the most significant features contributing to the accuracy of their ML model predictions.[69] By selecting ML model features to account for the nuanced and complex molecular interactions within complex solvent systems, including ILs and DESs, researchers can enhance the accuracy of their screening predictions, driving the discovery and development of novel solvents with customizable properties.

However, achieving a balance between the model complexity and interpretability is crucial. ML models, often perceived as "black boxes", are commonly subjected to a variety of intricate decision-making and extensive data analysis tasks. This inherent complexity, arising from a multitude of features, parameters, and algorithm layers, can pose challenges in elucidating the specific factors influencing a model's predictions. In solvent screening, a comprehensive understanding of a solvent's structure−property relationship and behavior is essential for guiding future research. Therefore, employing methods, such as feature importance analyses, are key to interpreting ML model predictions and gaining mechanistic insights.

Notably, the SHAP (SHapley Additive exPlanations) method has emerged as a popular feature importance technique in ML. In contrast to conventional feature importance methods commonly found in RF, DT, and gradient boosting algorithms, SHAP calculates and averages numerical values, known as Shapley values, for each model feature. This approach considers all possible feature combinations, resulting in a fair and uniformly distributed measurement of a specific feature's impact on the output of an ML model, allowing an enhanced degree of versatility and applicability across a wide range of ML model types. Within the solvent screening space, Lei et al. employed SHAP to elucidate the importance of their ML model features in predicting the surface tension and viscosity of IL-organic solvent binary systems.[103] Their findings revealed that the IL mole fraction, followed by the presence of −OH functional groups in organic solvent, had the greatest impact in predicting the surface tensions of IL-organic systems, while the IL mole fraction, followed by the presence of $-CH_2$ groups in the IL, exhibited the greatest impact on viscosity prediction. Similarly, a study by Liu et al. utilized SHAP to quantify the impact of each structural descriptor and parameter in the density and heat capacity of IL-organic solvent binary systems.[104] The SHAP results indicated that the IL mole fraction exhibited the greatest influence in predicting the densities of IL-organic systems, followed by the presence of an IL anion, Tf_2N, and the presence of $-CH_3$ groups in the organic solvent. For heat capacity prediction, SHAP revealed the IL mole fraction as the greatest contributing feature, followed by the presence of $-CH_3$ groups on ring structures. Moreover, the SHAP methodology has been utilized for the synthesis of DESs.[110] The investigation by Shi et al. leveraged the SHAP framework to methodologically quantify and rank the influence of HBA and HBD structures on the viscosity properties of DESs (Figure 2). Intriguingly, Shi et al. articulated a comprehensive analysis whereby the SHAP values were correlated with various parameters, including temperature, the molar mass of HBA and HBD, and the molar ratio of HBA and HBD. Through this approach, the study has significantly contributed to the advancement of a more rationalized design and synthesis process for DESs, enhancing the precision in tailoring their physicochemical properties.

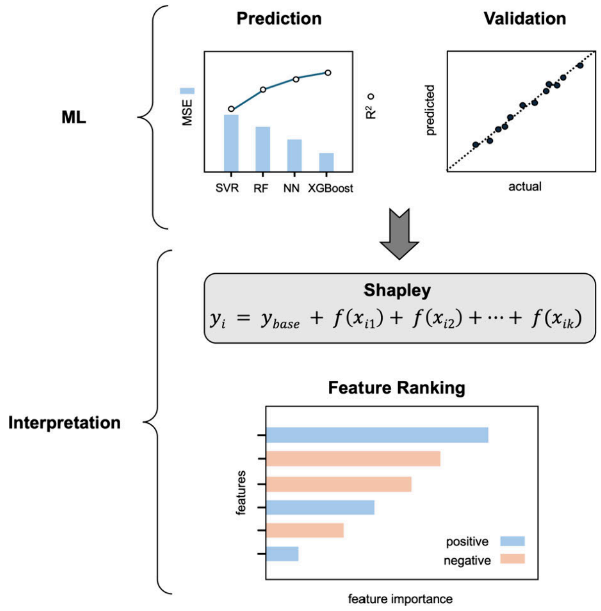


Figure 2. ML workflow with model interpretation via Shapley methodology.

2.3. Extraction Process Design and Optimization via ML Models. The application of ML models can surpass the general prediction of solvents' physicochemical properties to a more focused investigation of the behavior and interactions of solvents in extraction processes.[119−121] The distribution or partitioning coefficient (K_d), often expressed in its logarithmic form (LogD), is the ratio of mass fractions in a mixture of two immiscible liquid phases at equilibrium. K_d is expressed as

$$K_d = m_{extract}/m_{raffinate} \tag{3}$$

where $m_{extract}$ is the mass fraction of a solute in the extract phase (solvent-rich) and $m_{raffinate}$ is the mass fraction of a solute in the raffinate phase (aqueous). Related to the distribution behavior is the separation factor (β), also known as the selectivity. β is expressed as

$$\beta = \frac{m_A^{extract}/m_B^{extract}}{m_A^{raffinate}/m_B^{raffinate}} \tag{4}$$

where $m_A^{extract}/m_B^{extract}$ is the ratio of mass fractions of solutes A and B in the extract phase and $m_A^{raffinate}/m_B^{raffinate}$ is the ratio of mass fractions of solutes A and B in the raffinate phase.

Table 2. Summary of Studies Using ML Models for Extraction Process Design and Optimization

Extraction process	Key parameters	No. of data points (training/test)	ML model	R^2	Ref.
IL/DES extraction of oil sludge	Conductivity, surface tension, pH, viscosity	Not specified	RR	0.96 (RMSE = 0.95)	119
			MLP	0.95 (RMSE = 1.01	
			SVR	0.95 (RMSE = 1.03)	
DES extraction of boron contaminants from aqueous media	Molar ratio, density, viscosity, water solubility, leachability	500/126	ANN	0.98	120
Ultrasonic DES extraction of date fruit sugars	Temperature, extraction time, solvent to date fruit ratio	Not specified	RSM	0.84	121
			ANN	0.97	
Ligand-assisted IL extraction of lanthanides	Molecular physicochemical descriptors, atomic extended-connectivity fingerprints	1085/117	Multilayer NN (DL)	0.85	122
IL extraction of metals	Cation/anion structures dictate metal selectivity, cation structure dictates eco-toxicity	186 (total)	RF	0.76	123
Hydrophobic DES extraction of furfural	Aromatic-based solvents containing phenolic hydroxyl groups	Not specified	MLR	0.92	126
Solvent extraction of water pollutants using emulsion liquid membranes	Extraction time, volume ratio of organic phase to internal phase, emulsification time, feed concentration	850/150	LR	0.60	125
			RF	0.92	
			XGB	0.94	
			ANN	0.82	
Cosolvent-assisted supercritical fluid extraction of date fruit sugars	Cosolvent selection, CO_2 flow rate, cosolvent ratio, temperature, pressure, time	Not specified (ratio 85/15)	MNLR	0.87	126
			ANN	0.99	
Process parameter evaluation and optimization for solvent extraction using pulsed disk and doughnut columns	Sauter mean droplet diameter (d_{32}), axial mixing coefficient continuous phase (E_c), height of a mass transfer unit continuous phase (H_{oc}), volume fraction holdup of dispersed phase (x_d)	d_{32} = 337 (total); E_c = 557 (total); H_{oc} = 260 (total); x_d = 203 (total)	RF	0.99 (d_{32}), 0.92 (x_d), 0.97 (E_c), 0.98 (H_{oc})	127
			SVM	0.98 (d_{32}), 0.96 (x_d), 0.95 (E_c), 0.97 (H_{oc})	
			ANN	0.98 (d_{32}), 0.93 (x_d), 0.97 (E_c), 0.98 (H_{oc})	
Hydrogen recovery using IL binary mixtures	Temperature, molar fraction, chemical functional groups, density, viscosity, surface tension, heat capacity		ANN	0.99	128

Table 3. Summary of Studies Using High-Throughput Screening for Chemical and Separation Applications

