WHY DOES THAT MOLECULE SMELL?

A PREPRINT

Aditi Seshadri*
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
aseshad4@u.rochester.edu

Heta A. Gandhi*
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
hgandhi@ur.rochester.edu

Geemi. P. Wellawatte
Department of Chemistry
University of Rochester
Rochester, NY, 14627
gwellawa@ur.rochester.edu

Andrew D. White†
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
andrew.white@rochester.edu

ABSTRACT

Learning structure-scent relationships is a complex challenge due to both the large chemical space of odorous molecules and the molecular biology of a smell. We empirically fit structure-scent relationships by training an accurate graph neural network and then explaining its predictions. We use counterfactuals and descriptor attribution to generate explanations for the 112 scents in the Leffingwell Odor Dataset (Sanchez-Lengeling et al., 2019). Then we use natural language processing to summarize the quantitative explanations into text. The complete process goes from data to a natural language explanation with the aim of determining structure-scent relationships.

1 Introduction

Humans are thought to be able to differentiate over 400,000 odorant molecules.[1] Odorant molecules are volatile compounds with a low molecular weight.[2–4] When an odorant binds to an olfactory receptor protein, a signal is produced and sent to the olfactory bulb, which is then sent to different areas of the brain and results in the perception of scent. The signal produced is different for each odorant molecule, allowing for humans to distinguish between various scents.[4–7] Although many theories attempt to explain how humans can differentiate between these molecules, the specific structural properties of odorants that determine their scent remains unknown.[4,8] For example, we currently do not know what structural properties make some molecules, such as vanillin, have a vanilla scent, while others, such as hexanoic acid, smell "goat-like".[2,9–11] Additionally, in some cases the concentration of molecules can alter the perceived scent, such as in the case of indole, which is floral at low concentrations and putrid at high concentrations.[2,12]

There are certain relationships between molecular structure and scent that are well-documented.[2,8,13–15] For instance, it is known that esters tend to have a fruity and floral scent while thiols often are described as having a rotten or alliaceous scent.[8,14] Even though both tert-amyl acetate and n-amyl acetate have the ester functional group and are very similar in structure, tert-amyl acetate has a camphoraceous scent while n-amyl acetate has a fruity scent.[14] This complex relation between structure and scent is not unique to esters. Both vanillin and isovanillin are aldehydes and are noted for their different scents. Vanillin has a strong vanilla scent, while isovanillin has a weak phenolic scent. Structurally, the only differences between the two molecules are the positions of the ring carbons to which the hydroxyl (OH) and methoxy (OCH_3) groups are bonded.[2,16] In addition, molecules that have drastically different structures can have similar scents. For instance, both muscone and musk ketone have a musky scent, despite musk ketone having an aromatic ring and muscone not having one.[8,14] Therefore, the complex relationship between structure and scent makes it a good candidate for machine learning.

*Denotes equal contribution
†Corresponding Author

Predicting the scent of molecules based on their structure can be considered as a supervised learning problem. Supervised learning involves first training the model on a set of labeled data, which consists of the input features ($\vec{x}$) and labels (y), and then predicting the $\hat{y}$ labels for different $\vec{x}$ inputs. Here the inputs are molecular graphs and the outputs are labels indicating each molecule's predicted scent. This can be treated as multi-class, multi-label, or a regression problem depending on the data and approach.[17,18]

One regression-based approach to scent prediction was the DREAM Olfaction Prediction Challenge created by Keller et al., which included having competitors develop models that predicted the intensity of different scents for a given molecule.[19] However, molecular scent prediction can also be characterized as a multi-label classification problem, where rather than predicting the intensity of every scent for every molecule, the scent classes that the molecule belongs to are predicted. For example, the molecule lyral could be described as having more than one scent – specifically a 'fresh,' 'floral,' 'muguet,' and 'sweet' scent.[11] The traditional approach to this type of classification problem would be to perform a regression analysis that takes in various features of the molecule, such as the number of hydroxyl (OH) groups or its molecular weight. For instance, Chacko et al.[2] used RDkit descriptors as the input to their model to classify whether molecules had 'sweet' or 'musky' scents. However, since the relationship between the structural properties of molecules and their scent is complex and not well understood, deep learning (DL) methods may be a better approach to tackling this problem as they can capture non-linear relationships.

There have been many efforts to develop DL models that predict molecular scent.[20,21] For example, Zhang et al.[22] used a Multilayer Perceptron (MLP) model to predict the scent of molecules based on their surface charge density profile. Shang et al.[23] compared the performance of support vector machines (SVM), random forest (RF) and extreme learning machine (ELM), for scent prediction using physicochemical properties of compounds. Nozaki and Nakamoto[24] used an MLP for scent prediction using mass spectra information of compounds. The authors later showed that this model did not perform well on validation data and proposed a new predictive model that uses a hierarchical clustering approach along with the mass spectra information as input to the MLP.[25] Sharma et al.[4] compared the scent prediction from an MLP, that took the physiochemical properties and molecular fingerprints as features, to a Convolutional Neural Network (CNN) which took the image of a molecular structure as input. This study found that an ensemble model that combined the predictions from the CNN and MLP provided the best results.[4] There are more direct ways to featurize molecules than using images for CNNs, such as using molecular graphs.[26] Sanchez-Lengeling et al.[11] treated predicting scents as a multi-label classification problem and applied Graph Neural Networks (GNNs) to classify the scents of small molecules. They also created a labeled dataset of odorants. Lee et al.[27] used the same GNN to create a scent map, which can be used in a manner similar to color maps, where molecules with similar scents appear closer together. They also demonstrated that their GNN was generalizable to a new dataset as well as to other scent-prediction tasks.[27] DeepSniffer is a framework, developed by Liu et al.[28], to classify scents in essential oils using an MLP based on k-shot meta-learning. Most of the described studies primarily focus on evaluating the empirical accuracy of different models and developing models with strong predictive capabilities for molecular scent prediction.[4] However, good predictive accuracy does not necessarily provide an understanding of why molecules have particular scents.

In this paper, the goal is to develop a scent prediction model and use explainable artificial intelligence (XAI) methods to understand why molecules have a certain smell in terms of molecular features. XAI is a field that is emerging to provide insight into how DL models work and why certain predictions are made. As DL is employed more and more for QSAR modeling, explainability has become important. Explainability is important when using DL models, as it can not only help with finding limitations in models and increase trust in model predictions, but can also aid in scientific developments.[29–31] Using Miller's definition, an explanation provides additional context or insight for why a model prediction is made.[32] A local explanation can be defined as one that explains the model prediction for a specific case, whereas global explanations provide a broader description of model behavior.[30]

Feature attribution is an XAI method that provides quantitative information on the influence different input features have on the model prediction. Some recent scent-prediction studies include feature importance rankings.[2,19,33] However, one drawback to using feature attribution methods in chemical research is that many molecular descriptors are not very intuitive.[34] For instance, it is not very intuitive as how one would experimentally evaluate results reported by Saini and Ramanathan[33] that the 'centered moreau-broto autocorrelation of lag 5 weighted by van der Waals volume' is important in predicting the scent of a molecule. This feature was computed using Mordred, a Python library for molecular descriptor calculation.[35]

According to Grisoni et al.[36], classical molecular descriptors encode a precise 'structural/chemical feature (or a set of features of different complexity) into one, single number.' Chemical fingerprints, on the other hand, provide information on the molecular structure in a sequence of bits.[36] For example, MACCS fingerprints are one type of substructure-based fingerprint that provide information about a molecule based on the presence or absence of different substructures, such as a ring or carbon-nitrogen single bond.[37–39] Extended connectivity fingerprints (ECFP) are a type of circular fingerprints that provide information on the specific substructures in a molecule.[38,40] An advantage to

using chemical fingerprints rather than other chemical descriptors is that they are often easier to interpret, as the value of a MACCS or ECFP fingerprints can be often determined simply by looking at the molecular structure.[41]

One type of explanation approach is to utilize a surrogate model that is more interpretable, such as a linear model, and fit it to the DL model.[31,34] Surrogate models are used in the Locally Interpretable Model Explanations (LIME) algorithm. LIME generates a sample space around a specific data point and uses the DL model to obtain predictions for each element in the space. A surrogate model is then trained on this space and used to obtain local explanations.[42,43] This study used a modified version of the descriptor explanation method developed by Gandhi and White[41] for identifying influential MACCS and ECFP fingerprints with LIME.

Counterfactuals are another method with which to interpret the predictions produced by a machine learning model.[44] A counterfactual explanation provides information as to the minimal number of changes that must be made to the input features for a specific case to alter the output prediction of the model.[34,45,46] Some recent uses of counterfactuals include crop growth prediction, and medical diagnosis.[47,48] A molecular counterfactual is the molecule that is most similar to the input but has a different predicted label. In the case of scent prediction, if a given molecule has a fruity scent, its corresponding counterfactual would be the most structurally similar molecule that is predicted to not have a fruity scent. In this study, we use the Python package exmol to generate molecular counterfactuals.[46]

In this article, we develop a model to predict scent of molecules and propose the relationship between molecular structure and scent using explainable artificial intelligence (XAI) methods. We build a GNN for our predictive model and show that it has comparable performance to existing scent prediction models in the literature. To explain the predictions and obtain structure-scent relationships, we use counterfactual analysis[46] and descriptor explanations.[41] The counterfactual approach provides insight into which structural groups on a specific molecule may influence its corresponding scent. The descriptor explanations allow for a broader exploration of the structure-scent relationship by identifying influential structures for different scent classes in general. The descriptor explanations are further used to obtain text explanations for the structure-scent relationship.

1.1 Related work

Recent efforts have been made to interpret the results of models used to predict scent, although few provide a thorough analysis of these results. Mayhew et al.[49] note that molecules that are volatile and hydrophobic are generally odorous, although they do not predict particular odors. Using their RF model, Saini and Ramanathan[33] identify two spatial autocorrelation features as being the most important for scent prediction in general. Keller et al.[19] report the top five molecular features used by a random-forest model to classify whether a molecule had 'fruit,' 'burnt,' and 'bakery' scents. Kowalewski et al.[50] found that an aggregation of SVM models performed better than RF models in predicting scent, but also use RF models to determine the importance of different chemical features in scent prediction. Similarly, Chacko et al.[2] report rankings of the different features used. They found that the presence of the ether functional group and molar refractivity were important in predicting whether a molecule had a sweet scent. Another study by Licon et al.[51] used a data mining technique to generate descriptive structure-scent relationships for molecules.[51] This study found that the scent of a molecule can be related to more than one class of physicochemical properties as opposed to a single property. Gupta et al.[52] use a transformer-CNN combined with a recurrent neural network to predict molecular scent and provide per atom attribution using integrated gradients for the predicted odor.

2 Methods

2.1 Dataset

The Leffingwell Odor Dataset[11] was used to train and test the Graph Neural Network (GNN) model and the logistic regression model. The logistic regression model was used as a baseline from which the performance of the GNN model could be compared against. Sanchez-Lengeling et al.[11] created the Leffingwell Odor Dataset using scent descriptions given in the Leffingwell PMP 2001 database. The Leffingwell Odor Dataset consists of 3523 molecules and covers 113 different scent classes. 70% of the data was used for training the model, 10% was used for validation and 20% was used for the test set. The splits used for the training and test sets were based on that used by Sanchez-Lengeling et al.[11] for their work with molecular scent prediction. Sanchez-Lengeling et al.[11] created these splits using IterativeStratification in scikit-multilearn.[11] The training set used by Sanchez-Lengeling et al.[11] was further split into a training and validation set for this study using IterativeStratification in scikit-multilearn.[53] A bar chart depicting the ten most common scents in the dataset and the number of molecules belonging to each scent class can be found below (Figure 1).

