Title: Filter feature selection for unsupervised clustering of designer drugs using DFT simulated IR spectra data

Kedan He
Eastern Connecticut State University, Physical Sciences, 83 Windham St, Willimantic, CT 06226, USA.
Email: hek@easternct.edu
Phone: 860-465-5759
ORCID: 0000-0002-1953-4956

Abstract (250 words).
The rapid emergence of novel psychoactive substances (NPS) poses new challenges and requirements for forensic testing/analysis techniques. This paper aims to explore the application of unsupervised clustering of NPS compounds' infrared spectra. Two statistical measures, Pearson and Spearman, were used to quantify the spectral similarity and to generate the affinity matrices for hierarchical clustering. The correspondence of spectral similarity clustering trees to the commonly used structural/pharmacological categorization was evaluated and compared to the clustering generated using 2D/3D molecular fingerprints. Hybrid model feature selections were applied using different filter-based feature ranking algorithms developed for unsupervised clustering tasks. Since Spearman tends to overestimate the spectral similarity based on the overall pattern of the full spectrum, the clustering result shows the highest degree of improvement from having the non-discriminative features removed. The loading plots of the first two principal components (PCs) of the optimal feature subsets confirmed that the most important vibrational bands contributing to the clustering of NPS compounds were selected using NDFS feature selection algorithms.

Keywords: IR spectra classification, Designer drugs, Machine learning, Feature selection, Feature reduction, Pattern recognition, Hierarchical clustering

Conflicts of interest/Competing interests
The authors declare no competing financial interest.

Data Availability
Supporting information is provided, and the code used is available upon request to the corresponding author.

Acknowledgement
Computational resources were provided in part by the MERCURY consortium (http://mercuryconsortium.org/) under NSF grants CHE-1229354, CHE-1662030, and CHE-2018427. Clemson University is acknowledged for generous allotment of compute time on Palmetto cluster.

Introduction

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.[1, 2] According to the United Nations Office on Drugs and Crime (UNODC), as of December 2020, 126 countries had reported a total of more than 1,047 NPS.[1] Forensic analysis of NPS faces challenges such as diverse samples of unknown nature, insufficient quantity of evidence, the need for protecting the integrity of materials for criminal investigations and legal disputes, and the demand for in-field testing. Nondestructive, low-cost, and relatively easy-to-use vibrational spectroscopy techniques such as infrared (IR) and Raman are used to characterize the structure of organic molecules.[3-6] The most common method of spectral identification is library search, in which an unknown sample is compared to each spectrum in the library and a list of the best hits is returned based on a similarity metric.[7] The quality of library reference spectra and the robustness of similarity metrics limit the quality of library searches. The library must be large enough to contain spectra of samples similar to the unknown compound, and there must be a high degree of structural similarity between the "unknown" and the library substances in order to identify the "unknown" compound with confidence. The rapidity of emergence and the often-transient nature of some NPS compounds make it difficult to obtain a comprehensive spectral library. An alternative approach to identify never-before-seen NPS is to classify them based on structural similarity, as structurally similar compounds are likely to exhibit similar biological activities and spectroscopic characteristics.[8-10] However, the increasing complexity and diversity of NPS prevent systematic classification with respect to their structural similarity by visual inspection alone. The similarity of chemical structures could be quantified using its 2D/3D molecular fingerprint representation to calculate the Tanimoto coefficient. Conversely, representing spectra as a linear vector of intensities allows the quantitative comparison using statistical correlation coefficients.

Pattern recognition leverages information extracted from training samples to assign an unknown sample to a given class or category. Hierarchical clustering analysis is an unsupervised technique that provides multilevel nested results that can be used to help guide the identification of drug compounds that share a common structural and spectral feature. The IR spectrum is represented as a vector in a multidimensional space, where each dimension (feature) corresponds to a certain wavenumber and the corresponding absorbance (intensity). However, with the existence of a large number of features, a learning model tends to overfit and their learning performance degenerates. It has been verified that for complex analytical systems like vibrational spectroscopy data, it is very important and essential to conduct feature selection to gain better prediction performance.[11] Dimension reduction approaches use low-dimensional space to substitute the original high-dimensional variable space. For example, projection methods, such as principal component analysis (PCA) and partial least squares (PLS) are used to reduce the impact of collinearity, band overlaps, and redundant noise irrelevant to the property of interest by replacing the original variables with a few latent variables or principal components of larger variance.[12, 13] However, the latent variables are hardly interpretable compared to original variables. In contrast, feature selection is based on the assumptions of choosing a small number of variables that can improve the prediction performance, and provide easier interpretation. Unsupervised feature selection methods don't utilize label information and can be classified into filter model, wrapper model, and hybrid model according to different selection strategies. Filter feature selection algorithms are computationally efficient as they evaluate the relevance of a feature using certain statistical criteria and are independent of any clustering algorithm. A wrapper model evaluates the candidate feature subsets by the quality of clustering and are more biased to the chosen clustering algorithm. To alleviate the computational costs and benefit from the efficient filtering criteria, the hybrid model bridges the gap between the filter and wrapper models by utilizes filtering criteria to

select the candidate feature subsets then evaluates the quality of clustering of each candidate subset.[14] The subset with the highest clustering quality will be selected.

The underlying idea of ensemble feature selection is to combining the subsets of several individual feature selection methods (feature selectors) to obtain better or comparable results than using a single feature selection approach. When the data dimensionality is very high but the number of samples is relatively small, ensemble feature selection is used to improve the stability. A more appropriate (stable) feature subset is obtained by combining the multiple feature subsets of the ensemble, as the aggregated result tends to obtain more accurate and stable results, reducing the risk of choosing an unstable subset. The other main motivation is to increase the diversity: different feature selectors provide different enough outputs on the same sample of data and decrease the chance of inaccurate prediction of samples. The main issues involved in the process are 1) the individual feature selection methods to be used; 2) the number of different feature selection methods to use; 3) the aggregation method for feature subset generation. Ensemble feature selection can typically be categorized into the combination of labeled prediction, the combination of subsets of features, and the combination of ranking of features, which depends on whether the feature selector returns a subset of relevant features or an ordered ranking of all the features according to their relevance. When filter methods are used which return an ordered list of all features, a threshold must be chosen to reduce the dimensionality of the problem, which can become computationally expensive. The combined feature subsets, on the other hand, are generated by computing the intersection or the union of the ranked features. The intersection consists in selecting only those features which are selected by all the feature selectors, whereas the union consists in combining all the features which have been selected by at least one of the feature selectors. The potential issue is that it can lead to very restrictive sets of features (an empty set) or to select even the whole set of features, respectively. To alleviate the problem, a simple approach is to include a

subset of ranked features into the final ensemble only if it contributes to improving the learning tasks. Lastly, the relevancy of the final selection of features needs to be evaluated, which is possible using synthetic data where label information is known.

In this study, we performed unsupervised clustering analysis of a set of the most common NPS compounds whose IR spectra were simulated using density functional theory (DFT). We compared the correspondence of hierarchical clustering of NPS compounds into structurally distinct groups using 2D and 3D binary molecular fingerprints with cluster labels assigned according to generally accepted chemical/pharmacological classifications. Similarly, the spectral similarity can be quantified by statistical measurements such as Pearson and Spearman correlation coefficients. The clustering performance was quantified using Silhouette score,[15] Adjusted Rand index,[16] and Normalized mutual information.[17] Four filter-based feature selection methods developed for clustering tasks were explored in this study: Spectral feature selection (SPEC), Laplacian score (LS), Unsupervised Discriminative Feature Selection (UDFS), and Nonnegative Discriminative Feature Selection (NDFS). The class distributions in the NPS dataset are highly imbalanced. The oversampling technique, Synthetic Minority Oversampling Technique (SMOTE), was implemented in the feature selection process and its efficiency in improving the clustering performance of imbalanced datasets was examined. Finally, aggregated feature subsets were generated using a fusion-based ensemble technique. The optimal feature subset of IR spectroscopy for NPS compound clustering was identified. When comparing the loading plots of the first two principal components of the full range and dimension reduced datasets, it can be confirmed that the most discriminative features are retained even after the feature reduction.

Methods

Calculation of Infrared Spectra

According to the UNODC report up to December 2020, the majority of synthetic NPS are stimulants, followed by synthetic cannabinoid receptor agonists and psychedelics with a notable increase in synthetic opioids.[1] This classification is broadly defined according to their pharmacological targets: 1) stimulants mediate the actions of dopamine, norepinephrine, and/or serotonin as reuptake transporter inhibitors;[18-22] 2) cannabinoids primarily interact with G protein-coupled receptors;[23] 3) serotonergic psychedelics are mainly mediated by 5-HT_{2A} receptor agonism;[24, 25] 4) synthetic opioids and fentanyl analogs interact with G protein-coupled opioid receptors as partial to full agonists. 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 tryptamine), 6 sedatives (benzodiazepines), and 7 cannabinoids. Ten conformers were downloaded from PubChem for each compound.[26] PubChem3D provides low-energy conformers from a conformer model that samples the energetically accessible and (potentially) biologically relevant conformations of chemical structures using the average atomic pair-wise RMSD.[27] The geometry optimizations were performed using the Gaussian 16 program[28] using B3LYP level of density functional theory in combination with 6-311++G(d, p) basis set. Redundant conformers converged to the same structure were eliminated from the dataset. The harmonic vibrational wavenumbers of all conformers were determined at the corresponding optimized structures, which were confirmed to be local minima by check that there were no imaginary frequencies. To offset the systematic errors due to basis set incompleteness, neglect of anharmonicity, and incomplete treatment of electron correlation, single scaling factors were applied.[29] To validate the quality of the DFT simulated IR spectra and the dependence on basis

set, four basis sets (6-31G(d), 6-31+G(d, p), 6-31++G(d, p), and 6-311++G(d, p) were used and scaling factors were chosen from NIST database.[30] The theoretical vibrational frequencies and intensities were convoluted with a Lorentzian distribution, centered at the frequency and multiplied by the intensity. The full width at half-maximum (fwhm) of each distribution was set to 24 cm^{-1} on the basis of the estimated bandwidth observed in the NIST database.[31, 32] Finally, the resulting spectra were normalized with respect to the area under the curve and scaled from 0 to 1 and truncated from 400 – 4000 cm^{-1} range with 2 cm^{-1} interval. Quasi-constant features were further removed using *VarianceThreshold* in *sklearn* package with threshold value of 9.88e–06, excluding wavenumbers in 1854 – 2686 cm^{-1}, 3188 – 3386 cm^{-1}, and above 3786 cm^{-1} range. The final dataset is with size n = 930 and m = 1181.

