Unsupervised Raw IR Spectra Images Recognition: Toward Chemist-Like Chemical Structure Classification.

Kentarou Fuku[a] and Takefumi Yoshida*[b]

[a] Dr. K. Fuku
Faculty of Advanced Engineering, Tokyo University of Science, Tokyo, 125-8585, Japan.

[b] Dr. T. Yoshida*
Cluster of Nanomaterials, Graduate School of Systems Engineering, Wakayama University. 930 Sakaedani, Wakayama, 640-8510, Japan.
E-mail: tyoshyd@wakayama-u.ac.jp

Abstract: The recent advancements in artificial intelligence have greatly improved spectral data analysis. Here, we explored using unsupervised machine learning to classify chemical compounds based on IR spectrum images without relying on prior chemical knowledge. The research demonstrated the potential of machine learning in chemical classification by extracting IR spectral images from the SDBS database and converting them into 218196-dimensional vector data. The hierarchical clustering of 227 compounds revealed distinct main clusters (**A**-**G**), each with specific subclusters showing higher intra-cluster similarity. Despite some challenges, such as sensitivity to spectral deviations and difficulty distinguishing delicate chemical structures in spectra with low transparency in the fingerprint area, the approach showed promise. The Tanimoto coefficient was used as a metric for molecular similarity, providing valuable insights, though sometimes diverging from chemist intuition. The study also highlighted that scaling composition formulas and molecular weights did not affect the classification results, as the high-dimensional features dominated the process. Overall, the research demonstrated the feasibility of using IR spectral image data in machine learning for chemical classification, offering a novel perspective that complements traditional methods, even though some classifications may not always align with chemist intuition.

Introduction

Materials informatics (MI) has been conducted to develop new functional materials. Particularly in the field of drug discovery, which has traditionally required enormous costs and labor, innovative advances are expected.[1] Additionally, attempts have been made to extract trends in the development of physical properties by using machine learning on spectral information such as XAFS (X-ray absorption fine structure) and XPS.[2] Timoshenko *et al.* have reported that the local structure of the heterogeneous materials is optional, whereas the EXAFS database reveals M−M or M−L distances or ratios of the structure using the neural network-based method.[2b,2c] Torrisi et al. have simulated the X-ray absorption spectroscopy (XAS) spectra using FEFF software and applied the random forest machine learning model to XAS spectra, where the spectra expected Bander charge.[2d] Thus, spectral analysis by artificial intelligence is compelling in visualizing features that scientists previously described by relying on intuition.

However, in fields where developed databases (such as CCDC[3] and SDBS[4]) have not been compiled so far, collecting data is a significant barrier to entry into the fields. Therefore, a currently popular method is to prepare spectral data through quantum chemical calculation simulations and use it for machine learning.[5] However, the problem with these methods is that they output data from simulation methods that are not fully established and that scientists use already known knowledge to perform machine learning. Furthermore, in methods that use neural networks and other methods, the process is a black box, making it difficult for scientists to derive causal relationships. In such cases, it is impossible to obtain completely unknown knowledge, and from the perspective of materials science, there is a fear that this will deviate from the actual material system.

Lu *et al.* reviewed the limitations of machine learning with small datasets and the development of data extraction and high-throughput computations.[6] Research is also being conducted on methods such as Bayesian optimization to derive correct solutions from small data sets. However, in any case, with a small data set, there is a possibility that you will not be able to get out of the box of existing solutions, and efforts are needed to increase the amount of data.

Swain *et al.* developed the ChemDataExtractor to extract chemical data from published papers.[7] Also, WebPlotDigitizer helps to extract plot data from figures.[8] The Starrydata project was achieved by extracting experimental data on several ten thousand thermoelectric materials.[9] However, these methods still require human resources and can be overwhelmed by the volume of papers produced worldwide, requiring a more mechanical process.

Machine learning using image recognition is applied in a wide range of fields and is expected to develop widely in the chemical field as well. Inokuma *et al.* reported mixing-ratio prediction for various solid mixtures using image-based ML, where the trained model with 300 images could solve the weight ratio of each sugar and dietary salt in their mixture.[10] Yanagida *et al.* reported image processing and machine learning to identify the

multivariate chemo-/biomarker features in chromatography–mass spectrometry.[11] These supervised learning experiments demonstrate that artificial intelligence can be used in chemistry using images. However, how artificial intelligence understands chemical space unsupervised is unclear.

IR spectroscopy, a cornerstone method for material characterization, visually represents chemical bond stretching modes. Artificial intelligence (AI) capability to discern organic compounds from spectra, aided by deep learning of extensive IR spectra databases, aligns with human chemists' abilities.[12] Unsupervised clustering with smaller datasets has also shown promise in organic compound classification, with applications extending to hyperspectral identification.[12c] This exemplifies the potential of image recognition analysis to complement traditional methods.

In this work, we reported that image data, particularly IR spectral data, can be directly utilized in machine learning to classify chemical structures. To validate unsupervised machine learning, we used IR spectral image data from the AIST spectrum database (SDBS),[4] where well-formatted spectral image data can be obtained. We extracted the spectral portion of the image, converted it into data with pixel-amounts dimensions vector, and performed hierarchical clustering.

