Journal of Computer-Aided Molecular Design
https://doi.org/10.1007/s10822-022-00488-9

Examining unsupervised ensemble learning using spectroscopy data of organic compounds

Kedan He[1] · Djenerly G. Massena[1]

Received: 1 September 2022 / Accepted: 3 November 2022

Abstract

One solution to the challenge of choosing an appropriate clustering algorithm is to combine different clusterings into a single consensus clustering result, known as cluster ensemble (CE). This ensemble learning strategy can provide more robust and stable solutions across different domains and datasets. Unfortunately, not all clusterings in the ensemble contribute to the final data partition. Cluster ensemble selection (CES) aims at selecting a subset from a large library of clustering solutions to form a smaller cluster ensemble that performs as well as or better than the set of all available clustering solutions. In this paper, we investigate four CES methods for the categorization of structurally distinct organic compounds using high-dimensional IR and Raman spectroscopy data. Single quality selection (SQI) forms a subset of the ensemble by selecting the highest quality ensemble members. The Single Quality Selection (SQI) method is used with various quality indices to select subsets by including the highest quality ensemble members. The Bagging method, usually applied in supervised learning, ranks ensemble members by calculating the normalized mutual information (NMI) between ensemble members and consensus solutions generated from a randomly sampled subset of the full ensemble. The hierarchical cluster and select method (HCAS-SQI) uses the diversity matrix of ensemble members to select a diverse set of ensemble members with the highest quality. Furthermore, a combining strategy can be used to combine subsets selected using multiple quality indices (HCAS-MQI) for the refinement of clustering solutions in the ensemble. The IR + Raman hybrid ensemble library is created by merging two complementary “views” of the organic compounds. This inherently more diverse library gives the best full ensemble consensus results. Overall, the Bagging method is recommended because it provides the most robust results that are better than or comparable to the full ensemble consensus solutions.

Keywords Clustering ensemble · Clustering ensemble selection · Bagging · Hierarchical cluster and selection · Normalized mutual information · Consensus function

Abbreviations

CE	Cluster ensemble
CES	Cluster ensemble selection
CSPA	Cluster-based Similarity Partitioning Algorithm
HBGF	Hybrid Bipartite Graph Formulation
SQI	Single Quality Index Selection
HCAS-SQI	Hierarchical Cluster and Select with Single Quality Index
HCAS-MQI	Hierarchical Cluster and Select with Multiple Quality Indices
DC	Direct combining
WC	Weighted combining
BC	Bagging combining

Introduction

In recent years, we have witnessed a dramatic explosion of chemical ‘big’ data from high-throughput screening (HTS), combinatorial synthesis, and theoretical simulations. Unsupervised machine learning, or pattern recognition, aims to extract useful information using unlabeled data, has become an indispensable tool for drug designers to mine chemical information from large compound databases. Cluster analysis is a type of unsupervised machine learning technique that divides unlabeled data objects into groups or clusters such that objects in the same cluster are more similar than

✉ Kedan He
hek@easternct.edu

[1] Department of Physical Sciences, School of Arts and Sciences, Eastern Connecticut State University, Willimantic, CT 06226, USA

Published online: 21 November 2022

objects belonging to different clusters [1, 2], and it has been applied to solve different research problems in various fields. PubChem, a public repository for information on small molecules and their biological activities, reported a structure–activity relationship (SAR) clustering approach to group non-inactive compounds according to their structural similarity and bioactivity similarity to facilitate hit exploration in the early stage of drug discovery [3]. Machine learning techniques are used increasingly in combination with quantum chemical calculations or molecular modeling to complement experiments for studying complex chemical systems. Clustering of molecular dynamics (MD) trajectories is a commonly used approach that can reveal, explain, and even predict the behavior of a particular experiment [4]. The analysis of datasets obtained from molecular simulations can often benefit from transforming the original features into a lower dimensional representation, which is referred to as dimensionality reduction [5]. The use of clustering in molecular simulations is very common because clustering can be seen as a way to compactly represent complex multidimensional probability distributions and is therefore used as a complement to other dimensionality reduction approaches. In addition, the clustering of gene expression data has been proven to be valuable in revealing the inherent structure in gene expression data, understanding gene functions, and understanding gene regulation [6].

The commonly used clustering algorithms can be divided into two classes: partitioning schemes and density-based schemes. Using a partitioning scheme, one can find any number of clusters from a set of data harvested from a uniform probability distribution. On the other hand, using a density-based scheme, one will obtain a single cluster since it corresponds to the peaks of the probability distribution. Which approach one should employ strongly depends upon the purpose of the analysis. The most commonly used partition-based clustering algorithms are *k*-means [7], spectral clustering [8], and hierarchical clustering [9]. The performance of most clustering algorithms is highly data-dependent, which means there is no single clustering method that can learn any data set according to Kleinberg's theorem [10]. In addition, several challenges are inherent to clustering algorithms [11, 12]. Different techniques discover different structures from the same set of data objects because each algorithm optimizes according to a specific criterion. A single clustering algorithm with different parameter settings can also reveal various structures in the same data set. How to validate clustering results without a labeled test dataset. Choosing a suitable clustering algorithm that can apply to all data sets is difficult. It is crucial to choose a clustering algorithm based on what is known about the dataset and what is expected about the result.

One solution to the challenge in choosing a proper clustering technique is by combining different clusterings into a single consensus clustering solution. Cluster ensemble [13] (CE), or consensus clustering, generates a consensus from multiple clustering solutions without using the base clustering algorithms or original data features [14, 15]. CE strategy is characterized by producing consensus partitions that are more robust, novel, stable, and flexible than the clusters produced by a single base clustering algorithm [16–18]. The advantage of the CE strategy is that it can handle multiple data sources or representations, with each model capturing the big picture and complementing each other. As shown in Fig. 1, CE has mainly two stages: (1) obtains a large library of clustering solutions which should be highly diverse, and then (2) combines these base partitions using a consensus

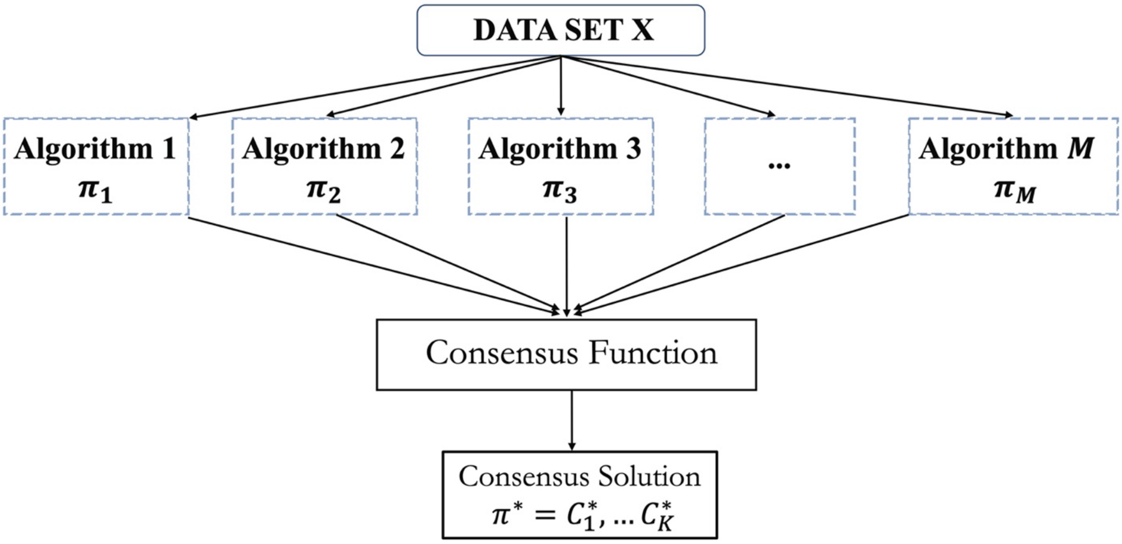


Fig. 1 Cluster ensemble framework

function. It has been shown that ensembles are more efficient when constructed from a set of clustering solutions with dissimilar errors [19, 20]. It is shown that the clustering solutions in the ensemble needed to be as diverse as possible to give more information about the underlying patterns in the data. A number of ensemble approaches have been proposed to generate multiple diverse base clustering solutions from a single dataset [21–23]. The second stage is difficult since there is no well-defined correspondence between the different clustering results. Traditionally, all clustering solutions (full ensemble library) are combined with the use of a consensus function to achieve the final solution. Various applications of CE have been implemented and deployed for drug discovery and design. For example, CE method is exploited to reduce cost and time for the process of High-Throughput Screening for combining multiple clusterings of chemical structures to enhance the ability of separating biologically active molecules from inactive ones in each cluster [24–26]. Chu et al. used CE methods on sets of chemical compounds represented by 2D fingerprints and concluded that consensus methods can outperform the standard clustering method for cheminformatics application [27].

Recently, various clustering ensemble selection (CES) strategies have been proposed, aiming to select a subset of clustering solutions from the ensemble library whose consensus solution outperforms that of the full ensemble [28–30]. Two critical questions need to be addressed in the CES process: (1) How to measure the quality of clustering solutions in the ensemble? (2) How to select a subset of base clustering solutions with high quality and eliminate the redundant base clustering solutions? Validation indices used for measuring the quality of a data partition can be categorized into two classes: internal and external indices. The internal validation indices are based on the information intrinsic to the data to assess the goodness of the clustering structure without external data labels, which are usually used to select the best clustering algorithm to be applied or the optimal number of clusters present on a given data set. On the other hand, the external validation indices measure the similarity between the output of the clustering algorithm and the correct partition of the data set (external data labels). Note that the consensus function does not use the original data features and that the dataset is unlabeled. Thus, the quality of each clustering solution using external validation indices can be defined as the pair-wise similarity among the ensemble members or the similarity to the consensus solution of the full ensemble. Specifically, an external validation index, such as normalized mutual information (NMI) or adjusted Rand index (ARI), is used to measure the shared information of a clustering solution (ensemble member) with the full ensemble. Past research has shown that when combining clustering solutions into a final partition, diversity and quality importantly impact the ensemble performance [28, 31–33]. Another question that needs to be addressed is what level of diversity would benefit the consensus solution, and different opinions on the effects of diversity would translate into different selection strategies in the cluster ensemble.

New psychoactive substances (NPS), also known as designer drugs, are compounds that alter the molecular structure of existing controlled substances to mimic their pharmacological effects and circumvent legislation [34, 35]. New NPS are emerging at an alarming rate and often without time for adequate experimental determination of their pharmacological profile. By definition, designer drugs are made up of chemical combinations that we have not seen before. They almost never match traditional databases, and chemists often don't know what they are looking for. In traditional drug testing, such as infrared (IR) and Raman, if a sample does not match any known substance, it does not yield a positive identification [36]. An unknown sample is assigned to a given class or category by leverages information extracted from training samples using pattern recognition techniques [37]. Most mature pattern recognition techniques are based on supervised learning such as partial least squares discriminate analysis (PLS-DA) and linear discriminant analysis (LDA) for classification, and calibration using partial least squares regression (PLSR) [37, 38]. However, unsupervised learning techniques (e.g., clustering analysis) have not reached the same level of maturity in chemometric analysis. Cluster analysis is particularly useful when the class structure of the data varies over time, or where the cost of acquiring classified (labeled) samples might be too costly to make it feasible to obtain the large data sets required for some supervised learning techniques, especially in the case of spectroscopic data [37].

The majority of the papers reporting the application of cluster analysis to spectroscopy data focused on demonstrating that an analytical testing technique, such as FTIR spectroscopy, paired with cluster analysis, can discriminate between different classes of materials. Some papers included a comparison of multiple clustering techniques, and a few presented an evaluation of new clustering algorithms [39–41]. There are scarce investigations of CES using spectroscopy data in pharmaceutical or forensic analysis of organic compounds. Designing a CES method that yields the best results is not trivial (if possible) considering the variety of options available. We report a comparative study of CES workflows for clustering IR and Raman spectroscopy data, with the aim of ensuring rigor and validity for future practitioners performing cluster analysis. We show techniques found in the contemporary pattern recognition and machine learning literature include similarity measures used for clustering, the clustering algorithm itself, how to choose the number of clusters, and how to evaluate and quantify the results. Furthermore, the workflow reported in this study can be applied to other datasets beyond spectroscopy data, since CE only need to access clustering solutions rather than original data.

The rest of this paper is arranged as follows: Section II presents the related works on CES strategies. Section III descries the different CES methods to be investigated in this paper. Section IV describes the series of experiments to evaluate the performance different CES strategies. Section V is the results and discussion. Section VI is the final conclusion.

