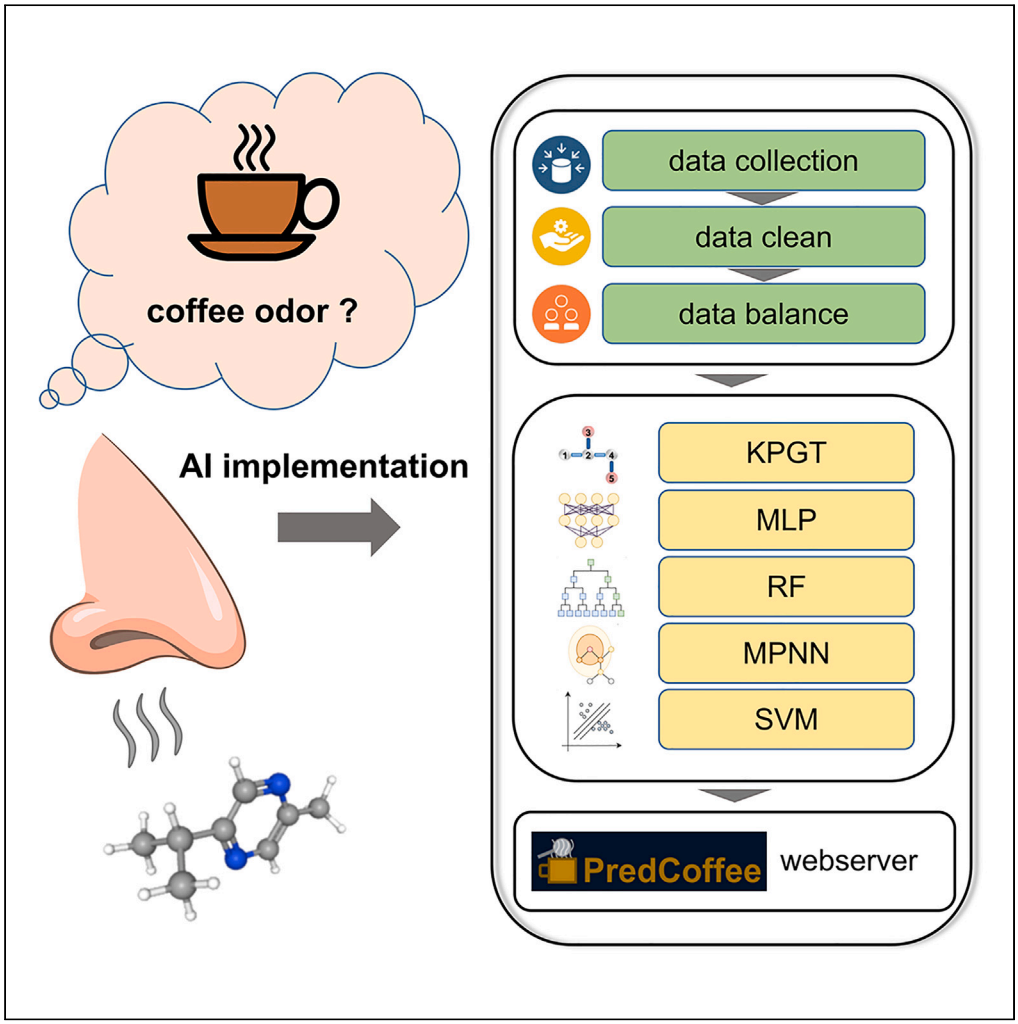
iScience

CellPress
OPEN ACCESS

Article

PredCoffee: A binary classification approach specifically for coffee odor



Yi He, Ruirui Huang, Ruoyu Zhang, Fei He, Lu Han, Weiwei Han

hefe@missouri.edu (F.H.)
luhan@jlu.edu.cn (L.H.)
weiweihan@jlu.edu.cn (W.H.)

Highlights

A dataset of coffee odor molecules was established

A molecular coffee odor predictor was developed using deep learning

The PredCoffee website was built

A systematic computational analysis of coffee odor molecules

He et al., iScience 27, 110041
June 21, 2024
https://doi.org/10.1016/j.isci.2024.110041

iScience

CellPress
OPEN ACCESS

Article

PredCoffee: A binary classification approach specifically for coffee odor

Yi He,[1,3] Ruirui Huang,[1,3] Ruoyu Zhang,[1] Fei He,[2,*] Lu Han,[1,*] and Weiwei Han[1,4,*]

SUMMARY

Compared to traditional methods, using machine learning to assess or predict the odor of molecules can save costs in various aspects. Our research aims to collect molecules with coffee odor and summarize the regularity of these molecules, ultimately creating a binary classifier that can determine whether a molecule has a coffee odor. In this study, a total of 371 coffee-odor molecules and 9,700 non-coffee-odor molecules were collected. The Knowledge-guided Pre-training of Graph Transformer (KPGT), support vector machine (SVM), random forest (RF), multi-layer perceptron (MLP), and message-passing neural networks (MPNN) were used to train the data. The model with the best performance was selected as the basis of the predictor. The prediction accuracy value of the KPGT model exceeded 0.84 and the predictor has been deployed as a webserver PredCoffee.

INTRODUCTION

Smell is the perception of volatile compounds by people. It occurs when volatile compounds combine with odorant molecules on the nasal epithelial cells, which in turn activate olfactory sensory neurons distributed in the olfactory epithelium through receptors expressed on those neurons. Olfactory neurons transmit information to the olfactory cortex located in the cerebral cortex. This pathway allows for the processing and interpretation of olfactory signals, leading to the formation of olfactory perception in animals.[1–3] The sense of smell allows individuals to gather substantial information about the external environment.[4] Typically, people label odors based on their subjective assessment. However, for certain odors, it may be difficult to provide a complete judgment or identify the type of odor. There are many kinds of odors that we come into contact with daily, such as fruit, tea, alcohol, coffee, wood, and so on. However, under the influence of some external factors, people cannot accurately identify certain odors according to their subjective perception. Moreover, for the same odor, different results will be identified due to different ages, gender, sensitivity, and some other personal factors.[5,6] The precise identification of odors is a crucial requirement in various industries including perfumery, flavoring, and the food industry.

As the pace of life gets faster, coffee is a very popular drink.[7] The popularity of coffee is mainly attributed to some of its properties and its unique smell. Studies have proved that moderate daily coffee consumption can reduce the occurrence of chronic diseases such as cardiovascular disease and diabetes.[8–10] The innovation and improvement of the odor of coffee drinks is a major selling point for merchants to sell products. Some individuals are unable to consume coffee due to their intolerance to caffeine or other components found in coffee. As an alternative, they opt for coffee-scented beverages that are caffeine-free. In addition to coffee, other natural product components have been identified to possess a coffee aroma, such as the extraction from chicory roots.[11–13] Natural products with coffee odor are important sources for making coffee odor beverages.

With the rapid advancement of the Internet and computer technology, utilizing data analysis and statistical methods has become highly practical and convenient for us to identify and summarize connections and variations within sample data especially in the fields of biology and medicine.[14] In the field of biological and chemical research, machine learning and data analysis are frequently employed for statistical analysis of substance structures and properties. These techniques aid in extracting meaningful insights and patterns from complex datasets, facilitating the understanding and prediction of various substance characteristics.[15] At present, many studies have applied computational model methods to conduct structural analysis and feature induction for a variety of odors or one odor and then formed a classifier.[16] These classifiers can make objective judgments on the odor of molecules better, or make odor predictions for molecules with unknown odors so that people do not have to go through the wet experiment process to identify the odor of a molecule, which greatly saves manpower and material costs and reduces time. The recent study has used graph neural networks (GNN) to generate a principal odor maps (POM) that retain perceptual relationships and are capable of odor quality prediction for previously uncharacterized odors.[17]

In this study, we collected 371 coffee-odor molecules and 9,700 non-coffee-odor molecules and performed data cleaning and balancing on these data. The information about these molecules can be obtained on the webserver (https://hwwlab/webserver/predcoffee). The

[1]Key Laboratory for Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, 2699 Qianjin Street, Changchun 130012, China
[2]Department of Electrical Engineer and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[3]These authors contributed equally
[4]Lead contact
*Correspondence: hefe@missouri.edu (F.H.), luhan@jlu.edu.cn (L.H.), weiweihan@jlu.edu.cn (W.H.)
https://doi.org/10.1016/j.isci.2024.110041