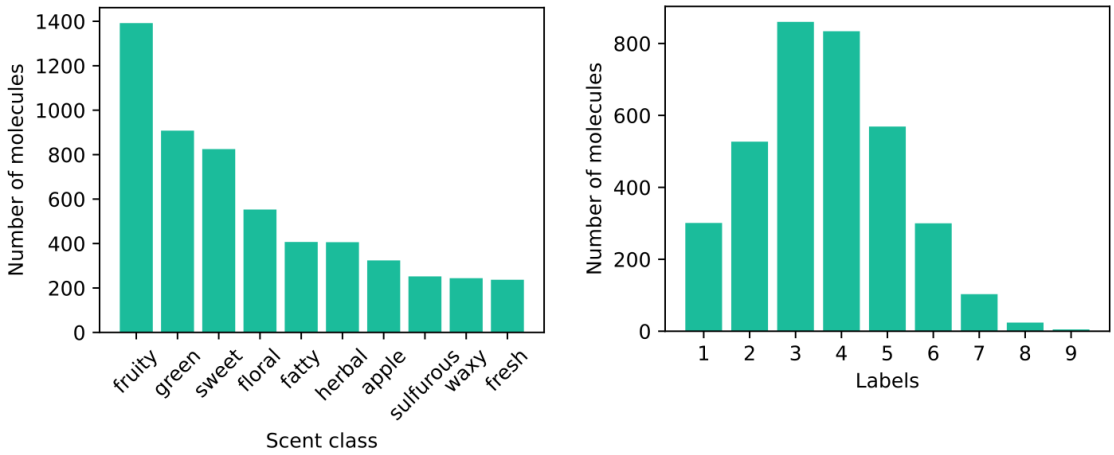

Figure 1: Overview of the Leffingwell Odor Dataset. A bar chart depicting the most common scent classes and distribution of number of scent labels per molecule is shown.

2.2 Logistic Regression and GNN Models

The output of both the logistic regression and GNN models is a 112 length vector, with each element corresponding to a specific scent class. A model prediction consisting of the zero vector corresponds to the odorless class.

First, a logistic regression model was created to serve as a baseline. The descriptors used as input to the logistic regression model were computed using Mordred.[35] The baseline model was trained for 1000 epochs, with a learning rate of 0.04 and used stochastic gradient descent.

Then the GNN model was created. Both the logistic regression and GNN models used cross-entropy loss. The final GNN model consisted of four GNN layers followed by two Linear layers. The model was trained for 138 epochs, with a learning rate of 10^{-5} and used the Adam optimizer. An image depicting the architecture of the GNN model created can be found in Figure 2.

The molecular graphs were generated from the input SMILES string for each molecule using one-hot encoding for node feature vectors and the adjacency matrix. The adjacency matrix only contained information on whether a bond was present between two atoms (nodes) or not. The molecular graphs were then passed through the GNN layers.

The Battaglia equations were used to define the GNN layers.[54] Only the node feature vector and graph feature vector were updated after each layer. It was found that using a leaky ReLU rather than ReLU resulted in a slightly higher mean AUROC (Area Under Receiver Operating Characteristic Curve) and slightly lower cross-entropy loss. The node feature vector and graph feature vector were updated using Equations 1 and 2, respectively.

$$\vec{v_i'} = \sigma(\mathbf{W_v}\bar{e_i'}) + \vec{v_i} \tag{1}$$

$$\vec{u'} = \sigma(\mathbf{W_u}\bar{v'}) + \vec{u} \tag{2}$$

Where $\vec{v_i'}$ represents the updated nodes, $\vec{v_i}$ is the original node feature vector, $\bar{e_i'}$ is the aggregated edge updates, $\mathbf{W_v}$ are the node weight vectors, and σ is a leaky ReLU. The node feature vectors were normalized after each update. The updated graph feature vector is represented by $\vec{u'}$, $\vec{u}$ is the original graph feature vector, $\bar{v'}$ is the sum of the node updates, $\mathbf{W_u}$ are the graph weight vectors, and σ is a leaky ReLU.

The graph feature vector outputted from the last GNN layer was then passed through two Linear layers, created using the Haiku Linear module.[55] The label vector with predictions for each scent class was taken to be the output of the final Linear layer. This output is a 112 length vector, where each entry corresponds to a different scent class. For each scent class, the threshold that maximized the F1 score for that class on the training and validation sets combined was selected. These thresholds were used when determining the predicted scent labels of the input molecules. Weights & Biases was used for tracking the performance of different models when varying the values of the hyperparameters.[56]

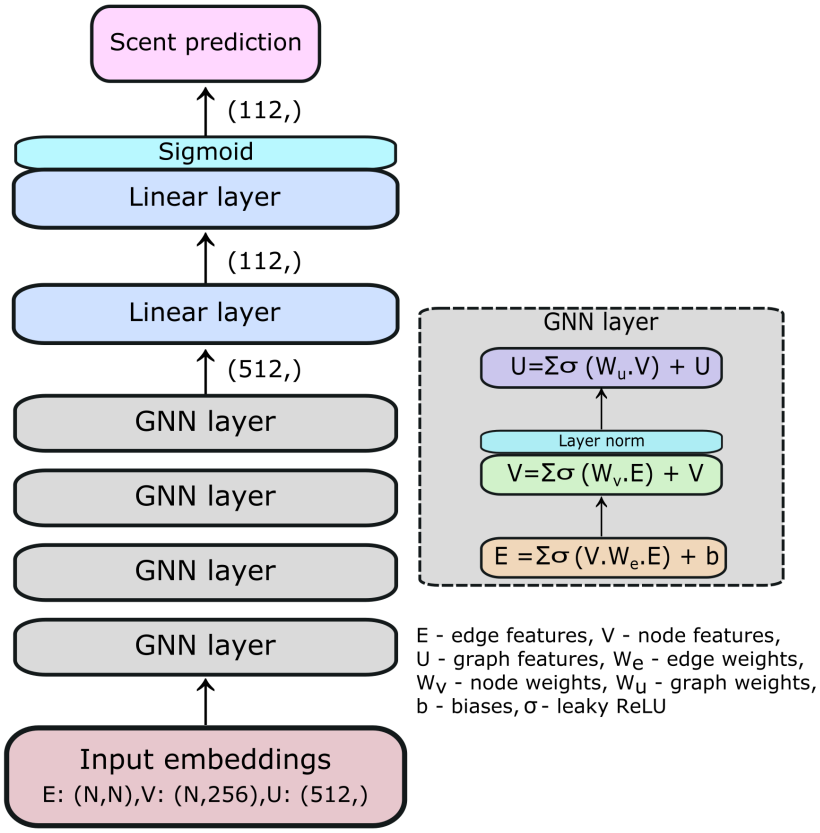

Figure 2: Architecture of GNN model used to predict molecular scent. The molecular graph representation of the molecule, which consists of the adjacency matrix and node feature vector is the model input. The model consists of four GNN layers followed by two Linear layers. The model output is a 112 length vector, with each entry corresponding to a scent class

2.3 Model Evaluation

Both the logistic regression and GNN models used in this study were trained by minimizing the cross-entropy loss. The macro-average, micro-average and weighted average AUROC, precision, recall and F1 scores were calculated. The macro-averaged values correspond to the average of the metric for each scent class. The micro-averaged values correspond to each example in the dataset having an equal contribution to the metric. For weighted-average metrics, the weighting corresponds to the number of positive examples in each scent class. Additionally, the median AUROC was calculated by taking the median of the individual AUROC values for each scent class. The threshold values for each scent class were optimized by taking the threshold value that maximized the F1 score on the training and validation sets combined. These computations were done using the scikit-learn Python library.[57]

To test whether the GNN model was learning information relating the structure of molecules and their associated scent, the GNN was also trained on a mislabeled dataset, where each molecule was incorrectly labeled as having the scent of the subsequent molecule in the original dataset. The mean AUROC of the model trained on the mislabeled dataset was much lower than that for the model trained on the original dataset, suggesting that the GNN model did indeed learn a relationship between molecular structure and scent.

The primary scents examined using XAI methods were selected as those with an AUROC on the test set above 0.8 as well as the GNN model correctly classifying over 70% of the positive examples for that scent class in the dataset. The latter condition was included because, as described in 2.5, for each scent, the spaces for the descriptor explanations were generated using all of the molecules with a positive label for that scent in the dataset.

2.4 Counterfactual Generation

Wellawatte et al.[46] developed a method, molecular model agnostic counterfactual explanations (MMACE), for counterfactual generation. We use this method to explain smells predicted for ethyl benzoate and 2,4 decadienal – molecules known for having 'fruity' and 'fatty' scents, respectively.[58–60] MMACE first uses the STONED-SELFIES method[61] to generate a local chemical space around the molecule of interest and then identifies counterfactual molecules from this space. There are a few parameters that can be passed in to MMACE for creation of the chemical space before selecting counterfactuals. These include number of mutations, type of mutations (called the "alphabet") and size of the chemical space. Here, the input parameters used for counterfactual generation were a sample size of 20,000 molecules, a maximum of two mutations, and a modified version of the basic alphabet. The basic alphabet was modified to remove boron, bromine, chlorine, fluorine and iodine as these elements were not present in any of the molecules in the dataset. Each molecule can have multiple scents. In order to get 'true' counterfactuals, we modify MMACE algorithm to pick molecules where the label for only the scent of interest is flipped. This ensures that the counterfactuals still have other scents attributed to them, and we explain just one scent.

2.5 Descriptor explanations and natural language explanations

Descriptor explanation was used to identify significant substructures that influence the prediction of whether a molecule has a certain scent or not. In a recent study, Gandhi and White[41] used LIME with MACCS fingerprints[37] and ECFP fingerprints[62] to explain which molecular substructures contribute to a certain model prediction. These are primarily local explanations. They also derive natural language explanations from these descriptor explanations. The procedure for creating a perturbed chemical space around the molecule of interest was the same as MMACE, using the STONED-SELFIES method.

To understand more generally why molecules might have a particular scent, we modify the method developed by Gandhi and White[41] to give global explanations. For each molecule that has a positive label for a scent, local chemical spaces are generated using the STONED-SELFIES algorithm. The input parameters used for generating each local sample space were a sample size of 200 molecules, a maximum of two mutations, and the modified version of the basic alphabet that was used for counterfactual generation. Then, these chemical spaces are combined to form a larger space, which is used for the descriptor explanation analysis. The same chemical space was used when identifying significant MACCS and ECFP descriptors. This procedure is used to get explanations for all of the scents.

3 Results and Discussion

3.1 Model Performance

The mean AUROC and F1 score for the logistic regression baseline model and GNN model are given in Table 1. The model results reported by Sanchez-Lengeling et al.[11] using the same Leffingwell Odor Dataset are also included along with the results reported by Sharma et al.[4]. Previous studies have reported mean AUROC scores generally ranging between 0.767 and 0.913.[4,11,63] As shown in Table 1, the model performance of the GNN that was used in this study for exploring structure-scent relationships is comparable to others in the literature. As model accuracy was not the primary objective of this study, the performance of the GNN model is acceptable. Additional model performance results, including the median AUROC and macro-average, micro-average, and weighted-average AUROC, F1, precision, and recall scores for the logistic regression and GNN models can be found in Table S3. The GNN AUROC score for each scent class can be found in Table S4.