Property	Application	HT method	Analytical method	Ref.
Solubility	Automated generation of empirical solubility database of various solute/organic solvent combinations	Unchained Laboratories robotic system	HPLC	135
	Automated determination of caffeine solubility in different organic solvents	Customized robotic system with web-camera/algorithm	N/A	136
	Automated salt screening of synthetic intermediates and ranking of counterion/solvent combinations for drug development	Automated birefringence imaging station	HPLC	137
	Automated determination of aqueous and nonaqueous solubility of redox active materials for battery applications	Big Kahuna robotic system; Unchained Laboratories robotic system	NMR UV−vis	138
Phase boundary	Automated interface detection of organic solvent and aqueous mixtures for drug discovery	Abbott liquid−liquid extraction station with refractive index detection	HPLC	139
	Automated partition coefficient measurement of organic solvent and aqueous mixtures	Unchained Laboratories Freeslate CM3 system; camera and MATLAB image analysis algorithm	HPLC NMR KF	140
	Automated phase separation dynamics of organic solvent and aqueous mixtures	Basler acA1300-39 μm scan monochrome camera for image analysis	N/A	141
	Automated phase detection or organic solvent and aqueous mixtures for downstream purification processes	Tecan Freedom EVO200 robotic system with TubeEyeX camera and software	LC	142

In ML-based solvent screening, both the distribution coefficient and separation factor can be predicted to evaluate the efficiency of a solvent extraction process, as shown in Table 2. For instance, Liu et al. constructed a DNN to predict the distribution coefficients of ligand-assisted solvent extraction of rare-earth elements.[122] Utilizing a combination of molecular physicochemical descriptors, generated from RDKit, and atomic extended-connectivity fingerprints, their DNN model predicted the distribution coefficients of four novel ligand structures for the liquid−liquid extraction of trivalent lanthanides from aqueous solution, which was successfully validated through experimentation. Another notable example is the study by Fajar et al. in which two RF algorithms were constructed using a classification data set of 76 ILs and a regression data set of 110 ILs to predict the selectivity and eco-toxicity values of ILs for metal extraction, respectively.[123] By leveraging the feature importance RF algorithm, Fajar et al. successfully identified the important factors determining IL metal selectivity and eco-toxicity: cation/anion pairing selection and cation selection (particularly the hydrocarbon region), respectively. Similarly, a study by Darwish et al. utilized a combination of COSMO-RS and ML modeling to prescreen DES constituents and investigate the effect of σ-profile on the predicted extraction behavior of DESs for furfural extraction.[124] Specifically, Darwish et al. successfully prescreened 108 DES constituents based on predicted distribution coefficient and selectivity, which resulted in the identification of five optimal constituents. Their DT and MLR algorithms indicated positive correlations between HBD and HBA σ-values, revealing aromatic-based DESs containing phenolic hydroxyl groups as the ideal solvents for the liquid−liquid extraction of furfural.

Furthermore, some research groups have expanded the application of ML models to predict the effects of external process parameters, such as operating temperature, pressure, and time, on extraction processes. For instance, Li et al. employed LR, RF, XGB, and ANN algorithms using a database comprising preparation condition and operating parameter data to simulate the extraction efficiency of emulsion liquid membrane-based extraction for contaminant removal.[125] They found that their XGB model showed the highest accuracy for predicting the extraction efficiency of their emulsion liquid membrane process. Furthermore, their models identified extraction time, volume ratio of the organic to internal phase, emulsification time, and feed concentration as significant parameters in tailoring the efficiency of their process. Similarly, a study by AlYammahi et al. employed MNLR and ANN models to explore the relationship between various process conditions and sugar extraction efficiency through supercritical CO_2 and cosolvents.[126] They found that their ANN model performed better than the MNLR in predicting the maximum total sugar content of the extraction process based on the solvent volume ratio, temperature, pressure, and cosolvent volume ratio. Furthermore, Su et al. developed three ML models (RF, SVM, and ANN) to predict the performance of pulsed disc and doughnut columns for liquid−liquid extraction.[127] By utilizing a comprehensive database comprising 1357 performance criteria, all three ML models predicted performance parameters, including the dispersed-phase holdup, drop size, axial diffusion coefficient, and height of the mass transfer unit. From their results, they observed that the RF model exhibited the highest accuracy in predicting the drop size, axial diffusion coefficient, and height of the mass transfer unit, while the SVM exhibited the highest accuracy in predicting the dispersed-phase holdup. Additionally, the RF feature importance algorithm was applied, which identified pulse intensity, dispersed phase velocity, continuous phase velocity, and continuous phase properties as the most important input features in determining the drop size, dispersed-phase holdup, axial diffusion coefficient, and mass transfer unit height, respectively. In summary, these examples have shown the extensive applicability of ML-based approaches in the optimization of process parameters in various extraction processes. It is also noted that, in many cases, the available training data set may only allow for simpler ML algorithms rather than neural networks. Nevertheless, these simpler ML models can still achieve satisfactory performance. Therefore, future investigations may benefit from starting with simpler ML algorithms before progressing to more complex deep learning methods, which require significantly larger data sets from experimental measurements under various operational conditions.

Last but not least, we highlight a recent successful workflow for IL−IL binary mixtures: from ML-assisted modeling, IL solvent tailoring, to process design and optimization, then to the application in hydrogen recovery from raw coke oven gas.[128] Since most published studies have focused on the property prediction of pure single-component IL, the property prediction of IL-containing mixtures is still rare. Chen et al. demonstrated an integrated multistep framework starting from raw data curation to a final optimized process for hydrogen recovery using IL mixtures. It involves the initial database development from public resources, descriptor generation and ML model development, mathematical modeling of IL mixtures with case-dependent constraints, and process design modeling and optimization with case-dependent constraints. The authors not only identified the components of a desired IL binary mixture but also suggested a series of key process parameters such as flow rate, adsorption column size, pressures, etc., for hydrogen recovery. Although the current framework lacks flexibility and transferability, it provides a valuable starting point for the solvent screening community to integrate the fundamental development of solvent compounds with the applied design of the solvent-based process in the early stages of research.

3. HIGH-THROUGHPUT SCREENING

Considering the potential inconsistency and low repeatability of reported solvent data, as well as the issue of low-quality training data sets for ML models for ILs/DESs, HT methods can effectively provide high quality data and reduce experimental costs.[129−131] The maturation of AI/ML, in recent years, has presented a unique opportunity to seamlessly integrate intelligent decision-making with HT workflows, thereby creating a synergistic platform for rapid and efficient data collection, data analysis, and process optimization in a wide range of materials, chemistry and biochemical research.[132−134] This integrated approach holds tremendous promise for driving innovation in separations, especially in solvent screening for extraction processes. Examples of HT studies for solvent-based chemical and separation applications are summarized in Table 3.

3.1. Solubility Determination. Thermodynamic solubility data are key to the design and optimization of separation processes. In solvent extraction, the solubility of a target compound plays a significant role in determining the rate and yield of its transfer across liquid phases. Traditionally, researchers have relied on hands-on experimentation using titration, gravimetric analysis, conductivity measurements, and shaking/stir-flask techniques to determine the empirical solubility of various solute−solvent systems. However, such methods are often time-consuming and resource intensive. In recent years, there has been great interest in leveraging HT workflows to accelerate the solubility screening process, specifically for chemical and separation applications. Notably, Qiu and Albrecht established a comprehensive data set comprising 1125 solubility screening panels of 905 solutes that were empirically collected using an Unchained Laboratories automated, HT workflow.[135] Their automated system contained a 96-well filter assembly with liquid transfer, mixing, and temperature control, which allowed for the preparation of various solute, temperature, and organic solvent combinations for high-performance liquid chromatography (HPLC) analysis. In a separate study, Shiri et al. developed a modular robotic platform integrated with computer vision as a proof-of-concept system for fully autonomous solubility screening.[136] Specifically, this approach utilized a robotic solid and liquid dosing system for automated solubility sample preparation as well as a webcam and corresponding algorithm to measure the turbidity of various solutions. Similarly, a study by Qiu et al. utilized a HT methodology involving an automated birefringence imaging station for salt screening on synthetic intermediates.[137] By combining their HT workflow with Z-score statistical analysis, Qiu et al. were able to systematically rank different acids, solvents, and acid/solvent combinations based on their respective solubility. The accuracy and precision of these HT solubility experiments, however, may vary based on various experimental parameters, determined by the specific researcher. To address this limitation, Liang et al. proposed a standardized HT screening workflow for the large-scale determination of redox-active molecule solubilities.[138] Specifically, Liang et al. evaluated the effects of the mixing time, solid−liquid phase separation process selection, analytical method selection, and volume transfer calibration on their solubility results. Using indol-3-carboxylic acid as a model compound, their optimized system successfully predicted its solubility under various conditions with high accuracy and precision, as validated through experimentation.

