

SCENTINFORMATICS: MINING OF STRUCTURE-ODOR RELATIONSHIPS AND
SCENT-RELATED MEDICAL EFFECTS FOR MONO-MOLECULAR ODORANTS

Andrew Joseph Thieme

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Division of Chemical Biology and Medicinal Chemistry in the Eshelman School of Pharmacy.

Chapel Hill
2022

Approved by:

Alexander Tropsha

V. Jo Davisson

Sheryl Moy

Ken Pearce

Bryan Roth

© 2022
Andrew Joseph Thieme
ALL RIGHTS RESERVED

ABSTRACT

Andrew Joseph Thieme: Scentinformatics: Mining of Structure-Odor Relationships and Scent-Related Medical Effects for Mono-Molecular Odorants
(Under the direction of Alexander Tropsha)

In this dissertation, we address the unique challenge of establishing predictive relationships between chemical structure and scent properties of monomolecular odorants, in order to support the discovery of new odorants with targeted properties. This challenge is both difficult and exciting because unlike traditional medicinal agents tested in biological assays, scent properties are characterized by verbal descriptors rather than traditional quantitative metrics such as binding constants or dose-response curves. Thus, the stated challenge requires novel ways of quantifying and harmonizing verbal scent descriptors of odorants to enable the use of cheminformatic techniques for scent research. In Chapter 1, we establish a natural language processing-based technique for harmonizing subjective scent perception-based data. In Chapter 2, we build and validate Quantitative Structure-Odor Relationship models to predict standardized scent profiles from chemical structures. In Chapter 3, we develop a knowledge graph database that integrates biomedical and scent-perceptual data linked to odorants, to enable the exploration of links between olfactory processes and biomedical phenomena. The processes detailed in the three chapters of this dissertation form a singular workflow designed to support odorant discovery research. The protocols developed in this thesis are made publicly available at https://figshare.com/projects/AJT_Dissertation UNC_CH_ESOP_CBMC_2022/137364.

This dissertation is dedicated to my wife, family, friends, mentors, and to God, all of whom have provided me with nurturing love and abundant support to pursue my goals, as well as to continually progress and develop on my path through life.

Words do little to express my gratitude.

With Love, Thank You.

ACKNOWLEDGEMENTS

I would like to formally thank those staff and educators who worked at Woodlin Elementary School in Silver Spring Maryland, Westland Middle School and Bethesda Chevy Chase High School in Bethesda Maryland, who all provided me with a strong foundation for my pursuits in higher education. I would like to acknowledge Dr. Miriam Peters, who taught me through direct exposure how meaningful an impact pharmacists can have on the lives of patients in their care, providing me with inspiration and knowledge of an inner will-power to help others through scientific discovery. I would also like to thank the School of Pharmacy at Purdue University at West Lafayette, Indiana, and especially Dr. V. Jo Davisson, for bringing me into the world of pharmaceutical research through his mentorship during my undergraduate studies.

I am deeply grateful to the Eshelman School of Pharmacy for the education I have received during my time at UNC Chapel Hill. I would like to thank Jon-Michael Beasley, Daniel Korn, Dr. Vinicius Alves, and Dr. Eugene Muratov, each of whom have contributed majorly to the accomplishments detailed in this dissertation. I would also like to thank my dissertation committee for providing me with critical feedback and guidance during the development and execution of my thesis aims. I would like to make a special acknowledgement to Dr. Alexander Tropsha, who has provided me with key insights about the nature of life and science, and who guided my transformation into a professional chem-informatician. Thank you all, as stated above, words do little to express my gratitude.

