Digital Chemical Engineering 14 (2025) 100207

Contents lists available at ScienceDirect

Digital Chemical Engineering

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

Original Article

Ensemble machine learning to accelerate industrial decarbonization: Prediction of Hansen solubility parameters for streamlined chemical solvent selection

Eslam G. Al-Sakkari [a,b], Ahmed Ragab [a,b], Mostafa Amer [c], Olumoye Ajao [d], Marzouk Benali [b,*], Daria C. Boffito [e], Hanane Dagdougui [a], Mouloud Amazouz [b]

[a] Department of Mathematics and Industrial Engineering, Polytechnique Montréal, succ. Centre-ville, Montréal, Québec, H3C 3A7, Canada
[b] Natural Resources Canada, CanmetENERGY, 1615 Lionel-Boulet Blvd, P.O. Box 4800, Varennes, Québec, J3 × 1P7, Canada
[c] Department of Electrical Engineering, Polytechnique Montréal, C.P. 6079, succ. Centre-ville, Montréal, Québec, H3C 3A7, Canada
[d] Natural Resources Canada, Clean Fuels Branch, Fuel Diversification Division, 580 Booth Street, Ottawa, K1A 0E4, Canada
[e] Department of Chemical Engineering, Polytechnique Montréal, succ. Centre-ville, Montréal, Québec, H3C 3A7, Canada

ARTICLE INFO

Keywords:
Machine learning
Biclustering algorithm
Decision fusion
Chemical analytics
Lignins
Smiles
Hansen solubility parameters

ABSTRACT

Several processes and strategies have been developed to promote the utilization of lignin and to facilitate its market adoption across a broad spectrum of applications within the expanding lignin bioeconomy. However, the inherent variability in lignin properties, resulting from diverse feedstock sources and varied recovery and downstream processing methods, remains a significant challenge. This highlights the critical need to investigate lignin's miscibility and reactivity with polymers and solvents, as most lignin valorization pathways involve mixing, blending, or solubilization. Accurate estimation of Hansen solubility parameters (HSP) is crucial for solvent selection in several fields such as polymer science, coatings, adhesives, lignin-based biorefineries and solvent-based carbon capture. Traditional methods for predicting HSP are time-consuming and involve complex experiments, especially in applications dealing with carbon dioxide and lignin solubility. This paper introduces a novel ensemble modeling methodology based on machine learning (ML) techniques for accurate HSP prediction using Simplified Molecular Input Line Entry System (SMILES) codes as entries. The methodology integrates different ML approaches, including deep and shallow learning, to enhance prediction accuracy. Decision fusion of individual ML models is achieved through a hybrid approach combining non-learnable and learnable methods, resulting in reduced errors and enhanced accuracy. The results highlight the effectiveness of the ensemble-based methodology, which achieved 99% accuracy in predicting dispersion solubility parameters, outperforming other individual ML techniques. The proposed generic methodology, from data preprocessing to decision fusion through diverse ML algorithms, can be applied to various chemical analytics beyond HSP prediction.

1. Introduction

The integration of artificial intelligence (AI), machine learning (ML), and AI-based predictive analytics is revolutionizing the fields of chemistry and process system engineering including biorefineries (Hashemi et al., 2024; Adeleke et al., 2024; Taqvi et al., 2021; Akinpelu et al., 2023; Emori et al., 2022; Pilario et al., 2022; Chmiela et al., 2023; Götz et al., 2023; York et al., 2024; Arias et al., 2023; Zeidler, 2024). Recent studies in prominent journals highlight the transformative potential of these digital technologies (Wen et al., 2022; Su et al., 2019; Zhang et al., 2023; Wang et al., 2022; Su et al., 2020; Wen et al., 2023). By harnessing the power of AI, researchers are advancing the understanding and prediction of molecular interactions that govern separation processes and the behavior of complex molecules and materials, enabling more accurate predictions and deeper insights into molecular structures (Ritt et al., 2022; Unke et al., 2024; Sanchez-Lengeling and Aspuru-Guzik, 2018; Zhang et al., 2022; Khan and Ammar Taqvi, 2023). Additionally, AI-driven models are being utilized to predict environmental impacts/properties and process emissions (Zhang et al., 2022), (Wang et al., 2020), (Wang et al., 2019), (Jablonka et al., 2023), enabling more sustainable and environmentally friendly industrial practices, as well as predicting catalyst selectivity (Zahrt et al., 2019). Moreover, the

* Corresponding author.
E-mail address: marzouk.benali@nrcan-rncan.gc.ca (M. Benali).

https://doi.org/10.1016/j.dche.2024.100207
Received 23 September 2024; Received in revised form 22 November 2024; Accepted 3 December 2024
Available online 13 December 2024
2772-5081/Crown Copyright © 2024 Published by Elsevier Ltd on behalf of Institution of Chemical Engineers (IChemE). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

acceleration of reaction condition identification through AI is significantly reducing the time and cost associated with experimental trials (Meuwly, 2021), (He et al., 2024).

It is in this context that accurate prediction of lignin properties and their functionality is crucial for the efficient design and operation of large-scale lignin-biorefineries (Ajao et al., 2021; O'Dea et al., 2022; Ginni et al., 2021; Li et al., 2022). The complexity and variability of lignin make it challenging to process, but understanding its properties can lead to significant advancements in producing high-value lignin derivatives (Li et al., 2023). AI and ML methods are essential in accelerating the development and evolution of such biorefineries (Arias et al., 2024), (Balsora et al., 2022). AI and ML can analyze vast amounts of data from various sources, including experimental results, literature, and operational data (Lofgren et al., 2022; Khashaba et al., 2022; Garcia et al., 2022; Ge et al., 2023). This analysis can uncover patterns and relationships that are not immediately apparent, leading to a deeper understanding of lignin properties and processing methods. ML algorithms can create predictive models that accurately forecast lignin behavior under different conditions, which requires advanced AI approach combining ensemble-based learning algorithms and fusion of data from diverse sources.

The reasons for the creation and use of an ensemble-based machine learning for decision fusion include statistical considerations, data-related limitations, and models/techniques-related constraints (Varshney et al., 2023), (He et al., 2022). From the statistical point of view and taking the neural network modeling for classification and regression as an example, it is common that the high accuracy of results during model training does not guarantee the same results when developed models are tested (Morimoto et al., 2022). This is well-known as the problem of AI/ML generalization to model unseen data where their ability to adapt properly to this new and previously invisible information can be hindered due to various aspects (Sester et al., 2018), (Hui et al., 2022). A possible cause is that the testing datasets may not be representative of the same distribution/range as that of the training dataset. However, employing different models complemented by results averaging or majority voting can decrease generalization risk (Tidriri et al., 2018). The data-related limitations are the variation between the processing/modeling of very large or limited available data (Wang et al., 2016; Brigato and Iocchi, 2021; Peng et al., 2018). In the case of very large datasets, it is better to split this dataset into smaller arrays to make the model training step more efficient. On the other hand, in the case of limited data, it is difficult to construct a representative dataset where the model can be successfully trained to catch the underlying distribution of the whole population. In this regard, data resampling and constructing other different data inventories from the parent dataset with limited or imbalanced inputs can be a good choice (Ghorbani and Ghousi, 2020). These datasets can be modeled using different ML techniques and their results can be subsequently combined/fused to obtain more accurate predictions. With respect to the models/techniques-related constraints, no single method or model performs well enough for all problems; instead, each has its merits and demerits as well as regions of best performance. This is well-known as the *no-free-lunch* theorem (Adam et al., 2019), (Wolpert, 2002). Hence, the diversification of methods and techniques is a key solution that can play an inevitable role in reaching the right decision (Ragab et al., 2022).

Given the above-mentioned needs and challenges, several methods have been proposed in literature for implementing ensemble-based learners, especially in the classification problems. These methods of ensemble creation include bagging (Sutton, 2005), (Breiman, 1996), boosting (Schapire, 2003), (Schapire and Freund, 2013), AdaBoost (Schapire, 2013) and stacked generalization (Wolpert, 1992), (Naimi and Balzer, 2018) mechanisms. AdaBoost refers to a particular method of training a boosted classifier. Combining the results of the ensembles takes several forms such as majority voting (Dietterich, 2000), (Atallah and Al-Mousa, 2019), weighted majority voting (weighted averaging) (Neloy et al., 2022), (Dogan and Birant, 2019) and behavior knowledge space (Ragab et al., 2022) for categorical output variables. Some combination methods are specific to continuous outputs, and they include averaging (Abba et al., 2020), weighted averaging (Mehta et al., 2019) and *Dempster-Shafer* based combination (Sentz and Ferson, 2002).

As a response to these needs and technical challenges of single ML-modeling, a new ensemble-based methodology is proposed for a decision fusion in regression problems seeking high predictability and generalization during testing and validation. This is inspired by the bagging and boosting methods/techniques that have been successfully used in classification problems as well as other ensemble-based modeling time series cases (Mian et al., 2024; Asri et al., 2024a; Asri et al., 2024b). To evaluate the performance and effectiveness of the proposed methodology, the prediction of the solubility of various solvents/substances was selected as a case study.

Hansen solubility parameters (HSPs) are well known as a unique predictive approach to quantify solubility considering the molecular permanent dipole and molecular hydrogen bonding interactions. The HSPs are firstly presented by Prof. *Charles M. Hansen* in his PhD thesis in 1967 (Hansen, 1967), (Hansen, 2007). They represent the combination of dispersion (δ_D), polarization (δ_P) and hydrogen bonding (δ_H) energies to explain the mechanism of solvation for single solvent or solvents mixture systems (Bapat et al., 2021), (Han et al., 2019). Unlike the *Hildebrand* solubility parameter (Sreekanth et al., 2012) which focuses mainly on the non-polar solvents, *Hansen* solubility parameters are suitable for describing both polar and non-polar solvents systems (Venkatram et al., 2019). This is because they consider polarization and hydrogen bonding energies as previously mentioned (Venkatram et al., 2019).

There are three main traditional approaches/methods for the calculation of different solubility parameters including *Hansen* solubility parameters (Ribeiro et al., 2020). These approaches/methods include the functional group contribution method (Stefanis and Panayiotou, 2008), (Sistla et al., 2012), the intrinsic viscosity method (Gharagheizi and Torabi Angaji, 2006) and the properties correlations or regression method (Tamura and Yamamoto, 2019), (Zhao et al., 2018). However, these methods have some limitations, including limited access to reliable data, reliance on simplified assumptions (e.g., ideal solution behavior), and the need for approximations. Additionally, they are time consuming as they primarily depend on several lab experiments (Ribeiro et al., 2020), (Ruwoldt et al., 2022). There is also a new trend to calculate these parameters using computational methods such as molecular dynamics, density functional theory (DFT) and other quantum mechanics-based calculations (Mohan et al., 2022), which are also computationally resource intensive and time consuming despite their lack of robustness. Therefore, new faster methods are needed and in response several data-driven methods have been proposed recently to overcome these limitations (Przybyłek et al., 2019). For instance, AI/ML-based approaches are garnering attention to calculate/predict *Hansen* solubility parameters as well as other molecular and thermodynamic properties in general (Jackson et al., 2019). These approaches have been investigated as a faster way to estimate different properties of interest with acceptable accuracies (Leonard et al., 2021). Based on their nonlinearity, these approaches create new QSPR/QSAR (Quantitative structure–property relationship/Quantitative structure–activity relationship) models with the above-mentioned merits (Jarvas et al., 2011). Examples of the ML-based models used in predicting *Hansen* solubility parameters and their space are neural networks (Chandrasekaran et al., 2020), (Perea et al., 2016), combined Gaussian processes & Bayesian ML approach (Sanchez-Lengeling et al., 2019), random forest (Obradović et al., 2018) and support vector machine/regression (Delbecq et al., 2020). It is worth noting that the support vector machine was also used for the prediction of *Hildebrand* solubility parameter (Rexhepi et al., 2023). The selection of an appropriate solvent is critical for applications involving polymers, where identifying suitable solvents and non-solvents is crucial. This necessity has driven the development of quantitative models of polymer-solvent compatibility based on the

principle that "like dissolves like." Furthermore, Gaussian Process Regression (GPR) has been employed to predict both the Hildebrand and Hansen solubility parameters for various polymers. These predictions facilitate the classification of materials as solvents or non-solvents for specific polymers (Venkatram et al., 2019). The accuracy of the models used in most of these studies ranged from around 50% to above 90%. As observed, some cases possess low prediction accuracy, which can affect the calculations done by experts in future applications such as solvent-polymer compatibility determination. In addition, in some cases the dataset utilized was relatively small, i.e. below 100 points, which affect model/method generalization. All these limitations/drawbacks along with some suggested solutions will be discussed in detail in the upcoming paragraphs.

To facilitate the processing of different chemical molecules and their related molecular properties historical data used for training, a new form for molecular structure representation, i.e. SMILES codes, is introduced and increasingly used as an open standard in chemistry for formula representation since 2007 (Alshehri et al., 2020). The SMILES expression stands for Simplified Molecular Input Line Entry System (US-Environmental Protection Agency). These codes introduce a simplified way for representing chemical formulas and structures easy to be used by computers that can be then converted to ML-readable descriptors/features (Fan et al., 2023).

The utilization of data-driven computational models presents a significant advantage in the prediction of solvent and material properties from their SMILES representations, alongside pertinent geometric and molecular descriptors (Pyzer-Knapp et al., 2022), (Chen and Qian, 2023). This approach notably reduces the duration traditionally required for experimental determination through labor-intensive laboratory procedures. Moreover, these models can be trained using historical datasets obtained from alternative computational methodologies, thereby serving as expedient surrogate models that can potentially supplant conventional techniques (Perea et al., 2016). Ensemble or combination of different techniques/methods in parallel and sequential ways can be a novel approach to overcome the problem of low accuracies of single models. This is the core of the proposed methodology that will be discussed in detail in the subsequent sections. Besides, the black box-based calculations by some techniques represent a significant obstacle to understanding the impact of each feature/descriptor on the output variables of interest, i.e., *Hansen* solubility parameters. Fortunately, automatic features selection of tree-based techniques including the ensemble ones can help in solving this problem. In addition, the employment of explainable AI (XAI) techniques will give a complete description of these impacts qualitatively and quantitively (Kobayashi and Alam, 2024). Hence, it is recommended to include these tree-based models in the new ensemble method followed by the employment of XAI.

The following points summarize the challenges, gaps, and needs that prompted this study and the development of the proposed methodology:

- Existing machine learning techniques exhibit suboptimal prediction accuracies.
- Current feature selection methods are often inefficient, relying on manual selection as observed in some literature.
- There is a prevalent issue of black-box modeling, leading to a lack of interpretability and explanation.
- There is an imperative for a precise tool to predict Hansen solubility parameters for various materials, including solvents, which is crucial for optimizing industrial processes such as lignin valorization, carbon capture, and polymer manufacturing.

This work aims to propose a novel framework for data preparation, feature selection, and decision fusion, enhancing prediction accuracy. Furthermore, we extend the application of various AI and ML techniques to achieve precise predictions of *Hansen solubility* parameters, serving as a practical case study. The scientific and technical contributions of this study are summarized as follows:

1. Investigate the use of bi-clustering as a powerful data mining approach (inspired by the successful work done in the field of biology and bioinformatics) to perform robust features selection and use advanced non-linear clustering techniques to optimize the parameters of bi-clustering graphically.
2. Develop a decision fusion method to enhance the prediction accuracy based on gathering different machine learning models, each having its own advantages, through different mechanisms.
3. Validate the developed methodology to perform highly accurate prediction of Hansen solubility parameters using SMILES codes as inputs. This will help in the accurate definition of the *Hansen* sphere to select the best solvents to dissolve a certain material, e.g. diverse lignins and carbon dioxide (CO_2).
4. Interpret the black-box models by incorporating XAI techniques to extract knowledge and actionable insights that help the users, e.g., material and process designers accelerate the solvent selection and design optimization procedures afterwards.

The proposed AI-powered tool will help discover new solvents for the materials of interest by exploring the predicted *Hansen* solubility parameters. In addition, it is the first step towards the acceleration of AI-assisted solvents material design. Where, there is no unique chemical structure (e.g., lignin chemical structure is feedstock and separation process dependent) or the new non-existing solvents will be generated/discovered and designed based on the desired *Hansen* solubility parameters as desired input properties. It can also be used for the identification of novel lignin-based products and potential derivatives.

2. Methods

The general approach for this work is summarized in Fig. 1. As depicted in this figure, the raw data is first collected and preprocessed to ensure its reliability and make it ready to perform the analysis. After that, the cleaned/preprocessed data is partitioned to training and validation sets based on the different case studies and human expertise. The training dataset is then utilized to train an ensemble of different ML techniques. The results are then fused to enhance the overall prediction accuracy. As is clearly stated in the figure, all the work is done under the supervision of human experts, where their expertise is applied in each step. This will ensure rationality and maximum possible modeling efficiency based on the data-human expertise-AI/simulation combined approach (Al-Sakkari et al., 2024), (Abdeldayem et al., 2022). In return, the results after fusion give new insights helping in increasing knowledge/expertise to tackle process improvements. The work should be interdisciplinary where the expertise in different fields will be exchanged and augmented to achieve high prediction accuracies.

2.1. Data cleaning, preprocessing and validation

It is well known that data is the foundation of AI modeling. Thus, careful preprocessing and preparation of the data is essential to ensure high prediction accuracy. The raw data can be in various forms such as texts/strings, tabular data, images, etc. As illustrated in Fig. 2, the first step is to clean the raw data. This can be done by removing any empty, missing, incomplete or erroneous values. Additionally, different validation techniques should be applied to further clean the data. Efficient computational tools should be utilized to test the different forms of available raw data to remove any invalid points, e.g., incorrect texts, invalid symbols, and any incorrect mix between numerical & categorical/alphabetical values. This step is key to ensure high accuracy and avoid confusion. After invalid data elimination, the valid ones are gathered in a new refined dataset ready for further processing. The cleaned data is then further processed to extract the most representative and informative features to be fed to the analysis and ML modeling.

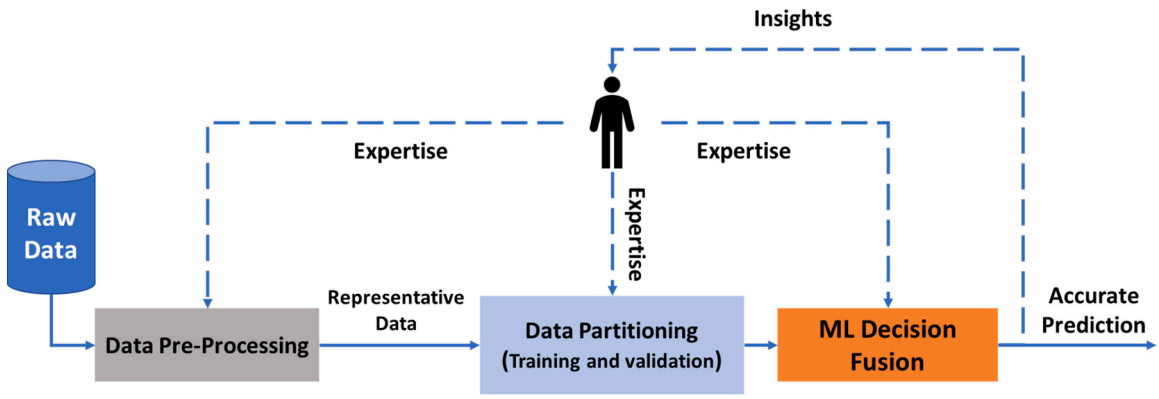

Fig. 1. General approach.

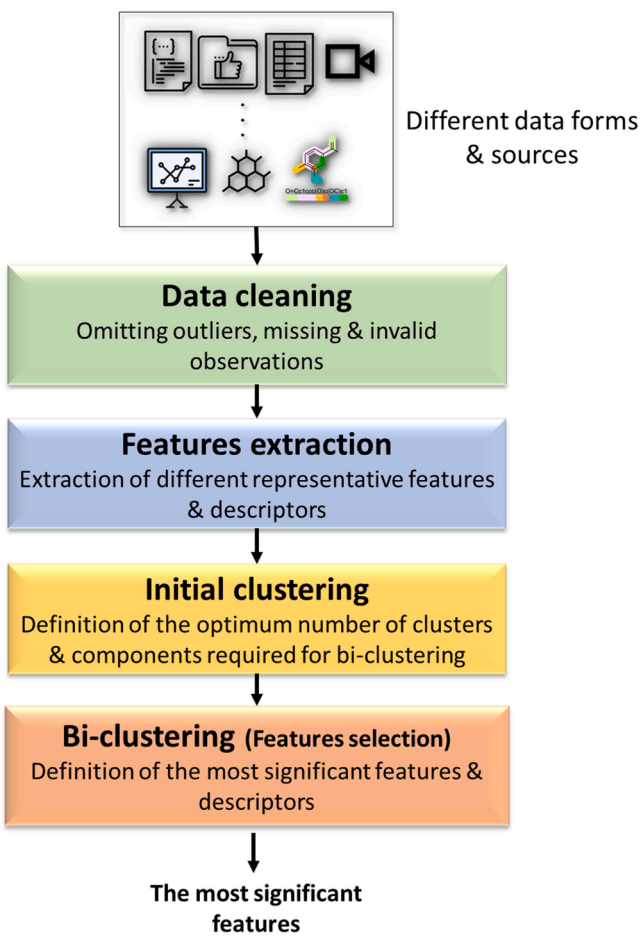

Fig. 2. Data preprocessing methodology.