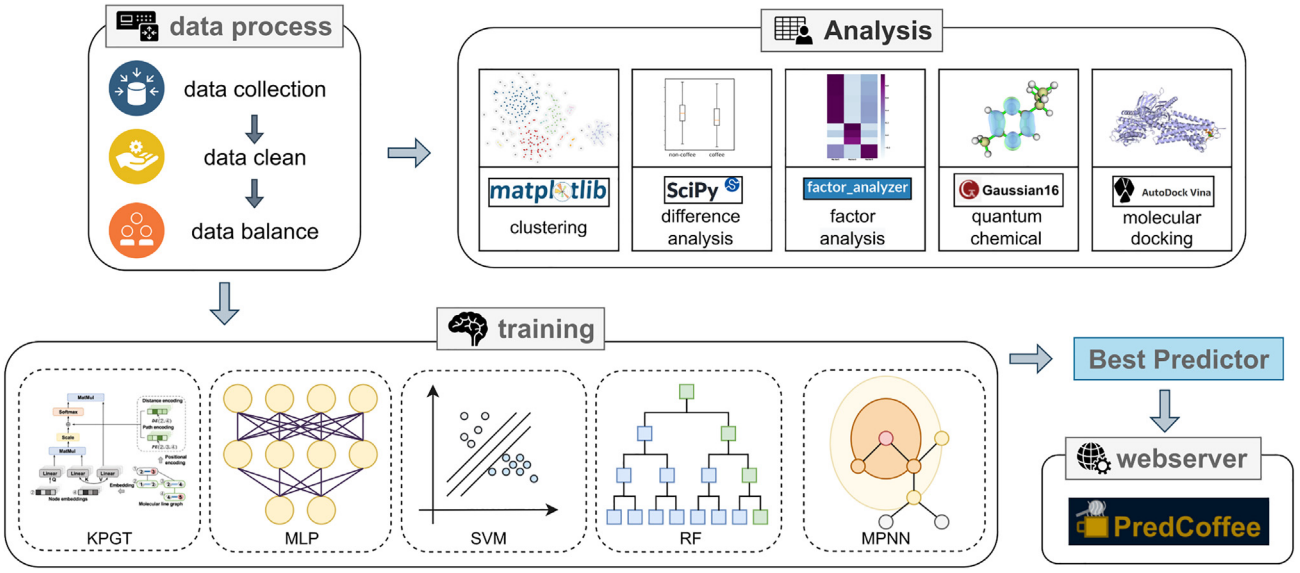


Figure 1. The workflow of our study

coffee odor dataset used in our study was sourced from some databases, including the LRI and Odor Database, and Leffingwell PMP 2001 and the training data provided by Phil Mennell during the "Learning Odors Challenge". Five models, KPGT,[18] SVM, RF,[19] MLP, and MPNN[20] that is one kind of GNN,[21] are used to predict and summarize the structure and property characteristics of coffee odor molecules. These algorithm models of machine learning are widely used in data analysis and processing in biology, medicine, and other industries, providing reliable data model support.[22–26] The optimal model is selected to build a binary classifier for predicting coffee odor, and it is used as the basis for building a website (Figure 1). This study provides a binary classification method, which provides a reliable basis for judging whether a molecule has coffee odor. This will provide a reference for the relation between molecular structure and biological activity or toxicity.

RESULTS AND DISCUSSION

Data

After the data cleaning, we got 271 coffee-odor molecules. We performed cluster analysis on 271 coffee-odor molecules, and it can be seen from Figure 2 that these 271 molecules are divided into different groups. By filtering out groups with fewer than 20 molecules, 205 of the 271 molecules were divided into four groups, and they were the 2-isopropyl-5-methylpyrazine group, (S)-2-methyl butyraldehyde group, difurfuryl disulfide group, hexahydrophenol group. We can see that the molecules grouped into the same group mostly have similar structural features. Next, we analyzed the specific structural characteristics of each group.

Difurfuryl disulfide is a chemical compound that belongs to the furan derivative group. It is composed of two furan rings connected by a disulfide bridge. Difurfuryl disulfide can exist in different forms, including derivatives such as ketones, ethers, and ester compounds. Some molecules within this group may also contain sulfur atoms or sulfhydryl groups. Compounds in the 2-isopropyl-5-methylpyrazine group are characterized by five-membered or six-membered ring structures. Most contain aliphatic side chains, including esters, ethers, and ketones. This group can be divided into two categories. The first is a compound with pyrazine as the skeleton, such as 2, 3-dimethyl-5-isopentylpyrazine, 2, 6-dimethylpyrazine, acetylpyrazine, 2,3-dimethylpyrazine, etc. The second category is thiazole-based compounds, such as benzothiazole, 2, 4-dimethylthiazole, 2, 4, 5-trimethylthiazole, etc.

(S)-2-methylbutyraldehyde group, which is the group containing the most molecules. This group mainly consists of two types of compounds, one of which is the non-cyclic chain compounds, including 3-mercapto-1-hexanol, 5-methyl-2-hepten-4-one, 4,6,9-trimethyl-3,5,8,10-tetraoxadodecane, etc. The other is cyclic compounds including dicyclohexyl disulfide, 3-mercapto-5-methyl-4, 5-dihydr-o-2(3H)-thio -phenone, etc. Most of the compounds in this group contain carbonyl, hydroxyl, or sulfhydryl groups. The hexahydrophenol group has a relatively uniform structure, with a benzene ring or thiophene as its structural skeleton, including benzyl mercaptan, 2-thiophenemethanethiol, and others.

Molecule structure and docking

We analyzed the interactions between the molecules (ligands) of the four clustered groups and protein (receptors), resulting in a heatmap (Figure S3). We found that the difurfuryl disulfide group, 2-isopropyl-5-methylpyrazine group, and hexahydrophenol group predominantly interact with the protein receptors through electrostatic and aromatic face-to-face interactions, while the

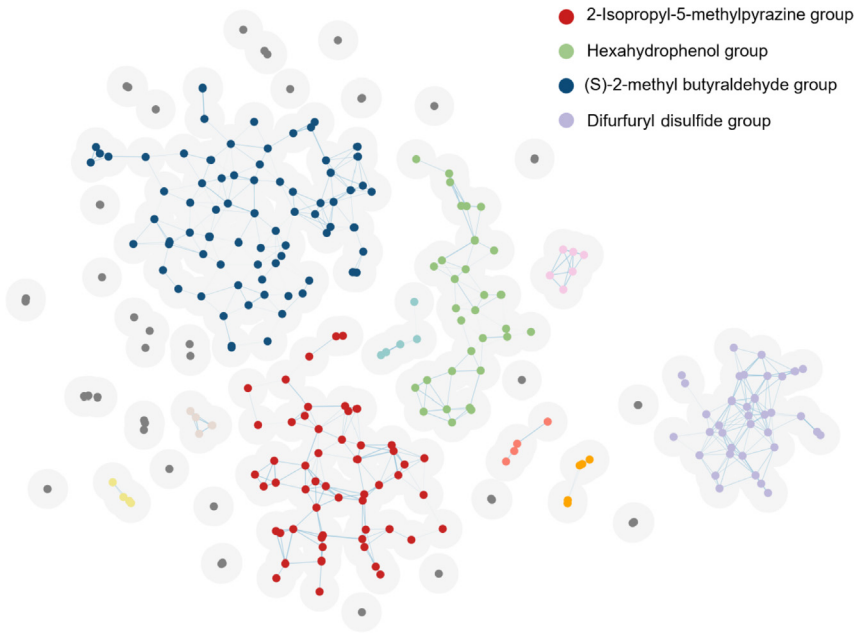


Figure 2. Clustering analysis of 271 coffee odor molecules

The light gray circles represent the threshold radius of molecules (1/48 of the distance between the farthest two molecules), and two molecules with intersecting radii are divided into a group. The different colors of the molecules represent different categories, groups containing fewer than 7 molecules are shown in gray, and groups containing more than 7 molecules are shown in color.

hexahydrophenol group mainly interacts with the protein receptor through hydrophobic interactions. During the interaction process between the molecular ligands of these four groups and the protein receptors, several amino acid residues such as Y197, T200, and A201 were involved in forming interactions, indicating that coffee flavor molecules primarily bind to the receptor through these amino acid residues.

Difurfuryl disulfide group

The HOMO and LUMO of difurfuryl disulfide are shown in Figure 3A, Difurfuryl disulfide has a furan ring on each side. The HOMO of the molecule falls on the carbon-carbon double bond and the disulfide bond in the middle of the furan ring, indicating that the molecule has strong reducibility and may lose electrons and form new chemical bonds. The molecule's LUMO falls on the carbon atoms on the left and right sides of the disulfide bond, which are each connected to two hydrogens, can receive foreign electrons, and has a certain oxidation property. Figure 4A shows the interaction of difurfuryl disulfide with the protein receptor OR51E2, PDB: 8F76. L185 is an alkaline amino acid that forms covalent bonds with the disulfide bond of the molecule. In addition, L192, M184, and V195 are hydrophobic amino acids that provide a hydrophobic environment for the binding of molecular ligands to protein receptors.