TABLE OF CONTENTS

LIST OF FIGURES	ix
LIST OF TABLES.....	x
LIST OF ABBREVIATIONS.....	xi
INTRODUCTION.....	1
CHAPTER 1: NOVEL CLASSIFICATION OF MONO-MOLECULAR ODORANTS USING STANDARDIZED SEMATIC PROFILES	7
1.1: INTRODUCTION	7
1.2: MATERIALS AND METHODS	12
1.2.0: <i>Data Collection</i>	12
1.2.1: <i>Dataset Curation</i>	12
1.2.2: <i>Structure-Odor Relationship Dataset Integration</i>	13
1.2.3: <i>Selection of the Primary IFRA Scent Ontology as the Target Scent Ontology</i>	14
1.2.4: <i>Generation of Semantic Embeddings for Verbal Scent Descriptors</i>	18
1.2.5: <i>Odorant Semantic Projection Calculation</i>	18
1.2.6: <i>Semantic Distance-Based Verbal Scent Descriptor Profile Prediction</i>	19
1.2.7: <i>VSD Profile Standardization Performance Evaluation</i>	20
1.2.8: <i>Semantic-space Visualization</i>	22
1.3: RESULTS AND DISCUSSION.....	22
1.3.0: <i>Construction of the structure-odor relationship dataset (SORD)</i>	22
1.3.1: <i>Standardized Verbal Scent Descriptor Profile Translation</i>	24
1.3.2: <i>Discussion</i>	35
1.4: CONCLUSIONS	38

CHAPTER 2: SCENT-INFORMATICS: DEVELOPMENT AND VALIDATION OF QUANTITATIVE STRUCTURE-ODOR RELATIONSHIP MODELS TO PREDICT STANDARDIZED SCENT PROFILES OF ODORANT MOLECULES.....	40
2.1: INTRODUCTION	40
2.2: MATERIALS AND METHODS	44
2.2.0: <i>Datasets</i>	44
2.2.1: <i>Scent Multi-Descriptor Read Across Modeling</i>	47
2.2.2: <i>External Cross-Validation</i>	49
2.2.3: <i>Additional External Validation</i>	49
2.2.4: <i>Statistical metrics</i>	50
2.2.5 <i>Virtual Screening</i>	50
2.3: RESULTS	51
2.3.0: <i>Training Set Curation from SORD</i>	51
2.3.1: <i>External Cross-Validation of QSOR MuDRA Models</i>	54
2.3.2: <i>Independent External Test Set Curation</i>	56
2.3.3: <i>Independent External Validation of QSOR MuDRA Models</i>	57
2.3.4: <i>Virtual Screening of SuperNatural II</i>	58
2.4: DISCUSSION	66
2.4.0: <i>Scent-MuDRA versus MuDRA</i>	67
2.4.1: <i>How to Use Scent MuDRA</i>	68
2.4.2: <i>Validation of QSOR MuDRA Models via Virtual Screening of SuperNatural II</i>	69
2.5: CONCLUSIONS	72
CHAPTER 3: SCENT-KOP: SUBJECTIVE SCENT-PERCEPTION AND BIOMEDICAL DATA KNOWLEDGE GRAPH	73
3.1: INTRODUCTION	73
3.2: MATERIALS AND METHODS	76
3.2.0: <i>Mono-molecular odorants and Verbal Scent Descriptors</i>	76
3.2.1: <i>Integrating Scent Datasets</i>	76
3.3: RESULTS AND DISCUSSION.....	79
3.3.1: <i>Description of Scent-KOP</i>	79

3.3.2: <i>Applications/Case Studies</i>	80
3.4: CONCLUSION.....	83
CONCLUSIONS	85
REFERENCES	87

LIST OF FIGURES

Figure 0.1. Overview schematic of the thesis project detailed in this dissertation..	6
Figure 1.1. Workflow schema for data collection from online sources, subsequent curation, and integration to form the SORD	14
Figure 1.2. Overlap analysis of curated data sources used in this study.	24
Figure 1.3. Box plot capturing distributions of distances between OSP_C vectors..	27
Figure 1.4. Percentage of odorants in the SOR Dataset compared to sets of nearest VSD terms, ranked by semantic distance.....	28
Figure 1.5. Heatmap summarizing VSD profile categorical harmonization results for SORD	31
Figure 1.6. Nearest neighbor Primary IFRA VSD term to the OSP_C vectors representative of VSD profiles, calculated via ranking of cosine distances	33
Figure 1.7. Cosine distances between selected Primary IFRA VSD terms (A. ‘camphoraceous’, B. ‘floral’, C. ‘musk-like’, D. ‘woody’) to OSP_C vectors representative of VSD profiles.	34
Figure 1.8. 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..	38
Figure 2.1. General workflow depicting the three major steps of the study design starting from the Structure-Odor Relationship Dataset (SOR)	44
Figure 2.2. Bar chart with counts of frequency of each unique term in the raw VSD profiles of SORD odorants.....	52
Figure 2.3. Heatmap showing the overlap between activity classes used to label mono-molecular odorants.	53
Figure 3.1. Resultant SCENT-KOP subgraph is shown in panel “B”, from a query linking the verbal scent descriptor term “peppermint” to the disease node “migraine disorder”.	81

