Journal of Nonlinear and Convex Analysis
Volume 26, Number 7, 2025, 2049–2060

Yokohama Publishers
ISSN 1880-5221 ONLINE JOURNAL


# ODOR RECOGNITION WITH MOLECULAR FINGERPRINTS AND GRAPHS THROUGH MACHINE LEARNING

RENJIE XIA, YUNSONG WANG, FENGFEI WANG, WEI LI, LIANGXU XIE*, AND XIAOJUN XU†

Abstract. Odor recognition holds significant academic importance and drives innovation and competitiveness across various commercial sectors. Machine learning technology has facilitated the exploration of the structure-activity relationship of different molecules. In this study, we employ machine learning techniques using molecular fingerprints and graphs for odor recognition, specifically examining the impact of imbalanced data on machine learning-base odor recognition. Results obtained from the Leffingwell dataset demonstrate that molecular fingerprints (MACCS, ECFP, and PaDEL-descriptor) and graphs effectively represent odorant molecules, enabling accurate odor classification. Further evaluation of a multimodal model reveals the potential of RGCN in capturing the complex structural characteristics of odor molecules. Although an increase in class imbalance adversely affects the overall accuracy of multi-label odor recognition, the predictive accuracy for individual labels remains remarkably robust. These findings underscore the importance of addressing data imbalance, and highlight the potential of machine learning and multimodal approaches for advancing our understanding of odor perception.

## 1. Introduction

The interaction between volatile odorant molecules in the atmosphere and olfactory receptors (ORs) within the nasal cavity facilitates our ability to detect a wide range of odors in our environment, from the distinct fragrance of flowers to the fruity aroma of fresh fruits, and beyond. Upon binding to ORs located on the walls of the nasal cavity, these molecules become activated, initiating the generation of electrophysiological signals. Subsequently, these signals are transmitted through neurons to the olfactory nerves within the nasal cavity. The olfactory nerve then conveys these signals to the olfactory bulb in the brain [1, 12], where they undergo analysis, comparison, and recognition, resulting in the formation of interpretations and perceptions regarding the nature of the odor [8, 13]. Human olfactory receptors encompass over 400 types, theoretically capable of distinguishing nearly a trillion different olfactory stimuli [2, 7, 16, 18].

Odor recognition holds significant importance, contributing to a deeper understanding of olfactory processing mechanisms, evolutionary communication, and interdisciplinary research spanning neuroscience, biology, and psychology. Moreover,

2010 *Mathematics Subject Classification.* 68T09, 92B20.

*Key words and phrases.* Odor recognition, molecular fingerprints, molecular graphs, machine learning.

This work was supported by the National Natural Science Foundation of China under Grant 12074151.

*†Corresponding authors.

this recognition serves as a pivotal technology across various commercial sectors, including food, beverages, and perfumes, driving the innovation of new products, enhancing quality control measures, and strengthening market competitiveness. In addition to empirical investigations, researchers have actively pursued the elucidation of the structure-odor relationship through theoretical calculations and simulations, aiming to refine odor recognition methodologies. Despite advancements in understanding the expression patterns of olfactory receptors, their signaling pathways, and the projection profiles of associated neurons, the precise matching relationship between olfactory receptors and odorant molecules remains largely elusive.

In recent years, the rapid advancement of AI technology, fueled by extensive datasets, has profoundly reshaped the landscape across various industries. Leveraging the sophisticated feature learning capabilities of machine learning algorithms, the field of odor recognition using machine learning has emerged as a pivotal research area. For instance, Saini et al. utilized the Random Forest algorithm to infer odorant molecules based on their structures [11]. Similarly, Chacko et al. pioneered data-driven odor prediction, employing algorithms such as Random Forest and XGBoost [3]. However, it is imperative acknowledge the significant imbalance present within the existing odorant molecule datasets. Taking the OlfactionBase dataset as an example [14], the distribution remains significantly imbalanced even among the top seven odor categories. Specifically, the 'floral' category encompasses 210 samples, 'fruity' is represented by 417 samples, 'minty', 'nutty', 'pungent', 'sweet' and 'woody' includes only 87, 130, 50, 187, and 191 samples, respectively. The Goodscents dataset also exhibits significant imbalance. The 'fruity' label appears in 2507 molecules, marking it the most prevalent label. In contrast, 'malty', the least frequent label, is observed in only 19 samples (less than 1% to 'fruity').