2-Isopropyl-5-methylpyrazine group

2-Isopropyl-5-methylpyrazine is the representative compound of its group, which contains a pyrazine ring with three methyl groups in the side chain (Figure 3B). The pyrazine ring has two carbon atoms attached to one hydrogen each, and these two carbon atoms can either gain electrons or lose electrons, both oxidizing and reducing. From the binding of this molecule to OR51E2 (Figure 4B), it can be seen that M184 and L185 can interact with methyl groups on their side faces to form covalent bonds, V195 interacts with the pyrazine ring in the form of Pi bonds, and T191 forms hydrogen bonds with nitrogen atoms on the pyrazine ring. Hydrogen bond is a common bond formed by the interaction between the pyrazine ring and protein receptor.[27]

(S)-2-Methyl butyraldehyde group

(S)-2-methyl butyraldehyde is an aldehyde with a relatively simple structure and no ring structure. From the HOMO orbit (left) and LUMO orbit (right) of this molecule (Figure 3C), it can be seen that the aldehyde group has both reducing and oxidizing properties. It can be oxidized to carboxylic acid or reduced to alcohol. From the interaction between (S)-2-meth-yl butyraldehyde and protein receptor, it can be seen that it interacts with T191 in the form of hydrogen bond, and V195, L185, and A187 also play an important role in the interaction between (S)-2-methyl butyraldehyde and protein.

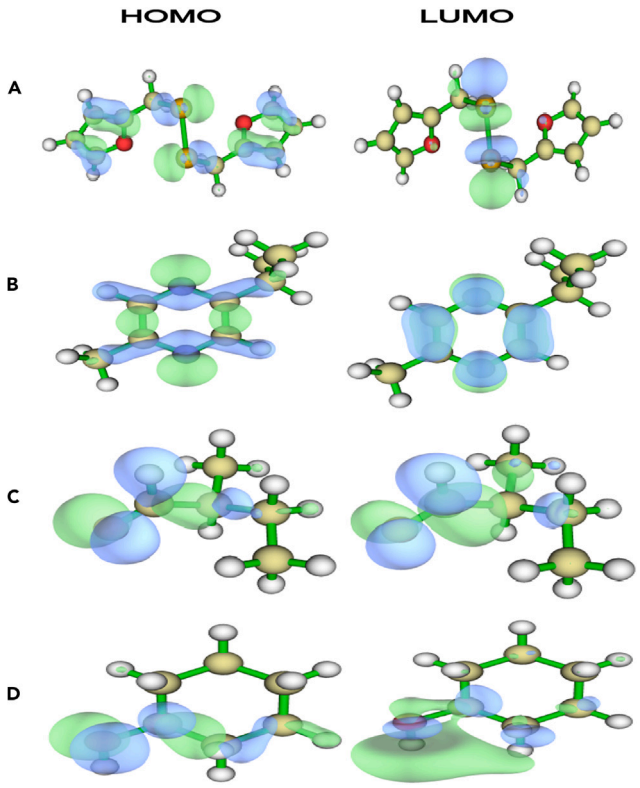


Figure 3. HOMO and LUMO of the four representative molecules

(A) Difurfuryl disulfide.

(B) 2-Isopropyl-5-methylpyrazine.

(C) (S)-2-methyl butyraldehyde.

(D) Hexahydrophenol.

Hexahydrophenol group

Hexahydrophenol is the representative molecule of this group. The chemical reaction occurs mainly in the phenolic hydroxyl group. In the process of its interaction with protein receptors, M184, and L185 have Alkyl, and A262 has hydrogen bond interactions with phenolic hydroxyl groups on small molecules.

Models

We tested the four selected models and represented the results in the form of a confusion matrix (Figure S2), about 45 molecules of the KPGT model were predicted to be coffee-odor molecules during the validation, the same as the real label, and about 9 non-coffee-odor molecules were predicted to be coffee-odor molecules. Similarly, it can be seen that the number of prediction errors in the remaining three models is between 10 and 20. In contrast, the prediction error of KPGT is less, indicating that its accuracy is better.

The performance of these five models in six metrics is shown in Figure 5, which shows each score of these four models (more details can be seen in Table S5). The KPGT had an accuracy of 0.84, which are significantly higher than that of the other four models, indicating that the KPGT has a good prediction effect on our data and is more suitable for being a prediction model.

Morgan fingerprints of molecules are easily visualized. To explore which Morgan fingerprints MLP learned were key to the molecule's coffee flavor, we analyzed the model using SHapley Additive exPlanations (SHAP). In Figure 6, we show the Morgan fingerprint substructure with the largest SHAP value with the largest number of 4 groups of coffee flavor molecules in the cluster (Figure 2). It is clear that the MLP model learns important structural features differently for different classes of molecules. For molecules in the Hexahydrophenol group (Figure 6D), the dominant Morgan fingerprint substructure is the conjugated structure on the phenyl group. For the Difurfuryl disulfide group (Figure 6A), the MLP considers the oxygen and carbon atoms of the furan group to be important. For the molecules of the 2-Isopropyl-5-methylpyrazine group (Figure 6B), it is important to have six-membered heterocyclic structures containing two nitrogen atoms, mainly pyrimidine and pyrazine structure in this group of molecules is crucial for coffee odor recognition. MLP believed that the oxygen atom or sulfhydryl group in the molecule of (S)-2-methyl butyraldehyde group and the surrounding carbon atom environment were the key substructural characteristics to distinguish coffee odor (Figure 6C). In summary, the MLP model learned key substructural features for each of the four classes of molecules that may be important for odor receptors to recognize coffee flavor molecules.

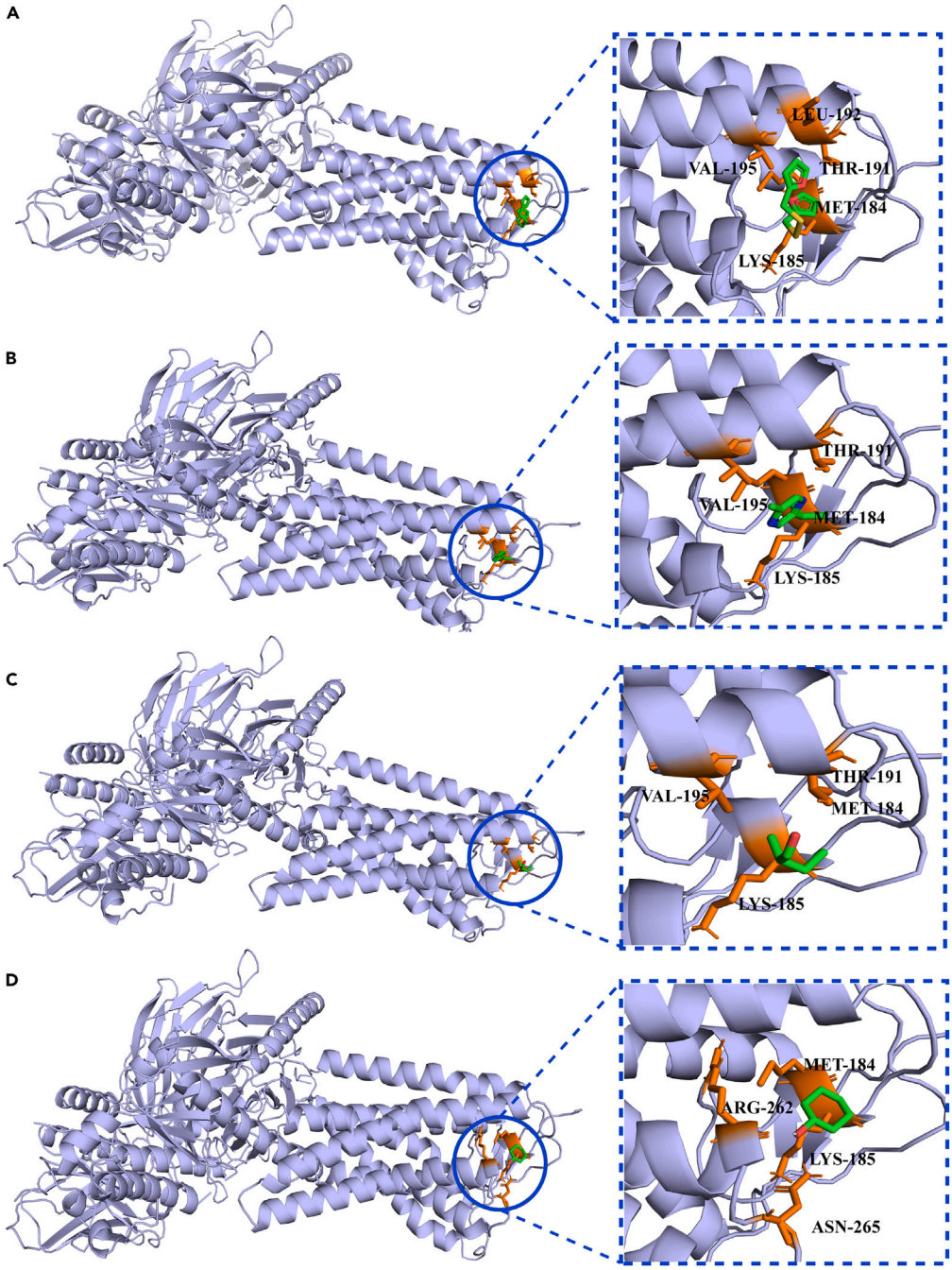


Figure 4. Molecular docking results of OR51E2 and four representative molecules

(A) Difurfuryl disulfide docking with OR51E2 and active residues around it.

