Chem Rev Lett 7 (2024) 346-358

Review Article

Chemical Review and Letters
journal homepage: www.chemrevlett.com
ISSN (online): 2645-4947 (print) 2676-7279

Materials alternative recommender using machine learning based on COSMO-SAC

Ameer Hassan Idan[1], Mustafa Humam Sami[2], Nahed Mahmood Ahmed Alsultany[3], Mohammed Ubaid[4], Rebaz Obaid Kareem[5]

[1]*Al-Zahrawi University College, Karbala, Iraq*
[2]*Department of Pharmacy, Al-Noor University College, Nineveh, Iraq*
[3]*Collage of Dentist, National University of Science and Technology, Dhi Qar, 64001, Iraq*
[4] *Medical technical college, Al-Farahidi University, Iraq*
[5] *Physics Department, College of Science, University of Halabja, 46018, Halabja, Iraq*

ARTICLE INFO

Article history:
Received 5 March 2024
Received in revised form 29 March 2024
Accepted 2 April 2024
Available online 26 April 2024

Keywords:
COSMO-SAC
Activity Coefficient
FreeSolv dataset
Machine learning
Density-based clustering
K-nearest neighbors

ABSTRACT

Finding alternative materials and solvents in a chemistry lab or the process of designing would be a time-consuming matter. The activity coefficient is one of the most important thermodynamic properties that could be used for this purpose. COSMO-SAC modeling is a reliable method to determine the activity coefficient of the mixtures and is used to find alternatives to the organic materials in the present study. A dataset of 96 organic molecules' activity coefficients in the different solvents (water, ethanol, methanol, toluene, and benzene) mixtures have been obtained in full range composition with COSMO-SAC. The created database has been merged with the FreeSolv dataset to extend the diversity of the properties to enrich the dataset for machine learning training. Unsupervised machine learning methods (clustering) including centroid-based and density-based clustering methods have been conducted to introduce the best alternatives for the studied 96 organic materials. Proper pre-processing for these methods has been utilized to evaluate the optimum parameters of the clustering methods including the elbow method for centroid-based clustering and k-nearest neighbors for the density-based clustering. The centroid-based clustering methods recommend a different variety of materials based on the cluster numbers and sorting the alternatives based on the nearest properties. However, the density-based method works with the optimum distance and the number of the k-nearest neighbors that were 0.08 and 7, respectively for the created dataset. Its results are exclusive and show that the clustering could be used to isolate the clusters based on the chemical families which were 5 clusters and 12 out layers. The out layers are important since no alternatives have been introduced for them in the trained dataset and should be considered as unique materials. The density-based clustering results were more promising using COSMO-SAC data for organic materials alternative recommender.

1. Introduction

Designing a process in chemical engineering such as separation, extraction, and azeotropic distillation on the academic and industrial scale could be a dead-end due to the lack of exact solvents and materials. Also, replacing a non-toxic material instead of toxic alternative might be necessary [1]. Knowledge about the alternatives for the solvents and industrial materials always was a necessity to keep the process going on and reduce the costs.

Corresponding author; e-mail: obedrebaz9@gmail.com
https://doi.org/10.22034/crl.2024.447155.1305

The thermodynamic properties of materials are key for alternating them with other chemicals in the process of designing, literally. The activity coefficient is one of the most important properties that could be used to compare the materials and alternating them.

The COSMO-SAC model is commonly used for the activity coefficient calculation of mixtures [2]. It works based on the statistical thermodynamics that gets σ-profiles from quantum mechanics calculations as input. Generally, $dmol^3$ was used for geometry optimization and minimization of molecule energy, and evaluation of σ-profiles [3]. Also, the COSMO-SAC model provides good results for the activity coefficient with a low deviation from experimental results. Indeed, it has a good reputation and is considered a reliable method in the prediction of the activity coefficient of organic materials. Also, it has been shown that the COSMO-SAC thermodynamic properties depends on the chemical family rather than the size of the molecule that makes it powerful tool for the purpose of this study [4].

The activity coefficient could be used to compare two materials in process engineering. However, comparing more than two materials could be impossible and time-consuming for engineers or academics. Also, dealing with more properties rather than activity coefficient would be harder to achieve accurate deduction. Machine learning could be used to overcome this problem and save time and costs to find an alternative chemical for process designing. Generally, machine learning is used for supervised, unsupervised, and regression problems [5]. Finding alternative materials is an unsupervised and clustering problem.

Machine learning has been used in chemical reactions as a predictor of the product, or determination of the reaction rate [6]. Machine learning can use different data for training with a large number of features as variables and a label for prediction. Also, it could be used for the classification as a supervised method with initial conditions. On the other hand, clustering is an unsupervised machine learning process to put different inputs to specific clusters. Different methods could be used for clustering that main methods are partitioning, density-based, distribution model-based, hierarchical, and Fuzzy clustering [7,8].

An activity coefficient dataset based on the COSMO-SAC modeling has been evaluated for 96 aqueous mixtures of organic solvents and merged with the FreeSolv dataset including the free energy of hydration for the same organic molecules, [9]. There are experimental, DFT calculations, and molecular dynamics data including different thermodynamic properties and molecular descriptive in the FreeSolv [9]. Also, the COSMO-SAC implementation by Bell et al has been used to calculate the activity coefficient of the mixture of these materials with different solvents [10]. A machine learning clustering with different algorithms has been conducted to provide a solution for alternating chemicals based on the dataset's thermodynamic properties and other features.

2. Results and discussion

COSMO-SAC

The activity coefficient of the organic compounds in different solvents has been predicted using the COSMO-SAC model, and the corresponding predicted data are given in supporting information in Tables S1 – S5. The predicted activity coefficient data for binary mixtures of thiophene in the studied solvents have been illustrated in Figure 1 as an example.