Table 1: Macro-average AUROC and F1 score for the logistic regression baseline and GNN models. Reported model performance metrics from other scent-prediction studies are also provided for comparison.

Model	Dataset	Macro-average AUROC	Mean F1
Logistic Regression Baseline	Leffingwell Odor Dataset	0.873	0.329
GNN Model	**Leffingwell Odor Dataset**	**0.885**	**0.308**
Sanchez-Lengeling et al.[11] GNN	Leffingwell Odor Dataset	0.913	0.406
Sharma et al.[4] CNN	Sharma et al.[4]	0.767	0.88

After filtering the scent classes down to those that had a test set AUROC value greater than 0.8 as well as the model correctly predicting, for that specific scent, over 70% of the positive examples in the dataset, nine scents remained. These nine scents were: 'alcoholic', 'apple', 'fatty', 'fruity', 'meaty', 'popcorn', 'roasted', 'sulfurous', and 'catty'. The

number of positive examples, percentage of correct predictions for the positive examples, and test set AUROC, F1, precision, and recall scores for these scents is given in Table 2. Since the 'catty' scent had much lower precision, recall, and F1 scores compared to the other scents, it was not included in the main analysis.

Table 2: Scents that had a test set AUROC above 0.8 and the correct model prediction for over 70% of the positive examples. For each scent, the number of positive examples, percentage of correct predictions for the positive examples, and test set AUROC, F1, precision, and recall scores are provided.

Scent	Number of Positive Examples	True Positives Rate (%)	AUROC	F1	Precision	Recall
Fruity	1392	84.7	0.882	0.759	0.698	0.832
Fatty	407	71.3	0.890	0.599	0.528	0.691
Sulfurous	252	72.2	0.972	0.558	0.592	0.527
Apple	239	70.7	0.916	0.492	0.403	0.633
Meaty	218	72.0	0.934	0.547	0.510	0.591
Roasted	195	70.8	0.955	0.605	0.553	0.667
Alcoholic	85	82.4	0.994	0.651	0.538	0.824
Popcorn	23	73.9	0.932	0.667	0.750	0.600
Catty	20	75.0	0.945	0.167	0.125	0.250

3.2 Counterfactual Analysis

We conducted counterfactual analysis for two molecules - ethyl benzoate and 2,4 decadienal to explain their scents. Ethyl benzoate, which is used in fragrances and as a flavoring compound, is known to have a 'fruity' scent and is found in many fruits, including species of apples, bananas, and strawberries.[58,59,64] One of the main aroma compounds in butter oil, 2,4 decadienal is well-documented as having a 'fatty' scent.[60,65]

Counterfactuals generated around the 'fruity' scent using ethyl benzoate as a base molecule and the 'fatty' scent using 2,4 decadienal as the base molecule are given in Figure 3. These counterfactuals correspond to molecules from the sampled chemical space with the highest Tanimoto similarity to ethyl benzoate or 2,4 decadienal, but lack the 'fruity' or 'fatty' scent, respectively.

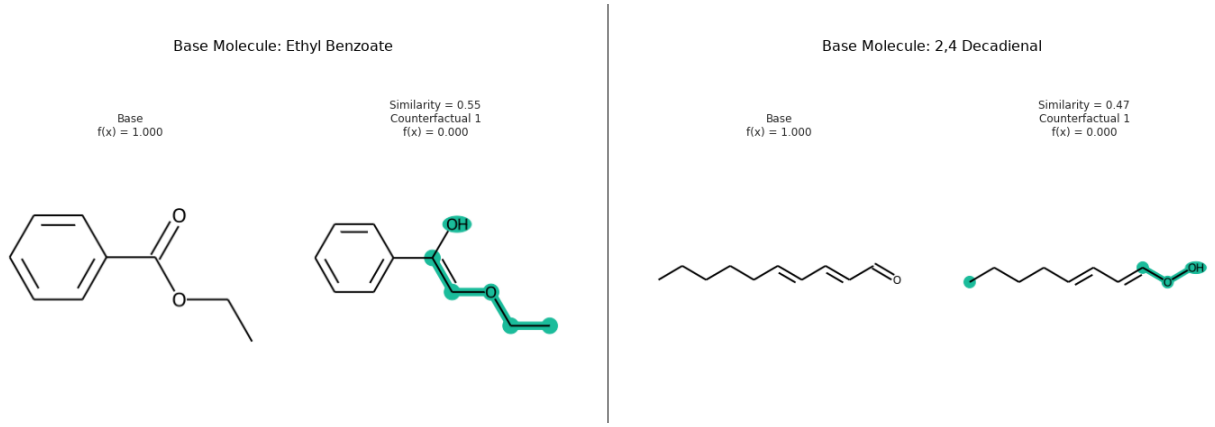

Figure 3: Counterfactuals generated around the 'fruity' and 'fatty' scents using ethyl benzoate and 2,4 decadienal as the base molecules, respectively. The counterfactuals indicate structural changes to the base molecule that would result in the model predicting the molecule to not have 'fruity' (for the ethyl benzoate counterfactual) or 'fatty' (for the 2,4 decadienal counterfactual) scents. The Tanimoto similarity between the counterfactuals and the base molecules is also provided.

As shown in Figure 3, altering the structure of ethyl benzoate such that it is no longer an ester would likely result in it no longer having a 'fruity' scent. This result is supported by the fact that esters are generally associated with 'fruity' scents.[8,14] As the 'fruity' counterfactual only differs from the base molecule by one structural change, it can also be considered as a matched molecular pair for ethyl benzoate.[66]

For 2,4 decadienal, it appears that replacing the carbonyl group with a hydroperoxide (-OOH) group and reducing the number of carbons in the chain from ten to eight may result in the molecule no longer having a 'fatty' scent (Figure 3). Increasing the carbon chain length is known to increase the 'fatty' scent for some molecules.[67] Additionally, Licon et al.[51] used a data-mining approach to identify structure-scent relationships and reported that most molecules with a 'fatty' scent had a "larger carbon-chain skeleton which is highly hydrop[h]obic with aldehyde or acid functions". These findings further support the result shown in Figure 3, since compared to 2,4 decadienal, the 'fatty' counterfactual has a shorter carbon chain and is no longer an aldehyde.

3.3 Descriptor Explanations

Descriptor explanations were found for 'fruity' and 'fatty' scents using the method described in section 2.5. This method is hypothesized to provide global explanations of why molecules may have a certain scent. MACCS fingerprints and ECFP fingerprints are used to represent molecules. The descriptor explanation results using MACCS and ECFP fingerprints for the 'fruity' and 'fatty' scents are given below in Figure 4.

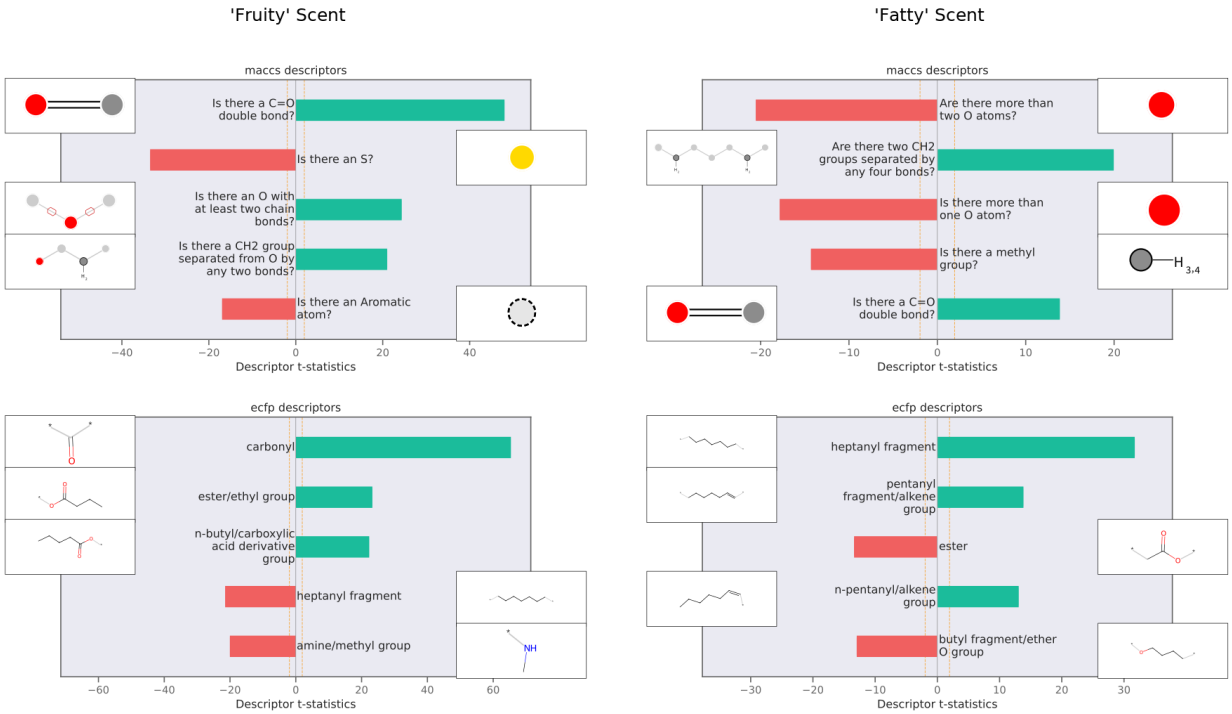

Figure 4: The five most significant ECFP and MACCS descriptors found using the sample space obtained from enumerating around all molecules in the Leffingwell Odor Dataset with 'fruity' and 'fatty' scents, respectively. The green and red bars show descriptors that influence predictions positively and negatively, respectively. Dotted yellow lines show significance threshold ($\alpha = 0.05$) for the t-statistic. MACCS substructures may either be present in the molecule as is or may represent a modification, and ECFP fingerprints are substructures in the molecule that affect the prediction.[68]

As shown in Figure 4, the presence of an ester group, carbonyl group, and an oxygen with two chain bonds positively influences the 'fruity' scent of a molecule. The combination of the latter two descriptors could also correspond to an ester group. Esters are often known for having a 'fruity' scent, but Jelen and Gracka[67] note that increasing the chain length can result in their scent becoming "more fatty, soapy or metallic".[67] This trend can support the result in Figure 4, where the fourth ECFP descriptor is a heptanyl fragment that appears to negatively influence the 'fruity' scent of a molecule.

Three of the top five most significant ECFP descriptors contain alkyl fragments with five or seven carbon atoms and positively influence the 'fatty' scent of a molecule (Figure 4). For aldehydes, alcohols, and esters, molecules with a

longer chain length can have more of a 'fatty' scent, which supports this result of alkyl chains positively influencing a molecule's 'fatty' scent.[67] The fifth MACCS descriptor, a carbon-oxygen double bond, is present in both aldehydes and carboxylic acids (Figure 4). As stated earlier, Licon et al.[51] found that molecules with a 'fatty' scent had a longer carbon chain and aldehyde or acid functional groups.[51] These results further support the conclusion from this study that long alkyl chains and the presence of a carbon-oxygen double bond are likely influential in determining a molecule's 'fatty' scent.

