Smell and Tell: The emergence of olfactory expertise in perfumery students

Authors: Saive Anne-Lise[1,2], Plailly Jane[3], Chambaron Stéphanie[4], Hourri Youness[1], Monin Carla[1], Nhouchi Zeineb[5], Belay Justine[5], Chalut Pauline[5], Hanaei Farnaz[5], Vallet Nadine[5]

Affiliations

[1]Lyfe Institute Research and Innovation Center, Écully, France

[2]Unique Center (Québec Neuro-AI Research Center), Montréal, Québec, Canada

[3]Centre de Recherche en Neurosciences de Lyon (Inserm U1028 - CNRS UMR5292 - UCBL) – CMO Team, Bron, France

[4] Centre des Sciences du Goût et de l'Alimentation, CNRS, INRAE, Institut Agro, Université de Bourgogne, F-21000 Dijon, France.

[5]Institut Supérieur International du Parfum, de la Cosmétique et de l'Aromatique alimentaire (ISIPCA), F-78000 Versailles, France

ORCID-IDs:

Anne-Lise Saive, https://orcid.org/0000-0001-7352-0312

Jane Plailly, https://orcid.org/0000-0003-4189-1631

Stéphanie Chambaron, https://orcid.org/

Zeineb Nhouchi, https://orcid.org/0000-0003-3509-373X

Nadine Vallet, https://orcid.org/0000-0001-8943-8620

Correspondence should be sent to Anne-Lise Saive, alsaive@institutlyfe.com

Abstract

Developing olfactory expertise is essential in professions like perfumery, where the ability to describe, categorize, and conceptualize odors is critical. This study investigates how academic training during a 1.5-year program at a perfumery school (ISIPCA) shapes olfactory expertise of perfumery students. Forty students were assessed at three time points, focusing on odor description, evocation, recognition, discrimination, and categorization tasks. Results show that training significantly enhanced language abilities related to odor description and categorization. Students developed a richer and more precise vocabulary to characterize

odors, aligning more closely with expert's terminology and contributing to the formation of a shared olfactory lexicon. Semantic similarity within and between students, as well as with expert references, increased, emphasizing the importance of consistent language use in expertise development. Advanced natural language processing and machine learning tools revealed that the richness of verbal descriptions and semantic similarity were strong predictors of expertise acquisition. In contrast, improvements in non-verbal tasks, such as odor discrimination and recognition, were more limited, suggesting that perceptual abilities may require more extensive training or specialized methods. Building on these results, we propose potential enhancements to olfactory training including reinforced language practice, mental imagery exercises, and sensory discrimination tasks, along with personalized training strategies. These findings highlight the central role of language in the emergence of olfactory expertise and the importance of computational methods for optimizing training programs and advancing educational practices in olfactory science.

Keywords: olfactory expertise, learning, perfume training, natural language processing, machine-learning

1. Introduction

Olfactory expertise is a distinctive and highly specialized skill, cultivated through intensive, deliberate practice and targeted training. This expertise is foundational in professions such as perfumers and sommeliers, where individuals demonstrate extraordinary abilities to recognize, discriminate, and describe a wide array of complex odor profiles. For perfumers, this expertise involves not only mastering the intricate combination of thousands of aromatic compounds to design sophisticated fragrances but also the exceptional ability to mentally construct and evaluate scent compositions without physical stimuli.

Achieving olfactory expertise requires both perceptual learning and the development of higher-level cognitive processes (for reviews, Parr, 2019; Royet et al., 2013). While olfactory training has been shown to improve sensitivity, these effects are often modest and odor-specific (Haehner et al., 2013; Al Aïn et al., 2019; Zambom-Ferraresi et al., 2021). Olfactory expertise is primarily characterized by enhancements in cognitive functions, including conceptualizing, categorizing, and memorizing olfactory information (Solomon, 1997; Hughson and Boakes, 2001; Plailly et al., 2012). Short-term training studies also support the role of language in olfactory expertise. Training laypeople to consistently label odors improves their ability to learn and categorize them, resulting in higher accuracy and faster learning (Fournel et al., 2017; Vanek et al., 2021). This suggests a causal link between verbal description and the formation of odor categories, which is particularly noteworthy in the olfactory domain due to the well-documented challenges of articulating smells through language (Herz, 2005; Olofsson and Gottfried, 2015). However, while short-term training enhances specific skills like odor labeling, its benefits may not extend to overall olfactory memory or more generalized tasks, as shown by Lesschaeve & Issanchou (1996) and Al Aïn et al. (2019). Together, these findings highlight the nuanced and task-specific nature of olfactory expertise development.

In professional perfumers, extensive training leads to superior olfactory mental imagery and richer, more precise verbal descriptions of odors. Plailly et al. (2012) demonstrated that perfumers outperform novices in generating odor mental imagery. This enhanced imagery ability is highly specialized and does not transfer across sensory modalities, suggesting the domain-specific nature of olfactory expertise (Bensafi et al., 2017). Perfumers use more detailed and precise language when describing odors, focusing on chemical and olfactory qualities rather than general hedonic terms, reflecting a shift in semantic odor processing with expertise (Sezille et al., 2014).

Similarly, wine experts build their olfactory expertise through improved discrimination, identification, and recognition of wine-related odors. Studies have shown that wine experts have a modest yet significant advantage in naming smells and flavors within their domain, attributable to both perceptual and linguistic training (Croijmans and Majid, 2016). Olfactory training has been shown to enhance odor identification abilities, even in novices. For instance, short-term training over five days allowed participants to perform as well as sommeliers in identifying single odorants (Poupon et al., 2018). However, this improvement did not extend to mixtures of two or more odorants, highlighting a limitation of short-term training for complex olfactory tasks. In contrast, other research has demonstrated that targeted olfactory training can improve novices' ability to detect individual components within a mixture, suggesting that with focused and potentially longer training, individuals can develop the skills to analyze odor mixtures (Morquecho-Campos et al., 2019). Wine experts also exhibit superior memory for wine odors, a skill that is highly specific to their domain of expertise and reflects the effects of specialized training (Parr et al., 2002; Croijmans et al., 2021). Training enhances their mental imagery for wine; the vividness of wine imagery increases only after specialized training, underscoring the role of targeted practice in building olfactory expertise (Croijmans et al., 2020). Additionally, wine experts use richer and more concrete language with distinct patterns

when describing wines, which facilitates the translation of nuanced sensory inputs into detailed verbal descriptions (Sezille et al., 2014; Croijmans et al., 2024).

Specialized training in perfumery and wine tasting induces both functional and structural brain changes associated with perceptual and cognitive enhancement. Functional imaging studies reveal that professional perfumers exhibit decreased activation in olfactory and memory regions during odor imagery tasks, indicating enhanced efficiency and reduced cognitive effort (Plailly et al., 2012). Structural brain changes include increased gray matter volume in areas surrounding the olfactory sulcus, such as the gyrus rectus and medial orbital gyrus, with this reorganization positively correlated with experience, even counteracting the effect of age, reflecting the brain's adaptability to olfactory expertise (Delon-Martin et al., 2013). Neural differences are observed in wine experts compared to novices. Wine sommeliers show heightened activation in brain regions responsible for gustatory-olfactory integration and high-level cognitive processing (Castriota-Scanderbeg et al., 2005; Pazart et al., 2014; Banks et al., 2016; Sreenivasan et al., 2017). Structural adaptations are also evident, with sommeliers exhibiting increased volume in sensory olfactory regions, including the olfactory bulb, right insula, and bilateral entorhinal cortex (Banks et al., 2016; Filiz et al., 2022). Enhanced connectivity between sensory and semantic networks further supports the translation of nuanced sensory inputs into detailed verbal descriptions (Carreiras et al., 2024).

Overall, these findings raise important questions about the extent to which olfactory training generalizes—whether its benefits are limited to specific odor labeling or extend to broader linguistic and cognitive processes associated with olfaction. Furthermore, little is known about the gradual emergence of expertise over time. Although Filiz et al. (2022) investigated how sommelier students develop expertise through a rigorous 1.5-year program involving extensive sensory training in educational settings, most studies still focus primarily on comparisons between experts and novices rather than the process of expertise formation.

The present study addresses these gaps by investigating the development of olfactory expertise in perfumery students over a 1.5-year training program. We assessed a wide range of cognitive competencies through tasks including odor categorization, odor description, odor evocation, odor recognition, and odor discrimination. In addition, we performed a detailed analysis of the language used by participants to describe odors using natural language processing (NLP) and machine learning (ML) tools. This research aims to provide a comprehensive understanding of how academic training enhances—or fails to enhance—sensory perception, memory, and related cognitive abilities, contributing to advancements in olfactory education and professional expertise.

2. Materials and methods

2.1. Participants

The olfactory experiments were conducted at ISIPCA (International School of Perfume, Cosmetics Products, and Food Flavor Formulation) between 2018 and 2021 by ISIPCA faculty and staff, integrating the experiments into the educational curiculum. Since the study was embedded within the school's pedagogical framework and conducted by instructors rather than external researchers, no ethical committee approval was required. The study involved 40 ISIPCA students, 35 women and 5 men, with an average age of 17.475 years (SD = 1.585). The gender distribution reflected the demographics of the school's student cohorts, a factor beyond our control.

2.2.Olfactory training

Participants were enrolled in the “Specialized Laboratory Technician” diploma, a two-year full-time program. This curriculum covered foundational knowledge in perfumery, cosmetics, and

food flavors, with olfaction as a major component. The training emphasized olfactory perception, description, and categorization through structured courses and hands-on sessions.

Perfumery Courses – Students studied 120 raw materials over the two-year program. Each raw material was smelled blindly, described verbally, and classified into an olfactory family based on Jean-Noël Jaubert's field of odors (1995). Students assigned three descriptive terms and created detailed sheets for each raw material. For homework, students assessed the volatility of raw materials using scent strips and containers prepared during class. Sessions lasted 3 hours and 30 minutes, including a minimum 20-minute break. The raw materials were organized by olfactory families to facilitate comparative learning, gradually introducing more intense odors as the course progressed. A final exam evaluated students' knowledge. They were required to identify 10 raw materials presented blindly, specifying the olfactory family, descriptors, volatility, and handling characteristics.

Food Flavors Courses – The training approach for food flavors followed the same principles as perfumery but included an additional retro-nasal testing step. Raw materials were diluted in water, tasted, and analyzed both ortho and retro-nasally. Students discussed descriptors until reaching a consensus for qualitative and quantitative evaluations. In the first year, sessions covered 6–8 raw materials, increasing to 10 raw materials in the second year. Each session focused on two distinct families (e.g., fruity and toasted). Evaluations progressed from less to more persistent materials. For example, fruity materials with minimal afterglow were evaluated first, followed by more persistent toasted materials. Sessions included one or two breaks. Exams, held two to three times per year, required students to identify 10 raw materials presented blindly, specifying descriptors, families, and persistence characteristics.

2.3.Olfactory Stimuli

The odorants consisted of monomolecular compounds, essential oils, homemade compositions, and perfumes, all sourced from the olfactory base-sample used in ISIPCA's courses. The monomolecular compounds and essential oils were diluted in alcohol to concentrations appropriate for olfaction, while the perfumes were pure. All stimuli were presented in brown 15ml bottles coded with a unique three-digit number. A total of 92 independent samples were selected, each assigned to specific tests: (i) 12 for the evocation and free description tasks, (ii) 6 odorants for the triangle test, (iii) 60 for the olfactory recognition test, and (iv) 14 for the categorization test (Table S1).

