Novel Classification of Mono-Molecular Odorants using Standardized Semantic Profiles

Andrew J. Thieme[1], Daniel Korn[2], Vinicius M. Alves[1], Eugene N. Muratov[1*], Alexander Tropsha[1*]

[1]UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA.

[2]Department of Computer Science, University of North Carolina, Chapel Hill, NC, USA.

*__Corresponding Authors__: murik@email.unc.edu, alex_tropsha@unc.edu.

Keywords: data curation, natural language processing, olfaction, semantic similarity, standardization, verbal scent descriptors.

Abstract

Odorants are typically classified by specially trained individuals using subjective verbal scent descriptors. Herein, we used natural language processing to develop standardized semantic profiles of mono-molecular odorants. We have (i) curated and integrated scent perception data for mono-molecular odorants from 4 online sources; (ii) represented verbal scent descriptors used in these sources as vectors in semantic space; (iii) calculated average semantic distances between vectors representing each mono-molecular odorant and each of the vectors for a set of 27 standard verbal scent descriptors to yield 27-dimensional harmonized odorant semantic profile; and (iv) applied dimensionality reduction techniques to these harmonized profiles, to visualize clustering of odorants with similar semantic profiles. This novel uniform representation of odorants can be employed to transform any subjective verbal description of any odorants into standardized semantic profiles that can facilitate automated classification, structure-odor relationship studies, and design of odorants with the desired scent.

Introduction

Mono-molecular odorants are volatile small molecules that can be perceived through the sense of smell when inhaled through the nose. These molecules should also bind and activate olfactory receptors expressed on the surface of sensory neurons in the olfactory epithelia to qualify as true odorants. Neuronal pathways and higher-order processes mediate scent perception in the brain downstream of activated olfactory receptors (Ache & Young, 2005). Mono-molecular odorants are employed as ingredients in scented products, such as perfumes, colognes, air fresheners, shampoos, soaps, deodorants, food products, aromatherapy products, and even fragrances designed to influence customer behavior in retail or culinary settings (Spence, 2020) but also commonly found outside of scent research or fragrance industry settings.

The global fragrance market had a revenue of over $50 million USD in 2021 (Statista, 2021). Innovative approaches for discovering new mono-molecular odorants with targeted properties should have a profound effect by reducing the cost of production, minimizing environmental impact, and improving toxicological safety profiles of scented products. For example, the replacement of natural mono-molecule '*musk-like*' odorants, which have been historically obtained from animals, with new synthetic molecules can serve both to protect endangered animal species from overhunting (Ahmed et al., 2018), meet increasingly stringent regulatory guidelines (Pistollato et al., 2021), and potentially lower the cost of production for scented product manufacturers.

The chemical structures of mono-molecular odorants determine their interactions with olfactory receptors. Therefore, the subjective scent qualities of mono-molecular odorants are objectively bound to their chemical structures, and the study of structure-odor relationships has long been a critical area of scent research (Rossiter, 1996). Annotation of odorant scent profiles is typically achieved via experimental scent perception-based surveys, where participants are requested to indicate the subjective quality of mono-molecular odorants. Findings from such studies have enabled structure-odorant relationships studies and guided the discovery of the next-generation odorants with targeted scent properties.

These experiments have shown complex results (Kaeppler & Mueller, 2013). To conceptualize the degree of this complexity, one may consider that the human sense of smell has been estimated to distinguish between 1 trillion discrete stimuli (Bushdid et al., 2014). Extensive differences have been observed between subjective ratings of odorant scent profiles. Often, different reviewers use different verbal descriptors of an odorant. More interestingly, the same reviewer may give different scent ratings in response to the same odorant across separate experiments. These scent rating differences are dependent on combinations of genetic, neurological, linguistic, and cultural factors; that influence the detection and description of scent percepts. Simply put, there is a high degree of intrinsic variability in representing scent perception-based data from studies where human subjects performed scent rating tasks (Kaeppler & Mueller, 2013).

Historically, many scent ontologies have been created to fully describe all possible scents. These ontologies were generated based on empirical observations of psychologists, data-driven observations of scent researchers, and the personal experiences and insights of professionally trained perfumers. Unfortunately, none of these ontologies serves as a universal, all-purpose ontology, which perfectly handles any situation (Kaeppler & Mueller, 2013). Instead, researchers select a collection of many different scent ontologies according to their specific interests and task. Recently, natural language processing (NLP) approaches have been used on problems of scent descriptors. For instance, Gutiérrez et al. employed natural language descriptors of mono-molecular odorants as inputs for machine learning algorithms trained to predict numerical descriptors of odorant scent profiles (Gutiérrez et al., 2018).

Indeed, NLP approaches present a natural avenue to standardizing scent perception where words and phrases, i.e., verbal scent descriptors (VSD), are used to indicate odorant scent qualities. Categorical VSD profiles are typically represented as lists of unique VSD terms reported by survey participants. These categorical profiles can include from one to over a dozen unique VSD terms per odorant, but most often are comprised of 3-5 unique terms (Rugard et al., 2021). Sometimes numerical descriptors indicating the relative intensity of verbal scent descriptor terms are used to construct continuous VSD profiles. Continuous VSD profiles can include over a dozen unique VSD terms per odorant and differ from categorical profiles such that numerical values are used to indicate the similarity of odorants to each of the VSD terms included in a given set of profiles, as opposed to sets of VSD terms themselves. It is important to distinguish between the two varieties of VSD profiles, as the use of VSD terms for categorical classification is the natural, and dominant, human mode of scent description; outside of work specifically focused on obtaining and/or analyzing continuous VSD profiles. Therefore, there is semantic information latent in virtually all subjective scent-based data, by the multifarious connections between scent perception and semantic processes (Iatropoulos et al., 2018).

Herein, we have developed and implemented an approach to the harmonization of online scent perception-based data using NLP techniques. More specifically, we have employed a set of 27 standard verbal scent descriptor and represented each odorant by a set of distances between its conventional VSD terms and each of these descriptors to yield harmonized verbal scent descriptor profiles. This novel standardized scent representation system enables straightforward quantitative analysis of scent similarity and further investigations into structure-odor relationships. The approach developed herein can be employed universally to harmonize any odorant VSD profile obtained from different sources, regardless of the idiosyncrasy of VSD terms included in categorical or continuous classifications of odorants.

Materials and Methods

Data Collection

Data sources were selected according to the following criteria: (i) public availability, (ii) inclusion of mono-molecular odorants, and (iii) use of categorical VSD terms to annotate odorant VSD profiles. Chemical names and VSDs assigned to mono-molecular odorants were collected from 4 different data sources: (i) FlavorNet (http://www.flavornet.org/flavornet.html), a database containing VSD profiles and physicochemical descriptors for 738 natural product odorants found in the human environment (Arn & Acree, 1998); (ii) SuperScent (http://bioinf-applied.charite.de/superscent/), a database that contains chemical structures and scent profile description of over 2100 volatile materials (Dunkel et al., 2009); (iii) the Sigma Aldrich Fragrances and Flavors Catalog (https://www.sigmaaldrich.com/industries/flavors-and-fragrances/learning-center/catalog-request.html) (Merck KGaA, Darmstadt, 2019); and (iv) the International Fragrance Association's Fragrance Ingredient Glossary (https://ifrafragrance.org/priorities/ingredients/glossary), which is provided by the International Fragrance Association (IFA), a global representative body of the fragrance industry that seeks to represent the collective interests of the industry (International Fragrance Association, 2020). The brief analysis and comparison between the data sources can be found in Results and Discussion section below.

Dataset Curation

All VSD terms collected were left unchanged, except for converting to lower case and stored as strings in comma-separated lists, such that the raw VSD profile for each odorant was a set of all unique VSD terms used to annotate each odorant. Specific VSD terms, such as *"green tea"* were also left unchanged and presented as phrases, not as single words. Chemical names for mono-molecular odorants obtained from the online sources were used to retrieve chemical structures by utilizing the Chemical Identifier Resolver (CIR) node in KNIME Analytics Platform (KNIME, 2020), which queries the CIR resource (https://cactus.nci.nih.gov), hosted by the National Cancer Institute/National Institutes of Health.

Odorants without defined corresponding mono-molecular chemical structures, such as "*botanical essential oils and extracts*" representing complex products without unique chemical identifiers, were excluded from our curated data tables. Organometallic, ionic, and multi-molecular compounds were also excluded. For the minority of odorant names that were not readily translated to SMILES strings by the CIR node, standard IUPAC names were identified via search on PubChem and used to retrieve SMILES strings.

Mono-molecular structures in the 4 collected datasets were thoroughly curated following the workflows previously developed by our group (Fourches et al., 2016). Chemical structures were standardized using ChemAxon Standardizer (ChemAxon, 2021). Briefly, counter ions were removed and specific chemotypes such as aromatic rings and nitro groups were standardized. Standardized structures were then subject to structure matching to deduplicate reoccurring odorants within each of the 4 data subsets. All the curated data used in this study are available in the Supplementary Material and FigShare (https://figshare.com/articles/software/VSD_Profile_Harmonization_Workflow/18624047).

Structure-Odor Relationship Dataset Integration

The 4 curated data sets described above were merged. Overlapping odorants were identified, and their verbal scent descriptor profiles were combined by concatenating all unique VSD terms used to annotate odorants across online sources. The resultant dataset, initially containing 2,819 unique mono-molecular odorants annotated with VSD profiles to be harmonized, each consisting of one or more of 422 unique VSD terms, is referred to herein as the structure-odor relationship dataset (SORD) (**Table S1**).