(B) 2-Isopropyl-5-methylpyrazine docking with OR51E2 and active residues around it.

(C) (S)-2-methyl butyraldehyde docking with OR51E2 and active residues around it.

(D) Hexahydrophenol docking with OR51E2 and active residues around it.

Difference analysis

We performed a factor analysis of 208 properties of coffee/non-coffee odor molecules. Factor 1 and factor 2 explain 72% of the total variance in our dataset (Table S2). The value of logP[28] mainly characterized the hydrophobicity of molecules. As can be seen from Figure 7A, factor 1, factor 2 and factor 3 have higher eigenvalues, and we choose these three factors as common factors. Figure 7B shows the factor loading matrix of factor 1, factor 2, and factor 3 on 12 descriptors (Table S3). It can be seen that factor 1 predominantly contributes to 7 properties, such as Heavy Atom Mol

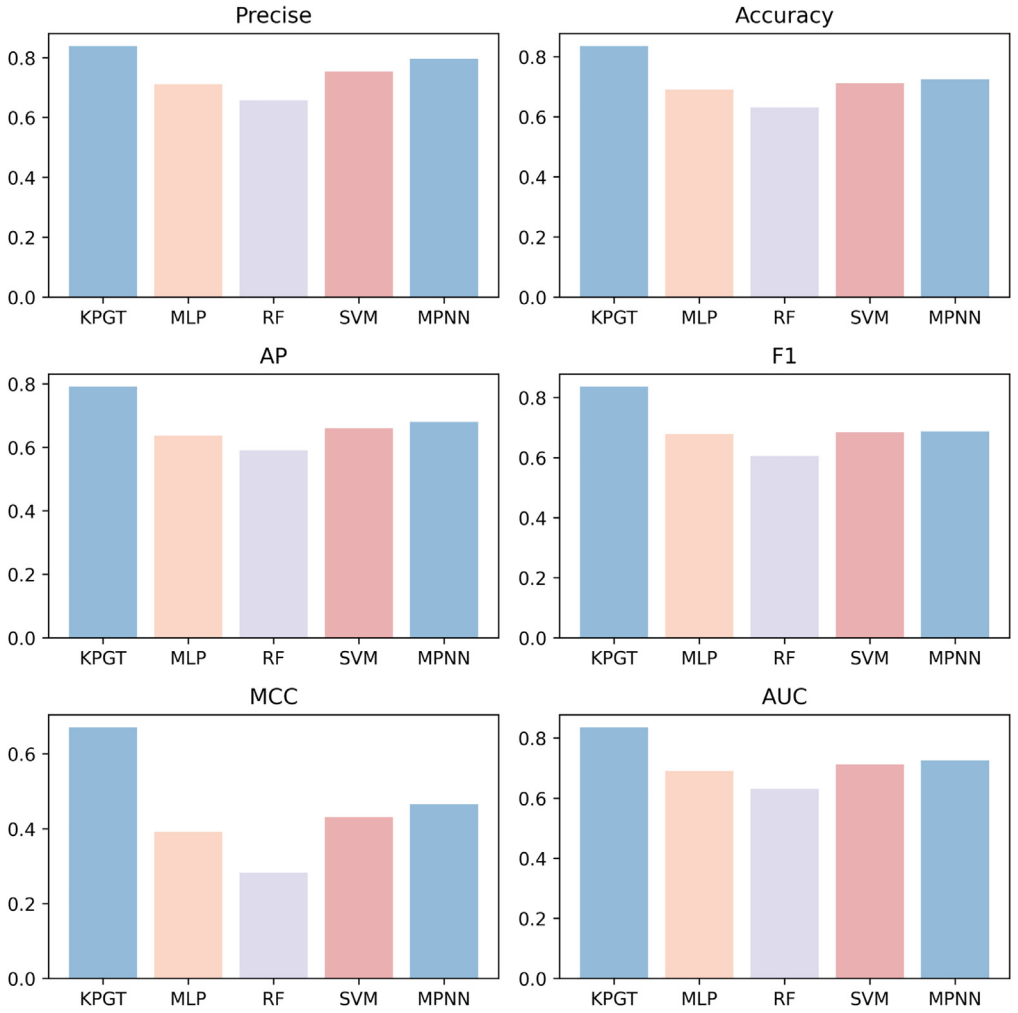


Figure 5. Performance of 5 models on 6 performance metrics

Wt and MW, while factor 2 primarily influences 3 properties, including the Morgan fingerprint of molecules. Factor 1 reflects the basic physical and chemical properties of molecules, such as size, shape, and electronic structure, which are fundamental factors that determine the behavior and function of molecules. Factor 2 represents the molecular fingerprint density, reflected the structural complexity and diversity of molecules, especially the potential for molecular recognition and intermolecular interactions. Factor 3 represents the lipid solubility of the molecule and is a measure of the lipid solubility of the molecule. We can see the correlation between the three common factors and the 12 descriptors (Table S3), the darker the color and the larger the value, the higher the correlation (Figure 7C). The performance of coffee odor and non-coffee odor molecules in 12 properties was normalized (Figure 7D), where 0 represents non-coffee odor molecules and 1 represents coffee odor molecules.

Specific numerical comparisons were made between coffee odor molecules and non-coffee odor molecules in 12 specific properties (Figure 8), and p-value <0.05 proved that coffee and non-coffee molecules were different in the corresponding properties. As can be seen from Figure 8, the p-values of coffee odor molecules and non-coffee odor molecules in 12 properties are all less than 0.05, indicating that there are significant differences between coffee and non-coffee molecules in these properties. Coffee-odor molecules have smaller molecular weights and volumes, more FpDensityMorgan 1–3, and are less hydrophobic than non-coffee odor molecules.

Webserver

Based on the selected KPGT model, we built a website named PredCoffee with the URL https://hwwlab.com/webserver/predcoffee. By entering the SMILES format of the molecule you want to query in the prediction option on the homepage of the website (Figure 9), you can find out whether the molecule is coffee-odor.

Our research outcomes are effectively disseminated through the platform of this website, offering enhanced practical utility and societal impact. This approach significantly economizes on human resources and material costs, considering the substantial investment of time, energy, and financial resources required to cultivate proficient odor assessors. Moreover, the inherent constraint of finite human labor hours, in

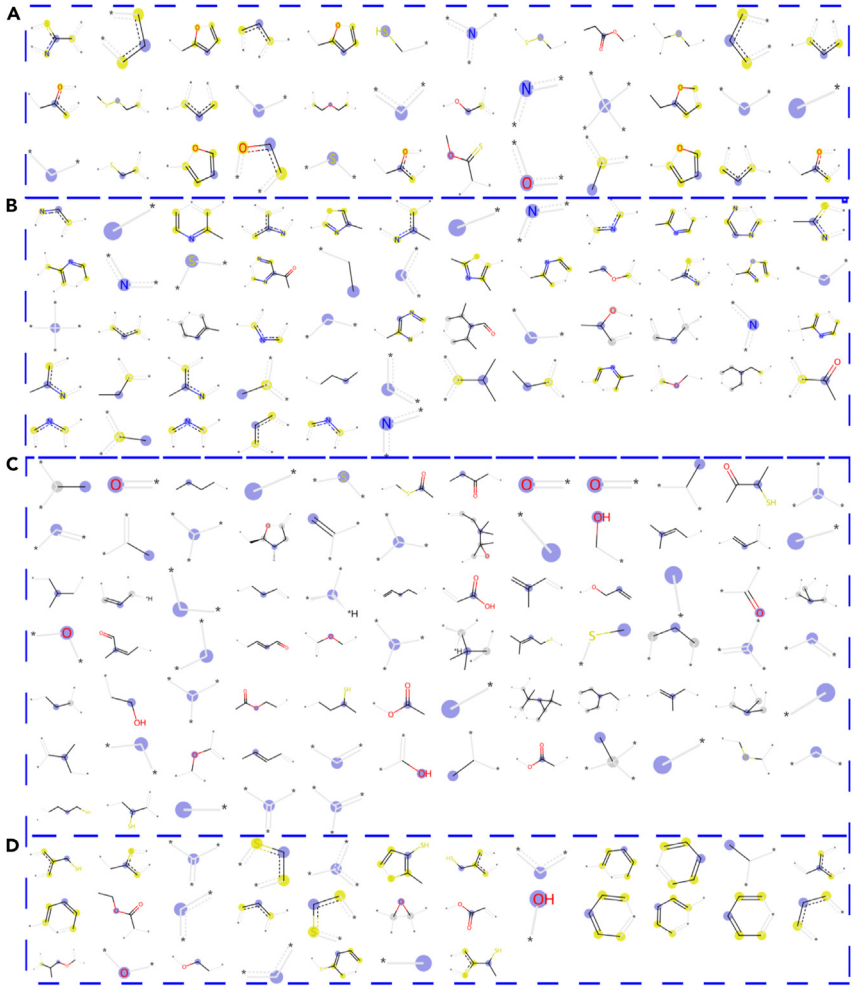


Figure 6. Most important Morgan fingerprint substructure that MLP learned of molecules

(A) Difurfuryl disulfide group.