3.2. Phase Detection. The phase behavior of a solvent system is vital in defining the overall distribution and partitioning interactions within an extraction process. Traditional evaluation of a multiphase solvent system's phase behavior and distribution coefficient has typically involved the manual preparation of liquid−liquid extraction samples, phase volume estimation, and manual extraction of aliquots from each phase for chromatographic analysis. To accelerate this process, various automated HT technologies focused on phase detection have been developed. One early example is the Abbott Liquid−Liquid Extraction Station, which featured a cylindrical robot capable of managing 80 sample vials paired with a refractive index phase detector.[139] In its initial 18 months of operation, the Abbott Station successfully processed over 6000 compounds, significantly improving workflow efficiency. More recently, researchers and engineers have integrated a variety of algorithms to advance these technologies. Duffield et al., for instance, customized their robotic system with a camera and MATLAB image analysis algorithm to visually quantify the phase volumes of organic−aqueous biphasic samples in equilibrium.[140] Notably, their automated image analysis HT workflow achieved highly accurate and robust partition coefficient measurements while using less than 1% of the typical reagent amounts and providing up to 94% of the time savings from previous manual experimental studies. A study by Daglish et al. expanded upon this workflow by incorporating an automated image analysis algorithm to measure the separation rate of biphasic samples over time.[141] In addition, a study by Sun et al. introduced a custom Python algorithm combined with their automated Tecan image analysis system for more rapid liquid chromatography data visualization and partition coefficient calculation.[142] In both solubility measurement and phase boundary identification, liquid chromatography and NMR have been employed in most cases, although potential issues, such as measurement efficiency and compatibility, still remain. We believe that further enhancement of these two methods and implementation of new characterization methods will attract more researchers in the near future.

3.3. Self-Driving Laboratories. While numerous automated HT solvent screening platforms have been successfully developed and commercialized, many of these systems still rely on manual data analysis and interpretation by researchers. This reliance has prompted growing interest in the advancement of

AI/ML-powered laboratories that can autonomously design, validate, and optimize processes. Readers can refer to several comprehensive review articles for AI-assisted self-driving laboratories in the field of chemistry, separation, and materials science.[39,143−145] For instance, Clayton et al. introduced a self-driven platform for optimizing multistep continuous reaction processes in pharmaceutical development.[146] By integrating multiobjective ML models like TSEMO with automated continuous flow reactor systems, Clayton et al. achieved efficient process optimization of a Sonogashira reaction within only 13 h. Moreover, their system markedly reduced the quantity of catalytic reagents typically required in manual process design and experimentation, saving both time and resources. Another noteworthy example is the study conducted by Pomberger et al., which integrated ML models with a liquid handling robotic platform for the autonomous preparation and adjustment of buffered polyprotic solutions with varied pH levels for formulation chemistry applications.[147] They found that their GP model, featured with chemical input data, achieved the highest accuracy in predicting pH, thereby enabling precise pH adjustment of solutions through automated robotic acid/base transfers. In the context of solvent screening, this study showcases an exciting opportunity to optimize separation processes in real-time, facilitating an accurate and robust closed-loop workflow that minimizes manual intervention and the use of experimental resource use.

In addition to a physical self-driving lab, the digital lab demonstrates another important type of research approach that is usually supported by cloud facilities.[148] One of the mostly studied types of digital laboratories is digital twin.[149−151] A digital twin is a virtual counterpart of a physical entity, designed for simulation, analysis, and evaluation of a real-world object.[152] It can bring numerous advantages, such as cost reduction, performance optimization, enhanced decision-making, and accelerated innovation. For instance, Wu et al. developed a digital twin model to characterize CO_2 capture by single atom solutions.[150] The model successfully revealed the thermoelectrical effect and CO_2 capture performance, and it suggested conditions that increased CO_2 desorption efficiency by 4.2% and reduced energy consumption by 35.4%. Several two-dimensional single atom solutions such as Cu/C, Cu/N, or Cu/Si were proposed as potential advanced materials for CO_2 capture. Pallavicini et al. applied digital twin modeling to validate a demo-scale plant of biogas scrubbing technology.[153] Through evaluating the H_2S absorption process in a biogas production plant, the model simulated operating conditions different from those of the demo-scale plant to identify the controlling variables for H_2S removal efficiency. The authors found soda flow rate and concentration as the top two influencing parameters, compared with H_2S concentration, temperature, and freshwater flow rate. Örs et al. also show that the AI-assisted operational digital twin can be promising and beneficial in the chemical process engineering.[151] Moreover, Lo et al. proposed the concept of a “frugal twin” and reviewed its application to low-cost self-driving laboratories for chemistry and materials science.[154] We believe that the digital twin approach, although not widely implemented in solvent screening and design yet, will attract more attention from scientists and engineers globally in the near future.

4. LIMITATIONS AND PROSPECTS

AI/ML has proven to be a powerful tool for screening, designing, and optimizing novel solvents, chemical reactions, and separation processes. However, it is crucial to acknowledge that the ease and accuracy of ML-based solvent screening strategies can vary based on several factors. The quantity and quality of an input data set, for instance, can significantly impact the accuracy and applicability of an ML model for different solvents, solutes, and processes. A consistent trend among organic solvent ML studies is the abundance of existing data readily accessible through open-source databases. Still, it is worth noting that the amount of data for organic solvents is far below those for aqueous solvation. Therefore, reliable data for organic solvent mixtures remain in short supply. Nonetheless, the well-established knowledge of organic solvents and their properties from the previous literature can provide valuable guidance for ML model development, deployment, and interpretation. On the contrary, newer types of solvents such as ILs and DESs may face challenges in ML applications due to the limited data and relatively modest understanding of their molecular interactions and properties. Moreover, the diverse array of cation/anion or HBA/HBD combinations necessitates the use of advanced ML techniques with additional experimental validation, potentially requiring extra time and effort from researchers.

Addressing these challenges, the synergy of AI/ML with HT technologies appears to be a promising solution. This combination excels at automating experimental tasks and swiftly generating comprehensive empirical data sets, offering a synergistic platform for quickly and efficiently identifying optimal solvent structures, properties, and conditions for chemical and separation processes. Recent research has verified the feasibility of automation laboratories built on this concept. Furthermore, the rapid advancement and adoption of large language models, such as ChatGPT, herald a new era in which these LLM tools serve as invaluable chemistry assistants for solvent screening. These innovations enable researchers, including those without programming skills, to perform complex text mining, standardize data formats, and execute detailed calculations based on published experimental data. However, it is important to note that general LLM models such as ChatGPT often fail in domain sciences such as chemistry or materials science, where its answers require careful examination by a domain expert.[155,156] A further refinement by domain knowledge is necessary to improve the reliability. By integrating real experimental outcomes with established empirical or semiempirical models, the accuracy of predictions can be significantly improved. Such pioneering efforts have been documented in fields like catalysis research,[157] material design,[158] drug discovery,[159] and water harvesting.[160] Although reports on solvent selection and extraction separation are currently sparse, it is anticipated that research in this domain will soon gain significant attraction.

Finally, we would like to address a powerful tool for inverse design, e.g., starting from the desired property to suggesting a possible chemical structure or material component. We have witnessed successes of Gen-AI achieved in numerous fields such as chemical science,[161] materials engineering,[162] and food design.[163] We believe the time has already come to move beyond showcasing the viability of Gen-AI and toward utilizing Gen-AI in practical scientific research. However, applying Gen-AI to separation science such as green solvent screening remains a less explored territory. Integrating self-driving experiments with Gen-AI for solvent selection or process optimization will undoubtedly accelerate extraction research in the near future.

5. CONCLUSION

The pivotal role of separations across industries, compounded with the drive toward sustainable practices, necessitates the evolution of more efficient and environmentally friendly solvent extraction processes. The integration of AI/ML and automated high-throughput (HT) technologies presents an exciting opportunity to expedite the discovery and development of novel solvents and process designs with enhanced specificity. These integrated methodologies exhibit substantial potential for overcoming bottlenecks associated with traditional solvent selection and experimentation. By enabling a more efficient, data-driven paradigm, the integration of AI/ML and HT technologies offers a promising path toward sustainable and economically viable extraction approaches, reshaping the landscape of industrial separation processes with unprecedented precision and effectiveness.