These features can be in the form of various descriptors, eigenvalues, eigenvectors, and fingerprints. Afterwards, they are gathered and combined with other variables to start assigning the inputs and outputs. Commonly, the features and data points in the available datasets are in large number, which can be computationally expensive and time consuming. Besides, some of the features in the raw data can be confusing and misleading thereby hindering the prediction accuracy upon ML modeling. Therefore, a dimensionality reduction step is essential at this stage for proper features selection. In this regard, employing biclustering is proposed as a powerful data mining and dimensionality reduction approach to perform this task. However, it is key to determine the optimum number of components where the data will be reduced to. This step can be time-consuming and computationally expensive, especially with larger datasets. Hence, the utilization of other clustering techniques for data visualization, in the initial clustering step, is suggested to determine the optimum number of components (latent dimensions) graphically. The models used for clustering is further discussed in Section 2.3.

2.2. Biclustering

Biclustering is a useful technique for dimensionality reduction and data mining that gathers both rows and columns of the input data simultaneously by identifying other submatrices with rational similarity patterns (Madeira and Oliveira, 2004), (Kalna et al., 2008). Within the context of ML, biclustering can be utilized for feature selection where the goal is to identify subsets of the most relevant features depending on the case study or the task (Farhan et al., 2016; Liu and Motoda, 1998; Dy, 2007). There are several methods to perform bi-clustering (Prelić et al., 2006) such as spectral bi-clustering (Kluger et al., 2003), *Plaid* models (Henriques and Madeira, 2015), *Bayesian* bi-clustering (Gu and Liu, 2008), (Meeds and Roweis, 2007), sparse singular value decomposition (SSVD) (Lee et al., 2010) and non-negative matrix factorization (NNMF) (Li and Ngom, 2013), (Carmona-Saez et al., 2006). In the proposed methodology, the use of NNMF is recommended due to its easy implementation, high accuracy, clear visualization of results, and its successful application in the biomedical/bioinformatics fields. Fig. 3 presents a simple schematic illustrating the general concept behind NNMF biclustering. Other schematics are presented in supplementary materials.

In the presented schematic, which is the most common one, the original data matrix (A) is decomposed into two new matrices. The first one is the feature matrix (W) and the second is the coefficient matrix (H). Where, m, n and k are the number of features, points/instances, and components of reduced data. For the mathematical background, types and algorithms of this method, readers can refer to (Lee and Seung, 2000; Wang and Zhang, 2012; Lee and Seung, 1999). The different methods of NNMF performance evaluation are discussed in Section 2.5. It is important to note that NNMF can be computationally expensive and complex, especially when optimizing its key hyperparameter—the number of components—and when working with very large datasets. The complexity of NNMF is primarily influenced by the dimensions of the original matrix ($m \times n$) and the number of components (k). Therefore, performing data clustering to reduce the problem's dimensionality before biclustering can offer a promising solution. Besides, in this study, we propose a promising and effective solution using clustering techniques to determine the optimal number of components, thereby improving the scalability of NNMF for larger datasets.

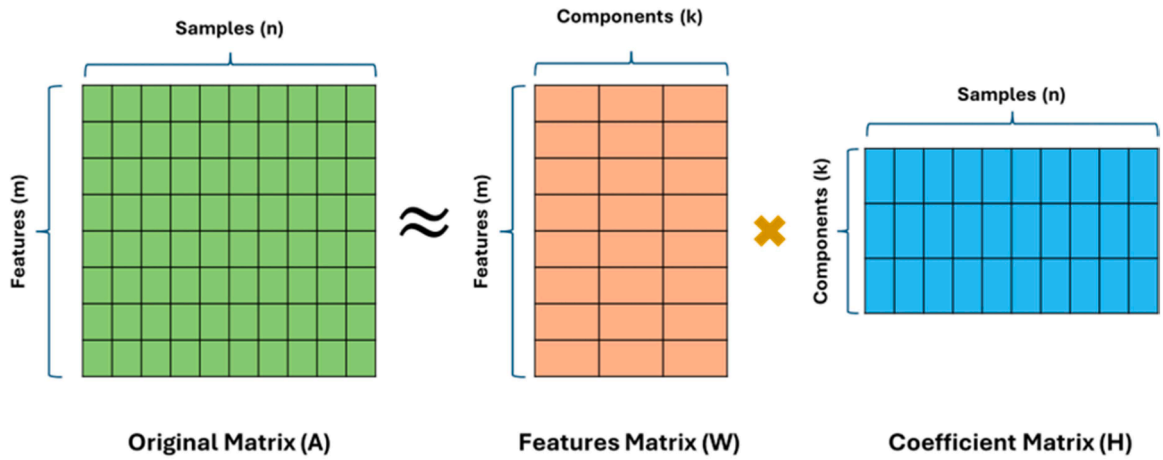

Fig. 3. NNMF concept simple schematic (adapted from (Non-Negative Matrix Factorization)).

2.3. Clustering

In the proposed methodology, uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), (Ghojogh et al., 2021) and t-distributed stochastic neighbor embedding (t-SNE) (Wang et al., 2023), (Poličar et al., 2019) combined with *k*-means algorithm (Sinaga and Yang, 2020) were chosen as efficient dimensionality reduction, clustering and data visualization techniques. There are many other dimensionality reduction techniques being widely used in literature including principal component analysis (PCA) (Reddy et al., 2020), (Hasan and Abdulazeez, 2021). However, these two techniques, i.e., UMAP and t-SNE, were selected because of their ease of handling non-linearity, preservation of data structure and flexible visualization (Wang et al., 2021), (Gisbrecht et al., 2015). For instance, one of the limitations of PCA is its linearity where, in contrast, UMAP and t-SNE are nonlinear techniques (Anowar et al., 2021). In addition, UMAP can preserve the local and global data structures (Choi et al., 2024). They are also flexible and have better visualization. Moreover, UMAP is relatively computationally inexpensive where its processing time is relatively short and can handle large datasets. For more information about the mathematical background and assumptions considered during the implementation of UMAP and t-SNE, readers can refer to (McInnes et al., 2018), (Anowar et al., 2021), (der Maaten and Hinton, 2008). The different methods of UMAP and t-SNE performances evaluation are discussed in Section 2.5.

2.4. ML modeling and decision fusion

The ML modeling and decision fusion methodology is illustrated in Fig. 4. As shown, after preparing the data, these data containing the combinations of the most representative features is fed to the selected ML algorithms. Before modeling, the data is divided into two parts, i.e., one used for training the models whereas the second part is used for validation. After modeling, the results of each optimized model are fused

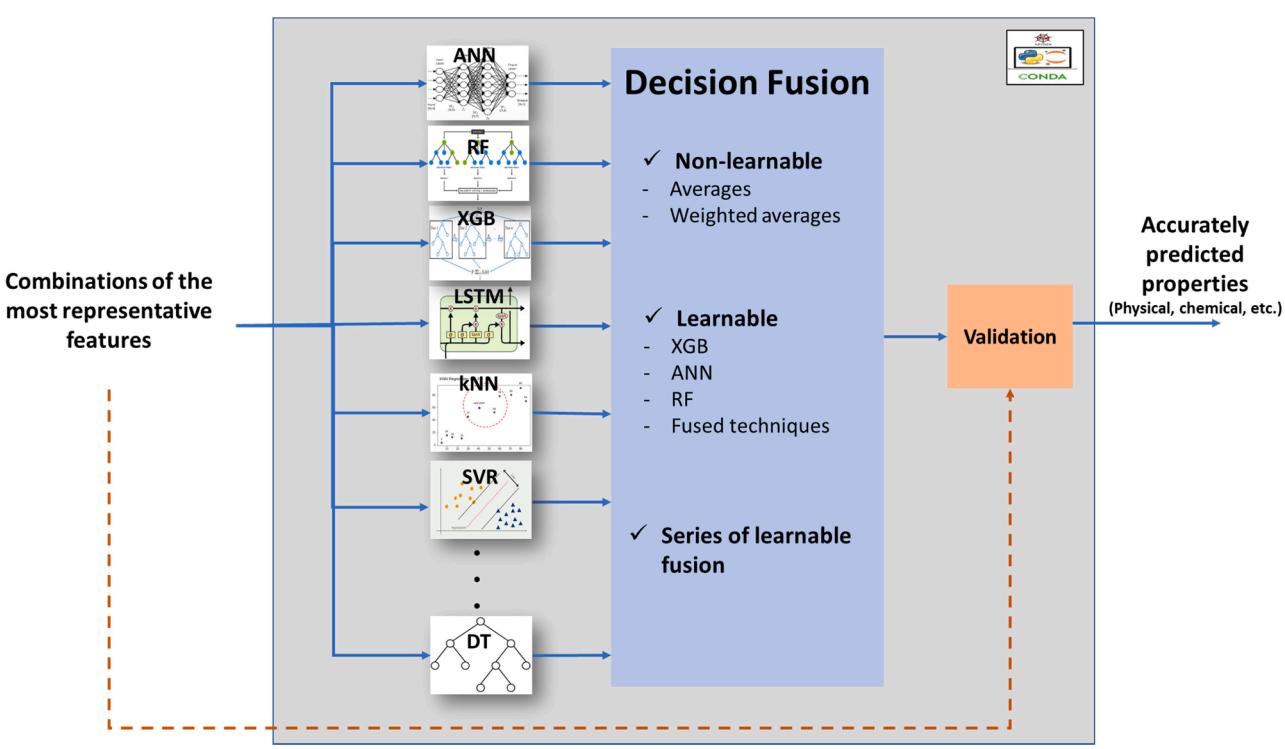

Fig. 4. Decision fusion methodology.

through different ways, i.e., learnable, and non-learnable fusion. The upcoming sections elaborate more on the data partitioning, ML models considered, and the different fusion methods proposed.

2.4.1. Data partitioning for model training

As previously mentioned, the analysis-ready data will be divided into two sets, one for training and the other one for validation. There are several factors that control the choice of the percentages of both training and validation subsets. These factors include but are not limited to the available dataset size, which determines the amount of information available for training, the complexity of the models to be developed and the models' hyperparameters and their number. Yet, based on the knowledge of the human experts, there are common values of these percentages typically employed for several machine learning modeling case studies. The related common values are as follows:

- **70 to 30 split ratio:** this case considers 70% of the input data to train the developed model, and the remaining 30% is used for model validation. This ratio is usually employed for relatively large datasets and relatively simple models.
- **80 to 20 split ratio:** This is a common split ratio in the case of reasonably large datasets and significantly complex models. Where, 80% of the data is dedicated for models training and the other 20% is devoted to validating the developed model.
- **90 to 10 split ratio:** when the input dataset is small and the model is simple, 90% of the input data is used for model training and only 10% of it is reserved for model validation.

In addition, the validation datasets are usually employed to facilitate hyperparameters tuning. Thus, based on the model complexity and the number of their hyperparameters to be tuned, the validation dataset size will vary significantly. Moreover, an additional set may be utilized in specific cases to evaluate/test the models' performance after their training and hyperparameters tuning. It is worth noting that the splitting of the data is done randomly to ensure generalization and avoid any bias during the training and validation.

2.4.2. ML techniques under consideration

Several ML techniques (shallow and deep learning) are proposed for being employed at the first stages of modeling; however, the most relevant ones are selected based on the determination coefficient (R^2) and mean squared error (MSE). These techniques are the random forest (RF), artificial neural network (ANN), support vector regression (SVR), optimized decision tree (DT), *k*-nearest neighbour (*k*-NN) regression, extreme gradient boost (XGB), long short-term memory (LSTM), Gaussian process regression (GPR) and convolutional neural network (CNN). As it can be deduced, this is a wide spectrum of techniques where shallow, deep learning, non-parametric and ensemble methods are employed. Each one of these techniques has its own merits and that is why the utilization of all these different techniques is proposed. It is worth mentioning that these models were selected based on an extensive literature review and the widely recognized best practices in the AI/ML community (Taqvi et al., 2021), (Wen et al., 2023), (Khan and Ammar Taqvi, 2023), (Al-Sakkari et al., 2024), (Hu et al., 2021). Table 1 summarizes the key merits of each individual model, which justify their selection and motivate the development of the *Ensemble-of-Ensembles* technique in this study. For an in-depth explanation and the mathematical foundation of each technique, readers are encouraged to consult the references listed in Table 1.

2.4.3. Hyperparameter optimization of ML techniques

Each technique (except the non-parametric *k*-NN lazy ML model) has hyperparameters that should be optimized to obtain the best possible prediction accuracy. In the current study, the utilization of *Grid Search* and *Bayesian Optimization* techniques are suggested to perform this hyperparameter optimization. Table 2 summarizes the different

Table 1
Overview of the different ML techniques and their key merits.

Model	Description	Merits	Reference(s)
***k*-NN**	A non-parametric supervised ML technique that does not assume certain distribution. It is widely used for both regression and classification based on the number of nearest similar neighbours.	Simple and fast model that requires inexpensive training.	(Zhang, 2016; Ray, 2019; Zhao et al., 2017)
DT	Supervised interpretable ML technique that performs classification and regression through splitting the data based on specific parameters (cut points). It is a flow chart-like algorithm.	Easy handling and interpretation. Gives interpretable patterns.	(Ray, 2019)
RF	An ensemble-based ML technique consists of multiple weak decision trees and represents an advanced form of bagging.	Easy handling and it can overcome the problem of overfitting.	(Liu et al., 2012; Schonlau and Zou, 2020; Biau and Scornet, 2016)
XGB	An ensemble of decision trees that considers the boosting mechanism and level-wise tree growth.	Fast and high accuracy predictions besides overcoming the overfitting problem and handling the missing data	(Chen et al., 2015), (Yu et al., 2020)
SVR	Supervised ML model having the objective of finding/creating a hyperplane that fits the training data while keeping a specific margin around this hyperplane. The regression model and its efficiency are determined by the closest points around this hyperplane which forms the "support vectors".	Easy tuning, memory efficiency, high performance even when the data has noise or outliers and the flexibility in handling both linear and non-linear relationships in the data.	(Zhang and O'Donnell, 2020; Smola and Schölkopf, 2004; Awad et al., 2015)
ANN	A class of ML models inspired by the human brain cells (neurons) and the structure of human brain that consists of three main types of layers, i.e. input layer, hidden layer(s) and output layer.	It can handle a large amount of structured and tabular data, possess high flexibility when dealing with complex and non-linear data, can adapt to new data points.	(Specht, 1991; Sen et al., 2023; Heiat, 2002)
LSTM	A type of recurrent neural network that is designed and used to handle/model sequential data.	It can handle large datasets, excellent performance when dealing with sequential data having long-term dependencies such as language structures, can avoid the vanishing gradient problem and can handle irregularly spaced data points (sequences with varying lengths).	(Sherstinsky, 2020), (Staudemeyer and Morris, 2019)
CNN	A special type of artificial neural networks developed to process and identify	It can capture both high and low levels of information from the available data. Further,	(Li et al., 2021; Gu et al., 2018; O'Shea and Nash, 2015; Albawi et al., 2017)

(*continued on next page*)

Table 1 (*continued*)

Model	Description	Merits	Reference(s)
	patterns in grid-like data and images. Its main components are convolutional, pooling, and fully connected layers.	it can handle large datasets and offers the advantage of transfer learning	
GPR	A non-parametric probabilistic machine learning model. It works according to the Gaussian processes concept representing distributions over functions. It models the entire space of the function and gives predictions in probability distributions form instead of estimating the parameters of a specific model.	It can handle noisy and limited data	(Schulz et al., 2018)

Table 2
Hyperparameters and their corresponding ranges.

Model	Hyperparameters	Ranges
DT	Depth	8–64
RF	Number of estimators (trees)	10–300
XGB	Number of estimators (trees)	10–300
SVR	C penalty of misclassification	1–500
	Gamma decision boundary range	*Scale, Float & Auto*
	Kernel function	*RBF, Sigmoid & Poly*
ANN	Learning rate	0.0005–0.1
	Batch size	1–128
	Number of hidden layers	1–5
	Number of neurons per hidden layer	20–200
	Activation function	*Tanh & ReLU*
LSTM	Number of LSTM layers	1–3
	Number of LSTM units in each layer	10–100
	Number of units in dense layers	1–3
	Batch size	8–64
	Number of epochs	20–200
CNN	Number of convolutional layers	1–3
	Number of filters per convolutional layers	16–256
	Number of dense layers	1–3
	Number of units per dense layers	16–256
	Kernel size	1–5
	Activation function	*Tanh & ReLU*
	Pooling size	1–5
GPR	Kernels	*RBF, Constant & White noise*

hyperparameters of each technique along with their corresponding proposed ranges. It is worth noting that the determination of optimum k, i.e., number of neighbours, in the case of k-NN is also essential. Hence, according to literature, the selected range for k is from 2 to 7 neighbours to avoid high bias as well as overfitting.

2.4.4. Decision fusion techniques

The techniques proposed for fusion in this study are divided into two main categories. The first one is the non-learnable fusion where the decision (the final prediction) is taken based on the average and weighted average of all the results of the ML techniques. In the average-based ensemble modeling, the final prediction output is the average of the outputs obtained from single models where all the values of each model have the same weight. In the case of weighted average, the weights are put based on the determination coefficient, mean squared error and a combination of both. The mathematical representations of the final output predictions based on the proposed average weights are introduced by the following formulas:

$$P_{if} = \frac{\sum P_{ij} \times R_j^2}{\sum R_j^2} \tag{1}$$

$$P_{if} = \frac{\sum P_{ij} \times \frac{1}{MSE_j}}{\sum \frac{1}{MSE_j}} \tag{2}$$

$$P_{if} = \frac{\sum P_{ij} \times \frac{R_j^2}{MSE_j}}{\sum \frac{R_j^2}{MSE_j}} \tag{3}$$

Where P_{if} is the final predicted value of instance/obervation (i) based on weighted average of models (j) results, P_{ij} is the predicted value of instance (i) based on model (j), R_j^2 is the determination coefficient of model (j) and MSE_j is the mean squared error of model (j) for all i,j.

The second fusion method is the learnable fusion where all the data points obtained from training and validating the ML techniques are gathered and used to train some other ML techniques inspired by staking-based ensemble modeling. In this case the inputs are the predicted values obtained from various single models, average-based ensemble and weighted average-based ensemble models whereas the outputs are the actual values (experimental or computational raw data of variable of interest). In addition, the techniques used are ANN, XGB, RF and their fusion. A new variant of the learnable fusion is the series of learnable fusions to maximize the accuracy of prediction as much as possible. This is done through continuous learning of the decision fusion results on multiple stages. In particular, the results from the first learnable fusion step are fed to another learnable fusion agent/model which in turn gives new results that are introduced to another fusion stage and so on. In the methodology in the current study, it is recommended to perform this series on three stages. Fig. 5 shows a simple schematic of the proposed series of learnable decision fusions.

It is worth noting that non-learnable fusion techniques are proposed and implemented as an intermediate stage, with their results used as inputs for the learnable fusion technique. The objective is to develop an ensemble-based approach that combines simple statistical methods (i.e. non-learnable techniques) with those based on ML concepts comprising shallow and deep learning. This multi-stage fusion of non-learnable and learnable techniques helps reduce error and improve prediction accuracy. However, we intentionally avoided relying heavily on other statistical techniques to prevent introducing pre-assumptions about the data. Our learnable fusion techniques are based on both shallow ML and deep learning concepts, aiming to capture the actual distribution without the need for approximations.