3.4 Natural language explanations

Following the procedure laid out by Gandhi and White[41] to obtain natural language explanations, we find text explanations for molecular scents from their MACCS and ECFP descriptor explanations. Table 3 contains text explanations for the eight scents that had an AUROC above 0.8 and had over 70% of their positive examples correctly classified by the GNN model. A complete list of explanations for all scents is given in the SI.

Table 3: Natural Language explanations generated using GPT-3 `text-davinci-003`[69,70] model for 'alcoholic', 'apple', 'fatty', 'fruity', 'meaty', 'popcorn', 'roasted', and 'sulfurous' scents. These scents were chosen as they had an AUROC above 0.8 and the GNN model correctly predicted over 70% of the positive examples had that scent

Scent	Natural Language Explanation
alcoholic	The molecular property "alcoholic scent" can be explained by the presence of an ethyl/ether O group and the absence of acetal like/methyl groups, two CH2 groups separated by any three bonds, an alkyne group, and an S. These are all very important for the property.
apple	The molecular property "apple scent" can be explained by the presence of an ethyl/ether O group, as well as the absence of an aromatic atom, hydroxy oxygen (OH), propyl fragment, and S. These structure-property relationships are very important for the property and explain why the molecule smells like apples.
fatty	The molecular property "fatty scent" can be explained by the presence of a heptanyl fragment, two CH2 groups separated by four bonds, and a C=O double bond, as well as the lack of more than one or two O atoms.
fruity	The molecular property "fruity scent" can be explained by the presence of a carbonyl, a C=O double bond, and an oxygen atom with at least two chain bonds. The lack of a heptanyl fragment and an S atom also contribute to this property.
meaty	The molecular property "meaty scent" can be explained by the presence of an S atom and the absence of an alkene, an O atom, a methyl group, and a CH2 group bonded to two neighbors by non-ring bonds.
popcorn	The molecular property "popcorn scent" can be explained by the presence of a carbonyl and a hetero N nonbasic/aromatic group. The absence of a carboxylic acid derivative/methyl group, an amine (NH2) group, and a heteroatom bonded to a methyl C was also important for the property.
roasted	The molecular property "roasted scent" can be explained by the presence of an S atom, a hetero N nonbasic/aromatic group, and an atom at a ring/chain boundary. The absence of an N bonded to at least one H, and a CH2 group bonded to two neighbors by non-ring bonds were also important in contributing to the molecular property.
sulfurous	The molecular property "sulfurous scent" can be explained by the presence of an S atom and a hydroxy oxygen (OH) group, as well as the lack of a CH2 group bonded to two neighbors by non-ring bonds, an O atom, and a methyl group.

Natural language explanations are a condensed plain text form of the descriptor explanations and we observe for many scents that text explanations agree with structure-scent relationships reported in previous studies. Many 'fruity' molecules are esters, but aldehydes, ketones, and lactones, which all have a C=O double bond, have also been found to contribute to a molecule's 'fruity' scent, supporting the result in Table 3 that a C=O double bond is important for 'fruity' molecules.[71] As mentioned previously, molecules with a 'fatty' scent tend to have a longer carbon chain, and this trend supports the result that a heptanyl fragment is very important in determining if a molecule has a 'fatty' scent (Table 3).[51,60,67]

The natural language explanations for the 'sulfurous' and 'meaty' scents indicate that the presence of sulfur positively influences these scents (Table 3). This result is supported in the literature as both 'sulfurous' and 'meaty' scents are associated with sulfur-containing compounds.[3,8,51,60,72] Heterocycles, especially heterocyclic rings containing sulfur are known to be important for 'meaty' scents.[72,73] However, an S-heterocycle was not found to be the most important. This result may be due to the diversity of 'meaty' molecules in the dataset used, as less than half of these molecules (103 out of 218) contained a heterocycle and only 52 out of the 218 molecules contained an S-heterocycle. Sulfur-containing compounds are also associated with 'bread', 'coffee', 'grapefruit', 'garlic' and 'potato' scents, and the importance of sulfur-related descriptors are found in the natural language explanations for each of these scents (Table S5).[3,74]

According to Rowe[60], many acetyl compounds are associated with 'popcorn' scents, which supports importance of the carbonyl group in the natural language explanation for the 'popcorn' scent as the acetyl group consists of a carbonyl (C=O) structure bonded to a methyl group. It is important to note that for the 'popcorn' scent, there were only 23 positive examples in the dataset, which may have influenced the results.

In some cases, text explanations failed to reasonably explain the scents. For example, although the model performed reasonably well for the 'apple' scent, and correctly predicted over 70% of the positive examples as having an 'apple' scent, the text explanation does not match the expected result that esters and alcohols are important (Table 3).[71] In the dataset used the majority of molecules with an 'apple' scent had an ester group. Some possible reasons for the fact that an ester group was not identified as being important include the fact that the fit of the ECFP explanatory model to the GNN model had a low correlation of 0.618 for the 'apple' scent, and that the dataset used was highly imbalanced. Conducting the same analysis using a more balanced dataset with a greater number of positive examples for each of the scent classes likely would result in the explanations more closely matching known structure-scent relationships and could aid in the proposal of new structure-scent relationships. The 'pineapple' scent has 139 positive examples and a correlation of 0.737 of the ECFP model with the GNN model. For this scent, the ester group, which is present in many pineapple volatile organic compounds, was found to be important in the natural language explanation (Table S5).[71,75]

In general, most of the MACCS descriptors were found to have higher t-statistics compared to ECFP descriptors for the same scent, which resulted in primarily MACCS descriptors appearing in the natural language explanations. This may be due to the MACCS descriptors matching more general substructures, while the ECFP descriptors correspond to more specific substructures. However, in some cases, like for 'pungent' the ECFP descriptors were able to capture the importance of structures such as isothiocyanate, which are thought to contribute to 'pungent' flavors (Table S5).[76] Conducting additional analysis with different descriptor types may result in a closer match between the natural language explanations and known structure-scent relationships.

4 Conclusions

We have shown how to go from a dataset to a natural language explanation of a structure-function relationship purely with artifical intelligence methods. We created a graph neural network that can accurately predict scents across a range of scent categories. Then we generated local chemical spaces around known positive molecules from the dataset and used those spaces to find counterfactual molecules that explain which parts of the structure contribute to scent for a specific example. The counterfactuals are local explanations. We then joined all these chemical spaces and fit surrogate models of ECFP/MACCS descriptors to give global explanations of the important molecular structures for each scent. Finally, these surrogate models were summarized into two sentence natural language summaries for the 112 scents.

The explanations can be useful for creating fragrances and flavor compounds, where knowing how to modify a structure can change or remove a scent. This can be done with counterfactuals. The global explanations can provide insight into biological mechanisms and also provide clues to biases in the dataset. For example, it is not clear if 'fatty' scented molecules require a carbonyl to remain soluble/volatile or for their scent. Future work can explore these unusual explanations and make use of the emerging larger Pyrfume Project dataset.[63]

5 Code Availability

All code for this work is available at `https://github.com/ur-whitelab/exmol` and the data is available in Sanchez-Lengeling et al.[11].

6 Acknowledgements

Research reported in this work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM137966. This work was supported by the NSF under awards 1751471

and 1764415. We thank the Center for Integrated Research Computing at the University of Rochester for providing computational resources.

References

[1] Kensaku Mori and Yoshihiro Yoshihara. Molecular recognition and olfactory processing in the mammalian olfactory system. *Progress in neurobiology*, 45(6):585–619, 1995.

[2] Rinu Chacko, Deepak Jain, Manasi Patwardhan, Abhishek Puri, Shirish Karande, and Beena Rai. Data based predictive models for odor perception. *Scientific reports*, 10(1):1–13, 2020.

[3] Karen J Rossiter. Structure- odor relationships. *Chemical reviews*, 96(8):3201–3240, 1996.

[4] Anju Sharma, Rajnish Kumar, Shabnam Ranjta, and Pritish Kumar Varadwaj. Smiles to smell: decoding the structure–odor relationship of chemical compounds using the deep neural network approach. *Journal of Chemical Information and Modeling*, 61(2):676–688, 2021.

[5] Valentina Villalobos Coa, Vito Lubes, Johannes Polster, Maria Monteiro de Araújo Silva, and Giuseppe Lubes. Relationship between structure and odor. In *Food Aroma Evolution*, pages 679–694. CRC Press, 2019.

[6] Anju Sharma, Rajnish Kumar, Imlimaong Aier, Rahul Semwal, Pankaj Tyagi, and Pritish Varadwaj. Sense of smell: structural, functional, mechanistic advancements and challenges in human olfactory research. *Current neuropharmacology*, 17(9):891–911, 2019.

[7] Heinz Breer. Olfactory receptors: molecular basis for recognition and discrimination of odors. *Analytical and bioanalytical chemistry*, 377(3):427–433, 2003.

[8] Manon Genva, Tierry Kenne Kemene, Magali Deleu, Laurence Lins, and Marie-Laure Fauconnier. Is it possible to predict the odor of a molecule on the basis of its structure? *International journal of molecular sciences*, 20(12): 3018, 2019.

[9] National Center for Biotechnology Information. Pubchem compound summary for cid 8892, hexanoic acid, . URL `https://pubchem.ncbi.nlm.nih.gov/compound/8892`. Accessed August 31, 2022.

[10] Nethaji J Gallage and Birger Lindberg Møller. Vanilla: The most popular flavour. In *Biotechnology of natural products*, pages 3–24. Springer, 2018.

[11] Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019.

[12] John C Leffingwell et al. Olfaction–update no. 5. *Leffingwell reports*, 2(1):1–34, 2002.

[13] D.G. Laing, P.K. Legha, A.L. Jinks, and I. Hutchinson. Relationship between Molecular Structure, Concentration and Odor Qualities of Oxygenated Aliphatic Molecules. *Chemical Senses*, 28(1):57–69, 01 2003. ISSN 0379-864X. doi:10.1093/chemse/28.1.57. URL `https://doi.org/10.1093/chemse/28.1.57`.

[14] Charles S Sell. On the unpredictability of odor. *Angewandte Chemie International Edition*, 45(38):6254–6261, 2006.

[15] Anne Tromelin, Florian Koensgen, Karine Audouze, Elisabeth Guichard, and Thierry Thomas-Danguin. Exploring the Characteristics of an Aroma-Blending Mixture by Investigating the Network of Shared Odors and the Molecular Features of Their Related Odorants. *Molecules*, 25(13), jul 2020. ISSN 14203049. doi:10.3390/MOLECULES25133032. URL `/pmc/articles/PMC7411594//pmc/articles/PMC7411594/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7411594/`.

[16] Luca Turin. A spectroscopic mechanism for primary olfactory reception. *Chemical senses*, 21(6):773–791, 1996.

[17] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[18] Andrew D White. Deep learning for molecules and materials. *Living Journal of Computational Molecular Science*, 3(1):1499, 2021. doi:10.33011/livecoms.3.1.1499. URL `https://dmol.pub`.

[19] Andreas Keller, Richard C Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, et al. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017.

[20] Jörn Lötsch, Dario Kringel, and Thomas Hummel. Machine Learning in Human Olfactory Research. *Chemical Senses*, 44(1):11–22, 10 2018. ISSN 0379-864X. doi:10.1093/chemse/bjy067. URL `https://doi.org/10.1093/chemse/bjy067`.