This study also explores the feasibility of employing image analysis techniques in materials science for machine learning, potentially replicating classifications made by chemists. Utilizing existing image recognition methods could enhance efficiency in data collection for exploring unknown materials. Furthermore, using images eliminates the need for processing such as data interpolation. Our proposal involves utilizing image texture analysis for spectral data analysis, a novel approach that prompts philosophical debates on the extent of supervision.

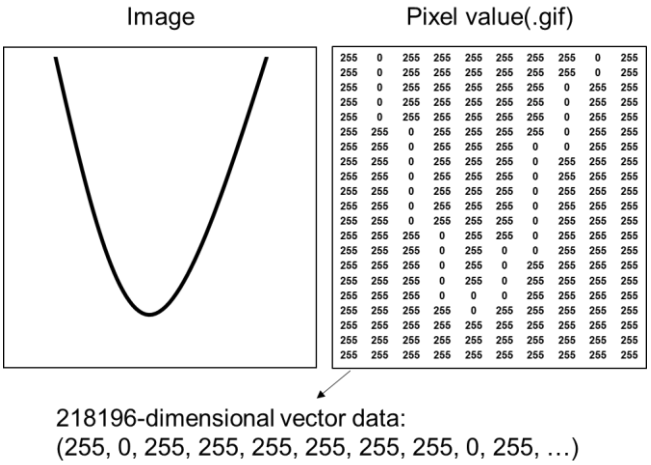


Figure 1. Schematic image of processing; the gray-scaled pixels of the image are stacked horizontally (the results are the same even in the vertical direction), resulting in 218196-dimensional vector data.

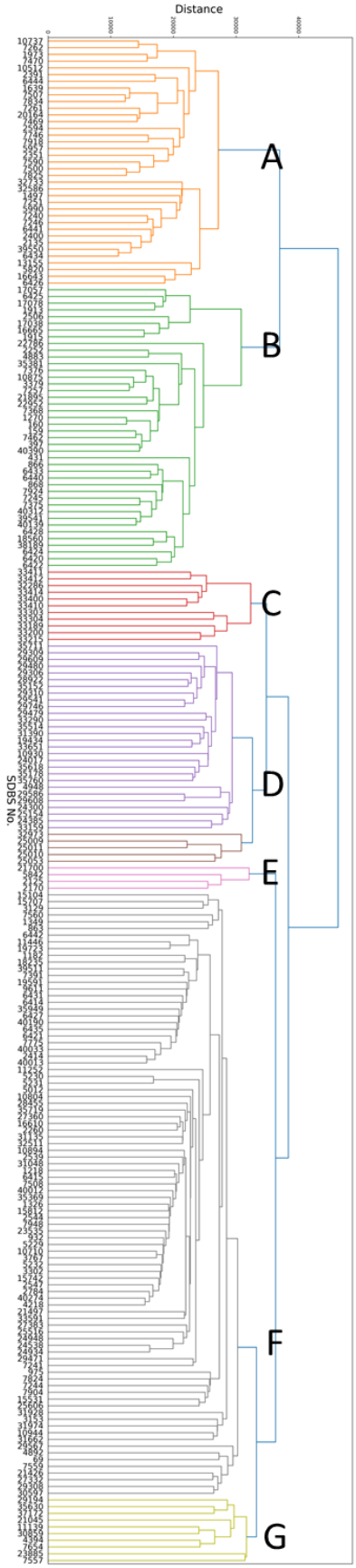


Figure 2. The result of hierarchical clustering result of 227 compounds was performed without separating organic and inorganic compounds.

Results and Discussion

Note. The AI does not know chemistry and only learns images. In addition, we evaluated the results and inferred what kind of decisions the AI made, but the actual decisions made by the AI are kept in a black box, as in other machine learning research. Also, technical issues with the measurement are not considered (this condition does not change even if it is plot data).

Data set and Processing. The target compounds were those with various functional groups ranging from low to medium organic molecules; to confirm that it works, we used 227 compounds (Some pairs have different SDBS numbers but the same compounds.), include inorganic compounds. Many of these molecules' structures are complicated, and it is assumed that it is relatively difficult to distinguish them by IR, even if chemists. A list is shown in Table S1. Although the data measured by the KBr method and the Nujol method were mixed, no difference was observed between the measurement methods except when measuring inorganic compounds.

All data analysis was performed on the Anaconda platform.[13] In particular, we used OpenCV for handling images.[14] The details of the script and data on GitHub.[15] We used the images from the SDBS database and trimmed regions where IR spectra were described in .gif. The gray-scaled pixels of the image are stacked horizontally (the results are the same even in the vertical direction), resulting in 218196-dimensional vector data. Machine learning was performed using the accompanying data: This method dramatically reduces calculation costs (especially RAM) compared to training while maintaining pixel coordinates and values. On the other hand, this method has the disadvantage of being intolerant of spectral deviations. SDBS number, IUPAC name, InChI, molecular formula, and molecular weight. For hierarchical clustering, we used the linkage function (method='ward', metric='euclidean').[16] Since this work used unsupervised machine learning, the AI classifies IR spectra from a different perspective than a chemist. Depending on the group, even if the ends are similar, some are not similar as a whole. In addition, some classification criteria are unclear, and improvements are needed. However, we showed that unsupervised machine learning, which uses images as they are, is generally applicable.