2.4.Visual stimuli in the visual recognition task

In the visual recognition task, 120 images were divided into six sets of 20 images each. These images were sourced from previously published datasets, including the DMS48 test (Barbeau et al., 2004). Each participant viewed a unique subset of images across three time points, ensuring that different images were presented in each session. This design minimized familiarity effects and provided a more accurate assessment of recognition memory over time. While unrelated to participants' olfactory training, this task allowed for the assessment of general recognition abilities in a domain not influenced by students' specialized curriculum.

2.5.Experimental tasks

All tests were carried out in the ISIPCA sensory laboratory. Testing and data collection took place in standard sensory booths, with white lighting and controlled temperature ($20 \pm 2°C$) and airflow conditions. The expert students participated at three different measurement times: T0 (on arrival at the school, before olfactory training started), T1 (6 months after T0), and T2 (18 months after T0, 12 months after T1).

To assess the effect of academic olfactory training on olfactory capacities, different tasks were conducted over four half-days on four separate days at each measurement time:

- **Day 1:** Odor Description, Evocation, and Discrimination tasks
- **Day 2:** Visual and Olfactory Encoding part of the recognition tasks
- **Day 3:** Visual and Olfactory Retrieval part of the recognition tasks
- **Day 4:** Odor Categorization and Categories Description task

These tasks were designed to evaluate various aspects of olfactory perception and cognitive processing across the three measurement times and were conducted as described below:

Odor description and evocation: In the description task, the objective was to generate terms in writing describing the odor by answering the question, "How would you objectively describe this odor?". In the evocation task, students were asked, "What spontaneously comes to mind when you smell this odor?" and were invited to complete the sentences: "This scent makes me want to..." and "This scent reminds me of...". For both tasks, each student was presented with two different olfactory stimuli at each measurement time, with up to two minutes allocated to complete both tasks for each odor. The selection of stimuli was randomized for each student but ensured that student evaluated each odor only once.

Odor discrimination: Students were presented with three strips in a randomized order: two strips contained the same odor, and one strip contained a different odor. They were asked to identify the odd odor. This task involved three trials with the following sets of stimuli: (i) Flowerbomb (*2) and La Vie Est Belle (*1), (ii) La Nuit Trésor (*2) and Black Opium (*1), and (iii) L'Homme (*2) and Pink Grapefruit (*1). The same stimuli were used across all measurement times.

Odor recognition*:* The test was conducted over two consecutive days. On the first day, students were presented with ten target olfactory stimuli, each labeled with a unique three-digit code

specific to Day 1 (Encoding) and different from Day 2 (Recognition). They were instructed to smell each stimulus and try to memorize them. On the second day, students were presented with a set of 20 olfactory stimuli, including the 10 target stimuli from the first day and 10 new stimuli (distractors). They were given 5 seconds to smell each stimulus and then indicate whether they had encountered the odor on the first day. In total, 60 odorants were used across all students and time measurements, with different subsets of odorants assigned for each time measurement per student. Responses were recorded using Fizz software during the recognition phase on Day 2, with no maximum response time set.

Visual recognition: This task followed the same paradigm as the odor recognition task, but with visual stimuli replacing olfactory stimuli. On Day 1 (Encoding), participants were presented with a series of pictures using a timed PowerPoint presentation, with each image displayed for 5 seconds. On Day 2 (Recognition), participants' responses were recorded using Fizz software, with no maximum response time set. The inclusion of a visual recognition task served as a control to assess participants' baseline cognitive abilities in a domain unrelated to their training. This distinction allows us to isolate the impact of olfactory training on performance and ensure that any differences observed are specific to the sensory modality of interest, rather than reflecting general improvements in recognition or memory processes.

Odor categorization task and categories description: This task aimed at investigating odor learning through categorization. Students were presented with a total of 15 odorants in a two-step test. In the first step, they were asked to group the odors they perceived as harmonious together, forming multiple groups ranging from a minimum of 2 to a maximum of 14. In the subsequent step, students explained their assessments by providing descriptive terms that capture the common attributes of odors within each group. One participant did not complete the task at T2 and therefore was removed from the analyses of this task.

2.6.Data analyses

Odor description and evocation: The verbal responses from both tasks were first independently preprocessed by removing non-alphanumeric characters and converting the text to lowercase before tokenization into individual words and eliminating stop words to normalize the data. The cleaned tokens were then assigned part-of-speech tags (*e.g.,* nouns, adjectives, verbs) using a SpaCy's French language model (`fr_core_news_sm`). Nouns were analyzed as reprensentations of olfactory sources, while adjectives were identified as descriptors of the olfactory sources. The total number of words (including nouns, adjectives, adverbs, and verbs) was also counted as a general proxy for verbal richness. To validate the accuracy of the automated part-of-speech tagging, a manual, exhaustive check was performed, ensuring the correct classification of tokens and refining the data where necessary. After tagging, tokens were lemmatized using SpaCy to reduce words to their root forms, ensuring consistency in lexical analysis. Lexical diversity was then measured using two complementary measures: type-token ratio (TTR) and cosine similarity ($cos\theta$) to capture both diversity and alignment in language use.

TTR is a measure of lexical diversity that compares the number of unique lemmas (types) to the total number of words (tokens) in a text ($TTR = \frac{Types}{Tokens}$). This metric provides insight into the variety of vocabulary used in the responses, serving as an important indicator of linguistic complexity and richness. TTR ratios were computed within each participant's response as well as across students' responses to assess the lexical diversity of the verbal responses both within and between students.

$cos\theta$ measures the semantic similarity between words represented in vector embeddings. A pre-trained French model from FastText (https://github.com/facebookresearch/fastText/) was used to provide a vectorial representation of word meanings. Cosine angle was then used to measure the cosine similarity between words vectors $= \frac{A.B}{\|A\|\|B\|}$, offering a measure of how similar two

words are in terms of their direction in the embedding space. We defined and calculated several specific similarity metrics, which were averaged across comparisons to generate a single similarity value per participant or odor:

- *Similarity Within Student*: This metric evaluates how similar the words used by a single participant were to each other, providing insights into the consistency of language within individual responses.
- *Similarity Between Students*: This metric compares the vocabulary used by different students, to identify common themes and vocabulary.
- *Similarity with an Expert Teacher*: This metric assessed how closely student language aligned with the target language, by comparing student responses to a reference set of words representing an expert teacher's vocabulary,.
- *Similarity by Odor*: This metric groups words by the odors they described, analyzing how consistently students described the same odor.

Odor and visual recognition tasks: Recognition memory performance was assessed using parameters derived from signal detection theory (Lockhart and Murdock, 1970). Four response categories were defined based on the experimental conditions (target or distractor items) and the students' behavioral responses (yes or no): Hit and Miss occurred when the target items were correctly recognized or incorrectly rejected, respectively, and correct rejection (CR) and false alarm (FA) occurred when the distractor items were correctly rejected or incorrectly recognized, respectively. Two parameters were calculated from the Hit and FA scores: a memory score (d'L) and a response bias score (CL). These scores were determined as follows:

$$d'L = ln\left(\frac{HR(1-FR)}{FR(1-HR)}\right)$$

$$CL = -0.5(ln\left(\frac{HR(1-HR)}{FR(1-FR)}\right))$$

Here, HR represents the Hit rate [$\frac{Hit+0.5}{N1+1}$], FR represents the false alarm rate [$\frac{FA+0.5}{N2+1}$], and N1 and N2 represent the number of target and distractor odorants, respectively, for which the students provided a response. Memory scores can be positive or negative, indicating students' good or poor performance in recognizing target items and rejecting distractor items, respectively. A *d'L* score of 0 indicates chance-level performance, where students cannot reliably distinguish between target and distractor items. Response bias scores indicate three individual attitudes: conservative (tending to respond "no" to an item), neutral (responding "yes" or "no" with equal probability), or liberal (tending to respond "yes"), with positive, neutral, or negative values, respectively (Snodgrass and Corwin, 1988).

Categorization task and Categories verbalization*:* To analyze the categorization task, we computed the number of groups created by students, as well as the agreement between cluster assignments made by students and an expert teacher. To evaluate the consistency and reliability of the clustering results across time points, two metrics were employed: Cohen's Kappa (κ) and Adjusted Rand Index (ARI). Cohen's Kappa measures the inter-rater agreement for categorical items, correcting for the agreement that could happen by chance. It is defined as $\kappa = \frac{po-pe}{1-pe}$ where p_o is the observed agreement and p_e is the expected agreement by chance. The Adjusted Rand Index measures the similarity between two data clusterings, adjusting for the chance grouping of elements. It is defined as $ARI = \frac{RI-E[RI]}{(RI)-E[RI]}$ where RI is the Rand Index, which measures the percentage of decisions that are correct, and $E[RI]$ is the expected value of the Rand Index. These metrics were calculated for each participant by comparing their cluster assignments with those of all other students and with the reference (expert teacher) at each time point.

The descriptions of each category created by students were analyzed using part-of-speech tagging, TTR scores and $cos\theta$ metrics as in the evocation and description task (see the extensive presentation of these scores above).

Odor discrimination: For each trio of odors, we determined whether the responses were correct or incorrect. These responses were then summed for each participant at each time point to yield a global score, referred to as the odor discrimination score. This score enabled to evaluate overall odor discrimination performance over time, with a random level baseline set at 33%, indicating chance-level performance.

Temporal dynamic of learning: In the final step of the analyses, we created the performance profile of each participant at each time measurement by combining their scores in all metrics significantly modulated by time. These profiles were normalized independently for each metric to ensure meaningful comparisons between metrics. We then computed the Euclidean distances between individual performance profiles across different time points (T0 vs. T1, T1 vs. T2, and T0 vs. T2) to quantify the progress in performance over time. For example, the Euclidean distance between participant profiles at time T0 and T1 containing three metrics (m1, m2, m3) is given by the formula:

$$d(T0,T1) = \sqrt{(m1_{T0} - m1_{T1})^2 + (m2_{T0} - m2_{T1})^2 + (m3_{T0} - m3_{T1})^2}$$

where $m1_{T0}$, $m2_{T0}$, $m3_{T0}$ and $m1_{T1}$, $m2_{T1}$, $m3_{T1}$ are the values of the three metrics at time points T0 and T1, respectively. Additionally, we calculated the Euclidean distance between each pair of students at each distinct time point (T0, T1, and T2) to assess the heterogeneity of profiles over time. For example, the Euclidean distance between the performance profiles containing three metrics (m1, m2, m3) of students J1 and J2 is given by the formula:

$$d(J1,J2) = \sqrt{(m1_{J1} - m1_{J2})^2 + (m2_{J1} - m2_{J2})^2 + (m3_{J1} - m3_{J2})^2}$$

2.7. Machine-learning analysis

In the final step of our analysis, we aimed to predict students' training levels (T0: no training, T1: 6 months of training, T2: 1.5 years of training) using students' performance profiles, combining their scores on metrics significantly affected by time.

Data Preprocessing: Each metric was independently scaled by removing the mean and scaling to unit variance, and missing values were imputed using SimpleImputer (https://scikit-learn.org/1.5/modules/generated/sklearn.impute.SimpleImputer.html) with a mean strategy. This preprocessing ensured a normalized and complete dataset suitable for machine learning analysis.