The large amount of data and its diversity is clear. Accordingly, no data validation is carried out. However, the pioneers and the developers of the COSMO-SAC implementation have shown that the model is quite reliable [2-4, 10-13]. On the other hand, the integrity of the evaluated data with the COSMO-SAC is more important rather than the accuracy of the data, and no data comparison with experimental results has been carried out. The COSMO-SAC uses the quantum mechanics data as primary data and evaluates chemical thermodynamic data [14]. These two types of data might be in contradiction due to their different microscopic and macroscopic approaches. Accordingly, the evaluated data has been used without validation in the machine learning process.

2.1 Machine learning clustering

The first step in machine learning clustering is to find the optimal number of clusters. Determination of the optimal number of the clusters depends on the utilized method.

Elbow and silhouette score methods are used in the centroid-based methods such as K-means [15]. Using both of the methods would be more reliable [16]. The elbow and silhouette score results for the K-means are given in Figure 2 for the evaluated dataset.

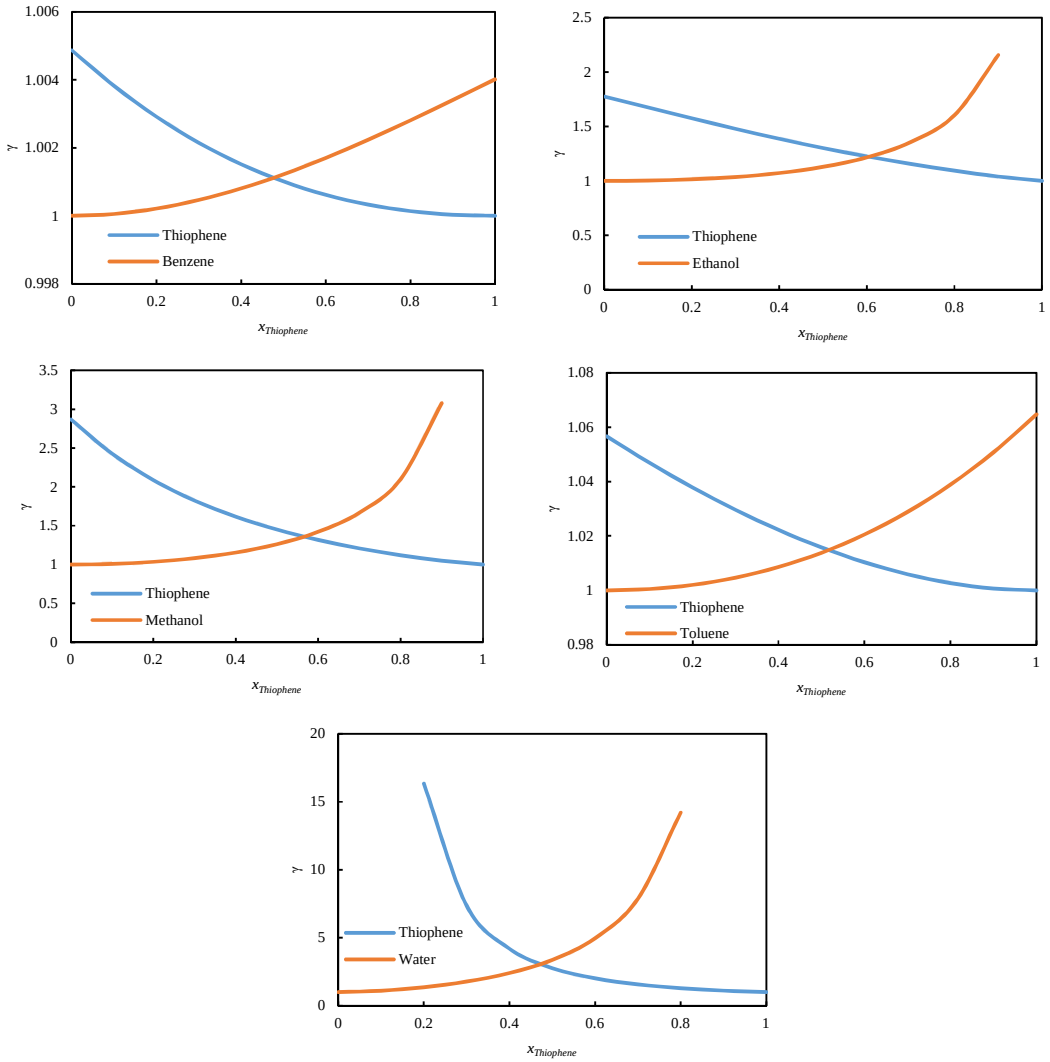

Fig. 1. The activity coefficients of the binary mixture's components including thiophene in different solvents (water, ethanol, methanol, benzene, and toluene) versus the mole fraction of thiophene using COSMO-SAC under 0.1 MPa pressure at 298.15 K.

b

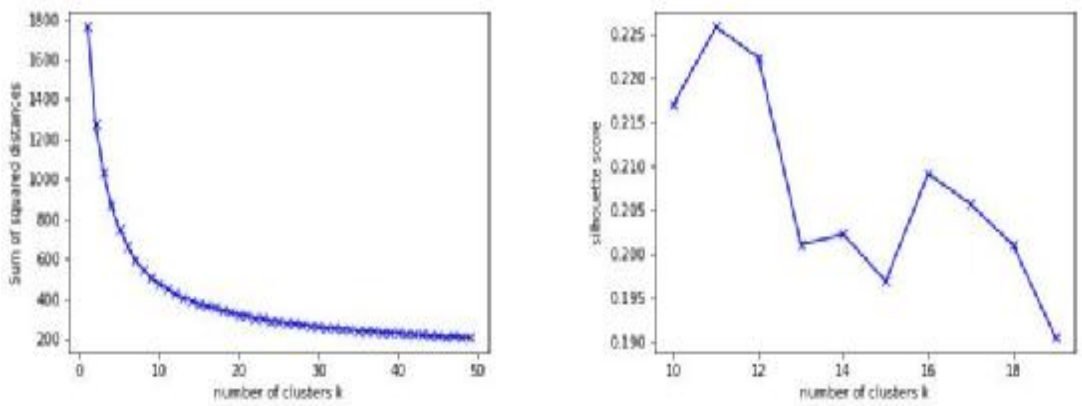

Fig. 2. a) The elbow inertia values versus the number of clusters, b) the silhouette score versus the number of clusters.