Related work

Given a dataset X with N data points $\left\{x_1,\ldots,x_N\right\}$, where each data point $x_p \in X$ is represented by a vector of D attribute values (or features), i.e., $x_p = (x_{p,1},\ldots,x_{p,D})$. Let $\Pi = \left\{\pi_1,\ldots,\pi_M\right\}$ be an ensemble library with M clustering solutions, each of which is referred to as an 'ensemble member'. Each clustering solution $\pi_i \in \Pi$ is represented as an N -dimensional vector, which denotes a set of cluster labels $\pi_i = \left\{C_1^i, C_2^i \ldots, C_{k_i}^i\right\}$ of N data points, where k_i is the number of clusters in the i-th run of the clustering. For each $x_p \in X$, $\pi_{i,p}$ denotes the cluster label of data point x_p in the i-th base clustering. The cluster ensemble selection problem is to find a new subset $\Pi^S = \{\pi_1,\ldots,\pi_J\}$ where $\pi_j \in \Pi$ and $J \le M$, then a consensus function generates the final consensus partition $\pi^* = C_1^*, \ldots C_K^*$, where K denotes the number of clusters in the final clustering partition result of a dataset X that summarizes the information from the cluster ensemble Π.

CES mainly includes three steps [22, 23]: (1) Generative mechanism: approaches that can produce diverse set of clustering solutions of a given data set [18, 20, 21, 42–52]. (2) Ensemble subset selection: select ensemble members that are differ from each other (diversity) and have a satisfactory quality [28–30, 32, 53–55]. (3) Consensus function: combine multiple clustering solutions into the final data partition without gaining access to the clustering algorithms or data features [13, 42, 56–60].

There are many ways to generate a diversity collection of ensemble members for a given dataset, and there are no restrictions on how the clustering solutions must be obtained. Homogeneous ensembles are created using a single clustering algorithm and run iteratively with several sets of parameters. The non-deterministic k-means clustering algorithm is commonly used with random initialization in this approach [18, 21, 61, 62]. Since the output of the clustering algorithm depends on the initial choice of the number of clusters k, each clustering run can use a randomly selected value of k from a pre-specified interval to increase the diversity of the ensemble [63–65]. As a rule-of-thumb, the maximum number of clusters should be greater than the expected number of clusters, that can be set as $\max(k) = \sqrt{N}$. [20, 21, 44, 45] Heterogeneous ensembles are created using different clustering algorithms to introduce diversity [50–52, 66]. Since each clustering algorithm has its own advantages and disadvantages, using multiple algorithms can provide different decisions for data partitioning and complement each other. Gionis et al. [66]. used hierarchical clustering with single, average, complete, and Ward's linkage as well as k -means to generate the ensembles used in the study.

Clustering of high-dimensional data faces additional challenges, such as poor discrimination of distance [67–69], redundant features [69], and irrelevant features [70]. Studies point out that as the dimensionality increases, the relative distance between the farthest and nearest points converges to zero. Traditionally, the vibrational spectrum of a sample is matched to the corresponding experimental reference in the database based on the deviation of the peak position. With the development of data-driven methods in chemical analysis, correlation coefficients, such as Pearson and Spearman, can be used as similarity measures for spectral data to enable automated processing of large numbers of spectra [71, 72]. Studies reported by van der Spoel et al. show that Spearman's correlation can better represent the approximate matching of frequency bands, while Pearson's correlation can better represent the consistency of the most dominant feature(s), and the two measures can be used in a complementary manner [71, 72]. Using either measure, the original data set can be converted into a $N \times N$ similarity matrix and subject to further clustering analysis.

The goal of the consensus function is to explore a clustering π^* that shares the highest amount of information regarding the ensemble $\Pi = \left\{\pi_1,\ldots,\pi_M\right\}$. Different consensus functions on the same ensemble diversity can results in different consensus solutions. Consensus function based on hypergraph partitioning, such as Cluster-based Similarity Partitioning Algorithm (CSPA), hypergraph Partitioning Algorithm (HGPA), and Meta-Clustering Algorithm (MCLA) are very popular and widely used [13]. CSPA builds a similarity matrix based on the clustering solutions in the ensemble, which measure for each pair of data points the frequency of them being clustered together in the ensemble, also referred to as the co-association matrix. However, these algorithms act properly just for balanced clusters, where cluster sizes are constrained to N/K. [73] Hybrid Bipartite Graph Formulation (HBGF) [74] is a graph-based hybrid method, which is introduced with the purpose to improve the previous models of CSPA and MCLA that considers only either the associations between data points or those amongst clusters.

Cluster ensembles are mainly applied to enhance the quality of single clustering results. Hence, a large library of clustering solutions is generated to form the ensemble. A more efficient consensus solution can be obtained if the ensemble members are different from each other (diversity) and have satisfactory quality [53, 75]. Especially when the ensemble size is small, combining identical clustering solutions leads to an inaccurate consensus solution [45]. In the supervised classification task, the classifiers are ranked based on their individual performance on a held-out test set, and the best ones are picked. On the contrary, in unsupervised clustering, the data sets are unlabeled, so it is impossible to estimate the quality of individual clustering solutions by computing their quality using the test set. This leads to unreliable clustering solutions in a large ensemble, so not all ensemble members are necessarily beneficial to the final consensus solution [55, 76, 77]. Most methods in existing literature work on the basis of label matching between two data partitions. Generally, when the labels of two partitions are not matched completely, then the two partitions are considered diverse. The ARI and NMI are widely used to measure the clustering solutions' diversity and quality. Diversity measures can be further divided into pair-wise and non-pair-wise fashions. Specifically, in the pair-wise diversity each ensemble member is chosen as a class label implicitly, and other ensemble members are measured by the chosen class label: $diversity\left(\pi_i, \pi_j\right) = 1 - quality(\pi_i, \pi_j)$. Based on an objective function introduced by Strehl and Ghosh [13], Lin and Fern used $SNMI\left(\pi_i, \Pi\right) = \sum_{j=1}^{M} NMI(\pi_i, \pi_j)$ that measure the information an ensemble member π_i shares with all the clustering solutions in the ensemble [28]. Naldi et al. [63] report a comparative study using different internal validation indices to select ensemble members and revealed that each index may be more suitable for a specific data conformation, on the basis of which they proposed a combination of indices in the selection process.

Only a few researches have focused on the way a subset of ensemble members must be chosen considering quality and diversity [28, 76, 78]. Hadjitodorov et al. [45] used four ARI-based diversity measures in the selection process, and the results showed that ensemble subsets with median diversity are usually significantly better than the subsets chosen at random. Fern and Lin [28] introduced the Cluster And Select (CAS) approach, which first divides all ensemble members into K groups based on their similarity, then selects the ensemble member with the highest quality from each group to be included in the ensemble subset for the final consensus solution. In this approach, the size of ensemble subset is arbitrarily determined. Based on the CAS framework, Akbari et al. [78] proposed Hierarchical cluster ensemble selection (HCES) that identifies the subsets of ensemble members considering both diversity and quality using hierarchical clustering techniques with different linkage methods. On the other hand, Jia et al. [65] present the Selective Spectral Clustering Ensemble (SELSCE) by applying the bagging technique to rank and evaluate the ensemble members. In the Hybrid clustering solution selection strategy (HCSS) proposed by Yu et al. [79], the problem of selection of ensemble members is converted to feature selection. They applied four feature selection strategies to create four ensemble subsets. After that a merged subset is selected on the basis of a weighting consensus function. Ma et al. [80] also used different combination strategies that combine different subsets obtained by several selection algorithms. A consensus matrix is then constructed and a normalized cut algorithm is then applied as the consensus function.

Cluster ensemble selection methods

The CES is divided into three stages: the generation of the base clustering ensemble library Π, the selection of the optimal ensemble subset Π^S, and the aggregated results using the consensus function with the ensemble subset. Table 1 provides a concise summary of the four CES methods (*Algorithm b-e*) evaluated in this paper, and a flow chart is also given in Fig. 2. Note that the performance of CES methods is evaluated by comparing to two references. It is reported in the literature that the aggregated consensus results should outperform the full ensemble. Therefore, the average performance of all ensemble members is used as Reference 1. *Algorithm a* is the traditional cluster ensemble approach in which the consensus solution is computed using the full ensemble and is referred as Reference 2.

Quality and diversity measures of clustering solution

In this work, we chose four quality indices to evaluate the ensemble members' quality and to enable the selection of the ensemble subset to generate the consensus solutions, individually (SQI method) and combined (HCAS-MQI method). The external validation index NMI is adopted to measure the similarity between a pair of clustering solutions. Let π_i and π_j $(i, j \in (1, \ldots, M))$ be two ensemble members with k_i clusters $C_i = \{C_1^i, C_2^i \ldots, C_{k_i}^i\}$ and k_j clusters $C_j = \{C_1^j, C_2^j \ldots, C_{k_j}^j\}$, respectively. NMI is defined as:

Table 1 CES Methods Evaluated in this paper

Reference 1. The average performance of all ensemble members

(Reference 2) ***Algorithm a.***

Traditional Cluster Ensemble: compute consensus solution using the full ensemble Π

Algorithm b. **Single Quality Index Selection (SQI):**

Ensemble members are ranked according to a single quality index:

Q_1-SNMI

Q_2-Silhouette score

Q_3-Calinski Harabasz

Q_4-Davies Bouldin

Algorithm c. **Bagging Selection:**

Ensemble members are ranked using the bagging technique by comparing to the consensus solution of bootstrapped subsets

Algorithm d. **Hierarchical Cluster and Select with Single Quality Index (HCAS-SQI):**

Consider both ensemble members' diversity and ranked according to a single quality index

Algorithm e. **Hierarchical Cluster and Select with Multiple Quality Indices (HCAS-MQI):**

Consider ensemble members' diversity and combined rankings of multiple quality indices using a combining strategy

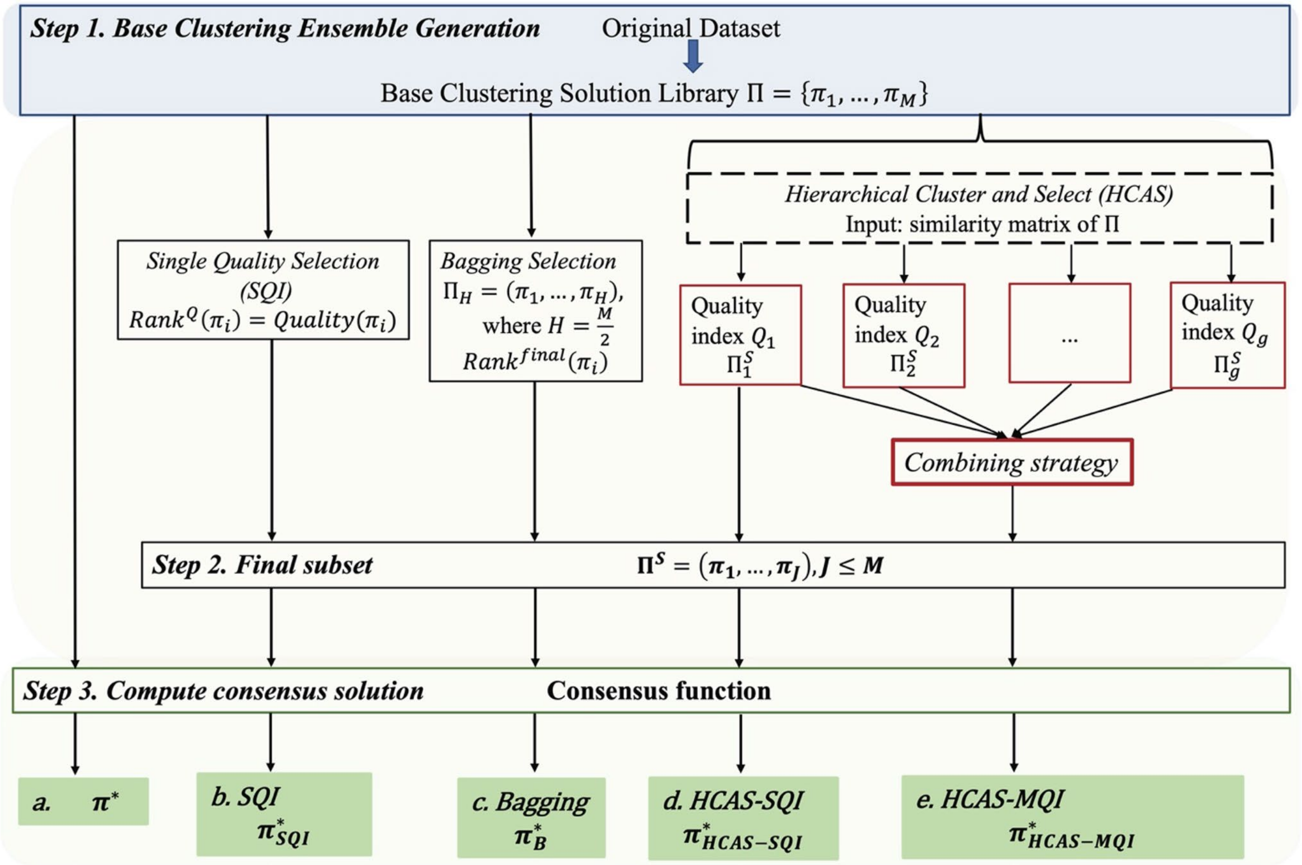

Step 1. Base Clustering Ensemble Generation
Original Dataset
Base Clustering Solution Library $\Pi = \{\pi_1, \ldots, \pi_M\}$
Hierarchical Cluster and Select (HCAS)
Input: similarity matrix of Π
Single Quality Selection (SQI)
$Rank^Q(\pi_i) = Quality(\pi_i)$
Bagging Selection
$\Pi_H = (\pi_1, \ldots, \pi_H)$, where $H = \frac{M}{2}$
$Rank^{final}(\pi_i)$
Quality index Q_1 Π_1^S
Quality index Q_2 Π_2^S
…
Quality index Q_g Π_g^S
Combining strategy
Step 2. Final subset
$\Pi^S = (\pi_1, \ldots, \pi_J), J \le M$
Step 3. Compute consensus solution
Consensus function
a. π^*
b. SQI π^*_{SQI}
c. Bagging π^*_B
d. HCAS-SQI $\pi^*_{HCAS-SQI}$
e. HCAS-MQI $\pi^*_{HCAS-MQI}$