Spectrum and chemical structure similarity measure

The similarity between two spectra, represented by vectors x_A and x_B, is characterized by two statistical measures. The Pearson's product moment correlation coefficient based on mean centered intensities:

$$r = \frac{\sum_i (x_{A,i} - \bar{x}_A)(x_{B,i} - \bar{x}_B)}{\sqrt{\sum_i (x_{A,i} - \bar{x}_A)^2}\sqrt{\sum_i (x_{B,i} - \bar{x}_B)^2}} \quad (1)$$

where $x_{A,i}$ and $x_{B,i}$ are the elements of the intensity vectors representing the spectra under comparison and $\bar{x}_A$ and $\bar{x}_B$ are the mean intensity values of spectra A and B, respectively. The Spearman's rank correlation coefficient is based on the mean of the intensity differences. A vector d is the difference between the ranks of $x_{A,i}$ and $x_{B,i}$ in their respective data set:

$$\rho = 1 - \frac{6\sum_i d_i^2}{n \cdot (n^2 - 1)} \quad (2)$$

where n is the number of elements in each vector. Both correlation coefficients range from –1 to +1. A positive Pearson score indicates that all data points are linearly associated, whereas two variables are monotonically related in Spearman correlation even if their relationship is not linear.

The degree to which two molecules are considered 'similar' depends on both their structural encoding and the similarity metric used. NPS compounds are often classified based on their pharmacological action, then further classified based on their common chemical scaffold, such as phenethylamines, piperazines, cathinones, and tryptamines, etc.[33] It is impractical to manually assign labels to NPS compounds as the number of them grows, as does their structural complexity. Molecular fingerprints that encode molecular structure as binary bit strings allow rapid scan for structural similarity/diversity using a bitwise comparison on pairs of molecules. Tanimoto coefficient is a widely used metric for molecular structural similarity quantification.[34] Two types of 2D molecular fingerprints were used in this study. Molecular ACCess System (*MACCS*) is a structural key fingerprint that encodes for the absence (0) and the presence (1) of a particular structural fragment, with the most commonly used being 166-bits long.[35] Morgan fingerprint is a circular fingerprint belong to the Extended Connectivity Fingerprints (ECFP) family of fingerprints that encodes heavy atoms into multiple circular layers up to a given diameter.[36] The RDKit implementation of Morgan with radius = 2 is roughly equivalent to ECFP4. An Extended Three-Dimensional FingerPrint (E3FP) is motivated by ECFP that draws concentrically larger shells and encodes the 3D atom neighborhood patterns from small to larger shells iteratively.[37] Lastly, the Maximum Common Substructure (MCS) similarity is calculated by identifying structural overlap by matching atomic elements and bond types.[38] Tanimoto values calculated using binary fingerprints will always have a value between 0 and 1, with 1 indicating identical and 0 indicating entirely different. Pertinent details can be found in Supporting Information. All fingerprints were calculated using RDKit software.[39]

Clustering Performance Measurement

The original dataset in (n, m) dimension was transformed into (n, n) affinity matrices using Pearson's or Spearman's correlation coefficient, where the value in *i*-th row and *j*-th column indicates the spectral similarity between sample *i* and *j*, and each sample is described by its similarity compared to all other samples. The affinity matrices were used as input and submitted to a Ward linkage clustering with Euclidean distance as the similarity metric for hierarchical clustering.

Silhouette score is an internal index in measuring the quality of a partition without external information. The optimal number of clusters *K* was determined by silhouette index (SI) analysis,[15] which is a measure of how well cluster members belong to their respective clusters, averaged over all samples:

$$SI_k = \frac{1}{n}\sum_{i=1}^{n}\frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{3}$$

where n is the total number of points, a_i is the average distance between point i and all other points in its own cluster, and b_i is the minimum of the average dissimilarities between i and points in other clusters.

External indices measure the similarity between the output of the clustering algorithm and the correct partitioning of the dataset. Different clustering trees were compared with each other using the adjusted Rand-Index (ARI).[16] Let $U = \{u_1, u_2, \dots, u_R\}$ and $V = \{v_1, v_2, \dots, v_C\}$ represent the external cluster label and that determined by the cluster algorithm, n_{ij} is the number of objects belonging to both subset , u_R and v_j, the ARI is calculated:

$$\mathrm{ARI} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right]}{\binom{n}{2}}}{{}^1\!/_2\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \frac{\left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right]}{\binom{n}{2}}} \tag{4}$$

When two sets of cluster labels have a perfect one-to-one correspondence, the ARI equal to unity. The normalized mutual information (NMI) quantifies the mutual dependence between two random variables based on concepts of information theory:

$$\mathrm{NMI}(C_i, C_j) = \frac{I(C_i, C_j)}{\sqrt{[H(C_i), H(C_j)]}} \quad (5)$$

where C_i and C_j are cluster assignments of the points generated from feature subsets of feature selector i and j, respectively. Mutual information $I(C_i, C_j)$ is given as $H(C_i) - H(C_i|C_j)$. $H(C)$ is the Shannon entropy of C, and $H(C_i|C_j)$ 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.

All hierarchical clusterings are generated using the *fcluster* and *dendrogram* in *scipy.cluster.hierarchy* package. Heatmaps are generated using *seaborn* package, SI, ARI, and NMI are computed using *sklean.metrics* package.

Filter Feature selection models and ensemble method

Four filter feature selection models were used to rank the features according to certain criteria. Spectral feature selection (SPEC) algorithm studies how to select features according to the structure of the adjacency matrix *W* and graph *G* induced from the samples' similarity matrix *S*.[40] The similarity matrix *S* is calculated uses the Radial-Bases Function as a similarity function between two samples x_i and x_j:

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$$

The main idea behind SPEC is that the features consistent with the graph structure are assigned similar values to instances that are near to each other in the graph. Therefore, these features should be relevant since they behave similarly in each similar group of samples.[14] Laplacian score (LS) is a

special case of SPEC that selects the features most consistent with the Gaussian Laplacian matrix use a different ranking function and very efficient with respect to the data size.[41] The Unsupervised Discriminative Feature Selection (UDFS) algorithm simultaneously exploit the discriminative information and feature correlation to select discriminative features in batch mode.[42] Lastly, the Nonnegative Discriminative Feature Selection (NDFS) algorithm utilizes spectral clustering to obtain cluster label indicators as well as a nonnegative constraint into the objective function.[43] The sparse feature selection matrix is formulated as $l_{2,1}$-norm minimization term and solved iteratively. All four feature filter models were implemented in *scikit-feature* package.[44]

The ensemble method used in this study can be summarized as follow:

1. Rank all the features using each feature selector.

2. Evaluate the clustering performance by increasing the size of top-ranked feature subsets.

3. Identify the feature subsets corresponding to the optimal clustering performance using each feature selector.

4. Identify the common overlap (intersect) feature ensembles using different feature selectors.

Table 1. Nomenclature

D	Dataset
n	Sample size
m	Number of features
x_j	j^{th} sample
f_i	i^{th} feature
F	Selected feature set
l	Number of selected features
K	Number of clusters
C_k	k^{th} cluster

Model training, validation, and performance evaluation

In this work, we used a class-imbalanced dataset to reflect the distribution of NPS in the real-world market. Changes in supply, manufacturing, and regulatory regulations all have an impact

on the market's continuously moving trend. Machine learning on class-imbalanced data, on the other hand, is biased in favor of the majority class, which is compounded by the high-dimensionality of the feature space. SMOTE (Synthetic Minority Oversampling TEchnique) is a popular oversampling technique that produces class-balanced data. In this study, we also looked into whether applying SMOTE improves clustering by computing the feature importance score with and without SMOTE.

The feature importance ranking as calculated by dividing the dataset by ten-folds and averaging the results over ten iterations. The scores were standardized to a range of 0 – 1 before being averaged for ease of comparison. Each iteration calculates the feature importance score using 10% of the dataset and is repeated 10 times using a different 10% of the dataset. The arithmetic means of the scores acquired from 10 iterations was used to establish the overall feature ranking. Similarly, the clustering evaluations were repeated 5 times, with a 5-fold split of the dataset for each feature subset chosen using different selectors, and the average ARI and AMI were calculated. The confidence intervals of ARI and AMI were determined by 250 bootstrap iterations using a sample that is 15% of the size of the dataset for the final ensemble feature subset comparison.

Results and Discussion

Comparison with experimental gas IR spectra using quantitative correlation measurements

To assess the quality of the DFT simulated IR spectra, and the dependence on the basis sets, the correlation coefficients are calculated from the comparison of DFT and experimental spectra. Usually the scaling factors are derived by minimizing the RMSD of peak wavenumbers, however, the exact combination of DFT/6-31++G(d, p) and DFT/6-311++G(d, p) are not available in the CCCBDB[30] database, hence the same scaling factor 0.964 for DFT/6-31+G(d, p) was used for the other two larger basis sets. The optimal scaling factors were determined by a maximization of the

correlation coefficient to assess the usefulness of using two correlation coefficients as a quantitative measure of spectral similarity. **Table 1** summarizes the average results for six compounds for which the experimental gas IR spectra are available at NIST.[45] **Figure 1** shows a visual spectra comparison of the IR spectra of the lowest-energy conformer of each compound. The corresponding compound-wise correlation coefficients are listed in **Table 2**. Spearman's correlation coefficient consistently gives a comparatively higher score with or without using the scaling factors. In terms of reproduction of the experimental spectra, the unscaled 6-311++G(d, p) spectra resulted in the highest spectral correlation coefficients ($r = 0.549$, $\rho = 0.779$). Applying scaling factors resulted in a somewhat larger increase for Pearson, indicating that Pearson is more sensitive to the exact wavenumbers of peak position. The optimization of both statistical measures gave rise to scaling factors that are only differed insignificantly from the literature and follows the general trends that larger the basis set higher the scaling factor using Spearman. At the same time, application of the optimized scaling factor improves the Pearson more significantly, in contrast to the marginal improvement of Spearman. Finally, the scaling factor 0.971 was used for generation of the IR spectra dataset using DFT/6-311++G(d, p) level of theory.

Table 1. Average correlation coefficient of DFT spectra compared to experimental data shown in Figure 1.