■ AUTHOR INFORMATION

Corresponding Authors

Difan Zhang − *Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99354, United States;* orcid.org/0000-0001-7530-2378; Email: difan.zhang@pnnl.gov

Dupeng Liu − *Advanced Biofuels and Bioproducts Process Development Unit, Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Emeryville, California 94608, United States;* orcid.org/0000-0002-6951-4064; Email: pengduliu@lbl.gov

Ning Sun − *Advanced Biofuels and Bioproducts Process Development Unit, Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Emeryville, California 94608, United States;* orcid.org/0000-0002-9689-9430; Email: nsun@lbl.gov

Authors

Justin P. Edaugal − *Advanced Biofuels and Bioproducts Process Development Unit, Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Emeryville, California 94608, United States*

Vassiliki-Alexandra Glezakou − *Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States;* orcid.org/0000-0001-6028-7021

Complete contact information is available at:
https://pubs.acs.org/10.1021/cbe.4c00170

Author Contributions

DZ and DL conceptualized the topic and determined the scope. JPE, DZ, and DL conducted the literature search and drafted the manuscript. NS and V-AG provided critical insights into the manuscript content and provided substantial revisions. All authors reviewed and approved the final manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors from the ABPDU acknowledge support from the U.S. Department of Energy's Bioenergy Technologies Office (BETO), which is part of the Office of Energy Efficiency and Renewable Energy (EERE), and funding from the American Recovery and Reinvestment Act. All authors acknowledge the financial support through BETO's Bioprocessing Separations Consortium. PNNL is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830. ORNL is operated by UT-Battelle for the U.S. Department of Energy under contract no. DE-AC05-00OR22725. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed or represents that its use would not infringe privately owned rights.

■ ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ARL	Association Rule Learning
ARM	Association Rule Mining
CML	Chemical Markup Language
CNN	Convolutional Neural Net
COSMO-RS	Conductor-like Screening Model for Realistic Solvation
COSMO-SAC	Conductor-like Screening Model for Segment Activity Coefficient
DNN	Deep Neural Network
DES	Deep Eutectic Solvent
DT	Decision Tree
ELM	Extreme Learning Machine
ET	ExtraTrees
GB	Gradient Boosting
GA-MLR	Genetic Algorithm Multilinear Regression
GC	Group Contribution
ΔG_{sol}	Gibbs Free Energy of Solvation
GPU	Graphics Processing Unit
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
HT	High-Throughput
HPLC	High-Performance Liquid Chromatography
IL	Ionic Liquid
InChI	International Chemical Identifier
K_d	Distribution Coefficient
KF	Karl Fischer Titration
LC	Liquid Chromatography
LLE	Liquid−Liquid Extraction
LLM	Large-Scale Language Model
LogReg	Logistic Regression
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multilinear Regression
MNLR	Multiple Nonlinear Regression
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
R^2	Coefficient of Determination
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SVM	Support Vector Machine
SVR	Support Vector Regression
SHAP	SHapley Additive exPlanations

SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
TSEMO	Thompson Sampling Efficient Multi-Objective Optimization
UV−vis	Ultraviolet−Visible Spectroscopy
UNIQUAC	Universal Quasi-Chemical Activity Coefficient
UNIFAC	Universal Quasi-Chemical Functional Group Activity Coefficient
XGB	eXtreme Gradient Boosting

■ REFERENCES

(1) Sholl, D. S.; Lively, R. P. Seven chemical separations to change the world. *Nature* **2016**, *532* (7600), 435−437.

(2) Zhang, Q.-W.; Lin, L.-G.; Ye, W.-C. Techniques for extraction and isolation of natural products: A comprehensive review. *Chinese medicine* **2018**, *13*, 1−26.

(3) Shao, L. Grand challenges in emerging separation technologies. *Frontiers Media SA* **2020**, *1*, 3.

(4) Janicka, P.; Płotka-Wasylka, J.; Jatkowska, N.; Chabowska, A.; Fares, M. Y.; Andruch, V.; Kaykhaii, M.; Gębicki, J. Trends in the new generation of green solvents in extraction processes. *Current Opinion in Green and Sustainable Chemistry* **2022**, *37*, No. 100670.

(5) Picot-Allain, C.; Mahomoodally, M. F.; Ak, G.; Zengin, G. Conventional versus green extraction techniques — a comparative perspective. *Current Opinion in Food Science* **2021**, *40*, 144−156.

(6) Afonso, J.; Mezzetta, A.; Marrucho, I. M.; Guazzelli, L. History repeats itself again: Will the mistakes of the past for ILs be repeated for DESs? From being considered ionic liquids to becoming their alternative: the unbalanced turn of deep eutectic solvents. *Green Chem.* **2023**, *25* (1), 59−105.

(7) Prabhune, A.; Dey, R. Green and sustainable solvents of the future: Deep eutectic solvents. *J. Mol. Liq.* **2023**, *379*, 121676.

(8) Płotka-Wasylka, J.; De la Guardia, M.; Andruch, V.; Vilková, M. Deep eutectic solvents vs ionic liquids: Similarities and differences. *Microchemical Journal* **2020**, *159*, 105539.

(9) Ventura, S. P.; e Silva, F. A.; Quental, M. V.; Mondal, D.; Freire, M. G.; Coutinho, J. A. Ionic-liquid-mediated extraction and separation processes for bioactive compounds: past, present, and future trends. *Chem. Rev.* **2017**, *117* (10), 6984−7052.

(10) Chen, Y.; Mu, T. Revisiting greenness of ionic liquids and deep eutectic solvents. *Green Chemical Engineering* **2021**, *2* (2), 174−186.

(11) Atashnezhad, A.; Scott, J.; Al Dushaishi, M. F. Environmental Implications of Ionic Liquid and Deep Eutectic Solvent in Geothermal Application: Comparing Traditional and New Approach Methods. *ACS Sustainable Chem. Eng.* **2024**, *12* (40), 14684−14693.

(12) de Jesus, S. S.; Maciel Filho, R. Are ionic liquids eco-friendly? *Renewable and Sustainable Energy Reviews* **2022**, *157*, 112039.

(13) González-Miquel, M.; Díaz, I. Green solvent screening using modeling and simulation. *Current Opinion in Green and Sustainable Chemistry* **2021**, *29*, No. 100469.

(14) Eckert, F.; Klamt, A. Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE J.* **2002**, *48* (2), 369−385.

(15) Zhang, X.; Wang, J.; Song, Z.; Zhou, T. Data-driven ionic liquid design for CO2 capture: Molecular structure optimization and DFT verification. *Ind. Eng. Chem. Res.* **2021**, *60* (27), 9992−10000.

(16) Sun, Q.; Chen, H.; Yu, J. Investigation on the lithium extraction process with the TBP−FeCl3 solvent system using experimental and DFT methods. *Ind. Eng. Chem. Res.* **2022**, *61* (13), 4672−4682.

(17) Kaim-Sevalneva, V.; Sariola-Leikas, E.; He, C. Highly selective extraction of scandium (III) from rare earth elements using quaternary ammonium based ionic liquids: Experimental and DFT studies. *Sep. Purif. Technol.* **2024**, *334*, 126038.

(18) Panda, D. K.; Bhargava, B. Intermolecular interactions in tetrabutylammonium chloride based deep eutectic solvents: Classical molecular dynamics studies. *J. Mol. Liq.* **2021**, *335*, 116139.

(19) Hao, H.; Liu, Y.; Yuan, J.; Dong, X.; Wang, Z.; Xu, C.; Chen, J. Redox assisted solvent extraction to enable highly efficient separation of cerium from other lanthanides: Experimental studies and DFT calculations. *Hydrometallurgy* **2024**, *225*, 106264.

(20) Liu, X.; Xing, J.; Sun, M.; Su, Z.; Chen, Z.; Wang, Y.; Cui, P. Phase behavior and extraction mechanism of methanol-n-hexane separation using choline-based deep eutectic solvent. *J. Mol. Liq.* **2022**, *345*, 118204.

(21) Atilhan, M.; Aparicio, S. Molecular dynamics study on the use of Deep Eutectic Solvents for Enhanced Oil Recovery. *J. Pet. Sci. Eng.* **2022**, *209*, 109953.

(22) Klamt, A. The COSMO and COSMO-RS solvation models. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1* (5), 699−709.

(23) Filly, A.; Fabiano-Tixier, A. S.; Fernandez, X.; Chemat, F. Alternative solvents for extraction of food aromas. Experimental and COSMO-RS study. *LWT-Food Science and Technology* **2015**, *61* (1), 33−40.