Fig. 2 Flowchart of cluster ensemble selection methods. *Algorithm a* Consensus solution computed using the full ensemble Π. *Algorithm b–e* are CES methods summarized in Table 1 and described in " CES-single quality index selection (SQI)"–"CES-Hierarchical cluster and selection using multiple quality indices (HCAS-SQI)" sections

$$NMI\left(\pi_i,\pi_j\right) = \mathrm{NMI}\left(C_i,C_j\right) = \frac{I\left(C_i,C_j\right)}{\sqrt{\left[H\left(C_i\right),H\left(C_j\right)\right]}}, \quad (1)$$

Mutual information $I(C_i, C_j)$ is given as $H\left(C_i\right) - H\left(C_i|C_j\right)$. $H(C)$ is the Shannon entropy of C, and $H\left(C_i|C_j\right)$ is the conditional entropy of C_i given C_j. NMI = 0 mean two partitions contain no information about one another, whereases NMI = 1 indicates two partitions contain perfect information about one another. The first quality index (Q_1-SNMI) is then defined as [28]:

$$Q_1\left(\pi_i\right) = \frac{1}{M}\sum_{m=1}^{M} NMI\left(\pi_i,\pi_m\right) \quad (2)$$

Intuitively, an ensemble member π_i maximizing Q_1 maximizes the information it shares with all the members in the ensemble, thus can be considered to best capture the general trend contained in the ensemble.

The Silhouette index (Q_2-Silhouette) assesses how well each data point x_p belongs to its assigned cluster C_p. [81] Each individual Silhouette number is evaluated as:

$$s^{(i)} = \frac{\overline{x}_{C_q}^{(i)} - \overline{x}_{C_p}^{(i)}}{max(\overline{x}_{C_q}^{(i)}, \overline{x}_{C_p}^{(i)})} \quad (3)$$

where C_q represents the closest cluster to each C_p. At each depth on the dendrogram, the average silhouette number is evaluated across all samples and calculated as:

$$Q_2\left(\pi_i\right) = \frac{1}{N}\sum_{i=1}^{N} s^{(i)} \quad (4)$$

The Calinski Harabasz index [82] (Variance Ratio Criterion) (Q_3-Calinski Harabasz) evaluates the quality of a data partition as:

$$Q_3\left(\pi_i\right) = \frac{trace(\mathbf{B})}{trace(\mathbf{W})} \times \frac{n-k}{k-1} \quad (5)$$

where **W** and **B** are the within-group and between-group dispersion matrices. The normalization term $(n-k)/(k-1)$ prevents this ratio to increase monotonically with the number of clusters.

Davies Bouldin index [83] (Q_4-Davies Bouldin) also based on a ratio involving within-group and between-group distances as follows:

$$Q_4\left(\pi_i\right) = \frac{1}{k}\sum_{l=1}^{k} D_l \quad (6)$$

where $D_l = max_{l \neq m}\{D_{l,m}\}$, term $D_{l,m}$ is the within-to-between cluster spread for the lth and mth clusters, hence D_l represents the worst case within-to-between cluster spread involving the lth cluster. Hence, good data partition composed of compact and separated clusters and distinguished by small values of Davies Bouldin index, and the minimum value is zero.

CES-single quality index selection (SQI)

In SQI method (*Algorithm b*), given Π, the ensemble members are ranked according to the chosen quality index and the selected ensemble subset is formed with the ensemble members with the highest quality. Noted that Q_4-Davies Bouldin distinguished better partitions by smaller values, the ensemble members are sorted in ascending order instead. This method does not consider the diversity of ensemble, hence redundant ensemble members can be included.

CES-bagging selection

The bagging technique, usually applied in supervised learning, can be used to evaluate the quality of ensemble members and does not require the use of external ground truth labels. Specifically, part of the ensemble is randomly sampled to get a consensus result and then compute the NMI between the consensus results and the ensemble members. Finally, the ensemble members are ranked by aggregating multiple NMI values. Given T rankings of M members in Π, a combination function is defined as a function mapping the T rankings of original Π members into a single combined ranking:

$$Com : \{RC^t(t = (1, T)\} \rightarrow RC^{final}, \quad (7)$$

where $RC^t = (Rank^t\left(\pi_1\right), Rank^t\left(\pi_1\right), \dots, Rank^t\left(\pi_M\right))$, and $Rank^t\left(\pi_i\right)$ is the rank of the ensemble member π_i in the ranking solution RC^t where $t = (1, \dots, T)$. The final ranking is the average ranking of all ranking solutions defined as:

$$RC^{final} = \frac{\sum_{t=1}^{T} RC^t}{T}, \quad (8)$$

The Bagging selection algorithm (*Algorithm c*) is shown in Fig. 3 and as follows:

```
*Algorithm c.* Bagging Selection
Input: $\Pi = \{\pi_1,\ldots,\pi_M\}$// the full ensemble of base clustering library

Step 1: Compute the ranking of selected ensemble members in $\Pi_{DC}^{S}$.
        Repeat
        For $t = (1, T)$:
        $\Pi^t \leftarrow$ randomly select $(\frac{H}{2})$ ensemble members from $\Pi_{DC}^{S}$ with replacement,
        $\pi'^{,t} \leftarrow$ obtaining a consensus partition using $\Pi^t$,
        $RC^t \leftarrow$ compute the similarity measure (NMI) between the consensus partition and the ensemble members in $\Pi_{DC}^{S}$,
        End
Step 2: $RC^{final} \leftarrow$ compute the final ranking using $T$ ranking solutions of ensemble members in $\Pi_{DC}^{S}$.
Step 3: Sort $RC^{final}$ in descending order.
Step 4: Select the first $J^*$ ensemble members as the reduced subset.

Output: $\Pi_{BC}^{S} = \{\pi_1,\ldots,\pi_{J^*}\}$//Final reduced subset
```

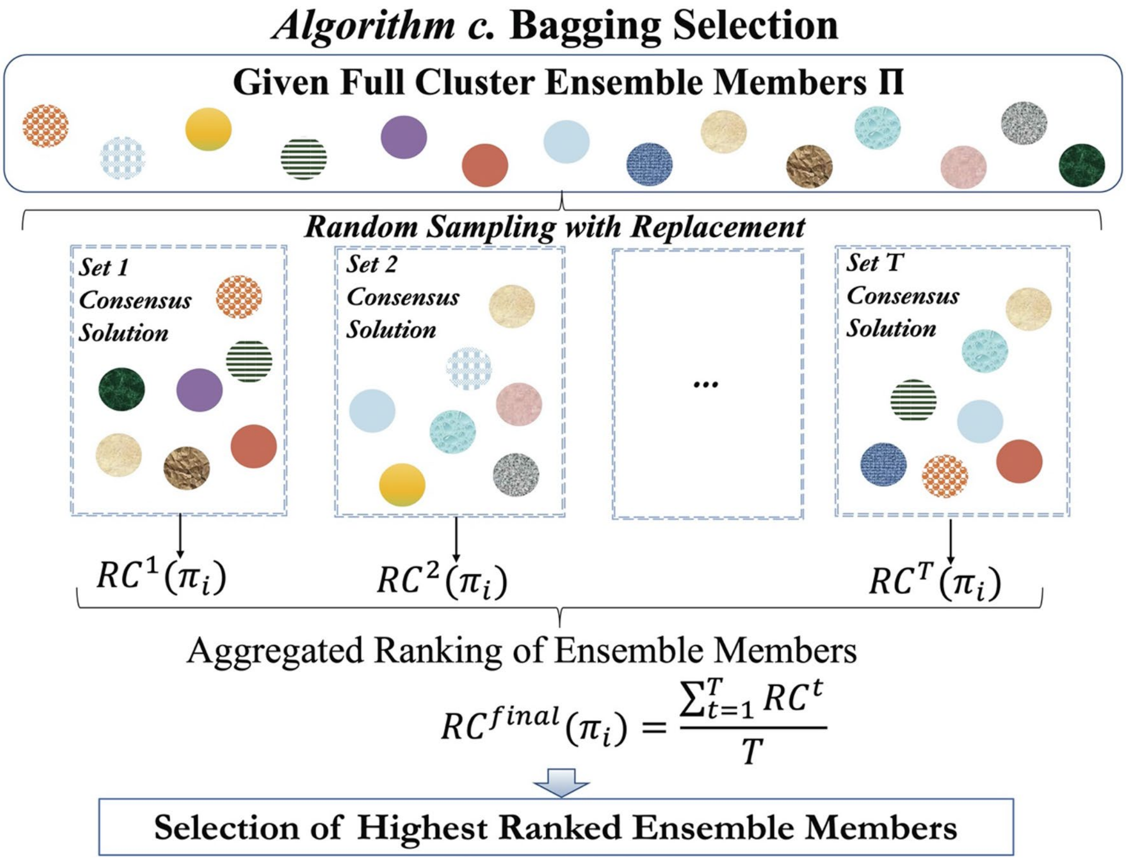
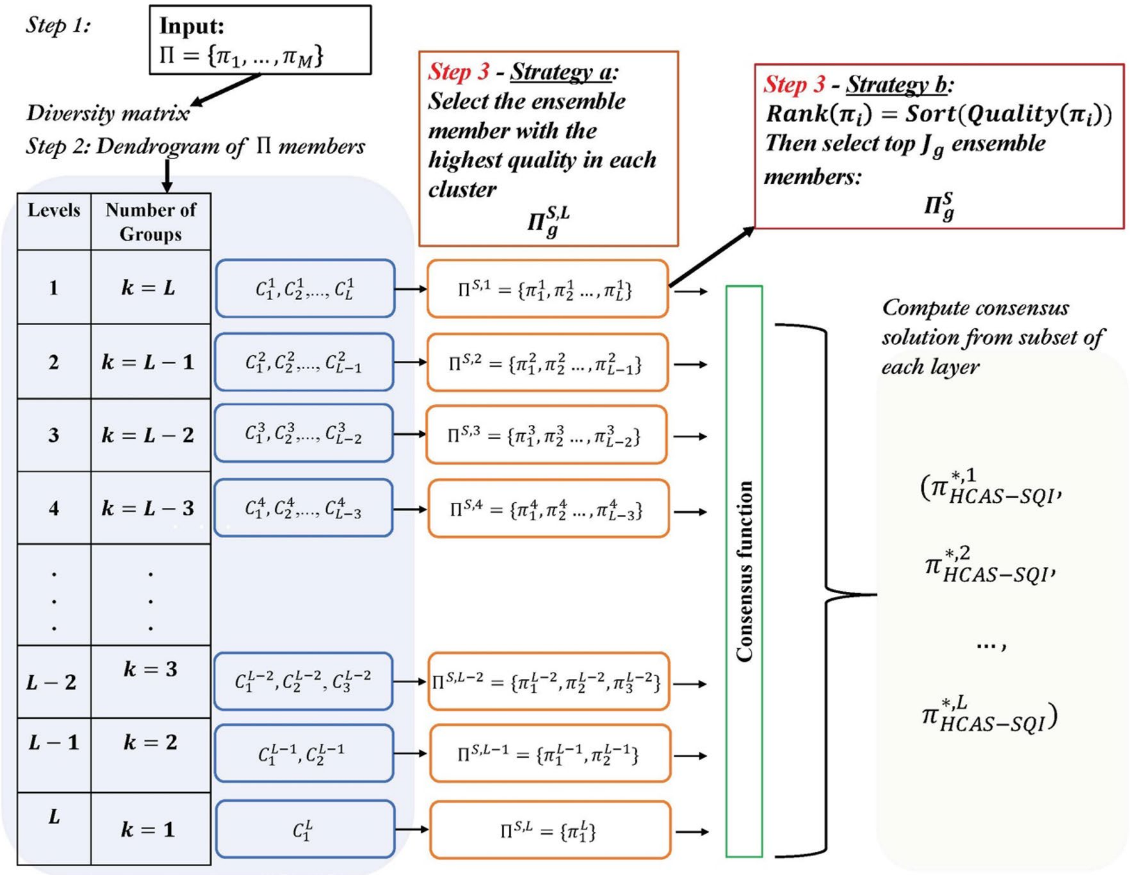