Data Splitting and Cross-Validation: To maintain data integrity and prevent leakage, a nested cross-validation procedure was employed using a 10-fold Stratified Group K-Fold (StratifiedGroupKFold; https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html) for both the inner loop, which focused on hyperparameter tuning and model selection, and the outer loop, which assesses the overall model performance, ensuring samples from the same individual (grouped by 'Person ID') were not split across folds. This process was repeated 10 times to obtain different random splits and provide a more robust estimate of the models' performance.

Models and Hyperparameter Tuning: We selected two machine learning models: Random Forest (RF) and Linear Support Vector Machine (LSVM). RF was chosen for its ability to handle complex feature interactions and provide valuable feature importance scores. LSVM is effective for high-dimensional spaces, robust against overfitting, and particularly suitable for small datasets, using a regularization parameter to manage noisy data. Hyperparameter tuning was performed using GridSearchCV (https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.GridSearchCV.html) with the same StratifiedGroupKFold cross-validation strategy to find the best hyperparameters for each

model. For RF, the hyperparameters tuned included n_estimators, max_depth, min_samples_split, and min_samples_leaf. For LSVM, the hyperparameters tuned included C and max_iter.

Model Evaluation: Each tuned model was evaluated using 10-fold StratifiedGroupKFold cross-validation on the training set to estimate generalization performance, with this process repeated 10 times for robustness. The models were then trained on the entire training set and tested on the hold-out test set, resulting in 100 total tests. Evaluation metrics, including accuracy (the overall percentage of correct predictions) and F1-score (a harmonic mean of precision and recall, where precision measures the proportion of correctly predicted instances of a class out of all predicted instances, and recall measures the proportion of correctly predicted instances of a class out of all actual instances of that class), were collected from the test sets of the outer cross-validation loop, along with confusion matrices to visualize performance. These metrics and the confusion matrix provide a comprehensive evaluation of the model's performance across different classes, highlighting both strengths and areas for improvement. The models were then trained on the entire training set and tested on the hold-out test set, resulting in 100 total tests. Evaluation metrics, including accuracy (the overall percentage of correct predictions) and F1-score (a harmonic mean of precision and recall, where precision measures the proportion of correctly predicted instances of a class out of all predicted instances, and recall measures the proportion of correctly predicted instances of a class out of all actual instances of that class), were collected from the test sets of the outer cross-validation loop, along with confusion matrices to visualize performance.

Feature Importance Analysis: To understand the contribution of each feature to the models' predictions, we extracted feature importances from the RF model and the absolute values of the LSVM coefficients from the models trained during the outer cross-validation splits. Feature importances were averaged, and standard deviations were computed to assess the models'

variability. This analysis identifies the most significant features contributing to predicting training levels across both models and assesses the consistency of the results.

2.8. Statistical data analysis

All reported metrics in Section 2.5 were computed independently for each measurement time (T0, T1, and T2). We used repeated measures ANOVAs to compare these metrics across the three time points, followed by Tukey post-hoc tests for significant results to identify specific differences. One-sample t-tests were used to test whether performance was significantly different from the chance level in recognition and discrimination tasks. In the machine-learning analysis, one-tailed significance tests were performed to estimate the likelihood that the chance level (33% for a 3-class problem) is the mean of the distribution of decoding accuracies from multiple train/test splits. Additionally, a paired t-test was conducted to compare the accuracies of the RF and LSVM models. On all reported-statistics, Bonferroni correction was applied to all p-values for multiple comparisons. A p-value < 0.050 was considered as demonstrating significant effect.

2.9. Software and libraries for data analysis

All data preprocessing and analysis were performed using Python (https://www.python.org/). The following packages and libraries were utilized: numpy (Harris et al., 2020) and pandas (McKinney, 2010; The pandas development team, 2024) for data handling and manipulation, SpaCy (Honnibal and Montani, 2017) for tokenization, Part-of-speech tagging, and lemmatization for lemmatization, matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) for creating visualizations, and Scikit-learn for scaling, normalizing and ML analyses (Pedregosa et al., 2011). These tools collectively facilitated comprehensive text processing, linguistic analysis, and visualization. Jamovi software (The jamovi project, 2024), a free open-source

statistical tool with a user-friendly graphical interface built on R, was used for statistical analyses.

3. Results

3.1. Description and Evocation tasks

Description task

The average number of words used in odor descriptions was assessed over three time periods (Figure 1A). We found a significant effect of time on the number of words used ($F(2, 78) = 40.30$, $p < 0.001$). Specifically, the average number of terms was lowest at T0 and progressively increased through T1 to T2. The number of words increased significantly from T0 to T1 (Mean Difference = -0.950, SE = 0.353, $p = 0.031$) and from T1 to T2 (Mean Difference = -1.713, SE = 0.242, $p < 0.001$), as well as from T0 to T2 (Mean Difference = -2.663, SE = 0.297, $p < 0.001$).

We then analyzed the average number of nouns, adjectives, and verbs in the odor descriptions over time (Figures 1B & 1C). All descriptions consisted of word lists, with no actual sentences present, resulting in the absence of any verbs. Similar to the trend observed with the number of words, there was an effect of time on the number of nouns ($F(2, 78) = 3.31$, $p = 0.042$), with a significant increase in the number of nouns from T0 to T2 (Mean Difference = -1.225, SE = 0.476, $p = 0.042$), but no significant differences between T0 and T1 (Mean Difference = -0.550, SE = 0.533, $p = 0.562$) and T1 and T2 (Mean Difference = -0.675, SE = 0.415, $p = 0.247$). The number of adjectives was also significantly influenced by time ($F(2, 78) = 48.50$, $p < 0.001$). An incremental increase in the use of adjectives over time was demonstrated with an increase from T0 to T1 (Mean Difference = -1.48, SE = 0.381, $p < 0.001$) and from T1 to T2 (Mean Difference = -2.52, SE = 0.407, $p < 0.001$), as well as from

T0 to T2 (Mean Difference = -4.00, SE = 0.443, $p < 0.001$). In conclusion, students used more words, nouns, and adjectives to describe odors as their learning progressed.

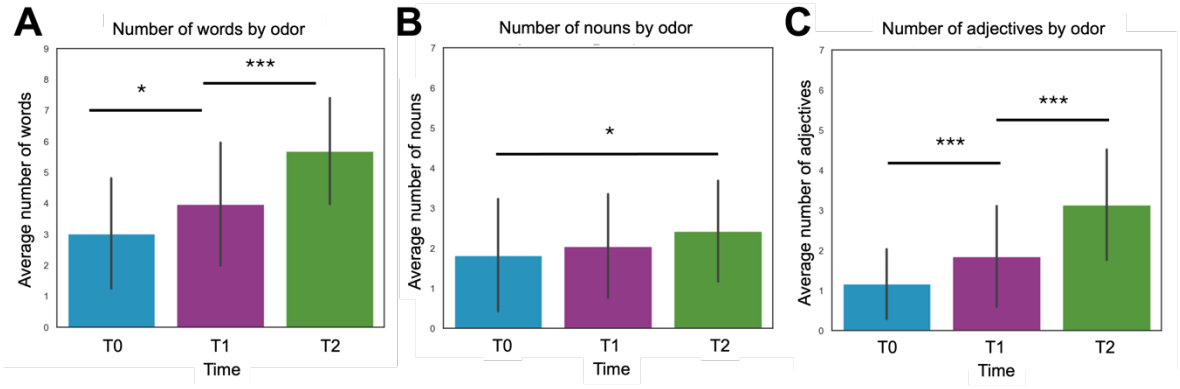

Figure1. Odor description task. Increases in the average number of words (A), nouns (B), and adjectives (C) over time (T0, T1 and T2) when describing odors. Error bars represent standard deviations. * indicates $p < 0.05$, and *** indicates $p < 0.001$.

In a second step, we explored the changes in lexical diversity within and between students' odor descriptions over time using the TTR diversity score and $cos\theta$ semantic similarity. While the within-students TTR remained close to 1, indicating maximal diversity, and was relatively stable over time ($F(2, 78) = 1.60$, $p = 0.209$; Figure 2A), within-participant semantic similarity significantly changed over time ($F(2, 74) = 14.90$, $p < 0.001$; Figure 2C). There were significant increases in semantic similarity between T0 and T1 (Mean Difference = -0.081, SE = 0.030, $p = 0.031$), T1 and T2 (Mean Difference = -0.078, SE = 0.029, $p = 0.031$), and T0 and T2 (Mean Difference = -0.160, SE = 0.029, $p < 0.001$). These results suggest that, over time, students used a distinct yet increasingly specific and semantically similar set of words to describe each odor.

The between-student TTR diversity score changed over time ($F(2, 78) = 5.23$, $p = 0.007$; Figure 2B), with a significant decrease in lexical diversity between T0 and T2 (Mean Difference = 0.146, SE = 0.042, $p = 0.004$), but no significant differences between T0 and T1 or T1 and T2 ($p > 0.245$). We also found a significant effect of time on between-participant semantic

similarity ($F(2, 78) = 90.90$, $p < 0.001$), with incremental increases across the three time points: T0 and T1 (Mean Difference = -0.082, SE = 0.013, $p < 0.001$), T1 and T2 (Mean Difference = -0.065, SE = 0.010, $p < 0.001$), and T0 and T2 (Mean Difference = -0.147, SE = 0.010, $p < .001$) (Figure 2D). This result reflects an increase in the specificity of words used by students over time, confirming greater homogeneity in their vocabulary.

In addition, we examined the semantic similarity between students' answers and an expert teacher and found a significant effect of time ($F(2, 78) = 5.82$, $p = 0.004$; Figure 2E), with significant increase between T0 and T2 (Mean Difference = -0.073, SE = 0.022, $p = 0.006$), but no significant differences between T0 and T1 ($p = 0.191$) and T1 and T2 ($p = 0.418$). Finally, the analysis of the semantic similarity of descriptions for each odor across the three time points demonstrated a significant effect of time ($F(2, 22) = 21.50$, $p < 0.001$; Figure 2F), with odors being more similarly described at T2 than T0 (Mean Difference = -0.164, SE = 0.027, $p < 0.001$) and T1 (Mean Difference = -0.154, SE = 0.034, $p = 0.002$), but no significant difference between T0 and T1 (Mean Difference = -0.010, SE = 0.022, $p = 1.000$). These results indicate that over time, students' descriptions become more aligned with the expert's descriptions and more consistent in how each odor is described, corroborating the increased precision and standardization in their vocabulary.

