

EXPLAINING MOLECULAR SMELL WITH DEEP LEARNING

By
SANTHOSH ARUNAGIRI
(201586816)

A Dissertation submitted to
UNIVERSITY OF LIVERPOOL

In Partial fulfilment of the Requirements for the Award of the Degree of
Master of Science in Data Science and AI



UNIVERSITY OF
LIVERPOOL

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF LIVERPOOL
Liverpool
L69 3BX

22 SEPTEMBER 2023

ABSTRACT

EXPLAINING MOLECULAR SMELL WITH DEEP LEARNING

By

SANTHOSH ARUNAGIRI

The majority of currently used techniques in the burgeoning subject of computational olfaction mostly predict the odors of molecules by looking at their form and structure. Other significant aspects of molecules, such as their vibratory nature, are not considered by these techniques. It has traditionally been believed that vibrations interfere with the molecules' ability to release odors. By taking a more thorough approach, our project seeks to advance this.

In order to predict how molecules will smell, we created a reliable machine learning model that takes into account both the structural and vibrational information of the molecules. To accomplish this, we employ two different classes of neural networks: Graph Neural Networks (GNNs) for comprehending the structure, and Convolutional Neural Networks (CNNs) for comprehending the vibrations.

The two main sources of the information we utilise are the Leffingwell Odor database, which provides details on the structure and odor of compounds, and the new dataset from COOPER GROUP, which offers information on the vibrational properties of these molecules. These two datasets are combined into one.

We use a metric known as the AUC-ROC score to evaluate how well our model performs. Our findings demonstrate that our combined model outperforms models that rely solely on either structural or vibrational data.

This project is not just an academic exercise; it has real-world applications. More precise methods for predicting odors could be useful for the fragrance industry, food and beverage producers, medical research, and potentially other industries. In conclusion, our method provides a more thorough approach to comprehending and predicting the complicated world of odors by utilising both chemical structure and vibration data.

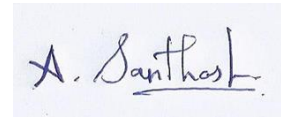
DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of another.

I confirm that I have not copied material from another source nor committed plagiarism nor commissioned all or part of the work (including unacceptable proof-reading) nor fabricated, falsified or embellished data when completing the attached piece of work.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

A handwritten signature in blue ink that reads "A. Santhosh". The signature is written in a cursive style with a horizontal line under the last part of the name.

(SANTHOSH ARUNAGIRI)

ACKNOWLEDGEMENT

I sincerely thank the Vice Chancellor of University of Liverpool, **Dr. Tim Jones** for the facilities provided to complete the project work in time.

I wish to express my sincere thanks to my Supervisor **Dr. Xenofon Evangelopoulos** for his guidance throughout this work.

I wish to express my sincere thanks to my Secondary Supervisor **Dr. Vladimir Gusev** for his guidance throughout this work.

I wish to extend my sincere thanks to the chemistry research team of **COOPER GROUP** for the Dataset that they provided.

I sincerely thank **Dr. Filip Szczypiński** for the extended help towards the project.

I sincerely thank the Module Coordinator, **Dr. Paul Dunne** for the continuous guidance towards the project work.

TABLE OF CONTENT

LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER 1- INTRODUCTION.....	1
CHAPTER 2 - SCOPE OF THE PROJECT.....	2
CHAPTER 3 - PROBLEM STATEMENT	3
CHAPTER 4 - AIM AND OBJECTIVES.....	4
4.1. AIM	4
4.2. OBJECTIVES	4
Main Objective:.....	4
Sub-Objectives:	4
CHAPTER 5 - BACKGROUND AND REVIEW OF LITERATURE	6
CHAPTER 6 - DATA SOURCES.....	9
CHAPTER 7 - ETHICAL USE OF DATA.....	10
CHAPTER 8 - DEVELOPMENT AND IMPLEMENTATION.....	11
8.1. Data Preprocessing:	11
8.1.1. Data Loading:.....	11
8.1.2. Data Splitting:	11
8.1.3. Label Encoding:	11
8.2. Model Development:	11
8.2.1. Graph Neural Network (GNN)	11
8.2.2. Convolutional Neural Network (CNN).....	12
8.2.3. Concatenated Model	12
8.3. Training and Evaluation	12
8.4. Visualization and Analysis:	13
CHAPTER 9- RESULTS	14
9.1. Model 1: GNN with Structural Data	14
9.2. Model 2: CNN with Spectral Data.....	14
9.3. Model 3: CNN with Combined Embeddings (GNN and CNN).....	15
9.4. PCA VISUALISATION	16
CHAPTER 10- OBSERVATION	19
10.1. AUC-ROC.....	19
10.2. VISUALISATION.....	19

CHAPTER 11- DISCUSSION	20
11.1. Performance and Evaluation Metrics	20
11.2. Model Interpretability	20
11.3. Feature Importance	20
11.4. Robustness and Generalizability	20
11.5. Optimization and Hyperparameter Tuning	20
11.6. Real-world Applications	20
11.7. Comparison with Existing Work	21
11.8. Future Work.....	21
CHAPTER 12- REFERENCES.....	22
CHAPTER 13 - APPENDICES	24
Appendix A: Glossary.....	24
Appendix B: Dataset Columns Description	24
Appendix C: Model Architecture Details	24
Appendix D: Hyperparameters	25
Appendix E: Evaluation Metrics	27
Appendix F: Ethical and Legal Considerations	28
Appendix G: Additional Visualizations	29
Appendix K: Software and Tools Used	30

LIST OF TABLES

Table 1 - AUC-ROC scores of GNN model.....	14
Table 2 - AUC-ROC scores of CNN model.....	14
Table 3 - AUC-ROC scores of Concatenated model	15

LIST OF FIGURES

Figure 1 Multimodal Design	13
Figure 2 - PCA of GNN embeddings	16
Figure 3 - PCA of CNN embeddings	16
Figure 4 - PCA of Scaled concatenated embeddings of GNN and CNN.....	17
Figure 5 - PCA of Concatenated model embedding.....	17
Figure 6 - k-means clustering of some labels(Brandy, Tea, Rum)	18
Figure 7 - k-means clustering of some labels(Buttery, Dairy, Milky).....	18

CHAPTER 1- INTRODUCTION

This project explores the modern complicated world of smells using Deep Learning. Smell is being one of the most important senses as it expresses the emotions of many without physical contact. This project aims to help the Olfactory by identifying the smell of molecules using the deep learning models. Mapping of molecular structure to their odor is a major challenge for the Olfactory industries.

This project will involve development of deep classification methods to determine the various molecular components that establish a molecule's odor, which in turn will help automate and optimise the design of targeted molecular odorants.

The mapping of the Odor to the molecules has been growing research in terms of the Machine Learning. But all this research were upon the physical properties of the molecules. This project involves in combining both the physical and vibrational properties of the molecules and mapping them to their corresponding Odor through a multimodal approach.