2.5. Results evaluation and validation

In the case of clustering, silhouette score and the silhouette plot analysis method are employed to enable the selection of optimum components and clusters numbers. Whereas, in the case of bi-clustering other additional metrics, i.e., explained variance (EV) and reconstruction error (RE), are evaluated. The explained variance can be referred to as the determination coefficient, but it is multiplied by the total variance. In addition, the results of machine learning modeling are evaluated mainly based on the model accuracy represented by determination coefficient (R^2) and mean squared error (MSE). The mathematical representations of these indicators are presented by the following equations:

$$EV = R^2 \times Total\ Varriance = \frac{SS_{exp.}}{n} \tag{4}$$

$$RE = \frac{1}{n} \times \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 \tag{5}$$

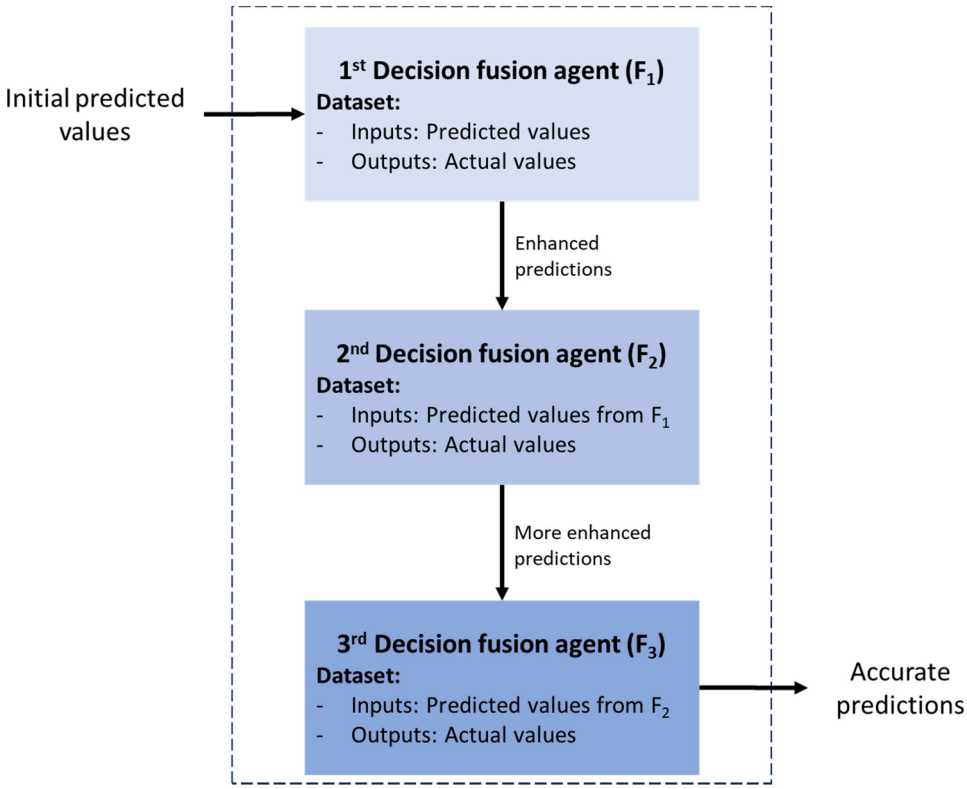

Fig. 5. Series of learnable decision fusions.

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_{i_{Predicted}} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_{i_{Actual}} - \bar{y}\right)^2} = \frac{SS_{\text{exp.}}}{SS_{tot.}} \tag{6}$$

$$MSE = \frac{1}{n} \times \sum \left(y_{i_{Actual}} - y_{i_{Predected}}\right)^2 \tag{7}$$

$SS_{exp.}$ is the explained sum of squares, n is the number of observations, $SS_{tot.}$ is the total sum of squares, x_i is the i th original data point and $\hat{x}_i$ is the i th reconstructed data point.

2.6. Models explainability

To avoid black-box modeling problem, an explainable AI (XAI) method was employed with the optimized ML models. This method is the calculation of SHAP values. By employing this method, the developed ensemble-based regression model can be considered a white box as it gives the weights related to the different input variables that define their relative impacts on the output variables. The weights are given in the form of relative magnitudes with their corresponding signs, i.e., positive, or negative, to totally define the impacts. Readers interested in more details about this method and its other possible applications can refer to this recently published study (Al-Sakkari et al., 2023).

3. Case study: Hansen solubility parameters estimation

In the present study, we focus on estimating the Hansen solubility parameters through a proposed case study. Initially, we introduce the data sources and the nature of the variables. Subsequently, we provide a detailed description of the preprocessing methodology.

3.1. Data source and type

The dataset in the current study is gathered from the literature and *HSPiP* software (Abbott and Hansen, 2008; Hansen Solubility Parameters in Practice (Official Web Page); de los R\'\ios and Hernández Ramos, 2020; De La Peña-Gil et al., 2016). The data essentially included solvents names, their SMILES codes, and thermodynamic properties. The properties covered are Antoine parameters, boiling points, melting points, flash points, vapor pressures, water solubility and Hansen solubility parameters. The variables initially selected for the HSP prediction are the SMILES codes and the three *Hansen* solubility parameters. Table 3 illustrates an excerpt of the data and variables used for ML modeling along with the corresponding compound's name of each SMILES code. SMILES codes were selected as the inputs instead of only the solvents names as they can be easily used to extract more features to well define the solvents. Features extraction will be discussed in more detail in the upcoming section. As can be observed, the data in the current state is heterogeneous, i.e., SMILES are alphabetical/categorical and hence the features extraction step is essential to convert SMILES code to numerical representations that are homogeneous with the HSP and easy to process for the prediction purpose.

HSP prediction/calculation for diverse materials and solvents is an essential step for different applications. For example, lignins fractionation and blinding with diverse polymers at high compatibility depend on the accurate calculation of their HSP (Ajao et al., 2021). However, as discussed in (Ajao et al., 2021), the common methods for identifying suitable solvents for different lignins are based on arbitrary selection of solvents and performing empirical laboratory procedures to assess their performance (Schieppati et al., 2023). Additionally, compatibility with polymers is assessed through an *Ad hoc* way that requires prior expertise. Therefore, this study (Ajao et al., 2021) suggested lignin' HSP calculation as a robust and systematic way to overcome the drawbacks of the other common methods. On the other hand, the traditional HSP calculation methods including experimentation and computational work suffer from being resource intensive and time consuming. Hence, the fast and easy predictions of these important parameters is crucial to accelerate the prediction of lignin computability with diverse polymers and technological adoption on large scale.

The compounds in this dataset represent a broad selection of known solvents used in several industrial applications, including lignin

Table 3
Dataset excerpt - Solvents with their SMILES codes and Hansen Solubility Parameters.

Solvent Name	Raw Input Variable	Output Variables		
	SMILES Codes	D	P	H
Acetone	CC(C)=O	15.5	10.4	7
Acetonitrile	CC#N	15.3	18	6.1
Benzyl Benzoate	O = C(OCC2=CC=CC=C2)C1=CC=CC=C1	20	5.1	5.2
tert-Butanol	O[C@](C)(C)C	15.2	5.1	14.7
n-Butyl benzoate	O = C(C1=CC=CC=C1)OCCCC	18.3	5.6	5.5
Butyldiglycol acetate	CCCCOCCOCCOC(=O)C	16	4.1	8.2
Dimethyl isosorbide (DMI)	O([C@@H]1[C@H]2OC[C@@H](OC)[C@H]2OC1)C	17.6	7.1	7.5
(±)-Limonene	CC1=CCC(CC1)C(=C)C	17.2	1.8	4.3
Methanol	OC[H][H][H]	14.7	12.3	22.3
Methyl Oleate	CCCCCCCC\C = C/CCCCCCCC(OC)=O	16.2	3.8	4.5
N-Methylpyrrolidone	CN1C(CCC1)=O	18	12.3	7.2
dichloromethane	ClCCl	17	7.3	7.1
Vinyl trifluoroacetate	C=COC(=O)C(F)(F)F	13.9	4.3	7.6
o-Vinyltoluene	CC1=CC=CC=C1C=C	18.6	1	3.8
Vinylsilicon	C = C[Si]	15.5	2.6	4
Vinyl S-ethyl mercapto ethyl ether	C=CSCCOCC	16.4	7	6
4-Vinylpyridine	C=CC1=CC=NC=C1	18.1	7.2	6.8
Vanillin	OC1=CC=C(C[H]=O)C = C1OC	19.4	9.8	11.2
Valeronitrile	CCCCC#N	15.3	11	4.8
L-(−)-Tyrosine	N[C@@H](CC1=CC=C(O)C = C1)C(O)=O	17.5	6.9	17.2
1-[2-(2-Methoxy-1-methylethoxy)-1-methylethoxy]-2-propanol	CC(COC(COC(COC)C)C)O	15.3	5.5	10.4
Phosphoric acid, triphenyl ester	O=[P](OC2=CC=CC=C2)(OC3=CC=CC=C3)OC1=CC=CC=C1	20.1	6.4	6.8
Trioctyl phosphate	CCCCCCCCOP(=O)(OCCCCCCCC)OCCCCCCCC	16.2	5.9	4.2
2,4,6-Trinitrotoluene	CC1=C(C = C(C = C1[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[O-]	19.5	10	4.5

valorization and carbon capture. The dataset was curated from the literature and *HSPiP* software, incorporating a wide range of chemical classes and solvent types to ensure diversity and relevance to the target applications of this study, particularly solvent compatibility assessments, which are crucial in lignin valorization and carbon capture. As shown in the simple data analysis performed on this constructed dataset, it includes heterocyclic compounds, amines, alcohols, and other chemical classes representing both polar and non-polar solvents, essential for capturing the full spectrum of solubility interactions. Furthermore, compounds were selected based on the availability and completeness of data, ensuring that each entry included corresponding dispersion, polarity, and hydrogen bonding solubility parameters, as well as a valid SMILES code.

3.2. SMILES-HSP data preprocessing

Preprocessing and preparation of the SMILES-HSP data were done to ensure a high prediction accuracy. In this study, the raw molecular representations used are the SMILES codes of different solvents in the dataset. Initially, the SMILES codes were cleaned by removing any empty or missing values. They were then validated using *RDKit*, a computational tool commonly used by experts, to filter out invalid codes from the raw dataset (Bento et al., 2020). *RDKit* was utilized through *Python* programing language for this essential task. After eliminating the invalid codes, the remaining valid ones were compiled into a refined dataset for further processing. Additionally, *RDKit* tool was used to extract more representative features, including molecular descriptors and Morgan fingerprints. These descriptors and fingerprints were then combined with the Hansen solubility parameters into a single dataset, where the features serve as the input values and the parameters as the output values. It was revealed that there are too many descriptors and using all of them or the ones having low significance can hinder the accuracy of the modeling afterwards. Accordingly, a features selection step is essential. Where, the biclustering data mining approach, particularly the non-negative matrix factorization algorithm (NNMF), was used to select the most significant descriptors. The NNMF requires determination of the number of components as an essential hyper-parameter. Thus, other clustering techniques, namely UMAP and tSNE, were employed to help in choosing the best number of components through data visualization. After features selection, the decoded SMILES (most significant descriptors and fingerprints) along with the Hansen solubility parameters were randomly divided into two sets, one for training and the other one for validation where their percentages are 85% and 15%, respectively. These percentages were selected upon trial and error until they reached satisfactory accuracy in both cases of training and testing. They are between the 80/20 and 90/10 common split ratios as the proposed ensemble method contains simple and complex models at the same time where the dataset is relatively large. 85% of the data was also used for the training to give the chance to the models to capture the true distribution of the available data.

3.3. ML modeling and decision fusion

The results of the optimized ensemble-based technique were gathered, evaluated, and benchmarked for the seven individual ML techniques selected and employed to predict HSP based on the key features extracted from the raw SMILES codes. In addition, all the fusion techniques were applied and compared to each other as well. Details of ML modeling and building the ensemble-based model are presented in Sections 2.4.2 **and** 2.4.4.

3.4. Models explainability in selected cases

The SHAP values as an informative XAI item were calculated for the developed ML ensemble-based regression model for decision fusion that calculates the different solubility parameters. The impacts of each descriptor on each solubility parameter were calculated in detail. In this case, the problem is a regression one where both input and output variables of interest are continuous. In addition, three other distinct cases were selected to show the importance of adding an XAI part to the developed high prediction accuracy ensemble-based model for decision fusion. These cases of materials that need to be dissolved are softwood-based kraft lignin, sugar cane bagasse-based lignin and CO_2. For validation purposes and to demonstrate the generalizability of the developed white-box methodology across various materials and systems, CO_2 was considered alongside different lignins. As a crucial material, selecting the appropriate solvents for CO_2 can significantly accelerate climate change mitigation efforts. After the calculation of each relative energy difference (RED) for each case, the solvents were classified as good and bad solvents taking the categorical labels of 1 and 0, respectively. The classification was based on the *Hansen* sphere concept; where, solvents scoring an RED equal to or higher than 1 are categorized as bad solvents, and the solvents with REDs lower than 1 are categorized as good solvents. The solubility parameters of the materials related to the three cases are summarized in Table 4. RED is calculated according to Eq. (8) (Duval et al., 2016).

$$RED = R_a/R_o \tag{8}$$

Where R_o is the experimental sphere radius of the material to be dissolved (m) and R_a is the radius of interaction between solvents

Table 4
Hansen solubility parameters of selected materials.

Material (Solute)	δ_D	δ_P	δ_H	R_o	Reference
Softwood-based kraft lignin	20.93	16.98	15.71	15	(Ajao et al., 2021)
Sugar Cane bagasse-based lignin	21.42	8.57	21.80	13.56	(Novo and Curvelo, 2019)
CO_2	15.6	5.2	5.8	4	(Williams, 2007), (Williams et al., 2004)

and the material to be dissolved. Ra is calculated according to Eq. (9) (Ruwoldt et al., 2022).

$$R_a = \sqrt{4 \times \left(\delta_D^m - \delta_D^{solvent}\right)^2 + \left(\delta_P^m - \delta_P^{solvent}\right)^2 + \left(\delta_H^m - \delta_H^{solvent}\right)^2} \quad (9)$$

It is worth noting that these two lignins have a wide range of applications. Our study focuses on the solvents' properties that make them suitable for dissolving these key materials, facilitating their efficient valorization in the future. Below are some specific applications that can be covered by these lignins:

- **Compatibility and blending with polymers:** Precise prediction of *Hansen* solubility parameters (HSP) enables the accurate selection of solvents for blending lignin with polymers, which can then be used to create composite materials.
- **Lignin valorization through fractionation:** Understanding the solubility behavior and the key solvent properties aids in the efficient fractionation of lignin into its simpler components/units. This will facilitate their utilization as precursors to other value-added products, including carbon fibers and bio-based resins as well as other intermediates or specialty chemicals.
- **Carbon capture and other environmental applications:** Solubility studies of sugarcane bagasse lignin (and even softwood-based lignin), can support its subsequent utilization in the synthesis of bio-based sorbents for carbon capture purposes, aligning with global Net-Zero emission goals. Identifying the most efficient solvents allows for the extraction of high purity lignins, which can be used to produce activated carbon-based adsorbents with tailored and desired capture capacities and selectivities (Zhao et al., 2021), (Barker-Rothschild et al., 2024). These lignins can also be utilized in other applications, such as the development of effective adsorbents for water and wastewater treatment (Al-Sakkari et al., Mar. 2020; Supanchaiyamat et al., 2019; Carrott and Carrott, 2007; Fu et al., 2013) or as catalysts in biofuels production (Naeem et al., 2021; Naeem et al., 2023; Dhawane et al., 2019).

3.5. Further case to validate model generalizability

To ensure that we built a generalizable ensemble model that avoids overfitting, it was further validated using a separate dataset. This new dataset contains over 1200 compounds with different distribution than the big dataset used for training. It includes experimental HSP data of these solvents, collected from various sources in the literature (Jeong et al., 2020; Subrahmanyam et al., 2015; Li et al., 2022). It is worth noting that cross validation was also considered during the training phase to overcome any kind of overfitting besides the dependence on models that avoid this problem from the beginning such as the Random Forest and XGB as mentioned in Section 2.4.2. From 5 to 10 *k*-fold cross validation were considered and tested in the current study. The *k*-fold ($k = 5$) was chosen and employed to avoid high consumption of time and computational resources.

4. Results and discussion

This section presents the results obtained by applying the developed methodology on the HSP-SMILES dataset to achieve high quality predictions of the different solubility parameters. It gives an overview about the distribution of the types of the solvents in this dataset where the most abundant solvents in the available repository are the heterocyclic ones (>3500). It also presents the results of clustering techniques and biclustering through NNMF to identify the key descriptors where they were reduced from 208 to 70. Afterwards, the results of ML modeling, through individual and ensemble models/techniques, are introduced and compared. Based on the developed ensemble modeling, the average prediction accuracy reached 99%.

4.1. Data analysis of HSP-SMILES dataset

Firstly, *RDKit* was used to validate the SMILES codes in the available dataset to clean data. As a result, about 170 invalid codes/data points were removed out of about 12,000 due to different errors, including syntax errors and other technical ones such as the atom bond kekulization (Hähnke et al., 2018). After obtaining the valid SMILES codes dataset, simple descriptive statistical analysis was done to determine the distribution of the available molecules of solvents/compounds. Fig. 6 presents the distribution of the different types of solvents included in the available data based on their SMILES codes. As can be observed, the most frequent type of solvent is the heterocyclic one with more than 3500 solvents. The second most common type is the amines followed by alcohol at the numbers of around 1800 and 1200 solvents, respectively.

Fig. 7 depicts the distribution of data for each of the three Hansen solubility parameters within the parent dataset, illustrating the frequency intervals for each parameter. Additionally, Table 5 provides a straightforward statistical summary of each variable in the HSP dataset, including the maximum, minimum, and average values of each parameter. The data for the dispersion solubility parameter appears to follow a nearly normal distribution, whereas the polarization and hydrogen bonding parameters exhibit noticeable skewness to the left.

On the other hand, *RDKit* was used to extract different descriptors based on the available SMILES codes. 208 descriptors in total were extracted to represent the SMILES code of each compound. However, as a first step, 18 common descriptors were chosen besides the *Morgan* fingerprints to represent each compound. These descriptors are molecular weight (MolWt), topological polar surface area (TPSA), number of heavy atoms (HeavyAtomCount), heavy atom molecular weight (HAMWt), molecular logarithm of the partition coefficient (LogP), number of rotatable bonds (NRB), number of hydrogen donors (NHD), number of hydrogen acceptor (NHA), *Hall-Kier* alpha shape index (HKA), Chi squared n descriptor (Chi2n), *Labute's* approximate surface area (LASA), number of radical electrons (NRE), 2D autocorrelation descriptor capturing spatial patterns and relationships within the molecule (Autoc), number of valence electrons (NVE), fingerprint density descriptor based on the *Morgan* algorithm with radius of 2 (FDFDM2), exact molecular weight (ExMwt), maximum absolute partial charges on the atoms in the molecule (MaxAPC) and minimum absolute partial charges on the atoms in the molecule (MinAPC). As previously mentioned, the most significant variables/descriptors will be selected to represent each compound. Hence, the correlation between each descriptor based on the *Pearson's* method was performed and presented in Fig. 8 to prove visually the need for reducing the number of descriptors to the most important and uncorrelated ones. This is a preliminary step before performing the biclustering for rigorous variables selection.

The numbers mentioned above correspond to different case studies or scenarios presented in the manuscript. In the first instance, the 18 components were selected as part of an initial example to demonstrate the effectiveness of NNMF and our combined feature selection approach in identifying impactful features or variables influencing the prediction

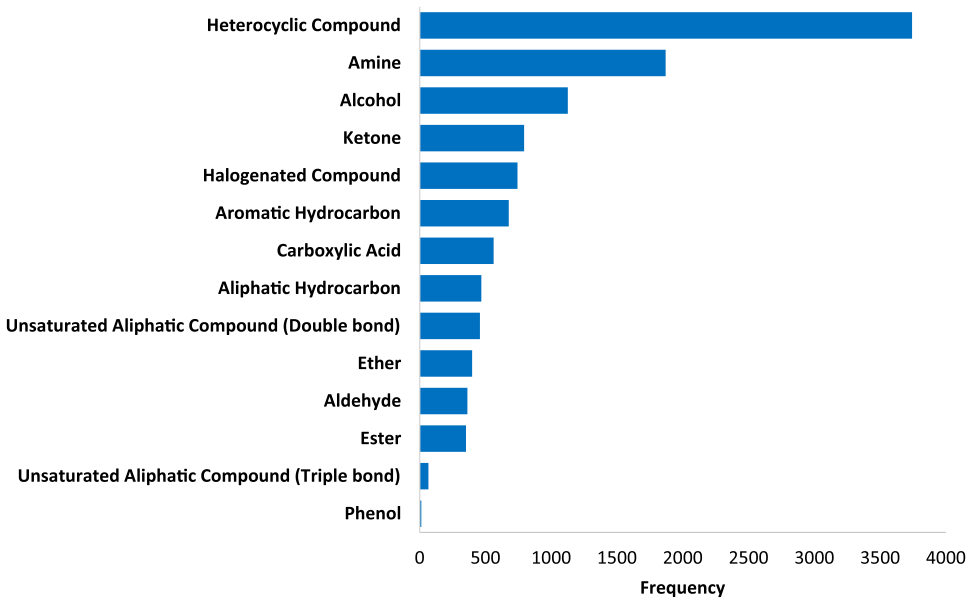

Fig. 6. Solvent type distribution.

of the desired outputs. These 18 variables, chosen randomly from a larger pool, were all highly relevant. However, to provide a more comprehensive evaluation, we extended the analysis to consider all available descriptors. In this extended example, the data was optimally grouped into three components, with each component representing a cluster of significant descriptors. Thus, for the initial case study, 18 components best described the data, while for the extended case study, three components offered a more balanced and efficient representation of the dataset. The final number of components selected for the study is three, as it provides an optimal trade-off between complexity and performance.