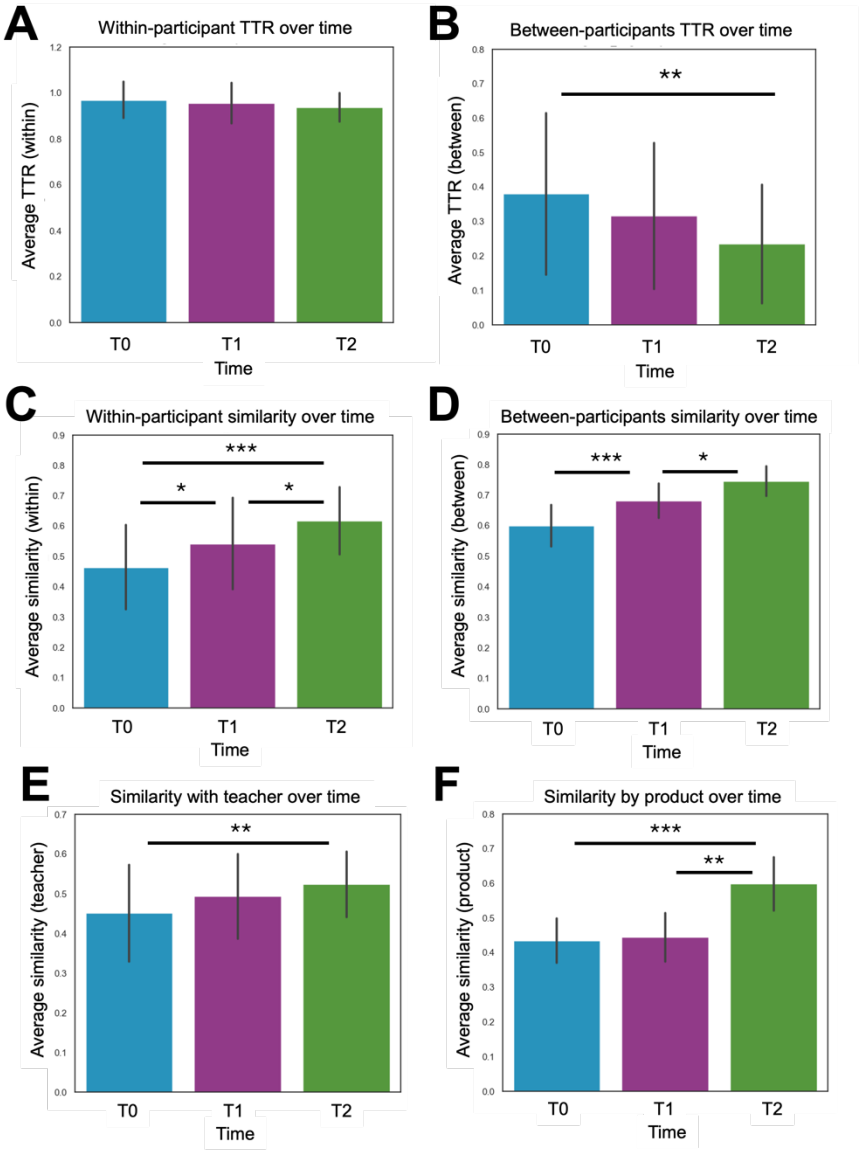

Figure 2. Odor description task. Changes in lexical diversity using Type-Token Ratio (TTR) within (A) and between (B) students, and changes in semantic similarity within (C) and between (D) students, with an expert teacher (E), and by odor (F) over time (T0, T1, T2) when describing odors. Error bars represent standard deviations. * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$.

Evocation task

No significant effect of time was found on the average number of words (Mean = 1.673, SE = 0.514, $F(2, 78) = 2.06$, $p = 0.134$), nouns (Mean = 1.228, SE = 0.452, $F(2, 78) = 0.47$, $p = 0.630$), adjectives (Mean = 0.290, SE = 0.221, $F(2, 78) = 0.83$, $p = 0.439$), and verbs (Mean = 0.111, SE = 0.206, $F(2, 78) = 1.43$, $p = 0.245$) in the evocation responses.

The analysis of lexical diversity within-students' odor evocations using TTR diversity score did not reveal significant differences of TTR over time (TTR_{within}: Mean = 0.97, SE = 0.07, F(2, 78) = 1.080, p = 0.346), but showed a significant effect of time on semantic similarity (F(2, 76) = 6.04, p = 0.004; Figure 3A), with similarity being higher at T2 than T0 (Mean Difference = -0.087, SE = 0.025, p = 0.004) but similar between T0 and T1 (Mean Difference = -0.031, SE = 0.024, p = 0.637) and T1 and T2 (Mean Difference = -0.056, SE = 0.026, p = 0.119). This result suggests that as students progressed in their training, the words they used became increasingly semantically similar.

When comparing between-participant lexical diversity, we found that while TTR diversity score showed no effect of time ($TTR_{between}$: Mean = 0.416, SE = 0.193, F(2, 78) = 1.080, p = 0.337), semantic similarity significantly changed over time (F(2, 78) = 35.00, p < 0.001; Figure 3B), with significant increases in similarity between T0 and T1 (Mean Difference = -0.026, SE = 0.009, p = 0.014), T1 and T2 (Mean Difference = -0.053, SE = 0.011, p < 0.001), and T0 and T2 (Mean Difference = -0.079, SE = 0.009, p < .001). This result indicates increased similarity in the students' vocabulary over time, reflecting greater semantic homogeneity in the words they used.

In addition, the semantic similarity of evocations for each odor significantly evolved with time (F(2, 22) = 11.20, p < 0.001; Figure 3C), with significant decreases between T0 and T1 (Mean Difference = 0.063, SE = 0.015, p = 0.005) and T2 and T1 (Mean Difference = -0.105, SE = 0.023, p = 0.002), but no difference between T0 and T2 (Mean Difference = -0.042, SE = 0.027, p = 0.462). These results indicate that over time, students' evocations were less consistent at T1 than at T0 and T2, suggesting an effect of time unrelated to their level of training. Finally, given that the evocation task is related to personal experiences and memories, we did not run similarity comparisons to a reference.

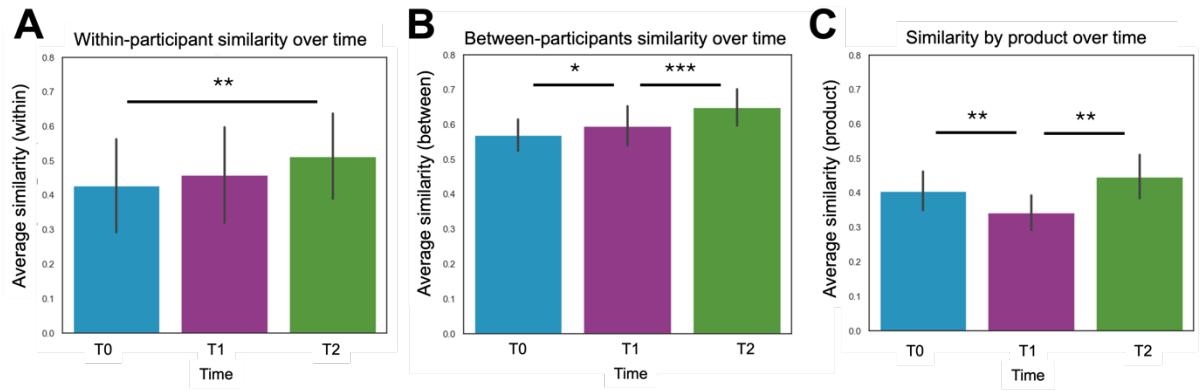

Figure3. Odor evocation task. Changes in cosine similarity ($\cos\theta$) within (A) and between (B) students and by odor (C) over time (T0, T1, T2) during odor-evoked evocations. Error bars represent standard deviations. * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$.

3.2.Discrimination task

The Odor Discrimination Score was evaluated at three time points (T0, T1, T2). At T0, the mean score was 45.00% (SD = 27.79%), which was significantly above the chance level of 33.33% ($t(39) = 2.66$, $p = 0.011$). At T1, the mean score was 59.17% (SD = 25.58%), also significantly above chance ($t(39) = 6.39$, $p < 0.001$). At T2, the mean score was 53.33% (SD = 28.04%), and it remained significantly above the chance level ($t(39) = 4.51$, $p < 0.001$).

In addition to these comparisons against chance level, the within-students effect of time was evaluated. The analysis revealed a marginally significant effect of time on performance ($F(2, 78) = 3.120$, $p = 0.050$), with no significant differences between T0 and T1 (Mean Difference = -0.425, SE = 0.189, $p = 0.091$), T0 and T2 (Mean Difference = -0.250, SE = 0.159, $p = 0.373$), and T1 and T2 (Mean Difference = 0.175, SE = 0.164, $p = 0.874$).

Overall, these results suggest that students' performance at each time point was significantly better than random guessing. However, while Odor Discrimination Scores tend to fluctuate over time, the differences between time points did not reach significance.

3.3. Olfactory and visual recognition tasks

Odor recognition task

The students were presented with target and distractor odors and were asked if they had smelled them the day before. Memory scores were high (d'L: Mean = 1.669; SE = 1.385), indicating proficiency in recognizing old odors and rejecting new ones at all measurement times (Figure 4A). d'L scores were significantly above chance at each measurement time: T0 (Mean = 1.324, SD = 1.281, $t(39) = 6.54$, $p < 0.001$), T1 (Mean = 1.631, SD = 1.497, $t(39) = 6.89$, $p < 0.001$), and T2 (Mean = 2.051, SD = 1.301, $t(39) = 9.97$, $p < 0.001$), demonstrating robust recognition performance across time points.

However, the increase in d'L over time did not reach statistical significance ($F(2,78) = 3.00$, $p = 0.056$). When examining correct response categories individually, a significant effect of time was found for correct rejections (CR) ($F(2,78) = 4.35$, $p = 0.016$; Mean = 6.858; SD = 1.789; Figure 4B), with a significant increase in the number of CRs between T0 and T2 (Mean Difference = -1.00, SE = 0.36, $p = 0.025$). No effect of time was observed on the number of Hits ($F(2,78) = 0.78$, $p = 0.460$; Mean = 6.875; SD = 1.771). The response bias was close to zero at all time points, indicating no specific response strategy [CL_{T0}: Mean = -0.055, SD = 0.525; CL_{T1}: Mean = -0.028, SD = 0.518; CL_{T2}: Mean = 0.085, SD = 0.624] and was not influenced by time ($F(2,78) = 0.89$, $p = 0.428$).

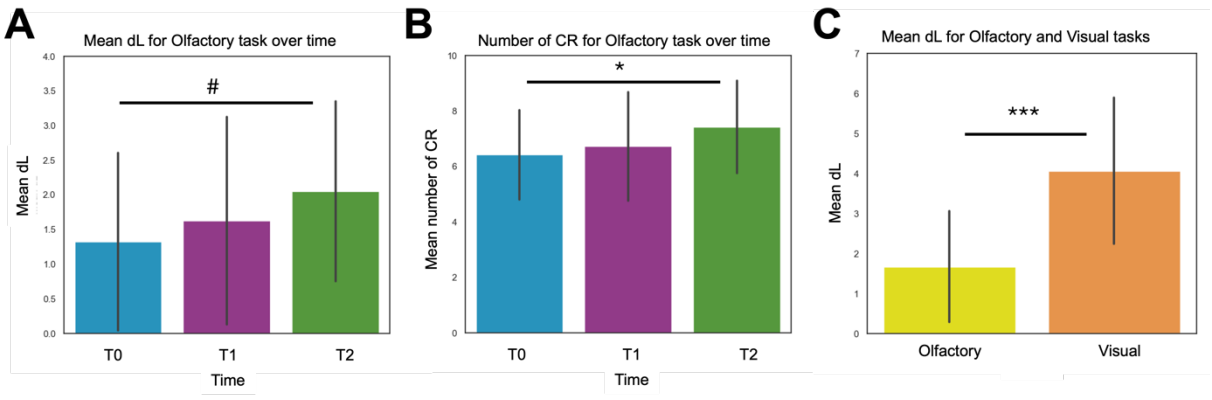

Figure4. Recognition Performance in Olfactory and Visual Tasks. (A) Mean d'L scores for the olfactory task over time. (B) Mean number of correct rejections (CR) for the olfactory task over time (10 distractors). (C) Mean d'L scores for olfactory and visual tasks. Error bars represent standard deviations. '#' indicates a trend ($p = 0.056$), * indicates $p < 0.05$, and *** indicates $p < 0.001$.