This work used the Tanimoto coefficient as an index to objectively express molecules' similarity.[17] However, this value often results in a low degree of similarity, even when a chemist judges the degree of similarity to be high. For example, dodecane, dodecan-1-ol, and dodecan-1-amine hydrochloride are highly similar because they are single-chain alkyl chains with functional groups attached to the ends. Still, the average Tanimoto coefficient between them is 0.44, less than 0.5.

Using RDKit,[18] we generated a structure from InChI and calculated the Tanimoto coefficient (hydrogen was excluded because its orientation was not determined). The radius of the Tanimoto coefficient was set to 2. This value was chosen because IR is particularly sensitive to functional groups. This gives partially high Tanimoto coefficients (*TC*) even if the long chain alkyl lengths differ.

Hierarchical Clustering. Hierarchical clustering of 227 compounds was performed without separating organic and inorganic compounds. They could be roughly divided into 7 main clusters. From the left end of the diagram, the 3rd generation cluster is named **A**, **B**, **C**, **D**, **E**, **F**, and **G**, and the 4th generation cluster and higher branches are named **A-1**, **A-2**, and **A-3**.

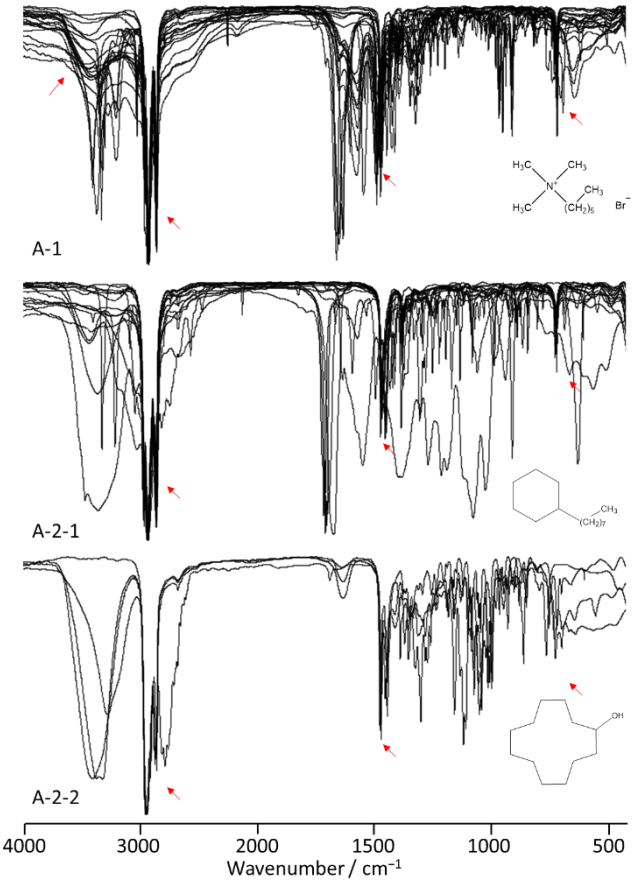


Figure 3. The image of the superposition for the **A** cluster.

The **A** cluster has 42 compounds and relatively long saturated hydrocarbons with no aromatic functional groups (except compound **10512**) and is finely branched with other functional groups. The image of the superposition is shown in Figure 3. The main common feature is sharp peaks derived from CH around 3000, 1500, and 700 cm^{-1}. Another factor is that the pattern of the fingerprint area is relatively simple. The average of whole pairs of the *TC* is 0.2865.

Regarding the **A-1** cluster (21 compounds, *TC* = 0.4275), they have a common feature of a relatively gentle peak derived from amines and the like around 3500 cm^{-1}. In addition, the valley between the peaks derived from CH (CH_3) around 2920 cm^{-1} and CH (CH_2) around 2850 cm^{-1} is characterized by being deeper than the **A-2** cluster, and

compounds **7500** and **7825** are exceptionally classified as the **A-1** cluster. Within the 5th-6th generation cluster, the structures were very similar, and the averaged *TC* were 1.0, 1.0, 0.3799 for **A-1-1** (long-alkyl-trimethyl-ammonium halide), **A-1-2-2** (long-alkyl-amine), and **A-1-2-3** (long-alkyl-amide, long-alkyl-nitrile etc.). Although the substructures are similar, compounds **10512** and **2351** are single branches from relatively far distances.

The **A-2** cluster (21 compounds, *TC* = 0.2395878) has no major differences from the **A-1** cluster other than the characteristic that the valley between the peaks derived from CH (CH_2) near 2850 cm^{-1} is shallower than that of the **A-1** cluster. Simple aliphatic hydrocarbons or aliphatic hydrocarbons with OH or double-bonded groups are classified. Since the **A-2-2** cluster has high similarity in the region 700-1500 cm^{-1} and is more complex than the **A-2-1** cluster, it has diverged from the **A-2-1** cluster by a long distance.