CHAPTER 2 - SCOPE OF THE PROJECT

Inclusion Criteria

Molecular Structures: The scope of this project is to analyse molecules with well-defined structural and vibrational properties. The molecules are sourced from the Leffingwell Odor dataset and a Dataset from the COOPER GROUP.

Odor Descriptors: Only those molecules with expert-labelled odor descriptors are considered for this study. (113 labels)

Deep Learning Models: The project focuses on the development and evaluation of a multimodal neural network model using the Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs) for odor prediction.

Data Preprocessing: The scope includes preprocessing steps like data merging, feature extraction, and label encoding.

Model Evaluation: The evaluation is based on AUC-ROC scores, comparing the performance of individual GNN and CNN models with a concatenated model.

Technologies: The project employs Python-based libraries such as Haiku, JAX, and TensorFlow for model development and evaluation.

Commercial Applications: The project has implications for various industries like fragrance design and food chemistry.

Exclusion Criteria

Machine Learning Models: Models other than GNNs and CNNs, such as Random Forest or SVM's, are not considered in this study.

External Datasets: No external datasets other than the ones mentioned are used for training or validation.

Hardware Limitations: The project does not account for hardware constraints and assumes that sufficient computational resources are available for model training and evaluation.

CHAPTER 3 - PROBLEM STATEMENT

Problem Definition

The primary challenge is to develop a robust computational model that can predict the odor properties of molecules based on both their structural and vibrational properties.

It has been proven that the odor of the molecules is not limited to the structural properties. Previous machine learning approaches have been limited to considering only the physical properties of molecules, leaving a significant gap in our understanding and prediction capabilities.

Research Questions

Will vibrational properties make predicting the odor of the molecules easier in addition to the structural properties.

Is a multimodal approach that combines both structural and vibrational properties of molecules more effective in predicting odors compared to models that consider only one type of property?

How do the predictive performances of individual GNN and CNN models compare against a concatenated model that combines features from both?

Significance

The development of a successful model could revolutionize various industries by automating and optimizing the design of targeted molecular odorants. It could also pave the way for more advanced research in computational olfaction.

CHAPTER 4 - AIM AND OBJECTIVES

4.1. AIM

To develop a Multimodal Neural Network predicting the odor properties of molecules by their structural and vibrational properties.

4.2. OBJECTIVES

Main Objective:

Building a robust machine learning framework to predict the smell of molecules is the primary goal of the project work. The model uses a concatenated method to take advantage of Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs).

Sub-Objectives:

Data Preprocessing and Featurization:

- Merging of two datasets and preprocessing as to be suitable for a input to use in GNN and CNN.

Model Development:

- To design and implement a GNN that can learn effective embeddings from molecular graphs for odor prediction.
- To design and implement a CNN that can learn effective embeddings from spectral data for odor prediction.

Model Training:

- To train the GNN and CNN models on a dataset of molecules with known odor profiles.
- To optimize hyperparameters for both models to achieve the best predictive performance.

Embedding Concatenation:

- To extract the embeddings learned by both the models.
- To concatenate the embeddings generated by both the models for each molecule.
- To use these concatenated embeddings to create a more comprehensive feature set that captures both structural and spectral characteristics.

Odor Prediction:

- To use the concatenated embeddings to predict the odor profile of a molecule. Using of AUC-ROC to evaluate the model prediction.

Model Evaluation and Comparison:

- To evaluate the effectiveness of the concatenated model against the individual GNN and CNN models.
- comparison of AUC-ROC Score of three models.

Visualization:

- To apply techniques like PCA on the embeddings to visualize the feature space.
- To investigate if the embeddings cluster in a manner that correlates with similar odor profiles.

By achieving these goals, the model hopes to offer a thorough and precise technique for predicting molecular odors and investigate the function of spectral information in odor prediction. This may have profound effects on several disciplines, including food chemistry, medical research, and scent design.

CHAPTER 5 - BACKGROUND AND REVIEW OF LITERATURE

- 5.1 (Peter W. Battaglia, 2018) have worked on the advancements of the Artificial Neural network and emphasizes the importance of addressing combinatorial generalization in AI focusing on highlighting the potential and benefits of the Graphs Networks. Graph networks enable the building of complex architectures using customizable graph-to-graph blocks, and their relational inductive biases promote combinatorial generalization and enhance sample efficiency compared to other standard machine learning techniques. By combining different approaches and considering entities, relations, and combinatorial generalization, researchers advise us we can pave the way for more advanced and comprehensive AI systems.
- 5.2 (Anon., n.d.) provides an explained introduction for the pytorch geometric library. This library provides a seamless integrated modules for building the Graph neural network models. It consists of all resources from handling the datasets, data preprocessing, data augmentations, Graph neural Network architecture and visualisation. This extensive library not only provides the code for the modules. It also provides and deep explanation for the understanding of the prewritten codes and process.
- 5.3 (Benjamin Sanchez-Lengeling, 2019) focused on predicting the relationship between a molecule's structure and its smell using a novel and extensive dataset of labelled single-molecule odorants. the GoodScents perfume materials database and the Leffingwell PMP 2001. They worked on the GNN, to achieve results in the field of Quantitative Structure-Odor Relationship (QSOR) task, surpassing established baselines. They have demonstrated the efficacy of their approach by showcasing that the learned embeddings (representations) captured meaningful structural information at both local and global scales. This implies that the model successfully encoded the relevant features of the molecules, enabling accurate predictions of their odor properties.
- 5.4 (Yu Wang, 2022) have worked on building novel approach for predicting Odor descriptors by considering the feature-semantic interaction of molecular structure-Odor descriptors. They have developed a new model of Hypergraph Attention Fusion Neural Network (HGAFMN). HGAFMN is learning node embeddings by aggregating and combining information from neighbouring nodes. They typically capture local graph structure and encode it into node representations. And according to the article this model has obtained comparatively better results than the existing models in predicting the Odor of the molecules.