The impact of data imbalance on machine learning classification is significant. It can lead classifiers to excessively favor frequent classes during training, thereby affecting their performance on minority classes. In order to investigate the impact of imbalanced distribution in odorant molecule datasets, this paper systematically studies odor recognition using machine learning techniques with molecular fingerprints and graphs. Specifically, we utilize three molecular fingerprints to characterize odorant molecules. Six machine learning models are employed to predict odors based on molecular fingerprints. Additionally, we represent odor molecules using molecular graphs and utilize four distinct graph neuron networks to capture the intricate relationships between nodes and edges for odor predictions from graphs.

## 2. Materials and methods

2.1. **The Leffingwell dataset.** The Leffingwell odor dataset, curated exclusively for olfaction research, consists of 3523 odorant molecules. Each molecule is represented in the SMILES format, and is associated with corresponding odor labels. These odor labels are annotated by olfactory experts and categorized as either singular or composite. The dataset encompasses 113 distinct odor categories, revealing a pronounced imbalance in odor distribution, as illustrated in Figure 1. Due to the exponential decline in distribution, particularly evident after the sixth ranked odor label, which corresponds to significantly fewer odorant molecules, we constructed a subset by selecting the top $N$ odors with the highest frequency of occurrence

from 113 odor labels (referred to as 'TOP-N'). By varying the value of $N$, we can systematically investigate the effect of imbalanced distribution on machine learning recognition of odor molecules. The Top-N subsets, with $N = 2$, 3, 4, 6, and 8, contain 1875, 2237, 2356, 2595, and 2759 molecules, respectively. All subsets were randomly divided into training and test sets using 4:1 ratio.

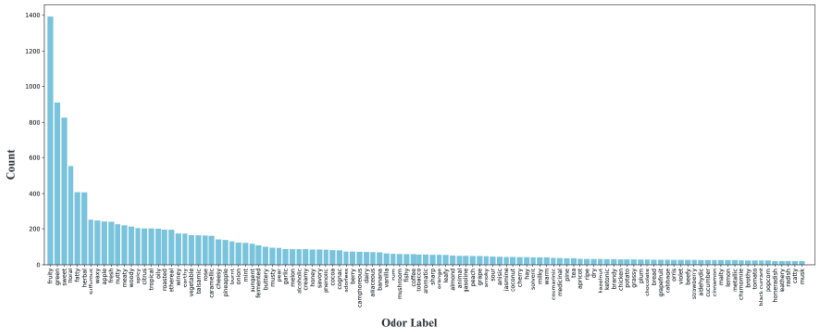


FIGURE 1. Odor distribution in the Leffingwell dataset

2.2. **Molecular fingerprints.** Molecular fingerprinting delineates the distinct identity of a molecule by translating its structural attributes into an array of binary or numerical codes. The MACCS key operates as a binary vector derived from chemical structures, comprising 166 binary bits [5]. Each bit indicates a specific structure or substructure within the molecule. ECFP fingerprint is also a binary vector, encoding various molecular fragments obtained by iteratively sampling all possible central atoms in the molecule and the radii up to the specified maximum radius [10]. While the prediction of molecular properties based on ECFP fingerprints usually depends on the selection of the maximum radius and vector length, we specifically focus on the default settings of ECFP fingerprints for this study. In addition, PaDEL-descriptor, a software for computing chemical molecular descriptors, supports the computation of 1875 molecular features, including 431 three-dimensional features and 1444 two- and one-dimensional features [21]. These descriptors provide a broader spectrum of chemical and physical properties such as atomic number, molecular mass, electron distribution, molecular shape and surface properties. The RDkit package (https://www.rdkit.org) is used to generate both MACCS and ECFP fingerprints. The PaDEL-descriptor is calculated using a Python package (https://github.com/ecrl/padelpy).

2.3. **Molecular graphs.** The graphical representation of molecules plays a crucial role in visualizing the atomic components and their connections, effectively depicting the molecule's topology through nodes and edges. As outlined in Table 1 [9,19], node attributes extend beyond elemental atom types to encompass chemical properties such as atom count, hybridization state, and aromaticity. Meanwhile, edge attributes include chemical bond type, conjugation status, bond stereochemistry, and interactions between atom pairs. Leveraging graph neural network models

with molecular graphs allows us to capture intricate relationships between nodes and edges, providing a more nuanced representation for molecule recognition.