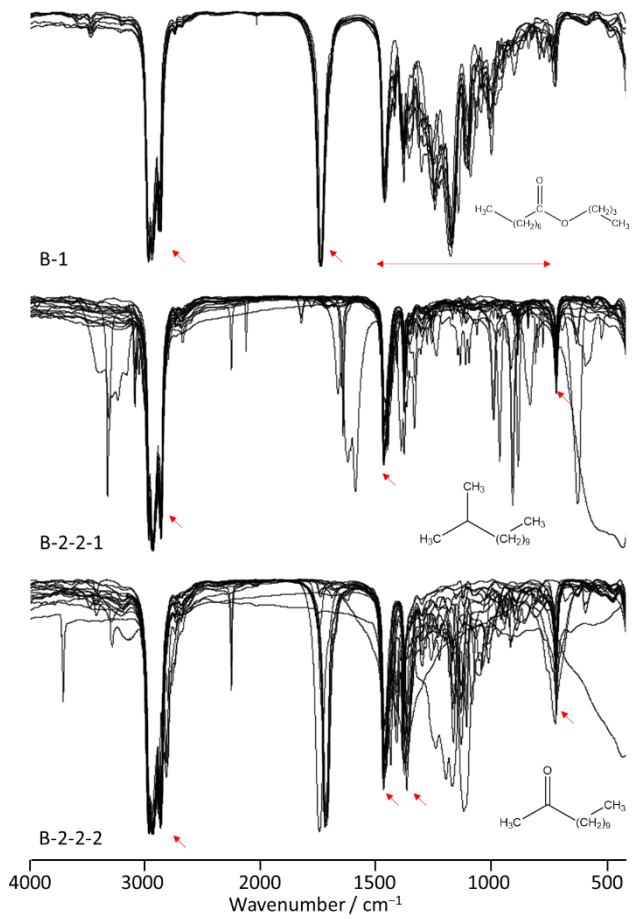


Figure 4. The image of the superposition for the **B** cluster.

The **B** cluster (38 compounds) has characteristics similar to the **A** cluster. It basically has no aromatic functional groups, has relatively long saturated hydrocarbons, and is finely branched with other functional groups. The image of the superposition is shown in Figure 4. The main common feature is sharp peaks derived from CH around 3000, 1500, and 700 cm^{-1}. Another factor is that the pattern of the fingerprint area is relatively simple. The difference with the **A** cluster is that the valley between the peaks derived from CH (CH_3) around 2920 cm^{-1} and CH (CH_2) around 2850 cm^{-1} is shallower than in the **A** (**A-2**) cluster. Furthermore, the peak derived from the bending vibration mode of C-CH_3 near 1350 cm^{-1} is sharper than that of the **A-2** cluster.

The **B-1** cluster (8 compounds) is assigned to a saturated fatty acid ester, with a characteristic ester peak around 1750 and 1150 cm^{-1} and a similar fingerprint region structure. The average *TC* is high at 0.7233 for the **B-1** cluster.

The **B-2** cluster (30 compounds) has a saturated aliphatic hydrocarbon as its main chain and alkenes, alkyne groups, CN, and ketones groups as brunch. In addition, inorganic compounds such as BN, MgO, and LiF are included in this cluster because signals of aliphatic hydrocarbons used in the Nujol method have been observed. A copper complex has also been assigned, but the CH stretching characteristic around 3000 cm^{-1} derived from the ligand matches that of the **B** cluster. In addition, the amino and carboxyl groups are different from those of other compounds, so they are on separate branches. The average *TC* is low at 0.2326 due to the inclusion of inorganic compounds and metal complexes.

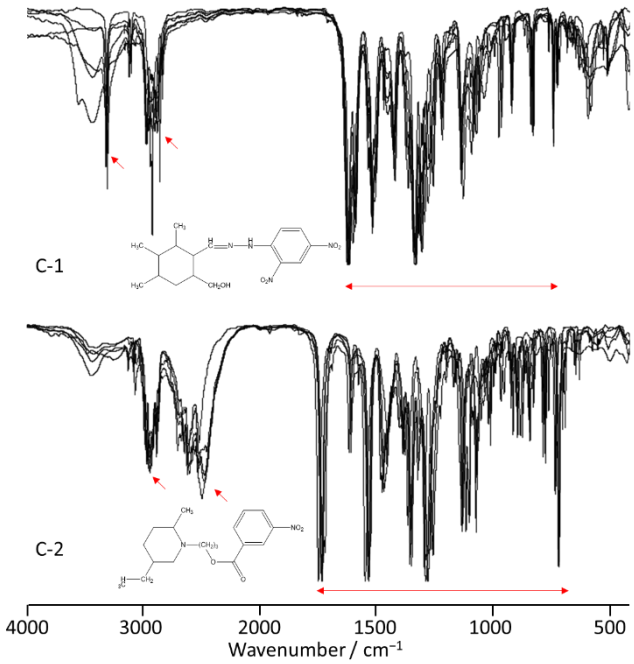


Figure 5. The image of the superposition for the **C** cluster.