Algorithm c. Bagging Selection
Given Full Cluster Ensemble Members Π
Random Sampling with Replacement
Set 1 Consensus Solution
Set 2 Consensus Solution
...
Set T Consensus Solution
$RC^1(\pi_i)$
$RC^2(\pi_i)$
$RC^T(\pi_i)$
Aggregated Ranking of Ensemble Members
$RC^{final}(\pi_i) = \frac{\sum_{t=1}^{T} RC^t}{T}$
Selection of Highest Ranked Ensemble Members

Fig. 3 Flowchart of bagging selection method (*Algorithm c*)


Step 1:
Input: $\Pi = \{\pi_1, \ldots, \pi_M\}$
Diversity matrix
Step 2: Dendrogram of Π members
$C_1^1, C_2^1, \ldots, C_L^1$
$C_1^2, C_2^2, \ldots, C_{L-1}^2$
$C_1^3, C_2^3, \ldots, C_{L-2}^3$
$C_1^4, C_2^4, \ldots, C_{L-3}^4$
$C_1^{L-2}, C_2^{L-2}, C_3^{L-2}$
C_1^{L-1}, C_2^{L-1}
C_1^L
$\Pi^{S,1} = \{\pi_1^1, \pi_2^1 \ldots, \pi_L^1\}$
$\Pi^{S,2} = \{\pi_1^2, \pi_2^2 \ldots, \pi_{L-1}^2\}$
$\Pi^{S,3} = \{\pi_1^3, \pi_2^3 \ldots, \pi_{L-2}^3\}$
$\Pi^{S,4} = \{\pi_1^4, \pi_2^4 \ldots, \pi_{L-3}^4\}$
$\Pi^{S,L-2} = \{\pi_1^{L-2}, \pi_2^{L-2}, \pi_3^{L-2}\}$
$\Pi^{S,L-1} = \{\pi_1^{L-1}, \pi_2^{L-1}\}$
$\Pi^{S,L} = \{\pi_1^L\}$
Step 3 - Strategy a: Select the ensemble member with the highest quality in each cluster $\Pi_g^{S,L}$
Step 3 - Strategy b: $Rank(\pi_i) = Sort(Quality(\pi_i))$ Then select top J_g ensemble members: Π_g^S
Consensus function
Compute consensus solution from subset of each layer
$(\pi_{HCAS-SQI}^{*,1}, \pi_{HCAS-SQI}^{*,2}, \ldots, \pi_{HCAS-SQI}^{*,L})$

Levels	Number of Groups
1	$k = L$
2	$k = L - 1$
3	$k = L - 2$
4	$k = L - 3$
⋮	⋮
$L - 2$	$k = 3$
$L - 1$	$k = 2$
L	$k = 1$

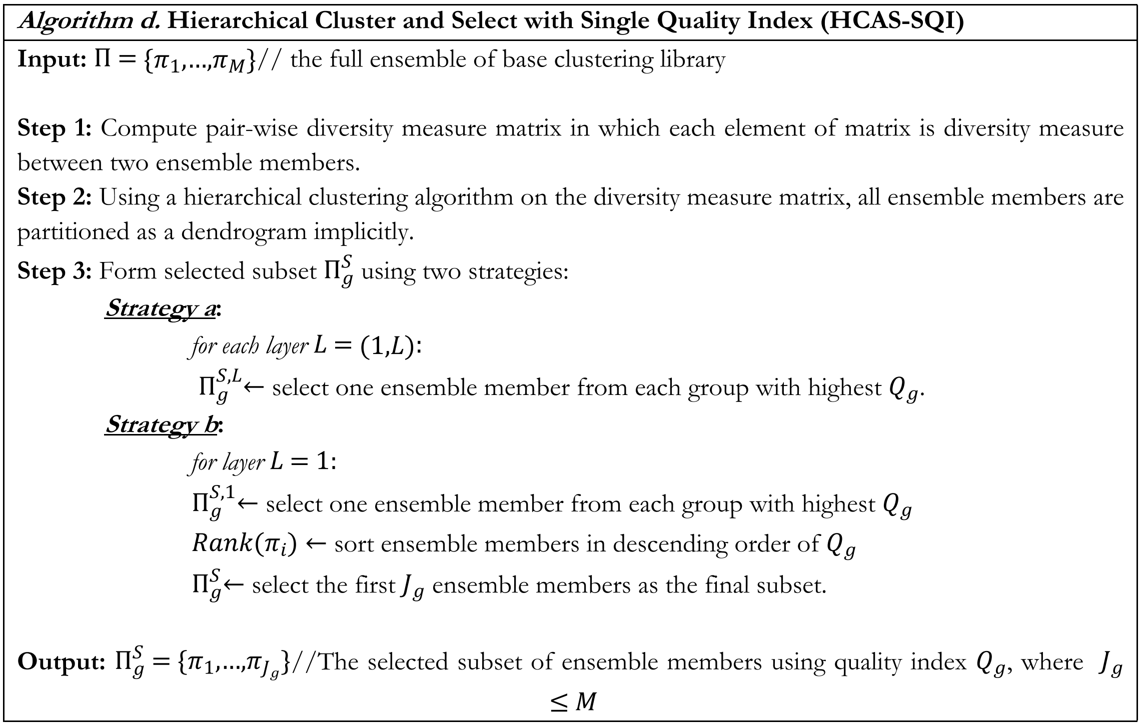
Fig. 4 Flowchart of HCAS-SQI method (*Algorithm d*)

CES—hierarchical cluster and select using single quality index (HCAS-SQI)

In HCAS-SQI method (*Algorithm d*), each ensemble member is considered as an entity (node in the dendrogram). The pair-wise diversity matrix is constructed use diversity measure defined as:

$$Diversity\left(\pi_i, \pi_j\right) = 1 - NMI\left(\pi_i, \pi_j\right) \tag{9}$$

The ensemble members are partitioned by a hierarchical clustering algorithm using the constructed diversity matrix. Different agglomerative hierarchical clustering linkage methods can be used such as single, average, and complete linkage. The results can be displayed as a dendrogram that includes nested partitions of all ensemble members. The final grouping of ensemble members is obtained by cutting the dendrogram at the proper layer. The subset Π^S is formed by select the highest quality member from each group, as described below and shown in Fig. 4:



Algorithm d. **Hierarchical Cluster and Select with Single Quality Index (HCAS-SQI)**

```
Input: $\Pi = \{\pi_1,\ldots,\pi_M\}$// the full ensemble of base clustering library

Step 1: Compute pair-wise diversity measure matrix in which each element of matrix is diversity measure between two ensemble members.
Step 2: Using a hierarchical clustering algorithm on the diversity measure matrix, all ensemble members are partitioned as a dendrogram implicitly.
Step 3: Form selected subset $\Pi_g^S$ using two strategies:
        Strategy a:
                for each layer $L = (1,L)$:
                $\Pi_g^{S,L}$← select one ensemble member from each group with highest $Q_g$.
        Strategy b:
                for layer $L = 1$:
                $\Pi_g^{S,1}$← select one ensemble member from each group with highest $Q_g$
                $Rank(\pi_i)$ ← sort ensemble members in descending order of $Q_g$
                $\Pi_g^S$← select the first $J_g$ ensemble members as the final subset.

Output: $\Pi_g^S = \{\pi_1,\ldots,\pi_{J_g}\}$//The selected subset of ensemble members using quality index $Q_g$, where $J_g \leq M$
```

Specifically, two strategies are used in Step 3 after the dendrogram is partitioned at a layer L. The first strategy (*strategy a*) is to simply select the highest quality ensemble member from each group. The number of ensemble members J_g in the selected subset Π^S is determined by the number of groups in this layer. From bottom layer 1 to top layer L, different subsets of ensemble members are chosen with L to 1 ensemble members, respectively. At layer L (the top layer), all ensemble members are included in one group (Fig. 3), in this condition, it is equivalent to choosing the ensemble member with the highest quality. However, in *strategy a*, low-quality ensemble members can be selected as long as they are diverse from the ensemble members in other groups. In *strategy b*, the dendrogram is partitioned at the bottom layer $L = 1$, then the highest quality ensemble members are selected from each group, and subsequently ranked according to their quality. The final ensemble subset is formed by including the J_g top-ranked ensemble members, where $J_g \leq M$.

CES-hierarchical cluster and selection using multiple quality indices (HCAS-MQI)

HCAS-MQI (*Algorithm e*), is based on HCAS-SQI (*Algorithm d*). After selecting subsets of Π_g^S ($g \ni 1, \ldots, G$), where G is the number of subsets from using different quality indices, a combining strategy can be used to generate the final reduced ensemble subset. The three different combining strategies evaluated in this study are:

a. Direct combining (DC): directly combining selected ensemble members in each subset Π_g^S to obtain a new subset Π_{DC}^S. Any ensemble member that is selected by one or more quality Indices.
b. Weighted combining (WC):

A unified weighting function, which takes into account both the weights of the subsets and those of the base clustering solution in each subset. Since each clustering solution π_i corresponds to a partition of the data, it is reasonable to adopt suitable criteria to measure the quality of the clusters and assign the weights of base clustering solutions. The Squared-Error Distortion (Distor) is designed to minimize the mean squared distance from all data points to their nearest cluster centroids, which is defined as follows:

$$\mathrm{Distor}\left(\pi_j\right) = \frac{1}{N}\sum_{p=1}^{N}\phi\left(x_p, U\right), \tag{10}$$

where ϕ is the distance function and U is the set of cluster centers.

Given a subset Π_g^S, which contains J_g clustering solutions selected from M full ensemble members of Π, the weight of selected ensemble member π_j in Π_g^S is computed as follows:

$$w_j^g = \exp\left(\frac{-\left(\mathrm{Distor}\left(\pi_j\right) - min_{j=1}^{J}\mathrm{Distor}\left(\pi_j\right)\right)}{\frac{1}{J}\sum_{j=1}^{J}\mathrm{Distor}\left(\pi_j\right)}\right), \tag{11}$$

The weights of the non-selected base clustering solutions are set to 0, which means $w_j^g = 0$ for $\pi_j \in \Pi - \Pi_g^S$.

The weight of a subset Π_g^S is determined by the weights of its selected base clustering solutions as follows:

$$\overline{w_g} = \frac{1}{J} \sum_{\pi_j \in \Pi_g^S} w_j^g, \tag{12}$$

where J is the number of chosen ensemble members by the g-th HCAS run. The weights of subsets $\overline{w_g}$ are then re-normalized:

$$\widetilde{w_g} = \exp\left(\frac{\overline{w_g} - max_{g=1}^{G}\overline{w_g}}{\frac{1}{G}\sum_{g=1}^{G}\overline{w_g}} \right), \tag{13}$$

The weighting function $\chi(\pi_i)$ of each base clustering solution π_i is calculated based on the weights of the subsets and individual w_i^g in each subset as follows:

$$\chi(\pi_i) = \sum_{g=1}^{G} \widetilde{w_g} \bullet w_i^g, \tag{14}$$

c. Bagging combining (BC):

The Bagging technique described in "CES-bagging selection" section is used, but Π_{DC}^S is used as the input instead.

The HCAS-MQI with different combining strategies are shown in Fig. 5 and as follows:

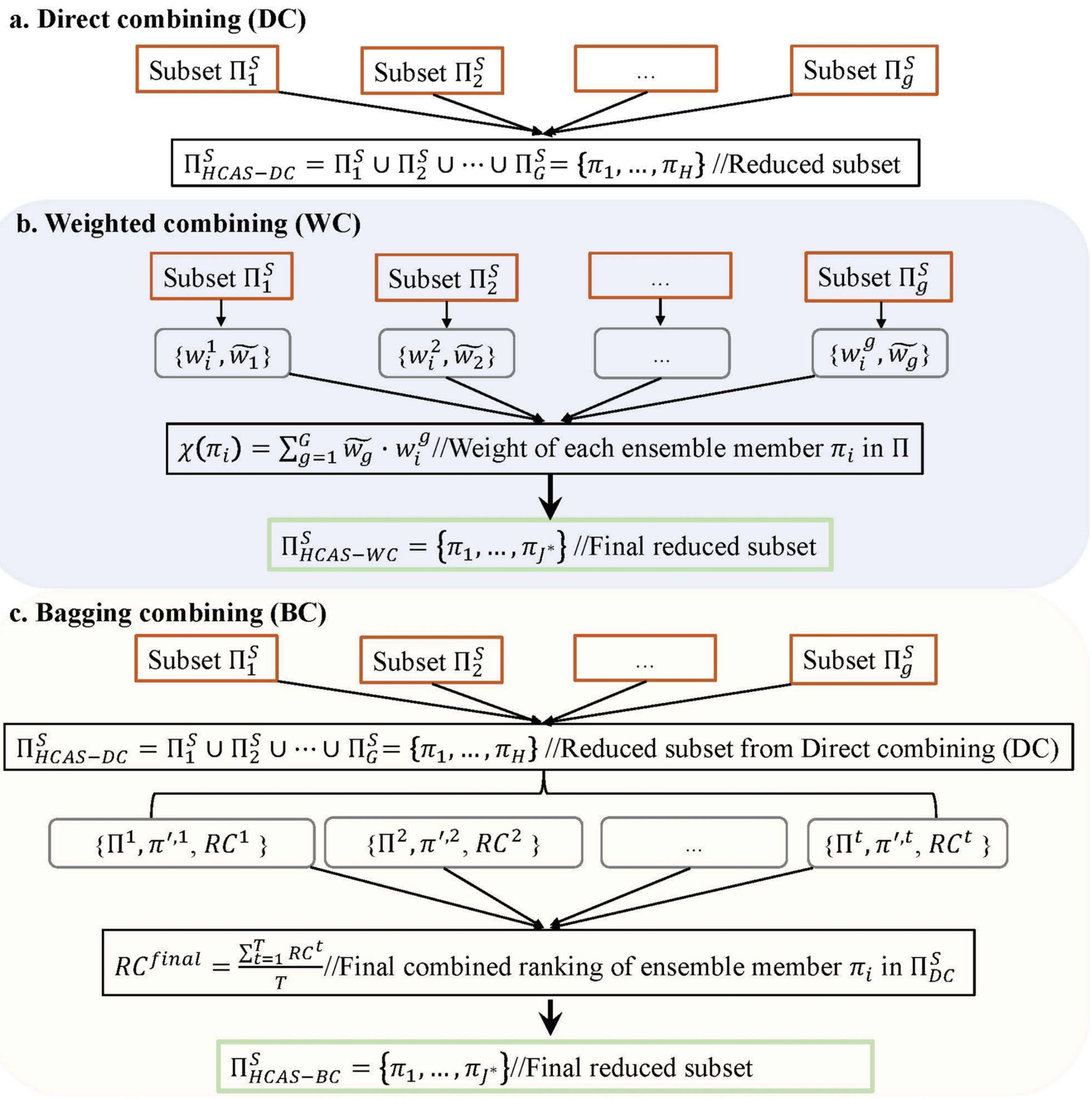

a. Direct combining (DC)
Subset Π_1^S
Subset Π_2^S
…
Subset Π_g^S
$\Pi_{HCAS-DC}^S = \Pi_1^S \cup \Pi_2^S \cup \cdots \cup \Pi_G^S = \{\pi_1, \ldots, \pi_H\}$ //Reduced subset
b. Weighted combining (WC)
Subset Π_1^S
Subset Π_2^S
…
Subset Π_g^S
$\{w_i^1, \widetilde{w_1}\}$
$\{w_i^2, \widetilde{w_2}\}$
…
$\{w_i^g, \widetilde{w_g}\}$
$\chi(\pi_i) = \sum_{g=1}^{G} \widetilde{w_g} \cdot w_i^g$ //Weight of each ensemble member π_i in Π
$\Pi_{HCAS-WC}^S = \{\pi_1, \ldots, \pi_{J^*}\}$ //Final reduced subset
c. Bagging combining (BC)
Subset Π_1^S
Subset Π_2^S
…
Subset Π_g^S
$\Pi_{HCAS-DC}^S = \Pi_1^S \cup \Pi_2^S \cup \cdots \cup \Pi_G^S = \{\pi_1, \ldots, \pi_H\}$ //Reduced subset from Direct combining (DC)
$\{\Pi^1, \pi'^{,1}, RC^1\}$
$\{\Pi^2, \pi'^{,2}, RC^2\}$
…
$\{\Pi^t, \pi'^{,t}, RC^t\}$
$RC^{final} = \frac{\sum_{t=1}^{T} RC^t}{T}$ //Final combined ranking of ensemble member π_i in Π_{DC}^S
$\Pi_{HCAS-BC}^S = \{\pi_1, \ldots, \pi_{J^*}\}$ //Final reduced subset

Fig. 5 Flowchart of HCAS-MQI method (*Algorithm e*) of different combining strategies

```
Algorithm e. Hierarchical Cluster and Select with Multiple Quality Indices (HCAS-MQI)
Input: $\Pi = \{\pi_1,\ldots,\pi_M\}$// the full ensemble of base clustering library

Step 1. Generation ensemble subset using HCAS-SQI as described in Algorithm d:
        $\Pi^S = \{\Pi_1^S,\ldots,\Pi_G^S\}$← $G$ subsets of selected ensembles using different quality index $Q_g$
Step 2. Generate combined final reduced subset:
        Direct combining (DC):
        $\Pi_{HCAS-DC}^S = \Pi_1^S \cup \Pi_2^S \cup \cdots \cup \Pi_G^S = \{\pi_1,\ldots,\pi_H\}$//Final reduced subset from DC

        Weighted combining (WC):
        For each subset $\Pi_g^S$:
                Step 2a. Distor($\pi_j$)← compute the Squared-Error Distortion of each ensemble member
                Step 2b. $w_j^g$← compute the weight of each ensemble member, where the weight of non-selected ensemble member is set to 0
                Step 2c. $\overline{w_g}$← compute the weight of each subset
        Step 2d. $\overline{w_g}$← compute the re-normalized. Weights of subsets
        Step 2e. $\chi(\pi_i)$← the weighting function for each ensemble member

        Bagging combining (BC):
        As described in section 3.3
        $\Pi_{HCAS-DC}^S = \Pi_1^S \cup \Pi_2^S \cup \cdots \cup \Pi_G^S = \{\pi_1,\ldots,\pi_H\}$//Final reduced subset from DC

Output: $\Pi_{HCAS-MQI}^S = \{\pi_1,\ldots,\pi_{J^*}\}$//Final reduced subset
```

Methods and materials

Dataset acquisition and curation

A total of 127 unique NPS compounds were selected from 16 major core chemical structure categories. These include 17 natural or synthetic opioids, 62 stimulants (piperidines, tropane alkaloids, amphetamines, cathinones, aminoindanes, and benzofurans), 35 hallucinogens (2C, 2C-B, and 2C-T series, and tryptamines), 6 sedatives (benzodiazepines), and 7 cannabinoids. A total of 10 low-energy conformers were generated by PubChem3D [84], which samples the energetically accessible and biologically relevant conformations of chemical structures using the average atomic pairwise rmsd. The geometry optimizations were performed using the Gaussian 16 program [85] using the B3LYP level of DFT in combination with the 6–311 + +G(d, p) basis set. Different basis sets (6-31G(d), 6–31 + +G(d, p), 6–311 + +G(d, p)) were used with B3LYP for the computation of the spectra as described in our previous study [86]. The results were compared to the experimental gas IR spectra available at the NIST [87] for six compounds, and the unscaled 6–311 + +G(d, p) spectra resulted in the highest spectral correlation coefficients. Redundant conformers converged to the same structure were eliminated from the dataset, thus leaving a total of 930 conformations. The harmonic vibrational wavenumbers of all conformers were determined at the corresponding optimized structures, which were confirmed to be local minima by checking that there were no imaginary frequencies. The dynamic Raman scattering activity was calculated with the polarizability gradient method with laser excitation wavelength set at 785 nm, which corresponds to a wavelength of 12,739 cm^{-1} and 0.0580 Hartree. The IR spectra and Raman spectra were truncated from 400 to 4000 cm^{-1} and 200 to 1800 cm^{-1}, respectively, with an interval of 2 cm^{-1}. Therefore, two separate datasets were obtained for the NPS compound set, using IR and Raman spectra as features, where the dimensions of the data sets are: (930×1801) and (930×801), respectively. See Supporting Document Appendix B and C for more details.

Generation of base clustering ensemble Π

Mixed heuristics were used to diversify the base clustering ensemble. First, two correlation coefficients are used as a spectral similarity measure to project the original spectral feature to N by N spectral similarity matrix, see Supporting Document Appendix D. This, including the original dataset, resulted in three data representations. Each of the following diversifying approaches was applied to all three data representations. Second, for each clustering run, the number of clusters k predicted for that run, is set by randomly drawing a number between 2 and c, where c is defined as $\sqrt{N}$. Third, the iterative clustering algorithm k-means is applied with different random initialization. Fourth, the deterministic clustering algorithm hierarchical clustering is used with five different linkage methods, which generate different dendrograms and subsequently different clustering solutions. The above settings are used to generate the base clustering ensemble Π with 300 ensemble members. All reported results are averaged across 20 evaluations.

Consensus functions

We experimented with the popular approaches include the Cluster-based Similarity Partitioning Algorithm (CSPA) [13] and Hybrid Bipartite Graph Formulation (HBGF) [74]. We apply both to produce a final partition of the data points into K clusters, where K is the number of known classes in the NPS dataset. Both CSPA and HBGF are implemented from using the ClusterEnsembles Python package [88].

Performance evaluation criterion

Since the ground truth label is unknown, the class labels of the NPS dataset are used as a surrogate. The degree to which two molecules are considered 'similar' depends on both their structural encoding and the similarity metric used. The class label of NPS compounds is assigned using the Maximum Common Substructure (MCS) similarity defined as:

$$T_{MCS} = \frac{N_C}{N_A + N_B - N_C}, \tag{15}$$

where N_C is the number of matched heavy atoms in MCS of molecule A and B, N_A and N_B are the number of heavy atoms in molecule A and B, respectively. T_{MCS} was calculated using the *rdFMCS* modules implemented in RDKit software [89]. The affinity matrices were used as input and submitted to a Ward linkage clustering with Euclidean distance as the similarity metric for hierarchical clustering. The optimal number of clusters K was determined by silhouette index (SI) analysis (see Supporting Document Appendix E, Figure S1).

The performance of CES methods and the traditional cluster ensemble approach is measured by the average value of the NMI between the predicted cluster labels and the ground truth labels after performing the evaluation 20 times.

Another measure, Dominant ratio (Γ) [79], is defined based on NMI to evaluate the effectiveness of clustering solution selection strategies as follows:

$$\Gamma = \frac{NMI(\mathrm{Y}_1^S, \mathrm{Y})}{NMI(\mathrm{Y}_2^S, \mathrm{Y})}, \tag{16}$$

where Y is the set of ground truth labels, Y_1^S is the set of predicted cluster labels derived from the selected ensemble subset Π^S by the CES strategy, and Y_2^S is the predicted cluster label set derived from the remaining unselected clustering solutions $\Pi - \Pi^S$. A better CES strategy will result in a higher value of the dominant ratio (Γ).

Results and discussion

The following results report the NMI and Γ on each CES method while varying the size of the ensemble subset from 5 to 200. Since the full ensemble consensus solution uses all ensemble members, it is not possible to calculate their Γ. In addition, we report the performance of a randomly selected strategy that forms an ensemble subset by randomly drawing from the library, which is repeated 10 times in each run. As mentioned earlier, each number reported is an average of 20 runs. The class label (ground truth) of the NPS data set is used only to evaluate the CES methods, and is not used in the CES process. All experiments were performed using both IR and Raman data sets, only unique results are shown. For the proposed CES methods and the Random method, each of their results was compared with the full ensemble, and those that are statistically superior to the full ensemble ($p < 0.05$, paired t-test) are shown in bold font. The performances of quality indices, consensus functions, and individual CES methods are shown in Figs. 6, 7, 8 and Tables 1, 2, 3. The subfigures on the right give the NMI values determined using the ground truth labels, whereas the sub-figures on the right show the dominant ratio Γ. The size of the selected subsets is plotted on the x-axis.

Comparison of quality indices and consensus function in SQI method

In this experiment, the four quality indices presented in " Quality and diversity measures of clustering solution" section were used in the SQI method. The ensemble subsets selected using SQI with a single quality index are referred to as SQI-Q_g (SQI-SNMI, SQI-Silhouette, SQI-Calinski Harabasz, and SQI-Davies Bouldin). The results obtained using the IR dataset are shown in Fig. 6. Subfigures (a) and (c) show the results of using the CSPA consensus function, and subfigures (b) and (d) show the results of using the HBGF consensus function. Table 2 report the NMI values for ensemble subset sizes of 30, 60, 90, 120, 150, and 180 obtained from each combination of data set, consensus