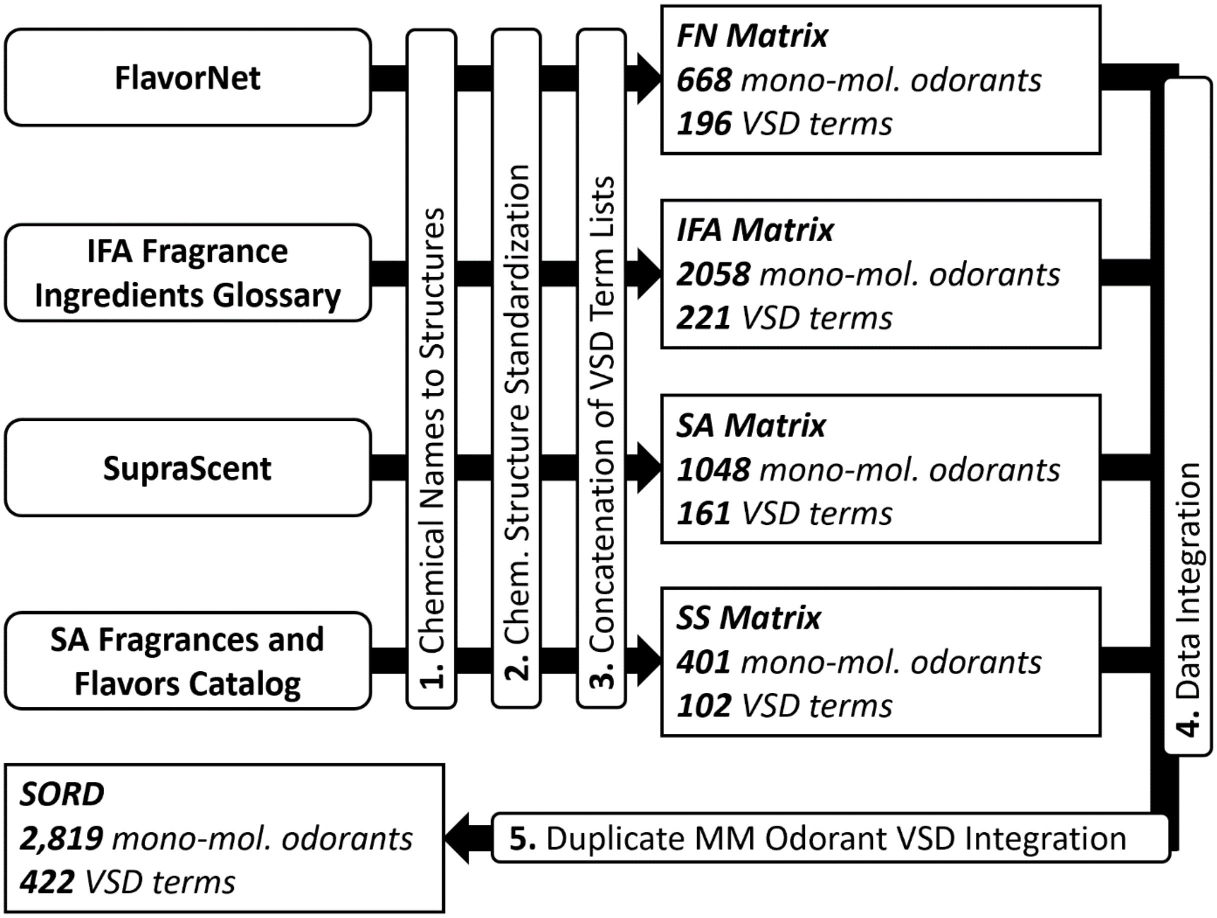


Figure 1. Workflow schema for data collection from online sources, subsequent curation, and integration to form the SORD, which contains raw VSD profiles to be harmonized following the protocol outlined below.

Selection of the Primary IFA Scent Ontology as the Target Scent Ontology

The Primary IFA ontology was created to categorize odorants featured in the IFA Fragrance Ingredients Glossary. Definitions for each of the 27 verbal scent descriptor terms featured in the Primary IFA scent ontology, provided to enhance clarity in the IFA Fragrance Ingredient Glossary, are reproduced below for reference:

1. **Acidic** – *"Acidic means a fragrance note that smells sharp and somewhat pungent. Acidic notes may help boost a citrus note or impart natural qualities."*
2. **Aldehydic** – *"Aldehydes vary: the more diluted they become, the greater the difference in smell. An overarching description is one of clean ironed linen. Aldehydes can be split into more specific profiles, such as citrus or ozonic. They are organic compounds found in natural oils (e.g., orange oil or rose oil) and are used at relatively low doses."*
3. **Amber** – *"Amber is used to describe a complex note in fragrances that are a mixture of warm, woody, and sweet notes that impart a rich and comforting character."*
4. **Animal-like** – *"Animal*-like *notes are important notes used in perfumery. They do not come from animals but are created to give what some would see as a faecal note or a musk note. In dilution, they might help to impart musk notes or floral notes like jasmin."*
5. **Anisic** – *"Anisic materials are those that smell similar to natural aniseed materials like tarragon or fennel."*
6. **Aromatic** – *"Aromatic notes are complex notes that are sometimes also described as having a diffusive aroma. They may be recognized in cooking as culinary herbs and spices, but they have a full fragrance quality."*

7. **Balsamic** – *“Ingredients that smell balsamic tend to have a delicate smell that is slightly sweet and woody and have been termed using natural resins and balsams exuded by some trees and shrubs.”*

8. **Camphoraceous** – *“A fresh, strong and diffusive smell that is characterized by natural camphor and other herbs such as rosemary or marjoram.”*

9. **Citrus** – *“Citrus notes are given by the smell of fruit from the citrus family – such as orange, lemon or grapefruit.”*

10. **Earthy** – *“Earthy notes are reminiscent of earth and mud. They are important when creating a fragrance that needs to impart the full character of a living flower or to give natural outdoor notes – allowing the creation of full landscape (e.g., a bed of roses on a wet day) as opposed to a single or specific smell.”*

11. **Floral** – *“Floral notes belong to the large floral family that includes notes such as rose, jasmin, narcissus, and others. Some fragrance materials have smells that are not one flower but multi-faceted, with a complex flowery character.”*

12. **Food-like** – *“Food-like describes food substances of a savoury or less specific character – such as the smell of roasted vegetables.”*

13. **Fruity** – *“Fruity notes belong to the non-citrus fruit family. This is a very large family that includes many fruit notes such as banana, apple and mango. Some fruit fragrance materials have smells that are note one fruit but multi-faceted, with a complex fruity character.”*

14. **Gourmand** – *“This very important fragrance group has been popular for a number of years – with a food-like smell that is sweet, sticky, or dessert-like. It includes caramel, fudge, chocolate, and meringue.”*

15. **Green** – *“Green is a broad descriptor that refers simply to those natural smell that are green – such as the distinctive scent of cut grass, hedgerow fruits flowers, and those green notes and many green materials that help impart natural smells in a more complex accord or mix of scents.”*

16. **Herbal** – *“Herbal notes include culinary herbs (e.g., thyme, rosemary) that often have a green note and impart fresh nuances to a complex fragrance.”*

17. **Honey** – *“Honey is used to describe materials that have honey characteristics – often sweet and cloying, but sometimes quite harsh and acidic.”*

18. **Marine** – *“Marine covers smells that you expect to find at the seashore – they tend to be fresh and sometimes ozonic, and often sea water-like.”*

19. **Minty** – *“These materials impart mint or menthol notes reminiscent of peppermint and spearmint.”*

20. **Musk-like** – *“These materials belong to an important fragrance note – while they are note obtained from animals, they are created to have an animal-like quality, often powdery and sometimes warm and sweet.”*

21. **Ozonic** – *“Ozonic notes are fresh-smelling materials that don’t have a more specific note but may remind you of a fresh windy day. Sometimes they have a weak, almost chlorine-like smell.”*

22. **Powdery** – *“Powdery fragrance ingredients are from a larger complex group that impart a warm, sometimes sweet or musky powdery smell.”*

23. **Smoky** – *“These ingredients have a smoked or phenolic quality, reminding you of the smell from a bonfire or the smell of food burning.”*

24. Spicy – *"These ingredients belong to a broad spicy family, characterized by many spicy notes from cinnamon to other culinary spices such as pepper, nutmeg, and clove. They sometimes have a sweet note and impart warm nuances to a complex fragrance."*

25. Sulfurous – *"Sulfurous materials have a distinctive smell, reminiscent of onion or garlic. Some sulfur materials may be very pungent and unpleasant at high levels, but when used in a fragrance they may impart citrus or floral notes."*

26. Tobacco-like – *"These ingredients are created to give a smell of tobacco before it has been lit of smoked. They tend to be sweet and warm notes, sometimes with the smell of dried fruit."*

27. Woody – *"Woody notes are part of a large odor family that includes woods such as sandalwood or cedarwood, sometimes with smoky or leather nuances. Often warm and dry notes, they impart a rich complexity that can help a fragrance last longer."*

Generation of Semantic Embeddings for Verbal Scent Descriptors

Semantic embeddings were generated using the ELMo model provided by Google (https://tfhub.dev/google/elmo/3), trained on a one billion word corpus (Peters et al., 2018). ELMo is a word embedding model trained on word vectors derived from a bidirectional long short-term memory, a type of recurrent neural network trained with a coupled language model. Each word is processed relative to other words in the corpus and the model is optimized so that words used in similar contexts have similar resultant descriptors in the embedded semantic space. This model provides a 1,024-dimensional embedding (vector) for every phrase included in its training set.