The **C** cluster (11 compounds) has aromatic, amine, nitro (NO_2) and aliphatic group. The image of the superposition is shown in Figure 5. The transmittance of the peaks derived from CH around 3000 cm^{-1} in the **C** cluster is weaker than that of the **A** and **B** clusters. They also feature fingerprint regions with characteristic shapes on individual branches. The **C-1** (6 compounds, *TC* = 0.7182) has a characteristic structure with peaks arranged at equal intervals around 1650-700 cm^{-1} derived from an

aromatic skeleton, hydrazine N─N bond, and nitro group. Furthermore, a sharp peak due to NH stretching is observed around 3300 cm^{-1}. **C-2** cluster (5 compounds, *TC* = 0.6918) also has a characteristic structure with peaks arranged at equal intervals around 1700-700 cm^{-1} derived from an aromatic skeleton and an ester (R─COO─R') or nitro (NO_2) group. The **C-2** cluster has a strong ester peak at 1700 cm^{-1} and instead has a weaker C=C or C=N double bond peak at around 1620 cm^{-1} compared to the **C-1** cluster. Furthermore, the peak derived from HCl near 3000 cm^{-1} is similar because the chemical structures are similar. A broad peak near 3500 cm^{-1} is likely to be derived from water.

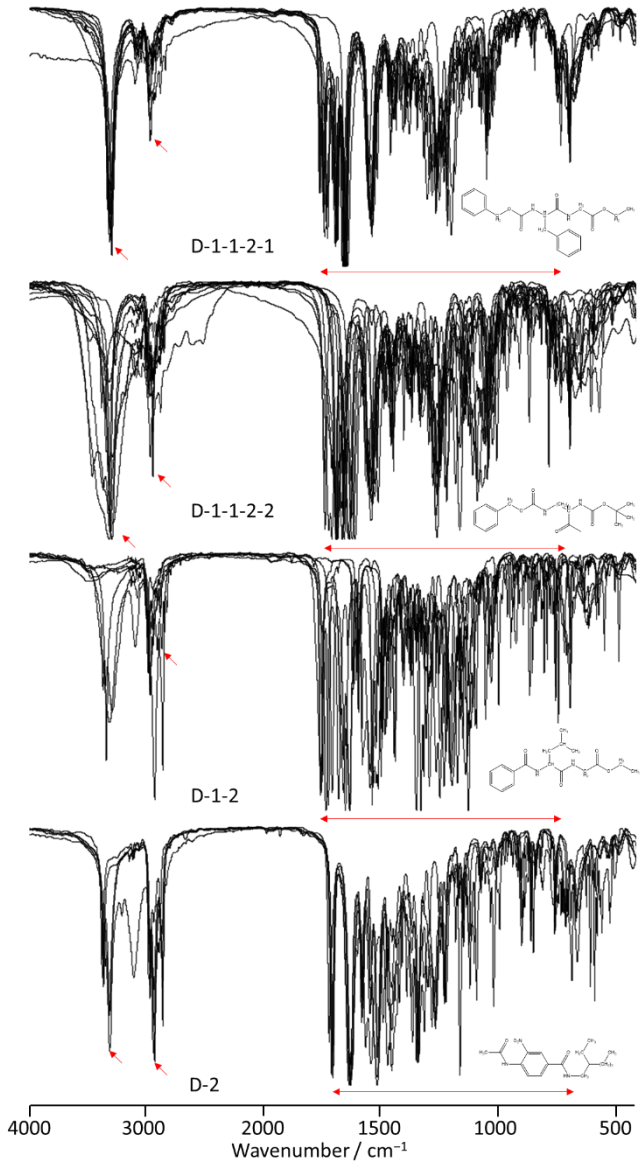


Figure 6. The image of the superposition for the **D** cluster.

The **D** cluster (33 compounds) is characterized by a strong NH peak of the amide bond around 3300 cm^{-1}, narrower CH peaks around 3000 cm^{-1}, and the structure of the C=O double bond and aromatic (exception: compounds **33290**, **33651**) fingerprint region around 1500-1000 cm^{-1} is similar. The CH stretching peak near 3000 cm^{-1} is modest. The image of the superposition is shown in Figure 6. The **D** cluster has functional groups similar to the **C** cluster, but due to its more complex structure, it has many fine peaks in the fingerprint region. **D-1-1-2-1** cluster (9 compounds, *TC* = 0.5555) contains amides (R─NH─CO─R') and esters and is characterized by a three-pronged peak in the 1750-1650 cm^{-1} range and a peak at 1520. The strength of the fingerprint region in the 1500-650 cm^{-1} range is also similar. The **D-1-1-2-2** cluster (11 compounds, *TC* = 0.2574) does not peak around 1750 cm^{-1} compared to the **D-1-1-2-1** cluster. Therefore, there are only two or three peaks between 1650 and 1500. The strength pattern in the fingerprint region is also different. **D-1-2** cluster (7 compounds, *TC* = 0.2599) has a higher baseline only in the range of 1500-1000 compared to **D-1-1-2**, which has a higher baseline in the 700-1000 cm^{-1} range. The **D-2** cluster (5 compounds, *TC* = 0.4775) has a higher peak of around 3000 cm^{-1} compared to the **D-1** cluster, and the baseline in the 1700-1000 cm^{-1} range is higher.