Visual recognition task

In the visual recognition task, memory scores were very high, indicating strong proficiency in recognizing old images and rejecting new ones at all measurement times (d'L: Mean = 4.068; SD = 0.167). d'L scores were significantly above chance for each measurement time: T0 (Mean = 3.789, SD = 1.898, $t(39) = 12.63$, $p < 0.001$), T1 (Mean = 4.403, SD = 1.429, $t(39) = 19.49$, $p < 0.001$), and T2 (Mean = 4.011, SD = 2.083, $t(39) = 12.18$, $p < 0.001$), highlighting consistently strong visual recognition performance.

Visual memory scores were significantly higher than olfactory memory scores ($F(1,39) = 131.54$, $p < 0.001$; Figure 4C), illustrating that the visual memory task was easier for students than the olfactory memory task. However, for the visual task, there was no effect of time on either d'L memory scores or CL individual response categories ($F's < 1.90$; $p's > 0.152$). Finally, the response bias was close to zero at all time points (CL_{T0}: Mean = -0.017, SD = 0.734; CL_{T1}: Mean = -0.014, SD = 0.508; CL_{T2}: Mean = -0.054, SD = 0.613), indicating no specific response strategy, and was not influenced by time ($F(2,78) = 0.22$, $p = 0.807$).

3.4. Odor categorization task and Categories description

Odor categorization task

No significant effect of time was shown on the number of groups created by students ($F(2, 76) = 1.65$, $p = 0.20$; Mean = 6.342; SD = 1.738). The analysis of inter-student agreement indicated no agreement regarding the categories of odors created [Cohen's Kappa: Mean = 0.008, SD = 0.024] and showed no significant effect of time on agreement ($F(2, 76) = 0.249$, $p = 0.780$). However, the similarity of odors pairwise clustering (ARI) between students significantly

varied with time ($F(2, 76) = 17.30$, $p < 0.001$; Figure 5A). There were significant increases between T0 and T1 (Mean Difference = -0.040, SE = 0.009, $p < .001$) and T0 and T2 (Mean Difference = -0.061, SE = 0.011, $p < 0.001$), but no significant difference between T1 and T2 (Mean Difference = -0.021, SE = 0.012, $p = 0.227$).

In addition, the analysis also revealed no agreement between students and the expert teacher regarding the categories of odors created [Cohen's Kappa: Mean = -0.034, SD = 0.100], and no significant effect of time was found ($F(2, 76) = 0.863$, $p = 0.426$). However, the similarity of odors pairwise clustering between students and the expert teacher significantly differed with time ($F(2, 76) = 3.260$, $p = 0.044$), with a significant increase between T0 and T1 (Mean Difference = -0.066, SE = 0.022, $p = 0.015$), but no significant differences between T0 and T2 (Mean Difference = -0.052, SE = 0.029, $p = 0.233$) and between T1 and T2 (Mean Difference = 0.014, SE = 0.030, $p = 1.000$).

Overall, while categorical agreement remained stable over time between students and the expert teacher, the pairwise similarity of clustering increased with training, both among students and between students and the expert teacher.

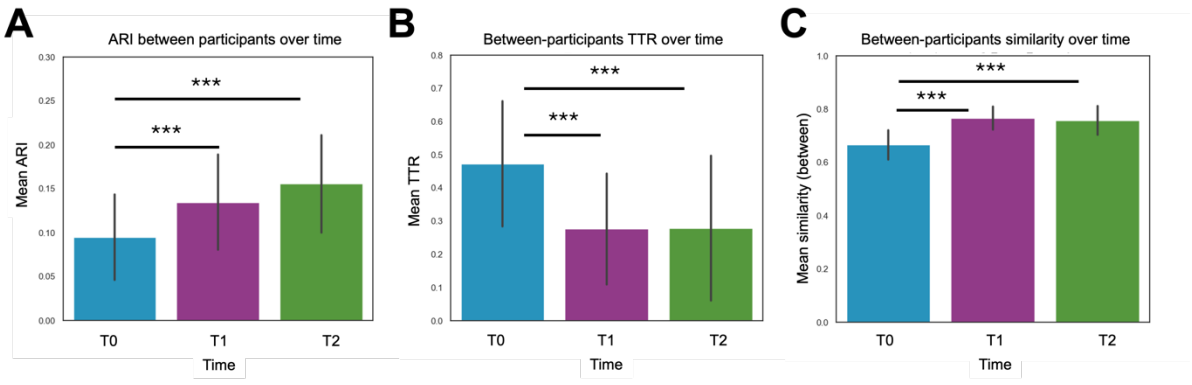

Figure5. Odor categorization task. Changes of between-students (A) pairwise odors clustering similarity (ARI), (B) lexical diversity (TTR), and (C) semantic similarity ($cos(\theta)$) scores between students over time (T0, T1, T2). Error bars represent standard deviations. *** indicates $p < 0.001$.

Description of odor categories

A significant effect of time on the average number of words was found ($F(2,76) = 3.36$, $p = 0.040$), showing a slight increase between T0 (Mean = 1.749, SD = 0.558) and T1 (Mean = 2.090, SD = 0.922) but no significant differences were found with T3 (Mean = 1.981, SD = 0.759; $p's > 0.25$). No effect of time was revealed for the average number of nouns (Mean = 3.949, SD = 2.137; $F(2,76) = 1.580$, $p = 0.213$), nor adjectives (Mean = 6.017, SE = 2.626; $F(2,76) = 0.514$, $p = 0.600$). No verb was found in the verbal descriptions.

The between-participant lexical diversity (TTR) and semantic similarity ($cos(\theta)$) significantly differed over time ($TTR_{between}$: $F(2,76) = 14.800$, $p < .001$; $cos(\theta)_{between}$: $F(2,76) = 50.800$, $p < .001$; Figure 5B-C). There were significant decreases of lexical diversity between T0 and T1 (TTR Mean Difference = 0.20, SD = 0.04, $p < 0.001$) and T0 and T2 (TTR Mean Difference = 0.196, SD = 0.041, $p < 0.001$) but not between T1 and T2 (TTR Mean Difference = -0.002, SD = 0.039, $p = 1.000$). Conversely, we found significant increases of semantic similarity between T0 and T1 (Mean Difference = -0.101, SD = 0.012, $p < 0.001$), T0 and T2 (Mean Difference = -0.091, SD = 0.011, $p < 0.001$), but not between T1 and T2 (Mean Difference = 0.009, SD = 0.009, $p = 0.984$). Given that the principle of the categorization task is to create different groups with specific descriptions, we did not run TTR and similarity comparisons within-students' categories.

To sum-up, while the average number of words in category descriptions slightly increased over time, no significant changes were observed in the use of nouns or adjectives. However, there was a significant decrease in lexical diversity and a significant increase in semantic similarity between students' verbalizations over time, indicating a trend towards more homogeneous and semantically aligned verbal descriptions.

3.5.Temporal dynamic of learning

Students' progress dynamics were analyzed over time. Students' performance profile was created by combining their scores on the 14 metrics significantly affected by time (Table S2). The within-students effect of time was found to be significant ($F(2, 76) = 10.66$, $p < 0.001$; Figure 6A), with a linear and consistent increase in performance over time. Significant differences in progress were observed between T0-T1 and T0-T2 (Mean Difference = -0.685, SE = 0.187, $p = 0.002$), and between T1-T2 and T0-T2 (Mean Difference = -1.027, SE = 0.210, $p < 0.001$). However, no significant difference was found between T0-T1 and T1-T2 (Mean Difference = 0.341, SE = 0.254, $p = 0.515$). Overall, these results suggest that students' progress was consistent from T0 to T1 and from T1 to T2.

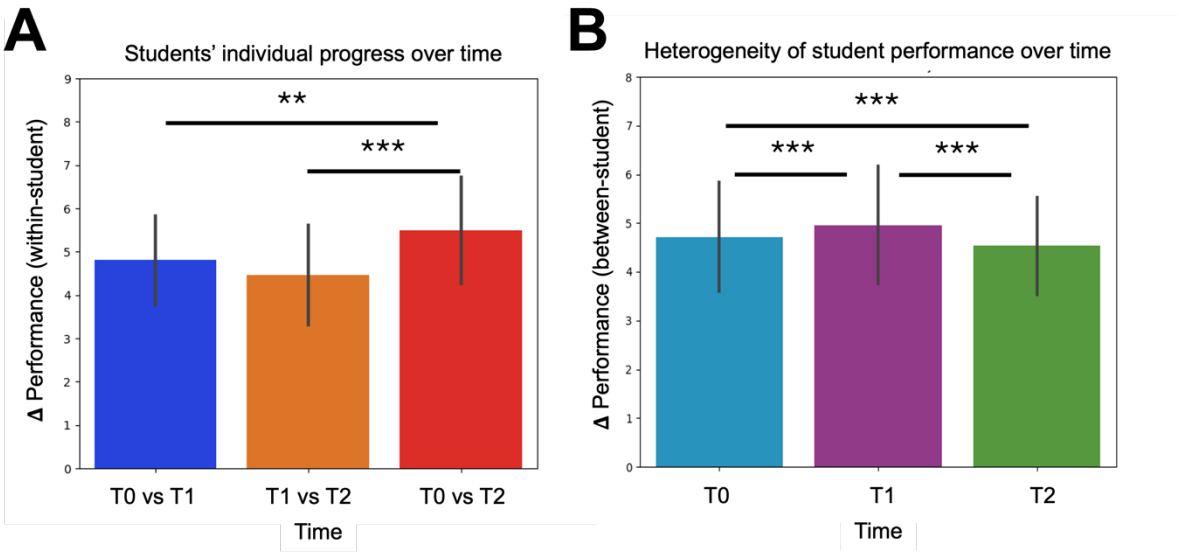

Figure6. Temporal dynamics of students' learning. (A) Changes of within-students' performance over reflecting their progress as their training progressed, and (B) Changes of between-students' performance over time showing the heterogeneity of performance between students at each time (T0, T1, T2). Error bars represent standard deviations. **, indicates $p < 0.01$; *** indicates $p < 0.001$

The heterogeneity of performance between students was evaluated based on the score differences between each pair of students at each time measurement (T0, T1, and T2). The effect of time on the heterogeneity of performance was significant ($F(2, 1480) = 29.00$, $p <$

0.001; Figure 6B), with a greater heterogeneity at T1 compared to T0 (Mean Difference = 0.241, SE = 0.064, $p < 0.001$) and T2 (Mean Difference = 0.429, SE = 0.054, $p < 0.001$). However, heterogeneity was lower at T2 than at T0 (Mean Difference = -0.188, SE = 0.050, $p < 0.001$). These results suggest that the initial heterogeneity of students' profiles at T0 increased by T1, indicating varying speeds of learning among students. By T2, the heterogeneity decreased to a level even below the initial point, reflecting a convergence in the learning levels of students.

3.6. Predicting students' training levels

Given the clear changes in performance over time, we explored whether it was possible to identify students' level of training with high accuracy based on their performance across tasks using Random Forest (RF) and Linear Support Vector Machine (LSVM) classifiers.