Here, all 422 unique VSDs featured in the SORD were used as input for ELMo. For each VSD term, ELMo generated a 1,024-dimensional descriptor vector, resulting in a *422 x 1,024* matrix, with 1 row per VSD term; where individual VSD terms (vsd_t) are represented as vectors

(vsd_v), and each column contains a co-ordinate for one of the 1,024-dimensions included in the semantic space occupied by generated embeddings (See **Equation 1**).

Equation 1. $vsd_v = ELMo(vsd_t)$

Similar approach was also employed for embedding the 27 verbal scent descriptor terms featured in the Primary IFA scent ontology, which enabled calculations described in the next section.

Odorant Semantic Projection Calculation

"Odorant semantic projections" (OSPs) mean vectors (osp_c) were calculated from the set of all terms used to represent each of the odorants captured in our integrated SORD. To calculate the osp_c for any odorant, we employed all VSD terms used to describe it in the SORD. We then run each VSD through ELMo, producing a respective VSD vector (See **Equation 1**), summed these vectors and divided the resulting vector by the number of VSD terms used for an odorant in SORD, producing an average osp_c vector (see **Equation 2**).

Equation 2. $osp_C = \frac{\sum_{t \in C} vsd_t}{\#\{t \in C\}}$,

where vsd_t is the ELMo vector representative of VSD term t which is one of 422 possible VSD terms used to describe the specific odorant, and C is the odorant being subjected to VSD profile standardization. This transformation enabled the representation of each odorant in SORD by a single mean osp_c vector in the embedded ELMo space.

Semantic Distance-Based Verbal Scent Descriptor Profile Prediction

The above process yields a single vector per odorant, allowing for representation of discrete, aggregate, and subjective scent perception-based data points associated with mono-molecular odorants into an objective semantic space. Thus, we used osp_C vectors representing 2,819 mono-molecular odorant VSD profiles to calculate the semantic distances between each odorant represented by osp_c (see **Equation 2**), and each of the 27 vsd_v vectors representing the terms included in the Primary IFA scent ontology (see **Equation 1**). The Euclidean ($euc_{matrix}[c,v]$) and cosine ($\cos_{matrix}[c,v]$) distances were calculated ($\boldsymbol{euc_{dist}}(A,B)$,$\boldsymbol{cos_{dist}}(A,B)$) (see **Equations 3-6**) between each odorant embedding and each of the 27 semantic embeddings used to represent the Primary IFA scent ontology. These transformations yield standardized and quantitative VSD profiles to describe the scent of each odorant in the dataset. All protocols employing these equations were implemented in a Python script inside a KNIME workflow (KNIME, 2020).

Equation 3.

$$euc_{dist}(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

Equation 4.

$$euc_{matrix}[c,v] = euc_{dist}(osp_C, vsd_v)$$

Equation 5.

$$cos_{dist}(A,B) = \frac{\sum_{i=1}^{n}(A_i * B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} * \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Equation 6.

$cos_{matrix}[c,v] = cos_{dist}(osp_C, vsd_v)$,

Where in **Equations 3-6,** c represents a unique mono-molecular odorant and v represents a unique VSD term.

VSD Profile Standardization Performance Evaluation

The performance of VSD profile standardization protocols detailed herein were evaluated by calculating the mean reciprocal rank (MRR) for each Primary IFA VSD term. Mono-molecular odorants featured in the SORD with one or more Primary IFA VSD terms included in their VSD profiles were identified. The distance matrices referenced above (see **Equations 3-6**) were then used to rank each of the 27 Primary IFA VSD terms from nearest to farthest to each mono-molecular odorant. For each Primary IFA VSD term, we identified all the odorants containing this term and the reciprocal value was calculated from that rank, from 1/1 to 1/27, where 1 is the nearest rank and 27 is the farthest rank. Taking the average of these reciprocal rank values gives a MRR value for that Primary IFA VSD term. This process was iterated for all 27 Primary IFA VSD terms, using both $\boldsymbol{euc_{matrix}}[\boldsymbol{c}, \boldsymbol{v}]$ and $\boldsymbol{cos_{matrix}}[\boldsymbol{c}, \boldsymbol{v}]$, separately, in order to compare the performance of both distance metrics for harmonization tasks.

Additionally, in order to simulate different scenarios under which VSD profile standardization might be attempted, 3 different types of osp_C vectors were used for MRR calculations. The first set was calculated from all VSD terms in VSD profiles (*'all-in'*). The second was calculated from all VSD terms in VSD profiles, excluding the term that was being evaluated (*'leave-one-term-out'*). The third set was calculated from VSD profiles where all 27 Primary IFA VSD terms were removed (*'all-out'*). As a negative control, the rank of Primary IFA VSD terms for each mono-molecular odorant in the SORD was randomly assigned, i.e., a rank-randomization was used to simulate random guessing by human subjects (*'rank-randomization'*).

Further, we assessed the robustness of this protocol for obtaining harmonized categorical VSD profiles, as opposed to the continuous profiles that result directly from semantic distance calculation (see **Equations 3-6**). This task was executed by establishing the relationship between the number of top-ranking (by semantic distance) Primary IFA VSD terms included in standardized VSD profiles and the percentage of odorants in the SORD with at least one exact match between VSD terms in raw and harmonized VSD profiles.

Semantic-space Visualization

Principal Component Analysis (PCA) (Jolliffe & Cadima, 2016) and *t*-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008) were employed to visualize the distribution of mono-molecular odorants represented by the osp_c vectors as well as 27 Primary IFA VSD terms represented by vsd_v vectors in the 1024-dimensional ELMO semantic space. Both operations were performed in Osiris DataWarrior (Sander et al., 2015), where matrices containing vsd_v and osp_C vectors were used as inputs for PCA and t-SNE, respectively.

Results and Discussion

Construction of the structure-odor relationship dataset (SORD)

In this study, we collected, curated, integrated, and harmonized publicly available scent perception-based data, to yield the structure-odor relationship dataset (SORD). SORD contains 2,819 unique odorant molecules, where odorants were annotated with 1-4 raw VSD profiles, as a function of the number of data sources a given odorant is included in. Each raw VSD profile consists of 1 or more of 422 unique VSD terms used to describe the odorants in the dataset. We observed that the raw VSD profile representative space is sparce. Most unstandardized VSDs present in the dataset describes less than 10 odorants. For instance, the term *"passionflower"* was only present in a single raw VSD profile.

The degree of overlap between unique odorants and verbal scent descriptors found between the 4 online sources is depicted in **Figure *2***. As one can see, the IFA Matrix is the largest, and the

SuperScent Matrix is the smallest, in terms of both mono-molecular odorants and VSD. There exists a degree of non-overlap and of overlap in terms of both unique odorants and VSD terms between all 4 matrices. In total, there were 106 unique mono-molecular odorants, and 41 unique VSD terms, that were commonly featured between all four data sources.

In **Figure *2*A**, it can be observed that each independent data source contributed its own set of unique mono-molecular odorants, which were not present in the other sources. Conversely, each source had a portion of mono-molecular odorants found in 1 or more of the 4 sources. In the first case, the addition of these unique odorants increases the size and chemical diversity of the SORD; along with an increase in coverage of semantic space by the set of all unique terms shared between their raw VSD profiles. In the second case, the combination of raw VSD profiles from multiple sources increases the breadth (coverage across semantic space) and depth (anchoring to key 'landmark' terms in semantic space) of description provided by profiles that are used to annotate mono-molecular odorants present in the multiple sources. In total, there were only 106 out of 2,819 total mono-molecular odorants in the SORD that were present in all 4 online sources.

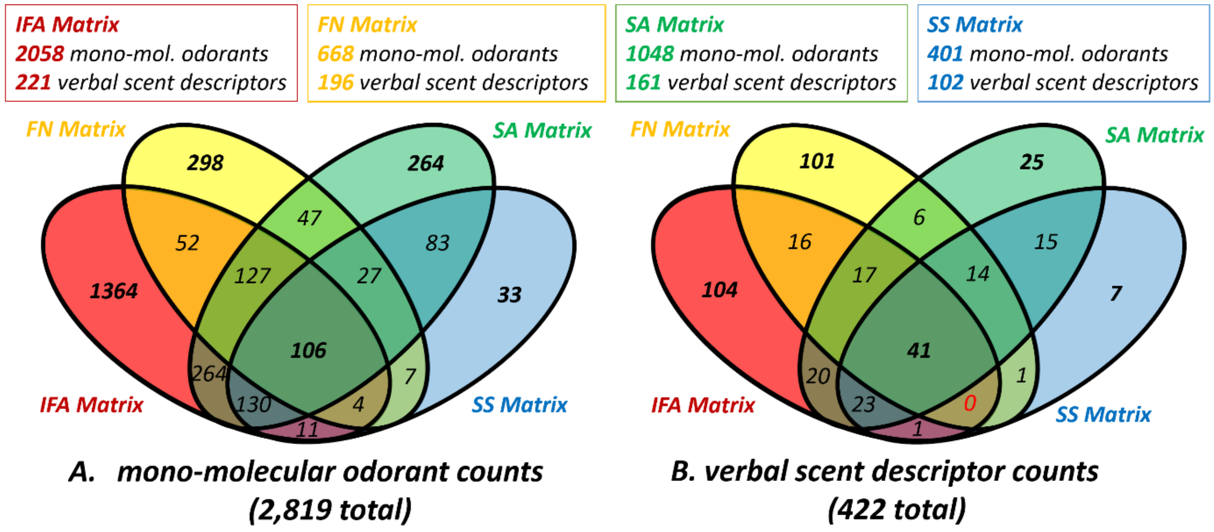


Figure 2. Overlap analysis of curated data sources used in this study and their respective contributions to the SOR Dataset. (A) Unique mono-molecular odorants and (B) unique verbal scent descriptors.