[21] Nicolas Armanino, Julie Charpentier, Felix Flachsmann, Andreas Goeke, Marc Liniger, and Philip Kraft. What's Hot, What's Not: The Trends of the Past 20 Years in the Chemistry of Odorants. *Angewandte Chemie International Edition*, 59(38):16310–16344, 2020. doi:https://doi.org/10.1002/anie.202005719. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202005719.

[22] Lei Zhang, Haitao Mao, Yu Zhuang, Lu Wang, Linlin Liu, Yachao Dong, Jian Du, Wancui Xie, and Zhihong Yuan. Odor prediction and aroma mixture design using machine learning model and molecular surface charge density profiles. *Chemical Engineering Science*, 245:116947, 2021.

[23] Liang Shang, Chuanjun Liu, Yoichi Tomiura, and Kenshi Hayashi. Machine-Learning-Based Olfactometer: Prediction of Odor Perception from Physicochemical Features of Odorant Molecules. *Analytical Chemistry*, 89(22): 11999–12005, 2017. doi:10.1021/acs.analchem.7b02389. URL https://doi.org/10.1021/acs.analchem.7b02389.

[24] Yuji Nozaki and Takamichi Nakamoto. Odor Impression Prediction from Mass Spectra. *PLOS ONE*, 11(6):1–15, 2016. doi:10.1371/journal.pone.0157030. URL https://doi.org/10.1371/journal.pone.0157030.

[25] Yuji Nozaki and Takamichi Nakamoto. Predictive modeling for odor character of a chemical using machine learning combined with natural language processing. *PLOS ONE*, 13(6):1–13, 2018. doi:10.1371/journal.pone.0198475. URL https://doi.org/10.1371/journal.pone.0198475.

[26] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics 2020 12:1*, 12(1):1–22, sep 2020. ISSN 1758-2946. doi:10.1186/S13321-020-00460-5. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00460-5.

[27] Brian K Lee, Emily J Mayhew, Benjamin Sanchez-Lengeling, Jennifer N Wei, Wesley W Qian, Kelsie Little, Matthew Andres, Britney B Nguyen, Theresa Moloy, Jane K Parker, Richard C Gerkin, Joel D Mainland, and Alexander B Wiltschko. A principal odor map unifies diverse tasks in human olfactory perception. *bioRxiv*, 2022. doi:10.1101/2022.09.01.504602. URL https://www.biorxiv.org/content/early/2022/09/06/2022.09.01.504602.

[28] Chuanjun Liu, Hitoshi Miyauchi, and Kenshi Hayashi. DeepSniffer: A meta-learning-based chemiresistive odor sensor for recognition and classification of aroma oils. *Sensors and Actuators B: Chemical*, 351:130960, 2022. ISSN 0925-4005. doi:https://doi.org/10.1016/j.snb.2021.130960. URL https://www.sciencedirect.com/science/article/pii/S0925400521015288.

[29] Richard Dybowski. Interpretable machine learning as a tool for scientific discovery in chemistry. *New Journal of Chemistry*, 44(48):20914–20920, 2020.

[30] Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith T Butler. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research*, 3(6):597–607, 2022.

[31] Carlos Zednik and Hannes Boelsen. Scientific exploration and explainable artificial intelligence. *Minds and Machines*, 32(1):219–239, 2022.

[32] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38, 2019.

[33] Kushagra Saini and Venkatnarayan Ramanathan. Predicting odor from molecular structure: a multi-label classification approach. *Scientific reports*, 12(1):1–11, 2022.

[34] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

[35] Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1):1–14, 2018.

[36] Francesca Grisoni, Viviana Consonni, and Roberto Todeschini. Impact of molecular descriptors on computational models. In *Computational chemogenomics*, pages 171–209. Springer, 2018.

[37] Liangxu Xie, Lei Xu, Ren Kong, Shan Chang, and Xiaojun Xu. Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. *Frontiers in Pharmacology*, 11:2148, 2020. ISSN 1663-9812. doi:10.3389/fphar.2020.606668. URL https://www.frontiersin.org/article/10.3389/fphar.2020.606668.

[38] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[39] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

[40] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[41] Heta A. Gandhi and Andrew D. White. Explaining structure-activity relationships using locally faithful surrogate models. *ChemRxiv*, 2022. doi:10.26434/chemrxiv-2022-v5p6m. URL https://doi.org/10.26434/chemrxiv-2022-v5p6m.

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[43] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

[44] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

[45] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[46] Geemi P. Wellawatte, Aditi Seshadri, and Andrew D. White. Model agnostic generation of counterfactual explanations for molecules. *Chemical Science*, pages –, 2022. doi:10.1039/D1SC05259D. URL http://dx.doi.org/10.1039/D1SC05259D.

[47] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.

[48] Mohammed Temraz, Eoin Kenny, Elodie Ruelle, Laurence Shalloo, Barry Smyth, and Mark T Keane. Handling climate change using counterfactuals: Using counterfactuals in data augmentation to predict crop growth in an uncertain climate future. *arXiv preprint arXiv:2104.04008*, 2021.

[49] Emily J Mayhew, Charles J Arayata, Richard C Gerkin, Brian K Lee, Jonathan M Magill, Lindsey L Snyder, Kelsie A Little, Chung Wen Yu, and Joel D Mainland. Transport features predict if a molecule is odorous. *Proceedings of the National Academy of Sciences*, 119(15):e2116576119, 2022. doi:10.1073/pnas.2116576119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2116576119.

[50] Joel Kowalewski, Brandon Huynh, and Anandasankar Ray. A system-wide understanding of the human olfactory percept chemical space. *Chemical senses*, 46, 2021.

[51] Carmen C Licon, Guillaume Bosc, Mohammed Sabri, Marylou Mantel, Arnaud Fournel, Caroline Bushdid, Jerome Golebiowski, Celine Robardet, Marc Plantevit, Mehdi Kaytoue, et al. Chemical features mining provides new descriptive structure-odor relationships. *PLoS computational biology*, 15(4):e1006945, 2019.

[52] Ria Gupta, Aayushi Mittal, Vishesh Agrawal, Sushant Gupta, Krishan Gupta, Rishi Raj Jain, Prakriti Garg, Sanjay Kumar Mohanty, Riya Sogani, Harshit Singh Chhabra, Vishakha Gautam, Tripti Mishra, Debarka Sengupta, and Gaurav Ahuja. OdoriFy: A conglomerate of Artificial Intelligence-driven prediction engines for olfactory decoding. *Journal of Biological Chemistry*, 297(2):100956, aug 2021. ISSN 1083351X. doi:10.1016/J.JBC.2021.100956/ATTACHMENT/53AFB23E-3D58-4054-8394-7A4D48D05B22/MMC3.XLSX. URL http://www.jbc.org/article/S0021925821007560/fulltexthttp://www.jbc.org/article/S0021925821007560/abstracthttps://www.jbc.org/article/S0021-9258(21)00756-0/abstract.

[53] Piotr Szymański and Tomasz Kajdanowicz. A scikit-based python environment for performing multi-label classification. *arXiv preprint arXiv:1702.01460*, 2017.

[54] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[55] Tom Hennigan, Trevor Cai, Tamara Norman, and Igor Babuschkin. Haiku: Sonnet for jax. *URL http://github.com/deepmind/dm-haiku*, 2020.

[56] Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[58] AM Api, D Belsito, D Botelho, M Bruze, GA Burton Jr, J Buschmann, ML Dagli, M Date, W Dekant, C Deodhar, et al. Rifm fragrance ingredient safety assessment, ethyl benzoate, cas registry number 93-89-0. *Food Chem. Toxicol*, 127(Suppl 1):S172–s8, 2019.

[59] National Center for Biotechnology Information. Pubchem compound summary for cid 7165, ethyl benzoate, . URL https://pubchem.ncbi.nlm.nih.gov/compound/Ethyl-benzoate. Accessed November 14, 2022.

[60] D Rowe. Aroma chemicals for savory flavors. *Perfumer and Flavorist*, 23:9–18, 1998.

[61] AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alan Aspuru-Guzik. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chemical science*, 2021.

[62] H L Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi:10.1021/c160017a018. URL https://doi.org/10.1021/c160017a018.

[63] Richard C Gerkin. Parsing sage and rosemary in time: The machine learning race to crack olfactory perception. *Chemical Senses*, 46, 2021.

[64] Lillian C 'Becker, Wilma F Bergfeld, Donald V Belsito, Ronald A Hill, Curtis D Klaassen, Daniel Liebler, James G Marks Jr, Ronald C Shank, Thomas J Slaga, Paul W Snyder, et al. Safety assessment of alkyl benzoates as used in cosmetics. *International journal of toxicology*, 31(6_suppl):342S–372S, 2012.

[65] Silvia Mallia, Felix Escher, and Hedwig Schlichtherle-Cerny. Aroma-active compounds of butter: a review. *European Food Research and Technology*, 226(3):315–325, 2008.

[66] Christian Tyrchan and Emma Evertsson. Matched molecular pair analysis in short: algorithms, applications and limitations. *Computational and structural biotechnology journal*, 15:86–90, 2017.

[67] Henryk Jelen and Anna Gracka. Characterization of aroma compounds: Structure, physico-chemical and sensory properties. *Flavour: From food to perception*, pages 126–153, 2016.

[68] SMARTS annotations for MACCS descriptors were created using SMARTSviewer (smartsview.zbh.uni-hamburg.de, Copyright: ZBH, Center for Bioinformatics Hamburg) developed by Schomburg et al.[77].

[69] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[70] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[71] Salma Mostafa, Yun Wang, Wen Zeng, and Biao Jin. Floral scents and fruit aromas: Functions, compositions, biosynthesis, and regulation. *Frontiers in plant science*, 13, 2022.

[72] Mónica Flores. Chapter 13 - the eating quality of meat: Iii—flavor. In Fidel Toldra´, editor, *Lawrie´s Meat Science (Eighth Edition)*, Woodhead Publishing Series in Food Science, Technology and Nutrition, pages 383–417. Woodhead Publishing, eighth edition edition, 2017. ISBN 978-0-08-100694-8. doi:https://doi.org/10.1016/B978-0-08-100694-8.00013-3. URL https://www.sciencedirect.com/science/article/pii/B9780081006948000133.

[73] Donald S. Mottram. Flavour formation in meat and meat products: a review. *Food Chemistry*, 62(4):415–424, 1998. ISSN 0308-8146. doi:https://doi.org/10.1016/S0308-8146(98)00076-4. URL https://www.sciencedirect.com/science/article/pii/S0308814698000764.

[74] Robert J McGorrin. The significance of volatile sulfur compounds in food flavors: An overview. *Volatile sulfur compounds in food*, pages 3–31, 2011.

[75] Yukiko Tokitomo, Martin Steinhaus, Andrea Büttner, and Peter Schieberle. Odor-active constituents in fresh pineapple (ananas comosus [l.] merr.) by quantitative and sensory evaluation. *Bioscience, Biotechnology, and Biochemistry*, 69(7):1323–1330, 2005.

[76] Monika Marcinkowska and Henryk H. Jeleń. Determination of the odor threshold concentrations and partition coefficients of isothiocyanates from brassica vegetables in aqueous solution. *LWT*, 131:109793, 2020. ISSN 0023-6438. doi:https://doi.org/10.1016/j.lwt.2020.109793. URL https://www.sciencedirect.com/science/article/pii/S0023643820307829.

[77] Karen Schomburg, Hans Christian Ehrlich, Katrin Stierand, and Matthias Rarey. From structure diagrams to visual chemical patterns. *Journal of Chemical Information and Modeling*, 50(9):1529–1535, sep 2010. ISSN 15499596. doi:10.1021/ci100209a. URL https://pubs.acs.org/doi/full/10.1021/ci100209a.