On average, the best performance was obtained with the RF classifier, which correctly identified the students' training level 78.9% of the time, with a margin of error of ±11.1% (Figure 7A). The statistical analysis showed that the accuracy score was highly significant ($t = 41.30$; one-tailed corrected p-value < 0.001), and the model achieved an F1-score of 0.783 ± 0.113, indicating strong performance in both precision (accuracy of positive predictions) and recall (ability to find all relevant instances). For the RF classifier, the most crucial features were the semantic similarities of odor descriptions among students (description_sim-BTW) and the semantic similarities between odor category verbalizations among students (categorisation_sim-BTW; Figure 7B).

The SVM classifier also performed well, correctly identifying the training level of students 74.2% of the time, with a margin of error of ± 6.9% ($t = 59.173$; one-tailed corrected p-value < 0.001; Figure 7A), and an F1-score of 0.728 ± 0.076, indicating good overall performance of the model across all classes. The feature importance analysis for the SVM classifier (Figure 7C) also highlighted the importance of the semantic similarities between students' responses

measured in the description and the verbalization part of the categorization tasks in determining the training level of students.

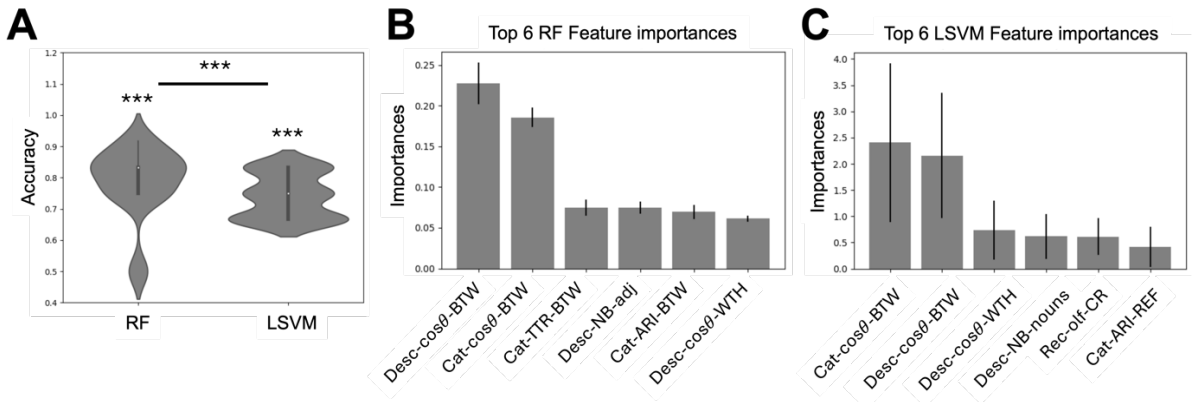

Figure7. Feature Importances and Model Accuracies of students' level of training predictions. (A) Test accuracies for the RF and LSVM models. (B) Top 6 feature importances for the Random Forest (RF) classifier. (C) Top 6 feature importances for the Linear Support Vector Machine (LSVM) classifier. Error bars represent standard deviations. *** indicates $p < 0.001$. Adj: adjectives; Cat: categorization task; Desc: description task; NB: number; Rec-olf: odor recognition task.

When directly comparing the performance of the RF and LSVM classifiers, the model comparison indicated that the RF model significantly outperformed the LSVM model ($t = 4.101$; $p < 0.001$). This suggests that the RF model, with its large number of estimators and fully grown trees, was better suited for capturing the underlying patterns in the perceptual and cognitive data, leading to its superior performance compared to the LSVM classifier. The full hyperparameters tuning results are summarized in Table S3.

4. Discussion

The primary objective of this research was to examine how academic olfactory training during a 1.5-year educational program impacts the development of olfactory expertise in perfumery students. The study focused on a broad range of tasks, including odor description, evocation, recognition, discrimination, and categorization. Our main findings reveal that while students demonstrated significant improvement in their ability to describe and categorize odors—

reflected in richer (words, nouns, adjectives count) and more consistent language (TTR scores and cosθ similarity), and its greater alignment with expert terminology (cosθ similarity)—their non-verbally mediated abilities, such as odor discrimination and recognition accuracies, did not show substantial enhancement over time.

Our results demonstrate that as students progressed through the training program, their vocabulary and descriptive abilities expanded, evidenced by an increased use of words, nouns, and adjectives to describe odors. Over time, however, lexical diversity within the student group decreased, while semantic similarity increased—both among students and between students and the expert teacher. This indicates a homogenization of vocabulary, as students adopted a more standardized language for odor description. This convergence suggests that the training not only enriched their descriptive repertoire but also guided them toward a shared, expert-aligned terminology. Interestingly, the number of words used in odor-evoked evocations did not change significantly over time, yet semantic similarity increased within and between students. This implies that while students did not elaborate more in terms of word count when recalling odor-associated memories, their language became more semantically aligned. This reflects the influence of training on the conceptualization and verbal encoding of olfactory experiences, facilitating more efficient retrieval and communication of olfactory information (Larsson and Willander, 2009). These findings align with previous research indicating that professional perfumers use more detailed and precise language focusing on chemical and olfactory qualities (Sezille et al., 2014) and that olfactory cognitive abilities can be improved with training (Royet et al., 2013).

In addition to improvements in descriptive abilities, our findings reveal significant changes in how students categorized odors over time. While the number of groups created and the categorical agreement among students did not change significantly, the similarity of pairwise odor clustering between students increased. This suggests that as students progressed through

686 the training, they began grouping odors in increasingly similar ways, reflecting a convergence
687 in their perceptual representation of odors. Moreover, although the average number of words
688 used in category descriptions increased slightly, there was a noticeable decrease in lexical
689 diversity and a significant increase in semantic similarity among students' verbalizations. In
690 other words, students used fewer unique terms and described odors in ways that were more
691 aligned with one another. The increased alignment in odor grouping and the standardization of
692 vocabulary suggest that students are developing a shared terminology for categorizing odors.

693 Our results highlight the central role of language in structuring olfactory knowledge, and align
694 with prior cross-cultural findings highlighting the variability of odor lexicons and the
695 importance of consistent linguistic frameworks for olfactory description (Majid, 2021). Studies
696 on languages such as Jahai and Maniq demonstrate that dedicated lexical fields for odors can
697 emerge when olfactory experience is consistently emphasized and linguistically encoded
698 (Majid and Burenhult, 2014). These findings challenge the notion proposed by Olofsson and
699 Gottfried that olfactory naming difficulties are primarily rooted in neural anatomical constraints
700 (Olofsson and Gottfried, 2015). Instead, our results support the idea that developing olfactory
701 expertise and a shared vocabulary for odors is not only possible but also heavily influenced by
702 academic training practices.

703 Despite the observed improvements in verbally-mediated and categorization tasks, our study
704 did not reveal significant enhancements in non-verbally mediated abilities such as odor
705 discrimination and recognition. The odor discrimination scores showed no significant
706 improvement over time, and while there was a trend toward better odor recognition
707 performance, it did not reach statistical significance. These findings are consistent with research
708 by Filiz et al. (2022), who found no substantial improvement in discrimination or recognition
709 abilities in wine students over a similar period of training. This lack of improvement could be
710 attributed to several factors. First, high-order non-verbally mediated tasks such as odor

discrimination and recognition may require extensive and more prolonged exposure to a wider variety of odors to see measurable changes (Royet et al., 2013), which might not have been fully achieved within the duration of our study. Second, prior research suggests that olfactory training effects are often task-specific and may not generalize across different odor categories (Bende and Nordin, 1997; Parr et al., 2002; Croijmans and Majid, 2016; Poupon et al., 2018). At ISIPCA, students are trained using a structured curriculum that focuses on the gradual introduction of raw materials classified by olfactory families, progressing from simpler to more complex odors. While this approach strengthens categorical organization and verbal descriptors, it may also reinforce expertise within specific odor categories rather than promoting broader generalization across categories. For instance, training sessions often emphasize identifying and describing fragrances or flavors within specific families, such as fruity, floral, or toasted, which could limit the transfer of perceptual skills to novel or less familiar odor groups. Another plausible factor is that the methods employed in this study, including task design and stimuli selection, may not have been sensitive enough to capture subtle changes in olfactory performance.

The use of natural language processing (NLP) and machine learning (ML) tools in this study provided valuable insights into olfactory expertise, reinforcing the centrality of language in its acquisition. These computational methods allowed us to capture subtle patterns in language use, uncovering links between linguistic descriptors and expertise that would have been overlooked with more traditional approaches. Features related to verbal tasks—such as the richness and precision of odor descriptions, the use of specific olfactory terminology, and semantic similarity measures in the description and categorization tasks—were the strongest predictors of students' performance and expertise levels. Without NLP and ML techniques, the effects of olfactory training on expertise would have gone undetected. This highlights the potential and importance

of NLP and ML methods for better characterizing olfactory language and expertise, as well as refining training methodologies.

Our effort to better characterize the semantic properties of olfactory language aligns with other computational approaches that have recently emerged to deepen our understanding of olfactory perception and its connection to language and expertise. For instance, Iatrapoulos et al. (2018) introduced new metrics—the Olfactory Association Index (OAI) and Olfactory Specificity Index (OSI)—to quantitatively assess how strongly words are associated with olfaction and how specifically they describe odors. Similarly, Hörberg et al. (2022) mapped the olfactory vocabulary of English using NLP techniques, identifying semantic dimensions such as valence, concreteness, and edibility that organize the olfactory lexicon. Another study by Hörberg et al. (2025) compared sensory vocabularies across domains such as wine, perfume, and food using computational methods, revealing critical differences in domain specificity, descriptor preferences, and semantic dimensions. These studies highlight how computational analyses of large-scale textual data can uncover the linguistic structures underlying sensory descriptions and expertise.

Collectively, these computational approaches to olfactory language underscore the usefulness of NLP and ML tools not only for characterizing olfactory expertise but also for advancing training methodologies in olfactory and sensory sciences. By leveraging these innovative methods, we can gain deeper insights into the verbal underpinnings of olfactory expertise, and develop more effective strategies for training and assessment.

The olfactory training program at ISIPCA demonstrates considerable effectiveness in enabling students to develop and utilize a precise, shared language for describing and categorizing odors. Through structured exercises, systematic exposure, and standardized verbalization practices, students progressively refine their vocabulary, achieving greater alignment with expert terminology. While these findings underscore the program’s proficiency in fostering linguistic

and conceptual frameworks critical for olfactory expertise, there remains scope for further refinement. Building upon this solid foundation, our results, supported by prior research, suggest three key areas that could further optimize training outcomes.

Firstly, emphasizing vocabulary specificity and encouraging the consistent use of standardized terminology can strengthen associative links between odors and their descriptions, facilitating memory and conceptualization. The structured language training followed at ISIPCA, where students repeatedly smelled, described, and classified raw materials using Jaubert's field of odors (1995), already demonstrated how systematic exposure and verbalization promote the development of a precise and shared olfactory lexicon. To further enhance this approach, the introduction of a shared textual vocabulary dictionary could ensure that all students have access to a unified reference for describing odors. This resource would serve as a foundational tool for communication and conceptual alignment, reinforcing the consistent use of shared language across both students and teachers.