As it is shown in Fig. 8, there are some highly correlated variables/descriptors such as the high positive correlation of heavy atom molecular weight (HAMwt) with both exact molecular weight (ExMwt) and molecular weight (MolWt). The quality of this correlation is very close to the unit, which means that only one descriptor can be used. This will lead to dimensionality reduction that facilitates computational work. To achieve this goal, a robust feature selection method should be employed and that is why the combined clustering/bi-clustering methodology is proposed to select the most significant descriptors.

4.2. Clustering results

The clustering methods were employed to give a preliminary recommendation about the best number of components graphically. This number is essential for performing the NNMF analysis to complete the features selection. Firstly, the best number of data clusters was determined for the solvents' available data according to silhouette test which determines the degree of data clusters separation. The clustering was performed based on *k*-means clustering method. In the current study, the optimal number of data clusters is 2 (Fig. 9(a)) where the separation of the clusters is significant and visible as indicated by the plot on the right side of the figure. In this case, the silhouette score is 0.2. Dividing the data into more than 2 clusters results in lower silhouette score values which means having inseparable clusters. This is further confirmed by the plot on the right side of Fig. 9(b). This result was consistent for both cases considered in this study: the first involving the randomly selected 18 common descriptors, and the second including all 208 physio-chemical descriptors.

After choosing the best number of clusters, the work was oriented to the determination of the best number of components (the input variables of the data after performing the dimensionality reduction). Both UMAP and tSNE combined with *k*-means were employed for this task. It should be noted that in this study, the selected number of components is used to guide the NNMF for subsequent features selection. This approach helps mitigate the issue of interpretability, as the NNMF determines the actual number of important variables, and the variables used as inputs afterwards are the most significant original variables/features extracted from the SMILES codes. To further reduce the computational complexity of NNMF, both UMAP and tSNE were incorporated into our novel methodology. UMAP outperformed tSNE in terms of both the computational time and graphical representation. In particular, the UMAP completed the dimensionality reduction in only 5 s while tSNE completed it in 112 s.

In the first case study, where the 18 common descriptors were randomly chosen to represent the available data, the best number of components that best represents the data is 18 according to silhouette score test and the graphical representations obtained by UMAP (Fig. 10 (a & c)). The 3D graph (Fig. 10 (c)) was added for more clarification of the obvious separation of the two clusters upon selecting the optimum number of components, i.e. 18. In the second case, upon extracting all the possible descriptors of each SMILES code using *RDKit*, the best number of components that the data can be reduced to was surprisingly found as 3. The graphical representation of 2D UMAP clustering in the case of extracting all the descriptors is provided by Fig. 10 (b). This step of extracting all the descriptors led to having more representative data. In conclusion, based on this advisory step, the data can be clustered into two main clusters, and its variables can be reduced to three primary components. These results were then used as inputs to NNMF, enabling robust features/variables selection without sacrificing the interpretability or the integrity of the original data.

4.3. Bi-clustering (NNMF) results

In the first case study, the best number of clustering components are 18 according to EV and RE tests in agreement with the results obtained in the proposed graphical method in the previous section. Where the EV and RE values were found as > 0.99 and 2.28E-5, respectively. EV and RE tests were conducted to validate the results obtained from the initial clustering step, which guided the NNMF modeling and features/variables selection process.

The relative importance of each descriptor in representing the data is

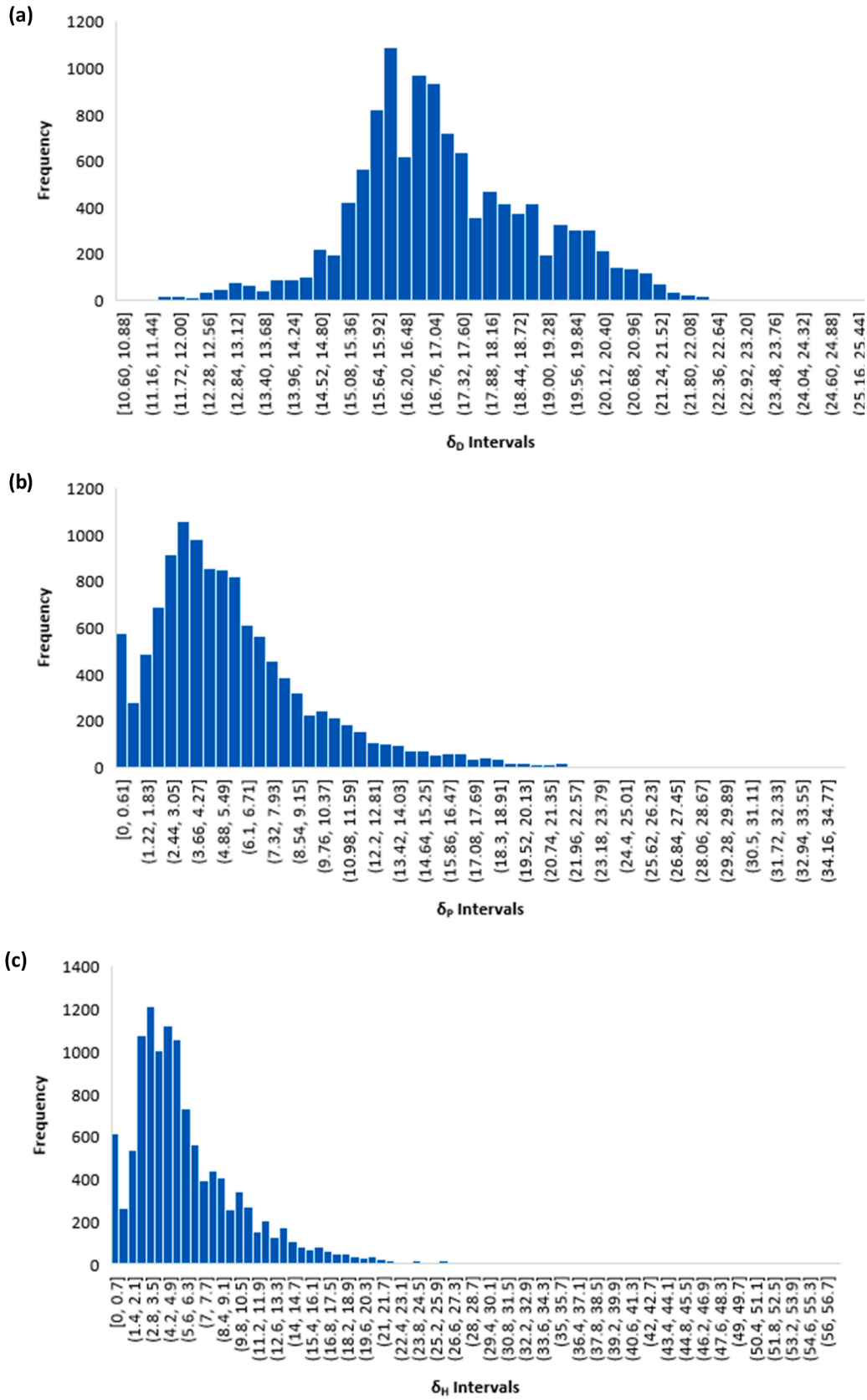

Fig. 7. Hansen solubility parameters (a) dispersion, (b) polarization and (c) hydrogen bonding distributions.

Table 5
Simple statistical description of *Hansen* solubility parameters data.

Statistical parameter	Variables		
	δ_D [MPa$^{0.5}$]	δ_P [MPa$^{0.5}$]	δ_H [MPa$^{0.5}$]
Maximum	25.3	35.2	57
Minimum	10.6	0	0
Average	17.14	5.69	6.05
Median	16.9	4.8	4.8
Standard Deviation	1.80	3.98	4.54
Mode	16	0.1	0.1
Skewness	0.26	1.49	1.83
Kurtosis	0.20	3.68	6.18

determined according to its score in each component. These scores are illustrated in the small squares (cells) depicted in the plot on the left side of Fig. 11. In particular, the higher the score the higher the relative significance of the descriptor in a specific data component. According to Fig. 11, the most significant descriptor for the first component is the number of heavy atoms (HeavyAtomCount) with a score of 6.74 followed by number of valence electrons (NVE) with the score of 6.64. The expert has an additional essential role in the interpretation of the NNMF results and the determination of the best number of significant descriptors per each component. For instance, only the highest score descriptor can be chosen or in other cases the most significant descriptors can be chosen as the best two descriptors with the highest scores. Trial and error can also assist the expert in determining this number. Through this trial-and-error analysis and due to the random selection of these 18 descriptors in the first case, all of them were found to be significant. This prompted us to extend our investigation to include all the possible physio-chemical descriptors, allowing us to select the most significant ones in a more robust manner.

In the case of extracting all the 208 descriptors, the best number of components was also 3 as the obtained preliminary graphical recommendation from the clustering method (UMAP). The EV and RE values at this number of components were > 0.99 and 0.0006, respectively. The figure of NNMF results in this case is provided in the supplementary files due to its large size. By trial-and-error approach, among the 208 descriptors and the 2048 bit-strings fingerprints, it was found that the first 25 descriptors per component with the highest scores are the most important ones to represent the available data. This led to having 70 unique descriptors due to the presence of 5 replicated descriptors that are significant in more than single components. The list of these 70 most significant descriptors is provided in the supplementary materials. According to these two case studies it can be observed that the use of graphical approach for determining the number of components is comparable with the quantitative analysis approach based on EV and RE. This also reinforces our intuition of using the initial clustering to guide the NNMF afterwards with the optimal number of data components, helping to mitigate its computational complexity, particularly with larger datasets.

4.4. Comparison of different ML modeling methods

As previously mentioned, the hyperparameters of each model were optimized to obtain the highest possible accuracy. In this regard, the optimum SVR model was obtained at the hyperparameters of C = 300 and a kernel type of radial basis function (RBF); while, the optimum number of estimators, i.e., trees, of the optimized XGB model was 200. In the case of ANN model, the optimum hyperparameters are learning rate of 0.001, *ReLU* activation function, 3 hidden layers and a batch size of

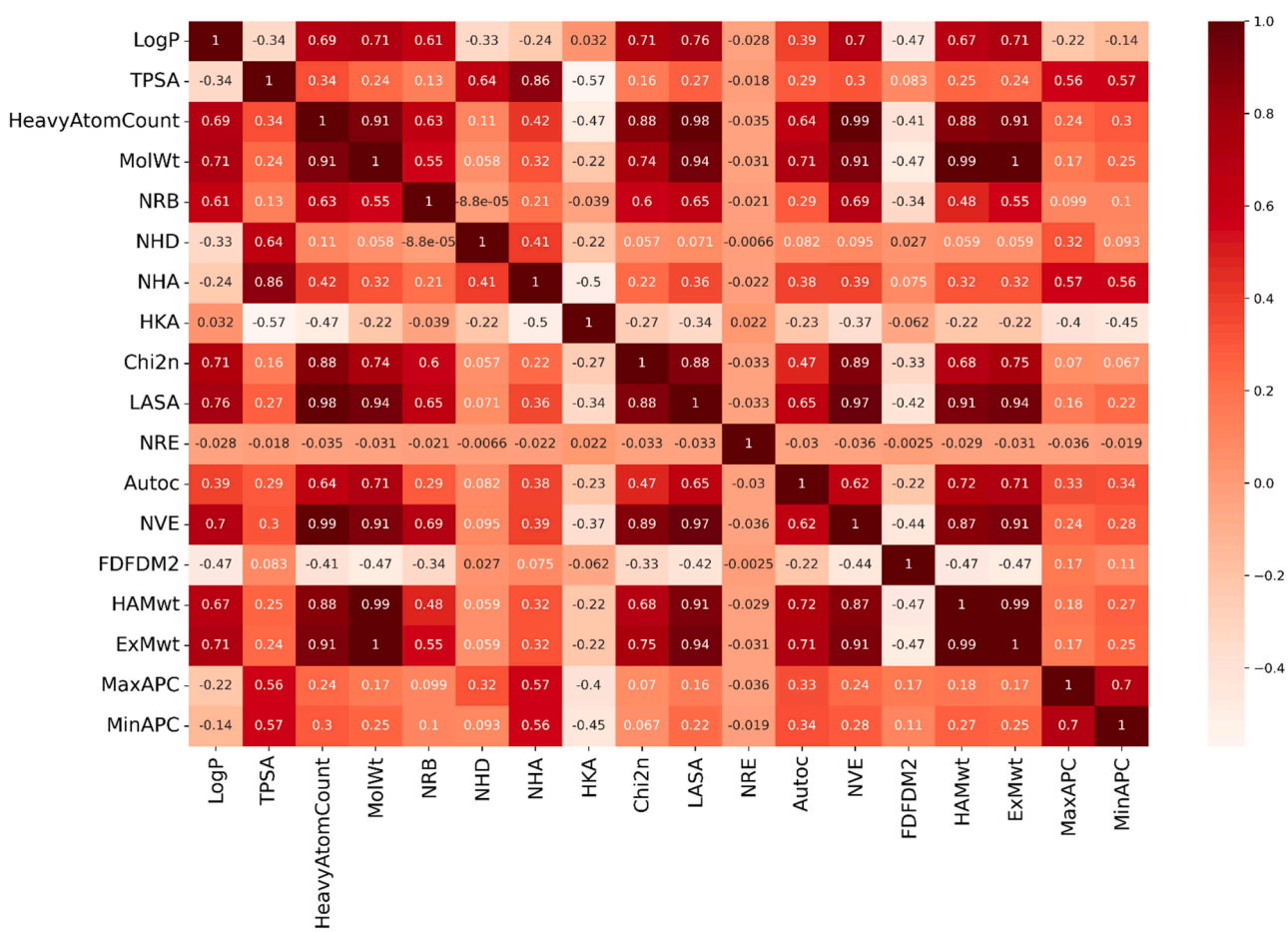

Fig. 8. Descriptors correlation matrix.

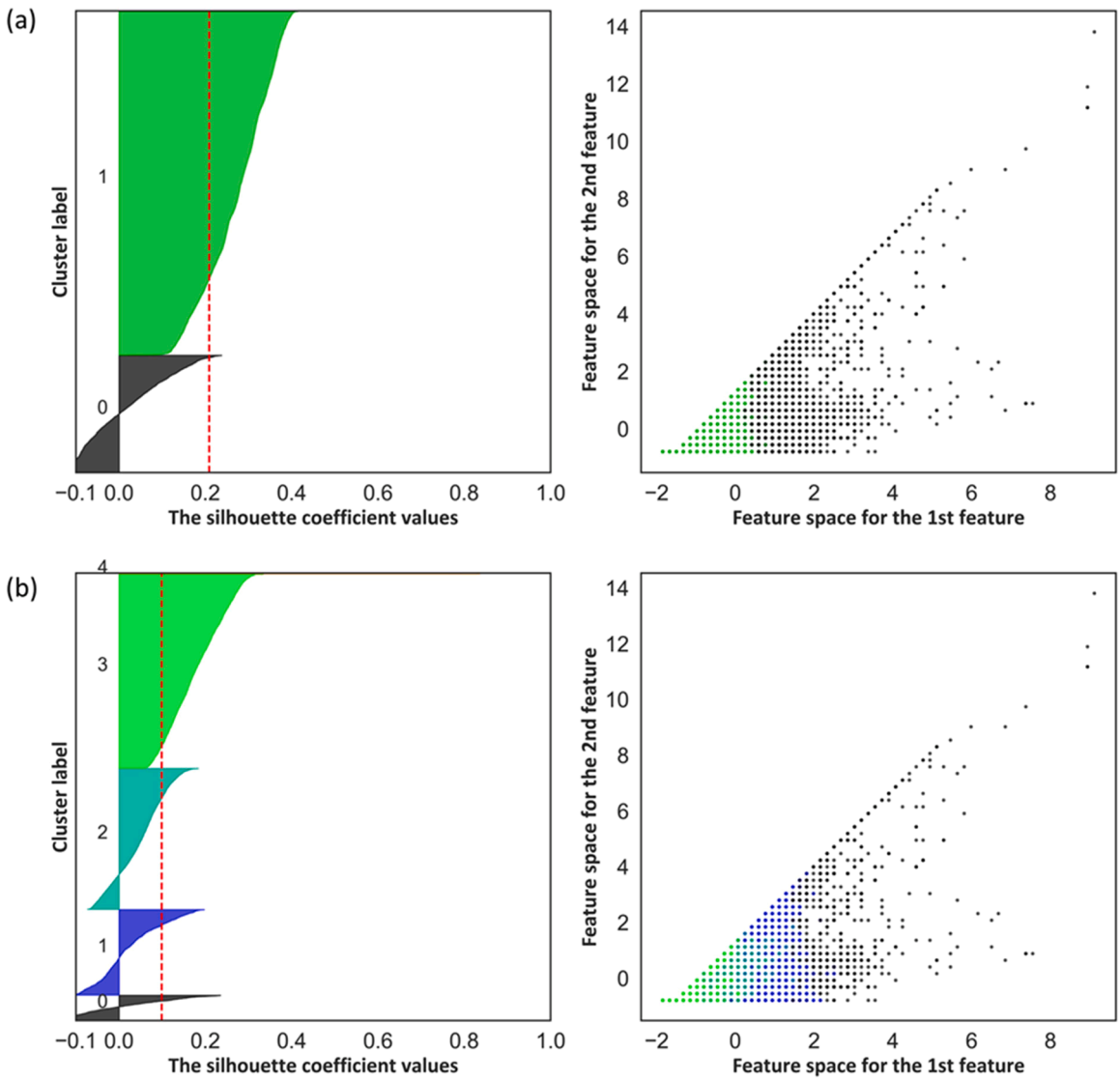

Fig. 9. Silhouette analysis plots for (a) the optimum number of clusters ($n_{descriptors}$ = 208 & $n_{clusters}$=2) (b) example of inseparable clusters ($n_{descriptors}$ = 208 & $n_{clusters}$=5).

16. The number of neurons of the first, second and third layers are 150, 100 and 50, respectively. Whereas the optimum RF model had 100 trees. In addition, the optimum number of neighbours (k) for *k*-NN model was 3. The optimum max depth of the optimized decision trees is 16. The optimum LSTM model contains one LSTM layer with 32 units and a dense layer with a single unit for the output. The optimum number of epochs and batch size were 200 and 16, respectively, where the "*Adam*" optimizer was adopted. It should be noted that both CNN and GPR techniques were also used. However, they gave lower accuracies, i.e. < 80%, and took relatively longer training times. This is due to the relatively large dataset used, which increases the computational complexity especially in the case of GPR modeling. Hence, their results were not included.

Table 6 summarizes the results obtained after performing the ML modeling using the optimized models. As demonstrated, SVR and XGB models have the highest accuracy and mean square error. This observation is extended to the cases of all three solubility parameters. The third best model was RF, as shown in Table 4, and this is due to the ability of RF to handle tabular numerical data. Being an ensemble-based technique, this enabled RF to outperform the optimized decision tree as illustrated in the table as well. Another important observation is that the LSTM model was a bit better than the ANN one. This is presumably due to the additional ability in its layers to have the seize sequential data and sequences in the available data.

After obtaining all the results from different models, they are fused to enhance the prediction accuracy. As previously mentioned, the fusion methods developed are categorized into two main categories, i.e., learnable, and non-learnable. The results of the fusion attempts are presented in Table 7. In the case of non-learnable fusion, both techniques based on MSE and combined R^2/MSE weighted averaging were better than those considering only simple averaging and R^2-based weighted averaging. In the case of learnable fusion, in general, it possessed higher accuracies; however, ANN did not show high accuracy such as XGB and SVR in this stage. This is the reason why it was not considered in the second and third fusion steps. As evident from the data of Table 7, the non-learnable decision based on simple and weighted averaging was inspired by the bagging mechanism. While the series learnable decision fusion was inspired by the boosting mechanism to minimize both variance and bias.