Table 1. The clustering results from k-means.

Cluster 1	ACETAMIDE, ACETONITRILE, AMMONIA, GLYCEROL, HYDRAZINE, METHANOL, N-METHYLACETAMIDE, NITROMETHANE, PHENOL, PIPERAZINE, PYRENE
Cluster 2	1-NITROBUTANE, 1-NITROPROPANE, 2-BUTOXYETHANOL, 2-METHYLPYRIDINE, 2-METHYLTHIOPHENE, 2-NITROPROPANE, 2-PHENYLETHANOL, 3-METHYLPYRIDINE, 4-METHYLPYRIDINE, ACENAPHTHENE, ACETALDEHYDE, ACETONE, ACETONITRILE, ANILINE, ANTHRACENE, BENZALDEHYDE, BENZONITRILE, BROMOBENZENE, CHLOROBENZENE, CHLOROFORM, CYCLOHEXANOL, CYCLOHEXANONE, CYCLOPENTANONE, DIBROMOMETHANE, DICHLOROMETHANE, DIIODOMETHANE, ETHANOL, FORMALDEHYDE, IODOBENZENE, M-CRESOL, METHANOL, MORPHOLINE, NAPHTHALENE, NITROBENZENE, NITROMETHANE, O-CRESOL, P-CRESOL, PHENANTHRENE, PIPERIDINE, PYRIDINE, PYRROLE, PYRROLIDINE, QUINOLINE, QUINONE, TETRAHYDROFURAN, THIOPHENE
Cluster 3	1-NITROBUTANE, 1-NITROPROPANE, 2-BUTOXYETHANOL, 2-ETHOXYETHANOL, 2-METHOXYETHANOL, 2-METHYLPYRIDINE, 2-NITROPROPANE, 2-PHENYLETHANOL, 3-METHYLPYRIDINE, 4-METHYLPYRIDINE, ACENAPHTHENE, ACETALDEHYDE, ACETONE, ACETONITRILE, ANILINE, ANTHRACENE, BENZALDEHYDE, BENZONITRILE, CHLOROFORM, CYCLOHEXANOL, CYCLOHEXANONE, CYCLOPENTANONE, DIBROMOMETHANE, DIIODOMETHANE, ETHANOL, FORMALDEHYDE, M-CRESOL, METHANOL, MORPHOLINE, NITROBENZENE, NITROMETHANE, O-CRESOL, P-CRESOL, PHENANTHRENE, PHENOL, PIPERAZINE, PIPERIDINE, PYRENE, PYRIDINE, PYRROLE, PYRROLIDINE, QUINOLINE, QUINONE, SULFOLANE, TETRAHYDROFURAN
Cluster 4	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 1-NITROBUTANE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACENAPHTHENE, ANTHRACENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CHLOROFORM, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, CYCLOPENTENE, ETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PHENANTHRENE, PROPANE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE
Cluster 5	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACENAPHTHENE, ANTHRACENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CYCLOHEXANE, CYCLOPENTANE, CYCLOPENTENE, DICHLOROMETHANE, ETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, FORMALDEHYDE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PHENANTHRENE,

	PROPANE, PYRENE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE
Cluster 6	1-NITROBUTANE, 1-NITROPROPANE, 2-BUTOXYETHANOL, 2-ETHOXYETHANOL, 2-METHOXYETHANOL, 2-METHYLPYRIDINE, 2-NITROPROPANE, 2-PHENYLETHANOL, 3-METHYLPYRIDINE, 4-METHYLPYRIDINE, ACENAPHTHENE, ACETALDEHYDE, ACETONE, ACETONITRILE, ANILINE, ANTHRACENE, BENZALDEHYDE, BENZONITRILE, BROMOBENZENE, CHLOROFORM, CYCLOHEXANOL, CYCLOHEXANONE, CYCLOPENTANONE, DIBROMOMETHANE, DICHLOROMETHANE, DIIODOMETHANE, ETHANOL, FORMALDEHYDE, IODOBENZENE, M-CRESOL, METHANOL, MORPHOLINE, NITROBENZENE, NITROMETHANE, O-CRESOL, P-CRESOL, PHENANTHRENE, PHENOL, PIPERIDINE, PYRIDINE, PYRROLE, PYRROLIDINE, QUINOLINE, QUINONE, TETRAHYDROFURAN, THIOPHENE
Cluster 7	2-BUTOXYETHANOL, 2-ETHOXYETHANOL, 2-METHOXYETHANOL, 2-PHENYLETHANOL, ACETAMIDE, AMMONIA, ANILINE, ETHANOL, GLYCEROL, HYDRAZINE, M-CRESOL, METHANOL, N-METHYLACETAMIDE, P-CRESOL, PHENOL, PIPERAZINE, PYRENE, PYRROLE, QUINONE, SULFOLANE
Cluster 8	2-METHYLHEXANE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACETAMIDE, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, ETHANE, HEXACHLOROBENZENE, ISOBUTANE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, OCTAFLUOROCYCLOBUTANE, PROPANE
Cluster 9	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 1-NITROBUTANE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 3-METHYLHEXANE, ACENAPHTHENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, CYCLOPENTENE, DICHLOROMETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PROPANE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE
Cluster 10	1-BROMOHEPTANE, 1-ETHYLNAPHTHALENE, 2-METHYLHEXANE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACENAPHTHENE, ANTHRACENE, CYCLOHEXANE, CYCLOPENTANE, ETHANE, GLYCEROL, HEXACHLOROBENZENE, HEXACHLOROETHANE, ISOBUTANE, ISOBUTYLBENZENE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, OCTAFLUOROCYCLOBUTANE, PHENANTHRENE, PYRENE, SEC-BUTYLBENZENE, TERT-BUTYLBENZENE
Cluster 11	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 1-NITROBUTANE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 2-NITROPROPANE, 3-METHYLHEXANE, ACENAPHTHENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CHLOROFORM, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, CYCLOPENTENE, DICHLOROMETHANE, ETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PROPANE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE

Table 2. The clustering results from birch.