Secondly, integrating mental imagery training into the curriculum could enhance cognitive skills crucial for perfumers. Research by Plailly et al. (2012) showed that olfactory mental imagery is central to perfumers' daily tasks and improves significantly with training. Croijmans et al. (2020) demonstrated that mental olfactory imagery can be developed through targeted practice rather than being an inherent ability. Similarly, Stevenson et al. (2007) found that training in odor naming enhances mental imagery by providing verbal anchors that strengthen cognitive associations. Given the structured exercises in ISIPCA training program, such as identifying raw materials blindly and describing their volatility, incorporating explicit mental imagery tasks could complement these activities and support students in imagining, recalling, and categorizing odors more effectively. Testing these processes in future studies would clarify how mental imagery integrates with verbal and perceptual learning to build olfactory expertise.

Thirdly, since improvements in perceptual abilities such as odor discrimination may be slower, training programs could incorporate prolonged exposure to a wider variety of odors and more intensive discrimination tasks. This could include differentiating subtle variations in odor mixtures or thresholds, as suggested by Morquecho-Campos et al. (2019) and Poupon et al. (2018). For instance, ISIPCA program already introduced students to a progression of odor intensities and complex combinations through structured sessions, but future designs might integrate targeted exercises to refine sensory acuity further. These could involve evaluating persistence and volatility or focusing more extensively on odor combinations to sharpen discrimination and perceptual organization.

Moreover, our study highlights the dynamics of learning over time and individual differences among students. Greater variability in performance observed at the midpoint of training (T1) suggests that individuals progress at different rates. By the end of the program (T2), performance levels converged, indicating that prolonged training promotes a more uniform level of expertise. These findings underscore the need for personalized training approaches that accommodate individual learning trajectories, tailoring intensity and content to optimize outcomes and help students reach their full potential.

Conclusion

In conclusion, our study highlights the pivotal role of language in the development of olfactory expertise among perfumery students. Over the course of the 1.5-year training program, students demonstrated significant improvements in their ability to describe and categorize odors, developing a richer and more precise olfactory lexicon. These linguistic advancements were accompanied by increased semantic alignment among students and with expert standards, underscoring the effectiveness of structured training in fostering a shared vocabulary essential for professional expertise. However, the findings reveal limitations in the enhancement of non-verbally mediated abilities, such as odor discrimination and recognition. This disparity suggests

that while acquiring a shared olfactory language is crucial for developing conceptual and descriptive skills, perceptual expertise may require more extensive or targeted sensory exercises. Our study highlights the need for complementary strategies, including prolonged exposure to diverse odor profiles, mental imagery training, and personalized training approaches tailored to individual learning trajectories, to fully support the multifaceted nature of olfactory expertise. Finally, our results also underscore the value of computational approaches, such as natural language processing (NLP) and machine learning (ML), in characterizing olfactory expertise and refining training methodologies. Future research should explore these approaches to deepen our understanding of how olfactory expertise emerges.

5. Conflict of interests

The authors declare no conflicts of interest related to this research.

6. Funding

The experimental costs associated with this research were fully covered by ISIPCA. The work of the researchers was supported by funding from their respective institutions. No additional external funding was received for this study.

7. Acknowledgments

We would like to express our sincere gratitude to all the perfumery students who participated in this study for their time and effort. We also extend our heartfelt thanks to our colleagues at ISIPCA for their invaluable support in facilitating the data acquisition process. Their assistance and collaboration were instrumental in the successful completion of this research.

8. Data availability

The data that support the findings of this study will be made publicly available in an anonymized format upon publication in the Open Science Framework (OSF).

9. References

Al Aïn, S., Poupon, D., Hétu, S., Mercier, N., Steffener, J., and Frasnelli, J. 2019. Smell training improves olfactory function and alters brain structure. NeuroImage. 189:45–54.

Banks, S.J., Sreenivasan, K.R., Weintraub, D.M., Baldock, D., Noback, M., Pierce, M.E., Frasnelli, J., James, J., Beall, E., Zhuang, X., et al. 2016. Structural and Functional MRI Differences in Master Sommeliers: A Pilot Study on Expertise in the Brain. Front Hum Neurosci. 10:414.

Barbeau, E., Didic, M., Tramoni, E., Felician, O., Joubert, S., Sontheimer, A., Ceccaldi, M., and Poncet, M. 2004. Evaluation of visual recognition memory in MCI patients. Neurology. 62:1317–1322.

Bende, M., and Nordin, S. 1997. Perceptual Learning in OlfactionProfessional Wine Tasters versus Controls. Physiol Behav. 62:1065–1070.

Bensafi, M., Fournel, A., Joussain, P., Poncelet, J., Przybylski, L., Rouby, C., and Tillmann, B. 2017. Expertise shapes domain-specific functional cerebral asymmetry during mental imagery: the case of culinary arts and music. Eur J Neurosci. 45:1524–1537.

Carreiras, M., Quiñones, I., Chen, H.A., Vázquez-Araujo, L., Small, D., and Frost, R. 2024. Sniffing out meaning: Chemosensory and semantic neural network changes in sommeliers. Hum Brain Mapp. 45:e26564.

Castriota-Scanderbeg, A., Hagberg, G.E., Cerasa, A., Committeri, G., Galati, G., Patria, F., Pitzalis, S., Caltagirone, C., and Frackowiak, R. 2005. The appreciation of wine by sommeliers: a functional magnetic resonance study of sensory integration. NeuroImage. 25:570–578.

853 Croijmans, I., Arshamian, A., Speed, L.J., and Majid, A. 2021. Wine experts' recognition of
854 wine odors is not verbally mediated. J Exp Psychol Gen. 150:545–559.

855 Croijmans, I., and Majid, A. 2016. Not All Flavor Expertise Is Equal: The Language of Wine
856 and Coffee Experts. PLOS ONE. 11:e0155845.

857 Croijmans, I., Pellegrino, R., and Janice Wang, Q. 2024. Demystifying wine expertise through
858 the lens of imagination: Descriptions and imagery vividness across sensory modalities. Food
859 Res Int Ott Ont. 182:114159.

860 Croijmans, I., Speed, L.J., Arshamian, A., and Majid, A. 2020. Expertise Shapes Multimodal
861 Imagery for Wine. Cogn Sci. 44:e12842.

862 Delon-Martin, C., Plailly, J., Fonlupt, P., Veyrac, A., and Royet, J.P. 2013. Perfumers' expertise
863 induces structural reorganization in olfactory brain regions. NeuroImage. 68:55–62.

864 Filiz, G., Poupon, D., Banks, S., Fernandez, P., and Frasnelli, J. 2022. Olfactory bulb volume
865 and cortical thickness evolve during sommelier training. Hum Brain Mapp. 43:2621–2633.

866 Fournel, A., Sezille, C., Licon, C.C., Sinding, C., Gerber, J., Ferdenzi, C., Hummel, T., and
867 Bensafi, M. 2017. Learning to name smells increases activity in heteromodal semantic areas.
868 Hum Brain Mapp. 38:5958–5969.

869 Haehner, A., Tosch, C., Wolz, M., Klingelhoefer, L., Fauser, M., Storch, A., Reichmann, H.,
870 and Hummel, T. 2013. Olfactory Training in Patients with Parkinson's Disease. PLOS ONE.
871 8:e61680.

872 Harris, C.R., Millman, K.J., Van Der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D.,
873 Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. 2020. Array programming with NumPy.
874 Nature. 585:357–362.

Herz, R. 2005. The unique interaction between language and olfactory perception and cognition.

Honnibal, M., and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hörberg, T., Kurfalı, M., and Olofsson, J.K. 2025. Chemosensory vocabulary in wine, perfume and food product reviews: Insights from language modeling. Food Qual Prefer. 124:105357.

Hörberg, T., Larsson, M., and Olofsson, J.K. 2022. The Semantic Organization of the English Odor Vocabulary. Cogn Sci. 46:e13205.

Hughson, A.L., and Boakes, R.A. 2001. Perceptual and cognitive aspects of wine expertise. Aust J Psychol. 53:103–108.

Hunter, J.D. 2007. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 9:90–95.

Iatropoulos, G., Herman, P., Lansner, A., Karlgren, J., Larsson, M., and Olofsson, J.K. 2018. The language of smell: Connecting linguistic and psychophysical properties of odor descriptors. Cognition. 178:37–49.

Jaubert, J.-N., Tapiero, C., and Dore, J.-C. 1995. The Field of Odors: Toward a Universal Language for Odor Relationships. Perfum Flavorist. 20.

Larsson, M., and Willander, J. 2009. Autobiographical odor memory. Ann N Acad Sci. 1170:318–23.

Lesschaeve, I., and Issanchou, S. 1996. Effects of panel experience on olfactory memory performance: influence of stimuli familiarity and labeling ability of subjects. Chem Senses. 21:699–709.

896 Lockhart, R., and Murdock, B. 1970. Memory and the theory of signal detection. Psychol Bull.
897 74:100–9.

898 Majid, A. 2021. Human Olfaction at the Intersection of Language, Culture, and Biology. Trends
899 Cogn Sci. 25:111–123.

900 Majid, A., and Burenhult, N. 2014. Odors are expressible in language, as long as you speak the
901 right language. Cognition. 130:266–270.

902 McKinney, W. 2010. Data Structures for Statistical Computing in Python. In: Austin, Texas.
903 pp. 56–61.

904 Morquecho-Campos, P., Larsson, M., Boesveldt, S., and Olofsson, J.K. 2019. Achieving
905 Olfactory Expertise: Training for Transfer in Odor Identification. Chem Senses. 44:197–203.

906 Olofsson, J.K., and Gottfried, J.A. 2015. The muted sense: neurocognitive limitations of
907 olfactory language. Trends Cogn Sci. 19:314–321.

908 Parr, W.V. 2019. Demystifying wine tasting: Cognitive psychology's contribution. Food Res
909 Int. 124:230–233.

910 Parr, W.V., Heatherbell, D., and White, K.G. 2002. Demystifying wine expertise: olfactory
911 threshold, perceptual skill and semantic memory in expert and novice wine judges. Chem
912 Senses. 27:747–755.

913 Pazart, L., Comte, A., Magnin, E., Millot, J.-L., and Moulin, T. 2014. An fMRI study on the
914 influence of sommeliers' expertise on the integration of flavor. Front Behav Neurosci. 8:358.

915 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
916 Prettenhofer, P., Weiss, R., Dubourg, V., et al. 2011. Scikit-learn: Machine Learning in Python.
917 J Mach Learn Res. 12:2825–2830.

918 Plailly, J., Delon-Martin, C., and Royet, J.P. 2012. Experience induces functional
919 reorganization in brain regions involved in odor imagery in perfumers. Hum Brain Mapp.
920 33:224–234.

921 Poupon, D., Fernandez, P., Archambault Boisvert, S., Migneault-Bouchard, C., and Frasnelli,
922 J. 2018. Can the Identification of Odorants Within a Mixture Be Trained? Chem Senses.
923 43:721–726.

924 Royet, J.P., Plailly, J., Saive, A.L., Veyrac, A., and Delon-Martin, C. 2013. The impact of
925 expertise in olfaction. Front Psychol. 4.

926 Sezille, C., Fournel, A., Rouby, C., Rinck, F., and Bensafi, M. 2014. Hedonic appreciation and
927 verbal description of pleasant and unpleasant odors in untrained, trainee cooks, flavorists, and
928 perfumers. Front Psychol. 5:12.

929 Snodgrass, J.G., and Corwin, J. 1988. Pragmatics of measuring recognition memory:
930 applications to dementia and amnesia. J Exp Psychol Gen. 117:34–50.