(B) 2-Isopropyl-5-methylpyrazine group.

(C) (S)-2-methyl butyraldehyde group.

(D) Hexahydrophenol group. Blue highlight indicated that the atomic or molecular fragment corresponds directly to the activated bit in Morgan's fingerprint. Yellow highlight was used to indicate atoms that are adjacent to blue highlighted atoms or that contribute to the generation of fingerprint sites but do not directly determine their activation. Uncolored atoms did not contribute directly to generating the Morgan fingerprint bit of the current focus.

contrast to the ceaseless operational capacity of machines, underscores the efficiency and advantage of utilizing our classifier for odor prediction. Additionally, the classifier's capability to transcend temporal and spatial restrictions contributes to its convenience and efficacy in odor assessment applications.

Limitations of the study

The binary classification machine learning for coffee odor molecules is based on ligand-based algorithms. In the future, consideration will be given to using receptor-based algorithms. Additionally, the scale of the coffee odor dataset trained in this study is too small, and efforts will be made to expand the coffee odor dataset.

Conclusions

In this study, cluster analysis and molecular docking analysis were carried out on 271 coffee odor molecules, corresponding data models were selected to predict coffee odor and non-coffee odor molecules, the most suitable model for the data in this study is KPGT, and a binary classifier was constructed. The prediction of whether a molecule has coffee odor has a high accuracy, which reached 0.84. Coffee-odor molecules have smaller molecular weights and volumes, more Morgan fingerprints 1–3, and are less hydrophobic than non-coffee odor molecules. The odor prediction of molecules is realized.

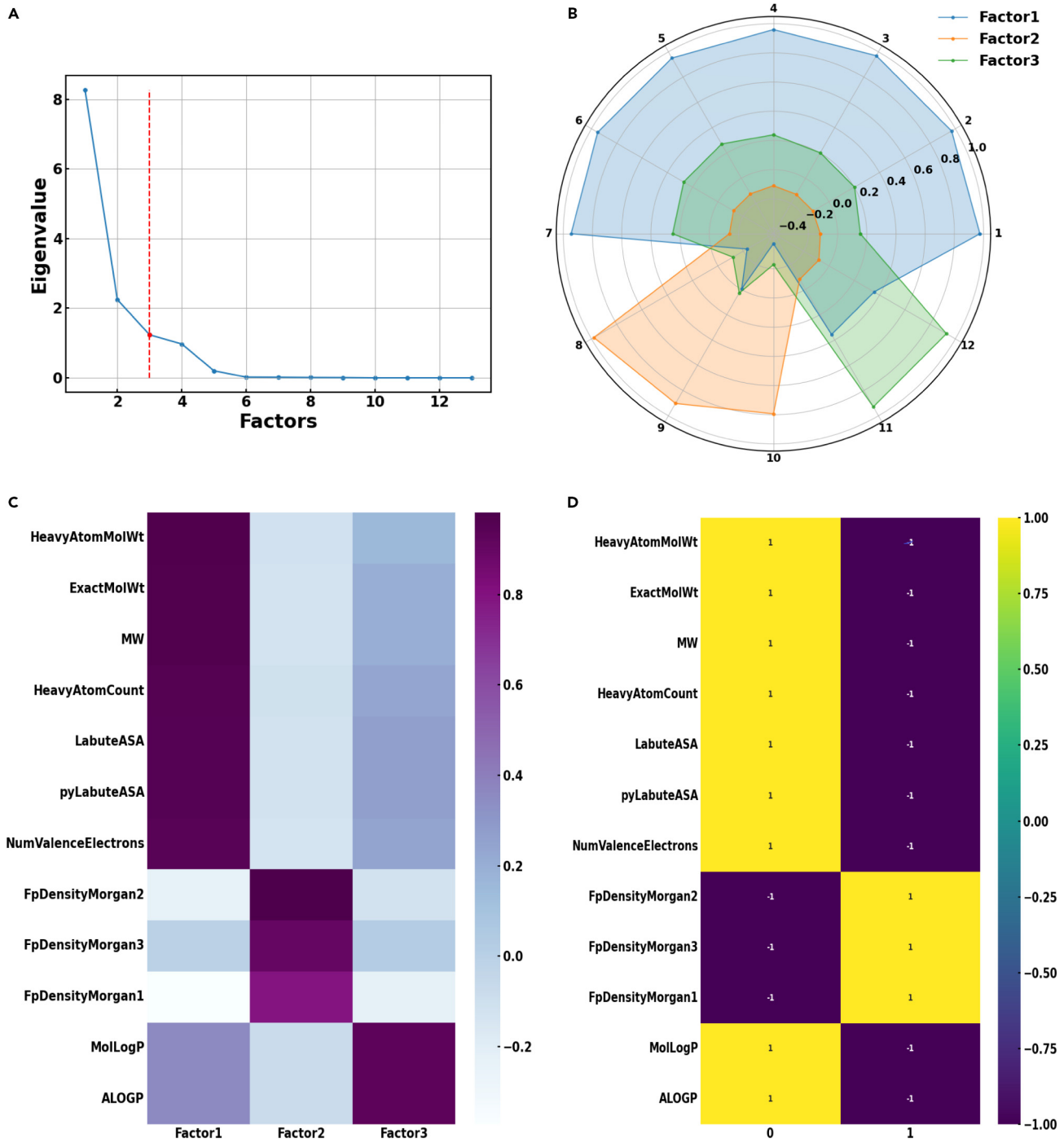


Figure 7. Factor analysis of the coffee/non-coffee dataset

(A) The scree plot for eigenvalue with factor.

(B) The radar chart of the screened descriptors contributing to 3 factors.

(C) The heatmap of factor loading matrix.

(D) Normalized 12 molecular properties of coffee and non-coffee.

The binary classifier constructed using machine learning methods not only predicts the odor of molecules but also estimates the intensity of the odor. It can efficiently perform odor prediction for a large number of molecules in a short period of time, which is difficult to achieve artificially. This improves the accuracy and efficiency of odor identification. Additionally, we have identified