	Pearson r			**Spearman ρ**		
Basis set	(unscaled)	(scaled)[a]	Optimized (*Scaling factor*[b])	(unscaled)	(scaled)[a]	Optimized (*Scaling factor*[b])
6-31G(d)	0.455	0.723	0.784 (*0.967*)	0.739	0.859	0.863 (*0.964*)
6-31+G(d, p)	0.497	0.676	0.798 (*0.965*)	0.676	0.860	0.865 (*0.967*)
6-31++G(d, p)	0.497	0.676	0.798 (*0.965*)	0.763	0.860	0.865 (*0.967*)
6-311++G(d, p)	0.549	0.643	0.798 (*0.977*)	0.779	0.857	**0.866 (*0.971*)**

[a] Scaling factor 0.960 and 0.964 are used for 6-31G(d) and 6-31+G(d, p) as reported in CCCBDB database. For larger basis sets, the scaling factor 0.964 was applied.
[b] The optimized scaling factors were obtained from a maximization of the statistical correlation coefficients. The scaling factor 0.971 was used for generation of the IR spectra dataset using 6-311++G(d, p) basis set.

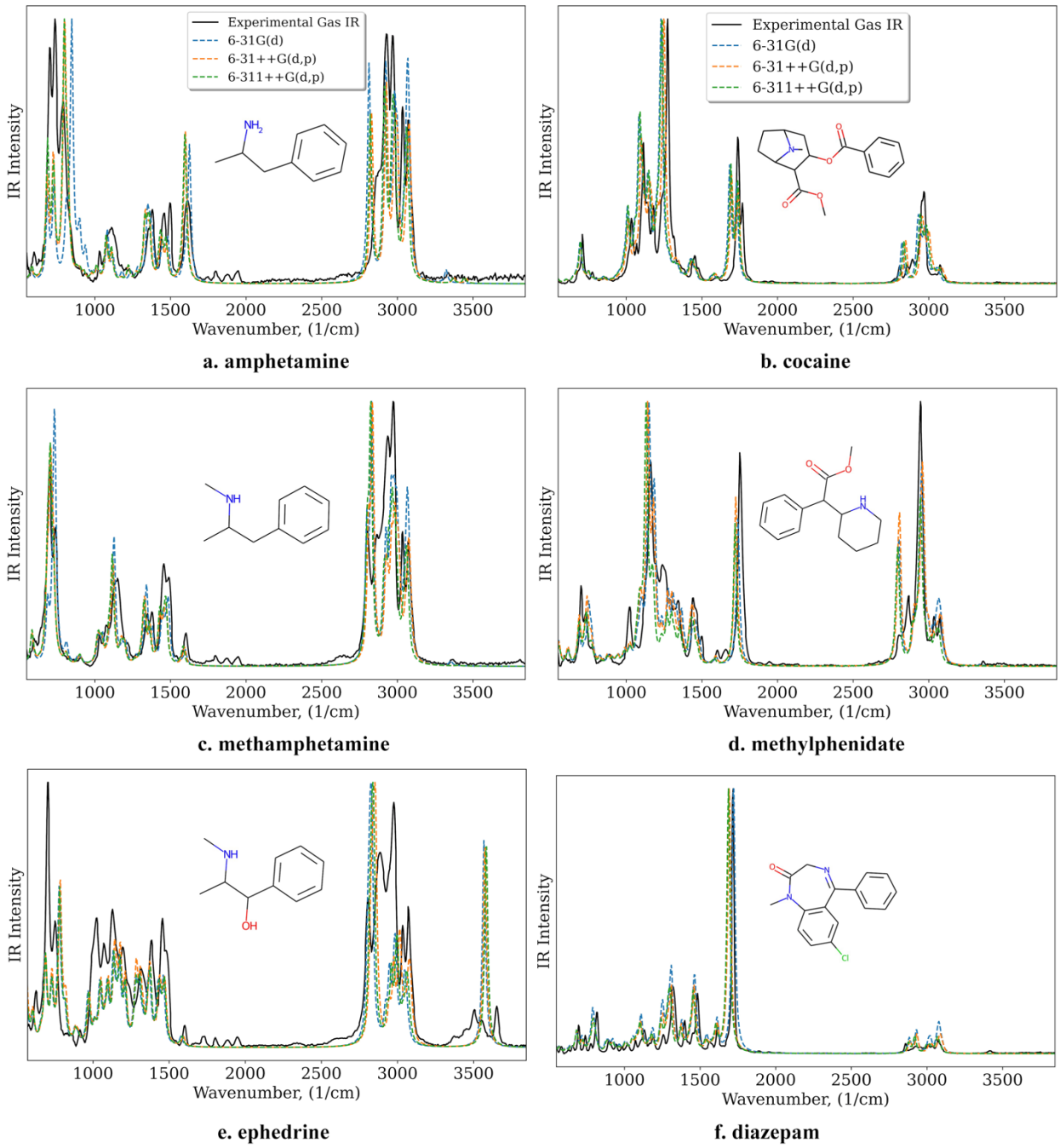

Figure 1. Infrared spectra of the lowest-energy conformers calculated by DFT using different basis sets in comparison with the experimental spectra.

Diazepam stands out as an interesting case that it resulted in the largest Pearson correlation coefficient and the lowest Spearman correlation coefficient among all compounds. From **Figure 1.f** it can be seen that the experimental spectrum of diazepam is dominated by carboxyl bond stretching

band around 1760 cm^{-1} followed by the benzene ring modes around 1500 cm^{-1} and 1350 cm^{-1}. The scaled DFT spectrum placed those bands at lower wavenumbers, and the optimization significantly increased the Pearson correlation by matching the most intense bands. It seems that reproduction of a dominant feature in the reference spectrum has a decisive impact on the Pearson correlation coefficient. Similarly, in the case of cocaine in **Figure 1.b**, the Spearman correlation is way above the average value of all compounds for the scaled spectrum, but the Pearson correlation is lowered compare to that of the unscaled spectrum. The DFT spectrum overestimated the intensity for the C-H bending bands below 1300 cm^{-1} and underestimated the intensity for C-H stretching band above 2800 cm^{-1}. Overall, it suggested that the Spearman correlation coefficient could provide a better estimate of the overall similarity of the spectra, whereas Pearson correlation coefficient is more sensitive to the peak position and intensity of the dominant features.

Table 2. Correlation coefficients of spectra shown in Figure 1 using 6-311++G(d, p) basis set.

	Pearson r			**Spearman ρ**		
Compounds	(unscaled)	(scaled)[a]	Optimized (*Scaling factor*[b])	(unscaled)	(scaled)[a]	Optimized (*Scaling factor*[b])
amphetamine	0.570	0.794	0.795 (*0.965*)	0.772	0.836	0.845 (*0.975*)
methamphetamine	0.657	**0.835**	0.843 (*0.962*)	0.803	0.858	0.866 (*0.972*)
ephedrine	0.473	0.575	<u>0.612</u> (*0.981*)	<u>0.673</u>	0.866	0.873 (*0.957*)
cocaine	**0.796**	0.638	0.885 (*0.990*)	**0.859**	**0.901**	**0.909** (*0.973*)
methylphenidate	0.541	0.678	0.704 (*0.983*)	0.821	0.869	0.882 (*0.974*)
diazepam	<u>0.255</u>	<u>0.337</u>	**0.951** (*0.981*)	0.745	<u>0.813</u>	<u>0.822</u> (*0.973*)

[a] Scaling factor 0.964 was used for DFT/6-311++G(d, p) calculated IR spectra.
[b] The optimized scaling factors were obtained from a maximization of the statistical correlation coefficients for quantitative quantification of spectral similarity. The highest and lowest correlation coefficients are marked in bold and underlined, respectively.

These two correlation coefficients also differ in terms of their sensitivity to conformational changes, as see in **Figure 2** and **Table 3**. Two conformers for which Pearson correlation

coefficients are distinctively different are shown for four compounds, along with the optimized structures. Conformational changes usually result in changes in band intensity, as see in **Figure 2.a** and **Figure 2.b**. For methamphetamine, the Pearson correlation increased from 0.767 to 0.860 but the Spearman correlation stays the same. In case of ephedrine in **Figure 2.c**, the rotation of hydroxyl group towards the benzene ring in conformer 2 decreased the O-H stretching band intensity and right-shifted to higher wavenumber, whereas the rotation of amine side chain right-shifted the benzene ring modes and left-shifted the N-H bending band. As for methylphenidate in **Figure 2.d**, the lower Pearson correlation of conformer 1 is caused by the poor match of stretching bands of ester functional group. However, these spectral differences are not reflected when comparing the Spearman correlation coefficients of the two conformers.

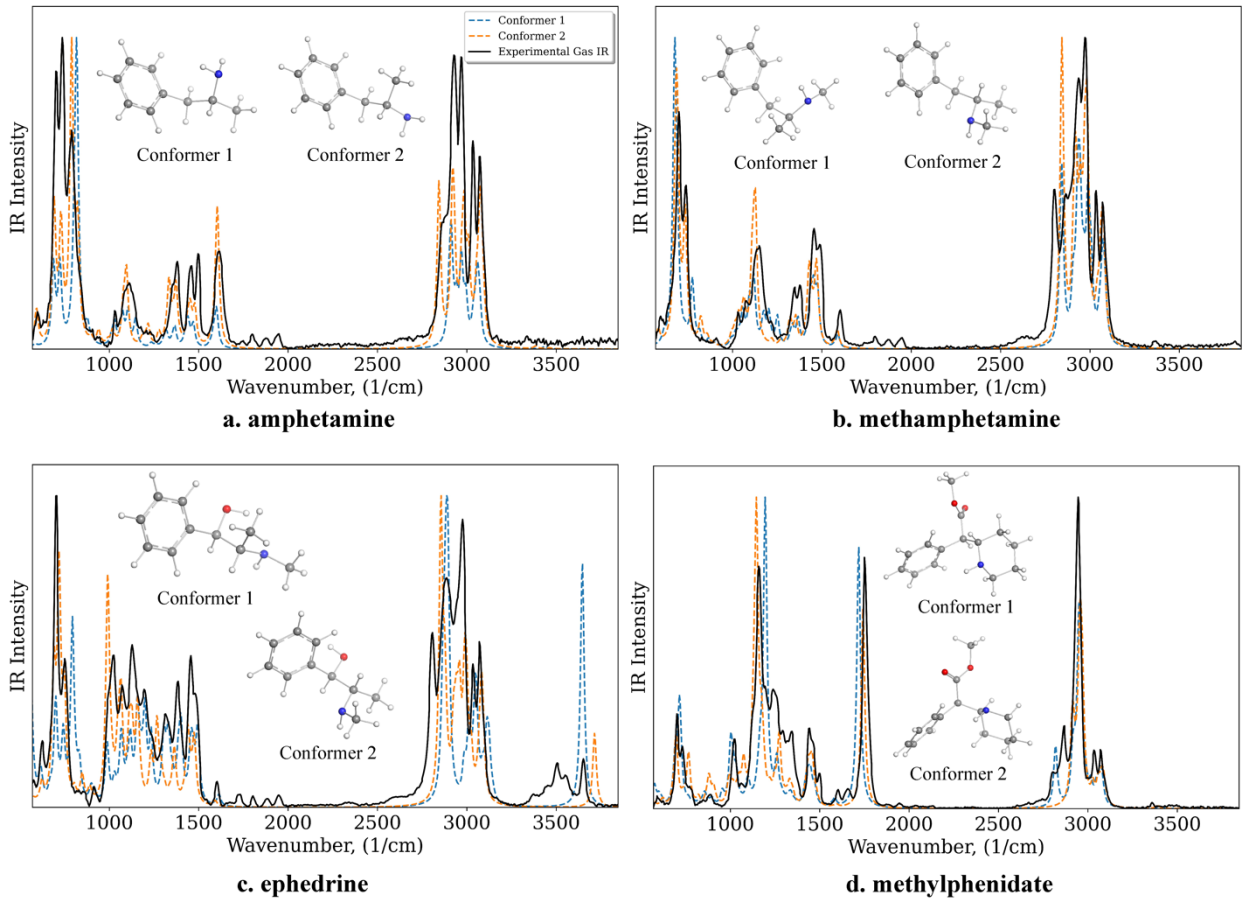

Figure 2. DFT/6-311++G(d, p) infrared spectra of two conformers in comparison with the experimental spectra.