In **Figure 2B**, we can observe a similar pattern for unique sets of VSD terms observed between the datasets. However, in this case, the addition of new unique VSD terms increases the diversity of terms featured in VSD profiles (increases descriptive breadth), while decreasing the conciseness (reduction in depth) of VSD profiles in SORD. In the case of reinforcement of non-unique VSDs, the effect is similar to the addition of non-unique odorants: an increase in the depth and breadth of VSD profiles already featured in the SORD. The above considerations emphasize the need for VSD profile harmonization during the curation and integration of scent perception-based data, as the natural state of these data are both sparse and discrete; and harmonization should both decrease sparsity of descriptor matrices and enable the clustering of unique mono-molecular odorants by their common features, at varying resolutions.

The use of large and unstructured scent ontologies, like those emergent from the raw VSD profiles in datasets like SuperScent and FlavorNet, can provide a high degree of specificity to profile odorants. This high descriptive specificity is valuable for comparing pairs of single odorants, especially in cases where there is partial overlap between scent profiles. Conversely, concise ontologies are useful for comparison between large datasets of odorants, including analysis of data generated in different scent-perception studies. For this reason, SORD employs relatively concise ontology (see Materials and Methods section) to enable the study of discrete aspects of scent perception. However, the task of translating raw VSD profiles to a *'target scent ontology'* is time consuming and requires extensive experience. For this reason, it is humanly impossible to be executed routinely, and NLP approaches were employed as an alternate means to enable the automation of this task.

Restricting VSD profiles to more concise ontologies allows (i) reduction in the number of VSD terms sparsely represented by chemicals in our dataset and (ii) generation of more practical rules and inferences from our model. Pruning terms with a few labeled chemicals increase the profiles' ability to be used more broadly. In this study, the *'target scent ontology'* (Primary IFA) was selected such that odorant VSD profile classification in the SORD is relevant to scent researchers working in academia, and within fragrance industry; as well as interested parties (such as our group) that have not received formal training in scent classification. The Fragrance Ingredients Glossary was generated with the careful attention of trained experts. In addition, their glossary represents the majority of unique odorants integrated into the SORD. For this reason, the Primary IFA scent ontology was selected as the *'target scent ontology'* for our study. This glossary is stated to be "the result of many months work by representatives of large, medium-sized, and small fragrance houses around the world, and was the subject of a global consultation among IFRA members" (International Fragrance Association, 2020).

Standardized Verbal Scent Descriptor Profile Translation

The specific goal of this study was to harmonize raw VSD profiles included in SORD to using sets of natural language descriptors. These descriptors indicate semantic distance from the osp_C vector calculated from arbitrary, unstandardized, VSD profiles used to annotate mono-molecular odorants, to each of the 27 terms defined within the Primary IFA scent ontology. In this way, the original unstandardized scent profiles of odorants used to build the SORD are given a uniform structure, in the context of their relationship to an established set of VSD terms that are of global relevance to fragrance industry, as discussed in the Materials and Methods section of this manuscript.

The accuracy of semantic distance-based calculations to translate VSD profiles to standardized ontologies was assessed by three approaches: *'all-in'*, *'leave-one-out'*, and *'all-out'* (see Materials and Methods section). The results of these accuracy assessments are captured in **Figure 3-*Figure 5*** Unsurprisingly, the performance of standardization using *"all-in"* osp_C vectors resulted in the highest MRR values for each Primary IFA VSD Term (see **Figure 3** and **Figure *4***), in all cases.

There are many instances where the performance of harmonization with *'leave-one-out'* osp_C vectors was lower than the performance of harmonization with *'all-out'* osp_C vectors (see Materials and Methods section). Both *'leave-one-out'* and *'all-out'* values were also equal to, or lower than, results generated randomly, in all cases. These results are important to note, as they indicate the limitations of this type of technique. The removal of specific VSD terms not only reduces the relative influence of that term on calculated osp_C vector outcomes, but also enhances the relative influence of the remaining terms on the resultant vector. In other words, selective removal of target information from input VSD profiles appears to heavily reduce the accuracy of this method. Therefore, it is important that scent ontologies featuring commonly used VSD terms, such as the Primary IFA scent ontology, are employed for the harmonization of VSD profiles.

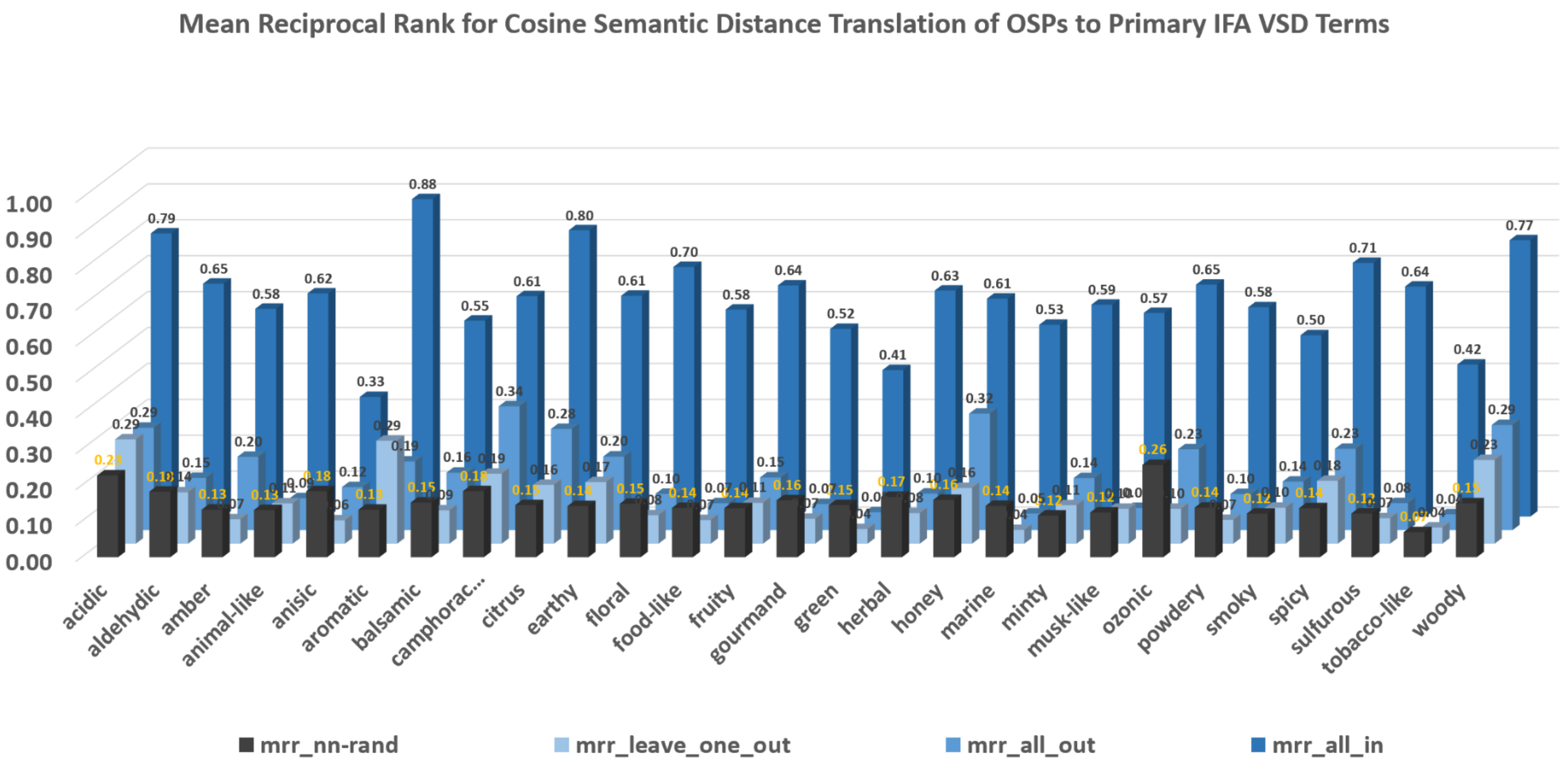


Figure 3. Bar chart depicting MRR analysis results, for cosine semantic distance-based translation of OSPs to each of the 27 Primary IFA VSD terms (See Materials and Methods Section).