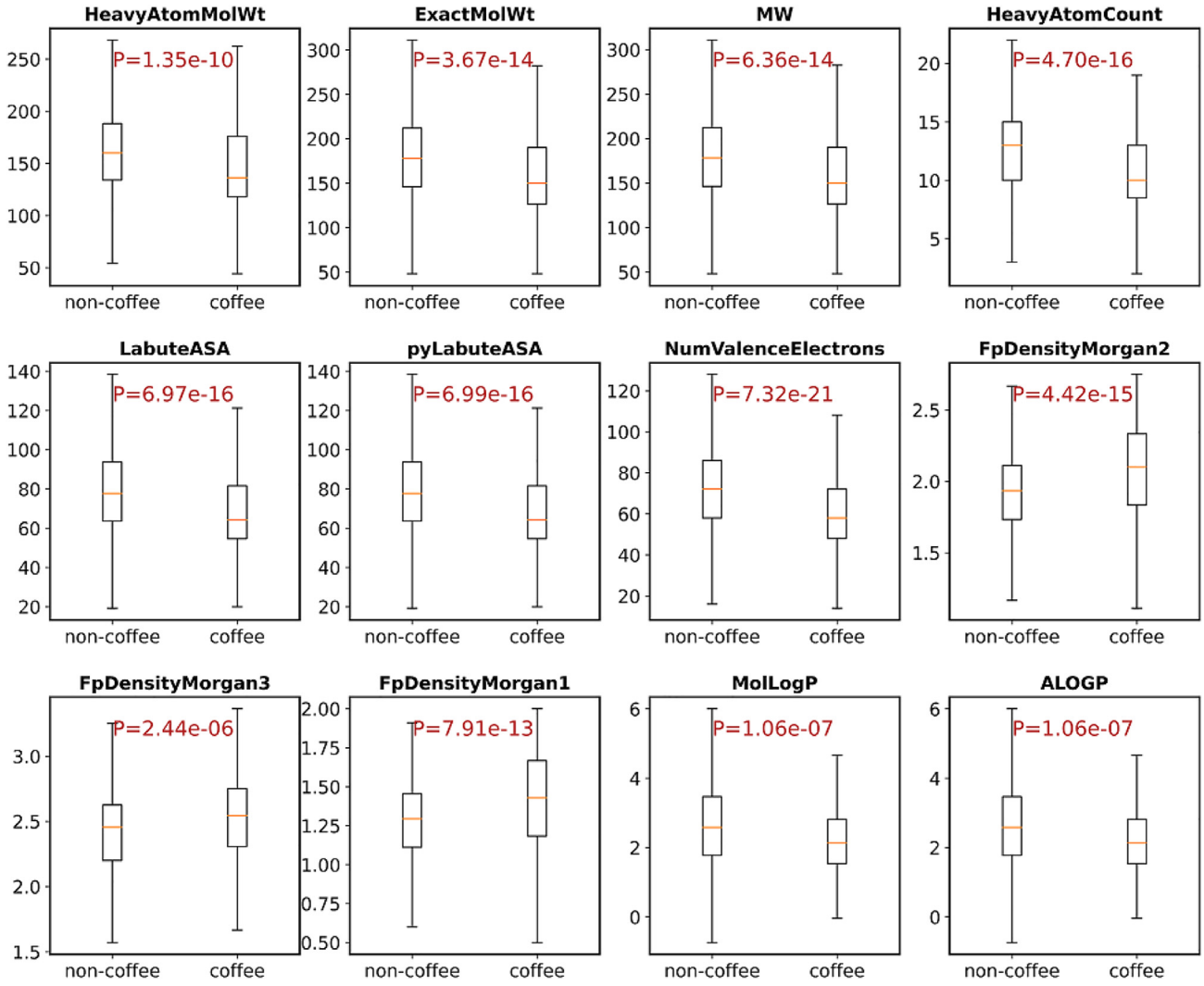


Figure 8. Significant difference analysis of 12 properties of coffee and non-coffee molecules

The *p*-value indicates the difference between two samples, the smaller the *p*-value, the more significant the difference.

characteristic structures of molecules with coffee aroma, providing a predictive basis and theoretical foundation for industries such as flavor formulation. For example, by predicting and screening certain components in plants, it is possible to formulate beverages with coffee odor but without caffeine, promoting product development in the food industry. It can also predict the odor of a molecule with a known chemical formula without the need for time-consuming wet lab experiments such as synthesis from scratch, significantly saving manpower and resources. The enhanced ability of AI in odor analysis holds significant value and significance in the fields of food, medicine, and chemical engineering.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS

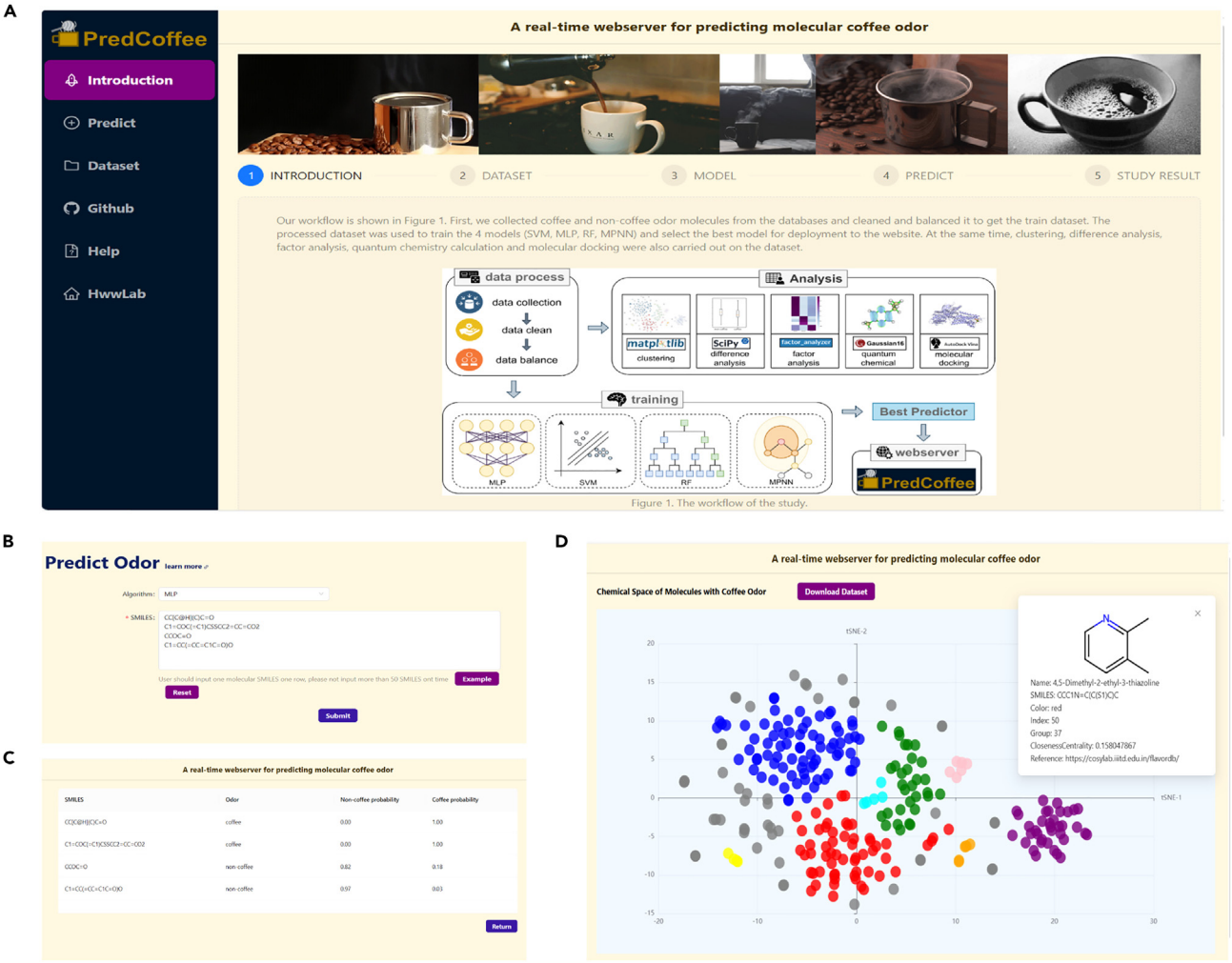


Figure 9. Webserver of PredCoffee

(A) Website homepage of PredCoffee.

(B) Submit page of PredCoffee.

(C) Result page of PredCoffee.

(D) Chemical space of coffee molecules.

- The workflow of the study
- The dataset processing
- The models for coffee odor prediction
- Clustering of the coffee odor molecules
- Factor analysis and difference analysis
- Molecular docking and quantum chemistry calculations
- Webserver

- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110041.

ACKNOWLEDGMENTS

This study was supported by the Key Project of the Jilin Education Department (Grant No. JJKH20231140KJ) and the Natural Science Foundation of Chongqing (Grant No. CSTB2022NSCQ-MSX1043).

AUTHOR CONTRIBUTIONS

Y.H.: Conceptualization, methodology, and software. R.H.: data collection, writing original draft, and visualization. R.Z.: Investigation. F.H.: Validation and supervision. L.H.: validation, supervision. W.H.: Validation and supervision.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: January 24, 2024
Revised: April 26, 2024
Accepted: May 16, 2024
Published: May 21, 2024