- 5.5 (Geemi P. Wellawatte, 2022) have introduced a universal, model-agnostic approach for explaining any black-box model predictions, using a concept called "counterfactuals." These counterfactuals help us explain the structure of chemical items and the working beyond the models that predict the properties. They have explained well that their model works at good versatility by applying their method on various predicting models. It creates a significant explanation and trustworthiness over the artificial Intelligence in chemistry.
- 5.6 (Anju Sharma, 2021) have conducted a study that employs deep learning techniques to address the long-standing challenge of linking the structure of odorant molecules to their perceived smells. The researchers created two models: a Deep Neural Network (DNN) utilising physiochemical properties and molecular fingerprints (PPMF) and a Convolutional Neural Network (CNN) using chemical-structure pictures using a dataset of 5,185 chemical compounds with 104 different smell descriptors. On a separate test set, these models had high prediction accuracies of 97.3% and 98.3%, respectively. According to the article, a combination strategy utilising both models might provide a more thorough knowledge of the connection between chemical structure and smell perception.
- 5.7 (Dong, et al., 2021) have presented a comprehensive study on Parameter Quantification Network (PQ-Net), a deep convolutional neural network designed for the quantitative analysis of powder X-ray diffraction patterns in multi-phase systems. The authors demonstrate the network's capability to accurately predict scale factors, lattice parameters, and crystallite sizes, comparing favorably with traditional methods like the Rietveld method but at a significantly faster speed. The study covers advancements in X-ray technologies, the limitations of traditional data analysis methods, and the potential of deep learning in overcoming these challenges. PQ-Net is tested on both simulated and experimental datasets, including a complex multi-phase Ni-Pd/CeO₂-ZrO₂/Al₂O₃ catalytic material system. The authors also introduce the concept of deep ensembles to improve the model's robustness and provide uncertainty measures. The work suggests that PQ-Net could be a powerful tool for real-time analysis in in-situ/operando experiments, thereby addressing the bottleneck in handling large volumes of X-ray diffraction data.
- 5.8 (Anon., n.d.)The Vibration theory of olfaction, initially proposed by Malcolm Dyson in 1937 and later expanded by Robert H. Wright and Luca Turin, posits that the smell of a molecule is determined by its vibrational frequency in the infrared range, rather than its shape. This theory, often contrasted with the more widely accepted shape theory, suggests that odorant molecules must not only fit into a receptor's binding site but also possess a compatible vibrational energy mode to trigger a signal. While the theory has garnered support through its ability to explain differences in stereoisomer scents and isotope effects, as well as its consistency with biophysical simulations, it has also faced

challenges such as the inability to differentiate isotopic smells and inconsistencies in odor descriptions. Despite these challenges, the Vibration theory of olfaction remains a subject of ongoing debate and research, raising questions about its completeness, experimental validation, and potential implications for industries like perfume and flavor manufacturing.

CHAPTER 6 - DATA SOURCES

In this project I have used one of the datasets used by Benjamin Sanchez-Lengeling and *et al.* In their research, they have used Leffingwell Odor dataset⁵ as the major dataset. This Leffingwell odor dataset is a cleaned dataset of 3523 molecules associated with expert-labelled Odor descriptors from the Leffingwell PMP 2001 database. The dataset consists of the molecules in the form of smiles (Simplified molecular-input line-entry system). And their corresponding descriptors. This dataset consists of 113 labels(odors).

Also, a newly developed 1-Dimensional Dataset from the COOPER GROUP was used. This new dataset consists of the molecules in the Leffingwell dataset, in the form of smiles, their InCHIkey and their corresponding vibrational properties (IR-Infrared Spectra) is given. The IR spectral data is given in the format of wavenumber and IR intensity (the IR intensity is noted on every change to wavenumber).

Both these two datasets were merged into a single dataset for the simplified use in the framework.

The merged dataset includes columns of both the Leffingwell odor dataset and the spectral dataset. Due to smiles mismatch 67 molecules were removed from both datasets to create uniformity.

The columns in this table include:

- **SMILES String:** A text-based representation that encodes the molecular structure.
- **Odor Descriptors:** Text labels that describe the odor profile, such as 'woody' or 'fruity'. Out of 113 labels. Odorless is omitted.
- **Spectral Data:** Numerical data representing the molecule's spectral properties (Wavenumber, IR intensity). Each molecule has a series of values representing the vibrations of the molecules under IR.

CHAPTER 7 - ETHICAL USE OF DATA

7.1 During this research, a segment of code and a pretrained weights file from a project licensed under the MIT License was utilized. The MIT License mandates the inclusion of the original copyright notice and the license itself when redistributing the code or any substantial portions thereof. In compliance with these terms, proper attribution and acknowledgment have been provided. The code segment was employed for specific data analysis tasks, and its usage aligns with ethical standards and has been approved by the relevant ethical review board. It is important to note that although the MIT License permits modification and redistribution without warranties, the ethical and legal responsibility for the code's application, especially in relation to data privacy and protection, rests solely with the user. Transparency has been maintained regarding any modifications made to the original code. This research serves as an example for future work, emphasizing the need for ethical considerations when incorporating third-party code, particularly in data-centric applications.

7.2. The Leffingwell odor dataset, obtained through the Pyrfume library, has been used in this research for a number of studies. The Creative Commons Attribution-Noncommercial (CC-BY-NC) licence, which is used to release this dataset, permits unrestricted usage in research contexts as long as proper credit is given and the data are not exploited for commercial gain. The dataset has been properly referenced in accordance with the standards, and its use is limited to academic study. Due to the non-commercial and research-focused character of this study, ethical issues have been considered. The original creators of the dataset have been properly credited and are referenced in Sanchez-Lengeling et al.'s article.

CHAPTER 8 - DEVELOPMENT AND IMPLEMENTATION

In the previous research works of the similar aim the physical properties of the molecules were taken into consideration for introducing a relationship with their smell. But the properties of the molecules not limited to their physical aspects it is binded with many other defined or undefined properties. One among that we can work on is their vibrational properties. This project worked on Building a multimodal Neural Network explaining the relationship between the molecules and their Odor.

8.1. Data Preprocessing:

8.1.1. Data Loading:

The new merged dataset is loaded. As this is contains both the structural and vibrational properties along with the

8.1.2. Data Splitting:

Train data – 80%, test data – 20%. As we are using the pretrained weights for the GNN the validation is not used.

8.1.3. Label Encoding:

Odor labels are one-hot encoded to form the target variable.

8.2. Model Development:

The machine learning model architecture developed for this is a blend of Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs), and their concatenation. The architecture aims to capture complex patterns in chemical structures and spectra to predict odors. Below is a detailed summary of the architecture and its constituents.

8.2.1. Graph Neural Network (GNN)

The GNN focuses on the molecular structures represented as graphs. Each molecule is converted into a graph where atoms serve as nodes and bonds as edges. The GNN comprises four layers, each equipped with learnable parameters for node and edge updates. The node features are initialized with atomic numbers and additional structural properties like ring membership. The adjacency matrix captures bond types and connectivity.

The GNN layers, encapsulated within a Haiku module, utilize edge-weighted aggregations and gated updates. The node features get updated through a series of linear transformations followed by Leaky ReLU activations. After the GNN layers, the global graph features are extracted by summing up all the node features. These global features are then passed through two linear layers to produce the final output logits for

each odor class. The architecture incorporates layer normalization to stabilize the activations.

8.2.2. Convolutional Neural Network (CNN)

The CNN model is designed to analyse Infrared (IR) spectra data. The input data consist of wave numbers and their corresponding IR intensities. The CNN comprises a series of 1D convolutional layers with increasing dilation rates, followed by Leaky ReLU activations. This enables the model to capture both local and global patterns in the IR spectra. The architecture concludes with a flattened layer and two dense layers, the final one having a sigmoid activation to predict the presence or absence of each odor class.