SUPPLEMENTAL INFORMATION: WHY DOES THAT MOLECULE SMELL?

S1 Hyperparameter Selection

The hyperparameters examined were the number of GNN layers, number of dense layers, GNN message feature length, GNN node feature length, GNN graph feature length, learning rate, the inclusion of edge updates in the GNN layers, and leaky ReLU (rather than ReLU) as the activation function. The hyperparameters used in the final GNN model can be found in Table S1. The training curves for each of the thirteen trials during the hyperparameter search can be found in Figure S1. The corresponding hyperparameter values for each trial can be found in Table S2.

Table S1: Optimal hyperparameter choices for GNN Model.

Hyperparameter	GNN Model
Number of GNN Layers	4
Number of Linear Layers	2
GNN Message Feature Length	256
GNN Node Feature Length	256
GNN Graph Feature Length	512
Learning Rate	1e-5
L2 Regularization Strength	1e-6
Early Stopping Patience	3
Number of Epochs	138

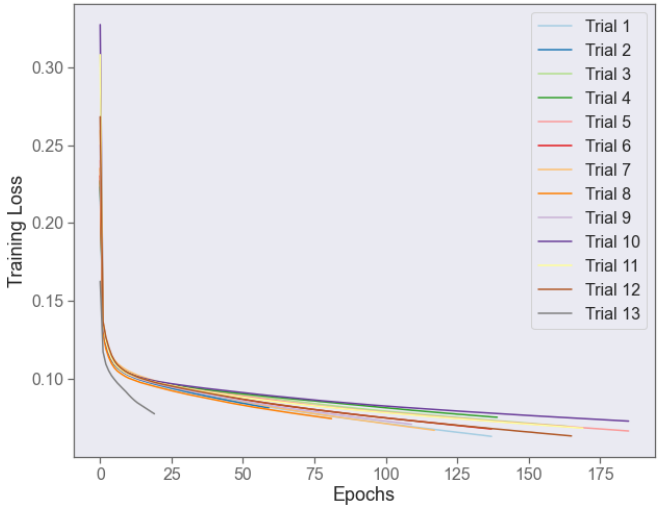

Figure S1: Training curves for each trial with early stopping. Trial 1 was the optimal trial, with the lowest training loss of 0.06323. The only difference between Trials 1 and 2 is that Trial 2 used ReLU rather than leaky ReLU as the activation function. The differences between the hyperparameter values for Trial 1 and all other trials, along with the final training losses, can be found in Table S2.

Table S2: Hyperparameters for different trials with early stopping.

Trial	Difference from Optimal Hyperparameters (Table S1)	Final Training Loss
1	Optimal	0.06323
2	ReLU activation function	0.08159
3	Edge updates in GNN layers	0.06959
4	Edge updates in GNN layers and ReLU activation function	0.07552
5	2 GNN layers	0.06671
6	3 GNN layers	0.06801
7	5 GNN layers, Regularization strength 10^{-5}	0.06715
8	1 Dense layer	0.07463
9	3 Dense layers	0.07098
10	message feature length = 128, node feature length = 128, graph feature length = 256	0.07303
11	message feature length = 128, node feature length = 256, graph feature length = 256	0.06851
12	message feature length = 256, node feature length = 256, graph feature length = 256	0.06352
13	learning rate = 1e-4	0.07772

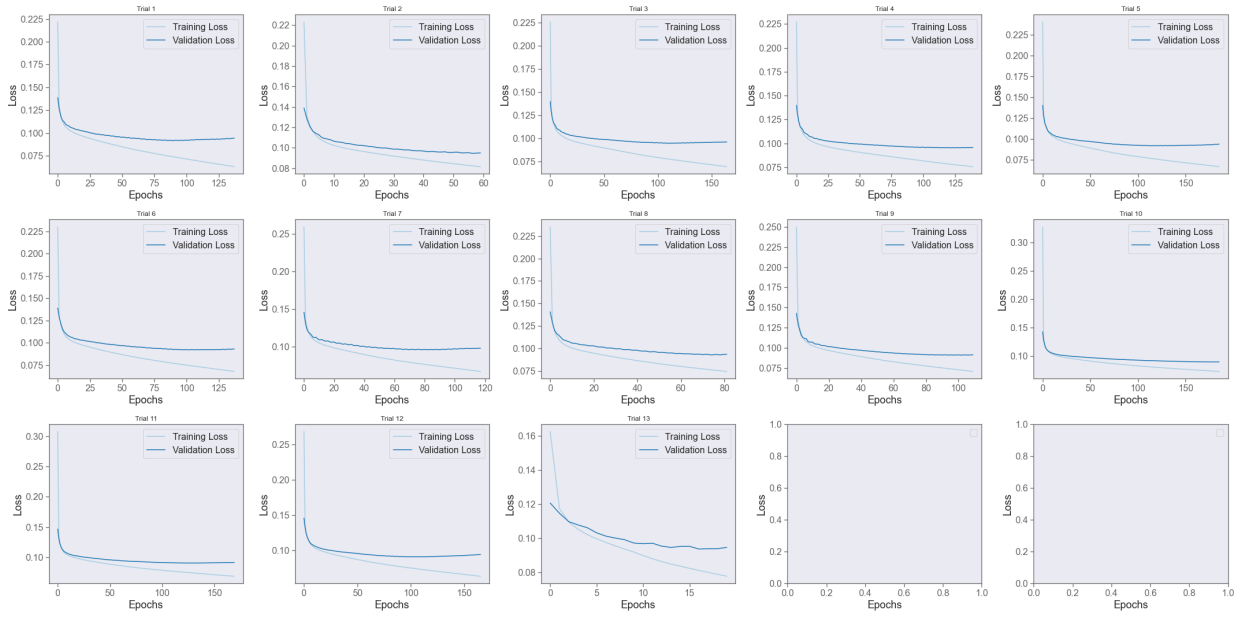

Figure S2: Training and validation loss versus epoch for each of the hyperparameter trials described in Table S2

S2 Additional Model Performance Results

Table S3: AUROC, precision, recall, and F1 score values for logistic regression and GNN models.

Performance Metric	Logistic Regression Baseline	GNN Model
Weighted average AUROC	0.865	0.865
Macro-average AUROC	0.873	0.885
Micro-average AUROC	0.915	0.925
Median AUROC	0.893	0.892
Weighted average Precision	0.435	0.432
Macro-average Precision	0.368	0.377
Micro-average Precision	0.447	0.451
Weighted Average Recall	0.467	0.446
Macro-average Recall	0.340	0.307
Micro-average Recall	0.467	0.446
Weighted Average F1	0.430	0.419
Macro-average F1	0.329	0.308
Micro-average F1	0.457	0.449

Table S4: AUROC value for each scent class

Scent Class	AUROC	Scent Class	AUROC
alcoholic	0.994	aldehydic	0.745
alliaceous	0.919	almond	0.846
animal	0.891	anisic	0.957
apple	0.916	apricot	0.863
aromatic	0.848	balsamic	0.953
banana	0.922	beefy	0.980
berry	0.780	blackcurrant	0.906
brandy	0.961	bread	0.891
brothy	0.932	burnt	0.844
buttery	0.928	cabbage	0.906
camphoreous	0.887	caramellic	0.852
catty	0.945	chamomile	0.965
cheesy	0.915	cherry	0.875
chicken	0.949	chocolate	0.900
cinnamon	0.886	citrus	0.881
cocoa	0.857	coconut	0.956
coffee	0.912	cognac	0.940
coumarinic	0.990	creamy	0.800
cucumber	0.948	dairy	0.807
dry	0.839	earthy	0.746
ethereal	0.943	fatty	0.890
fermented	0.847	fishy	0.932
floral	0.855	fresh	0.737
fruity	0.882	garlic	0.979
gasoline	0.966	grape	0.915
grapefruit	0.895	grassy	0.848
green	0.768	hay	0.838
hazelnut	0.977	herbal	0.776
honey	0.914	horseradish	0.923
jasmine	0.964	ketonic	0.988
leafy	0.827	leathery	0.870
lemon	0.854	malty	0.752
meaty	0.934	medicinal	0.920

Continued on next page

Table S4 – continued from previous page

Scent Class	AUROC	Scent Class	AUROC
melon	0.901	metallic	0.779
milky	0.809	mint	0.923
mushroom	0.830	musk	0.665
musty	0.786	nutty	0.825
oily	0.905	onion	0.971
orange	0.900	orris	0.775
peach	0.819	pear	0.870
phenolic	0.923	pine	0.913
pineapple	0.873	plum	0.853
popcorn	0.932	potato	0.943
pungent	0.883	radish	0.844
ripe	0.936	roasted	0.955
rose	0.892	rum	0.882
savory	0.918	sharp	0.950
smoky	0.970	solvent	0.979
sour	0.850	spicy	0.873
strawberry	0.877	sulfurous	0.972
sweet	0.741	tea	0.715
tobacco	0.900	tomato	0.888
tropical	0.910	vanilla	0.988
vegetable	0.860	violet	0.890
warm	0.737	waxy	0.928
winey	0.854	woody	0.867

Table S5: Natural Language explanations generated using GPT-3 `text-davinci-003` model for scents observed in the Leffingwell Odor Dataset.

Scent	**Why the scent?**
alcoholic	The molecular property "alcoholic scent" can be explained by the presence of an ethyl/ether O group and the absence of acetal like/methyl groups, two CH2 groups separated by any three bonds, an alkyne group, and an S. These are all very important for the property.
aldehydic	The molecular property "aldehydic scent" can be explained by the presence of an oxygen atom, a lack of an oxygen atom bonded to a secondary carbon atom, the absence of an aromatic/ether oxygen group, the absence of more than two oxygen atoms, and the lack of a sulfur atom.
alliaceous	The molecular property "alliaceous scent" can be explained by the presence of more than one CH2 group that is bonded to two neighbors, one of which is a heteroatom, an atom bonded to three other atoms, one of which is an S, and the presence of an S. The lack of an O and an oxymethylene group (-CH2O-) is also very important for the property.
almond	The molecular property "almond scent" can be explained by the presence of an oxygen atom, an aldehyde/aromatic group, and the absence of an atom bonded to two methyl groups, a CH2 group bonded to two neighbors by non-ring bonds, and an S atom. These are particularly important structure-property relationships for the almond scent.