LIST OF TABLES

Table 1.1. Primary IFRA Terms, Definitions, and Counts in online sources (from the IFRA Fragrance Ingredients Glossary (International Fragrance Association, 2020))	15
Table 1.2. MRR for cosine and Euclidean distance based ranking of Primary IFA Terms.....	26
Table 2.1. Summary of results from 5-fold external cross validation	55
Table 2.2. Summary of results from external test set validation using the ' <i>SMILES-to-smell</i> ' dataset.....	58
Table 2.3. Summary of virtual screening results, 1 'hit' compound selected per VSD term in the Primary IFA scent ontology.	60

LIST OF ABBREVIATIONS

ELMo	Embeddings from Language Models
FN	FlavorNet
IFRA	International Fragrance Association
MRR	Mean Reciprocal Rank
NLP	Natural Language Processing
OSP	Odorant Semantic Projection
QSAR	Quantitative Structure Activity Relationship
QSOR	Quantitative Structure Odor Relationship
ROBOKOP	Reasoning Over Biomedical Objects linked in Knowledge Oriented Pathways
SA	Sigma-Aldrich
SORD	Structure Odor Relationship Dataset
SS	SuperScent
S_i	Pairwise Tanimoto Similarity
SRI	Standards and Reference Implementations Component
VSD	Verbal Scent Descriptor

INTRODUCTION

There are many different discrete applications for scented products made available to consumers. Frequent examples of such products include deodorants, soaps, shampoos, colognes, perfumes, food and drink products, air fresheners, scented candles, and aromatherapy products. Each product has a specific scent profile that elicits a certain response from perceivers. Although industrial-scale production of scented products dates back to ancient civilizations, there is still ongoing demand for new odorant materials today. The continued success of the global fragrance industry is enabled by the discovery of new odorants with targeted scent properties.

Historically, discovering new odorant materials has been empirical, and perfumers have operated mainly using heuristic techniques inspired by personal experience and professional training. Currently, a growing body of scent perception-based data is available to researchers. If raw scent perception-based data can be properly collected, curated, and integrated, it can be used to support future odorant discovery. Methods initially developed for applications outside of scent research can be reappropriated to examine, elucidate, and forecast structure-odor relationships. In search of such methods, we used natural language processing, quantitative structure-odor relationship models, and knowledge graph databases to explore structure-odor relationships for mono-molecular odorants, in support of odorant discovery research as the focus of this dissertation.

Chemical structures can be represented numerically using chemical descriptors, enabling their manipulation with cheminformatic techniques and tools. Predicting the activity of chemicals can be more cost and time effective than running physical experiments in different areas of industry. Many predictive cheminformatic techniques, such as quantitative structure-activity relationship (QSAR) modeling, have become a common feature of drug discovery pipelines in the pharmaceutical industry. QSAR models predict chemical activities from chemical structures, functioning as mathematical representations of the medicinal chemist's mantra: "*structure dictates function*". Computational tools like QSAR models are often used for the virtual screening of large libraries of chemical structures to prioritize compounds for experimental testing.