Table 3. Correlation coefficients[a] of the spectra shown in Figure 2 using 6-311++G(d, p) basis set.

Compounds	Conformer 1		Conformer 2	
	Pearson r	**Spearman ρ**	**Pearson r**	**Spearman ρ**
amphetamine	0.674	0.829	0.820	0.848
methamphetamine	0.767	0.853	0.860	0.853
ephedrine	0.612	0.873	0.802	0.849
methylphenidate	0.638	0.888	0.808	0.882

[a] Optimized scaling factor 0.971 was used. Full list of conformers results is available in supporting information.

Correspondence of structural and spectral similarity

Clustering is used to find groups of objects that are more similar to each other than to other clusters. There are many clustering algorithms available, choosing a clustering technique is often a trial-and-error process that is very dependent on the data set.[46-51] However, determining the best clustering algorithm for NPS IR data is outside the scope of this study, and hierarchical clustering is used for its ease of visualization of cluster relationships on a dendrogram and heatmap. The Ward linkage method forms clusters by combining two clusters that result in the least increase in variance from an iterative ANOVA test.

The correspondence of the structural and spectral similarities of NPS compounds with respect to the manually assigned class label is investigated. Because conformational changes have no effect on the compound's 2D SMILES, only the lowest-energy conformer is used to generate the 3D description in this analysis. The chemical structural diversity of the 127 unique NPS compounds can be characterized by the calculation of the Tanimoto similarity score for the 8001 pairs ($n(n-1)/2$). **Figure 3** shows the right-skewed distributions of all structural similarities, and in contrary, the spectra similarities exhibit left-skewed distributions, with 50% of the compound pairs having a Spearman score equal to or greater than 0.745. There are also significant discrepancies amongst structural fingerprint approaches, with Morgan and E3FP being even more right-skewed, with mean similarity scores of 0.170 and 0.132, respectively. **Figure 4** compares the use of structural and

spectral similarity methods in retrieving a query compound's nearest neighbors (hits 1-5). The accuracy ratio is calculated by dividing the total number of retrieved compound pairs with the same assigned class label by the total number of compound pairs. Overall, 2D molecular fingerprints provide consistent better performance even when more top hits are chosen, whereas spectral similarity search delivers inferior and rapidly declining performance in its ability to retrieve compounds from the same class. The 3D description E3FP does not enhance the identification of structurally similar substances specified by commonly accepted NPS categorization.

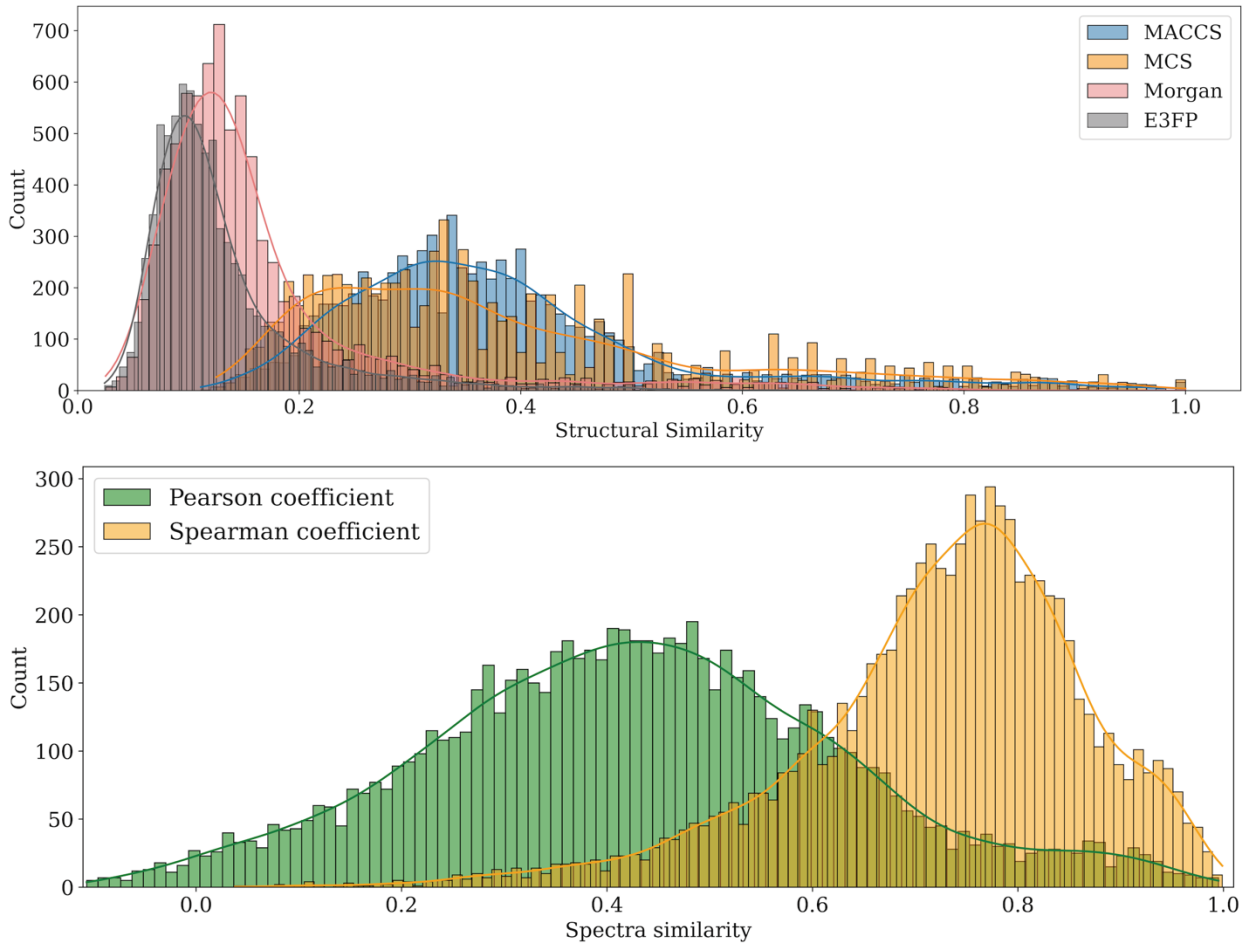

Figure 3. Frequency distribution of Tanimoto coefficient and spectral correlation coefficients of pair-wise comparison of drug compounds.

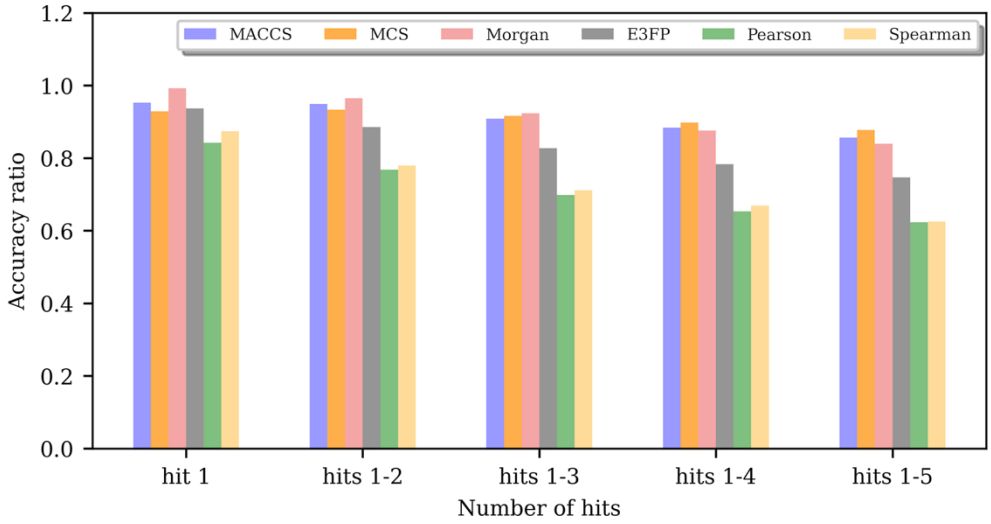

Figure 4. Accuracy ratio of retrieving compound(s) of the same class.

As two examples, **Figure 5 and 6** display the top-hits from the spectral similarity and structural similarity searches for the query molecule THC (delta-9 tetrahydrocannabinol) and MDA (3,4-Methylenedioxyamphetamine). THC is the primary component of the marijuana plant that produces psychoactive effects and a Schedule II substance. The structural and spectral similarity searches return the same top-three hits: DMHP and Synhexyl, synthetic analogues of THC, and CBN (Cannabinol), a derivative of THC. Despite a relatively high MACCS similarity score (0.690), CBD (cannabidiol), a major non-psychotropic constituent of cannabis, was assigned low similarity scores based on MCS (0.353) and both spectral correlation coefficients (r = 0.500, ρ = 0.528). It is evident that when a structural change causes a significant departure of the most intense bands, the spectral correlation coefficients can distinguish the spectral differences properly. On the other hand, Spearman severely overestimate the spectral similarity and subsequently retrieve incorrect hits as shown in **Figure 6.b**, that 2C-T-30 was assigned the highest similarity value (ρ = 0.963) compare to that of MBDB (*N*-methyl-1,3-benzodioxolylbutanamine) and MDMA (3,4-Methylenedioxy methamphetamine), two stimulants of the amphetamine family.