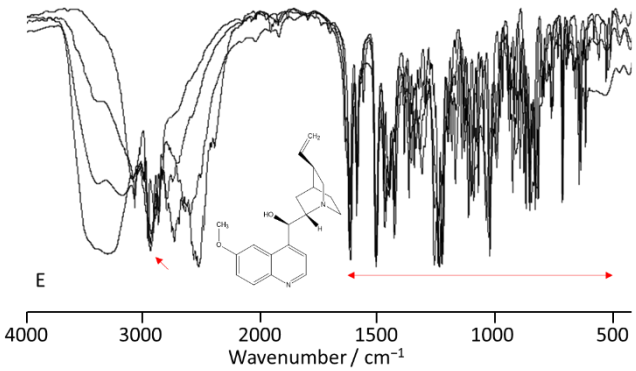


Figure 7. The image of the superposition for the **E** cluster.

The **E** (4 compounds, *TC* = 0.9731) is classified as quinine and quinidine. The image of the superposition is shown in Figure 7. It is characterized by the characteristic broad and extensive peak of the OH group, the fingerprint region around 1500 and 1620 cm^{-1} derived from the amine structure of Quinoline and Quinuclidine, and the alkene peak also around 1620 cm^{-1} (compound **21700** has no alkene).

The **F** cluster, excluding inorganic substances that are often assigned to it, often has a structure with aliphatic hydrocarbons similar to **A** and **B** clusters. It is characterized by a region with a relatively high baseline compared to **A** and **B** clusters.

The **F-1-1-1-1** cluster is a quaternary alkyl ammonium salt and has characteristics similar to those of the **A-1** cluster. However, while the **A-1** cluster is trimethyl, the **F-1-1-1-2** cluster has three or four long alkyl chains, and the peak derived from CH_2 is particularly strong in the fingerprint region. The **F-1-1-1-2** cluster is similar because Acenaphthene and Acenaphthylene have very similar chemical structures, so the aromatic peaks and baselines

overlap. On the other hand, the reason why 4,4'-dibutoxyazoxybenzene was included seems to be that there were many overlapping areas, such as the sides of the peaks. **F-1-1-2-1-2-1** cluster is a long alkyl chain containing an ester or tertiary amine, has characteristic CH peaks at 3000, 1500, and 700 cm^{-1}, and is very similar to **the B** cluster. It has the following characteristics. On the other hand, it is considered that the background of the fingerprint area was smaller than the **B-1** cluster and larger than the **B-2** cluster and was determined as a separate group. The **F-1-1-2-1-2-2-1** cluster has a high spectral baseline match and low similarity in chemical structure. The **F-1-1-2-1-2-2-2-1** cluster has a low similarity to all other spectra; some just happen to overlap. The **F-1-1-2-1-2-2-2-2** cluster is a saturated hydrocarbon, highly similar to **A** and **B** clusters, and also contains inorganic compounds measured by the Nujol method. The difference between **A** and **B** clusters is that the top of the 3000 cm^{-1} peak is flat. The **F-1-1-2-1-2-2-1** cluster is assigned to the hydrochloride and hydrobromide salts of hydrazine. These are reasonable results because the spectra themselves are highly similar. The **F-1-1-2-1-2-2-2-2-2-1** cluster has an ester structure and a chemical structure very similar to the **D** cluster, but except for the CH peak around 3000 cm^{-1}, it is clustered. There are no particularly common peaks within the range. The **F-1-1-2-1-2-2-2-2-2-2-1-1** cluster had no image resemblance that humans could discern. Items with similar characteristics, such as a broad peak near 3500 cm^{-1} due to water and quaternary alkyl ammonium, are classified but not necessarily on adjacent branches. The **F-1-1-2-1-2-2-2-2-2-2-1-2** cluster has an expected broad carboxylic acid peak around 1700 cm^{-1}, but other characteristics are similar. The **F-1-1-2-1-2-2-2-1-2-2-2** cluster has peaks derived from CH, NH, and C=O and has a chemical structure similar to that of the **D** cluster. However, the most significant difference is that it does not have an aromatic group, so the peaks in the fingerprint region are much weaker. The **F-1-1-2-1-2-2-2-2** cluster has a peak of around 3300 cm^{-1} for NH derived from the amide structure, a peak of around 1700 cm^{-1} for CO, and a peak around 3000 cm^{-1} derived from CH is also characteristic. The difference with the **D** cluster is that the peak in the fingerprint area is simple. The **F-1-1-2-2** cluster has no particular similarity except for the CH-derived peak near 3000 cm^{-1}, with only the sides of the peaks in the fingerprint region overlapping each other. **F-1-2** cluster has peaks derived from CH, NH, and C=O, and has a chemical structure similar to that of **D** and **E** clusters. However, the most significant difference is that it has large background in the fingerprint region. Also, quinine di-1,1'-oxide is properly assigned to a different cluster than quinine. The **F-2** cluster has no particular similarity with only the sides of the peaks in the fingerprint region overlapping each other.