Continued on next page

Scent	Why the scent?
animal	The molecular property "animal scent" can be explained by the presence of an atom in a ring, a C=O double bond, and the lack of more than two O atoms, an N separated from an O by any 4 bonds, and an S. These are all very important for the property.
anisic	The molecular property "anisic scent" can be explained by the presence of more than one O atom, an alkylarylether/aromatic group, and the absence of a C bonded to two O atoms, an alkene bond, and an S.
apple	The molecular property "apple scent" can be explained by the presence of an ethyl/ether O group, as well as the absence of an aromatic atom, hydroxy oxygen (OH), propyl fragment, and S. These structure-property relationships are very important for the property and explain why the molecule smells like apples.
apricot	The molecular property "apricot scent" can be explained by the presence of a CH2 group involved in a double bond, an atom bonded to a CH3CH2- group, and the lack of ether, N, and S groups.
aromatic	The molecular property "aromatic scent" can be explained by the presence of more than one O atom, an ether O/methyl group, and the absence of a C bonded to two O atoms, a methyl group, and an S. These structural attributes are all very important for the property.
balsamic	The molecular property "balsamic scent" can be explained by the presence of an atom in a ring, an aromatic atom, a cis double bond/aromatic group, and a C=O double bond. The lack of an S atom is also important in contributing to the property.
banana	The molecular property "banana scent" can be explained by the presence of a C=O double bond and a carboxylic acid derivative/methyl group, as well as the lack of an alkyne group, more than two oxygen atoms, and an S atom.
beefy	The molecular property "beefy scent" can be explained by the presence of an S atom, a hetero O/aromatic group, the lack of a pentanyl fragment, the lack of an O atom with at least two chain bonds, and the lack of an ether O/alkane group. These structure-property relationships are important for understanding the beefy scent of molecules.
berry	The molecular property "berry scent" can be explained by the presence of a C=O double bond and an ethyl/ether O group, as well as the absence of an N atom separated from an O atom by three bonds, more than one O atom, and multiple N atoms.
blackcurrant	The molecular property "blackcurrant scent" can be explained by its structure, which includes an S atom, a heteroatom separated from a CH2 group by two bonds, and ethyl. The lack of an S atom involved in a double bond and an atom bonded to another atom by an aromatic bond and bonded to an S by a non-aromatic bond are also important for the blackcurrant scent.
brandy	The molecular property "brandy scent" can be explained by its structure, specifically the presence of a C=O double bond, an O with at least two chain bonds, and the lack of more than three O atoms, more than two O atoms, and an S atom.
bread	The molecular property "bread scent" can be explained by the presence of multiple heterocyclic atoms, an S, and a C=O double bond. The absence of multiple methyl groups and a tertiary alcohol are also important for this property.

Continued on next page

Scent	Why the scent?
brothy	The molecular property "brothy scent" can be explained by the presence of hetero O/aromatic/primary carbon groups and an S atom, as well as the lack of an atom bonded to a CH3CH2- group, more than three O atoms, and an O atom.
burnt	The molecular property "burnt scent" can be explained by the presence of multiple O atoms that are involved in double bonds, an S atom, and multiple aromatic rings. The lack of an O atom and the lack of more than one O atom are also very important for this molecular property.
buttery	The molecular property "buttery scent" can be explained by the presence of more than one oxygen atom, a CH2 group separated from O by any two bonds, multiple O atoms involved in double bonds, and more than three O atoms. The absence of an S atom is also very important for this property.
cabbage	The molecular property "cabbage scent" can be explained by the presence of an S atom, disulfide/methyl group, and a heteroatom bonded to a methyl C. The lack of a CH2 group bonded to two neighbors by non-ring bonds and a methyl group are also important for this property.
camphoreous	The molecular property "camphoreous scent" can be explained by the presence of bridged rings/alkane groups and carbonyl groups, as well as the lack of a 6M ring, a CH2 group bonded to two neighbors by non-ring bonds, and an S. These structure-property relationships are very important for the molecular property.
caramellic	The molecular property "caramellic scent" can be explained by the presence of a carbonyl and more than one O atom, as well as the absence of multiple heteroatoms bonded to at least one H atom, more than three O atoms, and an S.
catty	The molecular property "catty scent" can be explained by the presence of an S atom, tertiary C, O bonded to a secondary C, and an oxymethylene group (-CH2O-). The lack of more than one CH2 group that is bonded to two neighbors, one of which is a heteroatom, is also important for this property.
chamomile	The molecular property "chamomile scent" can be explained by the presence of a michael acceptor/methyl group and a propyl fragment/ether O group, as well as the absence of two CH2 groups separated by any three or four bonds and the absence of an S.
cheesy	The molecular property "cheesy scent" can be explained by the presence of a C atom bonded to two O atoms, the lack of an O atom with at least two chain bonds, the lack of a butyl fragment/ether O group, the lack of more than one O atom, and the lack of an ether O/methyl group.
cherry	The molecular property "cherry scent" can be explained by the presence of an aromatic/methyl group and an aldehyde/aromatic group. The lack of two CH2 groups separated by any four bonds, a CH3 group separated from a CH2 group by any two bonds, and an S are all important for this property.
chicken	The molecular property "chicken scent" can be explained by the presence of an atom bonded to three other atoms, one of which is an S, as well as the lack of multiple methyl groups, an S, a heptanyl fragment, and an ether O/alkane group.

Continued on next page

Scent	Why the scent?
chocolate	The molecular property "chocolate scent" can be explained by the presence of multiple heterocyclic atoms, an atom in a ring, and a hetero N nonbasic/heteroaromatic/aromatic/primary carbon group. The lack of an iso-butyl/carboxylic ester group and an iso-butyl/ether O group are also important for this property.
cinnamon	The molecular property "cinnamon scent" can be explained by the presence of an oxygen atom, an alkene bond, and the absence of a 5M ring, more than one 6M ring, and an S atom.
citrus	The molecular property "citrus scent" can be explained by the presence of an oxygen atom, secondary carbon, and the lack of a C=O double bond and more than two oxygen atoms, as well as the lack of a sulfur atom.
cocoa	The molecular property "cocoa scent" can be explained by the presence of an oxygen atom, more than one oxygen atom, a lack of a charged atom/group, a lack of a sulfur atom and a lack of a carbon-oxygen double bond.
coconut	The molecular property "coconut scent" can be explained by the presence of an n-butyl group and a carbonyl group, and the lack of a CH2 group bonded to two neighbors (at least one of which is a heteroatom), a propyl fragment, and an S.
coffee	The molecular property "coffee scent" can be explained by the presence of an S atom, an aromatic ring, and the lack of an N atom bonded to at least one H, more than one 6M ring, and a secondary carbon/alkane group.
cognac	The molecular property "cognac scent" can be explained by the presence of more than one oxygen atom, an oxygen atom, and the lack of a C=O double bond, an alkyne group, and an S atom.
coumarinic	The molecular property "coumarinic scent" can be explained by the presence of an oxygen atom, the absence of a propyl fragment, the absence of an aromatic/primary carbon group, the absence of an alkene bond, and the absence of an sulfur atom.
creamy	The molecular property "creamy scent" can be explained by the presence of a carbonyl and an oxygen, as well as the absence of an S-heterocycle, heterocyclic, and a CH2 group bonded to two neighbors, at least one of which is a heteroatom.
cucumber	The molecular property "cucumber scent" can be explained by the structure of the molecule, with the presence of a secondary carbon, an O atom, and the absence of more than two O atoms, multiple methyl groups, and an S atom being the most important factors.
dairy	The molecular property "dairy scent" can be explained by the presence of a C atom bonded to two O atoms, an O atom that is separated from another O atom by any 3 bonds, and the lack of a CH2 group bonded to two neighbors, at least one of which is a heteroatom, as well as the lack of more than one or two O atoms.
dry	The molecular property "dry scent" can be explained by the presence of a C atom bonded to two O atoms, as well as a C=O double bond. The lack of more than one O atom, more than two O atoms, and an S atom are also very important for this property.

Continued on next page

Scent	Why the scent?
earthy	The molecular property "earthy scent" can be explained by the presence of an alkene bond and a hetero N nonbasic/aromatic group, as well as the lack of a butyl fragment/ether O group, more than one O atom, and an ester.
ethereal	The molecular property "ethereal scent" can be explained by the presence of an O atom, an ether O/methyl group, and the absence of an alkyne group, two CH2 groups separated by any four bonds, and an S atom.
fatty	The molecular property "fatty scent" can be explained by the presence of a heptanyl fragment, two CH2 groups separated by four bonds, and a C=O double bond, as well as the lack of more than one or two O atoms.
fermented	The molecular property "fermented scent" can be explained in part by the presence of an atom bonded to two methyl groups, a C-O single bond, and the absence of a carbonyl group, a CH3 group separated from a CH2 group by any four bonds, and a C=O double bond.
fishy	The molecular property "fishy scent" can be explained by the presence of an alkene bond and an N atom, as well as the absence of more than one O atom, more than two O atoms, and an ether O/alkane group.
floral	The molecular property "floral scent" can be explained by the presence of an O atom bonded to a secondary C, more than two methyl groups, an atom in a ring, the absence of a C bonded to two O atoms, and the absence of an S atom. These structure-property relationships are all important for the presence of the floral scent.
fresh	The molecular property "fresh scent" can be explained in terms of its structure, specifically the presence of an alkene bond and an O atom which are both very important for the property. The lack of an atom in a ring, an atom bonded to a CH3CH2- group, and an S atom are also very important for the property.
fruity	The molecular property "fruity scent" can be explained by the presence of a carbonyl, a C=O double bond, and an oxygen atom with at least two chain bonds. The lack of a heptanyl fragment and an S atom also contribute to this property.
garlic	The molecular property "garlic scent" can be explained by the presence of sulfur atoms, disulfide/methyl and disulfide/alkene groups, and atoms bonded to three other atoms, one of which is a sulfur atom. The absence of multiple methyl groups is also a contributing factor.
gasoline	The molecular property "gasoline scent" can be explained by looking at the structural attributes of the molecule, such as the presence of a methyl group, a CH2 group bonded to two neighbors by non-ring bonds, the absence of a CH2 group separated from O by two bonds, the absence of an ether O/methyl group, and the absence of an alkyne group. These attributes are all very important for the property.
grape	The molecular property "grape scent" can be explained by the presence of a primary amine, an O bonded to a secondary C, and a lack of a butyl fragment/ether O group, a heteroatom bonded to at least two CH2 carbons, and two CH2 groups separated by any four bonds.

Continued on next page

Scent	Why the scent?
grapefruit	The molecular property "grapefruit scent" can be explained in part by the presence of an S atom, an O atom bonded to a secondary C atom, and an oxymethylene group (-CH2O-). The absence of more than two O atoms and more than one O atom also contributes to this property.
grassy	The molecular property "grassy scent" can be explained by the presence of an oxygen atom, the lack of hetero N nonbasic/ketone/aromatic groups, multiple methyl groups, a C=O double bond and sulfur. These attributes of structure are the most important for the property.
green	The molecular property "green scent" can be explained by the presence of an O atom, the lack of a carbonyl, heptanyl fragment, C=O double bond, and S atom. These are all very important for this property.
hay	The molecular property "hay scent" can be explained by the presence of a heterocycle and a C=O double bond, as well as the absence of an N, more than two O atoms, and an S.
hazelnut	The molecular property "hazelnut scent" can be explained by the presence of hetero N nonbasic/aromatic/primary carbon groups, ethyl/aromatic groups, multiple methyl groups, and multiple N atoms. The lack of more than one O atom could have a relevent effect on the property.
herbal	The molecular property "herbal scent" can be explained by the presence of a C=O double bond, C-O single bond, atoms in a ring, more than two methyl groups, and the absence of more than one O atom.
honey	The molecular property "honey scent" can be explained by the presence of an ester, an atom in a ring, an aromatic atom, and a C bonded to two O atoms. Notably, the absence of an S atom is also very important for this property.
horseradish	The molecular property "horseradish scent" can be explained by the presence of an S atom, isothiocyanate, an atom bonded to both S and N, and a CH2 group bonded to two neighbors, at least one of which is a heteroatom. The lack of hetero O/aromatic groups is also an important factor for the property.
jasmine	The molecular property "jasmine scent" can be explained by the presence of more than two methyl groups, an atom in a ring, an oxymethylene group (-CH2O-), and the absence of an S and a CH2 group bonded to two neighbors, at least one of which is a heteroatom.
ketonic	The molecular property "ketonic scent" can be explained by the presence of an O bonded to a secondary C and iso-propyl, as well as the lack of a primary carbon, an alkene bond, and an S. These structural attributes are especially important for this property.
leafy	The molecular property "leafy scent" can be explained by the presence of more than one O atom, an alkene bond, and an O atom in the molecule, and the absence of a C=O double bond and an S atom.
leathery	The molecular property "leathery scent" can be explained by the presence of phenol, tert-butyl/aromatic groups, multiple methyl groups, and a hydroxy oxygen (OH). Of these, the presence of phenol is the most important for the leathery scent.