REFERENCES

1. Hatt, H. (2004). Molecular and cellular basis of human olfaction. Chem. Biodivers. *1*, 1857–1869. https://doi.org/10.1002/cbdv.200490142.
2. Menini, A., Lagostena, L., and Boccaccio, A. (2004). Olfaction: from odorant molecules to the olfactory cortex. News Physiol. Sci. *19*, 101–104. https://doi.org/10.1152/nips.1507.2003.
3. Rinaldi, A. (2007). The scent of life. The exquisite complexity of the sense of smell in animals and humans. EMBO Rep. *8*, 629–633. https://doi.org/10.1038/sj.embor.7401029.
4. Brookes, J.C. (2010). Science is perception: what can our sense of smell tell us about ourselves and the world around us? Philos. Trans. A Math. Phys. Eng. Sci. *368*, 3491–3502. https://doi.org/10.1098/rsta.2010.0117.
5. Braun, T., Doerr, J.M., Peters, L., Viard, M., Reuter, I., Prosiegel, M., Weber, S., Yeniguen, M., Tschernatsch, M., Gerriets, T., et al. (2022). Age-related changes in oral sensitivity, taste and smell. Sci. Rep. *12*, 1533. https://doi.org/10.1038/s41598-022-05201-2.
6. Weiffenbach, J.M., and Bartoshuk, L.M. (1992). Taste and smell. Clin. Geriatr. Med. *8*, 543–555.
7. Butt, M.S., and Sultan, M.T. (2011). Coffee and its consumption: benefits and risks. Crit. Rev. Food Sci. Nutr. *51*, 363–373. https://doi.org/10.1080/10408390903586412.
8. O'Keefe, J.H., Bhatti, S.K., Patil, H.R., DiNicolantonio, J.J., Lucan, S.C., and Lavie, C.J. (2013). Effects of habitual coffee consumption on cardiometabolic disease, cardiovascular health, and all-cause mortality. J. Am. Coll. Cardiol. *62*, 1043–1051. https://doi.org/10.1016/j.jacc.2013.06.035.
9. O'Keefe, J.H., DiNicolantonio, J.J., and Lavie, C.J. (2018). Coffee for Cardioprotection and Longevity. Prog. Cardiovasc. Dis. *61*, 38–42. https://doi.org/10.1016/j.pcad.2018.02.002.
10. Sartorelli, D.S., Fagherazzi, G., Balkau, B., Touillaud, M.S., Boutron-Ruault, M.C., de Lauzon-Guillain, B., and Clavel-Chapelon, F. (2010). Differential effects of coffee on the risk of type 2 diabetes according to meal consumption in a French cohort of women: the E3N/EPIC cohort study. Am. J. Clin. Nutr. *91*, 1002–1012. https://doi.org/10.3945/ajcn.2009.28741.
11. Fadel, H.H.M., Abdel Mageed, M.A., and Lotfy, S.N. (2008). Quality and flavour stability of coffee substitute prepared by extrusion of wheat germ and chicory roots. Amino Acids *34*, 307–314. https://doi.org/10.1007/s00726-006-0434-7.
12. Street, R.A., Sidana, J., and Prinsloo, G. (2013). Cichorium intybus: Traditional Uses, Phytochemistry, Pharmacology, and Toxicology. Evid. Based. Complement. Alternat. Med. *2013*, 579319. https://doi.org/10.1155/2013/579319.
13. Wu, T., and Cadwallader, K.R. (2019). Identification of Characterizing Aroma Components of Roasted Chicory "Coffee" Brews. J. Agric. Food Chem. *67*, 13848–13859. https://doi.org/10.1021/acs.jafc.9b00776.
14. Greener, J.G., Kandathil, S.M., Moffat, L., and Jones, D.T. (2022). A guide to machine learning for biologists. Nat. Rev. Mol. Cell Biol. *23*, 40–55. https://doi.org/10.1038/s41580-021-00407-0.
15. Dey, T.K., Mandal, S., and Mukherjee, S. (2022). Gene expression data classification using topology and machine learning models. BMC Bioinf. *22*, 627. https://doi.org/10.1186/s12859-022-04704-z.
16. Saini, K., and Ramanathan, V. (2022). Predicting odor from molecular structure: a multi-label classification approach. Sci. Rep. *12*, 13863. https://doi.org/10.1038/s41598-022-18086-y.
17. Lee, B.K., Mayhew, E.J., Sanchez-Lengeling, B., Wei, J.N., Qian, W.W., Little, K.A., Andres, M., Nguyen, B.B., Moloy, T., Yasonik, J., et al. (2023). A principal odor map unifies diverse tasks in olfactory perception. J. Science *381*, 999–1006. https://doi.org/10.1126/science.ade4401.
18. Li, H., Zhang, R., Min, Y., Ma, D., Zhao, D., and Zeng, J. (2023). A knowledge-guided pre-training framework for improving molecular representation learning. Nat. Commun. *14*, 7568. https://doi.org/10.1038/s41467-023-43214-1.
19. Rigatti, S.J. (2017). Random Forest. J. Insur. Med. *47*, 31–39. https://doi.org/10.17849/insm-47-01-31-39.1.
20. Das Sarma, S., Deng, D.L., and Duan, L.M. (2019). Machine learning meets quantum physics. Phys. Today *72*, 48–54. https://doi.org/10.1063/pt.3.4164.
21. Bacciu, D., Errica, F., Micheli, A., and Podda, M. (2020). A gentle introduction to deep learning for graphs. Neural Network. *129*, 203–221. https://doi.org/10.1016/j.neunet.2020.06.006.
22. Asadi, S., Roshan, S., and Kattan, M.W. (2021). Random forest swarm optimization-based for heart diseases diagnosis. J. Biomed. Inf. *115*, 103690. https://doi.org/10.1016/j.jbi.2021.103690.
23. Ghosh, D., and Cabrera, J. (2022). Enriched Random Forest for High Dimensional Genomic Data. IEEE ACM Trans. Comput. Biol. Bioinf. *19*, 2817–2828. https://doi.org/10.1109/tcbb.2021.3089417.
24. Rodríguez-Pérez, R., and Bajorath, J. (2022). Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. J. Comput. Aided Mol. Des. *36*, 355–362. https://doi.org/10.1007/s10822-022-00442-9.
25. Yan, J., Wang, X., Cai, J., Qin, Q., Yang, H., Wang, Q., Cheng, Y., Gan, T., Jiang, H., Deng, J., and Chen, B. (2022). Medical image segmentation model based on triple gate MultiLayer perceptron. Sci. Rep. *12*, 6103. https://doi.org/10.1038/s41598-022-09452-x.
26. Zhang, Z., Chen, L., Zhong, F., Wang, D., Jiang, J., Zhang, S., Jiang, H., Zheng, M., and Li, X. (2022). Graph neural network approaches for drug-target interactions. Curr. Opin. Struct. Biol. *73*, 102327. https://doi.org/10.1016/j.sbi.2021.102327.
27. Juhás, M., and Zitko, J. (2020). Molecular Interactions of Pyrazine-Based Compounds to Proteins. J. Med. Chem. *63*, 8901–8916. https://doi.org/10.1021/acs.jmedchem.9b02021.
28. Erös, D., Kövesdi, I., Örfi, L., Takács-Novák, K., Acsády, G., and Kéri, G. (2002). Reliability of logP predictions based on calculated molecular descriptors:: A critical review. Curr. Med. Chem. *9*, 1819–1829. https://doi.org/10.2174/0929867023369042.
29. Scalfani, V.F., Patel, V.D., and Fernandez, A.M. (2022). Visualizing chemical space networks with RDKit and NetworkX. J. Cheminf. *14*, 87. https://doi.org/10.1186/s13321-022-00664-x.
30. He, Y., Liu, K., Han, L., and Han, W. (2022). Clustering Analysis, Structure Fingerprint Analysis, and Quantum Chemical Calculations of Compounds from Essential Oils of Sunflower (*Helianthus annuus* L.) Receptacles. Int. J. Mol. Sci. *23*, 10169. https://doi.org/10.3390/ijms231710169.
31. Billesbølle, C.B., de March, C.A., van der Velden, W.J.C., Ma, N., Tewari, J., Del Torrent, C.L., Li, L., Faust, B., Vaidehi, N., Matsunami, H., and Manglik, A. (2023). Structural basis of odorant recognition by a human odorant receptor. Nature *615*,

742–749. https://doi.org/10.1038/s41586-023-05798-y.

32. Glezer, I., and Malnic, B. (2019). Olfactory receptor function. Handb. Clin. Neurol. *164*, 67–78. https://doi.org/10.1016/b978-0-444-63855-7.00005-8.
33. Sharma, A., Kumar, R., Aier, I., Semwal, R., Tyagi, P., and Varadwaj, P. (2019). Sense of Smell: Structural, Functional, Mechanistic Advancements and Challenges in Human Olfactory Research. Curr. Neuropharmacol. *17*, 891–911. https://doi.org/10.2174/1570159x17666181206095626.
34. Touhara, K. (2002). Odor discrimination by G protein-coupled olfactory receptors. Microsc. Res. Tech. *58*, 135–141. https://doi.org/10.1002/jemt.10131.
35. Istyastono, E.P., Radifar, M., Yuniarti, N., Prasasty, V.D., and Mungkasi, S. (2020). PyPLIF HIPPOS: A Molecular Interaction Fingerprinting Tool for Docking Results of AutoDock Vina and PLANTS. J. Chem. Inf. Model. *60*, 3697–3702. https://doi.org/10.1021/acs.jcim.0c00305.
36. Saikia, S., and Bordoloi, M. (2019). Molecular Docking: Challenges, Advances and its Use in Drug Discovery Perspective. Curr. Drug Targets *20*, 501–521. https://doi.org/10.2174/1389450119666181022153016.
37. Lu, T., and Chen, F. (2012). Multiwfn: a multifunctional wavefunction analyzer. J. Comput. Chem. *33*, 580–592. https://doi.org/10.1002/jcc.22885.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Receptor for molecular docking	RCSB PDB Code: 8F76	https://www.rcsb.org/structure/8F76
Molecualar dataset	This paper	https://hwwlab.com/webserver/predcoffee/dataset
Source Code	This paper	https://github.com/heyigacu/predcoffee
All data	This paper	https://zenodo.org/records/11067667
Software and algorithms		
KPGT	Li et al.[1]	https://github.com/lihan97/KPGT
MPNN	DGL-LifeSci	https://arxiv.org/abs/2106.14232
SVM and RF	sklearn 1.4.2	https://scikit-learn.org/stable/index.html
MLP	pytorch 2.1.2 + cuda 11.8	https://pytorch.org/
tSNE-CSN	Istyastono et al.[2]	https://github.com/heyigacu/DsitanceClustering
Vina	Eberhardt et al.[3]	RRID:SCR_011958; https://vina.scripps.edu/
Gaussian16	Gaussian16	RRID:SCR_014897; https://gaussian.com/gaussian16/
GaussView6	GaussView6	https://gaussian.com/gaussview6/
PyPLIF-HIPPOS	Istyastono et al.[2]	https://github.com/radifar/PyPLIF-HIPPOS
RDKit	RDKit 2023.9.5	RRID:SCR_014274; https://www.rdkit.org/docs/index.html
Open Babel	Open Babel 3.0.1	RRID:SCR_014920; https://openbabel.org
SHAP	SHAP-0.45.0	RRID:SCR_021362; https://github.com/shap/shap
PyMOL	PyMOL 3.0	RRID:SCR_000305; https://pymol.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Weiwei Han (weiweihan@jlu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Coffee-odor molecular dataset can be downloaded from https://hwwlab.com/webserver/predcoffee/dataset. All data reported in this paper will can be available at https://zenodo.org/records/11067667. All of the code in this article, including machine learning and analysis, is available at https://github.com/heyigacu/predcoffee. Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study is computational, so we omit this section.