The **G** cluster tends to have a high background, especially in the fingerprint region. It has aromatic and amide or ester groups and is characterized by a complex structure. The image of the superposition is shown in Figure 8. Pairs of individual branches have similar chemical structures, but the commonality between pairs of branches that are far apart is small. In particular, The **G-1-2-2** cluster is classified as brucine, which has an amide structure. The fingerprint region has a high similarity, NH stretching around 3500 cm^{-1} and CH stretching around 3000 cm^{-1}.

The IUPAC name, molecular weight, and molecular formula were normalized and combined as feature quantities. However, these did not affect hierarchical clustering results. This is thought to be because these features were monotonic and did not have modulation of the 218196-dimensional vector features in the IR spectrum image, and chemists only use methods such as decision tree judgments based on their intuition. It could not be reproduced.

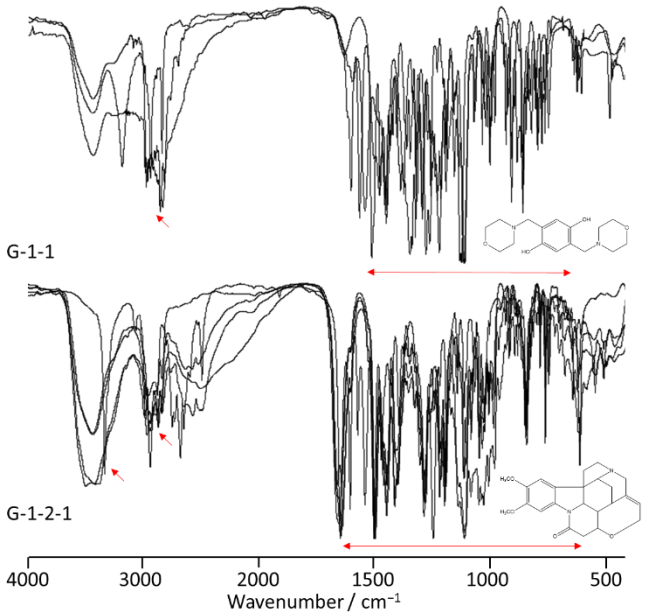


Figure 8. The image of the superposition for the **G** cluster.

Conclusion

This study aimed to use unsupervised machine learning without prior knowledge of chemistry and to separate compounds with similar chemical structures using IR spectrum raw images. The study demonstrates the feasibility of using image data of IR spectra in machine learning for chemical classification. The dataset preparation involved extracting spectral images from the SDBS database and converting them into high-dimensional vector data. Despite some limitations, such as sensitivity to spectral deviations and the need for methodological refinements, unsupervised machine learning shows promise in classifying IR spectra from a novel perspective. In spectra with low transparency in the fingerprint area and spectrum images with large broadened peaks, only the features were emphasized, making it difficult to separate the delicate chemical structures. In addition, chemists make identifications from other incidental information, but scaling the composition

formula and molecular weight and adding them to the elements did not change the results. This is because even if scaling is done, the 218196-dimensional vector features determine the features, even if a one-dimensional feature change is added to the displacement of the 218196-dimensional vector features. Additionally, the Tanimoto coefficient provides a useful, though sometimes divergent, metric for assessing molecular similarity. The hierarchical clustering of IR spectral data demonstrated that unsupervised machine learning could classify chemical structures based on spectral features. The main clusters (A-G) show distinct patterns and characteristics, with specific subclusters revealing higher intra-cluster similarity. The approach highlights the potential of using machine learning for chemical classification, even though some classifications may not align with traditional chemist intuition. This study also explores the feasibility of employing image analysis techniques in materials science for machine learning, potentially replicating classifications made by chemists. Utilizing existing image recognition methods could enhance efficiency in data collection for exploring unknown materials. Furthermore, using images eliminates the need for processing such as data interpolation. Our proposal involves utilizing image texture analysis for spectral data analysis, a novel approach that prompts philosophical debates on the extent of supervision.

Acknowledgements
This work was partially supported by JSPS KAKENHI Grant No. JP20K15293 (TY) and JP23K13761 (TY). IR spectrum images obtain from AIST: Spectral Database for Organic Compounds, SDBS.

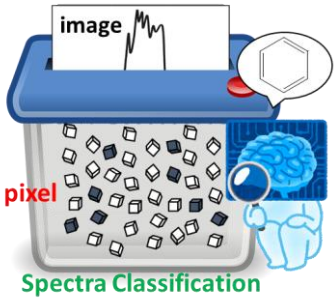
Conflict of interest
The authors declare no conflict of interest.