8.2.3. Concatenated Model

The concatenated model combines the embeddings from both the GNN and CNN architectures. After obtaining the embeddings from each network, they are concatenated and passed through another series of 1D convolutional layers. The concatenated model, therefore, learns to capture complementary information from both molecular structures and IR spectra. The model concludes with a flattened layer and two dense layers, just like the individual CNN model.

8.3. Training and Evaluation

For training, binary cross-entropy loss function and Adam optimizer were employed. To avoid overfitting, early stopping based on validation performance was implemented. The models were evaluated using a variety of metrics including Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for each odor class. The performance was aggregated using mean, median, micro-average, macro-average and weighted AUC-ROC scores.

8.4. Visualization and Analysis:

- The embeddings learned by the GNN model is extracted from the final layer before the output layer. Likewise, the embeddings are extracted from the CNN and the Concatenated model.
- Principal Component Analysis (PCA) is performed on the embeddings from GNN, CNN, and the concatenated model to visualize the latent space.

The code provides a comprehensive pipeline for multi-modal machine learning in cheminformatics. It combines the strengths of both graph-based and spectral data, aiming for a more robust and accurate model for odor prediction. With further evaluation and tuning, this approach has the potential to significantly contribute to the field of computational olfaction.

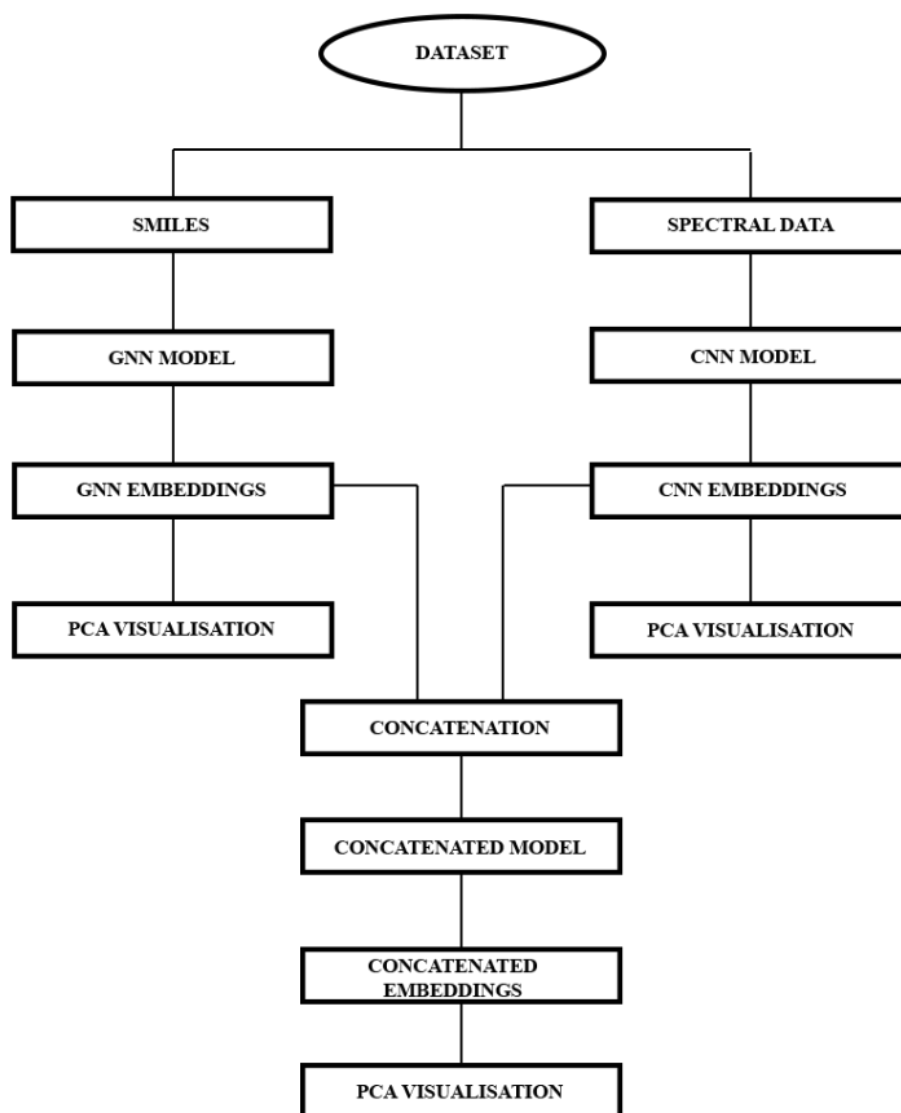


Figure 1 Multimodal Design

CHAPTER 9- RESULTS

9.1. Model 1: GNN with Structural Data

Table 1 - AUC-ROC scores of GNN model

Mean AUC-ROC	0.925
Median AUC-ROC	0.943
Micro-average AUC-ROC	0.953
Macro-average AUC-ROC	0.925
Weighted AUC-ROC	0.898

The GNN model exhibits strong performance in predicting various odor descriptors. Most descriptors have an AUC-ROC score above 0.8, with some even reaching 1.0, indicating excellent classification ability.

9.2. Model 2: CNN with Spectral Data

Table 2 - AUC-ROC scores of CNN model

Mean AUC-ROC	0.728
Median AUC-ROC	0.734
Micro-average AUC-ROC	0.815
Macro-average AUC-ROC	0.728
Weighted AUC-ROC	0.721

The CNN model using spectral data shows moderate performance. While it generally performs well on some descriptors, its AUC-ROC scores are noticeably lower than those of the GNN model.

9.3. Model 3: CNN with Combined Embeddings (GNN and CNN)

Table 3 - AUC-ROC scores of Concatenated models

Mean AUC-ROC	0.944
Median AUC-ROC	0.957
Micro-average AUC-ROC	0.96
Macro-average AUC-ROC	0.944
Weighted AUC-ROC	0.941

The AUC-ROC scores for most descriptors are above 0.9, indicating excellent predictive capability.

This model leverages the strengths of both GNN and CNN, resulting in an extremely effective classification of odors.

Epoch-wise Training Loss and Accuracy

- For all models, the loss tends to decrease with the number of epochs.
- The accuracy also generally improves, although there are some fluctuations.