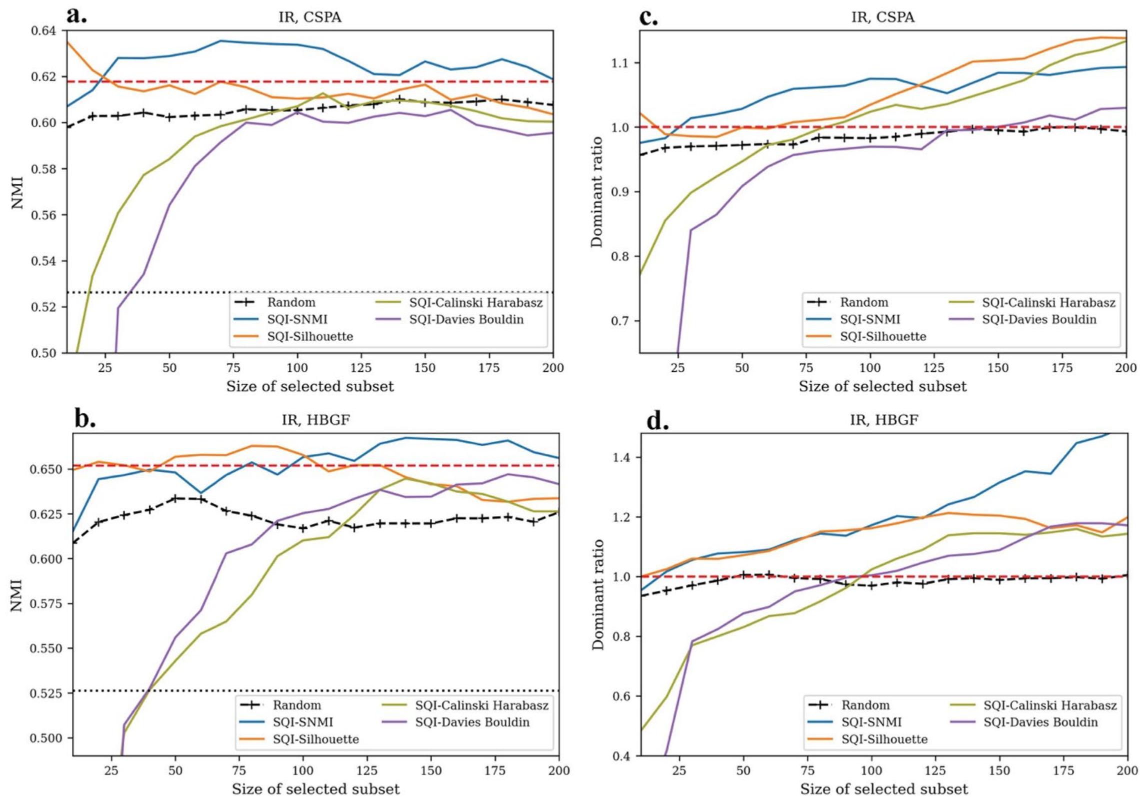


Fig. 6 Performance of SQI using different quality indices and consensus function. In (**a**) and (**b**), the red dashed line indicates the consensus solution using the full ensemble Π, and the black dotted line represents the average NMI of all ensemble members. In (**c**) and (**d**), the dominant ratio Γ = 1.0 is shown as a red dashed line

function, and SQI method, as well as the NMI values of the full ensembles.

To understand whether the full ensemble consensus solution is influenced by the quality of ensemble members, we compared the average and NMI distributions of the ensemble members generated using the IR and Raman datasets. As shown in Fig S3 in Appendix F of the Supporting Document, the average NMI of ensemble members using the Raman dataset is 0.608, while the average NMI using the IR data set is only 0.526. As expected, the quality of ensembles generated of the NPS compounds using IR and Raman datasets is different. However, it is uncertain whether this discrepancy is due to the lower dimensionality of the Raman data set. As shown in Table 2 and Fig. 6, the quality of the ensemble has a positive effect on the full ensemble consensus solution, for which the full ensemble consensus solution using Raman dataset is superior, regardless of the consensus function used.

The Random method was included to ensure that the performance improvement observed with the CES methods could not have been achieved by chance. The Random method represents the CE approach that does not take into account the quality or diversity of the ensemble members. The results shown in Fig. 6 confirm this, as the Random selection method is generally worse than the full ensemble consensus solution across different subset sizes. We also see that the quality indices SNMI and Silhouette can effectively improve consensus performance. On the other hand, subsets selected using Calinski Harabasz or Davies Bouldin only perform comparably to the Random method for increasingly larger subset sizes. It is interesting to note that for the Raman data set, as can be seen in Appendix Figure S4, the Random method performed respectably well in comparison to that of the full ensemble. This suggests that there exists larger amount of redundancy in the libraries.

Evaluate the effect of diversity in HCAS-SQI

The HCAS method reduces the redundancy in the ensemble library by hierarchically partitioning the ensemble members into different groups using the pairwise diversity matrix of the ensemble. The HCAS-SQI method described

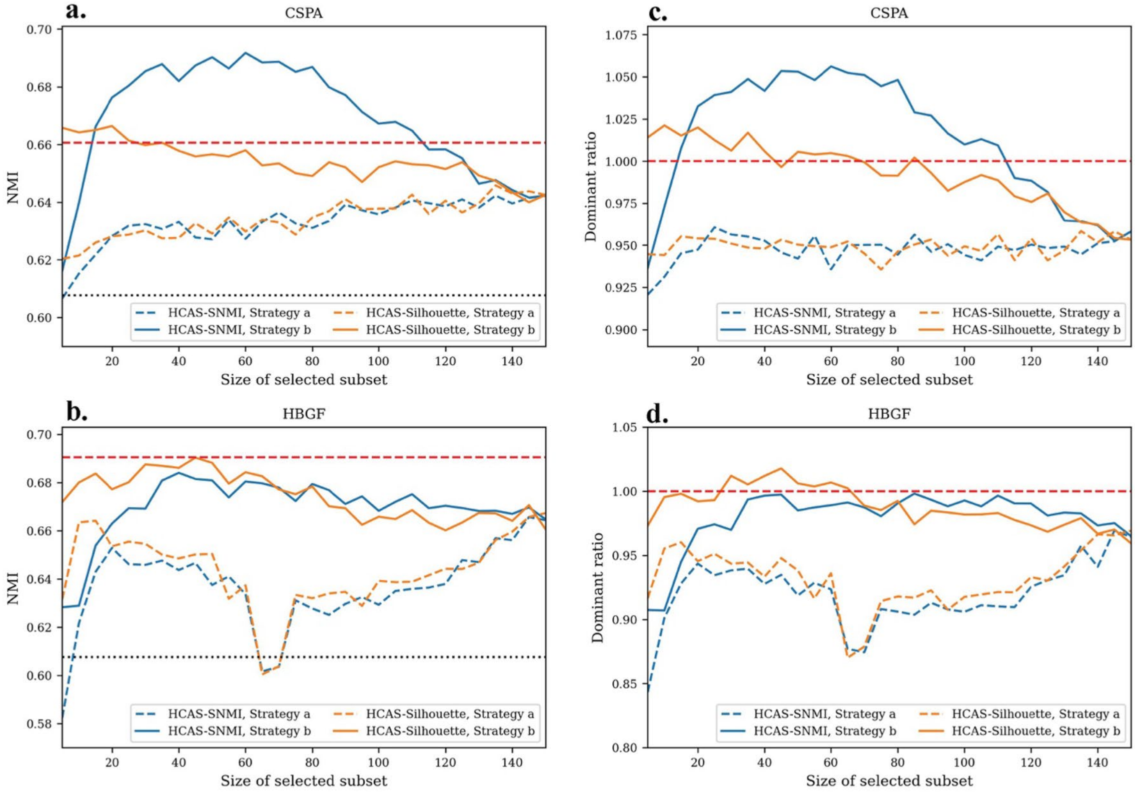


Fig. 7 Performance of HCAS-SQI using hierarchical clustering with Ward linkage. In (**a**) and (**b**), the red dashed line indicates the consensus solution using the full ensemble Π, and the black dotted line represents the average NMI of all ensemble members (Reference 1). In (**c**) and (**d**), the dominant ratio Γ = 1.0 is shown as a red dashed line

in "CES—hierarchical cluster and select using single quality index (HCAS-SQI)" section was tested with a focus on assessing which selection strategy is more effective. To simplify the analysis, only SNMI and Silhouette quality indices were used, and the ensemble subsets selected using HCAS-SQI with a single quality index are referred to as HCAS-Q_g (HCAS-SNMI and HCAS-Silhouette). From the discussion in "Comparison of quality indices and consensus function in SQI method " in "CES—hierarchical cluster and select using single quality index (HCAS-SQI)" section, we see that although the ensemble quality is on average higher using the Raman data set, there is greater redundancy in the ensemble library. Therefore, the Raman data set library may benefit more from a CES method that takes into account the diversity among ensemble members. The results obtained using the Raman data set using the CSPA and HBGF consensus functions are shown in Fig. 7. Different clustering linkage methods were used in this experiment, and since there is no statistical difference among the results of these linkage methods, only the results obtained using Ward's linkage method are shown here.

When constructing a tree diagram of ensemble members, the size of the clusters is requested to be 150. *Strategy a* is equivalent to cutting the dendrogram from the top layer to the bottom layer and gradually generating more distinct groups towards the lower layer. At each layer, a subset is formed by selecting the highest quality ensemble members from each group. At the top layer, all ensemble members are in one group, so the *strategy a* selects one ensemble member with the highest quality from the library, which is equivalent to that of the SQI method. However, starting from the second-top layer, it starts to diverge from the SQI method, as equally high-quality but redundant ensemble members will not be selected. Thus, while trying to increase diversity, a subset of the ensemble of the same size contains more low-quality members. As can be seen in Fig. 7, this CES strategy did not lead to improvements.

Strategy b can be considered a special case of *strategy a*. At the bottom layer $L = 1$, all ensemble members are divided into 150 diverse groups, and the highest quality members in each group are first selected and then sorted again according to their quality. Then, subsets of different sizes are formed by including the top-ranked ensemble members. As supported by the results, the advantage of *strategy b* is obvious because it can achieve better performance at a smaller subset size when using the CSPA consensus function. The subset that achieved the maximal performance in HCAS-SNMI

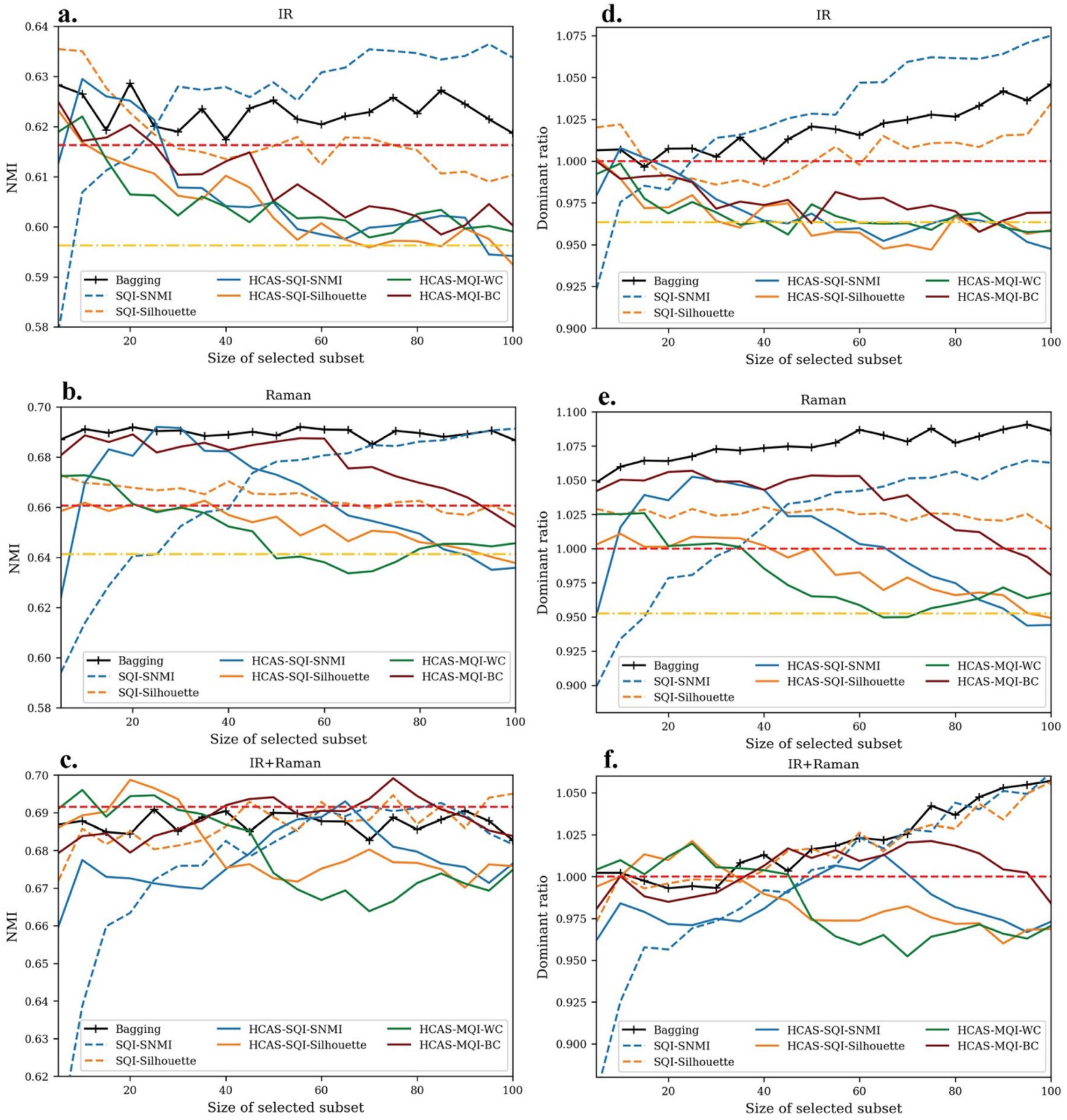


Fig. 8 Final comparison of all CES method. CSPA consensus function. HCAS using Ward linkage and strategy b. In **a–c** the red dashed line indicates the consensus solution using the full ensemble Π, and the yellow dotted line represents HCAS-MQI-DC subsets. In **d–f** the dominant ratio Γ = 1.0 is shown as a red dashed line

has about 60 members, which is half of the best-performing subset in SQI-SNMI. Of course, as the subset size increases, more low-quality members are selected, so it performs worse and eventually converges with *strategy a*.