(24) Wojeicchowski, J. P.; Ferreira, A. M.; Abranches, D. O.; Mafra, M. R.; Coutinho, J. A. Using COSMO-RS in the design of deep eutectic solvents for the extraction of antioxidants from rosemary. *ACS Sustainable Chem. Eng.* **2020**, *8* (32), 12132−12141.

(25) Diedenhofen, M.; Klamt, A. COSMO-RS as a tool for property prediction of IL mixtures—A review. *Fluid Phase Equilib.* **2010**, *294* (1), 31−38.

(26) Birajdar, S. D.; Padmanabhan, S.; Rajagopalan, S. Rapid solvent screening using thermodynamic models for recovery of 2, 3-butanediol from fermentation by liquid−liquid extraction. *Journal of Chemical & Engineering Data* **2014**, *59* (8), 2456−2463.

(27) Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density profiles for COSMO-RS(Ol). *AIChE J.* **2007**, *53* (12), 3231−3240.

(28) Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density distribution of molecules for COSMO-SAC. *AIChE J.* **2009**, *55* (12), 3298−3300.

(29) Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density profiles for COSMO-RS (Ol). *AIChE journal* **2007**, *53* (12), 3231−3240.

(30) Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density distribution of molecules for COSMO-SAC. *AIChE journal* **2009**, *55*, 3298−3300.

(31) Dong, Y.; Huang, S.; Guo, Y.; Lei, Z. COSMO-UNIFAC model for ionic liquids. *AIChE J.* **2020**, *66* (1), e16787.

(32) Zhu, R.; Taheri, M.; Zhang, J.; Lei, Z. Extension of the COSMO-UNIFAC thermodynamic model. *Ind. Eng. Chem. Res.* **2020**, *59* (4), 1693−1701.

(33) Liu, Q.; Zhang, L.; Liu, L.; Du, J.; Liang, X.; Mao, H.; Meng, Q.; Eden, M. R.; Ierapetritou, M. G.; Towler, G. P. GC-COSMO based Reaction Solvent Design with New Kinetic Model using CAMD. *Comput.-Aided Chem. Eng.* **2018**, *44*, 235−240.

(34) Peng, D.; Zhang, J.; Cheng, H.; Chen, L.; Qi, Z. Computer-aided ionic liquid design for separation processes based on group contribution method and COSMO-SAC model. *Chem. Eng. Sci.* **2017**, *159*, 58−68.

(35) Casas, A.; Rodríguez-Llorente, D.; Rodríguez-Llorente, G.; García, J.; Larriba, M. Machine learning screening tools for the prediction of extraction yields of pharmaceutical compounds from wastewaters. *Journal of Water Process Engineering* **2024**, *62*, 105379.

(36) Chung, Y.; Green, W. H. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. *Chemical Science* **2024**, *15* (7), 2410−2424.

(37) Orlov, A. A.; Valtz, A.; Coquelet, C.; Rozanska, X.; Wimmer, E.; Marcou, G.; Horvath, D.; Poulain, B.; Varnek, A.; de Meyer, F. Computational screening methodology identifies effective solvents for CO2 capture. *Communications Chemistry* **2022**, *5* (1), 37.

(38) Zhou, T.; McBride, K.; Linke, S.; Song, Z.; Sundmacher, K. Computer-aided solvent selection and design for efficient chemical processes. *Current Opinion in Chemical Engineering* **2020**, *27*, 35−44.

(39) Abolhasani, M.; Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis* **2023**, *2* (6), 483−492.

(40) Shiri, P.; Lai, V.; Zepel, T.; Griffin, D.; Reifman, J.; Clark, S.; Grunert, S.; Yunker, L. P. E.; Steiner, S.; Situ, H.; et al. Automated solubility screening platform using computer vision. *iScience* **2021**, *24* (3), No. 102176.

(41) Brunton, S. L.; Noack, B. R.; Koumoutsakos, P. Machine Learning for Fluid Mechanics. *Annu. Rev. Fluid Mech.* **2020**, *52* (1), 477−508.

(42) Sarker, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science* **2021**, *2* (3), 160.

(43) Ahmed, S. F.; Alam, M. S. B.; Hassan, M.; Rozbu, M. R.; Ishtiak, T.; Rafa, N.; Mofijur, M.; Shawkat Ali, A. B. M.; Gandomi, A. H. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review* **2023**, *56* (11), 13521−13617.

(44) Yenduri, G.; Ramalingam, M.; Selvi, G. C.; Supriya, Y.; Srivastava, G.; Maddikunta, P. K. R.; Raj, G. D.; Jhaveri, R. H.; Prabadevi, B.; Wang, W.; et al. GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access* **2024**, *12*, 54608−54649.

(45) Liu, D.; Sun, N. Prospects of artificial intelligence in the development of sustainable separation processes. *Frontiers in Sustainability* **2023**, *4*, 1210209.

(46) Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M.-H. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* **2024**, *56* (4), 105.

(47) Salakhutdinov, R. Learning Deep Generative Models. *Annual Review of Statistics and Its Application* **2015**, *2* (1), 361−385.

(48) Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111−132.

(49) Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey. *ACM Comput. Surv.* **2023**, *55* (6), 109.

(50) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1* (1), 140022.

(51) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864−2875.

(52) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102−D1109.

(53) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100−D1107.

(54) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977−2980.

(55) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123−1124.

(56) Linstrom, P. J.; Mallard, W. G. The NIST Chemistry WebBook: A Chemical Data Resource on the Internet. *Journal of Chemical & Engineering Data* **2001**, *46* (5), 1059−1063.

(57) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data* **2019**, *6* (1), 143.

(58) Onken, U.; Rarey-Nies, J.; Gmehling, J. The Dortmund Data Bank: A computerized system for retrieval, correlation, and prediction of thermodynamic properties of mixtures. *Int. J. Thermophys.* **1989**, *10* (3), 739−747.

(59) Groom, C. R.; Allen, F. H. The Cambridge Structural Database in Retrospect and Prospect. *Angew. Chem., Int. Ed.* **2014**, *53* (3), 662−671.

(60) Saini, V. Machine learning prediction of empirical polarity using SMILES encoding of organic solvents. *Molecular diversity* **2023**, *27* (5), 2331−2343.

(61) Rajan, K.; Zielesny, A.; Steinbeck, C. STOUT: SMILES to IUPAC names using neural machine translation. *Journal of Cheminformatics* **2021**, *13* (1), 34.

(62) Das, M.; Ghosh, A.; Sunoj, R. B. Advances in machine learning with chemical language models in molecular property and reaction outcome predictions. *J. Comput. Chem.* **2024**, *45* (14), 1160−1176.

(63) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12* (5), e1603.

(64) Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics* **2020**, *12* (1), 43.

(65) Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discovery Today* **2022**, *27* (11), No. 103356.

(66) Ignacz, G.; Alqadhi, N.; Szekely, G. Explainable machine learning for unraveling solvent effects in polyimide organic solvent nanofiltration membranes. *Advanced Membranes* **2023**, *3*, No. 100061.

(67) Ferraz-Caetano, J.; Teixeira, F.; Cordeiro, M. N. D. S. Explainable Supervised Machine Learning Model To Predict Solvation Gibbs Energy. *J. Chem. Inf. Model.* **2024**, *64* (7), 2250−2262.

(68) Aghaie, M.; Zendehboudi, S. Estimation of CO2 solubility in ionic liquids using connectionist tools based on thermodynamic and structural characteristics. *Fuel* **2020**, *279*, 117984.

(69) Amar, Y.; Schweidtmann, A. M.; Deutsch, P.; Cao, L.; Lapkin, A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chemical science* **2019**, *10* (27), 6697−6706.

(70) Maran, U.; Sild, S.; Tulp, I.; Takkis, K.; Moosus, M.; Mekenyan, O. Molecular Descriptors from Two-Dimensional Chemical Structure. In *In Silico Toxicology*; Cronin, M., Madden, J., Eds.; The Royal Society of Chemistry: 2010; p 0.

(71) Miyao, T.; Funatsu, K. Two- and Three-Dimensional Molecular Representations in Ligand-Based Approaches. In *Drug Development Supported by Informatics*; Satoh, H., Funatsu, K., Yamamoto, H., Eds.; Springer Nature: Singapore, 2024; pp 175−186.