Cluster 1	ACETAMIDE, ACETONITRILE, AMMONIA, GLYCEROL, HYDRAZINE, METHANOL, N-METHYLACETAMIDE, NITROMETHANE, PHENOL, PIPERAZINE, PYRENE, PYRROLE, SULFOLANE
Cluster 2	1-NITROBUTANE, 1-NITROPROPANE, 2-BUTOXYETHANOL, 2-METHYLPYRIDINE, 2-METHYLTHIOPHENE, 2-NITROPROPANE, 2-PHENYLETHANOL, 3-METHYLPYRIDINE, 4-METHYLPYRIDINE, ACENAPHTHENE, ACETALDEHYDE, ACETONE, ACETONITRILE, ANILINE, ANTHRACENE, BENZALDEHYDE, BENZONITRILE, BROMOBENZENE, CHLOROBENZENE, CHLOROFORM, CYCLOHEXANOL, CYCLOHEXANONE, CYCLOPENTANONE, DIBROMOMETHANE, DICHLOROMETHANE, DIIODOMETHANE, ETHANOL, FORMALDEHYDE, IODOBENZENE, M-CRESOL, METHANOL, MORPHOLINE, NAPHTHALENE, NITROBENZENE, NITROMETHANE, O-CRESOL, P-CRESOL, PHENANTHRENE, PIPERIDINE, PYRIDINE, PYRROLE, PYRROLIDINE, QUINOLINE, QUINONE, TETRAHYDROFURAN, THIOPHENE
Cluster 3	1-NITROBUTANE, 1-NITROPROPANE, 2-BUTOXYETHANOL, 2-ETHOXYETHANOL, 2-METHOXYETHANOL, 2-METHYLPYRIDINE, 2-NITROPROPANE, 2-PHENYLETHANOL, 3-METHYLPYRIDINE, 4-METHYLPYRIDINE, ACENAPHTHENE, ACETALDEHYDE, ACETONE, ACETONITRILE, ANILINE, ANTHRACENE, BENZALDEHYDE, BENZONITRILE, CHLOROFORM, CYCLOHEXANOL, CYCLOHEXANONE, CYCLOPENTANONE, DIBROMOMETHANE, DIIODOMETHANE, ETHANOL, FORMALDEHYDE, M-CRESOL, METHANOL, MORPHOLINE, NITROBENZENE, NITROMETHANE, O-CRESOL, P-CRESOL, PHENANTHRENE, PHENOL, PIPERAZINE, PIPERIDINE, PYRENE, PYRIDINE, PYRROLE, PYRROLIDINE, QUINOLINE, QUINONE, SULFOLANE, TETRAHYDROFURAN
Cluster 4	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 1-NITROBUTANE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACENAPHTHENE, ANTHRACENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CHLOROFORM, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, CYCLOPENTENE, ETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PHENANTHRENE, PROPANE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE
Cluster 5	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACENAPHTHENE, ANTHRACENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CYCLOHEXANE, CYCLOPENTANE, CYCLOPENTENE, DICHLOROMETHANE, ETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, FORMALDEHYDE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PHENANTHRENE, PROPANE, PYRENE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE

Cluster 6	1-NITROBUTANE, 1-NITROPROPANE, 2-BUTOXYETHANOL, 2-ETHOXYETHANOL, 2-METHOXYETHANOL, 2-METHYLPYRIDINE, 2-NITROPROPANE, 2-PHENYLETHANOL, 3-METHYLPYRIDINE, 4-METHYLPYRIDINE, ACENAPHTHENE, ACETALDEHYDE, ACETONE, ACETONITRILE, ANILINE, ANTHRACENE, BENZALDEHYDE, BENZONITRILE, BROMOBENZENE, CHLOROFORM, CYCLOHEXANOL, CYCLOHEXANONE, CYCLOPENTANONE, DIBROMOMETHANE, DICHLOROMETHANE, DIIODOMETHANE, ETHANOL, FORMALDEHYDE, IODOBENZENE, M-CRESOL, METHANOL, MORPHOLINE, NITROBENZENE, NITROMETHANE, O-CRESOL, P-CRESOL, PHENANTHRENE, PHENOL, PIPERIDINE, PYRIDINE, PYRROLE, PYRROLIDINE, QUINOLINE, QUINONE, TETRAHYDROFURAN, THIOPHENE
Cluster 7	2-BUTOXYETHANOL, 2-ETHOXYETHANOL, 2-METHOXYETHANOL, 2-PHENYLETHANOL, ACETAMIDE, AMMONIA, ANILINE, ETHANOL, GLYCEROL, HYDRAZINE, M-CRESOL, METHANOL, N-METHYLACETAMIDE, P-CRESOL, PHENOL, PIPERAZINE, PYRENE, PYRROLE, QUINONE, SULFOLANE
Cluster 8	2-METHYLHEXANE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACETAMIDE, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, ETHANE, HEXACHLOROBENZENE, ISOBUTANE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, OCTAFLUOROCYCLOBUTANE, PROPANE
Cluster 9	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 1-NITROBUTANE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 3-METHYLHEXANE, ACENAPHTHENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, CYCLOPENTENE, DICHLOROMETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PROPANE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE
Cluster 10	1-BROMOHEPTANE, 1-ETHYLNAPHTHALENE, 2-METHYLHEXANE, 3-METHYLHEPTANE, 3-METHYLHEXANE, ACENAPHTHENE, ANTHRACENE, CYCLOHEXANE, CYCLOPENTANE, ETHANE, GLYCEROL, HEXACHLOROBENZENE, HEXACHLOROETHANE, ISOBUTANE, ISOBUTYLBENZENE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, OCTAFLUOROCYCLOBUTANE, PHENANTHRENE, PYRENE, SEC-BUTYLBENZENE, TERT-BUTYLBENZENE
Cluster 11	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 1-NITROBUTANE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLTHIOPHENE, 2-NITROPROPANE, 3-METHYLHEXANE, ACENAPHTHENE, BENZENE, BROMOBENZENE, CHLOROBENZENE, CHLOROFORM, CYCLOHEXANE, CYCLOHEXENE, CYCLOPENTANE, CYCLOPENTENE, DICHLOROMETHANE, ETHANE, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, HEXACHLOROBENZENE, HEXACHLOROETHANE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-XYLENE, METHANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, NAPHTHALENE, O-XYLENE, OCTAFLUOROCYCLOBUTANE, P-XYLENE, PROPANE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, THIOPHENE, TOLUENE