In general, SVR is the best choice when the number of features is near to or higher than the number of observations. In the current case, when fusion goes on to the second and third stages, the ratio between number of observations and variables decreases. This made SVR the best candidate for completing the third learnable fusion and the results in

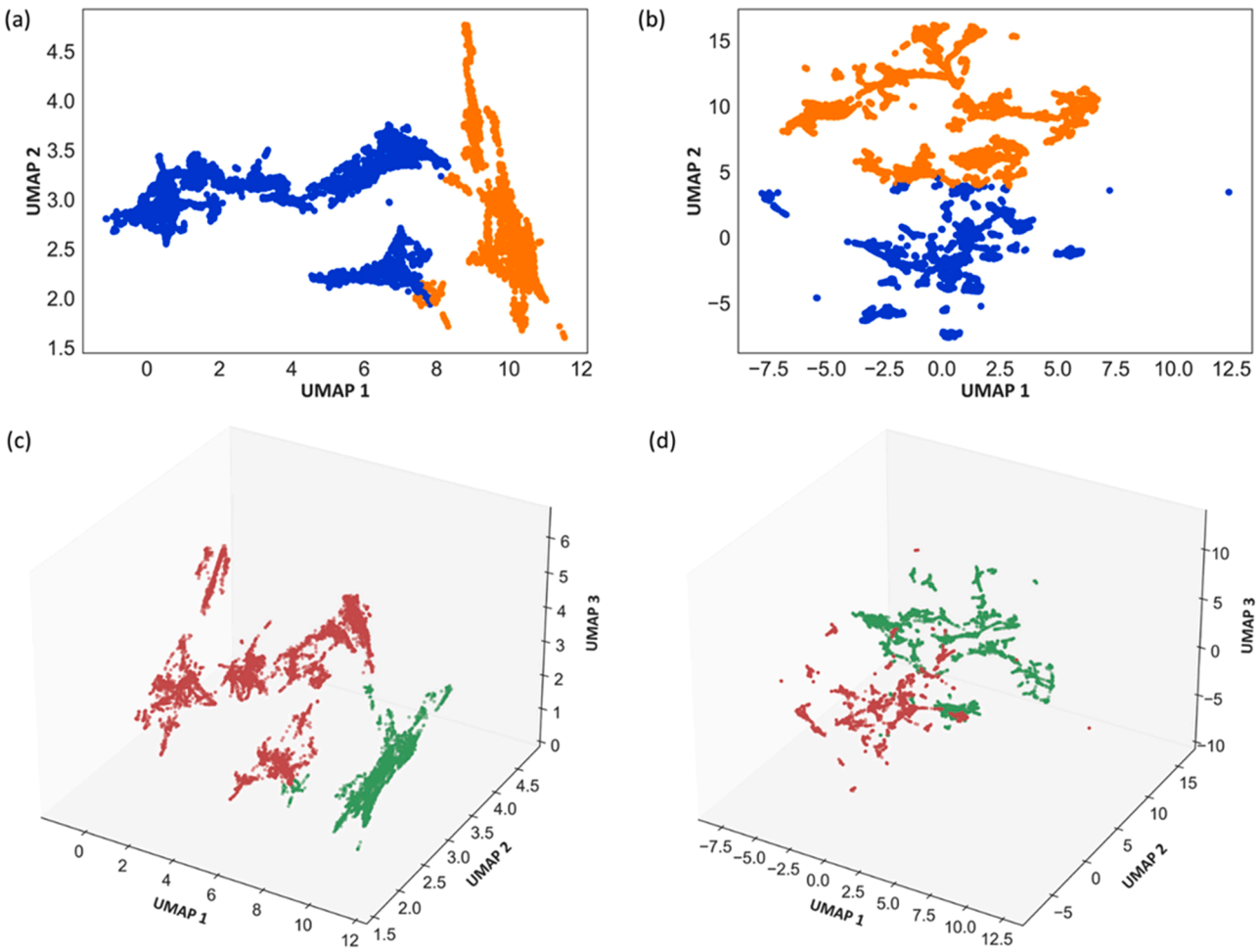

Fig. 10. Graphical representations of data using UMAP dimensionality reduction method for (a) ($n_{descriptors} = 18$ & $n_{components} = 18$) and (b) ($n_{descriptors} = 208$ & $n_{components} = 3$); (c) and (d) are the 3D graphs for both cases, respectively.

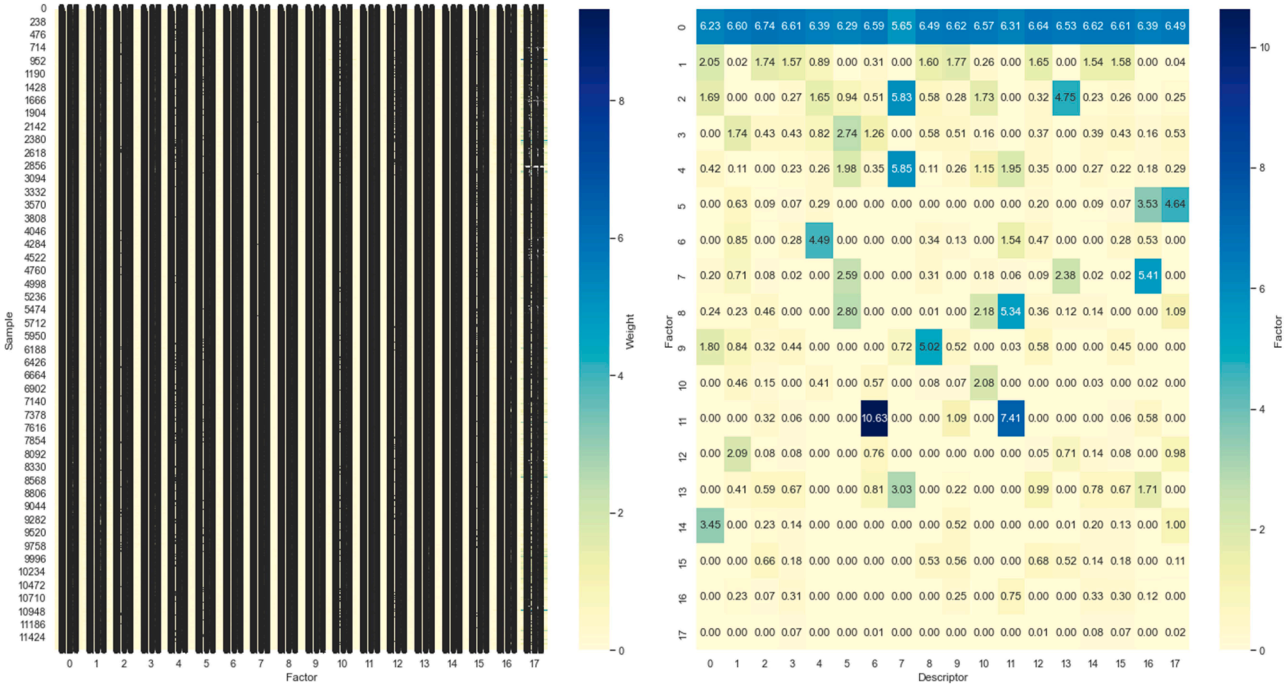

Fig. 11. NNMF results for ($n_{descriptors} = 18$ & $n_{components} = 18$).

Table 6
ML modeling results.

Variable	*k*-NN		DT		RF		XGB		SVR		ANN		LSTM	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Solubility1 (δ_D)	0.92	0.26	0.92	0.28	0.95	0.16	0.96	0.15	0.96	0.14	0.92	0.26	0.94	0.19
Solubility2 (δ_P)	0.89	1.72	0.81	2.96	0.91	1.37	0.93	1.06	0.93	1.11	0.90	1.49	0.91	1.44
Solubility3 (δ_H)	0.90	2.17	0.89	2.22	0.95	1.14	0.95	0.96	0.96	0.76	0.93	1.40	0.93	1.49

Table 7
Decision fusion results (Solubility_1 "δ_D").

Fusion No.	Non-learnable Fusion								Learnable Fusion					
	Average-based		R^2-based		MSE-based		R^2/MSE-based		XGB		ANN		SVR	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Fusion 1	0.97	0.12	0.97	0.12	0.97	0.11	0.97	0.11	0.97	0.12	0.96	0.14	0.97	0.12
Fusion 2	0.97	0.11	0.97	0.11	0.97	0.11	0.97	0.11	0.97	0.10	——	——	0.98	0.06
Fusion 3	0.98	0.05	0.98	0.05	0.98	0.04	0.99	0.03	——	——	——	——	0.99	0.02

* All the values are rounded to the second decimal.

Table 6 are in good agreement with this conclusion.

Figs. 12-14 were designed to visualize the comparison among the different methods. From these figures it can be observed that the decision fusion has a very significant effect on reducing the errors of individual techniques. The proposed decision fusion strategy/methodology combines the advantages of each individual regression technique which is excellent prediction model at certain regions. As a result, the decision fusion-based regression model avoids the high error of some techniques at the regions of their poor performance and follows the high predictability of the models with excellent performance at these regions. This is the reason why different shallow, deep, parametric, and non-parametric machine learning models were selected in the current study. The remaining results related to the other two solubility parameters are provided in the supplementary materials for more information. Fig. 12 represents a sample of the data and the results. As shown, individual ML models are compared to each other and to our proposed ensemble-based methodology. For example, in Fig. 12 (b) some of these models were selected and added for the sake of comparison, demonstrating the capability of our developed fusion (ensemble) technique to adapt and deliver high performance in predicting the desired output, i.e. HSPs in this case. However, Fig. 13 depicts the whole experimental data available versus the predicted values of the dispersion solubility parameter after applying all the decision fusion steps. As shown, the developed model based on the proposed methodology possesses high prediction accuracy of 99.54% compared to almost 96% obtained by both individual SVR and XGB models.

LSTM was included alongside XGB and ANN for a comprehensive comparison and to demonstrate the versatility and robustness of our developed fusion technique. The inclusion of LSTM highlights the ensemble method's ability to integrate models with varying individual accuracies while still achieving high overall performance in predicting the desired outputs, such as HSPs in this study. This approach underscores the adaptability of the ensemble method in leveraging diverse model characteristics to enhance predictive accuracy.

4.5. Effect of dataset size on prediction accuracy

This section discusses the effect of training dataset size on the prediction accuracy of the ML models. Fig. 14 presents the illustration of this effect for the dispersion solubility as a relevant parameter. As shown, there is a general trend of accuracy increase and mean squared error decrease upon increasing the training dataset size. This is presumably due to giving more chances to the developed model to capture and learn all the possible combinations and data/variables distribution. However, it can be deduced that sizes above 70% to 90% are sufficient and give high prediction accuracy. This also gives an indication about the amount of data from the same distribution required to train the developed ML models. For instance, in the current case of predicting HSP based on SMILES codes, it is recommended to work on representative datasets containing more than 8500 data points to have prediction accuracies higher than 90%. These results are in good agreement with those stated in previous research studies (Ethier et al., 2022).

As previously mentioned, the results in Fig. 14 are related to the prediction of the dispersion solubility parameter. The same trends were observed in the case of the other two solubility parameters, i.e., polarity and hydrogen bonding solubility parameters. These results are documented in the supplementary materials.

4.6. Model explanation

The results after the addition of the XAI part are summarized in the following figures (Figs. 15-17). In these figures, the magnitude of significance of each descriptor is presented on the left-hand side, whereas the right one gives the information about the sign of this magnitude, i.e., positive, or negative. With respect to the sign, for a certain input variable, if there is a high concentration of red dots on the right side of the SHAP value scale (right of the zero-value locus), then this means that the input variable positively affects the output variable. The vice versa is true. A red dot means that the value of the output variable is high at this point and SHAP value. In the left-hand side of the graphs, the average value of the SHAP values for all the studied points for each variable is presented to indicate the global effect of the input variables on the output ones. The following are examples of the explanation of these graphs to breakdown the extracted knowledge from them as a powerful XAI technique to overcome the problems of black-box modeling. For instance, according to Fig. 15, the most significant factor/descriptor influencing the first solubility parameter related to the dispersion of energy is the number of aromatic rings in the molecule, whereby this impact is a positive one. Elsewise the most significant factor in the case of polarization solubility parameter is the molecular logarithm of the partition coefficient (MolLogP) and it has a negative impact (Fig. 16). The negative impact of MolLogP extends to both dispersion and hydrogen bonding solubility parameters. As seen, each parameter has its own set of significant descriptors. This means that all different relations should be considered. Accordingly, the designer should take care about any conflict that can lead to competing multi-objective optimization problems. This information will help the designer or process operator in selecting or producing the optimum solvent(s). Specifically, if operators aim to select a solvent or group of solvents that possess high dispersion solubility parameter, then, based on the findings of this explainability

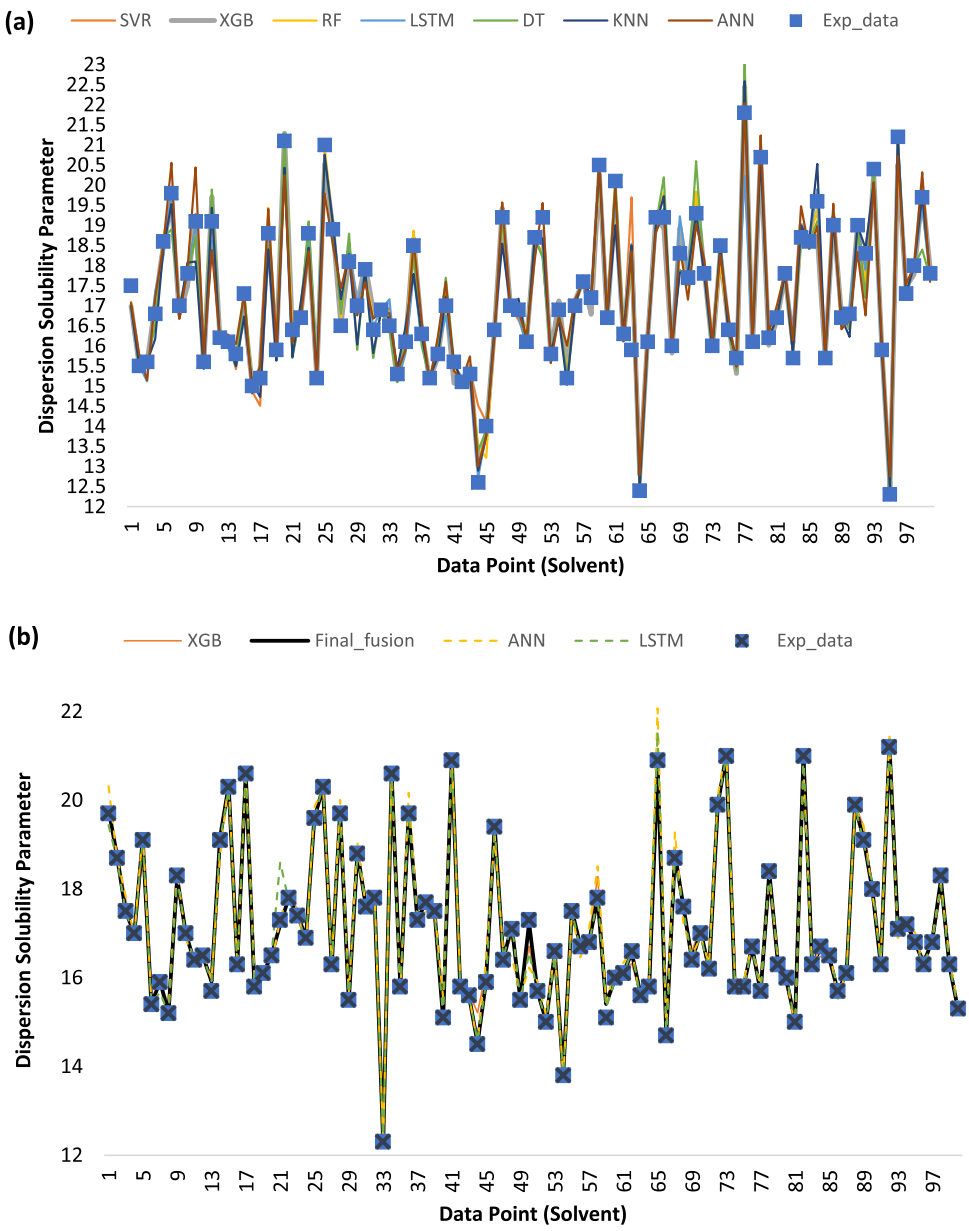

Fig. 12. Results of (a) different individual techniques, (b) final decision fusion vs. selected different individual techniques.

study, they should choose the solvents with higher number of aromatic rings in their structure. Besides, if it is desired to have a specific solvent with a lower hydrogen bond or polarization solubility parameter, then the operators should select the group of solvents with higher MolLogP values. These guidelines can also be followed in the case of designing new solvents tailored to a specific application.

As previously stated, other specific case studies related to the selection of suitable solvents based on the calculations of *Hansen* sphere are also selected. These cases include different lignin materials, i.e., kraft lignin from a softwood feedstock and sugar cane bagasse lignin, besides CO_2.

For softwood-based kraft lignin, the polarization energy parameter exerts the greatest influence, with all three parameters positively affecting lignin dissolution (Fig. 19). In this case, a robust interpretable machine learning (IML) model was developed to capture the key patterns that determine desired solvents based on their descriptors. The IML model was optimized to achieve the highest possible classification accuracy, with an average testing accuracy of 95%, a recall of 0.91, and an *F1*-score of 0.95. The training accuracy was nearly 100%. At the descriptor level (Fig. 18), TPSA emerges as the most significant factor positively influencing the dissolution of this type of lignin. Conversely, for bagasse-based lignin, the hydrogen bonding parameter is the most influential, followed by the positive effect of the dispersion energy parameter. This aligns well with Figure S11 in the supplementary materials, which highlights the number of hydrogen donors as a significant descriptor positively impacting bagasse-based lignin dissolution. These findings indicate that different lignin materials exhibit varied behaviors, necessitating more comprehensive future studies to explore the interactions between various solvents and the distinct functional groups and monomeric units of lignin materials. From these examples, it is evident that, generally, higher hydrogen bonding and dispersion solubility parameters are required for a solvent to be effective for lignin materials, regardless of their composition. Based on these findings, researchers, designers and process operators can narrow their search when selecting a group of optimum solvents for these lignins, focusing on those with higher hydrogen bonding and dispersion solubility parameters. In addition, drawing from the results of the first part of this explainability study, these solvents can be tailored to serve this

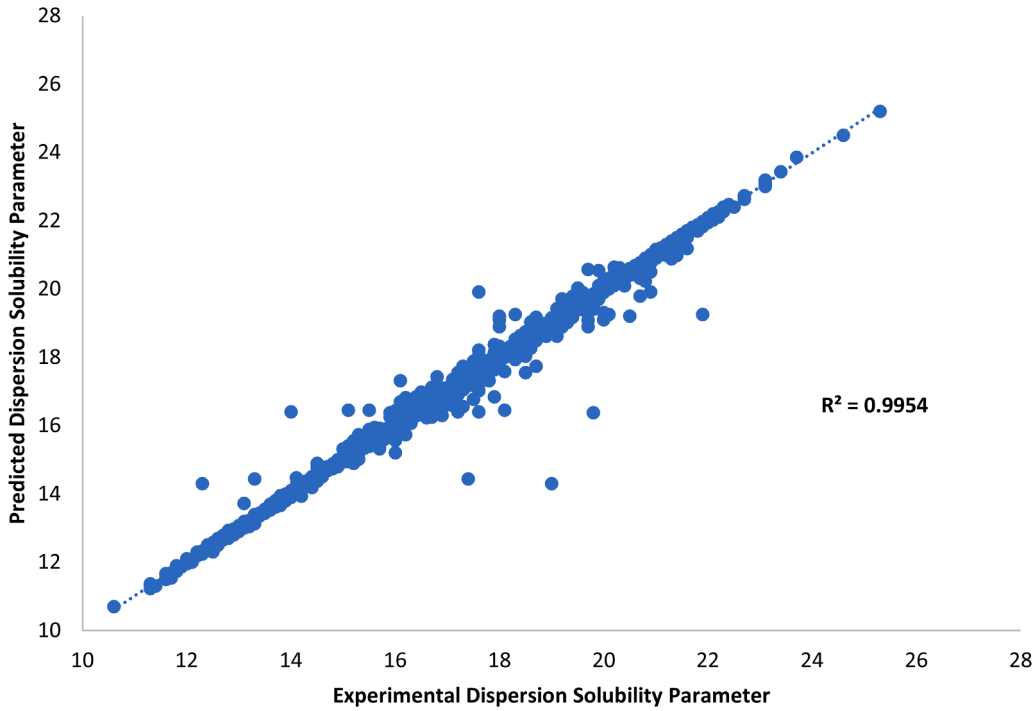

Fig. 13. Predicted Vs. experimental values of dispersion solubility parameter (final decision fusion results).

application through selecting/designing these solvents with lower MolLogP values and a higher number of aromatic rings.

According to **Fig. S12**, the 6th order electro-topological state of a fragment within a molecule (VSA_EState6) is the most significant feature that affects the solubility of CO_2 in a certain solvent, whereby this impact is mainly a negative one. The number of hetero atoms is, too, a very significant feature/descriptor compared to the others where this impact is also a negative one. This means that considering decreasing those features upon designing a new solvent will give a better performance to dissolve and capture CO_2. On the solubility parameters level, **Fig. S13** suggests that the dispersion solubility parameter/energy is the most significant one among the three parameters, whereby this impact is also negative, which means that selecting or designing a solvent having lower dispersion can be beneficial in the case of CO_2 dissolution/capture. Upon comparing these selected cases, the lignin problem/case is more complex and different compared to the CO_2 case study.