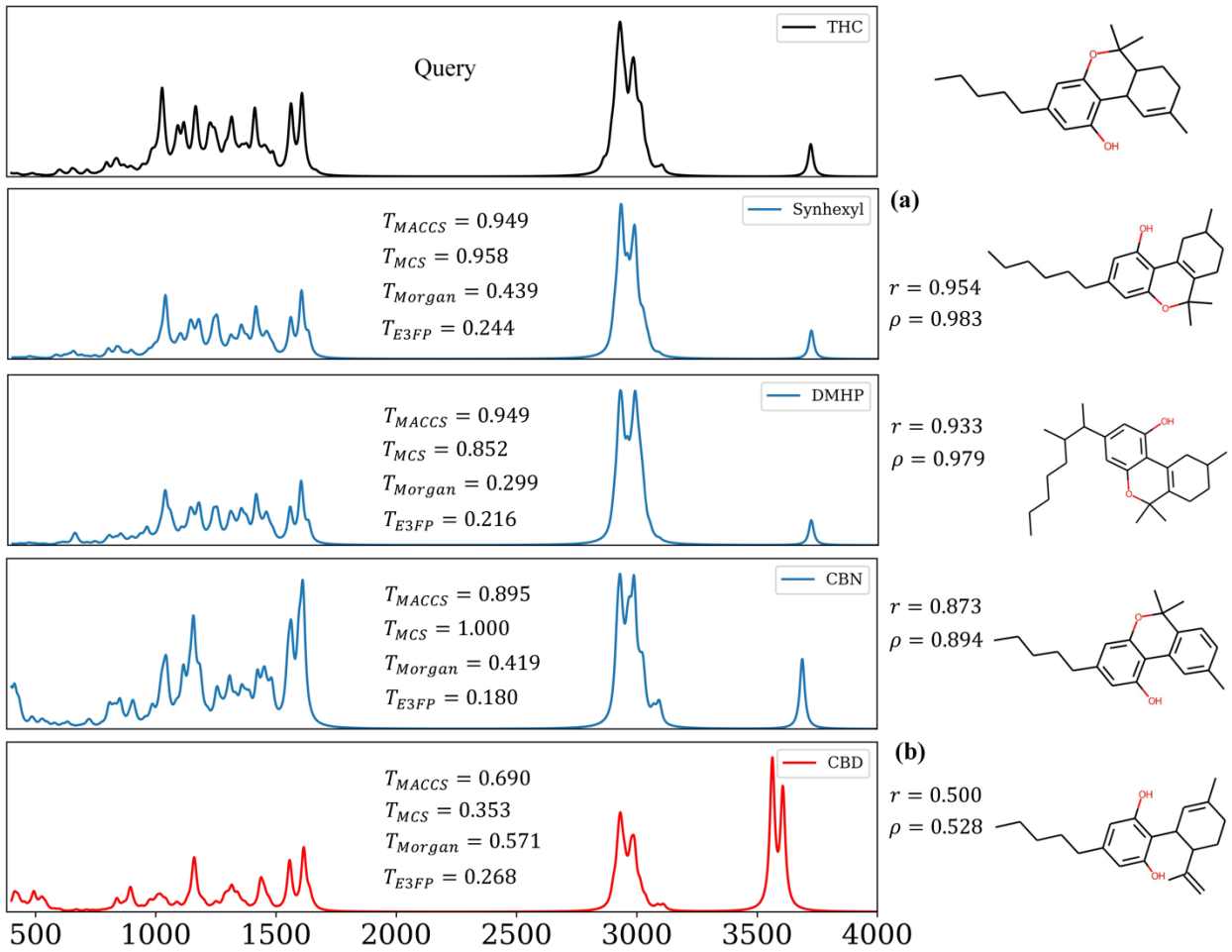

Figure 5. Search results for query compound THC. (a). Query and top three hits from spectra similarity search. (b). Spectrum of CBD and similarity scores.

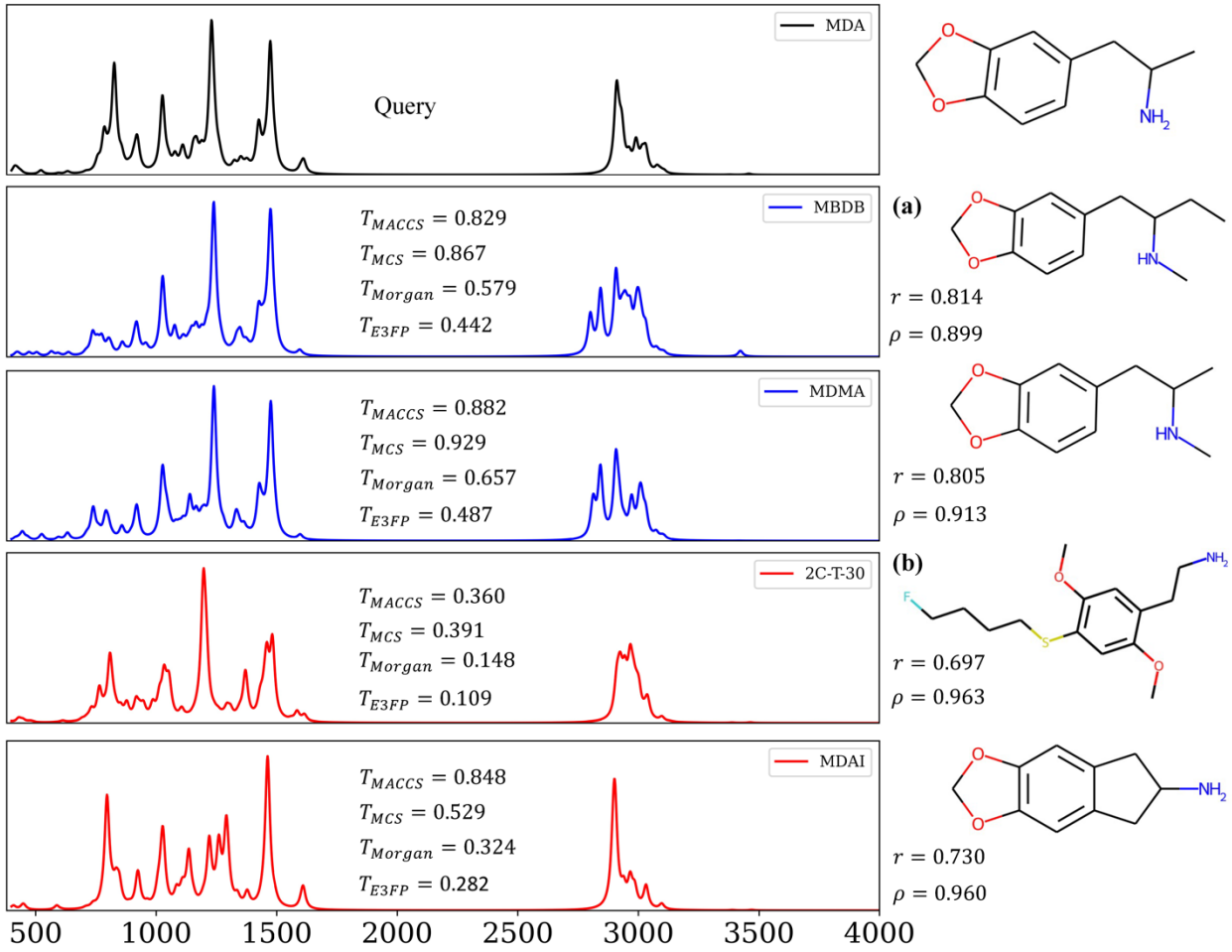

Figure 6. Search results for query compound MDA. (a). Top-two hits using Pearson spectra similarity. (b). Top-two hits using Spearman spectra similarity.

Finally, all clustering trees were compared using three measures: silhouette scores based on internal proximity of information intrinsic to the data, ARI and NMI assessed by comparing clustering partitions with external class label. **Figure 7** shows the silhouette scores of Pearson and Spearman clustering trees calculated for *K*=2 to *K*=50 clusters. The silhouette score increases as the intra-cluster distance decreases and the inter-cluster distance increases, so that the optimal number of clusters K will correspond to the highest silhouette value. The drawback of the Spearman correlation coefficient is also manifested in the highest silhouette score when dividing all samples into two clusters, as the Spearman similarity score is frequently exaggerated for many compound pairs. The silhouette analysis plots and heatmaps of all clustering trees are available in Supporting

information. According to all three clustering measures show in **Figure 8**, the MCS similarity clustering tree based on shared core chemical fragments outperformed all other clustering trees, followed closely by the substructural key fingerprint MACCS clustering tree. Consistent with the preceding analysis, the 3D description E3FP could not be used to sufficiently classify NPS substances based on their pharmacological/structural classification, indicated by the lowest silhouette score of 0.21. With a little higher silhouette scores than that of E3FP but a lower ARI, 0.36 and 0.32, respectively, there is no significant difference between the two spectral similarity clusterings. Because Spearman correlation scores appear to be more influenced by the general pattern of the IR spectra instead of the most intense bands, this clustering should benefit more from feature selection by removing non-discriminative features.