(72) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb Chem. High Throughput Screen* **2000**, *3* (5), 363−372.

(73) Hutter, M. C. Molecular Descriptors for Chemoinformatics (2nd ed.). By Roberto Todeschini and Viviana Consonni. *ChemMedChem* **2010**, *5* (2), 306−307.

(74) Zhang, D.; Wu, H.; Smith, B.; Glezakou, V.-A. Chapter Five - Harness the power of atomistic modeling and deep learning in biofuel separation. In *Annual Reports in Computational Chemistry*, Vol. *19*; Elsevier: 2023; pp 121−165.

(75) Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular Descriptors for Structure−Activity Applications: A Hands-On Approach. In *Computational Toxicology: Methods and Protocols*; Nicolotti, O., Ed.; Springer: New York, 2018; pp 3−53.

(76) Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A. J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science*; Berry, M. W., Mohamed, A., Yap, B. W., Eds.; Springer International Publishing: 2020; pp 3−21.

(77) Kumar, M.; Khan, S. A.; Bhatia, A.; Sharma, V.; Jain, P. Machine Learning Algorithms: A Conceptual Review. In *2023 1st International Conference on Intelligent Computing and Research Trends (ICRT)*, Feb 3−4, 2023; pp 1−7. DOI: 10.1109/ICRT57042.2023.10146678.

(78) Singh, A.; Thakur, N.; Sharma, A. A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), March 16−18, 2016; pp 1310−1315.

(79) Choudhary, R.; Gianey, H. K. Comprehensive Review On Supervised Machine Learning Algorithms. In *2017 International Conference on Machine Learning and Data Science (MLDS)*, Dec. 14−15, 2017; pp 37−43. DOI: 10.1109/MLDS.2017.11.

(80) Naeem, S.; Ali, A.; Anam, S.; Ahmed, M. M. An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems* **2023**, *13*, 911.

(81) Wang, W. Unsupervised Learning Paradigm. In *Principles of Machine Learning: The Three Perspectives*; Wang, W., Ed.; Springer Nature: Singapore, 2025; pp 291−324.

(82) Ghahramani, Z. Unsupervised Learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*; Bousquet, O., von Luxburg, U., Rätsch, G., Eds.; Springer: Berlin, Heidelberg, 2004; pp 72−112.

(83) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121* (16), 9722−9758.

(84) Shrestha, A.; Mahmood, A. Review of Deep Learning Algorithms and Architectures. *IEEE Access* **2019**, *7*, 53040−53065.

(85) Dhall, D.; Kaur, R.; Juneja, M. Machine Learning: A Review of the Algorithms and Its Applications. In *Proceedings of ICRIC 2019*; Singh, P. K., Kar, A. K., Singh, Y., Kolekar, M. H., Tanwar, S., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp 47−63.

(86) Koutsoukos, S.; Philippi, F.; Malaret, F.; Welton, T. A review on machine learning algorithms for the ionic liquid chemical space. *Chemical science* **2021**, *12* (20), 6820−6843.

(87) Krogh, A. What are artificial neural networks? *Nature biotechnology* **2008**, *26* (2), 195−197.

(88) Mouellef, M.; Vetter, F. L.; Strube, J. Benefits and limitations of artificial neural networks in process chromatography design and operation. *Processes* **2023**, *11* (4), 1115.

(89) Bejani, M. M.; Ghatee, M. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review* **2021**, *54* (8), 6391−6438.

(90) Balabin, R. M.; Lomakina, E. I. Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? *Phys. Chem. Chem. Phys.* **2011**, *13* (24), 11710−11718.

(91) Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189−215.

(92) Sarker, I. Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* **2021**, *2*, 160.

(93) Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55* (10), 78−87.

(94) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020**, *11* (1), 5753.

(95) Saini, V.; Kumar, R. A machine learning approach for predicting the empirical polarity of organic solvents. *New J. Chem.* **2022**, *46* (35), 16981−16989.

(96) Saini, V.; Singh, H. Predicting the ET (30) parameter of organic solvents via machine learning. *Chem. Phys. Lett.* **2023**, *826*, 140672.

(97) Lim, H.; Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chemical Science* **2019**, *10* (36), 8306−8315.

(98) Lim, H.; Jung, Y. MLSolvA: solvation free energy prediction from pairwise atomistic interactions by machine learning. *Journal of Cheminformatics* **2021**, *13* (1), 56.

(99) Wang, Z.; Song, Z.; Zhou, T. Machine learning for ionic liquid toxicity prediction. *Processes* **2021**, *9* (1), 65.

(100) Danush, S.; Dutta, A. Machine learning-based framework for predicting toxicity of ionic liquids. *Materials Today: Proceedings* **2023**, *72*, 175−180.

(101) Acar, Z.; Nguyen, P.; Lau, K. C. Machine-Learning model prediction of ionic liquids melting points. *Applied Sciences* **2022**, *12* (5), 2408.

(102) Boualem, A. D.; Argoub, K.; Benkouider, A. M.; Yahiaoui, A.; Toubal, K. Viscosity prediction of ionic liquids using NLR and SVM approaches. *J. Mol. Liq.* **2022**, *368*, 120610.

(103) Lei, Y.; Shu, Y.; Liu, X.; Liu, X.; Wu, X.; Chen, Y. Predictive modeling on the surface tension and viscosity of ionic liquid-organic solvent mixtures via machine learning. *Journal of the Taiwan Institute of Chemical Engineers* **2023**, *151*, 105140.

(104) Liu, X.; Gao, J.; Chen, Y.; Fu, Y.; Lei, Y. Machine learning-assisted modeling study on the density and heat capacity of ionic liquid-organic solvent binary systems. *J. Mol. Liq.* **2023**, *390*, 122972.

(105) Song, Z.; Shi, H.; Zhang, X.; Zhou, T. Prediction of CO2 solubility in ionic liquids using machine learning methods. *Chem. Eng. Sci.* **2020**, *223*, 115752.

(106) Cao, L.; Zhu, P.; Zhao, Y.; Zhao, J. Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids. *Journal of hazardous materials* **2018**, *352*, 17−26.

(107) Can, E.; Jalal, A.; Zirhlioglu, I. G.; Uzun, A.; Yildirim, R. Predicting water solubility in ionic liquids using machine learning towards design of hydro-philic/phobic ionic liquids. *J. Mol. Liq.* **2021**, *332*, 115848.

(108) Baskin, I.; Epshtein, A.; Ein-Eli, Y. Benchmarking machine learning methods for modeling physical properties of ionic liquids. *J. Mol. Liq.* **2022**, *351*, No. 118616.

(109) Lemaoui, T.; Boublia, A.; Darwish, A. S.; Alam, M.; Park, S.; Jeon, B.-H.; Banat, F.; Benguerba, Y.; AlNashef, I. M. Predicting the surface tension of deep eutectic solvents using artificial neural networks. *ACS omega* **2022**, *7* (36), 32194−32207.

(110) Shi, D.; Zhou, F.; Mu, W.; Ling, C.; Mu, T.; Yu, G.; Li, R. Deep insights into the viscosity of deep eutectic solvents by an XGBoost-based model plus SHapley Additive exPlanation. *Phys. Chem. Chem. Phys.* **2022**, *24* (42), 26029−26036.

(111) Roosta, A.; Haghbakhsh, R.; Duarte, A. R. C.; Raeissi, S. Machine learning coupled with group contribution for predicting the density of deep eutectic solvents. *Fluid Phase Equilib.* **2023**, *565*, 113672.

(112) Lavrinenko, A. K.; Chernyshov, I. Y.; Pidko, E. A. Machine learning approach for the prediction of eutectic temperatures for metal-free deep eutectic solvents. *ACS Sustainable Chem. Eng.* **2023**, *11* (42), 15492−15502.

(113) Lemaoui, T.; Abu Hatab, F.; Darwish, A. S.; Attoui, A.; Hammoudi, N. E. H.; Almustafa, G.; Benaicha, M.; Benguerba, Y.; Alnashef, I. M. Molecular-based guide to predict the pH of eutectic solvents: promoting an efficient design approach for new green solvents. *ACS Sustainable Chem. Eng.* **2021**, *9* (17), 5783−5808.

(114) Sun, J.; Sato, Y.; Sakai, Y.; Kansha, Y. A review of ionic liquids and deep eutectic solvents design for CO2 capture with machine learning. *Journal of Cleaner Production* **2023**, *414*, No. 137695.