Currently, the optimization of techniques for collection, curation, and integration of subjective scent perception-based data is an active area of research. In contrast to objective biomedical experimental 'activities' often encountered in datasets relevant to pharmaceutical science, 'activity' annotations in odorant scent profiles are largely subjective survey-based data harvested from experiments with human participants. The verbal nature of subjective scent-based data underscores the need for the use of semantic approaches, such as natural language processing (NLP) techniques, as a part of data curation workflows. NLP approaches serve to transform subjective and qualitative verbal data into a quantitative format, enabling the harmonization of raw and unstructured verbal scent descriptors into standardized numerical profiles. Such standardization approaches allow for subsequent cheminformatic investigation of scent perception-based data, where the data can be treated more similarly to traditional biological assay data. This translation from qualitative to quantitative description is a critical feature of the ongoing transition from the empirical discovery of fragrances to the rational design of mono-

molecular odorants, and represents a significant and exciting challenge in the field of cheminformatics.

Core cheminformatics approaches initially developed for drug discovery applications have been adapted for scent research tasks. One key example of such a translation is seen in quantitative structure-*odor* relationship (QSOR) modeling, via translation of QSAR techniques. QSOR models can be used for virtual screening in odorant discovery campaigns, which is analogous to the use of QSAR models in virtual screening during drug discovery campaigns. While the adaptation of certain computational techniques to scent research applications has been straightforward, others have required more fine tuning. Two primary issues, (i) the standardization of subjective scent-based data and (ii) the prediction of standardized odorant profiles from chemical structure data are addressed in Chapters 1 and 2 of this dissertation, respectively.

In Chapter 1, we propose a novel method for harmonizing non-standardized scent profiles of mono-molecular odorants. The proposed method relies on the calculation of semantic similarity between verbal scent descriptor terms. This method allows users to translate non-synchronous verbal scent descriptor labels on odorants into a user-defined set of targeted verbal scent descriptor terms. Herein, we used this method to produce a dataset called the “Structure Odor-Relationship Dataset” (SORD), which contains standardized verbal scent descriptor profiles for 2,819 mono-molecular odorants. After establishing a method for standardization of raw scent perception-based data, it became possible to assign scent class labels to odorants systematically.

Assigning odorants to specific standardized scent classes enabled the assembly of training sets for QSOR modeling. In Chapter 2 of this dissertation, we build a QSOR model that

uses chemical similarity to predict standardized scent profiles of mono-molecular odorants from chemical structures. The models described make predictions based on chemical similarity across multiple chemical descriptor spaces. After external validation of QSOR models built using the SORD, models were used for virtual screening of SuperNatural II, a database containing the chemical structures of over 300,000 natural products. Virtual screening resulted in prioritized sets of small molecules likely to possess predicted scent qualities as indicated by standard verbal scent descriptor terms such as *'amber'*, *'anistic'*, *'floral'*, *'gourmand'*, *'green'*, and *'herbal'*.

To properly utilize discovered odorants for specific applications, it is necessary to investigate how their scent properties relate to a wide variety of discrete phenomena, such as cost and mode of production, emotional or physiological reactions to scent, settings in which they might be perceived, how they might highlight a culinary experience, how they might be used as cosmetic products, how they could provide a warning of nearby danger such as a gas leak, or even how such odorants are related to human health and illness. In Chapter 3 of this dissertation, we develop the SCENT-KOP, a knowledge graph that integrates scent perception-based data with biomedical knowledge. Before mono-molecular odorants can be incorporated safely as 'active' ingredients into scented products, candidate compounds must be screened for human and environmental toxicity. This process can be supported through the bridging of biomedical knowledge and perceptual information about odorant compounds. Formation of the SCENT-KOP knowledge graph enables the exploration of relationships seen between scent profiles and biomedical properties observed in mono-molecular odorants.

In summary, this dissertation exists in a conceptual space, at the interface between cheminformatics and olfaction, wherein we have combined data science, chemistry, scent perception-based data, and biomedical knowledge; to yield a workflow for the discovery of new

odorants with targeted properties (See **Figure 0.1**). In Chapter 1, we establish a natural language processing-based technique for the harmonization of subjective scent perception-based data. In Chapter 2, we build and validate quantitative structure-odor relationship models to predict standardized scent profiles from chemical structures. In Chapter 3, we construct SCENT-KOP, a knowledge graph database that integrates biomedical and scent-perceptual data linked to odorants to enable exploration of relationships between odorants, olfaction, human physiology and disease. The protocols developed in this thesis are publicly available at https://figshare.com/projects/AJT_Dissertation UNC_CH_ESOP_CBMC_2022/137364, along with figures, tables, and supplementary data.