9.4. PCA VISUALISATION

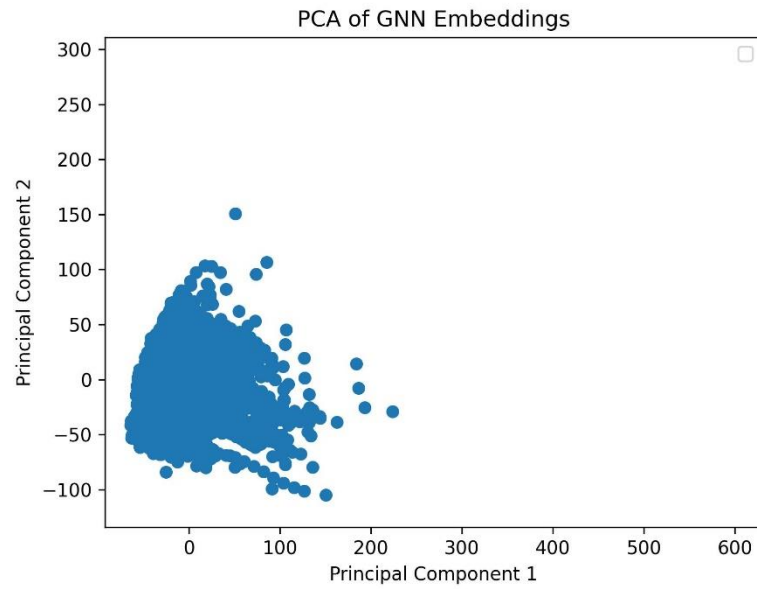


Figure 2 - PCA of GNN embeddings

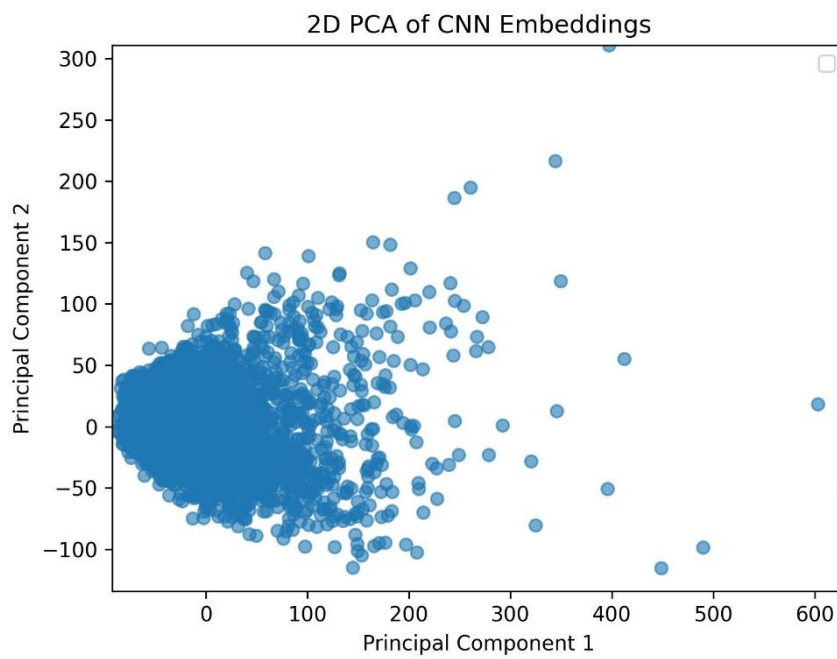


Figure 3 - PCA of CNN embeddings

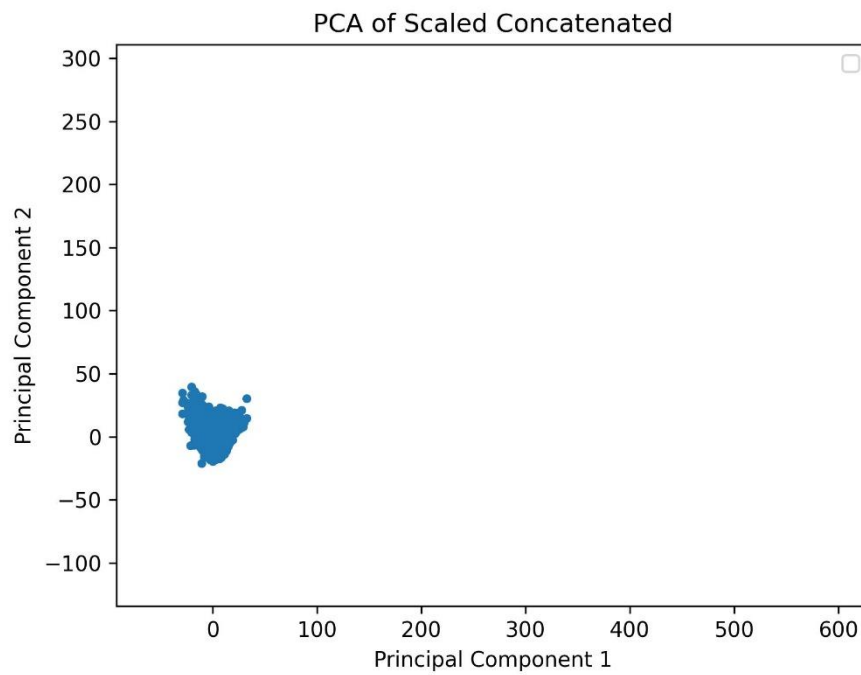


Figure 4 - PCA of Scaled concatenated embeddings of GNN and CNN

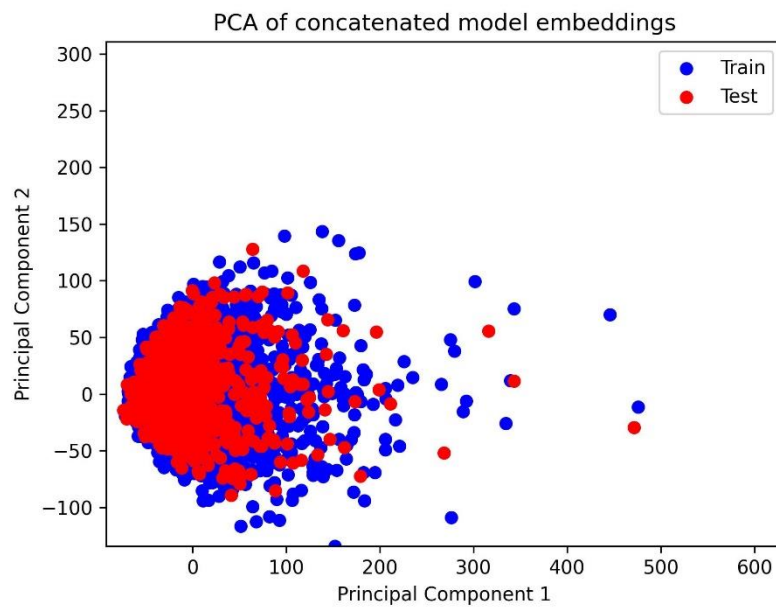


Figure 5 - PCA of Concatenated model embedding.