5. Model validation and generalization

The dataset employed in this validation case exhibits the chemical types of distribution illustrated in Fig. 20(a), based on the available SMILES codes. As shown, this distribution differs somewhat from the larger dataset used for training the developed ensemble model. For example, this dataset lacks phenols, and both ketones and carboxylic acids are significantly underrepresented compared to the training dataset. Additionally, the numerical distribution of the Hansen solubility parameters (HSPs) varies. Fig. 20(b) displays the frequency of dispersion parameter values (intervals) in the validation dataset, where the maximum value is 22 and the average is 16.9.

Regarding ensemble-based modeling, the developed model demonstrated its efficiency in predicting HSPs for the validation dataset, achieving high prediction accuracy across all three parameters. For instance, Fig. 20(c) compares the predicted and experimental values of the dispersion Hansen solubility parameter, showing a determination coefficient of 93% and a mean squared error below 0.24. These results confirm that the developed ensemble-modeling technique can be effectively generalized to other cases, enabling accurate determination of HSPs.

6. Comparison with literature, impacts and remarks

Recently, there is a considerable number of studies that investigate the application of AI/ML to predict various solubility parameters/representations including HSPs, molar equilibrium solubilities and *Hildebrand* solubility parameter (Nagulapati et al., 2022; Liu et al., 2024; Hsiao and Chang, 2024). **Table S4** (in supplementary materials) summarizes different selected attempts in the literature related to the AI-assisted solubility parameters prediction to present their performances and compare them with the current study. Some of these studies focus on relatively small datasets or specialized datasets with a very narrow and focused spectrum of chemical classes, which may limit the model's ability to generalize. These limited datasets do not capture the full spectrum of chemical classes or structures. In addition, as mentioned in the introduction section, most of these studies rely on black-box ML modeling, which hinders the understanding of the impact of various input features/variables on the target outputs (solubility parameters). For more details on the data types, number of observations and models used in each study, readers can refer to **Table S4** in the supplementary materials. According to the summary in the table, our methodology has several advantages such as high predictability and explainability. The use of larger dataset enhances the model generalization where larger true distribution is captured during training. The high predictability comes from the combination of several different ML techniques and architectures through various learnable and non-learnable fusion techniques as well. The dependence on the molecular structure and the corresponding molecular descriptors besides the use of large dataset of diverse solvents/materials ensure generalization and make the developed method suitable for a wide range of materials including ionic liquids and organic solvents. With respect to features selection, our methodology introduces a robust way depending on NNMF technique instead of relying on manual selection as most of the studies in literature. On the technological and managerial sides, the developed methodology and the accompanying tool for HSPs prediction will offer the following:

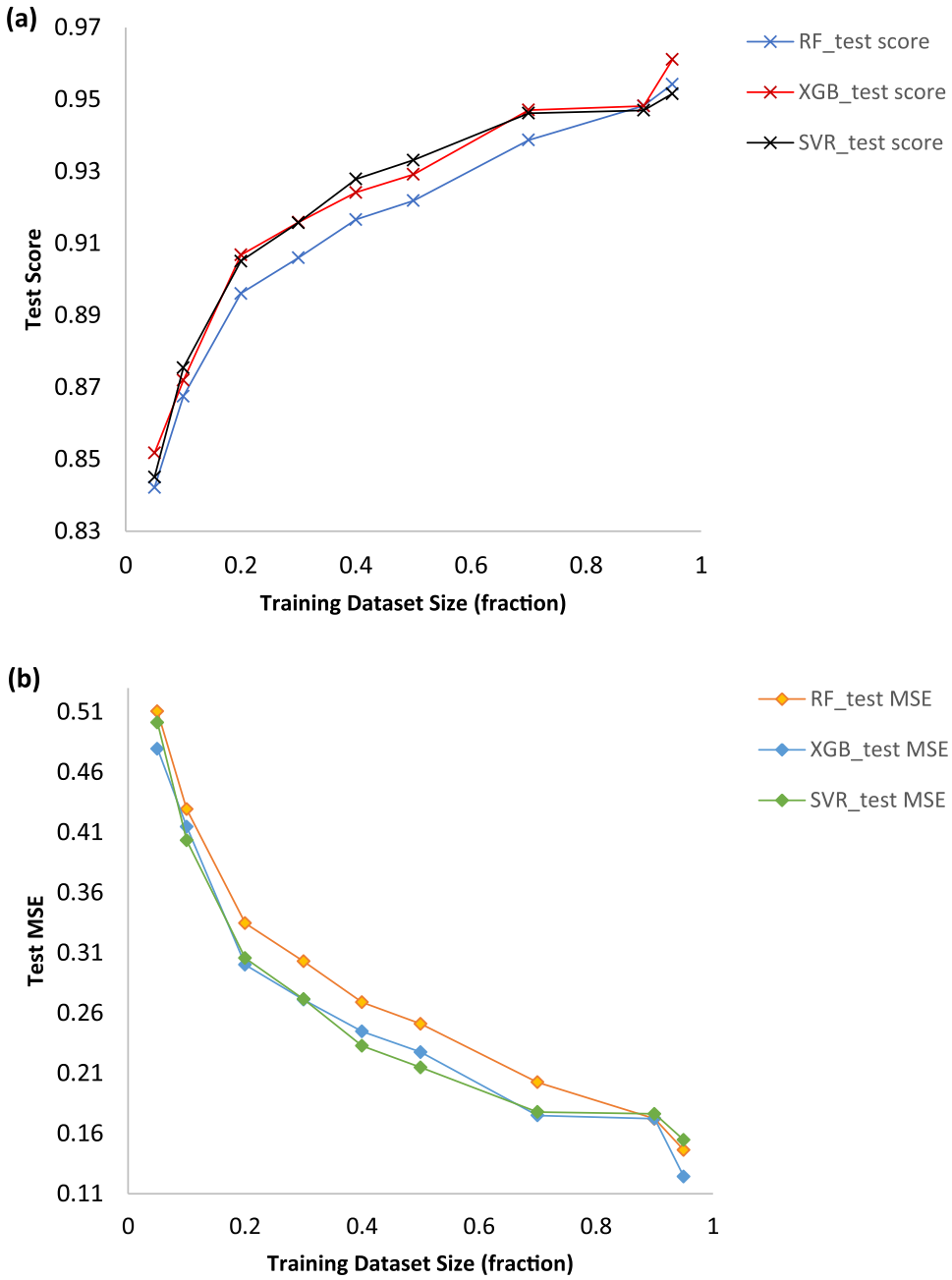

Fig. 14. Effect of dataset size on (a) the accuracy/score and (b) mean squared error of ML modeling using RF, XGB and SVR for Dispersion solubility parameter.

- Prediction of HSPs of new materials or solvents based on their molecular structure, represented by their SMILES codes, and the corresponding molecular descriptors with high accuracy in fractions of second. This will waive the burden of experimental procedures needed to evaluate these important parameters for new materials/solvents
- Streamlined assessment and screening of solvents based on their HSPs to be used for dissolving materials of interest such as various lignins and CO_2 which is a current hot real-world application related to *Net-Zero* strategies. This can lead to the suggestion of new solvents to perform specific environmental tasks that were not their primary field of application.
- Easy evaluation of compatibility between different materials, e.g. lignins and other polymers, during mixing based on the selection of the group of optimum (good) solvent for both polymeric materials.
- Accelerated decision making can be performed by different stakeholders based on the fast and accurate predictions offered by the developed methodology and tool to select the best solvent among a group of optimum solvents from a technical point of view which can be associated with economic one in future research.

It is worth noting that there are other studies that used HSPs as inputs to predict different important variables such as gelation properties (Delbecq et al., 2020), and polymers' cloud point (as an indicator of phase behavior of polymer solution system) (Ethier et al., 2022). Another application is the prediction of power conversion efficiency of solar cells system (specifically non-halogenated green solvent-processed organic solar cells) (Lee, 2023). Accordingly, our developed methodology and tool can play a key role in future research in these fields through providing accurate predictions of HSPs of new materials. These predicted HSPs can be then fed as inputs to have a reasonable estimation of

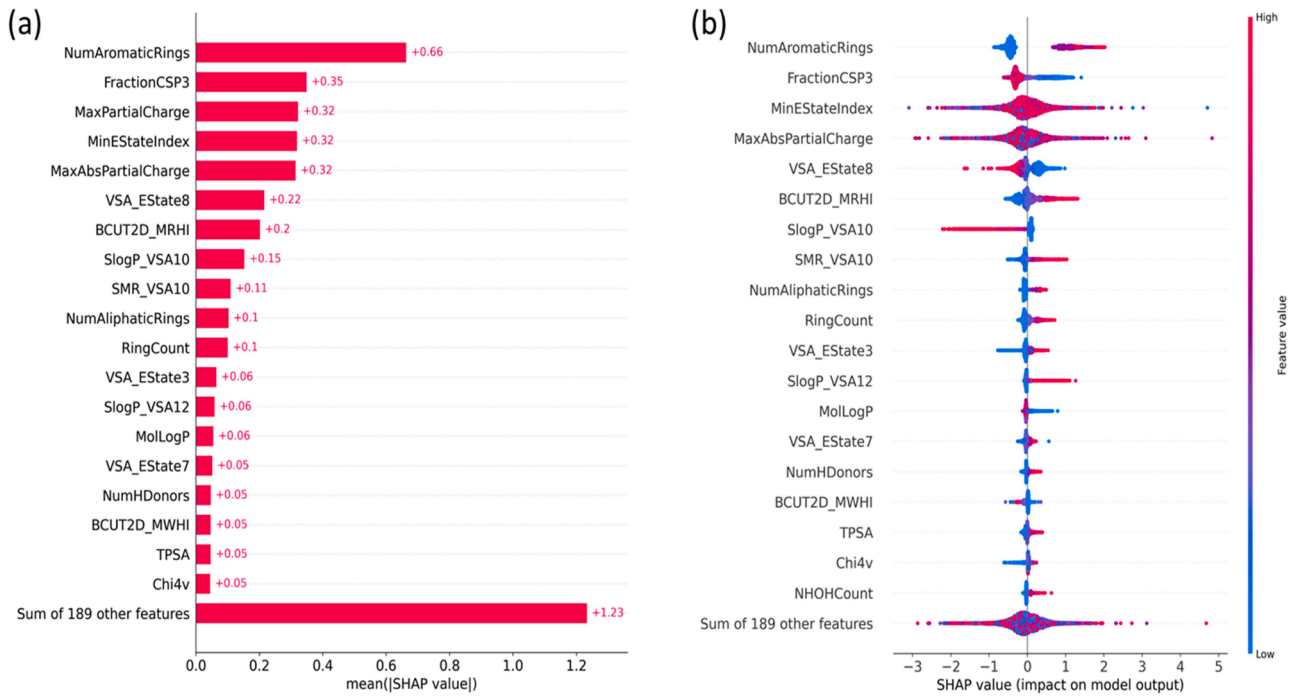

Fig. 15. SHAP values for the Dispersion solubility parameter.

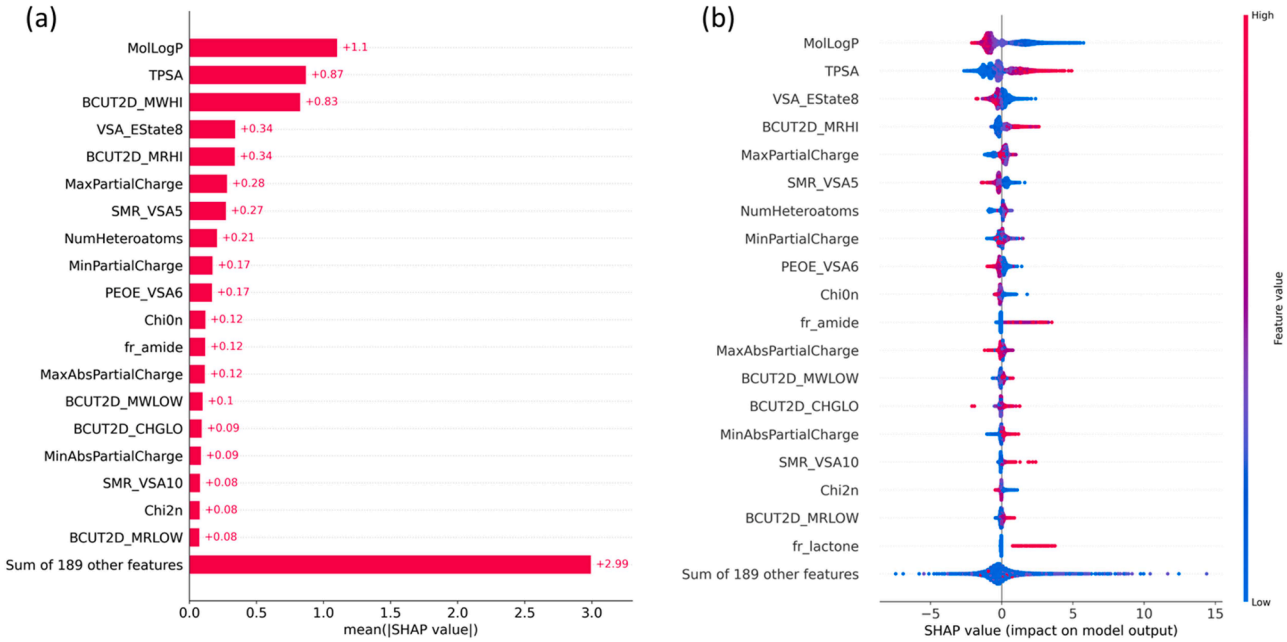

Fig. 16. SHAP values for the Polarization solubility parameter.

the abovementioned variables of interest.Combining high prediction accuracy and XAI has an additional important advantage where its results will be provided to the material designers to accelerate the solvents design/discovery for specific cases such as those presented in the current study. The results can also be used by the process designers to accelerate the selection of the most suitable solvents to make their process more feasible and greener. According to the presented case studies, accelerating the solvent design as well as the process design will accelerate the adoption of *Net-Zero* processes including carbon capture and lignin valorization biorefineries. Therefore, the future work will focus on the generative deep leaning to discover new materials, e.g., lignin and CO_2 solvents, based on their desired properties. The input desired properties will be mainly the Hansen solubility parameters sought. Language models including the generative pre-trained transformers (GPTs) will play a key role in this future research. Another direction for future research is the data augmentation for different types of SMILES codes for various chemical compounds and the corresponding properties for further enhancement of data quality. Furthermore, advanced data imputation using generative deep learning methods also serves in data quality enhancement and this will increase prediction accuracy. Besides, considering data/models uncertainty quantification will increase the confidence of the obtained results. As a side note, SMILES codes alone

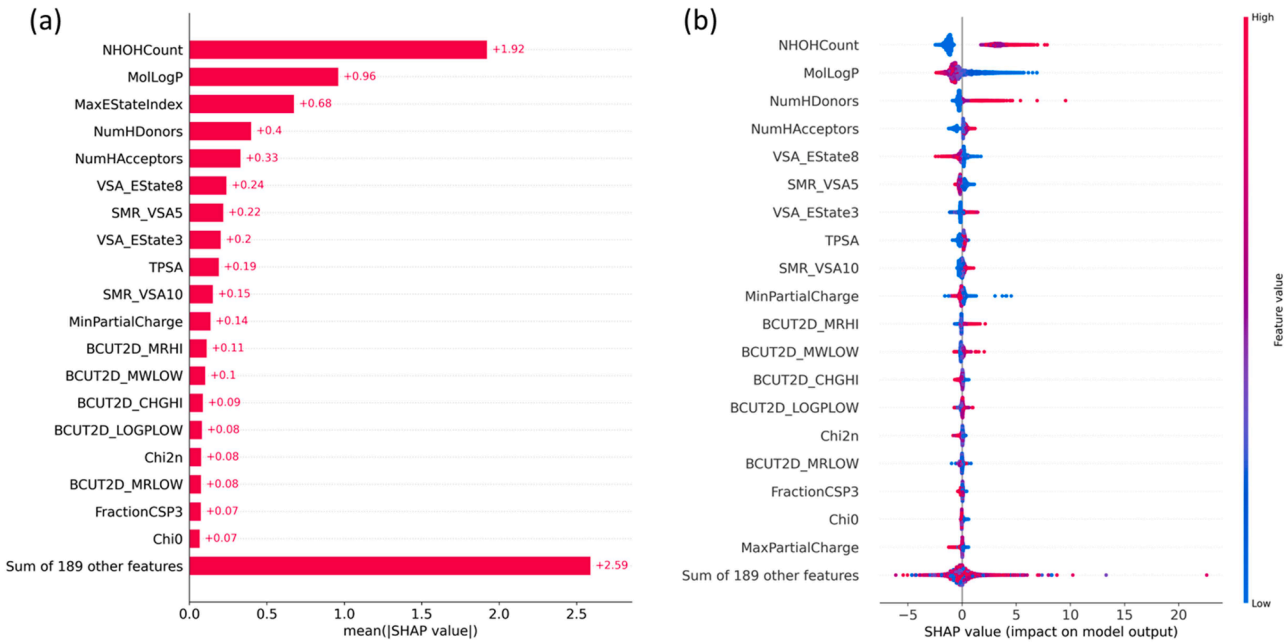

Fig. 17. SHAP values for the Hydrogen Bonding solubility parameter.

have some limitations in fully representing the solvents/compounds of interest. To address this, we proposed using both physio-chemical descriptors and *Morgan* fingerprints for a more comprehensive representation of each compound or solvent. As mentioned earlier, these descriptors and fingerprints were extracted using *RDKit*. However, other physical/chemical properties or descriptors such as phase change points (melting or boiling points), refractive index and heat of vaporization could be incorporated as inputs in future research to further enhance the prediction accuracy and confidence. Additionally, other software packages can be used to extract more informative descriptors, avoiding the limitations that arise when relying on a single library. For future work, we also plan to conduct an extensive comparison between our developed methodologies and other traditional methods, such as functional group contribution methods. This is to provide a robust quantitative assessment in terms of both time efficiency and accuracy along with reasonable interpretability. As we believe, this methodology is generalizable, we plan to extend its use to other applications such as the characterization of hydro-char and pyrolysis products, e.g. biochar, and predicting their yields. Preliminary results have been obtained from this extension, and they are promising.

7. Conclusions and prospects

This research work is a contribution to the development of Green and Sustainable Science and Engineering, where green chemistry and analytical chemistry combined with artificial intelligence methods play an increasingly important role in accurately characterizing the properties of biobased materials for their valorization or conversion into high value-added products. This paper introduces a new AI-methodology to enhance chemical compound identification and to predict accurately the properties of lignin solubility and solvent selection using ensemble machine learning methods for decision fusion. This methodology relies on the creation of an ensemble of different shallow and deep machine learning models having various architectures and algorithms followed by decision fusion through weighted averaging and other learnable pathways. As a result, upon applying the methodology for *Hansen* solubility parameters prediction, the prediction accuracy reached 99% with a mean squared error of 0.02. With respect to the feature selection step, NNMF was successfully used to reduce the number of input variables from 208 descriptors and 2048-bit strings fingerprints to only 70 most significant variables/descriptors. This helped to largely reduce the computational requirements for the training-step. This also increased the prediction accuracy afterwards after eliminating unnecessary input variables that can deteriorate modeling efficiency. The addition of the XAI part to the developed models gave a clearer picture of the impacts of different variables on the predicted outputs that overcame the obstacle of black-box predictions. Clearly, the utilization of an ensemble-based ML approach emerges as a critical strategy in accurately predicting the functional properties of diverse materials including lignin and its derivatives, as well as CO_2 liquid sorbents. This methodology proves reliable to assess lignin compatibility with diverse solvents and polymers, addressing the challenges faced by traditional models and individual ML techniques. By integrating various algorithms and decision fusion methods, our approach significantly enhances prediction accuracy, providing a robust tool for researchers and practitioners in fields such as polymer science, coatings, and biorefineries. The versatility of this ensemble-based ML methodology extends its applicability to a wide range of chemical analytics, reinforcing its importance in advancing the understanding and utilization of lignin in various industrial applications, as well as discovering new efficient solvents for carbon capture. The techniques proposed herein, whereby CO_2 and lignin dissolution represent the two case-studies, have the potential to accelerate the design of green process, thus contributing to *net-zero* emissions targets.

Data availability

Experimental data on solubility of diverse lignins and solvents are used. Such data have been generated by Natural Resources Canada. More data will be available on request.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors state that neither generative nor AI-assisted artificial intelligence technologies were used in writing this manuscript.