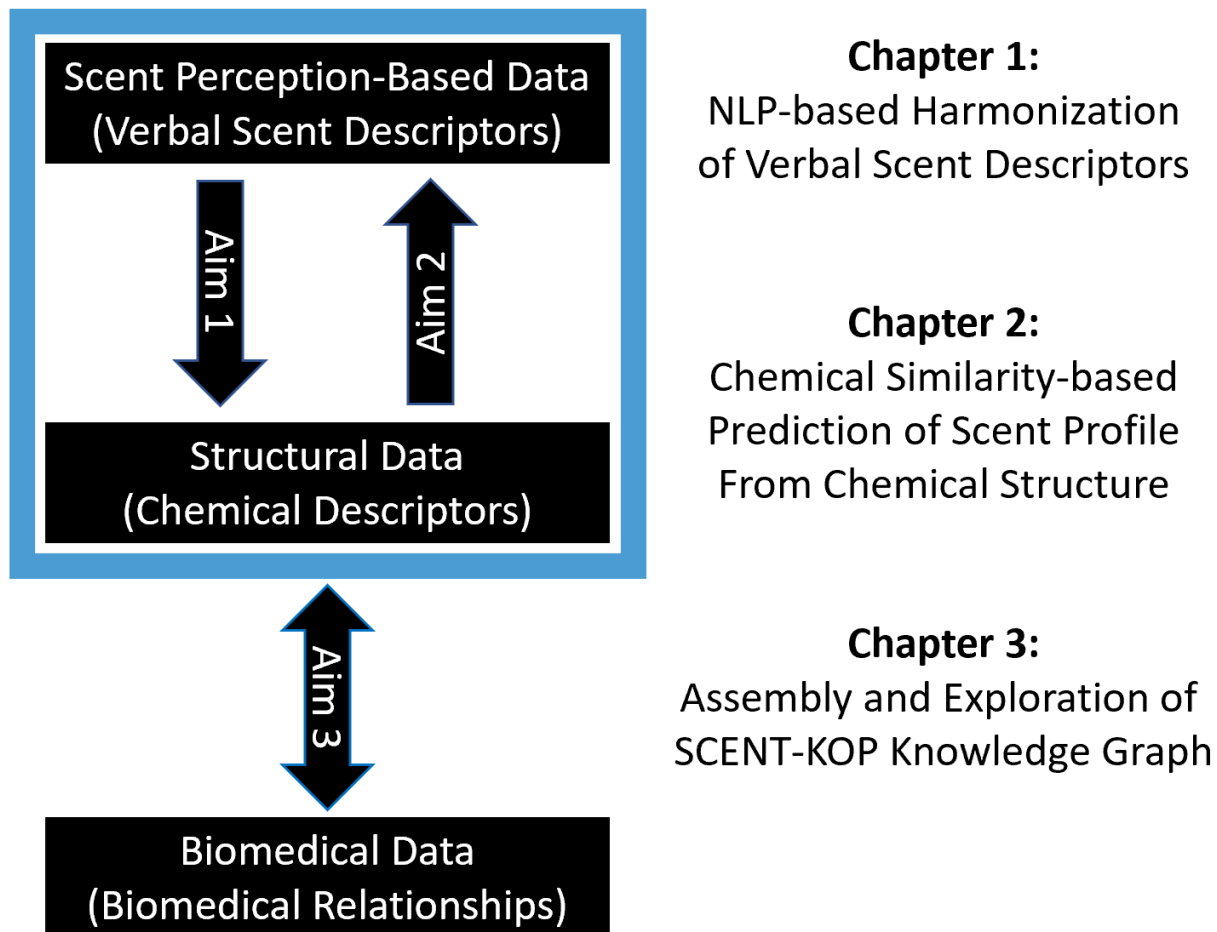


Figure 0.1. The above figure provides an overview schematic of the thesis project detailed in this dissertation. Briefly, Aim 1 is described in Chapter 1, where we outline a natural language processing based technique for the harmonization of raw scent perception-based data used to annotate mono-molecular chemical structures. Then, the pursuit of Aim 2, prediction of standardized verbal scent descriptor profiles from chemical structures, is detailed in Chapter 2. Finally, Chapter 3 highlights the construction of the SCENT-KOP knowledge graph, which is explored to gain deeper understanding and insight into the crossroads between odorants, olfaction, and disease. Overall, the 3 aims of this thesis project are oriented toward new odorant discovery.