(115) Velez, C.; Acevedo, O. Simulation of deep eutectic solvents: Progress to promises. *WIREs Computational Molecular Science* **2022**, *12* (4), No. e1598.

(116) Kovács, A.; Neyts, E. C.; Cornet, I.; Wijnants, M.; Billen, P. Modeling the Physicochemical Properties of Natural Deep Eutectic Solvents. *ChemSusChem* **2020**, *13* (15), 3789−3804.

(117) Abdollahzadeh, M.; Khosravi, M.; Hajipour Khire Masjidi, B.; Samimi Behbahan, A.; Bagherzadeh, A.; Shahkar, A.; Tat Shahdost, F. Estimating the density of deep eutectic solvents applying supervised machine learning techniques. *Sci. Rep.* **2022**, *12* (1), 4954.

(118) Low, K.; Kobayashi, R.; Izgorodina, E. I. The effect of descriptor choice in machine learning models for ionic liquid melting point prediction. *J. Chem. Phys.* **2020**, *153* (10), 104101.

(119) Hu, C.; Fu, S.; Zhu, L.; Dang, W.; Zhang, T. Evaluation and prediction on the effect of ionic properties of solvent extraction performance of oily sludge using machine learning. *Molecules* **2021**, *26* (24), 7551.

(120) Awaja, N. E.; Almustafa, G.; Darwish, A. S.; Lemaoui, T.; Benguerba, Y.; Banat, F.; Arafat, H. A.; AlNashef, I. Molecular-based

artificial neural networks for selecting deep eutectic solvents for the removal of contaminants from aqueous media. *Chemical Engineering Journal* **2023**, *476*, 146429.

(121) AlYammahi, J.; Darwish, A. S.; Almustafa, G.; Lemaoui, T.; AlNashef, I. M.; Hasan, S. W.; Taher, H.; Banat, F. Natural deep eutectic solvents for Ultrasonic-Assisted extraction of nutritious date Sugar: Molecular Screening, Experimental, and prediction. *Ultrasonics Sonochemistry* **2023**, *98*, 106514.

(122) Liu, T.; Johnson, K. R.; Jansone-Popova, S.; Jiang, D.-e. Advancing rare-earth separation by machine learning. *JACS Au* **2022**, *2* (6), 1428−1434.

(123) Fajar, A. T.; Hartono, A. D.; Moshikur, R. M.; Goto, M. Ionic liquids curated by machine learning for metal extraction. *ACS Sustainable Chem. Eng.* **2022**, *10* (38), 12698−12705.

(124) Darwish, A. S.; Lemaoui, T.; AlYammahi, J.; Taher, H.; Benguerba, Y.; Banat, F.; AlNashef, I. M. Molecular insights into potential hydrophobic deep eutectic solvents for furfural extraction guided by COSMO-RS and machine learning. *J. Mol. Liq.* **2023**, *379*, 121631.

(125) Li, H.; Wang, Y.; Wang, Y. Machine learning for predicting the dynamic extraction of multiple substances by emulsion liquid membranes. *Sep. Purif. Technol.* **2023**, *313*, 123458.

(126) AlYammahi, J.; Darwish, A. S.; Lemaoui, T.; AlNashef, I. M.; Hasan, S. W.; Taher, H.; Banat, F. Parametric analysis and machine learning for enhanced recovery of high-value sugar from date fruits using supercritical CO2 with co-solvents. *Journal of CO2 Utilization* **2023**, *72*, 102511.

(127) Su, Z.; Wang, Y.; Tan, B.; Cheng, Q.; Duan, X.; Xu, D.; Tian, L.; Qi, T. Performance prediction of disc and doughnut extraction columns using bayes optimization algorithm-based machine learning models. *Chemical Engineering and Processing-Process Intensification* **2023**, *183*, 109248.

(128) Chen, Y.; Ma, S.; Lei, Y.; Liang, X.; Liu, X.; Kontogeorgis, G. M.; Gani, R. Ionic liquid binary mixtures: Machine learning-assisted modeling, solvent tailoring, process design, and optimization. *AIChE J.* **2024**, *70* (5), No. e18392.

(129) Potyrailo, R.; Rajan, K.; Stoewe, K.; Takeuchi, I.; Chisholm, B.; Lam, H. Combinatorial and High-Throughput Screening of Materials Libraries: Review of State of the Art. *ACS Comb. Sci.* **2011**, *13* (6), 579−633.

(130) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1* (11), 882−894.

(131) Selekman, J. A.; Qiu, J.; Tran, K.; Stevens, J.; Rosso, V.; Simmons, E.; Xiao, Y.; Janey, J. High-Throughput Automation in Chemical Process Development. *Annu. Rev. Chem. Biomol. Eng.* **2017**, *8* (1), 525−547.

(132) Eyke, N. S.; Koscher, B. A.; Jensen, K. F. Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends in Chemistry* **2021**, *3* (2), 120−132.

(133) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121* (16), 9927−10000.

(134) McCullough, K.; Williams, T.; Mingle, K.; Jamshidi, P.; Lauterbach, J. High-throughput experimentation meets artificial intelligence: a new pathway to catalyst discovery. *Phys. Chem. Chem. Phys.* **2020**, *22* (20), 11174−11196.

(135) Qiu, J.; Albrecht, J. Solubility correlations of common organic solvents. *Org. Process Res. Dev.* **2018**, *22* (7), 829−835.

(136) Shiri, P.; Lai, V.; Zepel, T.; Griffin, D.; Reifman, J.; Clark, S.; Grunert, S.; Yunker, L. P.; Steiner, S.; Situ, H. Automated solubility screening platform using computer vision. *iScience* **2021**, *24* (3), 102176.

(137) Qiu, J.; Patel, A.; Stevens, J. M. High-throughput salt screening of synthetic intermediates: effects of solvents, counterions, and counterion solubility. *Organic process research & development* **2020**, *24* (7), 1262−1270.

(138) Liang, Y.; Job, H.; Feng, R.; Parks, F.; Hollas, A.; Zhang, X.; Bowden, M.; Noh, J.; Murugesan, V.; Wang, W. High-throughput solubility determination for data-driven materials design and discovery in redox flow battery research. *Cell Reports Physical Science* **2023**, *4* (10), 101633.

(139) Maslana, E.; Schmitt, R.; Pan, J. A fully automated liquid−liquid extraction system utilizing interface detection. *Journal of Analytical Methods in Chemistry* **2000**, *22* (6), 187−194.

(140) Duffield, S.; Da Vià, L.; Bellman, A. C.; Chiti, F. Automated High-Throughput Partition Coefficient Determination with Image Analysis for Rapid Reaction Workup Process Development and Modeling. *Org. Process Res. Dev.* **2021**, *25* (12), 2738−2746.

(141) Daglish, J.; Blacker, A. J.; de Boer, G.; Crampton, A.; Hose, D. R.; Parsons, A. R.; Kapur, N. Determining Phase Separation Dynamics with an Automated Image Processing Algorithm. *Org. Process Res. Dev.* **2023**, *27* (4), 627−639.

(142) Sun, A. C.; Jurica, J. A.; Rose, H. B.; Brito, G.; Deprez, N. R.; Grosser, S. T.; Hyde, A. M.; Kwan, E. E.; Moor, S. Vision-Guided Automation Platform for Liquid−Liquid Extraction and Workup Development. *Org. Process Res. Dev.* **2023**, *27* (11), 1954−1964.

(143) Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab. *Acc. Chem. Res.* **2022**, *55* (17), 2454−2466.

(144) Bennett, J. A.; Abolhasani, M. Autonomous chemical science and engineering enabled by self-driving laboratories. *Current Opinion in Chemical Engineering* **2022**, *36*, No. 100831.

(145) Bayley, O.; Savino, E.; Slattery, A.; Noël, T. Autonomous chemistry: Navigating self-driving labs in chemical and material sciences. *Matter* **2024**, *7* (7), 2382−2398.

(146) Clayton, A. D.; Schweidtmann, A. M.; Clemens, G.; Manson, J. A.; Taylor, C. J.; Niño, C. G.; Chamberlain, T. W.; Kapur, N.; Blacker, A. J.; Lapkin, A. A. Automated self-optimization of multi-step reaction and separation processes using machine learning. *Chemical Engineering Journal* **2020**, *384*, 123340.