Continued on next page

Scent	Why the scent?
lemon	The molecular property "lemon scent" can be explained by the presence of more than two methyl groups, an alkene/methyl group, and the absence of more than two O atoms, an atom at an aromatic/non-aromatic boundary, and an S atom.
malty	The molecular property "malty scent" can be explained by the presence of an atom bonded to two methyl groups, more than one O atom, and the lack of a C=O double bond, an S atom, and an alkyne group.
meaty	The molecular property "meaty scent" can be explained by the presence of an S atom and the absence of an alkene, an O atom, a methyl group, and a CH2 group bonded to two neighbors by non-ring bonds.
medicinal	The molecular property "medicinal scent" can be explained by the presence of phenol and the presence of a hydroxy oxygen (OH). The lack of an O with at least two chain bonds, the lack of more than two O atoms, and the lack of an S are all important structure-property relationships for this property.
melon	The molecular property "melon scent" can be explained by the presence of an O atom and an alkene bond, as well as the absence of multiple methyl groups, more than two O atoms, and an S atom.
metallic	The molecular property "metallic scent" can be explained by the presence of an aromatic atom and a C=O double bond, and the lack of an S-heterocycle and more than one O atom. These structure-property relationships are very important for the property.
milky	The molecular property "milky scent" can be explained by the presence of a molecule containing a C atom bonded to two O atoms. The lack of an alkyne group, tert-butyl/aromatic group, propyl fragment, and alkene bond are all important for the property.
mint	The molecular property "mint scent" can be explained by the presence of a carbonyl, a methyl group, and a C bonded to two O atoms, and the absence of an S atom and an aromatic atom. These are the most important structure-property relationships for this property.
mushroom	The molecular property "mushroom scent" can be explained by the presence of an oxygen atom bonded to a secondary carbon atom, as well as the presence of a dialkylthioether. The lack of a butyl fragment/ether oxygen group and more than one or two oxygen atoms was also important for the property.
musk	The molecular property "musk scent" can be explained by the presence of an atom in a ring and secondary carbon, and the absence of an aromatic atom, an alkyne group, and a 8M - 14M Ring.
musty	The molecular property "musty scent" can be explained by the presence of an O atom, a heteroatom bonded to a methyl C, and the lack of an atom bonded to O by a non-aromatic bond and bonded to another atom aromatically, multiple methyl groups, and an S atom.
nutty	The molecular property "nutty scent" can be explained by the presence of a hetero N nonbasic/aromatic group, the absence of a propyl fragment, multiple aromatic rings, two CH2 groups separated by any three bonds, and an alkene bond.

Continued on next page

Scent	Why the scent?
oily	The molecular property "oily scent" can be explained by the presence of a heptanyl fragment, an oxygen with at least two chain bonds, and a hexanyl fragment/ether oxygen group. The absence of more than two oxygen atoms and an S atom also contributes to the property.
onion	The molecular property "onion scent" can be explained by the presence of an S and a disulfide/methyl group. The lack of an N, multiple methyl groups, and a single methyl group are also important factors in contributing to the property.
orange	The molecular property "orange scent" can be explained by the presence of two CH2 groups separated by any four bonds and an O atom, as well as the lack of an S atom, an O atom bonded to a secondary C atom, and more than two O atoms.
orris	The molecular property "orris scent" can be explained by the presence of tertiary carbon, and the absence of pentanyl fragment, N, more than two O atoms, and S.
peach	The molecular property "peach scent" can be explained by the presence of an alkene bond, a carbonyl group, a five-membered ring, a carbon atom bonded to two oxygen atoms, and the absence of multiple heterocyclic atoms. These structural features are all important for the property.
pear	The molecular property "pear scent" can be explained by the presence of a C=O double bond, an ethyl/ether O group, and the lack of an atom in a ring, an S atom, and more than two O atoms. These structural attributes are all very important for this property.
phenolic	The molecular property "phenolic scent" can be explained by the presence of phenol, an aromatic/methyl group, and a hydroxy oxygen (OH). The lack of a C=O double bond and an O with at least two chain bonds are also important for the property.
pine	The molecular property "pine scent" can be explained by the presence of an atom in a ring and the lack of an alkene/secondary carbon group, a C=O double bond, a charged atom/group, and an S in the molecule.
pineapple	The molecular property "pineapple scent" can be explained by the presence of an ester, ethyl/ether O group, alkene/ether O group, and a C=O double bond. The absence of an aromatic atom is also significant for this property.
plum	The molecular property "plum scent" can be explained by the presence of a carbonyl, a C=O double bond, an oxymethylene group (-CH2O-), and the lack of an S atom and multiple N atoms. These attributes of structure are important for the property.
popcorn	The molecular property "popcorn scent" can be explained by the presence of a carbonyl and a hetero N nonbasic/aromatic group. The absence of a carboxylic acid derivative/methyl group, an amine (NH2) group, and a heteroatom bonded to a methyl C was also important for the property.
potato	The molecular property "potato scent" can be explained by the presence of sulfur, hetero nitrogen nonbasic/heteroaromatic/aromatic groups, dialkylthioether, and ethyl/aromatic groups in the molecule. The lack of an acetal-like group is also important for the property.

Continued on next page

Scent	Why the scent?
pungent	The molecular property "pungent scent" can be explained by the presence of an isothiocyanate group, an oxygen atom, and an atom bonded to both sulfur and nitrogen, as well as the absence of an nitrogen atom seperated from an oxygen atom by either three or four bonds.
radish	The molecular property "radish scent" can be explained by the presence of an S atom, dialkylthioether, and the absence of dialkylether/ether O group, ethyl/alkene group, and more than two O atoms.
ripe	The molecular property "ripe scent" can be explained by the presence of an S atom, an N atom, and the absence of hetero O/aromatic groups, an N atom separated from an O atom by 3 bonds, and an N atom separated from an O atom by 4 bonds.
roasted	The molecular property "roasted scent" can be explained by the presence of an S atom, a hetero N nonbasic/aromatic group, and an atom at a ring/chain boundary. The absence of an N bonded to at least one H, and a CH2 group bonded to two neighbors by non-ring bonds were also important in contributing to the molecular property.
rose	The molecular property "rose scent" can be explained by the presence of a C=O double bond and more than two methyl groups, and the absence of more than one O atom, more than two O atoms, and an S atom. These structural features are all very important for the property.
rum	The molecular property "rum scent" can be explained by the presence of an oxygen atom with at least two chain bonds, the absence of carboxylic acid, two CH2 groups separated by any four bonds, more than three oxygen atoms, and the absence of sulfur. These structure-property relationships are very important for understanding the molecular property.
savory	The molecular property "savory scent" can be explained by the presence of an S atom, an atom bonded to three other atoms, one of which is an S, and the absence of more than two O atoms, multiple methyl groups, and a methyl group.
sharp	The molecular property "sharp scent" can be explained by the presence of an alkene bond and a C=O double bond, as well as the absence of more than three O atoms and more than two O atoms, and the absence of an S atom.
smoky	The molecular property "smoky scent" can be explained by the presence of phenol, aromatic/methyl groups and hydroxy oxygen (OH) in the structure. The lack of carbonyl and a C=O double bond is also important for this property.
solvent	The molecular property "solvent scent" can be explained by the presence of an oxygen atom, oxygen atoms separated by three bonds, an alkyne group, two CH2 groups separated by four bonds, and an sulfur atom in the molecule.
sour	The molecular property "sour scent" can be explained by the presence of a C atom bonded to two O atoms, an atom bonded to a CH3CH2- group, and the lack of a CH3 group separated from a CH2 group by any four bonds, more than two O atoms, and more than one CH2 group that is bonded to two neighbors, one of which is a heteroatom.

Continued on next page

Scent	Why the scent?
spicy	The molecular property "spicy scent" can be explained by the presence of an ether O/methyl group, an alkene bond, and a heteroatom bonded to a methyl C. The lack of multiple N atoms and an N atom are also very important for this property.
strawberry	The molecular property "strawberry scent" can be explained by the presence of a C-O single bond and an O with at least two chain bonds. Additionally, the lack of alkylthiol, an O bonded to a secondary C, and an S is important for this property.
sulfurous	The molecular property "sulfurous scent" can be explained by the presence of an S atom and a hydroxy oxygen (OH) group, as well as the lack of a CH2 group bonded to two neighbors by non-ring bonds, an O atom, and a methyl group.
sweet	The molecular property "sweet scent" can be explained by the presence of a C=O double bond and the absence of a CH3 group separated from a CH2 group by any four bonds, a heptanyl fragment, an alkene bond, and an S.
tea	The molecular property "tea scent" can be explained by the presence of a C=O double bond and carbonyl, and the lack of an S atom and more than two or three O atoms.
tobacco	The molecular property "tobacco scent" can be explained by the presence of multiple aromatic rings and a C atom bonded to two O atoms. The lack of an 8M - 14M ring, more than two O atoms, and an S atom is also important for the property.
tomato	The molecular property "tomato scent" can be explained by the presence of an S atom, primary alcohol, and the lack of an aromatic and propyl fragment, as well as a CH3 group separated from a CH2 group by four bonds.
tropical	The molecular property "tropical scent" can be explained by the presence of an S atom and a C atom bonded to two O atoms, as well as an alkene bond. The lack of an O bonded to a secondary C atom and an aromatic atom also contributes to this property.
vanilla	The molecular property "vanilla scent" can be explained by the presence of more than one oxygen atom, an ether O/methyl group, more than two oxygen atoms, and an atom in a ring. The lack of a carbon atom bonded to two oxygen atoms is also important for the property.
vegetable	The molecular property "vegetable scent" can be explained by the presence of an S atom and an O atom in the molecule, as well as the absence of an S atom involved in a double bond, an ester, and a C=O double bond.
violet	The molecular property "violet scent" can be explained by the presence of a C=O double bond and an alkyne group, as well as the lack of an S atom, more than three O atoms, and more than two O atoms.
warm	The molecular property "warm scent" can be explained by the presence of an ether O/methyl group, an aromatic atom, and more than one O atom. The lack of a hydroxy oxygen (OH) and an S atom also contribute to the property.
waxy	The molecular property "waxy scent" can be explained by the presence of an n-heptanyl and heptanyl fragment, and the lack of a CH2 group bonded to two neighbors (at least one of which is a heteroatom), more than two O atoms, and an S.

Continued on next page

Scent	Why the scent?
winey	The molecular property "winey scent" can be explained by the presence of a C-O single bond, a C=O double bond, and an O with at least two chain bonds. The lack of a 6M Ring and an S is also important for this molecular property.
woody	The molecular property "woody scent" can be explained by the presence of a C=O double bond, an atom in a ring, and the lack of more than one O atom, an O atom, and an S atom. These attributes are all very important for this property.