TABLE 1. Description of node and edge features

| **Node features** | **Description** |
|---|---|
| Atom types | One-hot (C, N, O, S, other) |
| Atom degree | One-hot (0, 1, 2, 3, 4, 5, others) |
| Atom hybridization | One-hot (SP, SP2, SP3, SP3D, SP3D2, other) |
| Hydrogen | One-hot (0, 1, 2, 3, 4) |
| Aromatic | One-hot (0/1) |
| Formal charge | Int |
| Radical charge | Int |
| Atom chirality | One-hot (0/1) |
| Chirality type | One-hot (R, S) |
| **Edge features** | **Description** |
| Bond types | One-hot (single, double ,triple, aromatic) |
| Conjugation | One-hot (0/1) |
| Ring | One-hot (0/1) |
| Stereo | One-hot (StereoNone, StereoAny, StereoZ, StereoE) |
| Atom pair | Int ([‘CC’], [‘CN’, ‘NC’], [‘ON’, ‘NO’], [‘CO’, ‘OC’], [‘CS’, ‘SC’], [‘SO’, ‘OS’], [‘NN’], [‘SN’, ‘NS’], [‘others’]) |

2.4. **Machine learning algorithms.** Various machine learning algorithms exhibit different capabilities in learning features. For molecular fingerprints, four machine learning models — Random Forest (RF), Decision Trees (DT), Extra Trees (ET), and Support Vector Machines (SVM) — are utilized for odor recognition based on molecular fingerprints [6]. Additionally, densely-connected neural networks (DNN) and recurrent neural networks (RNN) are employed for the same purpose. For ECFP and PaDEL-descriptor, the DNN architecture comprises two hidden layers, with 256 and 16 neurons, respectively [20, 22]. Since MACCS only consists of 166 bits, the DNN architecture features a single hidden layer with 32 neurons. The RNN architecture incorporates one layer of bidirectional gated recurrent units, enabling the model to capture sequential correlations within the input features effectively.

For graphs, four distinct GNN models — Graph Convolutional Network (GCN) [15, 24], Graph Attention Network (GAT) [24, 25], Relational Graph Convolutional Network (RGCN) [17, 24] and Gated Graph Neural Network (GGNN) [4, 24] — are employed for odor recognition based on molecular graphs. GCN excels in recognizing local connectivity patterns, while GAT introduces an attention mechanism for dynamic weighting of neighboring nodes during feature aggregation. RGCN extends the framework of GCN to accommodate multiple relationship types in a graph, and GGNN incorporates Gated Recursive Units (GRUs) to modulate information flow across the graph. These diverse GNN architectures offer distinct approaches to capturing intricate relationships within molecular graphs.

For the multi-label classification task of odor recognition, we employ metrics such as accuracy, precision, recall, and the F1-score to provide a comprehensive assessment of the model performance in identifying odorant molecules.

## 3. Results and discussion

As depicted in Figure 2, odor molecules undergo featurization using molecular fingerprints and graphs. Various machine learning models are employed to extract features for odor recognition. This comprehensive approach facilitates exploration of diverse feature representations and learning architectures, thereby enhancing the accuracy and robustness. All calculations were performed on an Intel Xeon Platinum 8255C 2.5GHZ PC with a single NVIDIA GeForce RTX 3070 GPU.