Final comparison of all CES methods

In addition to improving the quality of clustering solutions, there are other motivations and benefits of using CE methods. It is well known that a single clustering algorithm might fail for certain datasets that do not match well with the modeling assumptions [90]. CE method that uses multiple base clustering algorithms applicable to various datasets can provide more robust performances. Strehl and Ghosh illustrated empirically the utility of CE as Feature distributed clustering, where different clustering solutions are built by selecting different subsets of the features while utilizing all the data points [13]. In the case of spectroscopy data of organic compounds, there are multiple aspects or "views" of the object to be clustered, as IR and Raman spectra of the same compound can complement each other as different features of the same data point. Here we use a simple strategy

Table 2 Results for SQI

Data	Consensus	SQI-Q_g	Size						Full
			30	60	90	120	150	180	
IR	CSPA	SQI-SNMI	0.628 (0.014)	**0.631** (0.007)	**0.634** (0.007)	**0.627** (0.014)	**0.626** (0.009)	**0.627** (0.011)	0.618 (0.016)
		SQI-Silhouette	0.616 (0.010)	0.612 (0.007)	0.611 (0.011)	0.612 (0.012)	0.616 (0.012)	0.608 (0.009)	
		Random	0.603 (0.006)	0.603 (0.006)	0.605 (0.007)	0.607 (0.007)	0.609 (0.005)	0.610 (0.007)	
	HBGF	SQI-SNMI	0.647 (0.015)	0.637 (0.015)	0.647 (0.011)	0.654 (0.007)	**0.667** (0.010)	**0.666** (0.012)	0.652 (0.015)
		SQI-Silhouette	0.652 (0.014)	0.658 (0.013)	**0.663** (0.010)	0.652 (0.011)	0.642 (0.011)	0.632 (0.011)	
		Random	0.624 (0.005)	0.633 (0.007)	0.619 (0.008	0.617 (0.006)	0.620 (0.007)	0.623 (0.008)	
Raman	CSPA	SQI-SNMI	0.652 (0.019)	**0.681** (0.012)	**0.689** (0.013)	**0.697** (0.010)	**0.694** (0.010)	**0.690** (0.010)	0.661 (0.012)
		SQI-Silhouette	0.668 (0.008)	0.662 (0.008)	0.657 (0.008)	0.655 (0.013)	0.669 (0.014)	0.667 (0.017)	
		Random	0.664 (0.005)	0.662 (0.004)	0.661 (0.004)	0.662 (0.003)	0.660 (0.004)	0.661 (0.004)	
	HBGF	SQI-SNMI	0.651 (0.013)	0.671 (0.010)	0.678 (0.012)	0.694 (0.020)	0.693 (0.010)	0.698 (0.010)	0.690 (0.016)
		SQI-Silhouette	0.692 (0.013)	0.698 (0.013)	**0.705** (0.013)	**0.708** (0.010)	**0.700** (0.010)	0.693 (0.013)	
		Random	0.673 (0.005)	0.672 (0.007)	0.672 (0.008)	0.672 (0.006)	0.676 (0.006)	0.679 (0.006)	

Results that are statistically superior to the full ensemble ($p<0.05$, paired t-test) are shown in bold font

to build a hybrid ensemble library (IR + Raman) combining clustering solutions built using IR and Raman datasets of NPS compound: 150 ensemble members were randomly selected with replacements from each of the IR and Raman generated ensembles, jointly creating an ensemble library with 300 members.

The last experiment compares all CES methods. As seen in Figs. 6 and 7, although HBGF gives better full ensemble results, neither SQI nor HCAS-SQI can further improve the ensemble subset performance. In this final comparison, only the CSPA consensus function was used. The Bagging method described in "CES-bagging selection" in "CES—hierarchical cluster and select using single quality index (HCAS-SQI)" section uses 50 bootstrap iterations to generate the ensemble subsets, and *Strategy b* and Ward linkage are used in HCAS-SQI. As described in "CES-hierarchical cluster and selection using multiple quality indices (HCAS-MQI)" section, the ensemble subsets selected using HCAS-SQI with SNMI or Silhouette (HCAS-SNMI and HCAS-Silhouette) were merged using three combining strategies (HCAS-MQI-DC, HCAS-MQI-WC, and HCAS-MQI-BC). Figure 8 shows the results obtained using the ensemble libraries generated from the IR and Raman datasets as well as the IR + Raman hybrid library. Table 3 provides the summary of the final comparison of all CES methods for the ensemble subset sizes of 20, 40, 60, 80, and 100, as well as the NMI values of the full ensemble consensus.

We first noticed that SQI-SNMI is very sensitive to ensemble subset sizes, and its performance is unfavorable for smaller set sizes. With the addition of more ensemble members, its performance gradually improves and surpasses all other CES methods. This sensitivity is less severe when diversity is also considered in the HCAS-SNMI approach, requiring only a smaller ensemble size to achieve its optimal performance. In contrast, the Bagging method is relatively more robust with regard to the ensemble subset sizes.

In terms of the impact of the composition of the ensemble library, we observe that the Raman responds most strongly to the CES methods, as all CES methods except HCAS-SQI-Silhouette and HCAS-MQI-BC result in improved performance compared to the full ensemble consensus solution. It is also interesting to observe that the IR + Raman hybrid library benefits the most from the CE approach, although none of the CES methods can further improve performance. This hybrid library that merges complementary "views" of the data objects improves the intrinsic diversity of the

Table 3 Comparison across all CES methods

Data	CES methods	Size					Max	Full
		20	40	60	80	100		
IR	SQI-SNMI	0.615 (0.012)	0.626 (0.011)	**0.631** (0.008)	**0.635** (0.007)	**0.636** (0.008)	0.638	0.618 (0.016)
	SQI-Silhouette	0.622 (0.012)	0.612 (0.012)	0.613 (0.007)	0.615 (0.009)	0.612 (0.013)	0.636	
	Bagging	**0.627** (0.013)	0.620 (0.010)	0.622 (0.009)	0.623 (0.012)	0.618 (0.013)	0.627	
	HCAS-SQI-SNMI	0.623 (0.010)	0.604 (0.010)	0.598 (0.010)	0.603 (0.011)	0.593 (0.009)	0.629	
	HCAS-SQI-Silhouette	0.611 (0.008)	0.610 (0.008)	0.598 (0.008)	0.596 (0.008)	0.594 (0.013)	0.625	
	HCAS-MQI-WC	0.607 (0.009)	0.603 (0.010)	0.600 (0.007)	0.602 (0.012)	0.599 (0.010)	0.621	
	HCAS-MQI-BC	0.621 (0.011)	0.613 (0.012)	0.604 (0.008)	0.600 (0.014)	0.600 (0.010)	0.624	
Raman	SQI-SNMI	0.641 (0.018)	0.659 (0.017)	**0.681** (0.012)	**0.686** (0.016)	**0.691** (0.010)	0.691	0.661 (0.012)
	SQI-Silhouette	0.668 (0.012)	**0.670** (0.011)	0.662 (0.008)	0.663 (0.012)	0.657 (0.010)	0.673	
	Bagging	**0.692** (0.016)	**0.689** (0.013)	**0.691** (0.010)	**0.690** (0.012)	**0.687** (0.010)	0.692	
	HCAS-SQI-SNMI	**0.681** (0.015)	**0.682** (0.012)	0.663 (0.011)	0.649 (0.010)	0.636 (0.012)	0.692	
	HCAS-SQI-Silhouette	0.661 (0.008)	0.657 (0.016)	0.653 (0.013)	0.646 (0.014)	0.638 (0.012)	0.663	
	HCAS-MQI-WC	0.661 (0.014)	0.652 (0.012)	0.638 (0.015)	0.643 (0.014)	0.646 (0.013)	0.673	
	HCAS-MQI-BC	**0.689** (0.010)	**0.683** (0.011)	**0.687** (0.013)	**0.670** (0.009)	0.652 (0.013)	0.689	
IR + Raman	SQI-SNMI	0.663 (0.024)	0.683 (0.017)	0.691 (0.011)	0.691 (0.010)	0.682 (0.014)	0.693	0.691 (0.020)
	SQI-Silhouette	0.685 (0.013)	0.686 (0.017)	0.693 (0.017)	0.687 (0.017)	0.695 (0.013)	0.695	
	Bagging	0.684 (0.020)	0.690 (0.017)	0.688 (0.015)	0.686 (0.013)	0.683 (0.015)	0.691	
	HCAS-SQI-SNMI	0.673 (0.017)	0.675 (0.012)	0.689 (0.023)	0.680 (0.014)	0.677 (0.013)	0.693	
	HCAS-SQI-Silhouette	0.699 (0.017)	0.675 (0.014)	0.675 (0.012)	0.677 (0.012)	0.676 (0.010)	0.699	
	HCAS-MQI-WC	0.694 (0.021)	0.687 (0.015)	0.667 (0.015)	0.671 (0.010)	0.675 (0.010)	0.696	
	HCAS-MQI-BC	0.679 (0.014))	0.692 (0.018)	0.690 (0.017)	0.694 (0.011)	0.684 (0.019)	0.699	

Results that are statistically superior to the full ensemble ($p<0.05$, paired t-test) are shown in bold font

The results of the CES method with the best maximum performance on each data set are underlined

clustering solutions, which we believe is the main reason for its superior performance.

Ward linkage and Strategy b were used in HCAS method when forming subset Π_g^S. The HCAS-MQI method merges the ensemble subsets created by HCAS-SQI-SNMI and HCAS-SQI-Silhouette. Only the CSPA consensus function was used in this analysis.

Conclusion

In this paper, we investigate the utility of cluster ensemble selection method (CES) in improving unsupervised learning tasks for high-dimensional spectroscopy data of organic compounds. Two complementary spectra of NPS compounds were used in this study, namely IR and Raman datasets, which were calculated using Gaussian16 at the B3LYP/6–311 + + G(d, p) level. The goal of the CES method is to select a subset of ensemble members from a large library of clustering solutions to form a consensus solution that achieves better performances than using the full ensemble. While ensemble learning algorithms in classification tasks, such as Bagging and Boosting, has become popular and widely used, unsupervised ensemble learning is much more difficult, and its application in high-dimensional spectroscopy data is worth investigating.

Four CES frameworks are proposed by incorporating commonly used clustering validation indices. The results presented in "Comparison of quality indices and consensus function in SQI method" section suggest that SQI method using SNMI and Silhouette can obtain consensus solutions with quality higher than or equivalent to that of the full ensemble. Interestingly, although consensus solutions obtained using HBGF gives better results, it also required larger ensemble sizes compare to that used by CSPA. The HCAS method aims at select ensemble subsets by considering the diversity and quality of the ensemble members. For libraries containing more redundant ensemble members, CES is more effective in further improving performance compared to the full ensemble consensus scheme. The IR + Raman hybrid ensemble library is created by merging two complementary "views" of the organic compounds. This inherently more diverse library gives the best full ensemble consensus results. Overall, the Bagging method is recommended because it provides the most robust results that are better than or comparable to the full ensemble consensus solutions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10822-022-00488-9.

Acknowledgements Thank the MERCURY Consortium for computing resources and technical support. Clemson University is acknowledged for generous allotment of compute time on Palmetto cluster.

Author contributions KH conceived and developed the presented idea, and were in charge of the overall direction and planning. Both KH and DM performed the computations and the data analysis.

Funding Computational resources were provided in part by the MERCURY consortium (http://mercuryconsortium.org/) under National Science Foundation grants CHE-1229354, CHE-1662030, and CHE-2018427.

Data availability The Supporting Document is provided. The datasets and python source code supporting the conclusions of this article are available in the GitHub repository, https://github.com/nina23bom/Cluster-Ensemble-Selection-Project

Declarations

Conflict of interest The authors declare no competing financial interest.

Consent to participate Not applicable.

Consent for publication Not applicable.

Ethical approval Not applicable.