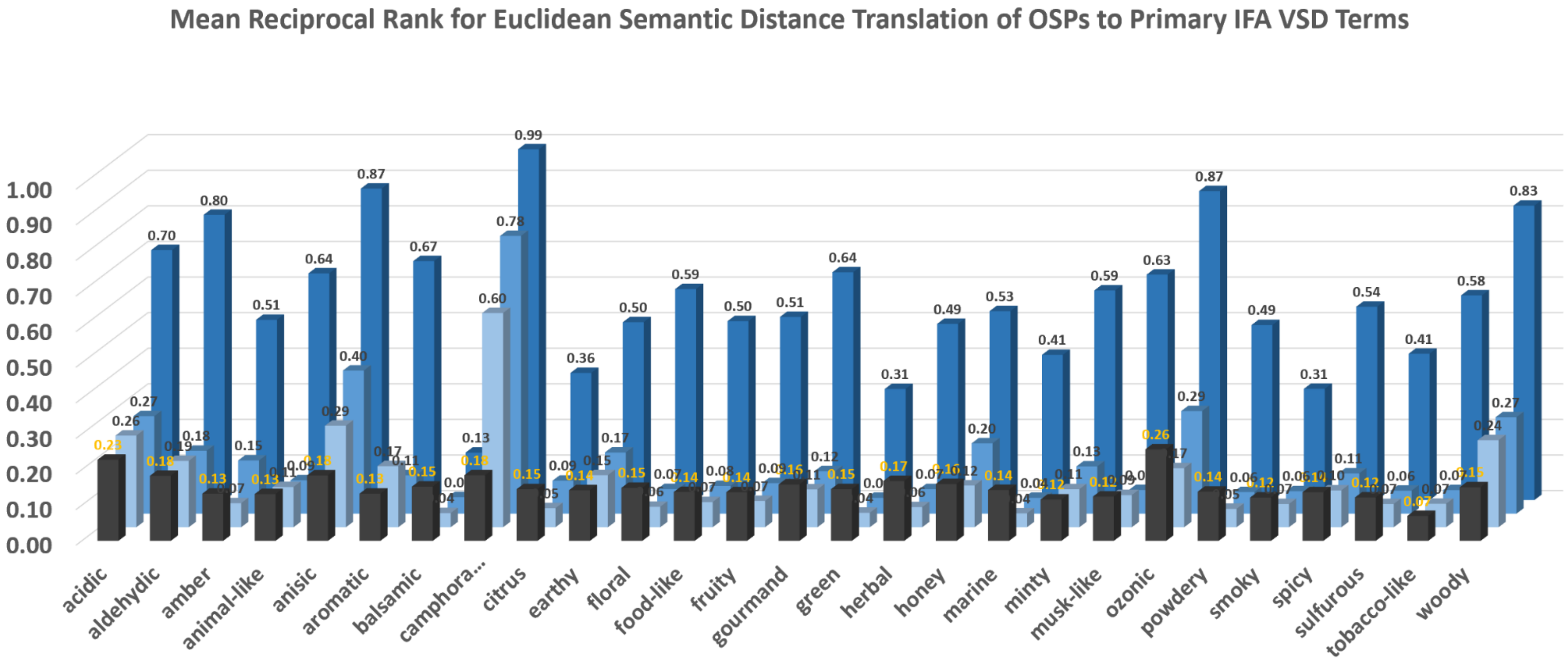


Figure 4. Bar chart depicting results of MRR analysis for Euclidean semantic distance-based translation of OSPs to each of the 27 Primary IFA VSD terms (See Materials and Methods Section).

In **Figure 5** we show the translation validation exercises that compared Euclidean and cosine distance-based VSD profile harmonization to random assignment (*'rank randomization'*), in the context of standardized categorical VSD profiles. Cosine distance outperformed Euclidean distance in this exercise, and it appears to achieve maximal performance within the first five nearest neighbor VSD terms, as opposed to 14 terms observed for the latter. Comparison of both Euclidean and cosine distance approaches to random guessing (randomized ranking, negative control) demonstrates that both methods produce non-random results. The lines in **Figure 5**. Percentage of odorants in the SOR Dataset compared to sets of nearest VSD terms, ranked by semantic distance. represent the percentage of odorants in the SORD with at least 1 exact match between nearest neighbor (Primary IFA) and target (original data) verbal scent descriptor lists, as a function of nearest neighbor list length. At 3 nearest neighbors, at least 75% of the odorant verbal scent descriptor profile translations were validated by the observed exact matches between nearest neighbor and known target terms for both Euclidean and cosine distance.

The total percentage of verifiable odorants, those which have been annotated using one of the 27 Primary IFA verbal scent descriptor terms in at least one of the 4 online resources used in this study, was 87%. The robustness of harmonization for the remaining 13% of odorants without target terms to be matched in this exercise in SORD cannot be known for certain but can be inferred by the performance of translation for odorants with known Primary IFA labels. The overall high prevalence of odorants containing Primary IFA terms in their raw VSD profiles in SORD is another indication that the selection of this ontology was appropriate for the task of VSD profile harmonization via semantic distance-based approach. At 3 nearest neighbors, 85% of the odorant verbal scent descriptor profile translations were validated by the observed exact matches between nearest neighbor and known target terms using cosine distances. Clearly, the use of cosine distances for translation outperforms Euclidean distance in this instance. Accordingly, the final version of the SORD built during this study utilized cosine distance-based translation (see **Table S1**).

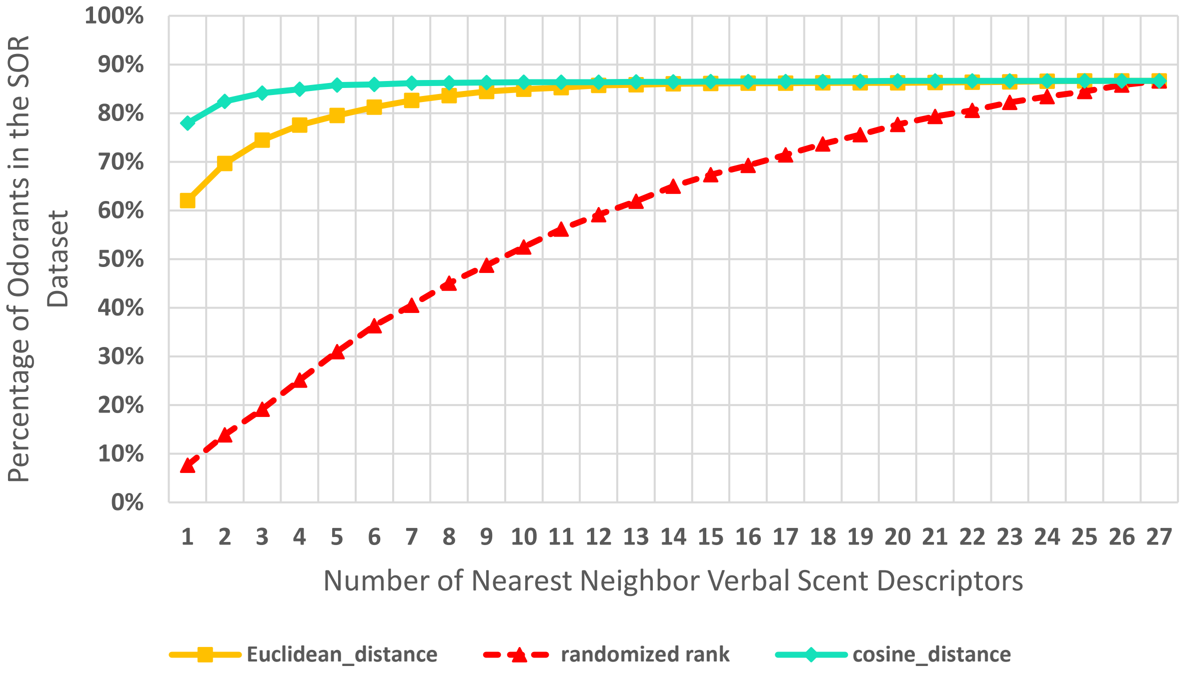
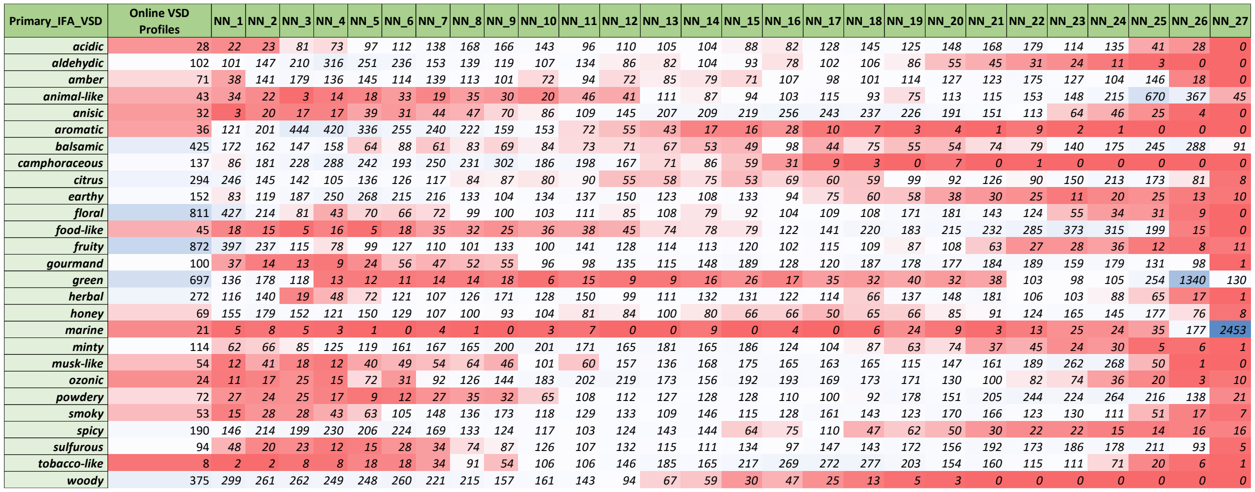


Figure 5. Percentage of odorants in the SOR Dataset compared to sets of nearest VSD terms, ranked by semantic distance.