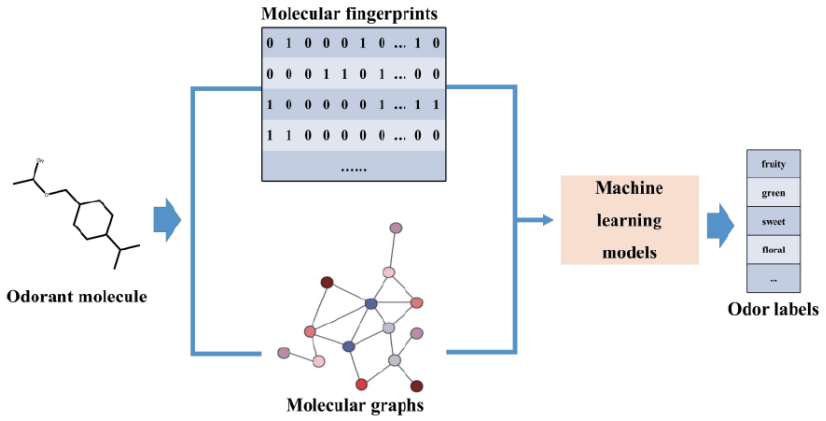


Figure 2. Schematic of odor recognition using molecular fingerprints and graphs

3.1. **Odor recognition using molecular fingerprints.** Figure 3 illustrates the performance comparison of various machine learning algorithms using three different molecular fingerprints. The results demonstrate a noticeable decline in the performance of the multi-label odor recognition model as the value of $N$ increases. This decline suggests that the model's performance is influenced by the incorporation of more frequently occurring odor types. Specifically, as $N$ increases, there is a corresponding escalation in the imbalance of odor types within the dataset. Such an imbalance may lead to better performance of the model on more common odor types but poorer performance on rare odor types, resulting in an overall decrease in performance. Therefore, for multi-label classification tasks, addressing the imbalanced distribution is crucial to avoid detrimental effects on model performance.

Regarding the performance of different machine learning algorithms with molecular fingerprints, the results indicates that MACCS and ECFP fingerprints exhibit similar performance across various machine learning models in odor recognition, with RNN slightly outperforming other models. Notably, ET performs weakest with MACCS, while RF shows the weakest performance with ECFP. SVM and DT fall between these two extremes. Conversely, for PaDEL-descriptor, RNN notably underperforms compared to the other four algorithms. These discrepancies in performance could be attributed to the diverse structural information captured by different fingerprint data and the varying feature learning capabilities of different algorithms. Further investigation into the specific structural characteristics captured

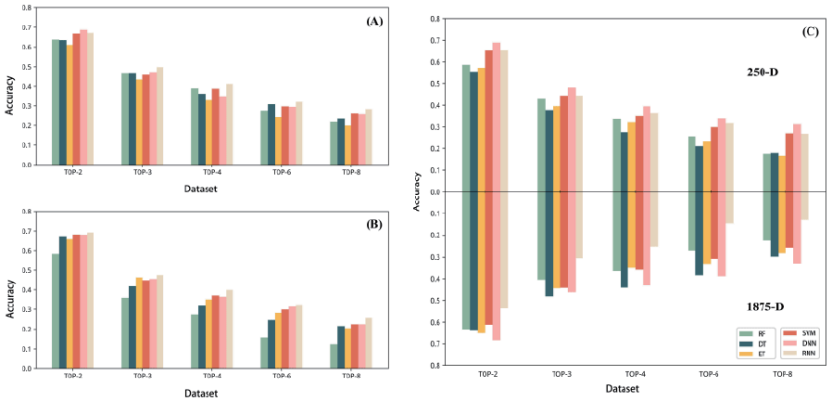


FIGURE 3. Performance comparison of machine learning algorithms using MACCS (A), ECFP (B), and PaDEL (C) fingerprints

by each fingerprint and the inherent strengths and weaknesses of different algorithms could provide deeper insights into these observed performance differences.

The PaDEL-descriptor comprises multiple descriptors, leading to potential redundancy in the encoded information. To address this, we employ Principal Component Analysis to reduce the dimensionality of the original 1875-dimensional PaDEL data [23]. This reduction effectively condenses the PaDEL data to 250 dimensions, with cumulative variance exceeding 95%. As shown in Figure 3C, regarding the performance of six models based on the 1875-D PaDEL data, we observe variations in the dimensionality reduction effects across different models. Both DT and ET experience performance degradation across different $N$ values, with DT showing the highest degradation by approximately 17%. Conversely, RF, SVM, and DNN exhibit fluctuations in performance across different $N$ values, with some improvements and declines, though not significant. In contrast, the RNN model exhibit notable performance improvement. Across tests at five $N$ values, RNN demonstrates an improvement exceeding 10%, particularly for Top-6, where its accuracy increased by approximately 17%. These outcomes underscore the effectiveness of RNN in feature learning following dimensionality reduction of high-dimensional data.

3.2. **Odor recognition using molecular graphs.** Figure 4 illustrates the performance comparison of different GNN models in odor recognition tasks. Observing the trend with the change of $N$ values, it aligns with the testing results of molecular fingerprints shown in Figure 3, indicating that as $N$ increases, the classification accuracy of GNN decreases. Evidently, this phenomenon is closely related to the imbalance in the data distribution. Further investigation into methods for mitigating the effects of data imbalance, such as oversampling, undersampling, or utilizing appropriate evaluation metrics, may be warranted to enhance the robustness of GNN models in odor recognition tasks.