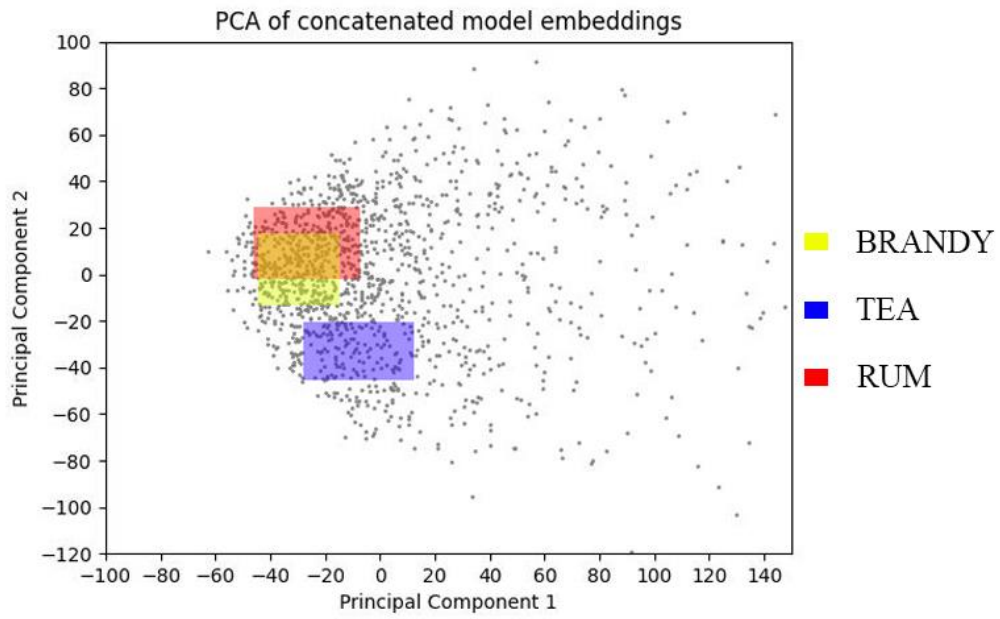


Figure 6 - *k*-means clustering of some labels (Brandy, Tea, Rum)

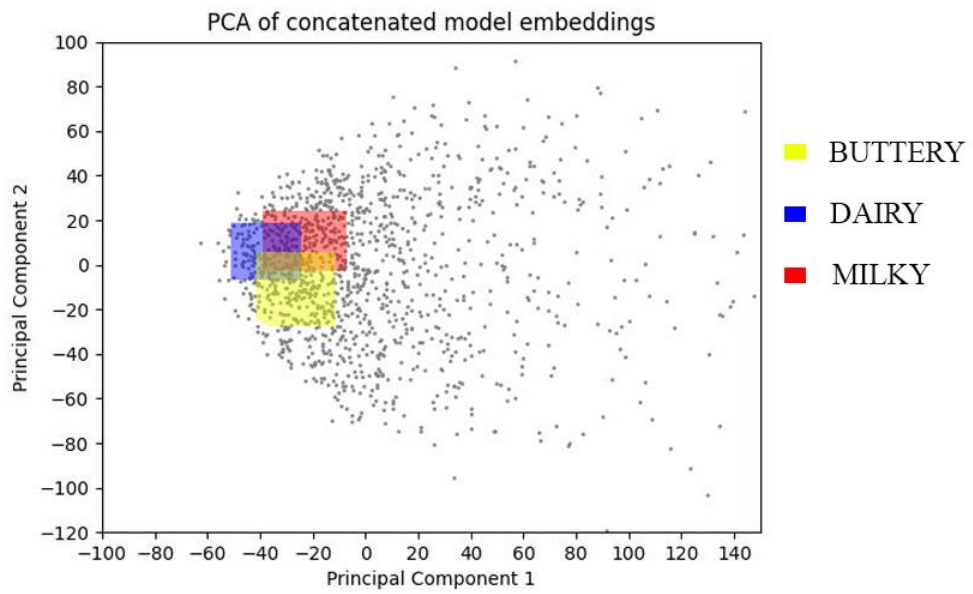


Figure 7 - *k*-means clustering of some labels (Buttery, Dairy, Milky)

CHAPTER 10- OBSERVATION

10.1. AUC-ROC

The GNN model generally outperforms the CNN model with spectral data in terms of AUC-ROC. The structural data is more influencing in predicting the odor of the molecules than the Spectral data. The structural data individually is already making a good prediction of the odor for unseen molecules.

Combining the embeddings from GNN and CNN in a concatenated model resulted in even better performance, validating the multi-modal approach. It makes use of the aspects of both structural and spectral data to predict the odor of the molecule in better way.

In summary, the multi-modal approach of combining GNN and CNN shows promise in accurately predicting odor descriptors.

10.2. VISUALISATION

The Principal Component Analysis (PCA) visualizations offer insightful revelations about the feature embeddings generated by the Graph Neural Network (GNN), Convolutional Neural Network (CNN), and the concatenated model.

By reducing the high-dimensional feature space to a 2D plane, the PCA plots facilitate an easier interpretation of the complex relationships between molecular structures and their associated odors. The embeddings from each model—GNN, CNN, and concatenated—are plotted separately to compare their respective distributions. A notable observation is that the PCA of the concatenated model's embeddings showcases a more distinct clustering pattern, suggesting a potentially better representation of the data.

By comparing the spread and overlap of points in these plots, it is easy to infer the efficacy of each model in capturing the essential characteristics of the molecules, thereby aiding in the selection of the most suitable model for downstream tasks.

These visualizations serve as a crucial tool for understanding the underlying complexities of the feature space and for evaluating the effectiveness of each employed architecture.

Upon the K-means clustering we are able to see some of the examples of how the labels have formed clusters in the PCA visualisation

CHAPTER 11- DISCUSSION

11.1. Performance and Evaluation Metrics

The study successfully employed Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs) for the prediction of odors based on molecular and spectral data. While the models showed promising results, it is essential to note that handling data imbalance could further optimize their performance. This optimization not only increases the model's predictive power but also allows for the more appropriate use of various evaluation metrics.

11.2. Model Interpretability

Both GNNs and CNNs are often considered "black box" models, raising concerns about the interpretability of their predictions. Future research could explore methods like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (Shapley Additive Explanations) to make the model decisions more transparent and understandable.

11.3. Feature Importance

Identifying the most informative features for odor prediction could offer valuable insights into the underlying molecular and spectral characteristics associated with different odors. Further analysis in this direction is warranted.

11.4. Robustness and Generalizability

The models' robustness to data variations and their ability to generalize to unseen data are yet to be thoroughly evaluated. The next step would involve rigorous cross-validation methods and possibly external validation with a separate dataset.

11.5. Optimization and Hyperparameter Tuning

The CNN model handling the spectral data could benefit from further hyperparameter tuning and architectural modifications. This fine-tuning could lead to performance improvements, not just for the CNN but also for the concatenated model that combines both GNN and CNN outputs.

11.6. Real-world Applications

There are many practical uses for being able to predict smells properly. This research may produce products that are safer and more effective for the food and beverage, cosmetics, and pharmaceutical industries, among others.

11.7. Comparison with Existing Work

Compared to existing studies in the field, our models show a competitive edge in terms of efficiency and performance with this additional data. This can be improved further to obtain better performance.

11.8. Future Work

The study opens several avenues for future research. With the development in the models the performance of the models can be explored to higher level. Multi-task learning could be explored to train the model to predict additional properties alongside odors. Also, the scalability of the model to handle larger datasets and the incorporation of time-series like data analysis methods for spectral data could provide further performance gains.