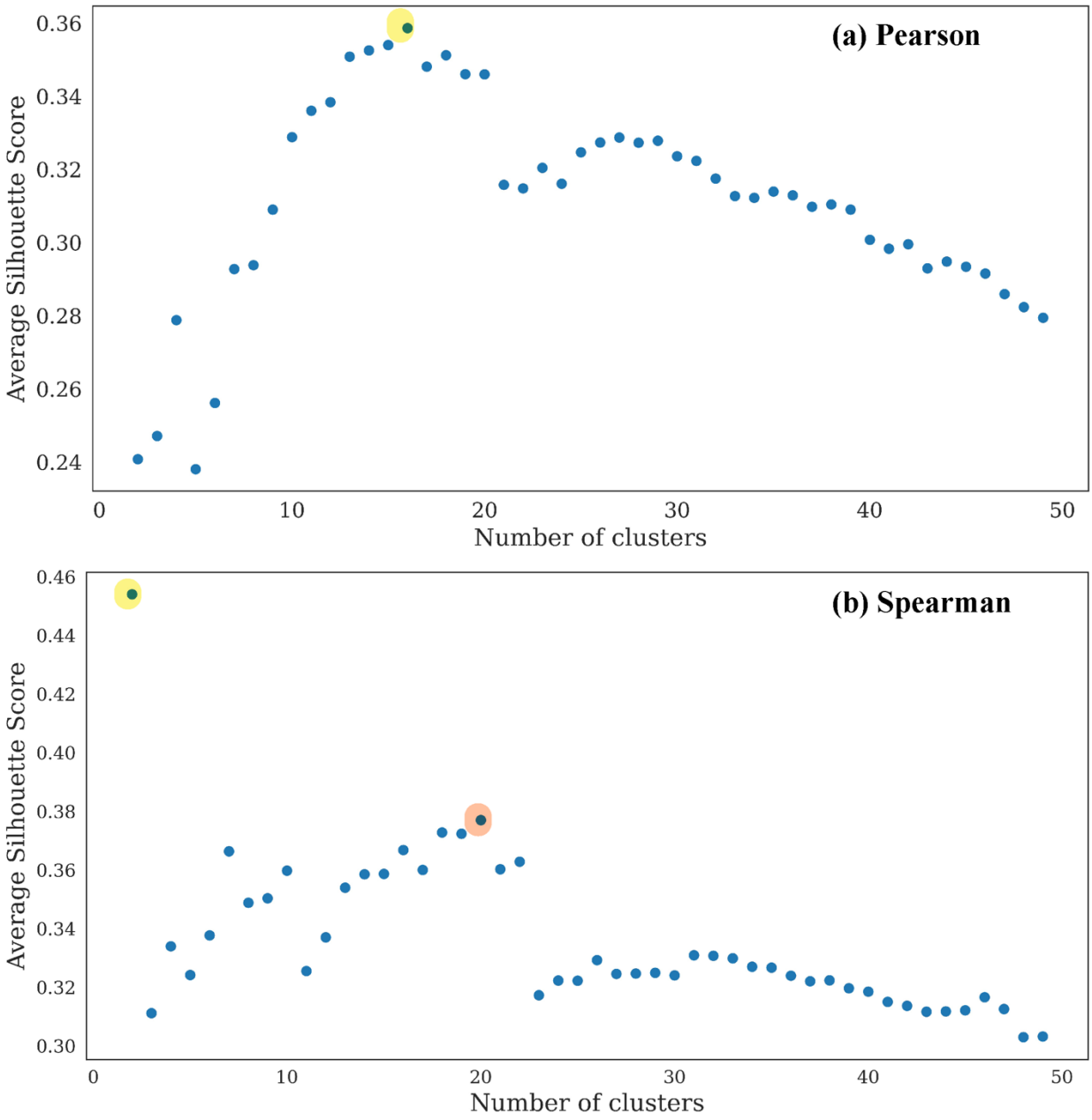

Figure 7. Silhouette scores as a function of cluster $2 < K < 50$ of spectral similarity clustering trees.

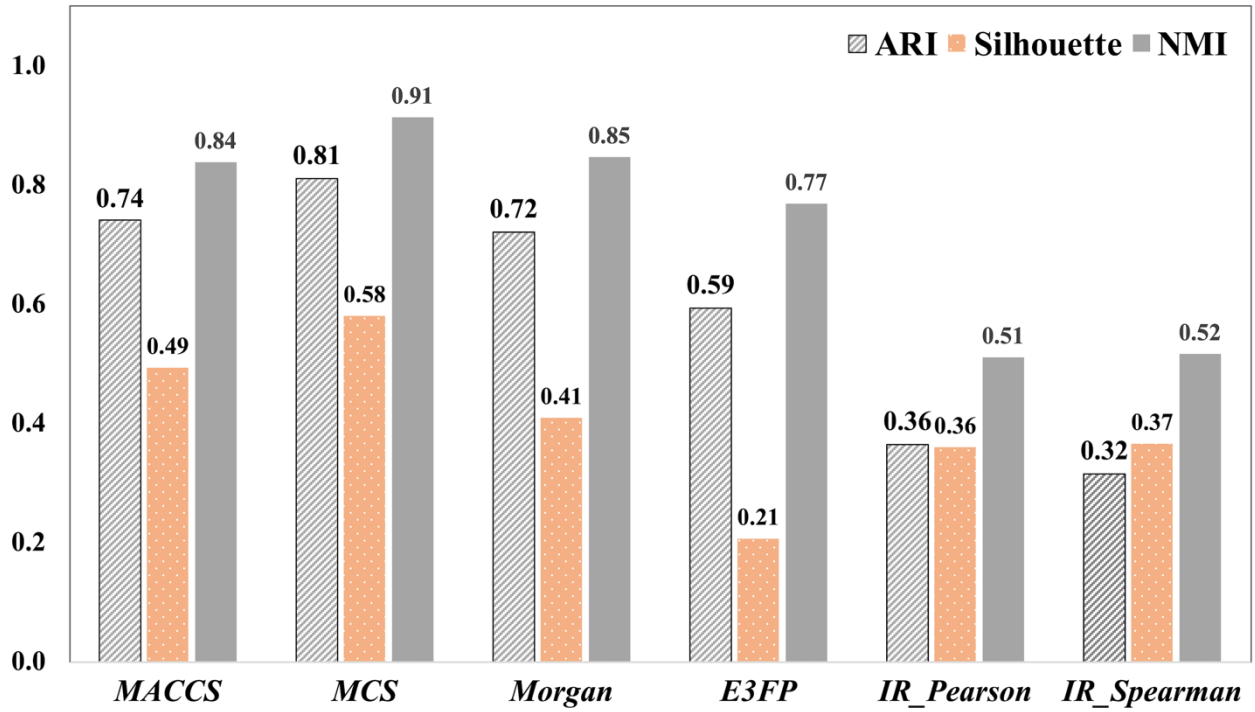

Figure 8. ARI, Silhouette score, and NMI of all clustering trees evaluated using external assigned class label.

Feature selection evaluated using hierarchical clustering

Figure 9 shows the scaled feature important score plots as a function of vibrational wavenumber to assist understand the differences in feature subsets generated by the four feature selectors and the effect of applying SMOTE. **Table 4** also list the top-ten features chosen by all four selectors. As shown in **Figure 9.a and b**, applying SMOTE produces the largest differences in the feature subsets identified using SPEC and LS. For example, features around 3646 - 3654 cm^{-1} are given higher importance scores compared to when SMOTE was not used. Another distinction is that different feature selectors analyze and sample features in varying manners. When the feature importance threshold is gradually lowered to subsequently include more top-ranked features, the first three feature selectors (SPEC, LS, and UDFS) sample a group of features in a localized fashion from one region to another. In contrary, the NDFS algorithm produces the most "sparse" selection

by assigning fewer features with very high importance, resulting in more scattered feature selection across the whole spectrum.

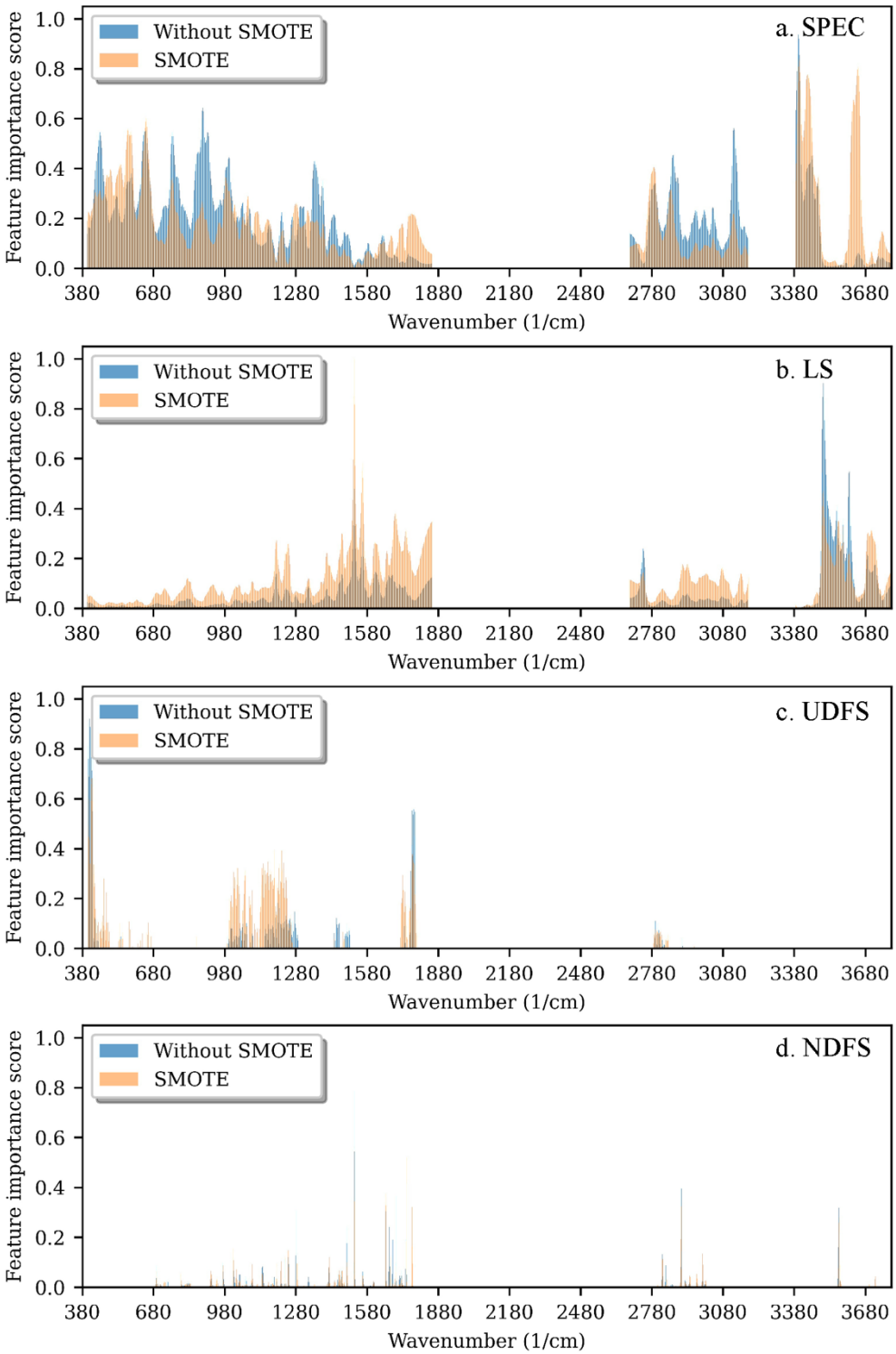

Figure 9. Scaled feature importance score from different feature selector, with or without applying SMOTE.