Comparing the performance across various GNN models, it is observed that for $N$ values ranging from 2 to 8, GCN consistently exhibits the weakest performance, followed by GAT. While GGNN slightly surpasses GAT, RGCN notably outperforms all other GNN models. Particularly noteworthy is the significant improvement in

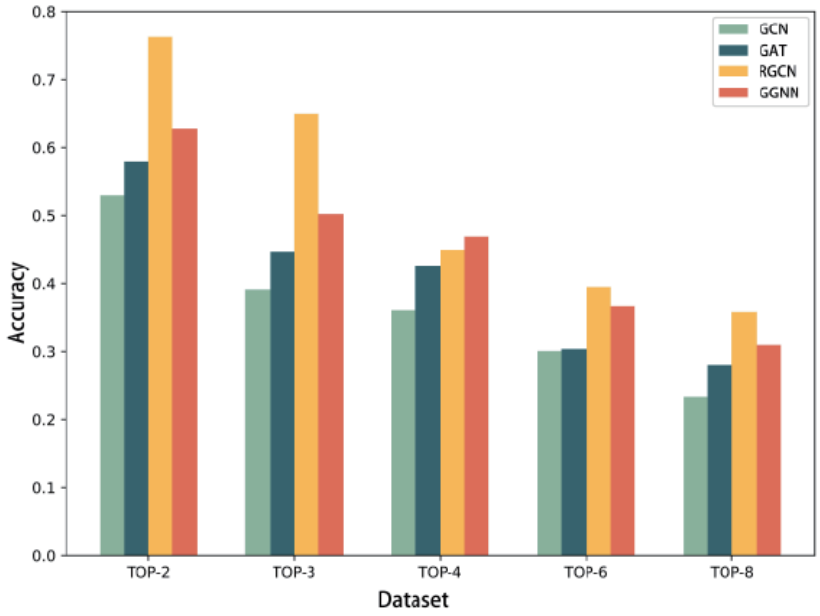


FIGURE 4. Performance comparison of GNN models using molecular graphs

accuracy, especially in tasks such as Top-2 and Top-3, where RGCN achieves up to 15% performance increase over GGNN, the second-ranked model. This comparative analysis underscores the importance of selecting an appropriate GNN model for odor recognition tasks, as different models exhibit varying degrees of effectiveness in capturing the intricate relationships within molecular graphs.

RGCN has demonstrated a significant advantage in accuracy, attributed to its ability to effectively manage multiple types of relationships within graphs. It proficiently distinguishes between various interatomic bonds by assigning distinct weights to each. This capability enables RGCN to provide an accurate portrayal of the intricate interactions within molecules. In contrast, while the GGNN model excels in handling dynamic or sequential data due to its gating mechanism, it struggles to capture the static nature of molecular structures, resulting in lower accuracy compared to RGCN. The GAT model, although not consistently outperforming in all datasets, maintains relatively stable results across different datasets, bolstered by its attention mechanism. This mechanism allows GAT to prioritize and integrate critical structural attributes that influence molecular characteristics, assigning weighted significance to pivotal nodes and edges within the molecular graph. This approach holds promise for enhancing the representation of structural details in molecular graphs, indicating that further exploration and refinement of the attention mechanism within GAT are warranted for complex molecular recognition tasks. Meanwhile, the performance of the GCN model is suboptimal, hindered by its uniform convolutional processing method, which lacks the adaptive tuning capability needed to effectively handle the data.

3.3. **Odor recognition using a multimodal model.** Multimodal models generally outperform single-modal approaches. We combine two single-modal models based on molecular fingerprints and one based on molecular graphs for multimodal fusion, as depicted in Figure 5A. Specifically, we fuse the outputs of ECFP+RNN, PaDEL+DNN and RGCN models and feed these fused features into a single fully connected layer to predict the multi-label odor types. Figure 5B illustrates a comparison between multimodal and single-modal models in odor recognition. Overall, the multimodal model demonstrates superior predictive accuracy compared to single-modal models across odor multi-label classification tasks for $N$ ranging from 2 to 8, highlighting the complementarity of features extracted from different modalities. However, the extent of enhancement is only marginally superior to that achieved by RGCN. Specifically, the improvements in prediction accuracy for Top-2, Top-3, Top-4, Top-6 and Top-8 tasks are 0.012%, 0.007%, 0.002%, 0.047% and 0.064%, respectively. This slight improvement suggests that the features learned by RGCN might already encompass much of the structural information encoded in ECFP and PaDEL molecular fingerprints. Thus, the additional fusion of multimodal features offers only marginal benefits at this stage.