931 Solomon, G.E.A. 1997. Conceptual Change and Wine Expertise. J Learn Sci. 6:41–60.

932 Sreenivasan, K., Zhuang, X., Banks, S.J., Mishra, V., Yang, Z., Deshpande, G., and Cordes, D.
933 2017. Olfactory Network Differences in Master Sommeliers: Connectivity Analysis Using
934 Granger Causality and Graph Theoretical Approach. Brain Connect. 7:123–136.

935 Stevenson, R.J., Case, T.I., and Mahmut, M. 2007. Difficulty in evoking odor images: the role
936 of odor naming. Mem Cognit. 35:578–589.

937 The jamovi project. 2024. Jamovi (Version 2.5) [Computer Software].

938 The pandas development team. 2024. pandas-dev/pandas: Pandas. Zenodo.

939 Vanek, N., Sóskuthy, M., and Majid, A. 2021. Consistent verbal labels promote odor category
940 learning. Cognition. 206:104485.

941 Waskom, M.L. 2021. seaborn: statistical data visualization. J Open Source Softw. 6:3021.

942 Zambom-Ferraresi, F., Zambom-Ferraresi, F., Fernández-Irigoyen, J., Lachén-Montes, M.,
943 Cartas-Cejudo, P., Lasarte, J.J., Casares, N., Fernández, S., Cedeño-Veloz, B.A., Maraví-
944 Aznar, E., et al. 2021. Olfactory Characterization and Training in Older Adults: Protocol Study.
945 Front Aging Neurosci. 13.

946

10. Figure Legends

Figure1. Odor description task. Increases in the average number of words (A), nouns (B), and adjectives (C) over time (T0, T1 and T2) when describing odors. Error bars represent standard deviations. * indicates $p < 0.05$, and *** indicates $p < 0.001$.

Figure 2. Odor description task. Changes in lexical diversity using Type-Token Ratio (TTR) within (A) and between (B) students, and changes in semantic similarity within (C) and between (D) students, with an expert teacher (E), and by odor (F) over time (T0, T1, T2) when describing odors. Error bars represent standard deviations. * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$.

Figure3. Odor evocation task. Changes in cosine similarity ($\cos\theta$) within (A) and between (B) students and by odor (C) over time (T0, T1, T2) during odor-evoked evocations. Error bars represent standard deviations. * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$.

Figure4. Recognition Performance in Olfactory and Visual Tasks. (A) Mean d'L scores for the olfactory task over time. (B) Mean number of correct rejections (CR) for the olfactory task over time (10 distractors). (C) Mean d'L scores for olfactory and visual tasks. Error bars represent standard deviations. '#' indicates a trend ($p = 0.056$), * indicates $p < 0.05$, and *** indicates $p < 0.001$.

Figure5. Odor categorization task. Changes of between-students (A) pairwise odors clustering similarity (ARI), (B) lexical diversity (TTR), and (C) semantic similarity ($cos(\theta)$) scores between students over time (T0, T1, T2). Error bars represent standard deviations. *** indicates $p < 0.001$.

Figure6. Temporal dynamics of students' learning. (A) Changes of within-students' performance over reflecting their progress as their training progressed, and (B) Changes of

between-students' performance over time showing the heterogeneity of performance between students at each time (T0, T1, T2). Error bars represent standard deviations. **, indicates $p < 0.01$; *** indicates $p < 0.001$

Figure7. Feature Importances and Model Accuracies of students' level of training predictions. (A) Test accuracies for the RF and LSVM models. (B) Top 6 feature importances for the Random Forest (RF) classifier. (C) Top 6 feature importances for the Linear Support Vector Machine (LSVM) classifier. Error bars represent standard deviations. *** indicates $p < 0.001$. Adj: adjectives; Cat: categorization task; Desc: description task; NB: number; Rec-olf: odor recognition task.

11. Supplementary materials

Table S1. Exhaustive list of olfactory stimuli

Note. Abs, absolute ; Ess, essence ; EO, essential oil. Perfumes are in italics. Hyperscript letters referred to odorants that are used together in a mixture

Tasks	Odorants	Reference
Evocation and description*	Vegetable Spicy grilled Pharmaceutical Floral Sun-soaked Sea scent Detergent Dried fruit honey Musky animal scent Late evening Green tea Fig Bin scent	Formulated at the ISIPCA using in-house recipes
Discrimination	*Flowerbomb* *La Vie Est Belle* *La Nuit Trésor* *Black Opium* *L'Homme* Pink Grape Fruit	Viktor & Rolf Lancôme Lancôme Yves Saint Laurent Yves Saint Laurent Hermès
Recognition T0	Ambrarome Thymol Borneol Nutmeg *Light blue*	Nactis Sigma Aldrich Prodasynth Robertet Dolce Gabbana

	Le REM	Reminiscence
	Accord EMMA	ISIPCA Formulation
	Aromatics in black	Clinique
	Aromatics in white	Clinique
	L'Instant Magic	Guerlain
	Eau d'Hadrien	A.Goutal
	Apogée	Vuitton
	Contre moi	Vuitton
	Dans la peau	Vuitton
	Matière noire	Vuitton
	Mille feux	Vuitton
	Rose des vents	Vuitton
	Turbulences	Vuitton
	Piment brûlant	Artisan Parfumeur
	Coach	New York
Recognition T1	Dew fruit[a]	Givaudan
	Limon ess[a]	Albert Vieille
	Phenylacetic acid[b]	Sigma Aldrich
	beta-naphthyl methyl ketone[b]	BLH PIM
	Chocovan[c]	GIVAUDAN
	Nutmeg ess[c]	Albert Vieille
	Damascenia[d]	FIRMENICH
	Nootkatone[d]	PRODASYNTH
	Exaltolide[d]	FIRMENICH

	Vertofix[e]	IFF
	Alpha-Ionone[e]	Sigma Aldrich
	Cashmeran[f]	IFF
	Ethyl linalool[f]	Givaudan
	Ginger ess[f]	Albert Vieille
	Russian leather[g]	NACTIS
	beta-naphthyl methyl ketone[g]	BLH PIM
	Nerol[h]	IFF
	Anisyl Acetate[h]	Givaudan
	Manzanate[i]	Givaudan
	Habanolide[i]	FIRMENICH
	Magnolia flowers ess oil[j]	LMR
	Cedramber[j]	IFF
	Fruit sec ©[k]	FIRMENICH
	Adoxal[k]	Givaudan
	Osmanthus abs[l]	BLH PIM
	Tamarine base[l]	FIRMENICH
	Verdox[m]	IFF
	Cashmeran[m]	IFF
	Veloutone[n]	FIRMENICH
	Viridine[n]	BLH PIM
	Sandalore[o]	Givaudan
	Iso E Super[o]	IFF
	Acetate cis 3 hexenyle[p]	Givaudan

	Rhodinol[p]	Givaudan
	Terebentine ess[q]	FIRMENICH
	Vanillin[q]	SOLVAY
	Scentenal[r]	FIRMENICH
	Miel Blanc ©[r]	PCW
	Bergamot ess[r]	Payan & Bertrand
	Gamma Nonalactone[s]	Sigma Aldrich
	Allyl amyl glycolate[s]	IFF
	Cyclotene[t]	Sigma Aldrich
	Cuminic aldehyde[t]	Givaudan
Recognition T2	Lime EO[u]	Payan & Bertrand
	Bitter orange EO[u]	Robertet
	Verdox[u]	IFF
	Menthanyl acetate[v]	FIRMENICH
	Vertofix[v]	IFF
	Decanal[v]	Givaudan
	Linalyl acetate[w]	Sigma Aldrich
	Opoponax EO[w]	FIRMENICH
	Cis-3-Hexenol[x]	PCW
	Ylang EO[x]	Albert Vieille
	Galaxolide[y]	IFF
	Heliotropex[z]	IFF
	Phenoxyethyl isobutyrate[z]	Givaudan
	Isoeugenol[z]	Sigma Aldrich

	Hedione[aa]	FIRMENICH
	Tagetes EO[aa]	BLH PIM
	Benzyl propionate[ab]	Sigma Aldrich
	Sandela[ab]	Givaudan
	Timberol[ab]	BLH PIM
	Terpineol[ac]	FIRMENICH
	Scentenal[ac]	FIRMENICH
	Benzyl salicylate[ac]	Sigma Aldrich
	Methyl salicylate[ad]	Robertet
	Linalyl propionate[ad]	Givaudan
	Phenoxanol[ad]	IFF
	Grapefruit EO[ae]	MPE
	Musk T[ae]	Sigma Aldrich
	Menthone[ae]	MANE
	Litsea cubeba EO[af]	Robertet
	Lilial[af]	PRODASYNTH
	L-Menthol[af]	BLH PIM
	Methyl naphthyl ketone[ag]	BLH PIM
	Ethyl vanillin[ag]	SOLVAY
	Tonalide[ag]	Sigma Aldrich
	Lemon petitgrain EO[ah]	Payan & Bertrand
	Ethyl maltol[ah]	MPE
	Ocimene[ah]	IFF
	Delta-octalactone[ai]	PCW

	Gamma undecalactone[ai]	Sigma Aldrich
	Undecavertol[ai]	Givaudan
	Myrrh EO[aj]	Robertet
	Muscenone[aj]	FIRMENICH
	Aldehyde C12 MNA[ak]	BLH PIM
	Lemon EO[ak]	Albert Vieille
	Linalyl acetate[ak]	Sigma Aldrich
	Coumarin[al]	Sigma Aldrich
	Cyclogalbanate[al]	BLH PIM
	Virginia cedarwood EO[am]	Albert Vieille
	Camphor[am]	Sigma Aldrich
	Cyclamen aldehyde[am]	Givaudan
	Hexyl salicylate[an]	PRODASYNTH
	Black pepper EO[an]	Albert Vieille
	Gamma-terpinene[ao]	PRODASYNTH
	Gelsol[ao]	IFF
Categorization	Cocoa abs	Robertet
	Tonka bean abs	BLH PIM
	Broom abs	Robertet
	Hay abs	Mane
	Noble laurel EO	Payan & Bertrand
	Clary sage EO	Albert Vieille
	Parmanthema K	Firmenich
	Mimosa abs	BLH PIM

	Tobacco abs	Mane
	Coffee abs	Robertet
	Guaiac wood EO	PCW
	Polysantol	Firmenich
	Diacetyl	Sigma Aldrich
	Glycolierral	Givaudan

Table S2. Students' performance profile. List of the 14 metrics significantly modulated by training that were selected from all tasks to create the performance profile of students at each time measurement independently. ADJ: adjectives; BTW: between-student comparisons; REF: reference or expert teacher-related metrics; sim: semantic similarity; WTH: within-student comparison.

Task	**Significant Metrics**
Description task	ADJ_count, NOUN_count, TTR_BTW, sim-WTH, sim-BTW, sim-REF
Evocation task	sim-WTH, sim-BTW
Visual recognition task	-
Odor recognition task	Correct Rejections
Categorization task	ARI-BTW, ARI-REF, words count, TTR-BTW, sim-BTW

Table S3. RF and LSVM hyperparameters tuning. The average value (Mean), the standard deviation (SD), and the most frequently occurring value (Mode) are presented for each parameter in both models.

Model	Parameter	Mean	SD	Mode
RF	n_estimators	190	94.34	100
	max_depth	-1	0	-1
	min_samples_split	4.7	3.58	2
	min_samples_leaf	2	1.34	1
SVM	C	13.51	29.12	1
	max_iter	1000	0	1000