Table 4. Top-10 features (in cm^{-1}) selected by four filter methods

SPEC		**LS**		**UDFS**		**NDFS**	
w/o SMOTE	SMOTE	w/o SMOTE	SMOTE	w/o SMOTE	SMOTE	w/o SMOTE	SMOTE
3398	3400	3502	1526	410	404	1526	1526
3400	3402	3504	1524	404	422	1528	1746
3396	3648	3500	1528	412	418	1746	1658
3394	3398	3506	1522	414	416	2904	1528
3402	3650	3508	1560	402	424	1702	1660
3392	3646	3498	1530	424	408	1658	2904
3404	3404	3510	1558	418	420	3568	1770
3390	3436	3512	1562	406	414	1282	3570
888	3438	1526	3502	420	402	1660	3568
3388	3440	3612	1556	408	406	1674	2902

The clustering results of different feature selectors with affinity matrices calculated using Pearson and Spearman correlation coefficients are summarized in **Figure 10**. From this analysis, we have the following observations. First, SMOTE appears to provide no significant improvement of the clustering tasks, as the ARI changes as number of features selected follows the same pattern for all four selectors with or without apply SMOTE. Second, feature selection using LS and NDFS effectively reduce the feature number and improved the clustering performance in comparison to the baseline that uses the full range features. The NDFS algorithm exploit discriminative information by evaluating features jointly results in higher ARI and NMI while using the smallest feature subset. Lastly, confirming the previously analysis, the Spearman clustering benefited more from using feature selectors.

The feature subsets that resulted in the highest ARI score of Spearman clustering trees using LS and NDFS algorithms are combined together by compute the union or intersection of both sets. The Ensemble 2 include the intersection of the optimal feature subsets from LS and NDFS with SMOTE incorporated. The number of features selected in each feature subset, clustering results evaluated using ARI and NMI are summarized in **Table 5**. Clustering using Ensemble 2 feature subset give the highest ARI and NMI among all four ensembles. However, from the number of

features in ensembles, it is clear that most features selected by NDFS are also selected by LS, and the ensemble feature subset clustering performance is comparable to that of the individual selectors. Given the above observation, the NDFS algorithm is able to select more informative features compare to all other filter-based feature selection models.

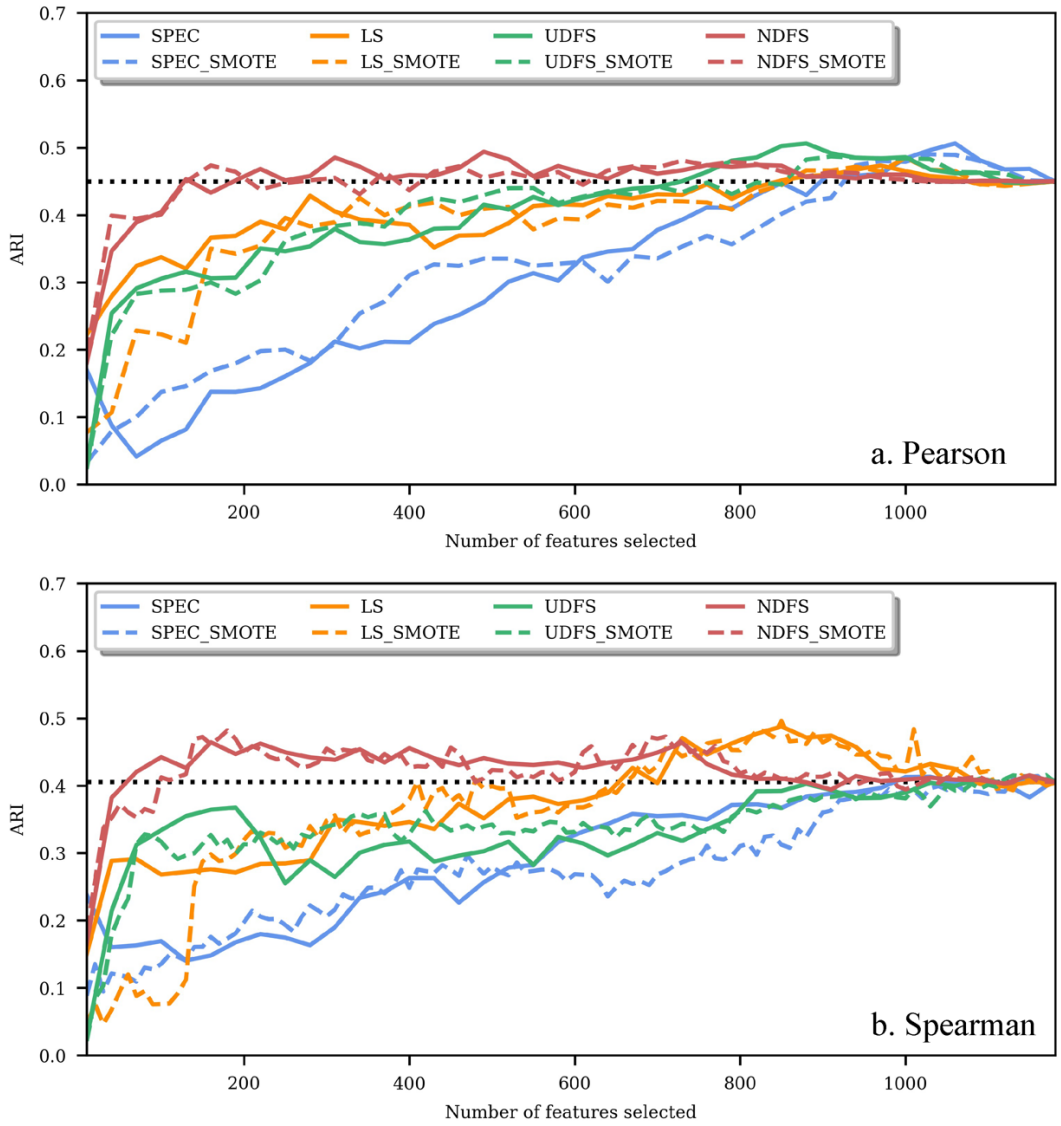

Figure 10. Clustering ARI when increasing number of top features using different feature selectors**.**

Table 5. Clustering results of different feature subsets using Spearman correlation coefficient for affinity matrix

		ARI (95% CI)			AMI (95% CI)		
	No. features	**mean**	**lower**	**upper**	**mean**	**lower**	**upper**
LS	850	0.470	0.362	0.600	0.623	0.546	0.710
LS_SMOTE	840	0.474	0.365	0.598	0.628	0.550	0.703
NDFS	180	0.460	0.348	0.584	0.623	0.551	0.691
NDFS_SMOTE	180	0.463	0.346	0.582	0.630	0.549	0.708
Ensemble1[a]	166	0.465	0.362	0.588	0.626	0.553	0.696
Ensemble2[a]	175	**0.472**	0.354	0.601	**0.637**	0.562	0.717
Ensemble3[b]	120	0.467	0.361	0.600	0.631	0.555	0.708
Ensemble4[c]	228	0.462	0.355	0.584	0.626	0.550	0.701
Baseline[d]	1181	*0.415*	0.309	0.542	*0.586*	0.504	0.677

[a] Ensemble 1 and 2 include the intersection of the optimal feature subsets from LS and NDFS without or with SMOTE applied, respectively. [b] Ensemble 3 is the intersection of ensemble 1 and 2. [c] Ensemble 4 is the intersection of the union feature subsets of LS and NDFS. [d] The baseline clustering trees use the full range features.

PCA was carried out to reduce the dimensionality of the full range (baseline) and the ensemble 2 subset datasets, and the loading plots of the first two principal components (PCs) are shown in **Figure 12**. When the ensemble 2 feature subset was used, the total explained variance (TEV) increased from 16.95% and 13.50% to 24.88% and 13.76% for PC1 and PC2, respectively. The same set of most important vibrational bands contributing to the clustering of NPS compounds was selected according to the PC1 and PC2 loadings plots of the baseline and ensemble 2 datasets.

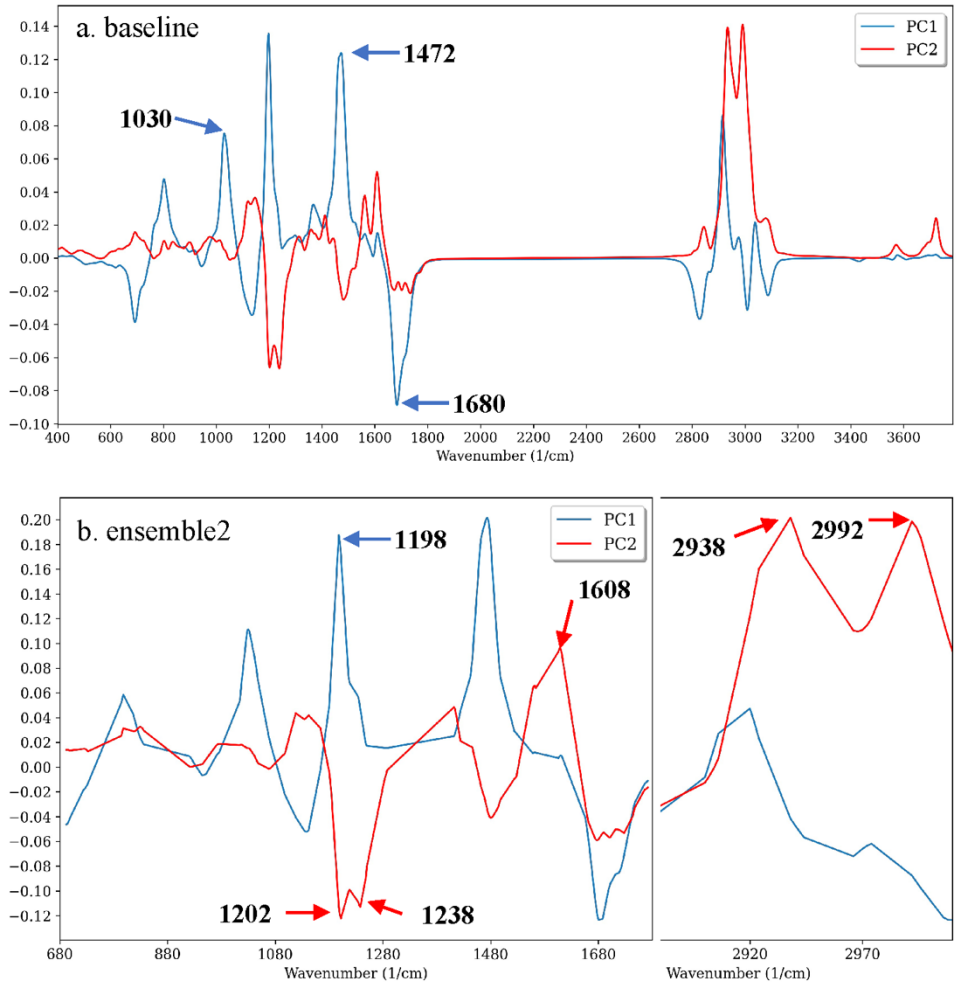

Figure 12. First and second PCs loadings plot using the full range and feature reduced datasets.