METHOD DETAILS

The workflow of the study

Our workflow is shown in Figure 1. We collected coffee and non-coffee odor molecules from the databases, and then cleaned and balanced them to get the training dataset. The processed dataset was used to train the 5 models (KPGT, SVM, MLP, RF, MPNN) and select the best model for deployment to the website. At the same time, clustering, difference analysis, factor analysis, quantum chemistry calculation, and molecular docking were also carried out on the dataset.

The dataset processing

As shown in Figure S1, we collected 371 coffee-odor molecules and 9,700 non-coffee-odor molecules from the databases. The data cleaning included removing molecules that were not recognized by RDKit[29] and DGL and deleting duplicate molecules with canonical SMILES respectively in coffee and non-coffee odor molecules, resulting in 271 coffee and 5,758 non-coffee odor molecules. To balance the dataset, 5758 non-coffee odor molecules were stratified and upsampled to 270 molecules to ensure that the ratio of coffee and non-coffee was about 1:1. The final coffee/non-coffee dataset was cross-validated with 5-folds and repeated 10 times.

The models for coffee odor prediction

We used the four models for training coffee odor predictors, detailed information of the models is as below.

(1) KPGT (Knowledge-guided Pre-training of Graph Transformer) is a self-supervised learning framework that can learn generalizable and robust molecular representations. The code for KPGT can be obtained at https://github.com/lihan97/KPGT, and it has been pre-trained on 2 million molecules. Here, we are using the odor dataset of coffee/non-coffee for fine-tuning.

(2) We constructed MLP with PyTorch (https://pytorch.org/). Firstly, the molecules of the dataset are converted into Morgan fingerprints with a radius of 2 and a length of 2048 bits by RDKit (https://www.rdkit.org/), and batch size was set to 1/16 of the total number of the molecules. Then the fingerprints was input into the MLP with the input layer containing 256 neurons, the hidden layer of 256 neurons, and the output layer of 2 neurons (refer to the MLP schematic diagram in Figure 1). Between each layer, there is a ReLU activation function and a 0.1 dropout. The trainer used the cross-entropy loss function, Adam optimizer, and a learning rate of 0.001, and the early stop strategy was used to stop the training when the number of times the loss no longer drops accumulates to 7.

(3) The SVM model was constructed with scikit-learn (https://scikit-learn.org/). The input for the SVM is the same as for the MLP above. The optimal parameters of the SVM were determined using a 5-fold cross-validated grid search method. The optimal parameters are {"C": 1, "gamma": 0.1, "kernel": "rbf", "probability": True}.

(4) The input for the RF model is the same as for the SVM above. The optimal parameters of the RF were determined using a 5-fold cross-validated grid search method. The optimal parameters are {"max_depth": 6, "max_features": "log2", "min_samples_leaf": 50, "min_samples_split": 2, "n_estimators": 100, "probability": True}.

(5) We built the MPNN with DGL-LifeSci (https://lifesci.dgl.ai/). The input for the MPNN is the DGL graphs of the molecules, and batch size was set to 1/16 of the total number of the molecules. Node feature and edge feature embedding using canonical atom and bond featurization that generated 74 one-hot coding features for atoms and 12 one-hot coding features for bonds (Table S4). The node output dimension and edge output dimension of MPNN were set to 64 and 128 respectively, and other parameters were set to default. Training parameters like early stopping, learning rate, loss function, and optimizer are the same as MLP.

Clustering of the coffee odor molecules

RDKit was used to calculate the Morgan fingerprint vector with a length of 2048 and a radius of the molecule, and then used t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce dimensionality to two-dimensional space. This space represents the chemical space of astringent molecules. Compared with the K-means method, we chose another clustering method that can clearly distinguish the margins of different classes, which considers that if the distance between two molecules is less than 1/24 of the distance between the farthest two molecules in the entire molecular set, it is considered a group, and then all the molecules are traversed to complete the clustering.

We visualize the chemical space network (CSN)[30] based on Matplotlib (https://matplotlib.org/) after clustering, including visualizing the node radius to characterize the distance cutoff and edge thickness to represent the Dice similarity coefficient between 2 molecules, the above clustering code can be obtained at https://github.com/heyigacu/DistanceClustering.

Factor analysis and difference analysis

We select 208 molecular property descriptors based on the computational chemistry package RDKit (https://www.rdkit.org) for factor analysis, about the meaning of those descriptors is shown in Table S5. After passing the KMO and Bartlett spherical tests, select the appropriate common factor number to obtain the component matrix and calculate the score of the descriptor. Descriptors with factor loading values greater than 0.75 in the component matrix are selected to test whether A and B are significantly different in distribution.

Regarding the significance difference test, first, use the Shapiro-Wilk test to check whether the two samples are normally distributed, and if not, use the Mann-Whitney U-test to directly test whether the two samples are different. If both samples conform to normality, we use the Levene test to check whether the two samples are homogeneous with variance, and if the variance is homogeneous, we carry out an independent T-test, otherwise, we use Welch T-test to determine whether there is a difference between two samples. The above test analysis applies to all of the significance difference tests in this paper, which are performed with SciPy (https://scipy.org/) and under p-value ≤ 0.01.

Molecular docking and quantum chemistry calculations

The molecule with a higher value of proximity centrality is considered more representative of the entire group. To identify representative molecules, we focus on the four clusters with the highest number of molecules. From each of these clusters, we select the molecule with the highest closeness centrality as the representative for that specific group (Table S1). Subsequently, we can proceed with molecular docking work to

investigate the interactions between these representative molecules and target protein. We chose a human olfactory receptor, OR51E2 (PDB code 8F76), as the receptor. The structure of OR51E2 is relatively conserved in the evolutionary process, and it is widely expressed in human olfactory cells, which is representative.[31] Most human olfactory receptors belong to G protein-coupled receptors.[32–34] The transmembrane structure of this protein is composed of seven alpha helices, and it is at this site where each molecule binds to it. Protein-ligands interaction fingerprints were constructed by PyPLIF HIPPOS.[35]

Next, we did molecular docking.[36] We use Autodocktools to process the small molecule ligands and protein receptors, find out the binding sites, and perform molecular docking through Autodovk vina software. After that, PyMOL and Discovery Studio were used to analyze the molecular docking results.

We used Gaussian16 (https://gaussian.com) to perform quantum chemistry calculations for the above molecules, and Multiwfn[37] to visualize the LUMO and HOMO orbitals.

Webserver

The front-end of the website uses the front-end language React (https://react.dev/) and user interface (UI) library Antd (https://ant.design), the back-end uses Django (https://www.djangoproject.com/) based on the model-view-controller (MVC) framework, and the server is real-time responsive.

QUANTIFICATION AND STATISTICAL ANALYSIS

Regarding the significance difference test, first, use the Shapiro-Wilk test to check whether the two samples are normally distributed, and if not, use the Mann-Whitney U-test to directly test whether the two samples are different. If both samples conform to normality, we use the Levene test to check whether the two samples are homogeneous with variance, and if the variance is homogeneous, we carry out an independent T-test, otherwise, we use Welch T-test to determine whether there is a difference between two samples. The above test analysis applies to all of the significance difference tests in this paper, which are performed with SciPy (https://scipy.org/) and under p-value ≤ 0.01.

ADDITIONAL RESOURCES

PredCoffee webserver can be available at https://hwwlab.com/webserver/predcoffee.