Here, focusing on the Top-8 dataset, we present a comprehensive analysis of the eight odor categories, detailed in Table 2. Across accuracy metrics, all labels surpass 0.73, with two achieving predictive accuracies exceeding 0.9. Similarly, the single-modal models also exhibit high accuracy for individual labels (data not shown), although slightly lower than the multimodal model. For example, the RGCN model demonstrates an accuracy ranges of 0.66 to 0.94. However, despite achieving high accuracies for individual labels, the overall predictive accuracy for multi-label classification, as depicted in Figure 5, remains approximately 0.4 (Top-8). This disparity stems from the stringent evaluation criteria of multi-label classification, which demands precise identification of all associated labels for a sample to be deemed correctly classified. Moreover, the variation in performance across different odor categories sheds light on the model's strengths and weaknesses in distinguishing between specific types of odors. For instance, labels such as 'fruity' and 'sulfurous' exhibit higher precision and recall, indicating the model's capability to accurately identify samples containing these odors. Conversely, labels like 'floral' and 'sweet' demonstrate lower performance, especially in recall, suggesting that the model struggles to accurately recognize these specific odors.

TABLE 2. Multimodal model performance for Top-8 odors

| **Odor** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **TP** | **TN** | **FP** | **FN** |
|---|---|---|---|---|---|---|---|---|
| fruity | 0.797 | 0.723 | 0.759 | 0.759 | 209 | 210 | 53 | 80 |
| green | 0.742 | 0.285 | 0.413 | 0.732 | 52 | 352 | 18 | 130 |
| sweet | 0.729 | 0.172 | 0.278 | 0.746 | 27 | 385 | 10 | 130 |
| floral | 0.591 | 0.121 | 0.202 | 0.813 | 13 | 436 | 9 | 94 |
| fatty | 0.842 | 0.511 | 0.636 | 0.901 | 48 | 449 | 9 | 46 |
| herbal | 0.606 | 0.225 | 0.328 | 0.851 | 20 | 450 | 13 | 69 |
| sulfurous | 0.782 | 0.632 | 0.699 | 0.944 | 36 | 485 | 10 | 21 |
| waxy | 0.738 | 0.449 | 0.559 | 0.911 | 31 | 472 | 11 | 38 |