Conclusion

The IR spectra of NPS compounds were calculated using DFT use various basis sets in this study. The spectral similarity was quantified using two statistical measures: Pearson's product moment correlation and Spearman's rank correlation coefficients. When using the gas-phase experimental spectra as a reference, it is shown that Pearson is more sensitive to the intensity and peak position of the most intense bands, and to the spectral changes caused by conformational changes. On the other hand, Spearman is better suited to describe the overall pattern of the full spectrum, but tends to overestimate the similarity of the spectra. The ability to retrieve compounds of the same structure/pharmacological class using spectral similarity searches was evaluated

compared to structural similarity searches using 2D/3D molecular fingerprinting. Hierarchical clustering using MCS similarity proved to be a suitable method to group NPS compounds into clusters with different maximum common substructures and gave the best partition based on ARI and NMI calculated using externally assigned class labels. The clustering trees generated using the two spectral similarities showed the lowest agreement with the external class labels. Since Spearman tends to overestimate the spectral similarity based on the overall pattern of the full spectrum, it is expected to benefit more from feature selection to remove nondiscriminatory features. Four filter-based feature ranking algorithms were evaluated and SMOTE was applied to balance the class distribution of the dataset in the feature importance calculation. When Spearman correlation coefficient was used in generating the affinity matrices for hierarchical clustering, the LS and NDFS algorithms were determined to provide the greatest improvement in clustering results. The NDFS feature selector is able to sample the entire spectrum by assigning a high feature importance score to a few features.

REFERENCE

1. UNODC Early Warning Advisory on New Psychoactive Substances. What are NPS? https://www.unodc.org/LSS/Home/NPS. (accessed Mar 2021).
2. "Title 21 United States Code (USC) Controlled Substances Act" United States Drug Enforcement Administration: https://www.dea.gov/controlled-substances-act. (accessed Mar 2021).
3. Roggo, Y.; Chalus, P.; Maurer, L.; Lema-Martinez, C.; Edmond, A.; Jent, N., A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis* **2007,** *44* (3), 683-700.
4. Wadood, S. A.; Boli, G.; Xiaowen, Z.; Hussain, I.; Yimin, W., Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchemical Journal* **2020,** *152*, 104295.
5. Zhu, M.-Z.; Wen, B.; Wu, H.; Li, J.; Lin, H.; Li, Q.; Li, Y.; Huang, J.; Liu, Z., The Quality Control of Tea by Near-Infrared Reflectance (NIR) Spectroscopy and Chemometrics. *Journal of Spectroscopy* **2019,** *2019*, 8129648.
6. Bel'skaya, L. V., Use of IR Spectroscopy in Cancer Diagnosis. A Review. *Journal of Applied Spectroscopy* **2019,** *86* (2), 187-205.
7. Luinge, H. J., Automated interpretation of vibrational spectra. *Vibrational Spectroscopy* **1990,** *1* (1), 3-18.
8. Zloh, M.; Samaras, E. G.; Calvo-Castro, J.; Guirguis, A.; Stair, J. L.; Kirton, S. B., Drowning in diversity? A systematic way of clustering and selecting a representative set of new psychoactive substances. *RSC Adv* **2017,** *7* (84), 53181-53191.
9. Varmuza, K.; Karlovits, M.; Demuth, W., Spectral similarity versus structural similarity: infrared spectroscopy. *Anal Chim Acta* **2003,** *490* (1), 313-324.
10. Muegge, I.; Mukherjee, P., An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin Drug Discov* **2016,** *11* (2), 137-148.
11. Yun, Y.-H.; Liang, Y.-Z.; Xie, G.-X.; Li, H.-D.; Cao, D.-S.; Xu, Q.-S., A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems. *Analyst* **2013,** *138* (21), 6412-6421.
12. Gemperline, P. J.; Salt, A., Principal components regression for routine multicomponent UV determinations: A validation protocol. *Journal of Chemometrics* **1989,** *3* (2), 343-357.
13. Geladi, P.; Kowalski, B. R., Partial least-squares regression: a tutorial. *Analytica Chimica Acta* **1986,** *185*, 1-17.
14. Alelyani, S.; Tang, J.; Liu, H., Feature Selection for Clustering: A Review. In *Data Clustering: Algorithms and Applications*, 2013; pp 29-60.
15. Rousseeuw, P. J., Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* **1987,** *20*, 53-65.
16. Hubert, L.; Arabie, P., Comparing partitions. *J Classif* **1985,** *2* (1), 193-218.
17. Strehl, A.; Ghosh, J., Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **2002,** *3*, 583-617.
18. Simmler, L. D.; Rickli, A.; Schramm, Y.; Hoener, M. C.; Liechti, M. E., Pharmacological profiles of aminoindanes, piperazines, and pipradrol derivatives. *Biochem Pharmacol* **2014,** *88* (2), 237-44.
19. Rickli, A.; Kopf, S.; Hoener, M. C.; Liechti, M. E., Pharmacological profile of novel psychoactive benzofurans. *Br J Pharmacol* **2015,** *172* (13), 3412-3425.

20. Luethi, D.; Kaeser, P. J.; Brandt, S. D.; Krähenbühl, S.; Hoener, M. C.; Liechti, M. E., Pharmacological profile of methylphenidate-based designer drugs. *Neuropharmacology* **2018,** *134* (Pt A), 133-140.
21. Luethi, D.; Kolaczynska, K. E.; Docci, L.; Krähenbühl, S.; Hoener, M. C.; Liechti, M. E., Pharmacological profile of mephedrone analogs and related new psychoactive substances. *Neuropharmacology* **2018,** *134*, 4-12.
22. Simmler, L. D.; Buser, T. A.; Donzelli, M.; Schramm, Y.; Dieu, L. H.; Huwyler, J.; Chaboz, S.; Hoener, M. C.; Liechti, M. E., Pharmacological characterization of designer cathinones in vitro. *Br J Pharmacol* **2013,** *168* (2), 458-70.
23. Mackie, K., Cannabinoid receptors as therapeutic targets. *Annu Rev Pharmacol Toxicol* **2006,** *46*, 101-22.
24. Rickli, A.; Moning, O. D.; Hoener, M. C.; Liechti, M. E., Receptor interaction profiles of novel psychoactive tryptamines compared with classic hallucinogens. *Eur Neuropsychopharmacol* **2016,** *26* (8), 1327-37.
25. Rickli, A.; Luethi, D.; Reinisch, J.; Buchy, D.; Hoener, M. C.; Liechti, M. E., Receptor interaction profiles of novel N-2-methoxybenzyl (NBOMe) derivatives of 2,5-dimethoxy-substituted phenethylamines (2C drugs). *Neuropharmacology* **2015,** *99*, 546-53.
26. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E., PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research* **2021,** *49* (D1), D1388–D1395.
27. Bolton, E. E.; Chen, J.; Kim, S.; Han, L.; He, S.; Shi, W.; Simonyan, V.; Sun, Y.; Thiessen, P. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H., PubChem3D: a new resource for scientists. *Journal of cheminformatics* **2011,** *3* (1), 32-32.
28. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; 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., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Wallingford, CT, 2016.
29. Scott, A. P.; Radom, L., Harmonic Vibrational Frequencies: An Evaluation of Hartree−Fock, Møller−Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *The Journal of Physical Chemistry* **1996,** *100* (41), 16502-16513.
30. Johnson, R. D., NIST Computational Chemistry Comparison and Benchmark Database. Release 20, August 2019 ed.; NIST Standard Reference Database No. 101: National Institute of Standards and Technology, 2017.
31. Mott, A. J.; Rez, P., Calculated infrared spectra of nerve agents and simulants. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2012,** *91*, 256-260.
32. Chu, P. M.; Guenther, F. R.; Rhoderick, G. C.; Lafferty, W. J., The NIST Quantitative Infrared Database. *J. Res. Natl. Inst. Stand. Technol.* **1999,** *104*, 59-81.

33. Luethi, D.; Liechti, M. E., Designer drugs: mechanism of action and adverse effects. *Arch Toxicol* **2020,** *94* (4), 1085-1133.
34. Bajusz, D.; Rácz, A.; Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015,** *7* (1), 20.
35. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G., Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* **2002,** *42* (6), 1273-1280.
36. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *J Chem Inf Model* **2010,** *50* (5), 742-754.
37. Axen, S. D.; Huang, X.-P.; Cáceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J., A Simple Representation of Three-Dimensional Molecular Structure. *Journal of Medicinal Chemistry* **2017,** *60* (17), 7393-7409.
38. Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J., Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *J Comput Aided Mol Des* **2015,** *29* (10), 937-950.
39. *RDKit: Open-source cheminformatics; http://www.rdkit.org*.
40. Zhao, Z.; Liu, H., Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, Association for Computing Machinery: Corvalis, Oregon, USA, 2007; pp 1151–1157.
41. He, X.; Cai, D.; Niyogi, P., Laplacian score for feature selection. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, MIT Press: Vancouver, British Columbia, Canada, 2005; pp 507–514.
42. Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; Zhou, X., l_{2,1}-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two*, AAAI Press: Barcelona, Catalonia, Spain, 2011; pp 1589–1594.
43. Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; Lu, H., Unsupervised feature selection using nonnegative spectral analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI Press: Toronto, Ontario, Canada, 2012; pp 1026–1032.
44. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H., Feature Selection: A Data Perspective. *ACM Comput. Surv.* **2017,** *50* (6), Article 94.
45. Linstrom, P. J.; Mallard, W. G., *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. National Institute of Standards and Technology, Gaithersburg MD, 20899.
46. Fraley, C.; Raftery, A. E., How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal* **1998,** *41* (8), 578-588.
47. Maulik, U.; Bandyopadhyay, S., Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2002,** *24* (12), 1650-1654.
48. Brohée, S.; van Helden, J., Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **2006,** *7* (1), 488.
49. de Souto, M. C. P.; Costa, I. G.; de Araujo, D. S. A.; Ludermir, T. B.; Schliep, A., Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* **2008,** *9* (1), 497.
50. Pirim, H.; Ekşioğlu, B.; Perkins, A. D.; Yüceer, Ç., Clustering of high throughput gene expression data. *Computers & Operations Research* **2012,** *39* (12), 3046-3061.

51. Kinnunen, T.; Sidoroff, I.; Tuononen, M.; Fränti, P., Comparison of clustering methods: A case study of text-independent speaker modeling. *Pattern Recognition Letters* **2011,** *32* (13), 1604-1617.