As could be seen, the number of clusters has been iterated up to 50 clusters to obtain the elbow plot. Consequently, the silhouette scores have been evaluated and plotted in the range that the elbow shows a curve. Accordingly, 11 clusters have been selected to be the optimal number of clusters for the evaluated dataset with K-means and Birch clustering. Briefly, the optimal number of the clusters should be considered as a breakpoint in the elbow while it is a maximum in the silhouette plot. The results of these methods have been evaluated and collected in Tables 1 and 2. There is a little difference between the two methods that could be seen in cluster 1 of the two methods. Also, the other clusters are identical with little difference in some cases. It means the optimum number of clusters is working very well for the utilized methods.

Generally, the first recommended materials for the evaluated clusters are from the same family for example in cluster 3 by K-means 1-NITROBUTANE and 1-NITROPROPANE are suggested as alternatives that are identical with cluster 3 of the Birch method. It means these methods give the best alternatives first, and the close results are given next. The initial and final materials in each cluster should be the best alternatives for the previous and next material. The studied materials are very well-known organic materials. At this point, the supporting information data could be used to compare the activity coefficient data of the recommended materials. It would be easier to compare the material recommended materials instead of the whole dataset.

Consider the Birch methods cluster 11 final materials, (TERT-BUTYLBENZENE, THIOPHENE, TOLUENE) where thiophene would be a good alternative for toluene and tert-butylbenzene. These materials are all aromatic solvents with similar properties, and their activity coefficient plots have been given in benzene in Figure 3 to compare the results. As could be seen the activity coefficient of the Benzene is identical for these materials.

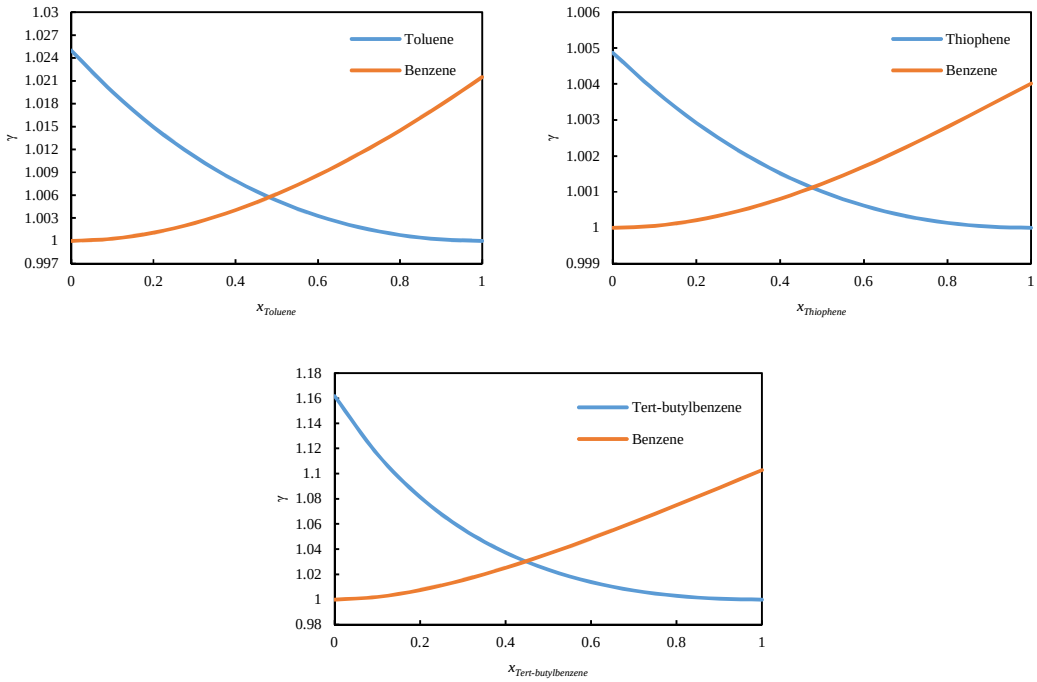

Fig. 3. The activity coefficient of the binary mixtures of final components (toluene, thiophene, and tert-butylbenzene) of cluster 11 with benzene at 298.15 K from COSMO-SAC calculation.