CHAPTER 1: NOVEL CLASSIFICATION OF MONO-MOLECULAR ODORANTS USING STANDARDIZED SEMATIC PROFILES

1.1: Introduction

Odorants are typically classified by specially trained individuals using subjective verbal scent descriptors. In this chapter, 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 profiles; 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 into standardized semantic profiles that can facilitate automated classification, structure-odor relationship studies, and design of odorants with the desired scent.

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 and 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). They are also commonly found outside of scent research or fragrance industry settings, as many mono-molecular odorants are naturally occurring.

The global fragrance market is projected to generate a revenue of roughly \$63 billion USD by 2025 (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 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 experiments 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 (Kaepler and 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 for the same 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 representation of scent perception-based data from studies where human subjects performed scent rating tasks (Kaepler and 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 serve as a universal, all-purpose ontology (Kaepler and Mueller 2013). Typically, 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). In contrast to categorical profiles, continuous VSD profiles use numerical values 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).

The use of large and unstructured scent ontologies, like those emergent from the raw VSD profiles in online structure odor datasets like SuperScent (Dunkel et al. 2009) and FlavorNet (Arn and Acree 1998), can provide a high degree of specificity to odorant profiles. 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. 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 increases the profiles' ability to be used more broadly. One well-known example of such concise ontology is

the Primary IFRA scent ontology described in International Fragrance Association's "*The Fragrance Ingredients Glossary*". This Glossary was generated with the careful attention of trained experts (International Fragrance Association 2020). In addition, their glossary represents the majority of unique odorants integrated into the SORD. 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).

For this study, the '*target scent ontology*' (Primary IFRA) was selected such that odorant VSD profile classification in our curated dataset, which we have named the "Structure Odor Relationship Dataset" (SORD), is of relevance 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. Herein, we have developed and implemented an approach to the harmonization of categorical scent perception-based data using NLP techniques. More specifically, we have employed a set of 27 standard verbal scent descriptors 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 classifications of odorants.

1.2: Materials and Methods

1.2.0: 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 and Acree 1998); (ii) SuperScent (Dunkel et al. 2009), 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 (IFRA), 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.

1.2.1: Dataset Curation

All VSD terms collected were left unchanged, except for being converted 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 curated data used in this study are available in the Supplementary Material and can be downloaded from FigShare (https://figshare.com/projects/AJT_Dissertation UNC_CH_ESOP_CBMC_2022/137364).

1.2.2: Structure-Odor Relationship Dataset Integration

The 4 curated data sets described above were integrated. 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) (See **Figure 1.1** and **Table S1**).