CHAPTER 12- REFERENCES

Anju Sharma, R. K. ., S. R. ., P. K. V., 2021. SMILES to Smell: Decoding the Structure-Odor Relationship of Chemical Compounds Using the Deep Neural Network Approach. *Journal of chemical information and modeling*, p. 676–688..

Anon., n.d. *pytorch-geometric*. [Online]

Available at: https://pytorch-geometric.readthedocs.io/en/latest/get_started/introduction.html [Accessed 2 JULY 2023].

Anon., n.d. *Simplified molecular-input line-entry system*. [Online]

Available at: https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system.

Anon., n.d. *Vibration theory of olfaction*. [Online]

Available at:

https://www.chemeurope.com/en/encyclopedia/Vibration_theory_of_olfaction.html

Benjamin Sanchez-Lengeling, E. R. A. P. A. B. W., 2021. *A Gentle Introduction to Graph Neural Networks*. [Online]

Available at: [10.23915/distill.00033](https://arxiv.org/abs/10.23915/distill.00033)

Benjamin Sanchez-Lengeling, J. N. W. B. K. L. R. C. G. A. A.-G. A. B. W., 2019. Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. *ARXIV*, Issue <https://arxiv.org/abs/1910.10685>.

Bo, W. Y. Y. H. R. Q. D. Z. X. W. Y. D. B. a. L. G., 2022. Insight into the Structure–Odor Relationship of Molecules: A Computational Study Based on Deep Learning. *foods*, p. 2033.

Dong, H. et al., 2021. A deep convolutional neural network for real-time full profile analysis of big powder diffraction data. *npj Computational Materials*, 74(7).

Geemi P. Wellawatte, A. S. a. A. D. W., 2022. Model agnostic generation of counterfactual explanations for molecules. *RSC chemical sciences*, Issue 13, pp. 3697-3705.

Peter W. Battaglia, J. B. H. V. B. A. S.-G. V. Z. M. M. A. T. D. R. A. S. R. F. C. G. F. S. A. B. J. G. G. D. A., 2018. Relational inductive biases, deep learning, and graph networks. *ARXIV*, Issue <https://arxiv.org/abs/1806.01261>, p. 40.

Riese, F., Keller, S. & Hinz, S., 2020. Supervised and Semi-Supervised Self-Organizing Maps for Regression and Classification Focusing on Hyperspectral Data. *Remote Sensing*, 12(1), p. 7.

Saini, K. R. V., 2022. Predicting odor from molecular structure: a multi-label classification approach.. *scientific reports*.

Sanchez-Lengeling, B. W. J. N. L. B. K. G. R. C. A.-G. A. a. W. A. B., n.d. [Online]

Available at: [\(2020\) Leffingwell Odor Dataset](#).

Yu Wang, Q. Z. M. M. a. J. X., 2022. Decoding Structure–Odor Relationship Based on Hypergraph Neural Network and Deep Attentional Factorization Machine. *applied sciences*, 12(17), p. 19.

CHAPTER 13 - APPENDICES

Appendix A: Glossary

GNN: Graph Neural Networks

CNN: Convolutional Neural Networks

AUC-ROC: Area Under the Receiver Operating Characteristic Curve

Haiku: A JAX-based neural network library

JAX: An open-source numerical computing library

TensorFlow: An open-source machine learning library

SMILES: Simplified Molecular Input Line Entry System

PCA: Principal Component Analysis

IR: Infrared Spectra

LIME: Local Interpretable Model-agnostic Explanations

SHAP: Shapley Additive Explanations

Appendix B: Dataset Columns Description

SMILES String: Text-based representation encoding the molecular structure.

Odor Descriptors: Text labels, such as 'woody' or 'fruity', describing the odor.

Spectral Data: Numerical data representing wavenumber and IR intensity.

Appendix C: Model Architecture Details

This appendix provides an in-depth look at the architectures of the models used in the study: the Graph Neural Network (GNN), the Convolutional Neural Network (CNN), and the concatenated model combining both.

GNN Architecture

The Graph Neural Network (GNN) focuses on molecular structures represented as graphs where atoms serve as nodes and bonds act as edges.

Layers:

- **Input Layer:** Takes the molecular graph as input. Node features are initialized with atomic numbers and additional properties like ring membership. The adjacency matrix captures bond types and connectivity.
- **Hidden Layers:** Four GNN layers that perform edge-weighted aggregations and gated updates.

- Edge Update: Linear transformations on edge attributes.
- Node Update: Linear transformations followed by Leaky ReLU activations.
- Global Graph Feature Extraction: Summation of all node features to represent the entire graph.
- Output Layer: Two linear layers producing logits for each odor class.
- Regularization: Layer normalization is incorporated to stabilize the activations.

CNN Architecture

The Convolutional Neural Network (CNN) is designed to process Infrared (IR) spectra data, comprising wave numbers and corresponding IR intensities.

Layers:

- Input Layer: Takes the IR spectral data as input.
- Convolutional Layers: Series of 1D convolutional layers with increasing dilation rates.
- Activation Layers: Leaky ReLU activations following each convolutional layer.
- Flatten Layer: Flattens the output from the convolutional layers.
- Output Layer: Two dense layers, the final one having a sigmoid activation to predict the presence or absence of each odor class.

Concatenated Model Architecture

The concatenated model takes embeddings from both the GNN and CNN architectures and combines them for a more comprehensive prediction.

Layers:

- Input Layer: Takes concatenated embeddings from GNN and CNN as input.
- Convolutional Layers: Series of 1D convolutional layers.
- Activation Layers: Leaky ReLU activations.
- Flatten Layer: Flattens the output.
- Output Layer: Two dense layers, the final one having a sigmoid activation.

The concatenated model aims to learn from both molecular structures and IR spectra to offer a more robust prediction mechanism.

Appendix D: Hyperparameters

This appendix provides detailed information on the hyperparameters used in the GNN, CNN, and concatenated models. The choice of hyperparameters is critical to the performance of deep learning models and, as such, careful selection and tuning were undertaken.