CRediT authorship contribution statement

Eslam G. Al-Sakkari: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data

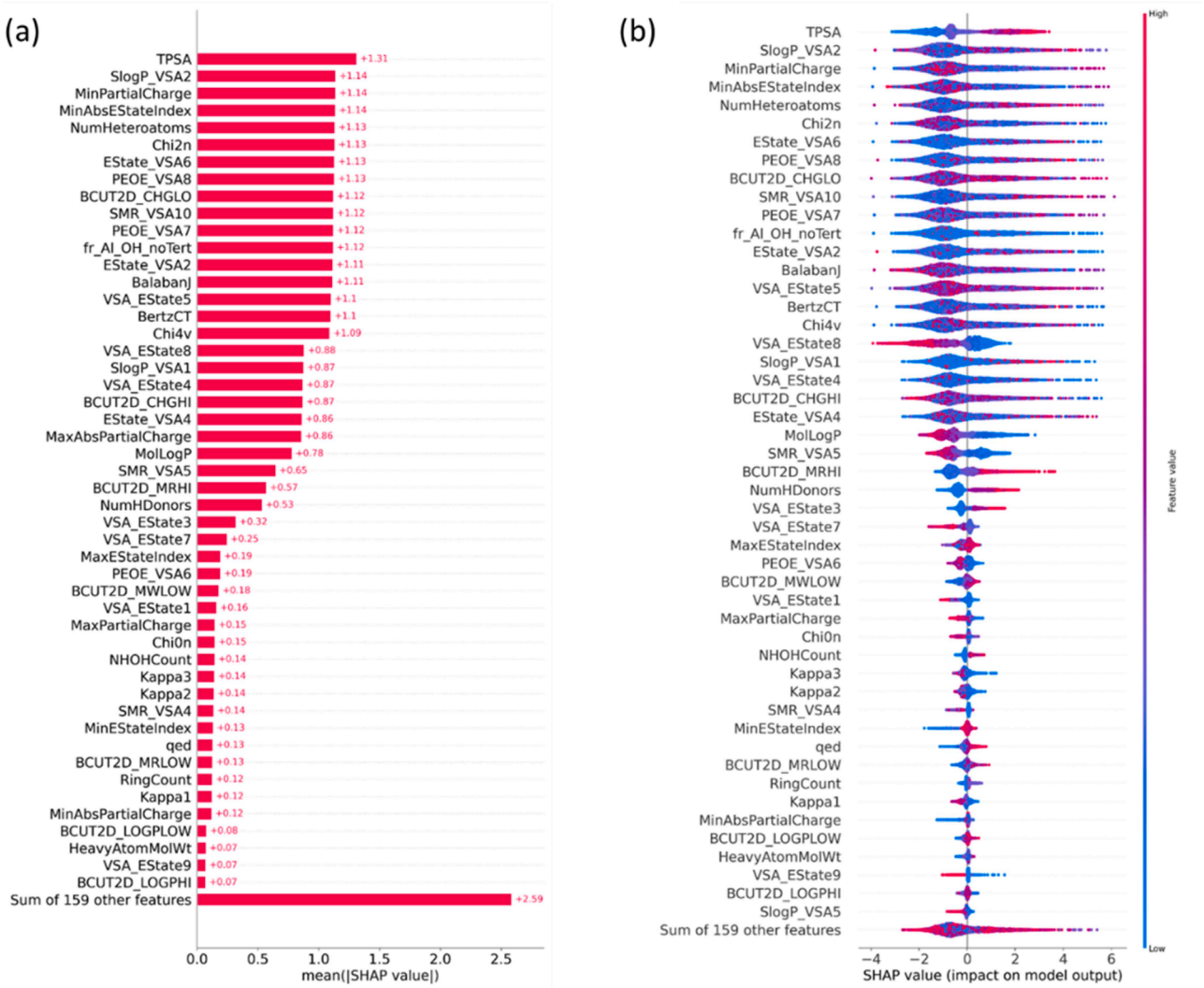

Fig. 18. SHAP values of softwood-based kraft lignin solvents classification based on RED (**Descriptors**).

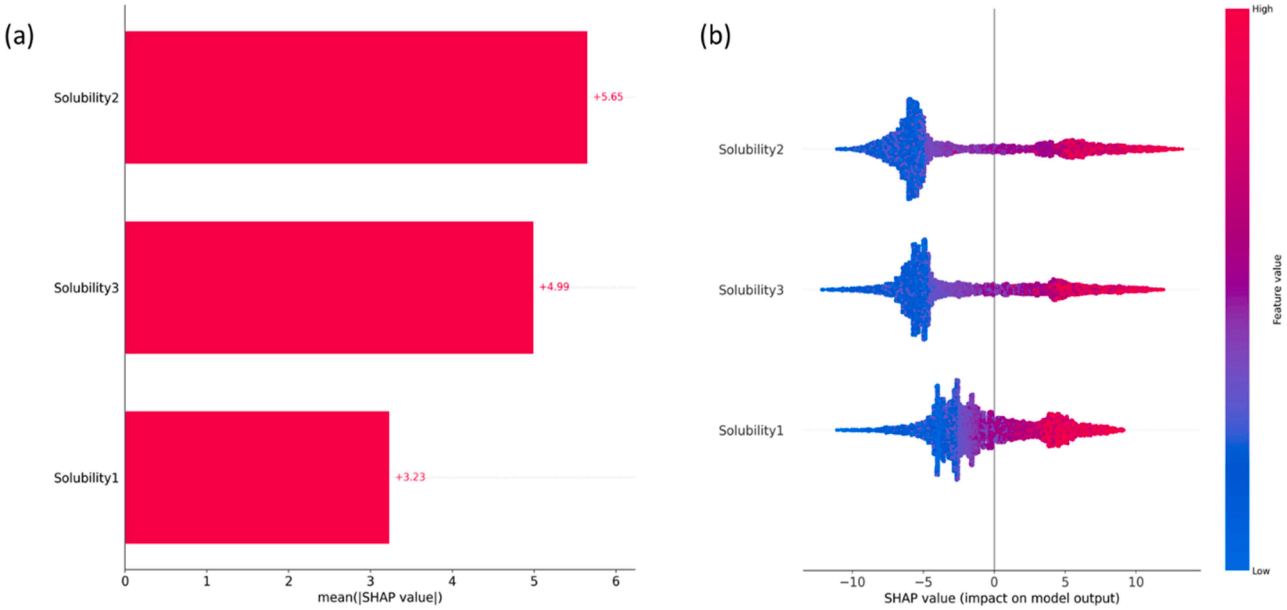

Fig. 19. SHAP values of softwood-based kraft lignin solvents classification based on RED (**Hansen solubility parameters**).

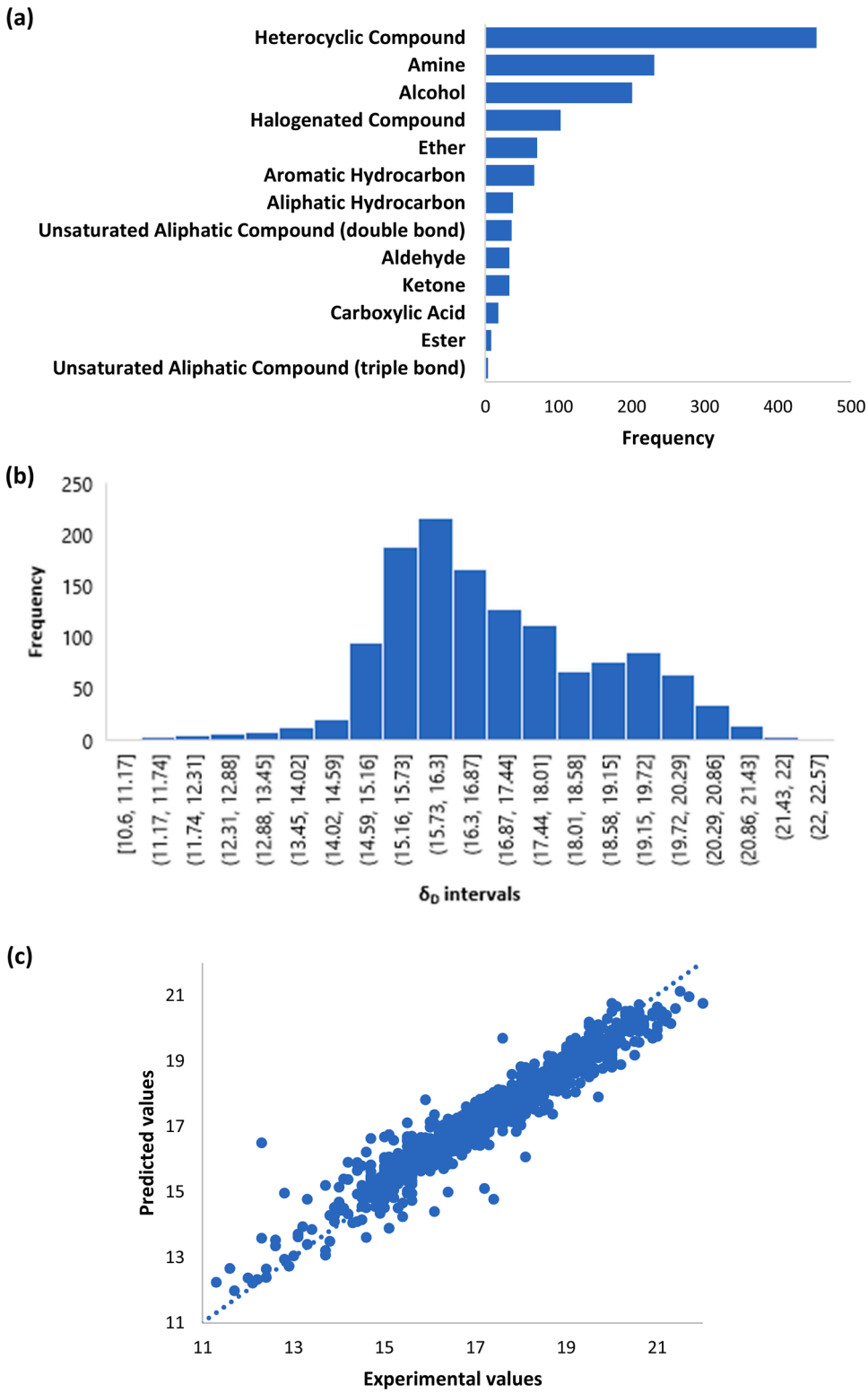

Fig. 20. (a & b) Distributions and (c) results of validation dataset.

curation, Conceptualization. **Ahmed Ragab:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Mostafa Amer:** Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Olumoye Ajao:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marzouk Benali:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Daria C. Boffito:** Writing – review & editing, Supervision. **Hanane Dagdougui:** Writing – review & editing, Supervision. **Mouloud Amazouz:** Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Marzouk Benali reports financial support was provided by Natural Resources Canada. Ahmed Ragab reports financial support was provided by Natural Sciences and Engineering Research Council of Canada (NSERC). Marzouk Benali reports a relationship with Natural Resources Canada that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors are grateful for the financial support received from the Program of the Office of Energy Research and Development and the Forest Innovation Program of the Canadian Forest Service, at Natural Resources Canada that is committed to help the Canadian forest industry make well-informed, stable and thriving transition toward a low-carbon and sustainable economy, as well as the financial support from Natural Sciences and Engineering Research Council of Canada (NSERC). The first author (E. G. Al-Sakkari) would like to thank Mr. Amr Gomaa for his technical support and advice during coding and debugging.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dche.2024.100207.

References

Abba, S.I., et al., 2020. Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration. IEEe Access. 8, 157218–157237.
Abbott, S., Hansen, C.M., 2008. Hansen Solubility Parameters in Practice. Hansen-Solubility.
Abdeldayem, O.M., et al., 2022. Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: a comprehensive review and outlook. Sci. Total Environ. 803, 149834.
Adam, S.P., Alexandropoulos, S.-A.N., Pardalos, P.M., Vrahatis, M.N., 2019. No free lunch theorem: a review. Approx. Optim. Algorithms, Complex. Appl. 57–82.
Adeleke, A.A., et al., 2024. Comparative studies of machine learning models for predicting higher heating values of biomass. Digit. Chem. Eng. 12, 100159.
Ajao, O., Benali, M., El Mehdi, N., 2021. Experimental and computer aided solubility quantification of diverse lignins and performance prediction. Chem. Commun. 57 (14), 1782–1785.
Akinpelu, D.A., Adekoya, O.A., Oladoye, P.O., Ogbaga, C.C., Okolie, J.A., 2023. Machine learning applications in biomass pyrolysis: from biorefinery to end-of-life product management. Digit. Chem. Eng. 8, 100103.
Al-Sakkari, E.G., et al., 2023. Machine learning-assisted selection of adsorption-based carbon dioxide capture materials. J. Environ. Chem. Eng., 110732
Al-Sakkari, E.G., Ragab, A., Dagdougui, H., Boffito, D.C., Amazouz, M., 2024. Carbon capture, utilization and sequestration systems design and operation optimization: assessment and perspectives of artificial intelligence opportunities. Sci. Total Environ., 170085
Al-Sakkari, E.G., et al., Mar. 2020. New alginate-based interpenetrating polymer networks for water treatment: a response surface methodology based optimization study. Int. J. Biol. Macromol. https://doi.org/10.1016/j.ijbiomac.2020.03.220.
Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. In: 2017 international conference on engineering and technology (ICET), pp. 1–6.
Alshehri, A.S., Gani, R., You, F., 2020. Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: state-of-the-art and future directions. Comput. \& Chem. Eng. 141, 107005.
Anowar, F., Sadaoui, S., Selim, B., 2021. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). Comput. Sci. Rev. 40, 100378.
Arias, A., Feijoo, G., Moreira, M.T., 2023. How could Artificial Intelligence be used to increase the potential of biorefineries in the near future? A review. Environ. Technol. \& Innov. 32, 103277.
Arias, A., Estévez-Rivadulla, S., Rebolledo-Leiva, R., Feijoo, G., González-Garc\`\ia, S., Moreira, M.T., 2024. Boosting the transition to biorefineries in compliance with sustainability and circularity criteria. J. Environ. Chem. Eng. 12 (5), 113361.
Asri, A.K., et al., 2024. A machine learning-based ensemble model for estimating diurnal variations of nitrogen oxide concentrations in Taiwan. Sci. Total Environ. 916, 170209.
Asri, A.K., et al., 2024. What is the spatiotemporal pattern of benzene concentration spread over susceptible area surrounding the Hartman Park community, Houston, Texas? J. Hazard. Mater., 134666
Atallah, R., Al-Mousa, A., 2019. Heart disease detection using machine learning majority voting ensemble method. In: 2019 2nd international conference on new trends in computing sciences (ictcs), pp. 1–6.
Awad, M., Khanna, R., Awad, M., Khanna, R., 2015. Support vector regression. Effic. Learn. Mach. Theor. concepts, Appl. Eng. Syst. Des. 67–80.
Balsora, H.K., et al., 2022. Machine learning approach for the prediction of biomass pyrolysis kinetics from preliminary analysis. J. Environ. Chem. Eng. 10 (3), 108025.
Bapat, S., Kilian, S.O., Wiggers, H., Segets, D., 2021. Towards a framework for evaluating and reporting Hansen solubility parameters: applications to particle dispersions. Nanoscale Adv. 3 (15), 4400–4410.
Barker-Rothschild, D., et al., 2024. Lignin-based porous carbon adsorbents for CO 2 capture. Chem. Soc. Rev.
Bento, A.P., et al., 2020. An open source chemical structure curation pipeline using RDKit. J. Cheminform. 12, 1–16.
Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Brigato, L., Iocchi, L., 2021. A close look at deep learning with small data. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2490–2497.
Carmona-Saez, P., Pascual-Marqui, R.D., Tirado, F., Carazo, J.M., Pascual-Montano, A., 2006. Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC. Bioinformatics. 7, 1–18.
Carrott, P.J.M., Carrott, M.M.L.R., 2007. Lignin–from natural adsorbent to activated carbon: a review. Bioresour. Technol. 98 (12), 2301–2312.
Chandrasekaran, A., Kim, C., Venkatram, S., Ramprasad, R., 2020. A deep learning solvent-selection paradigm powered by a massive solvent/nonsolvent database for polymers. Macromolecules. 53 (12), 4764–4769.
Chen, X., Qian, Q., 2023. subGE: enhancing the subgraph representation of molecular compounds structure–activity relationship discovery. Eng. Appl. Artif. Intell. 119, 105727.
T. Chen et al., "Xgboost: extreme gradient boosting," *R Packag. version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.
Chmiela, S., et al., 2023. Accurate global machine learning force fields for molecules with hundreds of atoms. Sci. Adv. 9 (2), eadf0873.
Choi, I., Koh, W., Koo, B., Kim, W.C., 2024. Network-based exploratory data analysis and explainable three-stage deep clustering for financial customer profiling. Eng. Appl. Artif. Intell. 128, 107378.
De La Peña-Gil, A., Toro-Vazquez, J.F., Rogers, M.A., 2016. Simplifying Hansen solubility parameters for complex edible fats and oils. Food Biophys. 11, 283–291.
de los R\`\ios, M., Hernández Ramos, E., 2020. Determination of the Hansen solubility parameters and the Hansen sphere radius with the aid of the solver add-in of Microsoft Excel. SN Appl. Sci. 2, 1–7.
Delbecq, F., Adenier, G., Ogue, Y., Kawai, T., 2020. Gelation properties of various long chain amidoamines: prediction of solvent gelation via machine learning using Hansen solubility parameters. J. Mol. Liq. 303, 112587.
der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).
Dhawane, S.H., Al-Sakkari, E.G., Halder, G., 2019. Kinetic Modelling of Heterogeneous Methanolysis Catalysed by Iron Induced on Microporous Carbon Supported Catalyst. Catal. Letters. https://doi.org/10.1007/s10562-019-02905-5. Jul.
Dietterich, T.G., 2000. Ensemble methods in machine learning. International Workshop On Multiple Classifier Systems, pp. 1–15.
Dogan, A., Birant, D., 2019. A weighted majority voting ensemble approach for classification. In: 2019 4th International Conference on Computer Science and Engineering (UBMK), pp. 1–6.
Duval, A., Vilaplana, F., Crestini, C., Lawoko, M., 2016. Solvent screening for the fractionation of industrial kraft lignin. Holzforschung. 70 (1), 11–20.
Dy, J.G., 2007. Unsupervised feature selection. Computational Methods of Feature Selection. Chapman and Hall/CRC, pp. 35–56.
Emori, E.Y., Ravagnani, M.A.S.S., Costa, C.B.B., 2022. Application of a predictive Q-learning algorithm on the multiple-effect evaporator in a sugarcane ethanol biorefinery. Digit. Chem. Eng. 5, 100049.
Ethier, J.G., et al., 2022. Predicting phase behavior of linear polymers in solution using machine learning. Macromolecules. 55 (7), 2691–2702.
Fan, D., et al., 2023. Deep learning model based on Bayesian optimization for predicting the infinite dilution activity coefficients of ionic liquid-solute systems. Eng. Appl. Artif. Intell. 126, 107127.
Farhan, A.A., Lu, J., Bi, J., Russell, A., Wang, B., Bamis, A., 2016. Multi-view bi-clustering to identify smartphone sensing features indicative of depression. In: 2016 IEEE first international conference on connected health: applications, systems and engineering technologies (CHASE), pp. 264–273.
Fu, K., Yue, Q., Gao, B., Sun, Y., Zhu, L., 2013. Preparation, characterization and application of lignin-based activated carbon from black liquor lignin by steam activation. Chem. Eng. J. 228, 1074–1082.
Götz, J., et al., 2023. High-throughput synthesis provides data for predicting molecular properties and reaction success. Sci. Adv. 9 (43), eadj2314.
Garcia, A.C., Shuo, C., Cross, J.S., 2022. Machine learning based analysis of reaction phenomena in catalytic lignin depolymerization. Bioresour. Technol. 345, 126503.
Ge, H., et al., 2023. Machine learning prediction of delignification and lignin structure regulation of deep eutectic solvents pretreatment processes. Ind. Crops Prod. 203, 117138.
Gharagheizi, F., Torabi Angaji, M., 2006. A new improved method for estimating Hansen Solubility Parameters of polymers. J. Macromol. Sci. Part B Phys. 45 (2), 285–290.