As evaluation of **Figure *5*** indicates that robustness of harmonization is near maximal at 3 nearest neighbors, and maximal at 5 nearest neighbor VSD terms, it is recommended to emphasize the top 3-5 nearest neighbor VSDs for odorants when interpreting standardized VSD profiles categorically. Compellingly, this finding is in congruence with the external observation by another research group that odorants are most commonly described using profiles consisting of 3-5 VSD terms, as opposed to single verbal scent descriptors (Rugard et al., 2021). Via establishment of a cut-off for ranked VSDs by an integer limit, continuous VSD profiles generated in this study are readily converted back to categorical format, as lists of verbal scent descriptors.

The results of categorical VSD profile semantic distance-based standardization on the SORD are summarized in **Figure 6.** Comparison between the frequency of Primary VSD terms in online VSD profiles and within sets of ranked terms for standardization indicates that use of the top 3-5 ranked terms for odorants does not have a significant impact on the relative number of odorants in the SORD annotated with each term overall. There is a slight variation between each of the 27 VSD term frequencies between the online and the standardized VSD profiles in the SORD, but the distributions are closely aligned.



Primary_IFA_VSD	Online VSD Profiles	NN_1	NN_2	NN_3	NN_4	NN_5	NN_6	NN_7	NN_8	NN_9	NN_10	NN_11	NN_12	NN_13	NN_14	NN_15	NN_16	NN_17	NN_18	NN_19	NN_20	NN_21	NN_22	NN_23	NN_24	NN_25	NN_26	NN_27
acidic	28	22	23	81	73	97	112	138	168	166	143	96	110	105	104	88	82	128	145	125	148	168	179	114	135	41	28	0
aldehydic	102	101	147	210	316	251	236	153	139	119	107	134	86	82	104	93	78	102	106	86	55	45	31	24	11	3	0	0
amber	71	38	141	179	136	145	114	139	113	101	72	94	72	85	79	71	107	98	101	114	127	123	175	127	104	146	18	0
animal-like	43	34	22	3	14	18	33	19	35	30	20	46	41	111	87	94	103	115	93	75	113	115	153	148	215	670	367	45
anisic	32	3	20	17	17	39	31	44	47	70	86	109	145	207	209	219	256	243	237	226	191	151	113	64	46	25	4	0
aromatic	36	121	201	444	420	336	255	240	222	159	153	72	55	43	17	16	28	10	7	3	4	1	9	2	1	0	0	0
balsamic	425	172	162	147	158	64	88	61	83	69	84	73	71	67	53	49	98	44	75	55	54	74	79	140	175	245	288	91
camphoraceous	137	86	181	228	288	242	193	250	231	302	186	198	167	71	86	59	31	9	3	0	7	0	1	0	0	0	0	0
citrus	294	246	145	142	105	136	126	117	84	87	80	90	55	58	75	53	69	60	59	99	92	126	90	150	213	173	81	8
earthy	152	83	119	187	250	268	215	216	133	104	134	137	150	123	108	133	94	75	60	58	38	30	25	11	20	25	13	10
floral	811	427	214	81	43	70	66	72	99	100	103	111	85	108	79	92	104	109	108	171	181	143	124	55	34	31	9	0
food-like	45	18	15	5	16	5	18	35	32	25	36	38	45	74	78	79	122	141	220	183	215	232	285	373	315	199	15	0
fruity	872	397	237	115	78	99	127	110	101	133	100	141	128	114	113	120	102	115	109	87	108	63	27	28	36	12	8	11
gourmand	100	37	14	13	9	24	56	47	52	55	96	98	135	115	148	189	128	120	187	178	177	184	189	159	179	131	98	1
green	697	136	178	118	13	12	11	14	14	18	6	15	9	9	16	26	17	35	32	40	32	38	103	98	105	254	1340	130
herbal	272	116	140	19	48	72	121	107	126	171	128	150	99	111	132	131	122	114	66	137	148	181	106	103	88	65	17	1
honey	69	155	179	152	121	150	129	107	100	93	104	81	84	100	80	66	66	50	65	66	85	91	124	165	145	177	76	8
marine	21	5	8	5	3	1	0	4	1	0	3	7	0	0	9	0	4	0	6	24	9	3	13	25	24	35	177	2453
minty	114	62	66	85	125	119	161	167	165	200	201	171	165	181	165	186	124	104	87	63	74	37	45	24	30	5	6	1
musk-like	54	12	41	18	12	40	49	54	64	46	101	60	157	136	168	175	165	163	165	115	147	161	189	262	268	50	1	0
ozonic	24	11	17	25	15	72	31	92	126	144	183	202	219	173	156	192	193	169	173	171	130	100	82	74	36	20	3	10
powdery	72	27	24	25	17	9	12	27	35	32	65	108	112	127	128	128	110	100	92	178	151	205	244	224	264	216	138	21
smoky	53	15	28	28	43	63	105	148	136	173	118	129	133	109	146	115	128	161	143	123	170	166	123	130	111	51	17	7
spicy	190	146	214	199	230	206	224	169	133	124	117	103	124	143	144	64	75	110	47	62	50	30	22	22	15	14	16	16
sulfurous	94	48	20	23	12	15	28	34	74	87	126	107	132	115	111	134	97	147	143	172	156	192	173	186	178	211	93	5
tobacco-like	8	2	2	8	8	18	18	34	91	54	106	106	146	185	165	217	269	272	277	203	154	160	115	111	71	20	6	1
woody	375	299	261	262	249	248	260	221	215	157	161	143	94	67	59	30	47	25	13	5	3	0	0	0	0	0	0	0

Figure 6. Heatmap summarizing VSD profile categorical harmonization results for SORD. Counts in the cells of each row represent the number of odorants with each Primary IFA VSD as its nearest neighbor VSD. The column "Online VSD Profiles" contains the frequency of each term in the online data used to build the SORD. Every other column represents a degree of nearest neighbor, ranging from one

to 27 nearest neighbors. The distribution of primary IFA VSDs with high values in cells within columns NN_1 through NN_5 bear resemblance to the original distribution of Primary IFA VSDs found in online sources.

In **Figure 7**, t-SNE plots generated from *'all-in'* osp_C vectors are used to visually demonstrate how the protocol described herein results in harmonization of VSD profiles. Each point in this space represents a unique mono-molecular odorant, and the proximity between odorants in this space can be used as a proxy for similarity between VSD profiles. In this way, the VSD profiles of mono-molecular odorants that were once annotated using arbitrary sets of unstandardized terms, are now projected into a space where equal comparisons can be drawn across all odorants within the SORD on the basis of the semantic information latent in their online VSD profiles. Importantly, **Figure 7** shows how the harmonization of online scent-perception based data into standardized VSD profiles with natural language descriptors enables discrete clustering of odorants according to their multi-dimensional scent profiles.