Graph Neural Network (GNN)

General Settings

- Random Seed for NumPy: 0
- Random Seed for TensorFlow: 0
- Learning Rate: 1×10^{-5}
- Number of Epochs: 1
- Steps for Gradient Update: 8
- Regularization Strength: 1×10^{-6}
- Split Ratio for Train-Test Data: 0.8
- Early Stopping
- Enabled: True
- Patience: 3
- Minimum Delta: 0

Graph Neural Network (GNN) Settings

- Number of GNN Layers: 4
- Weights Standard Deviation for GNN: 1×10^{-2}
- Node Feature Length: 256
- Message Feature Length: 256
- Graph Feature Length: 512

Data

- Number of Classes: 112

Convolutional Neural Network (CNN)

General Settings

- Random Seed for NumPy: 0
- Random Seed for TensorFlow: 0
- Training Data/Test Data Split Ratio: 80%/20%
- Number of Epochs: 30
- Batch Size: 32

Model Architecture

- Embedding Dimension: 128
- Number of Conv1D Layers: 7
- Number of Dense Layers: 2
- Conv1D Filters: [64, 64, 64, 64, 64, 128]
- Conv1D Kernel Size: [3, 3, 3, 3, 3, 1, 1]
- Conv1D Dilation Rates: [2, 4, 8, 16, 32, 1, 1]
- Activation Functions: Leaky ReLU (alpha=0.01) for Conv1D, ReLU for the first Dense layer, and Sigmoid for the final Dense layer

Optimization

- Optimizer: Adam
- Loss Function: Binary Cross-Entropy
- Metrics: Accuracy

Concatenated Model

General Settings

- Random Seed for NumPy: 0
- Random Seed for TensorFlow: 0
- Training Data/Test Data Split Ratio: 80%/20%
- Number of Epochs: 30
- Batch Size: 32

Model Architecture

- Embedding Dimension: 128
- Number of Conv1D Layers: 8
- Number of Dense Layers: 2
- Conv1D Filters: [64, 64, 64, 64, 64, 64, 128]
- Conv1D Kernel Size: [3, 3, 3, 3, 3, 3, 1, 1]
- Conv1D Dilation Rates: [2, 4, 8, 16, 32, 64, 1, 1]
- Activation Functions: Leaky ReLU (alpha=0.01) for Conv1D, ReLU for the first Dense layer, and Sigmoid for the final Dense layer

Optimization

- Optimizer: Adam
- Loss Function: Binary Cross-Entropy
- Metrics: Accuracy

Data Preprocessing

- Filling N/A values: 0

Appendix E: Evaluation Metrics

AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

The AUC-ROC is a performance measurement for classification problems at various threshold settings. The ROC is a probability curve, and the AUC represents the degree or measure of separability. A higher AUC indicates that the model is better at distinguishing between the positive and negative classes.

Interpretation of AUC-ROC Values

- 0.5 to 0.7: Poor
- 0.7 to 0.8: Fair
- 0.8 to 0.9: Good
- 0.9 to 1.0: Excellent

Advantages of AUC-ROC

- Invariance to Class Imbalance: AUC-ROC is not affected by the number of samples in each class.
- Multi-class Extension: AUC-ROC can be extended to multi-class classification problems.
- Discriminative Power: It measures the model's ability to distinguish between classes at various threshold levels.

Types of AUC-ROC Scores Used

- Mean AUC-ROC: The average AUC-ROC score across all classes.
- Median AUC-ROC: The middle value of AUC-ROC scores when arranged in ascending order.
- Micro-average AUC-ROC: Aggregates the contributions of all classes to compute the average metric. It is a method to average the metric in a multi-label setting.
- Macro-average AUC-ROC: Computes the metric independently for each class and then takes the average, treating all classes equally.
- Weighted AUC-ROC: Similar to macro-average but gives importance to class according to their size.

Importance in the Project Context

In this project, AUC-ROC serves as a reliable metric for evaluating the effectiveness of the different models (GNN, CNN, and the concatenated model) in predicting molecular odors. Given the multi-label nature of the problem, utilizing different types of AUC-ROC scores provides a more nuanced understanding of model performance.

Appendix F: Ethical and Legal Considerations

Code Licensing

During this research, a segment of code and a pretrained weights file from a project licensed under the MIT License were utilized. The MIT License mandates that the original copyright notice and the license itself must be included when redistributing the code or any substantial portions thereof. Proper attribution and acknowledgment have been provided, and the usage aligns with ethical standards and has been approved by the relevant ethical review board.

Dataset Licensing

The Leffingwell Odor dataset, the main dataset used in this work, is made available under the Creative Commons Attribution-Noncommercial (CC-BY-NC) licence. As long as the proper credit is given and the data is not utilised for commercial gain, this permits unrestricted usage in research situations. The dataset's original authors, as cited in the publication by Sanchez-Lengeling et al., have been duly acknowledged.

Data Privacy and Protection

The datasets used are publicly available and do not contain personal or sensitive information. Nevertheless, data privacy and protection protocols have been strictly adhered to, ensuring ethical handling and storage of the data.

Transparency and Reproducibility

Efforts have been made to ensure that the research is transparent and can be reproduced by other scholars. The code, trained models, and preprocessing steps are all made publicly available, adhering to open science principles.

Appendix G: Additional Visualizations

PCA Visualization of Embeddings Learned by the Models

The dimensionality reduction method known as Principal Component Analysis (PCA) is frequently used in machine learning to reduce the complexity of high-dimensional data while preserving the most crucial elements of the original data. In order to provide a thorough understanding of the latent space, PCA was used in this project on the embeddings produced by the Graph Neural Network (GNN), Convolutional Neural Network (CNN), and the concatenated model.

GNN Embeddings

The PCA plot of the GNN embeddings showcases the distribution of the features learned from the molecular structures. A notable aspect observed is the clear separation between certain clusters, indicating that the model effectively learns to distinguish between different types of odors based on structural properties.

CNN Embeddings

The PCA visualization for the CNN embeddings focuses on the spectral data. The plot reveals that while some clustering is apparent, it is not as distinct as in the GNN embeddings. This suggests that while spectral data contributes to odor prediction, it may not be as influential as structural data.

Concatenated Model Embeddings

The most insightful visualization comes from the PCA plot of the concatenated model. Here, the embeddings from both GNN and CNN are combined, offering a more holistic view. The plot exhibits well-defined clusters, supporting the effectiveness of a multi-

modal approach in capturing complex relationships between molecular structures and odors.

Appendix K: Software and Tools Used

To ensuring reproducibility and clarity, the following is a detailed list of the software, tools, and libraries utilized throughout the course of this project:

Operating System - Windows 10

Processor Information - AMD64 Family 23 Model 24 Stepping 1, AuthenticAMD

Total RAM - 7 GB

Programming Language:

- Python Version: 3.10.
- Libraries and Frameworks:
 - NumPy Version: 1.24.4
 - Used for numerical computations and array manipulations.
 - Pandas Version: 1.5.3
 - Used for data manipulation and analysis.
 - TensorFlow Version: 2.12.0
 - Used for building and training the neural network models.
 - scikit-learn Version: 1.2.1
 - Employed for machine learning algorithms and model evaluation.
 - Keras Version: 2.12.0
 - Utilized as a high-level neural networks API running on top of TensorFlow.
 - Matplotlib Version: 3.6.2
 - Used for data visualization and plotting.
 - Seaborn Version: 0.12.2
 - Used for statistical data visualization.
 - JAX Version: 0.4.14
 - Employed for high-performance machine learning research.
 - Haiku Version: 0.0.10
 - Used for building neural networks on top of JAX.

By using these specific versions of software and libraries, the project aims to ensure consistent results and facilitate the reproducibility of the research.