The centroid methods are efficient in sorting various alternatives from the trained dataset. However, the recommended materials are a different variety of components. The density-based methods for the clustering are working in a different way rather than the centroid-based clustering. The dbscan algorithm is one of the well-known density-based clustering methods. There are two main

parameters for optimum results of the dbscan namely epsilon, the distance parameters between the data points in the dataset, and a minimum number of samples or nearest neighbor points to consider a region as a dense region [17]. This method works based on the density of the accumulated data in the dataset determines the number of clusters based on the radius of the accumulated data in a mean dataset plot and limits the data in the range of the predetermined radius.

Determination of the epsilon and the minimum number of samples could be carried out using the k-nearest neighbors' module from the scikit learn library. An iteration on the number of neighbors up to 20 neighbor points has been carried out to achieve the optimum epsilon and the minimum number of samples based on the mean distance for k-nearest neighbors. The results are given in Figure 4 for different numbers of the neighbors and the mean value of the points.

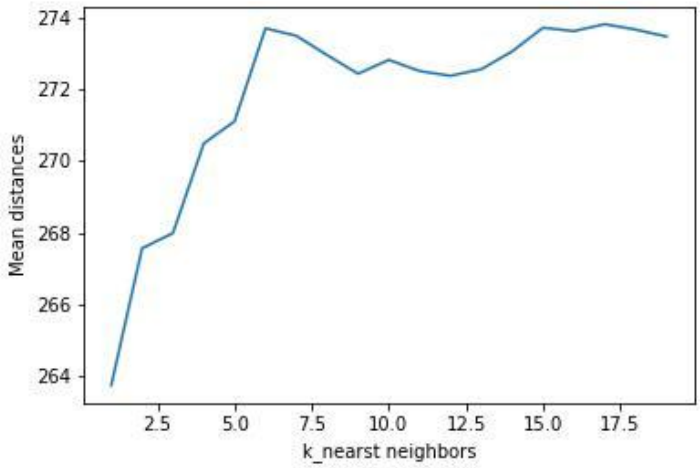

Fig. 4. The mean distance values versus k-nearest neighbors plot for the merged FreeSolv and COSMO-SAC dataset for 96 common organic materials between the two datasets.

The optimum Euclidean distance has been evaluated at about 0.08 for this dataset for dbscan clustering with 7 nearest neighbors. The method is quite simple and similar to the elbow method with some differences. The optimum number of neighbors considered as the point at the mean distance value changes dramatically and a breaking point in the plot would be observed which could be seen in Figure 4. However, the epsilon value would be calculated for this point with a mean Euclidean distance that is different from the information illustrated in Figure 4. In this respect the, string type data were removed from the dataset and the Euclidean distance was calculated for the dataset reported final value was obtained.

The obtained parameters were used for the dbscan clustering and the obtained clusters and out layers have been collected in Tables 3 and 4 for the clusters and out layers, respectively. As could be seen 5 clusters and 12 out layers were obtained for the trained dataset. The results are quite better than the centroid methods since the clusters contain materials with similar chemical families for example cluster 5 including CHLOROFORM, DIBROMOMETHANE, DICHLOROMETHANE, and DIIODOMETHANE are halomethane families that the structure and chemical formula of these materials are given in Figure 5. It is a proven fact the COSMO-SAC results are affected by the chemical family [4]. Also, the out layers mean these materials have no proper alternatives in the trained dataset that should be considered unique materials.

The results of the clustering based on the COSMO-SAC data could be extended to any warehouse and chemistry or materials science lab. The results of the presented work could be used in an organic chemistry lab or chemical engineering lab limited to the materials included in the supporting information. Also, it is possible to follow the instructions presented in this work to implement the workflow to larger datasets for example the petrochemical industries for different purposes such as the alternation of a toxic solvent with a proper and greener one.

Table 3. The clusters of the dbscan clustering.

Cluster 1	3-METHYLHEPTANE, ACETAMIDE, CYCLOHEXENE, HEXACHLOROBENZENE, HEXACHLOROETHANE, METHANE, OCTAFLUOROCYCLOBUTANE
Cluster 2	1-BROMOBUTANE, 1-BROMOHEPTANE, 1-BROMOPROPANE, 1-CHLOROBUTANE, 1-ETHYLNAPHTHALENE, 1-METHYLNAPHTHALENE, 1-NITROBUTANE, 1-NITROPROPANE, 2-BROMOPROPANE, 2-CHLOROBUTANE, 2-METHYLHEXANE, 2-METHYLPYRIDINE, 2-METHYLTHIOPHENE, 2-NITROPROPANE, 3-METHYLHEPTANE, 3-METHYLHEXANE, 3-METHYLPYRIDINE, 4-METHYLPYRIDINE, ACENAPHTHENE, ACETALDEHYDE, ACETONE, ACETONITRILE, ANILINE, ANTHRACENE, BENZALDEHYDE, BENZENE, BENZONITRILE, BROMOBENZENE, CHLOROBENZENE, CYCLOHEXANE, CYCLOHEXANOL, CYCLOHEXANONE, CYCLOHEXENE, CYCLOPENTANE, CYCLOPENTANONE, CYCLOPENTENE, ETHANE, ETHANOL, ETHYLBENZENE, ETHYLENE, FLUOROBENZENE, FORMALDEHYDE, INDANE, IODOBENZENE, ISOBUTANE, ISOBUTYLBENZENE, M-CRESOL, M-XYLENE, METHANE, METHANOL, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, MORPHOLINE, N-BUTANE, N-PENTANE, NAPHTHALENE, NITROBENZENE, NITROMETHANE, O-CRESOL, O-XYLENE, P-CRESOL, P-XYLENE, PHENANTHRENE, PHENOL, PIPERIDINE, PROPANE, PYRENE, PYRIDINE, PYRROLIDINE, QUINOLINE, QUINONE, SEC-BUTYLBENZENE, STYRENE, TERT-BUTYLBENZENE, TETRAHYDROFURAN, THIOPHENE, TOLUENE
Cluster 3	2-BUTOXYETHANOL, 2-ETHOXYETHANOL, 2-METHOXYETHANOL
Cluster 4	2-METHYLHEXANE, 3-METHYLHEPTANE, 3-METHYLHEXANE, CYCLOHEXANE, CYCLOPENTANE, ETHANE, ISOBUTANE, METHYLCYCLOHEXANE, METHYLCYCLOPENTANE, N-BUTANE, N-PENTANE, PROPANE
Cluster 5	CHLOROFORM, DIBROMOMETHANE, DICHLOROMETHANE, DIIODOMETHANE

Table 4. The out layers of the dbscan clustering.