(147) Pomberger, A.; Jose, N.; Walz, D.; Meissner, J.; Holze, C.; Kopczynski, M.; Müller-Bischof, P.; Lapkin, A. Automated ph adjustment driven by robotic workflows and active machine learning. *Chemical Engineering Journal* **2023**, *451*, 139099.

(148) Xie, C.; Li, C.; Ding, X.; Jiang, R.; Sung, S. Chemistry on the Cloud: From Wet Labs to Web Labs. *J. Chem. Educ.* **2021**, *98* (9), 2840−2847.

(149) Tao, F.; Xiao, B.; Qi, Q.; Cheng, J.; Ji, P. Digital twin modeling. *Journal of Manufacturing Systems* **2022**, *64*, 372−389.

(150) Wu, Y.; Zhou, C.; Li, Y.; Zhang, C.; Yu, Y.; Wang, G. Characterizing the 2D single atom solutions to capture CO2 by the digital twin model. *Chemical Engineering Journal* **2024**, *493*, No. 152584.

(151) Örs, E.; Schmidt, R.; Mighani, M.; Shalaby, M. A Conceptual Framework for AI-based Operational Digital Twin in Chemical Process Engineering. In *2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, June 15−17, 2020; pp 1−8. DOI: 10.1109/ICE/ITMC49519.2020.9198575.

(152) Jones, D.; Snider, C.; Nassehi, A.; Yon, J.; Hicks, B. Characterising the Digital Twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology* **2020**, *29*, 36−52.

(153) Pallavicini, J.; Fedeli, M.; Scolieri, G. D.; Tagliaferri, F.; Parolin, J.; Sironi, S.; Manenti, F. Digital twin-based optimization and demo-scale validation of absorption columns using sodium hydroxide/water mixtures for the purification of biogas streams subject to impurity fluctuations. *Renewable Energy* **2023**, *219*, No. 119466.

(154) Lo, S.; Baird, S. G.; Schrier, J.; Blaiszik, B.; Carson, N.; Foster, I.; Aguilar-Granda, A.; Kalinin, S. V.; Maruyama, B.; Politi, M.; et al. Review of low-cost self-driving laboratories in chemistry and materials science: the "frugal twin" concept. *Digital Discovery* **2024**, *3* (5), 842−868.

(155) Leon, A. J.; Vidhani, D. ChatGPT Needs a Chemistry Tutor Too. *J. Chem. Educ.* **2023**, *100* (10), 3859−3865.

(156) Deb, J.; Saikia, L.; Dihingia, K. D.; Sastry, G. N. ChatGPT in the Material Design: Selected Case Studies to Assess the Potential of ChatGPT. *J. Chem. Inf. Model.* **2024**, *64* (3), 799−811.

(157) Su, Y.; Wang, X.; Ye, Y.; Xie, Y.; Xu, Y.; Jiang, Y.; Wang, C. Automation and Machine Learning Augmented by Large Language Models in Catalysis Study. *Chemical Science* **2024**, *15*, 12200.

(158) Zheng, Z.; Zhang, O.; Nguyen, H. L.; Rampal, N.; Alawadhi, A. H.; Rong, Z.; Head-Gordon, T.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science* **2023**, *9* (11), 2161−2170.

(159) Juhi, A.; Pipil, N.; Santra, S.; Mondal, S.; Behera, J. K.; Mondal, H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus* **2023**, *15* (3), e36272.

(160) Zheng, Z.; Alawadhi, A. H.; Chheda, S.; Neumann, S. E.; Rampal, N.; Liu, S.; Nguyen, H. L.; Lin, Y.-h.; Rong, Z.; Siepmann, J. I. Shaping the water-harvesting behavior of metal−organic frameworks aided by fine-tuned GPT models. *J. Am. Chem. Soc.* **2023**, *145* (51), 28284−28295.

(161) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145* (16), 8736−8750.

(162) Liu, Y.; Yang, Z.; Yu, Z.; Liu, Z.; Liu, D.; Lin, H.; Li, M.; Ma, S.; Avdeev, M.; Shi, S. Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. *Journal of Materiomics* **2023**, *9* (4), 798−816.

(163) Al-Sarayreh, M.; Gomes Reis, M.; Carr, A.; Reis, M. M. d. Inverse design and AI/Deep generative networks in food design: A comprehensive review. *Trends in Food Science & Technology* **2023**, *138*, 215−228.

(164) Pathak, Y.; Mehta, S.; Priyakumar, U. D. Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks. *J. Chem. Inf. Model.* **2021**, *61* (2), 689−698.

(165) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal* **2021**, *418*, No. 129307.

(166) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy. *J. Chem. Inf. Model.* **2022**, *62* (3), 433−446.

(167) Gebhardt, J.; Kiesel, M.; Riniker, S.; Hansen, N. Combining Molecular Dynamics and Machine Learning to Predict Self-Solvation Free Energies and Limiting Activity Coefficients. *J. Chem. Inf. Model.* **2020**, *60* (11), 5319−5330.

(168) Venkatraman, V.; Evjen, S.; Knuutila, H. K.; Fiksdahl, A.; Alsberg, B. K. Predicting ionic liquid melting points using machine learning. *J. Mol. Liq.* **2018**, *264*, 318−326.

(169) Datta, R.; Ramprasad, R.; Venkatram, S. Conductivity prediction model for ionic liquids using machine learning. *J. Chem. Phys.* **2022**, *156* (21), No. 214505.

(170) Dhakal, P.; Shah, J. K. Developing machine learning models for ionic conductivity of imidazolium-based ionic liquids. *Fluid Phase Equilib.* **2021**, *549*, No. 113208.

(171) Dhakal, P.; Shah, J. K. A generalized machine learning model for predicting ionic conductivity of ionic liquids. *Molecular Systems Design & Engineering* **2022**, *7* (10), 1344−1353.

(172) Mohan, M.; Jetti, K. D.; Smith, M. D.; Demerdash, O. N.; Kidder, M. K.; Smith, J. C. Accurate Machine Learning for Predicting the Viscosities of Deep Eutectic Solvents. *J. Chem. Theory Comput.* **2024**, *20* (9), 3911−3926.

(173) Nagulapati, V. M.; Raza Ur Rehman, H. M.; Haider, J.; Abdul Qyyum, M.; Choi, G. S.; Lim, H. Hybrid machine learning-based model for solubilities prediction of various gases in deep eutectic solvent for rigorous process design of hydrogen purification. *Sep. Purif. Technol.* **2022**, *298*, No. 121651.

(174) Wang, J.; Song, Z.; Chen, L.; Xu, T.; Deng, L.; Qi, Z. Prediction of CO2 solubility in deep eutectic solvents using random forest model based on COSMO-RS-derived descriptors. *Green Chemical Engineering* **2021**, *2* (4), 431−440.

(175) Lemaoui, T.; Boublia, A.; Lemaoui, S.; Darwish, A. S.; Ernst, B.; Alam, M.; Benguerba, Y.; Banat, F.; AlNashef, I. M. Predicting the CO2 Capture Capability of Deep Eutectic Solvents and Screening over 1000 of their Combinations Using Machine Learning. *ACS Sustainable Chem. Eng.* **2023**, *11* (26), 9564−9580.

(176) Mohan, M.; Demerdash, O.; Simmons, B. A.; Smith, J. C.; Kidder, M. K.; Singh, S. Accurate prediction of carbon dioxide capture by deep eutectic solvents using quantum chemistry and a neural network. *Green Chem.* **2023**, *25* (9), 3475−3492.

(177) Makarov, D. M.; Fadeeva, Y. A.; Golubev, V. A.; Kolker, A. M. Designing deep eutectic solvents for efficient CO2 capture: A data-driven screening approach. *Sep. Purif. Technol.* **2023**, *325*, No. 124614.

(178) Zhu, X.; Khosravi, M.; Vaferi, B.; Nait Amar, M.; Ghriga, M. A.; Mohammed, A. H. Application of machine learning methods for estimating and comparing the sulfur dioxide absorption capacity of a variety of deep eutectic solvents. *Journal of Cleaner Production* **2022**, *363*, No. 132465.

(179) Zhou, F.; Shi, D.; Mu, W.; Wang, S.; Wang, Z.; Wei, C.; Li, R.; Mu, T. Machine learning models accelerate deep eutectic solvent discovery for the recycling of lithium-ion battery cathodes. *Green Chem.* **2024**, *26* (13), 7857−7868.