Keywords: IR spectra • machine learning • artificial intelligence • hierarchical clustering • image recognition

[1] a) Aspuru-Guzik et al., *ACS Cent. Sci.* **2018**, *4*, 268; b) B. Cuevas-Zuviría, L. F. Pacios, *J. Chem. Inf. Model.,* **2021**, *61*, 2658−2666; c) T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, K. I. Shimizu, *ACS Catal.,* **2020**, *10*, 2260–2297; d) E. Gawehn, J. A. Hiss, G. Schneider, *Mol. Inform.* **2016**, *35*, 3–14.
[2] a) C. D. Rankine, M. M. M. Madkhali, T. J. Penfold, *J. Phys. Chem. A*, **2020**, *124*, 4263−4270; b) J. Timoshenko, A. I. Frenkel, *ACS Catal.*, **2019**, *9*, 10192−10211; c) J. Timoshenko, F. T. Haase, S. Saddeler, M. Rüscher, H. S. Jeon, A. Herzog, U. Hejral, A. Bergmann, S. Schulz, B. R. Cuenya, *J. Am. Chem. Soc.,* **2023**, *145*, 4065–4080; d) S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram, L. Hung, *npj Comput. Mater.*, **2020**, *6*, 109; e) Z. Chen, N. Andrejevic, N. C. Drucker, T. Nguyen, R. P. Xian, T. Smidt, Y. Wang, R. Ernstorfer, D. A. Tennant, M. Chan, M. Li, *Chem. Phys. Rev.*, **2021**, *2*, 031301.
[3] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.,* **2016**, *72*, 171–179.
[4] SDBSWeb : https://sdbs.db.aist.go.jp (National Institute of Advanced Industrial Science and Technology, 2023/6).
[5] M. R. Carbone, M. Topsakal, D. Lu, S. Yoo, *Phys. Rev. Lett.* **2020**, *124*, 156401.
[6] P. Xu, X. Ji, M. Li, W. Lu, *npj Comput. Mater.*, **2023**, *9*, 42.
[7] M. C. Swain, J. M. Cole, *J Chem Inf Model* **2016**, *56*, 1894–1904.
[8] Rohatgi, A. (2017) Web Plot Digitizer. https://automeris.io/WebPlotDigitizer
[9] Y. Katsura, M. Kumagai, T. Kodani, M. Kaneshige, Y. Ando, S. Gunji, Y. Imai, H. Ouchi, K. Tobita, K. Kimura, K. Tsuda, *Sci Technol Adv Mater* **2019**, *20*, 511–520.
[10] Y. Ide, H. Shirakura, T. Sano, M. Murugavel, Y. Inaba, S. Hu, I. Takigawa, Y. Inokuma, *Ind Eng Chem Res* **2023**, *62*, 13790–13798.
[11] C. Jirayupat, K. Nagashima, T. Hosomi, T. Takahashi, W. Tanaka, B. Samransuksamer, G. Zhang, J. Liu, M. Kanai, T. Yanagida, *Anal Chem* **2021**, *93*, 14708–14715.
[12] a) W. Zhang, L. C. Kasun, Q. JieWang, Y. Zheng. Z. Lin, *Sensors*, **2022**, *22*, 9764; b) H. Ren, H. Li, Q. Zhang, L. Liang, W. Guo, F. Huang, Y. Luo, J. Jiang, *Fundamental Research*, **2021**, *1*, 488–494; c) Q. Zhong, C. Yang, F. Großerüschkamp, A. Kallenbach-Thieltges, P. Serocka, K. Gerwert, A. Mosig, *BMC Bioinformatics,* **2013**, *14*, 333; d) W. Fu, W. S. Hopkins, *J. Phys. Chem. A*, **2018**, *122*, 167–171; e) K. He, *ACS Omega,* **2021**, *6*, 32151−32165; f) M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci.*, 2017, 8, 6924; g) J. A. Fine, A. A. Rajasekar, K. P. Jethava, G. Chopra, Chem. Sci., 2020, 11, 4618; h) J. L. Lansford, D. G. Vlachos, *Nat. Commun.*, 2020, *11*, 1513.
[13] "Anaconda Software Distribution," can be found under https://anaconda.com, **2023/6/1.**
[14] G. Bradski, *Dr. Dobb's Journal of Software Tools* **2000**.
[15] https://github.com/yoshyd/IR_open
[16] a) H.-J. Mucha, H. Sofyan, in *XploRe*® - *Application Guide*, Springer Berlin Heidelberg, Berlin, Heidelberg, **2000**, pp. 239–279; b) J. H. Ward, *J Am Stat Assoc* **1963**, *58*, 236–244; c) F. Pedregosa FABIANPEDREGOSA, V. Michel, O. Grisel OLIVIERGRISEL, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot andÉdouardand, andÉdouard Duchesnay, Fré. Duchesnay EDOUARDDUCHESNAY, *Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot*, **2011**.
[17] T. T. Tanimoto, "An Elementary Mathematical theory of Classification and Prediction". *Internal IBM Technical Report.* 1957.

[18] “Open-source cheminformatics” can be found under https://www.rdkit.org, **2023/6/1.**

Entry for the Table of Contents



Research explored using unsupervised machine learning on IR spectrum images for chemical compound classification. Challenges included sensitivity to spectral deviations and difficulty in discerning structures in opaque spectra. The study underscores spectral image potential in machine learning for chemical classification, complementing traditional methods while acknowledging discrepancies with chemist intuition.