Out layer 1	PIPERAZINE
Out layer 2	PYRROLE
Out layer 3	ACETAMIDE
Out layer 4	2-PHENYLETHANOL
Out layer 5	HYDRAZINE
Out layer 6	HEXACHLOROETHANE
Out layer 7	HEXACHLOROBENZENE
Out layer 8	AMMONIA
Out layer 9	GLYCEROL
Out layer 10	SULFOLANE
Out layer 11	OCTAFLUOROCYCLOBUTANE
Out layer 12	METHYLACETAMIDE

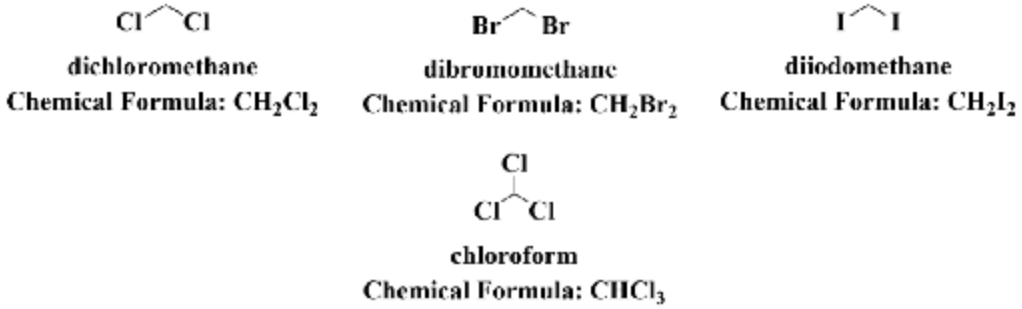

Fig. 5. The components of cluster 5 include halomethanes from dbscan clustering.

2.2. Materials and methods

Workstation

The procedures have been implemented with python in the Jupyter environment. Different two PCs with different configurations have been used to evaluate the results, and the results were identical in the two configurations which are important for the repeatability of the process. Accordingly, the FreeSolv dataset and VT2005 σ-profiles dataset has been used as initial data [9,10]. It should be noted that there were 96 exact matches according to the IUPAC names of the materials between the two datasets, and it was a limitation of this work.

The activity coefficients of 96 organic materials with different solvents such as methanol, ethanol, benzene, toluene, and water in full range composition (mole fractions of solute = 0, 0.1, …, 0.9, 1) at 298.15 K have been calculated by the open-source benchmark of the COSMO-SAC implemented by Bell et al. A detailed information is available in the corresponding paper. Also, it is accessible from the GitHub repository [18]. A considerable dataset of activity coefficients about 9500 data points are evaluated in total.

Finally, the datasets including the FreeSolv dataset that contains MD results for various organic materials' free energy of hydration, the VT2005 dataset that contains σ-profiles of different materials calculated using $dmol^3$ and used as COSMO-SAC inputs, and a dataset containing different organic materials activity coefficient in the full-range concentration of the mixtures at 298.15 K. These datasets have been combined using the Pandas library of Python based on the IUPAC names of the materials.

Machine learning clustering

Different clustering methods were used to create the clusters and alternate the organic materials according to the activity coefficient of the 96 materials in different solvents. The Gower module has been used to convert the string data to a numerical value. The elbow method and silhouette methods were used for the determination of the parameters in the centroid-based clustering methods [19,20]. Also, the k-nearest neighbor's method was used for the determination of the parameters of the density-based clustering method [3,21-23]. Three clustering methods including K-means and birch as centroid-based methods and dbscan as a density-based method from the sklearn library of python were used to create the alternative recommender [24-29].

3. Conclusion

The COSMO-SAC model has been used to evaluate the full range activity coefficient for 96 organic materials that were common between the FreeSolv and VT2500 datasets by their IUPAC name. The information of the two datasets was combined and pre-processed for machine learning clustering. The combined dataset has been used to implement an alternative material recommender by different machine learning clustering methods. In this respect, centroid-based methods including K-means and Birthc methods and density-based method dbsacn were used. According to the results, the centroid-based methods recommend a variety of materials for a component by sorting it from nearest to the possible far component in the cluster radius. However, the density-based clustering recommends the alternative materials based on the k-nearest neighbors with a determined radius. The results for

density-based clustering were more promising for the presented dataset obtained from COSMO-SAC calculations.

Supporting Information

A full-range of concentration dataset of the activity coefficient of the studied materials in different solvents (water, ethanol, methanol, benzene, and toluene) (Docx).

Data and Software Availability

1. The datasets analyzed during the current study are available in the [MobleyLab/FreeSolv] repository, [https://github.com/MobleyLab/FreeSolv]
2. The free benchmark implementation of COSMO-SAC model usnistgov/COSMOSAC repository, [https://github.com/usnistgov/COSMOSAC].

References

[1]. J.F. Jenck,., F, Agterberg, . & M.J. Droescher, . Products and processes for a sustainable chemical industry. a review of achievements and prospects. *Green Chem.* **6**, (2004) 544–556.

[2]. R. Xiong., I.S. Sandler,. & R.I. Burnett,. An Improvement to COSMO-SAC for Predicting Thermodynamic Properties. *Ind. Eng. Chem. Res.* **53 (**2014**)**, 8265–8278.