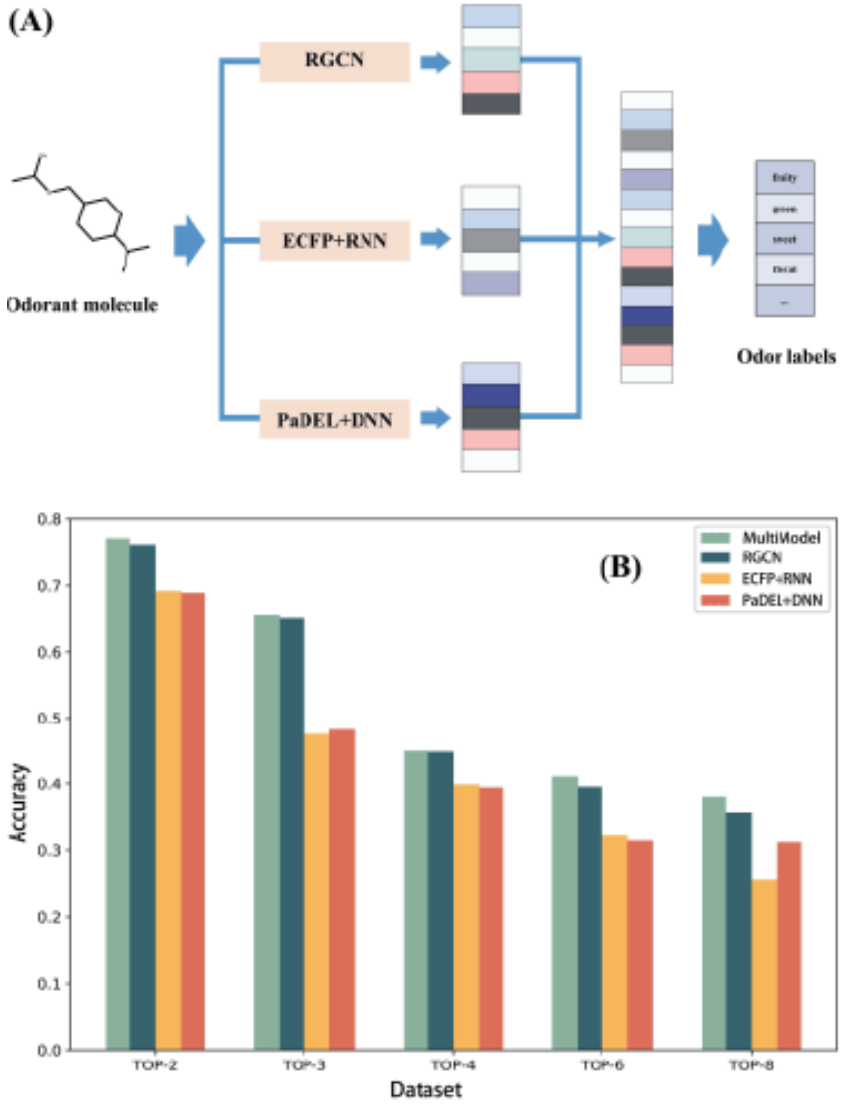


FIGURE 5. (A) Multimodal model for odor recognition. (B) Performance comparison between the multimodal and single-modal models

## 4. CONCLUSIONS

In this study, we conduct odor recognition using machine learning based on molecular fingerprints and graphs, and specifically investigate the impact of imbalanced distribution on machine learning-base odor recognition. For molecular fingerprints, the results suggest that odor recognition using MACCS and ECFP shows similar performance, implying that their encoded structural information plays a comparable role in odor recognition. However, there is some variance in the performance of the PaDEL-descriptor across the tested algorithms compared to the other two

fingerprints. This suggests that the structural information gathered by PaDEL-descriptor contributes slightly differently to odor recognition. For molecular graphs, the results demonstrate that RGCN exhibits a clear performance advantage over the other tested models in odor recognition. This superiority can be attributed to its proficiency in managing multiple relationship types within graphs and its capacity to assign specific weights to different edge types enhancing the granularity of representation in molecular graphs. Compared to single-modal models, the multi-modal model enhances odor recognition although the improvement margin is less pronounced compared to RGCN. This implies that the features learned by RGCN might already capture a significant portion of the structural information encoded in fingerprints. Further analysis of predictions for individual odors reveals that, while the overall accuracy of multi-label odor recognition declines noticeably across different models as the imbalance increases, the predictive accuracy for individual labels remains robust. Therefore, addressing these imbalances through specialized techniques becomes crucial to ensure equitable and comprehensive odor recognition across all categories in multi-label classification tasks.

## Acknowledgement

R. Xia and Y. Wang contributed equally to this work.

## References

[1] R. Achebouche, A. Tromelin, K. Audouze and O. Taboureau, *Application of artificial intelligence to decode the relationships between smell, olfactory receptors and small molecules*, Scientific Reports **12** (2022): 18817.

[2] C. Bushdid, M. O. Magnasco, L. B. Vosshall, and A. Keller *Humans can discriminate more than 1 trillion olfactory stimuli*, Science **343** (2014), 1370–1372.

[3] R. Chacko, D. Jain, M. Patwardhan, A. Puri, S. Karande and B. Rai *Data based predictive models for odor perception*, Scientific Reports **10** (2020): 17136.

[4] Z. Deng, C. Sun, G. Zhong and Y. Mao, *Text classification with attention gated graph neural network*, Cognitive Computation **14** (2022), 1464–1473.

[5] J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *Reoptimization of MDL keys for use in drug discovery*, Journal of Chemical Information and Computer Sciences **42** (2002), 1273–1280.

[6] B. Li, Y. Wang, Z. Yin, L. Xu, L. Xie and X. Xu, *Decision tree-based identification of important molecular fragments for protein-ligand binding*, Chemical Biology & Drug Design **103** (2024): e14427.

[7] B. Malnic, J. Hirono, T. Sato and L. B. Buck *Combinatorial receptor codes for odors*, Cell **96** (1999), 713–723.