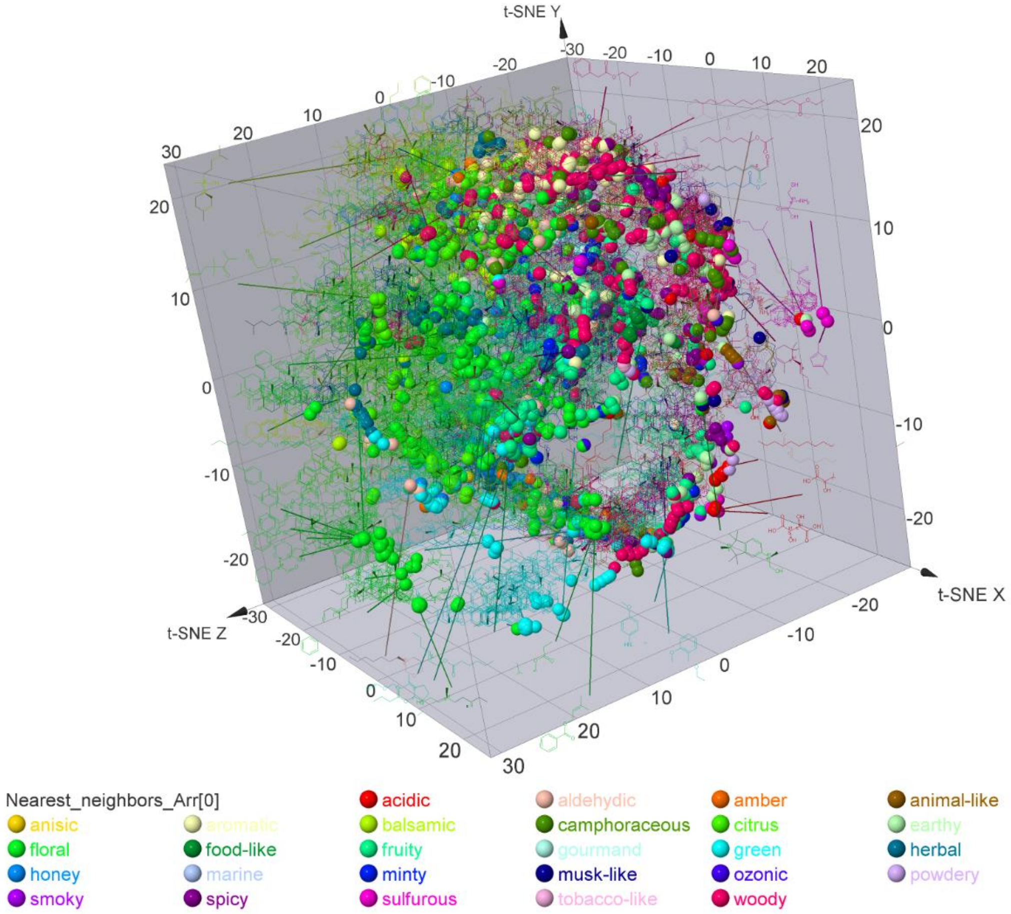


Figure 7. 3D t-SNE plot representation of SORD, where 'all-in' osp_C vectors were used as inputs. Colors of points and chemical structures included in the above plot correspond to the nearest neighbor Primary IFA VSD term to the $\mathrm{OSP_C}$ vectors representative of VSD profiles, calculated via ranking of cosine distances. This categorical labeling scheme shows how the harmonization of online scent-

perception based data into standardized VSD profiles with natural language descriptors enables discrete clustering of mono-molecular odorants according to their multi-dimensional scent profiles. Plots were generated with Osiris DataWarrior, See Materials and Methods Section.

Discussion.

Scent perception-based datasets, like those included in this study, can feature hundreds of different unique VSD terms, such as ‘pungent’, ‘moldy’, ‘warm’, ‘spicy”, ‘cinnamon’, ‘cut grass’, and ‘refreshing’. As a result, VSD profiles are frequently categorical. Alternatively, some studies have generated scent perception-based data, where numerical values within a set range are used to indicate the relative intensity of indicated VSD terms (Dravnieks, 1985).

Restricting odorant description to a limited set of terms is necessary for scent perception-based data integration and grouping of odorants according to a set of formalized, recognizable, scent qualities. This task has proven challenging, and it requires harmonizing scent perception-based data through translation of raw VSD profiles, such that profiles directly reference standardized scent ontologies (Wise et al., 2000). To harmonize raw scent perception based-data, it is necessary to select a fixed set of terms as a *‘target scent ontology’* to limit the resolution of standardized VSD profiles for practical purposes in the context of scientific research. Ideally, the *‘target scent ontology’* is oriented toward broad odorant classification, instead of specific scent descriptions for unique odorants, as this should enable more direct comparison between odorant scent profiles, as they no longer contain specialized or idiosyncratic VSD terms

For example, when studying links between olfaction and the perception of rewarding scent qualities (Haddad et al., 2010; Khan et al., 2007), a highly restricted two-term ontology comprised of the VSDs *“pleasant”* and *“unpleasant”* might be sufficient. However, this two-term ontology would not be adequate for discrete or specific aspects of odorant scent profiles (Manuel Zarzo, 2008, 2012). For example, scent ontologies have been proposed to describe wine aromas. One such case is the work of Dr. Ann Noble (Noble, 2022), who used over 100 unique VSD terms arranged in a hierarchical structure to develop the "wine aroma wheel", specifically designed to describe wine scent profile (Lehrer, 2009). In addition, a historical review of structure-odor relationship studies conducted by Rossiter provides insight into how such efforts fall into two categories; (1) broad or (2) focused in terms of specificity in scent qualities assessed (Rossiter, 1996).

Studies have also been devoted strictly to statistical analysis of the semantic space used to occupied by VSD terms used to indicate odorant scent quality. One study showed that the semantic space of scent description might provide insight into the neurological and psychological structure of olfactory mechanisms in the human brain (M. Zarzo, 2015). The term "semantic space" is used to describe how words relate to each other as vectors in a high-dimensional space, such that the quantifiable proximity between pairs of semantic entities such as words, phrases, sentences, and larger bodies of text, in this space corresponds with their closeness in meaning.

Inspection of PCA plots used to visualize the region of semantic space occupied by the 388 unique VSD terms in raw VSD profiles within the SORD (See **Figure 8**), shows the VSD terms **smoky** and **spicy** very close to each other, while terms like **honey** and **sulfurous** are far apart. Although the compression of high-dimensional space into principal components obscures discrete variations between vectors, trends in similarity across principal components are still observed. The PC1 appears to have a negative correlation with hard and potentially irritating scents (*'acidic'*, *'animal-like'*, *'smoky'*, *'spicy'*, *'sulfurous'*), with gradually more *'fresh'* scents going in the positive direction (*'herbal'*, *'citrus'*, *'balsamic'*) (Manuel Zarzo, 2012). Evidently, it is possible to span regions of semantic space containing hundreds of VSD terms using a rationally selected limited selection of VSD terms like those featured in the Primary IFA scent ontology. Overall, we observe that each term has a distinct set of discrete relationships to other terms.

Modern NLP approaches can revolutionize scent research. Recently, a study reported the development of an automated translation from experimental VSD profiles from a historical study featuring dozens of unique terms, to a secondary set of profiles employing a restricted ‘target scent ontology’ that included 19 terms (Gutiérrez et al., 2018). Authors computed semantic embeddings for experimental VSD profiles and used these embeddings to train a model using elastic net regression algorithms. This model reached an accuracy higher than 70% to predict continuous VSD profiles representative of a scent ontology oriented for structure-odor relationship analysis, for 53 of the 58 odorants used for validation in their study. Ultimately, the authors established a reproducible framework for the accurate translation and harmonization of experimentally obtained scent perception-based data (Gutiérrez et al., 2018).

This study was undertaken to assess the utility of NLP in the harmonization of categorical VSD profiles, which might include anywhere from one to over a dozen unique VSD terms per odorant. Categorical VSD terms are the natural human mode of scent description. In other words, semantic information is latent in virtually all subjective scent-based data under the manifold connections between scent perception and semantic processes (Iatropoulos et al., 2018). Therefore, representation of odorants as entities in semantic space for the harmonization of unstandardized scent perception-based data to standardized VSD profiles should always be feasible. In principle, such a system should harmonize any odorant VSD profile obtained from different sources, regardless of the idiosyncrasy of VSD terms included in both categorical and continuous classifications of odorant scent profiles.

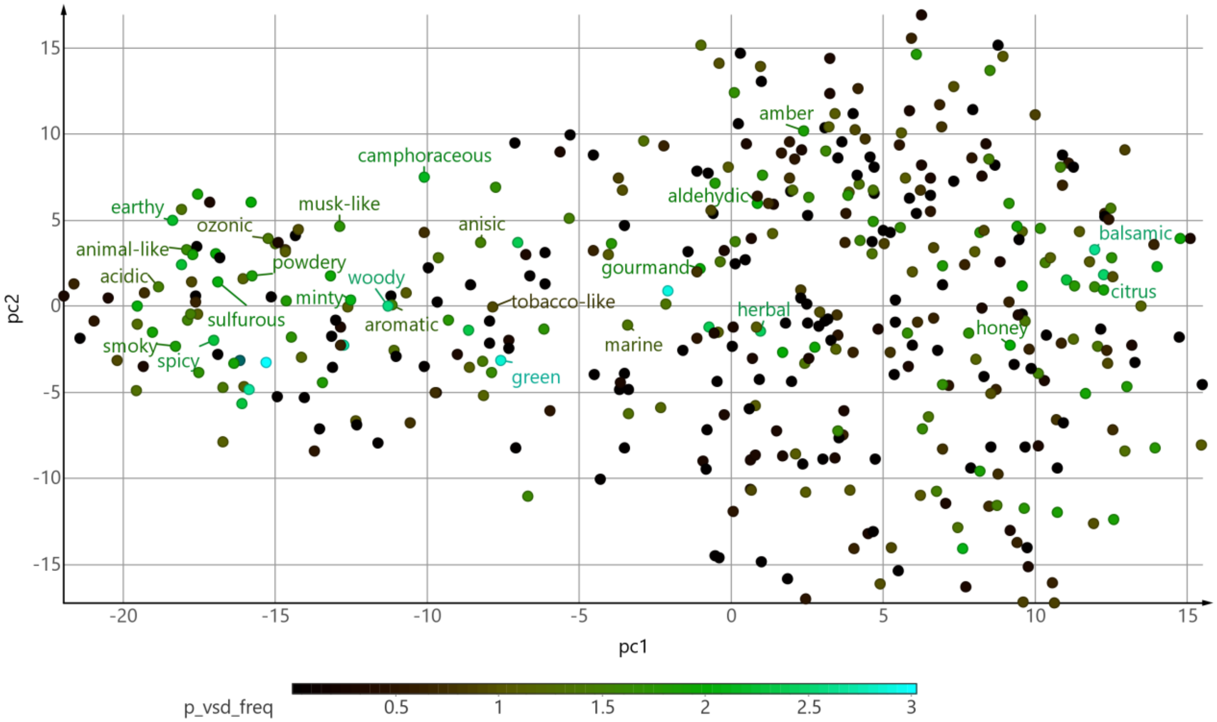


Figure 8. Semantic space analysis of 422 unique VSD terms observed in this study using PCA on semantic vectors 1,024-dimensional space representative of ELMo embeddings. For visualization purposes, we limited labels for points corresponding to the 27 terms included in the Primary IFA scent ontology. Point color corresponds to the log frequency at which each term occurs in the SORD.