[3]. S. Wang., S.I. Sandler,. & C. Chen, C. Refinement of COSMO−SAC and the Applications. *Ind. Eng. Chem. Res.* **46**, (2007) 7275–7288.

[4]. R. Fingerhut,. *et al.* Comprehensive Assessment of COSMO-SAC Models for Predictions of Fluid-Phase Equilibria. *Ind. Eng. Chem. Res.* **56**, (2017.) 9868–9884.

[5] K. El Bouchefry,. &R.S. de Souza, . Chapter 12 - Learning in Big Data: Introduction to Machine Learning. in *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (eds. Škoda, P. & Adam, F.) (2020) 225–249.

[6]. M. Meuwly,. Machine Learning for Chemical Reactions.. *Chem. Rev.* **121**, (2021)10218–10239 .

[7]. R. Ahuja., A. Chug., S. Gupta., P. Ahuja,. &S. Kohli,. Classification and Clustering Algorithms of Machine Learning with their Applications. in *Nature-Inspired Computation in Data Mining and Machine Learning* (eds. Yang, X.-S. & He, X.-S.) (2020) 225–248 (Springer International Publishing,

[8]. M. Serra-Burriel,. & C. Ames,. Machine Learning-Based Clustering Analysis: Foundational Concepts, Methods, and Applications. in *Machine Learning in Clinical Neuroscience* (eds. Staartjes, V. E., Regli, L. & Serra, C.) (2022) 91–100 (Springer International Publishing.

[9]. D.L. Mobley, & J.P . Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.* **28**, (2014) 711–720.

[10]. I. Bell,. H. *et al.* A Benchmark Open-Source Implementation of COSMO-SAC. *J. Chem. Theory Comput.* **16**, (2020) 2635–2646.

[11]. T.C. Liu,. & S.T. Lin. A new approach for developing exact local composition models for lattice fluids. *J. Taiwan Inst. Chem. Eng.* **96**, (2019) 63–73.

[12]. S. Balchandani,. & R. Singh,. Thermodynamic analysis using COSMO-RS studies of reversible ionic liquid 3-aminopropyl triethoxysilane blended with amine activators for CO2 absorption. *J. Mol. Liq.* (2021) **324**, 114713.

[13]. W. Hu. *et al.* Solubility of benorilate in twelve monosolvents: Determination, correlation and COSMO-RS analysis. *J. Chem. Thermodyn.* **152**,(2021) 106272.

[14]. M. R. Shah,. & G.D. Yadav,. Prediction of Liquid–Liquid Equilibria for Biofuel Applications by Quantum Chemical Calculations Using the Cosmo-SAC Method. *Ind. Eng. Chem. Res.* **50**, (2011) 13066–13075.

[15]. D.M. Saputra., D. Saputra,. & L.D. Oswari,. Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method.2020. in 341–346.

[16]. K. Matsuo, K. Mitsugi., A. Toyama., E. Kulla, & L.A. Barolli, Simulation System for Optimal Positions of MOAP Robots Using Elbow and Silhouette Theories: Simulation Results Considering Minimum Transmission Power of MOAP Robots.. in *Advances on Broad-Band Wireless Computing, Communication and Applications* (ed. Barolli, L.) (2022) 321–332.

[17]. S. Bhardwaj., A. Pandey, & S. Dahiya. Review based on Variations of DBSCAN algorithms.2022. in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)* 733–739 .

[18]. *COSMO-SAC*. (National Institute of Standards and Technology, 2022).

[19]. D. Marutho., S. Hendra Handaka,., E. Wijaya, , & Muljono. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News.. in *2018 International Seminar on Application for Technology of Information and Communication* (2018) 533–538.

[20]. H.B.. Zhou, & J.T. Gao. Automatic Method for Determining Cluster Number Based on Silhouette Coefficient. *Adv. Mater. Res.* **951**, (2014) 227–230.

[21]. Kareem, R.O., Kebiroglu, H. and Hamad, O.A., Investigation of Electronic and Spectroscopic Properties of Phosphosilicate Glass Molecule (BioGlass 45S5) and Ti-BioGlass 45S5 by Quantum Programming. Journal of Chemistry Letters, 4(4) (2024) 200-210.

[22]. Hamad, O., Kareem, R.O. And Kaygili, O.. Density Function Theory Study Of The Physicochemical Characteristics Of 2-Nitrophenol. Journal Of Physical Chemistry And Functional Materials, 6(1), (2023) 70-76.

[23]. A. Sharma. & A. Sharma,. KNN-DBSCAN: Using k-nearest neighbor information for parameter-free density based clustering. in *2017 International*

*Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (*2017*)* 787–792 .
[24]. A. Likas., N. Vlassis,. & J. Verbeek,. The global k-means clustering algorithm.2003. *Pattern Recognit.* **36**, 451–461.
[25] Hamad, O.A., Kareem, R.O. and Omer, P.K.,. Properties, Characterization, and Application of Phthalocyanine and Metal Phthalocyanine. Journal of Chemical Reviews, 6(1).
[26]. B. Lorbeer,. *et al.* Variations on the Clustering Algorithm BIRCH.. *Big Data Res.* (2018) **11**, 44–53.
[27]. K. Khan., S,U. Rehman., K. Aziz., S. Fong. & S. Sarasvady. DBSCAN: Past, present and future. in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)* 232–238.
[28]. I.S. Hasan, A.A. Majhoo, M.H. Sami. and A.K.O. Aldulaimi., Predicting Hydration Enthalpy of Low Molecular Weight Organic Molecules using COSMO-SAC Modeling. Chemical Review and Letters, 6(1), (2022).86-94.
[29]. B.M. Ali, and M. Akkaş. Assessing the Impact of Data Sciences and Smart Technologies in Air Conditioning Project Management: A Delphi Method Analysis within the Construction Industry. Buildings, 13(10), (2023).2581.