[8] C. A. March, S. Ryu, G. Sicard, C. Moon and J. Golebiowski, *Structure odour relationships reviewed in the postgenomic era*, Flavour and Fragrance Journal **30** (2015), 342–361.

[9] E. Poivet, N. Tahirova, Z. Peterlin, L. Xu, D. J. Zou, T. Acree and S. Firestein, *Functional odor classification through a medicinal chemistry approach*, Science Advances **4** (2018), 295-305.

[10] D. Rogers and M. Hahn, *Extended-connectivity fingerprints*, Journal of Chemical Information and Modeling, **50** (2010), 742–754.

[11] K. Saini and V. Ramanathan, *Predicting odor from molecular structure: a multi-label classification approach*, Scientific Reports **12** (2022): 13863.

[12] L. Sela and N. Sobel, *Human olfaction: a constant state of change-blindness*, Experimental Brain Research **205** (2010), 13–29.

[13] A. Sharma, R. Kumar, I. Aier, R. Semwal, P. Tyagi and P. Varadwaj, *Sense of smell: structural, functional, mechanistic advancements and challenges in human olfactory research*, Current Neuropharmacology **17** (2019), 891–911.

[14] A. Sharma, B. K. Saha, R. Kumar and P. K. Varadwaj *OlfactionBase: A repository to explore odors, odorants, olfactory receptors and odorant-receptor interactions*, Nucleic Acids Research **50** (2022), D678–D686.

[15] M. Shi, Y. Tang, X. Zhu and J. Liu, *Multi-label graph convolutional network representation learning*, IEEE Transactions on Big Data **8** (2022), 1169–1181.

[16] S. Takane and J. B. O. Mitchell, *A structure-odor relationship study using EVA descriptors and hierarchical clustering*, Organic & Biomolecular Chemistry **2** (2004), 3250–3255.

[17] T. Thanapalasingam, L. van Berkel, P. Bloem and P. Groth, *Relational graph convolutional networks: a closer look*, PeerJ Compututer Science **8** (2022): e1073.

[18] C. Trimmer, A. Keller, N. R. Murphy, L. L. Snyder, J. R. Willer, M. H. Nagai, N. Katsanis, L. B. Vosshall, H. Matsunami and J. D. Mainland, *Variation across the human olfactory receptor repertoire alters odor perception*, Proceedings of the National Academy of Sciences USA **116** (2019), 9475–9480.

[19] Z. Wu, D. Jiang, J. Wang, C. Y. Hsieh and T. Hou, *Mining toxicity information from large amounts of toxicity data*, Journal of Medicinal Chemistry, **64** (2021), 6924–6936.

[20] L. Xie, L. Xu, R. Kong, S. Chang and X. Xu, *Improvement of prediction performance with conjoint molecular fingerprint in deep learning*, Frontiers in Pharmacology **11** (2020): 606668.

[21] C. W. Yap, *PaDel-descriptor: An open source software to calculate molecular descriptors and fingerprints*, Journal of Computational Chemistry, **32** (2011), 1466–1474.

[22] Z. Yin, W. Song, B. Li, F. Wang, L. Xie and X. Xu, *Neural networks prediction of the protein-ligand binding affinity with circular fingerprints*, Technology and Health Care **31** (2023), 487–495.

[23] C. Yoo, and M. Shahlaei, *The applications of PCA in QSAR studies: A case study on CCR5 antagonists*, Chemical Biology & Drug Design **91** (2018), 137–152.

[24] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li and M. Sun, *Graph neural networks: A review of methods and applications*, AI open **1** (2020), 57–81.

[25] W. Zhou, Z. Xia, P. Dou, T. Su and H. Hu *Double attention based on graph attention network for image multi-label classification*, ACM Transactions on Multimedia Compututing, Communications and Applications **19** (2023): Article No. 18.

*Manuscript received March 25, 2024*
*revised October 24, 2024*

R. Xia
Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, China
*E-mail address*: XiaRJ2000@outlook.com

Y. WANG
School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, China
*E-mail address*: dxwys2@jsut.edu.cn

F. Wang
School of Mathematics and Physics, Jiangsu University of Technology, Changzhou, China
*E-mail address*: phywff@jsut.edu.cn

W. Li
Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, China
*E-mail address*: liwei4sci@163.com

L. Xie
Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, China
*E-mail address*: xieliangxu@jsut.edu.cn

X. Xu
Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, China
*E-mail address*: xuxiaojun@jsut.edu.cn