Conclusions

Harmonization of raw VSD profiles into a standardized natural language descriptor format enables unified description of mono-molecular odorants. Fortunately, NLP techniques now serve to answer the unmet needs of researchers engaged in on-the-fly collection, curation, and integration of online scent perception-based data into standardized structure-odor relationship datasets, for analytic and predictive chemoinformatic modeling. By relying on objective NLP techniques, such as contextual semantic embedding and distance calculation, the process outlined in this study enables researchers who are not themselves adequately qualified to make classified judgements on the basis of unstandardized VSD profiles, to collect, curate, and integrate online scent perception-based data that will yield new SORD-like datasets standardized for scientific use. In this manner, researchers should be better able to utilize the findings of others in their own studies, despite discrete differences between scent ontologies employed by themselves and others. The framework provided by the process detailed in this can be used by independent researchers to obtain similar results using SORD or any dataset where odorants are characterized by VSD. As a cautionary note, results can vary as a function of the type of semantic embedding, target scent ontology, and distance algorithm selection.

Prospectively, we intend to demonstrate further use the SORD, to develop a QSOR model for oriented toward the discovery of odorants with targeted scent properties. We have provided the entirety of SORD in the supplementary materials section of this manuscript as **Table S1**, which contains both the online and standardized verbal scent descriptor profiles for 2,819 unique mono-molecular odorants. Additionally, the python script used to harmonize online verbal scent descriptor profiles of odorants to user-defined scent ontologies has been provided at https://figshare.com/articles/software/VSD_Profile_Harmonization_Workflow/18624047, as a tool to researchers interested in performing their own scent perception-based data harmonization.

Acknowledgments

This work was supported in part by the NIH grant 3OT2TR003441.

Conflicts of Interest:

AT, VMA, and ENM are co-founders of Predictive, LLC, which develops computational methodologies and software for toxicity prediction. All other authors declare they have nothing to disclose.

References

Ache, B. W., & Young, J. M. (2005). Olfaction: Diverse Species, Conserved Principles. *Neuron*, *48*(3), 417–430. https://doi.org/10.1016/j.neuron.2005.10.022

Ahmed, L., Zhang, Y., Block, E., Buehl, M., Corr, M. J., Cormanich, R. A., Gundala, S., Matsunami, H., O'Hagan, D., Ozbil, M., Pan, Y., Sekharan, S., Ten, N., Wang, M., Yang, M., Zhang, Q., Zhang, R., Batista, V. S., & Zhuang, H. (2018). Molecular mechanism of activation of human musk receptors OR5AN1 and OR1A1 by (R)-muscone and diverse other musk-smelling compounds. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(17), E3950–E3958.

https://doi.org/10.1073/pnas.1713026115

Arn, H., & Acree, T. E. (1998). Flavornet: A database of aroma compounds based on odor potency in natural products. In *Developments in Food Science* (Vol. 40, p. 27). Elsevier. https://doi.org/10.1016/S0167-4501(98)80029-0

Bushdid, C., Magnasco, M. O., Vosshall, L. B., & Keller, A. (2014). Humans can discriminate more than 1 trillion olfactory stimuli. *Science (New York, N.Y.)*, *343*(6177), 1370–1372. https://doi.org/10.1126/science.1249168

ChemAxon. (2021). *ChemAxon - Software Solutions and Services for Chemistry & Biology*. https://chemaxon.com/

Dravnieks, A. (1985). *Atlas of odor character profiles*.

Dunkel, M., Schmidt, U., Struck, S., Berger, L., Gruening, B., Hossbach, J., Jaeger, I. S., Effmert, U., Piechulla, B., Eriksson, R., Knudsen, J., & Preissner, R. (2009). SuperScent--a database of flavors and scents. *Nucleic Acids Research*, *37*(Database), D291–D294. https://doi.org/10.1093/nar/gkn695

Fourches, D., Muratov, E., & Tropsha, A. (2016). Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. In *Journal of Chemical Information and Modeling* (Vol. 56, Issue 7, pp. 1243–1252). American Chemical Society. https://doi.org/10.1021/acs.jcim.6b00129

Gutiérrez, E. D., Dhurandhar, A., Keller, A., Meyer, P., & Cecchi, G. A. (2018). Predicting natural language descriptions of mono-molecular odorants. *Nature Communications*, *9*(1), 4979. https://doi.org/10.1038/s41467-018-07439-9

Haddad, R., Medhanie, A., Roth, Y., Harel, D., & Sobel, N. (2010). Predicting Odor Pleasantness with an Electronic Nose. *PLoS Computational Biology*, *6*(4), e1000740. https://doi.org/10.1371/journal.pcbi.1000740

Iatropoulos, G., Herman, P., Lansner, A., Karlgren, J., Larsson, M., & Olofsson, J. K. (2018). The language of smell: Connecting linguistic and psychophysical properties of odor descriptors. *Cognition*, *178*, 37–49. https://doi.org/10.1016/j.cognition.2018.05.007

International Fragrance Association. (2020). *IFRA Fragrance Ingredient Glossary*. IFRA Fragrance Ingredient Glossary. https://ifrafragrance.org/priorities/ingredients/glossary

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kaeppler, K., & Mueller, F. (2013). Odor Classification: A Review of Factors Influencing Perception-Based Odor Arrangements. *Chemical Senses*, *38*(3), 189–209. https://doi.org/10.1093/chemse/bjs141

Khan, R. M., Luk, C.-H., Flinker, A., Aggarwal, A., Lapid, H., Haddad, R., & Sobel, N. (2007). Predicting Odor Pleasantness from Odorant Structure: Pleasantness as a Reflection of the Physical World. *Journal of Neuroscience*, *27*(37), 10015–10023. https://doi.org/10.1523/JNEUROSCI.1158-07.2007

KNIME. (2020). *KNIME | Open for Innovation*. https://www.knime.com/

Lehrer, A. (2009). Aromas and Wine Wheels. In *Wine and Conversation* (pp. 42–50). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195307931.003.0004

Merck KGaA, Darmstadt, G. and/or its affiliates. (2019). *Flavors & Fragrances Catalog | Sigma-Aldrich*. Sigma Aldrich Website. https://www.sigmaaldrich.com/industries/flavors-and-fragrances/learning-center/catalog-request.html

Noble, A. (2022). *Discover Ann Noble's Aroma Wheel*. https://www.winearomawheel.com/ann-noble-aroma-wheel.html

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, *1*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pistollato, F., Madia, F., Corvi, R., Munn, S., Grignard, E., Paini, A., Worth, A., Bal-Price, A., Prieto, P., Casati, S., Berggren, E., Bopp, S. K., & Zuang, V. (2021). Current EU regulatory requirements for the assessment of chemicals and cosmetic products: challenges and opportunities for introducing new approach methodologies. *Archives of Toxicology*, *95*(6), 1867–1897. https://doi.org/10.1007/s00204-021-03034-y

Rossiter, K. J. (1996). Structure−Odor Relationships. *Chemical Reviews*, *96*(8), 3201–3240. https://doi.org/10.1021/cr950068a

Rugard, M., Jaylet, T., Taboureau, O., Tromelin, A., & Audouze, K. (2021). Smell compounds classification using UMAP to increase knowledge of odors and molecular structures linkages. *PLOS ONE*, *16*(5), e0252486. https://doi.org/10.1371/journal.pone.0252486

Sander, T., Freyss, J., Von Korff, M., & Rufener, C. (2015). DataWarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling*, *55*(2), 460–473. https://doi.org/10.1021/ci500588j

Spence, C. (2020). Using Ambient Scent to Enhance Well-Being in the Multisensory Built Environment. In *Frontiers in Psychology* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2020.598859

Statista. (2021). *Global: fragrance care market revenue 2012-2025 | Statista*. https://www.statista.com/forecasts/1268484/worldwide-revenue-fragrance-care-market van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2606. https://www.jmlr.org/papers/v9/vandermaaten08a.html

Wise, P. M., Olsson, M. J., & Cain, W. S. (2000). Quantification of Odor Quality. *Chemical Senses*, *25*(4), 429–443. https://doi.org/10.1093/chemse/25.4.429

Zarzo, M. (2015). A Sensory 3D Map of the Odor Description Space Derived from a Comparison of Numeric Odor Profile Databases. *Chemical Senses*, *40*(5), 305–313. https://doi.org/10.1093/chemse/bjv012

Zarzo, Manuel. (2008). Relevant psychological dimensions in the perceptual space of perfumery odors. *Food Quality and Preference*, *19*(3), 315–322. https://doi.org/10.1016/j.foodqual.2007.10.007

Zarzo, Manuel. (2012). What is a fresh scent in perfumery? Perceptual freshness is correlated with substantivity. *Sensors (Basel, Switzerland)*, *13*(1), 463–483. https://doi.org/10.3390/s130100463