References

1. Duda RO, Hart PE, Stork DG (2012) Pattern Classification. Wiley, New York
2. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
3. Kim S, Han L, Yu B, Hähnke VD, Bolton EE, Bryant SH (2015) PubChem structure-activity relationship (SAR) clusters. J Cheminform 7:33
4. González-Alemán R, Hernández-Castillo D, Caballero J, Montero-Cabrera LA (2020) Quality threshold clustering of molecular dynamics: a word of caution. J Chem Inf Model 60(2):467–472
5. Glielmo A, Husic BE, Rodriguez A, Clementi C, Noé F, Laio A (2021) Unsupervised learning methods for molecular simulation data. Chem Rev 121(16):9722–9758
6. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, Achas M, Adebiyi E (2016) Clustering algorithms: their application to gene expression data. Bioinform Biol Insights 10:237–253
7. MacQueen J (1967) In Some methods for classification and analysis of multivariate observations
8. von Luxburg U (2007) A tutorial on spectral clustering. Statist Comput 17(4):395–416
9. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ (2006) Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J Math Model Algorithms 5(4):475–504
10. Kleinberg J (2002) An impossibility theorem for clustering. Adv Neural Inform Process Syst 15:16
11. Hennig C (2015) What are the true clusters? Pattern Recognit Lett 64:53–62
12. Jain AK, Duin RPW, Jianchang M (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22(1):4–37
13. Strehl A, Ghosh J (2002) Cluster ensembles: a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
14. Ghosh J, Acharya A (2011) Cluster ensembles. Wiley Interdiscip Rev 1(4):305–315
15. Ghaemi R, Sulaiman NB, Ibrahim H, Mustapha N (2011) A review: accuracy optimization in clustering ensembles using genetic algorithms. Artif Intell Rev 35(4):287–318

16. Ayad HG, Kamel MS (2007) Cumulative voting consensus method for partitions with variable number of clusters. IEEE Trans Pattern Anal Mach Intell 30(1):160–173
17. Fred A, Lourenço A (2008) Cluster ensemble methods: from single clusterings to combined solutions. In Supervised and unsupervised ensemble methods and their applications, Springer, pp 3–30
18. Topchy A, Jain AK, Punch W (2003) In *Combining multiple weak clusterings*, Third IEEE international conference on data mining. IEEE: pp 331–338
19. Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239
20. Kuncheva LI, Vetrov DP (2006) Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Trans Pattern Anal Mach Intell 28(11):1798–1808
21. Fred ALN, Jain AK (2005) Combining multiple clusterings using evidence accumulation. IEEE Trans Pattern Anal Mach Intell 27(6):835–850
22. Boongoen T, Iam-On N (2018) Cluster ensembles: a survey of approaches with recent extensions and applications. Comput Sci Rev 28:1–25
23. Golalipour K, Akbari E, Hamidi SS, Lee M, Enayatifar R (2021) From clustering to clustering ensemble selection: a review. Eng Appl Artif Intell 104:104388
24. Saeed F, Salim N, Abdo A (2012) Voting-based consensus clustering for combining multiple clusterings of chemical structures. J Cheminf 4(1):37
25. Saeed F, Salim N, Abdo A (2013) Information Theory and voting based consensus clustering for combining multiple clusterings of chemical structures. Mol Inform 32(7):591–598
26. Saeed F, Ahmed A, Shamsir MS, Salim N (2014) Weighted voting-based consensus clustering for chemical structure databases. J Comput Aided Mol Des 28(6):675–684
27. Chu C-W, Holliday JD, Willett P (2012) Combining multiple classifications of chemical structures using consensus clustering. Bioorg Med Chem 20(18):5366–5371
28. Fern XZ, Lin W (2008) Cluster ensemble selection. Stat Anal Data Min 1(3):128–141
29. Abbasi S-O, Nejatian S, Parvin H, Rezaie V, Bagherifard K (2019) Clustering ensemble selection considering quality and diversity. Artif Intell Rev 52(2):1311–1340
30. Shi Y, Yu Z, Chen CLP, You J, Wong HS, Wang Y, Zhang J (2020) Transfer Clustering Ensemble Selection. IEEE Trans Cybern 50(6):2872–2885
31. Kuncheva LI, Hadjitodorov ST (2004) In *Using diversity in cluster ensembles*, 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), pp 1214–1219
32. Alizadeh H, Minaei-Bidgoli B, Parvin H (2014) To improve the quality of cluster ensembles by selecting a subset of base clusters. J Exp Theor Artif Intell 26(1):127–150
33. Minaei-Bidgoli B, Parvin H, Alinejad-Rokny H, Alizadeh H, Punch WF (2014) Effects of resampling method and adaptation on clustering ensemble efficacy. Artif Intell Rev 41(1):27–48
34. UNODC Early Warning Advisory on New Psychoactive Substances. What are NPS? https://www.unodc.org/LSS/Home/NPS. (Accessed Mar 2021).
35. "Title 21 United States Code (USC) Controlled Substances Act" United States Drug Enforcement Administration: https://www.dea.gov/controlled-substances-act. (Accessed Mar 2021).
36. Luinge HJ (1990) Automated interpretation of vibrational spectra. Vib Spectrosc 1(1):3–18
37. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature 559(7715):547–555
38. Biancolillo A, Marini F (2018) Chemometric methods for spectroscopy-based pharmaceutical analysis. Front Chem 6:576
39. Wang X-Y, Garibaldi J (2005) Simulated annealing fuzzy clustering in cancer diagnosis. Informatica 29:61–70
40. Wu X, Wu B, Sun J, Yang N (2017) Classification of apple varieties using near infrared reflectance spectroscopy and fuzzy discriminant C-means clustering model. J Food Process Eng 40(2):e12355
41. Haixia R, Weiqi L, Weimin S, Qi S (2013) Classification of edible oils by infrared spectroscopy with optimized k-means clustering by a hybrid particle swarm algorithm. Anal Lett 46(17):2727–2738
42. Fred ALN, Jain AK (2002) In *Data clustering using evidence accumulation*, 2002 International Conference on Pattern Recognition, pp 276–280
43. Ana LNF, Jain AK (2003) In *Robust data clustering*, 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. pp II–II.
44. Iam-on N, Boongoen T, Garrett S (2008) *Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations*. Springer, Berlin, pp 222–233
45. Hadjitodorov ST, Kuncheva LI, Todorova LP (2006) Moderate diversity for better cluster ensembles. Inf Fusion 7(3):264–275
46. Fern XZ, Brodley CE (2003) Random projection for high dimensional data clustering: a cluster ensemble approach. In *Proceedings of the twentieth international conference on international conference on machine learning*, AAAI Press: Washington, DC; pp 186–193
47. Fischer B, Buhmann JM (2003) Bagging for path-based clustering. IEEE Trans Pattern Anal Mach Intell 25(11):1411–1415
48. Dudoit S, Fridlyand J (2003) Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19(9):1090–1099
49. Minaei-Bidgoli B, Topchy AP, Punch WF (2004) In *A comparison of resampling methods for clustering ensembles*, IC-AI
50. Ayad H, Kamel M (2003) Finding natural clusters using multiclusterer combiner based on shared nearest neighbors. Springer, Berlin, pp 166–175
51. Hu X, Yoo I (2004) Cluster ensemble and its applications in gene expression analysis.
52. Law MHC, Topchy AP, Jain AK (2004) In *Multiobjective data clustering*, In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004; pp II–II.
53. Lu X, Yang Y, Wang H (2013) Selective clustering ensemble based on covariance. Springer, Berlin
54. Yousefnezhad M, Reihanian A, Zhang D, Minaei-Bidgoli B (2016) A new selection strategy for selective cluster ensemble based on Diversity and Independency. Eng Appl Artif Intell 56:260–272
55. Azimi J, Fern X (2009) Adaptive cluster ensemble selection. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc.: Pasadena, pp 992–997.
56. Faceli K, Carvalho ACPLFD, Souto MCPD (2006) In *Multi-Objective Clustering Ensemble*, 2006 Sixth International Conference on Hybrid Intelligent Systems (HIS'06). pp 51–51
57. Yu Z, Chen H, You J, Han G, Li L (2013) Hybrid fuzzy cluster ensemble framework for tumor clustering from biomolecular data. IEEE/ACM Trans Comput Biol Bioinform 10(3):657–670
58. Li F, Qian Y, Wang J, Liang J (2017) Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method. Inf Sci 378:389–409
59. Wu X, Ma T, Cao J, Tian Y, Alabdulkarim A (2018) A comparative study of clustering ensemble algorithms. Comput Electr Eng 68:603–615
60. Hamidi SS, Akbari E, Motameni H (2019) Consensus clustering algorithm based on the automatic partitioning similarity graph. Data Knowl Eng 124:101754

61. Ayad HG, Kamel MS (2010) On voting-based consensus of cluster ensembles. Pattern Recognit 43(5):1943–1953
62. Bagherinia A, Minaei-Bidgoli B, Hosseinzadeh M, Parvin H (2021) Reliability-based fuzzy clustering ensemble. Fuzzy Sets Syst 413:1–28
63. Naldi MC, Carvalho ACPLF, Campello RJGB (2013) Cluster ensemble selection based on relative validity indexes. Data Min Knowl Discov 27(2):259–289
64. Alizadeh H, Minaei-Bidgoli B, Parvin H (2014) Cluster ensemble selection based on a new cluster stability measure. Intell Data Anal 18(3):389–408
65. Jia J, Xiao X, Liu B, Jiao L (2011) Bagging-based spectral clustering ensemble selection. Pattern Recognit Lett 32(10):1456–1467
66. Gionis A, Mannila H, Tsaparas P (2007) Clustering aggregation. ACM Trans Knowl Discov Data 1(1):4
67. Hinneburg A, Aggarwal CC, Keim DA (2000) What is the nearest neighbor in high dimensional spaces? In *Proceedings of the 26th international conference on very large data bases*, Morgan Kaufmann Publishers Inc.: pp 506–515
68. Houle ME, Kriegel HP, Kröger P, Schubert E, Zimek A (2010) Can shared-neighbor distances defeat the curse of dimensionality? In: Ludäscher B (ed) Gertz M. Scientific and Statistical Database Management, Springer, Berlin pp, pp 482–500
69. Aggarwal CC (2001) Re-designing distance functions and distance-based applications for high dimensional data. SIGMOD Rec 30(1):13–18
70. Elghazel H, Aussem A (2015) Unsupervised feature selection with ensemble learning. Mach Learn 98(1):157–180
71. Henschel H, van der Spoel D (2020) An intuitively understandable quality measure for theoretical vibrational spectra. J Phys Chem Lett 11(14):5471–5475
72. Henschel H, Andersson AT, Jespers W, Mehdi Ghahremanpour M, van der Spoel D (2020) Theoretical infrared spectra: quantitative similarity measures and force fields. J Chem Theory Comput 16(5):3307–3315
73. Topchy A, Jain AK, Punch W (2004) A mixture model for clustering ensembles. In *Proceedings of the 2004 SIAM international conference on data mining (SDM)*, pp 379–390
74. Fern XZ, Brodley CE (2004) Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning*, Association for Computing Machinery: Banff, Alberta p 36
75. Yang F, Li X, Li Q, Li T (2014) Exploring the diversity in cluster ensemble generation: Random sampling and random projection. Expert Syst Appl 41(10):4844–4866
76. Hong Y, Kwong S, Wang H, Ren Q (2009) Resampling-based selective clustering ensembles. Pattern Recognit Lett 30(3):298–305
77. Li F, Qian Y, Wang J, Dang C, Jing L (2019) Clustering ensemble based on sample's stability. Artif Intell 273:37–55
78. Akbari E, Mohamed Dahlan H, Ibrahim R, Alizadeh H (2015) Hierarchical cluster ensemble selection. Eng Appl Artif Intell 39:146–156
79. Yu Z, Li L, Gao Y, You J, Liu J, Wong H-S, Han G (2014) Hybrid clustering solution selection strategy. Pattern Recognit 47(10):3362–3375
80. Ma T, Yu T, Wu X, Cao J, Al-Abdulkarim A, Al-Dhelaan A, Al-Dhelaan M (2020) Multiple clustering and selecting algorithms with combining strategy for selective clustering ensemble. Soft Comput 24(20):15129–15141
81. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
82. Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3(1):1–27
83. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell PAMI 1(2):224–227
84. Bolton EE, Chen J, Kim S, Han L, He S, Shi W, Simonyan V, Sun Y, Thiessen PA, Wang J, Yu B, Zhang J, Bryant SH (2011) PubChem3D: a new resource for scientists. J Cheminf 3(1):32–32
85. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Petersson GA, Nakatsuji H, Li X, Caricato M, Marenich AV, Bloino J, Janesko BG, Gomperts R, Mennucci B, Hratchian HP, Ortiz JV, Izmaylov AF, Sonnenberg JLW, Ding F, Lipparini F, Egidi F, Goings J, Peng B, Petrone A, Henderson T, Ranasinghe D, Zakrzewski VG, Gao J, Rega N, Zheng G, Liang W, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Throssell K, Montgomery Jr JA, Peralta JE, Ogliaro F, Bearpark MJ, Heyd JJ, Brothers EN, Kudin KN, Staroverov VN, Keith TA, Kobayashi R, Normand J, Raghavachari K, Rendell AP, Burant JC, Iyengar SS, Tomasi J, Cossi M, Millam JM, Klene M, Adamo C, Cammi R, Ochterski JW, Martin RL, Morokuma K, Farkas O, Foresman JB, Fox DJ (2016) *Gaussian 16*, Wallingford, CT
86. He K (2021) Filter feature selection for unsupervised clustering of designer drugs using DFT simulated IR spectra data. ACS Omega 6(47):32151–32165
87. Linstrom PJ, Mallard WG, *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards and Technology, Gaithersburg MD, 20899.
88. Sano T (2021) *ClusterEnsembles*, https://github.com/tsano430/ClusterEnsembles, 2021–08–05.
89. *RDKit:* Open-source cheminformatics; http://www.rdkit.org
90. Karypis G, Eui-Hong H, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. Computer 32(8):68–75

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.