B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Uniform manifold approximation and projection (UMAP) and its variants: tutorial and survey," *arXiv Prepr. arXiv2109.02508*, 2021.
Ghorbani, R., Ghousi, R., 2020. Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEe Access. 8, 67899–67911.
Ginni, G., et al., 2021. Valorization of agricultural residues: different biorefinery routes. J. Environ. Chem. Eng. 9 (4), 105435.
Gisbrecht, A., Schulz, A., Hammer, B., 2015. Parametric nonlinear dimensionality reduction using kernel t-SNE. Neurocomputing. 147, 71–82.
Gu, J., Liu, J.S., 2008. Bayesian biclustering of gene expression data. BMC. Genomics. 9 (1), 1–10.
Gu, J., et al., 2018. Recent advances in convolutional neural networks. Pattern. Recognit. 77, 354–377.
Hähnke, V.D., Kim, S., Bolton, E.E., 2018. PubChem chemical structure standardization. J. Cheminform. 10, 1–40.
Han, R., et al., 2019. Predicting physical stability of solid dispersions by machine learning techniques. J. Control. Release 311, 16–25.
"Hansen solubility parameters in practice (official web page)." https://www.hansen-solubility.com/HSPiP/(accessed Aug. 07, 2023).
Hansen, C.M., 1967. The three dimensional solubility parameter. Danish Tech. Copenhagen 14.
Hansen, C.M., 2007. Hansen Solubility parameters: a User's Handbook. CRC press.
Hasan, B.M.S., Abdulazeez, A.M., 2021. A review of principal component analysis algorithm for dimensionality reduction. J. Soft Comput. Data Min. 2 (1), 20–30.
Hashemi, S.H., Besharati, Z., Hashemi, S.A., 2024. Salicylic acid solubility prediction in different solvents based on machine learning algorithms. Digit. Chem. Eng., 100157
He, Y., Lou, R., Wang, Y., Wang, J., Fang, X., 2022. A dual attribute weighted decision fusion system for fault classification based on an extended analytic hierarchy process. Eng. Appl. Artif. Intell. 114, 105066.
He, L., et al., 2024. Reaction condition-and functional group-specific knowledge discovery: data-and computation-based analysis on transition-metal-free transformation of organoborons. Artif. Intell. Chem. 2 (1), 100034.
Heiat, A., 2002. Comparison of artificial neural network and regression models for estimating software development effort. Inf. Softw. Technol. 44 (15), 911–922.
Henriques, R., Madeira, S.C., 2015. Biclustering with flexible plaid models to unravel interactions between biological processes. IEEE/ACM Trans. Comput. Biol. Bioinforma. 12 (4), 738–752.
Hsiao, Y.-D., Chang, C.-T., 2024. Joint incremental learning network for flexible modeling of carbon dioxide solubility in aqueous mixtures of amines. Sep. Purif. Technol. 330, 125299.
Hu, P., Jiao, Z., Zhang, Z., Wang, Q., 2021. Development of solubility prediction models with ensemble learning. Ind. \& Eng. Chem. Res. 60 (30), 11627–11635.
L. Hui, M. Belkin, and P. Nakkiran, "Limitations of neural collapse for understanding generalization in deep learning," *arXiv Prepr. arXiv2202.08384*, 2022.
Jablonka, K.M., et al., 2023. Machine learning for industrial processes: forecasting amine emissions from a carbon capture plant. Sci. Adv. 9 (1), eadc9576.
Jackson, N.E., Webb, M.A., de Pablo, J.J., 2019. Recent advances in machine learning towards multiscale soft materials design. Curr. Opin. Chem. Eng. 23, 106–114.
Jarvas, G., Quellet, C., Dallos, A., 2011. Estimation of Hansen solubility parameters using multivariate nonlinear QSPR modeling with COSMO screening charge density moments. Fluid. Phase Equilib. 309 (1), 8–14.
Jeong, H., Kim, S., Gil, M., Song, S., Kim, T.-H., Lee, K.J., 2020. Preparation of poly-1-butene nanofiber mat and its application as shutdown layer of next generation lithium ion battery. Polymers. (Basel) 12 (10), 2267.
Kalna, G., Vass, J.Keith, Higham, D.J., 2008. Multidimensional partitioning and bi-partitioning: analysis and application to gene expression data sets. Int. J. Comput. Math. 85 (3–4), 475–485.
Khan, N., Ammar Taqvi, S.A., 2023. Machine learning an intelligent approach in process industries: a perspective and overview. ChemBioEng Rev 10 (2), 195–221.
Khashaba, N.H., Ettouney, R.S., Abdelaal, M.M., Ashour, F.H., El-Rifai, M.A., 2022. Artificial neural network modeling of biochar enhanced anaerobic sewage sludge digestion. J. Environ. Chem. Eng. 10 (4), 107988.
Kluger, Y., Basri, R., Chang, J.T., Gerstein, M., 2003. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. 13 (4), 703–716.
Kobayashi, K., Alam, S.B., 2024. Explainable, interpretable, and trustworthy AI for an intelligent digital twin: a case study on remaining useful life. Eng. Appl. Artif. Intell. 129, 107620.
Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.
Lee, D., Seung, H.S., 2000. Algorithms for non-negative matrix factorization. Adv. Neural Inf. Process. Syst. 13.
Lee, M., Shen, H., Huang, J.Z., Marron, J.S., 2010. Biclustering via sparse singular value decomposition. Biometrics 66 (4), 1087–1095.
Lee, M.-H., 2023. Interpretable machine-learning for predicting power conversion efficiency of non-halogenated green solvent-processed organic solar cells based on Hansen solubility parameters and molecular weights of polymers. Sol. Energy 261, 7–13.
Leonard, K.C., Hasan, F., Sneddon, H.F., You, F., 2021. Can artificial intelligence and machine learning be used to accelerate sustainable chemistry and engineering? ACS Sustainable Chemistry \& Engineering 9 (18), 6126–6129. ACS Publications.
Li, Y., Ngom, A., 2013. The non-negative matrix factorization toolbox for biological data mining. Source Code Biol. Med. 8 (1), 1–15.
Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans. neural networks Learn. Syst.
Li, M., Liu, X., Sun, C., Stevens, L., Liu, H., 2022. Synthesis and characterization of advanced bio-carbon materials from Kraft lignin with enhanced CO2 capture properties. J. Environ. Chem. Eng. 10 (3), 107471.
Li, M., et al., 2022. New parameter derived from the hansen solubility parameter used to evaluate the solubility of asphaltene in solvent. ACS. Omega 7 (16), 13801–13807.
Li, R., et al., 2023. Selective value-added conversion of lignin derivatives over heterogeneous catalysts of TEMPO-functionalized metal-organic frameworks. J. Environ. Chem. Eng. 11 (3), 109700.
Liu, H., Motoda, H., 1998. Computational methods of feature selection. Chapman \&.
Liu, Y., Wang, Y., Zhang, J., 2012. New machine learning algorithm: random forest. In: Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3, pp. 246–252.
Liu, H., et al., 2024. A generic machine learning model for CO2 equilibrium solubility into blended amine solutions. Sep. Purif. Technol. 334, 126100.
Lofgren, J., Tarasov, D., Koitto, T., Rinke, P., Balakshin, M., Todorovic, M., 2022. Machine learning optimization of lignin properties in green biorefineries. ACS Sustain. Chem. \& Eng. 10 (29), 9469–9479.
Madeira, S.C., Oliveira, A.L., 2004. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. Comput. Biol. Bioinforma. 1 (1), 24–45.
L. McInnes, J. Healy, and J. Melville, "Umap: uniform manifold approximation and projection for dimension reduction," *arXiv Prepr. arXiv1802.03426*, 2018.
E. Meeds and S. Roweis, "Nonparametric bayesian biclustering," 2007.
Mehta, S., Rana, P., Singh, S., Sharma, A., Agarwal, P., 2019. Ensemble learning approach for enhanced stock prediction. In: 2019 twelfth international conference on contemporary computing (IC3), pp. 1–5.
Meuwly, M., 2021. Machine learning for chemical reactions. Chem. Rev. 121 (16), 10218–10239.
Mian, Z., et al., 2024. A literature review of fault diagnosis based on ensemble learning. Eng. Appl. Artif. Intell. 127, 107357.
Mohan, M., et al., 2022. Prediction of solubility parameters of lignin and ionic liquids using multi-resolution simulation approaches. Green. Chem. 24 (3), 1165–1176.
Morimoto, M., Fukami, K., Zhang, K., Fukagata, K., 2022. Generalization techniques of neural networks for fluid flow estimation. Neural Comput. Appl. 1–23.
Naeem, M.M., Al-Sakkari, E.G., Boffito, D.C., Gadalla, M.A., Ashour, F.H., 2021. One-pot conversion of highly acidic waste cooking oil into biodiesel over a novel bio-based bi-functional catalyst. Fuel 283. https://doi.org/10.1016/j.fuel.2020.118914. Jan.
Naeem, M.M., Al-Sakkari, E.G., Boffito, D.C., Rene, E.R., Gadalla, M.A., Ashour, F.H., 2023. Single-stage waste oil conversion into biodiesel via sonication over bio-based bifunctional catalyst: optimization, preliminary techno-economic and environmental analysis. Fuel 341, 127587.
Nagulapati, V.M., Rehman, H.M.R.U., Haider, J., Qyyum, M.A., Choi, G.S., Lim, H., 2022. 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. 298, 121651.
Naimi, A.I., Balzer, L.B., 2018. Stacked generalization: an introduction to super learning. Eur. J. Epidemiol. 33, 459–464.
Neloy, M.A.I., Nahar, N., Hossain, M.S., Andersson, K., 2022. A weighted average ensemble technique to predict heart disease. In: Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021, pp. 17–29.
"Non-negative matrix factorization." https://www.geeksforgeeks.org/non-negative-matrix-factorization/(accessed May 23, 2023).
Novo, L.P., Curvelo, A.A.S., 2019. Hansen solubility parameters: a tool for solvent selection for organosolv delignification. Ind. \& Eng. Chem. Res. 58 (31), 14520–14527.
O'Dea, R.M., et al., 2022. Ambient-pressure lignin valorization to high-performance polymers by intensified reductive catalytic deconstruction. Sci. Adv. 8 (3), eabj7523.
K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv Prepr. arXiv1511.08458*, 2015.
Obradović, D., Andrić, F., Zlatović, M., Agbaba, D., 2018. Modeling of Hansen's solubility parameters of aripiprazole, ziprasidone, and their impurities: a nonparametric comparison of models for prediction of drug absorption sites. J. Chemom. 32 (4), e2996.
Peng, L., Peng, M., Liao, B., Huang, G., Li, W., Xie, D., 2018. The advances and challenges of deep learning application in biological big data processing. Curr. Bioinform. 13 (4), 352–359.
Perea, J.D., et al., 2016. Combined computational approach based on density functional theory and artificial neural networks for predicting the solubility parameters of fullerenes. J. Phys. Chem. B 120 (19), 4431–4438.
Pilario, K.E.S., Ching, P.M.L., Calapatia, A.M.A., Culaba, A.B., 2022. Predicting drying curves in algal biorefineries using Gaussian process autoregressive models. Digit. Chem. Eng. 4, 100036.
Poličar, P.G., Stražar, M., Zupan, B., 2019. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. bioRxiv., 731877
Prelić, A., et al., 2006. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 22 (9), 1122–1129.
Przybyłek, M., Jeliński, T., Cysewski, P., 2019. Application of multivariate adaptive regression splines (MARSplines) for predicting hansen solubility parameters based on 1D and 2D molecular descriptors computed from SMILES string. J. Chem. 2019.
Pyzer-Knapp, E.O., et al., 2022. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput. Mater. 8 (1), 1–9.
Ragab, A., Ghezzaz, H., Amazouz, M., 2022. Decision fusion for reliable fault classification in energy-intensive process industries. Comput. Ind. 138, 103640.

Ray, S., 2019. A quick review of machine learning algorithms. In: 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), pp. 35–39.
Reddy, G.T., et al., 2020. Analysis of dimensionality reduction techniques on big data. IEEe Access. 8, 54776–54788.
Rexhepi, F., Woolever, M., Nabity, J., Banerjee, S., 2023. Metal oxide solvation with ionic liquids: a solubility parameter analysis. J. Mol. Liq., 122314
Ribeiro, W.C.O., Martinez, P.F.M., Lobosco, V., 2020. Solubility parameters analysis of Eucalyptus urograndis kraft lignin. BioResources 15 (4), 8577.
Ritt, C.L., Liu, M., Pham, T.A., Epsztein, R., Kulik, H.J., Elimelech, M., 2022. Machine learning reveals key ion selectivity mechanisms in polymeric membranes with subnanometer pores. Sci. Adv. 8 (2), eabl5771.
Ruwoldt, J., Tanase-Opedal, M., Syverud, K., 2022. Ultraviolet Spectrophotometry of Lignin Revisited: exploring Solvents with Low Harmfulness, Lignin Purity, Hansen Solubility Parameter, and Determination of Phenolic Hydroxyl Groups. ACS. Omega 7 (50), 46371–46383.
Sanchez-Lengeling, B., Aspuru-Guzik, A., 2018. Inverse molecular design using machine learning: generative models for matter engineering. Science (80-.) 361 (6400), 360–365.
Sanchez-Lengeling, B., Roch, L.M., Perea, J.D., Langner, S., Brabec, C.J., Aspuru-Guzik, A., 2019. A Bayesian approach to predict solubility parameters. Adv. Theory Simulations 2 (1), 1800069.
Schapire, R.E., Freund, Y., 2013. Boosting: foundations and algorithms. Kybernetes 42 (1), 164–166.
Schapire, R.E., 2003. The boosting approach to machine learning: an overview. Nonlinear Estim. Classif. 149–171.
Schapire, R.E., 2013. Explaining adaboost. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Springer, pp. 37–52.
Schieppati, D., et al., 2023. Chemical and biological delignification of biomass: a review. Ind. \& Eng. Chem. Res. 62 (33), 12757–12794.
Schonlau, M., Zou, R.Y., 2020. The random forest algorithm for statistical learning. Stata J. 20 (1), 3–29.
Schulz, E., Speekenbrink, M., Krause, A., 2018. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. J. Math. Psychol. 85, 1–16.
Sen, S., Singh, K.P., Chakraborty, P., 2023. Dealing with imbalanced regression problem for large dataset using scalable Artificial Neural Network. New Astron 99, 101959.
K. Sentz and S. Ferson, "Combination of evidence in Dempster-Shafer theory," 2002.
Sester, M., Feng, Y., Thiemann, F., 2018. Building generalization using deep learning. ISPRS-International Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII-4 42, 565–572.
Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys. D Nonlinear Phenom. 404, 132306.
Sinaga, K.P., Yang, M.-S., 2020. Unsupervised K-means clustering algorithm. IEEe Access. 8, 80716–80727.
Sistla, Y.S., Jain, L., Khanna, A., 2012. Validation and prediction of solubility parameters of ionic liquids for CO2 capture. Sep. Purif. Technol. 97, 51–64.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.
Specht, D.F., 1991. A general regression neural network. IEEE Trans. neural networks 2 (6), 568–576.
Sreekanth, T.V.M., Ramanaiah, S., Lee, K.D., Reddy, K.S., 2012. Hansen solubility parameters in the analysis of solvent–solvent interactions by inverse gas chromatography. J. Macromol. Sci. Part B 51 (6), 1256–1266.
R.C. Staudemeyer and E.R. Morris, "Understanding LSTM–a tutorial into long short-term memory recurrent neural networks," *arXiv Prepr. arXiv1909.09586*, 2019.
Stefanis, E., Panayiotou, C., 2008. Prediction of Hansen solubility parameters with a new group-contribution method. Int. J. Thermophys. 29, 568–585.
Su, Y., Wang, Z., Jin, S., Shen, W., Ren, J., Eden, M.R., 2019. An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures. AIChe J. 65 (9), e16678.
Su, Y., Jin, S., Zhang, X., Shen, W., Eden, M.R., Ren, J., 2020. Stakeholder-oriented multi-objective process optimization based on an improved genetic algorithm. Comput. \& Chem. Eng. 132, 106618.
Subrahmanyam, R., Gurikov, P., Dieringer, P., Sun, M., Smirnova, I., 2015. On the road to biopolymer aerogels—Dealing with the solvent. Gels. 1 (2), 291–313.
Supanchaiyamat, N., Jetsrisuparb, K., Knijnenburg, J.T.N., Tsang, D.C.W., Hunt, A.J., 2019. Lignin materials for adsorption: current trend, perspectives and opportunities. Bioresour. Technol. 272, 570–581.
Sutton, C.D., 2005. Classification and regression trees, bagging, and boosting. Handb. Stat. 24, 303–329.
T. Tamura and H. Yamamoto, "Calculation of Hansen solubility parameters based on solvatochromic dye," 2019.
Taqvi, S.A.A., Zabiri, H., Tufa, L.D., Uddin, F., Fatima, S.A., Maulud, A.S., 2021. A review on data-driven learning approaches for fault detection and diagnosis in chemical processes. ChemBioEng Rev 8 (3), 239–259.
Tidriri, K., Tiplica, T., Chatti, N., Verron, S., 2018. A generic framework for decision fusion in fault detection and diagnosis. Eng. Appl. Artif. Intell. 71, 73–86.
Unke, O.T., et al., 2024. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. Sci. Adv. 10 (14), eadn4397.
US-Environmental Protection Agency, "SMILES Tutorial." https://archive.epa.gov/med/med_archive_03/web/html/smiles.html (accessed Aug. 11, 2023).
Varshney, S., Lakshmi, C.V., Patvardhan, C., 2023. Madhubani art classification using transfer learning with deep feature fusion and decision fusion based techniques. Eng. Appl. Artif. Intell. 119, 105734.
Venkatram, S., Kim, C., Chandrasekaran, A., Ramprasad, R., 2019. Critical assessment of the Hildebrand and Hansen solubility parameters for polymers. J. Chem. Inf. Model. 59 (10), 4188–4194.
Wang, Y.-X., Zhang, Y.-J., 2012. Nonnegative matrix factorization: a comprehensive review. IEEE Trans. Knowl. Data Eng. 25 (6), 1336–1353.
Wang, W., Zhang, M., Chen, G., Jagadish, H.V, Ooi, B.C., Tan, K.-L., 2016. Database meets deep learning: challenges and opportunities. ACM Sigmod Rec 45 (2), 17–22.
Wang, Z., et al., 2019. Predictive deep learning models for environmental properties: the direct calculation of octanol–water partition coefficients from molecular graphs. Green. Chem. 21 (16), 4555–4565.
Wang, Z., et al., 2020. A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. Green. Chem. 22 (12), 3867–3876.
Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y., 2021. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. J. Mach. Learn. Res. 22 (1), 9129–9201.
Wang, Z., et al., 2022. Insights into ensemble learning-based data-driven model for safety-related property of chemical substances. Chem. Eng. Sci. 248, 117219.
Wang, Y., Gao, X., Ru, X., Sun, P., Wang, J., 2023. Using feature selection and Bayesian network identify cancer subtypes based on proteomic data. J. Proteomics 280, 104895.
Wen, H., et al., 2022. A systematic modeling methodology of deep neural network-based structure-property relationship for rapid and reliable prediction on flashpoints. AIChe J. 68 (1), e17402.
Wen, H., et al., 2023. A systematic review on intensifications of artificial intelligence assisted green solvent development. Ind. \& Eng. Chem. Res. 62 (48), 20473–20491.
Williams, L.L., Rubin, J.B., Edwards, H.W., 2004. Calculation of Hansen solubility parameter values for a range of pressure and temperature conditions, including the supercritical fluid region. Ind. \& Eng. Chem. Res. 43 (16), 4967–4972.
L.L. Williams, "10 determination of hansen solubility parameter values for carbon dioxide," 2007.
Wolpert, D.H., 1992. Stacked generalization. Neural networks 5 (2), 241–259.
Wolpert, D.H., 2002. The supervised learning no-free-lunch theorems. Soft Comput. Ind. Recent Appl. 25–42.
York, D., Vidal-Daza, I., Segura, C., Norambuena-Contreras, J., Martin-Martinez, F.J., 2024. Data-driven representative models to accelerate scaled-up atomistic simulations of bitumen and biobased complex fluids. Digit. Discov. 3 (6), 1108–1122.
Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L., Qin, H., 2020. Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. J. Hydrol. 582, 124293.
Zahrt, A.F., Henle, J.J., Rose, B.T., Wang, Y., Darrow, W.T., Denmark, S.E., 2019. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science (80-.). 363 (6424), eaau5631.
Zeidler, V.Zuin, 2024. Digitalization paving the ways for sustainable chemistry: switching on more green lights. Science (1979) 384 (6701), eadq3537.
Zhang, F., O'Donnell, L.J., 2020. Support vector regression. Machine Learning. Elsevier, pp. 123–140.
Zhang, J., Wang, Q., Shen, W., 2022. Message-passing neural network based multi-task deep-learning framework for COSMO-SAC based $σ$-profile and VCOSMO prediction. Chem. Eng. Sci. 254, 117624.
Zhang, J., et al., 2022. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. AIChe J. 68 (6), e17634.
Zhang, J., Wang, Q., Eden, M., Shen, W., 2023. A deep learning-based framework towards inverse green solvent design for extractive distillation with multi-index constraints. Comput. \& Chem. Eng. 177, 108335.
Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbors. Ann. Transl. Med. 4 (11).
Zhao, Y., Qian, Y., Li, C., 2017. Improved KNN text classification algorithm with MapReduce implementation. In: 2017 4th International Conference on Systems and Informatics (ICSAI), pp. 1417–1422.
Zhao, G., Ni, H., Jia, L., Ren, S., Fang, G., 2018. Quantitative analysis of relationship between Hansen solubility parameters and properties of alkali lignin/acrylonitrile–butadiene–styrene blends. ACS. Omega 3 (8), 9722–9728.
Zhao, B., et al., 2021. Lignin-based porous supraparticles for carbon capture. ACS. Nano 15 (4), 6774–6786.