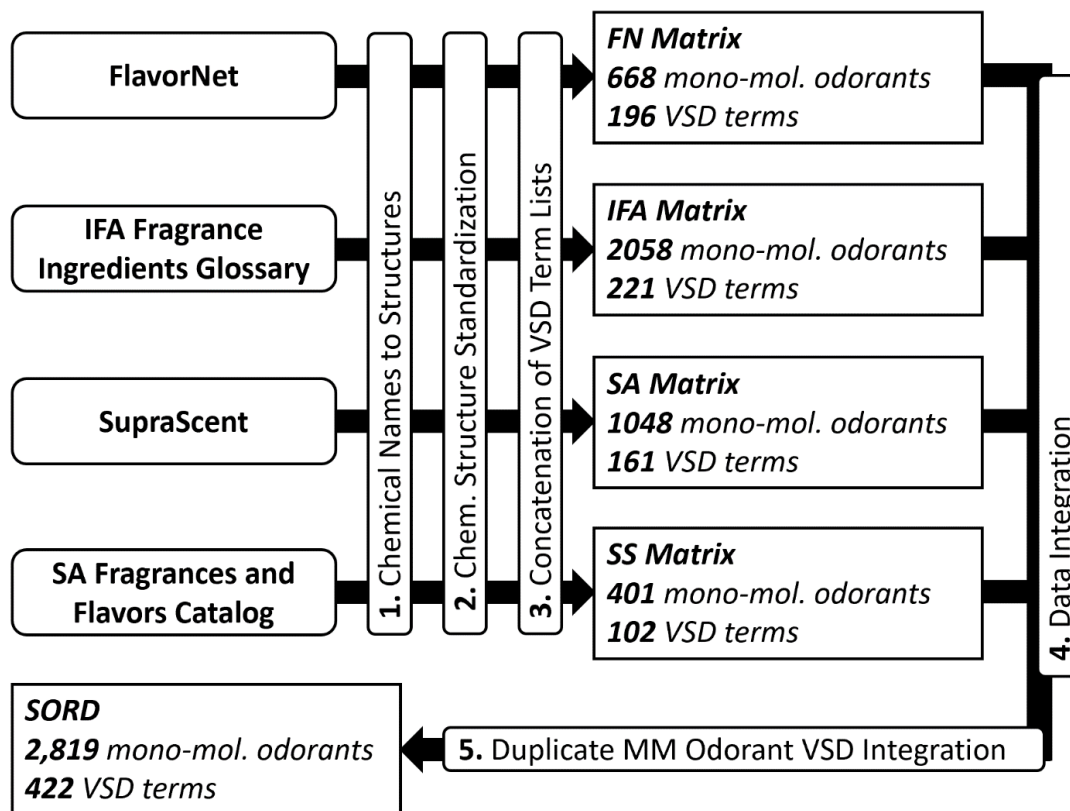


Figure 1.1. Workflow schema for data collection from online sources, subsequent curation, and integration to form the SORD, which contains raw VSD profiles to be harmonized following the protocol outlined below.

1.2.3: Selection of the Primary IFRA Scent Ontology as the Target Scent Ontology

The Primary IFRA ontology was created to categorize odorants featured in the IFRA Fragrance Ingredients Glossary. Definitions for each of the 27 verbal scent descriptor terms featured in the Primary IFRA scent ontology are reproduced below for reference in Table 1.1. Table 1.1 also features 5 additional columns that provide counts of the frequency of each Primary IFRA VSD term in the 4 sources (FlavorNet, FN; Sigma Aldrich, SA; SuperScent, SS; IFRA Glossary, IFRA) and the SORD as a whole.

Table 1.1. Primary IFRA Terms, Definitions, and Counts in online Sources (from the IFRA Fragrance Ingredients Glossary (International Fragrance Association 2020)).

Primary IFRA Term	Primary IFRA Term Definition	FN	SA	SS	IFRA	SORD
Acidic	<i>“Acidic means a fragrance note that smells sharp and somewhat pungent. Acidic notes may help boost a citrus note or impart natural qualities.”</i>	0	6	0	24	30
Aldehydic	<i>“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.”</i>	0	0	0	102	102
Amber	<i>“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.”</i>	0	0	0	71	71
Animal-like	<i>“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.”</i>	0	0	0	43	43
Anisic	<i>“Anisic materials are those that smell similar to natural aniseed materials like tarragon or fennel.”</i>	0	0	0	32	32
Aromatic	<i>“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.”</i>	0	0	0	36	36
Balsamic	<i>“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.”</i>	17	327	0	110	454
Camphoraceous	<i>“A fresh, strong and diffusive smell that is characterized by natural camphor and other herbs such as rosemary or marjoram.”</i>	0	89	8	49	146
Citrus	<i>“Citrus notes are given by the smell of fruit from the citrus family – such as orange, lemon or grapefruit.”</i>	20	93	0	207	320
Earthy	<i>“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.”</i>	0	102	0	58	160
Floral	<i>“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.”</i>	0	200	0	702	902

Primary IFRA Term	Primary IFRA Term Definition	FN	SA	SS	IFRA	SORD
Food-like	<i>“Food-like describes food substances of a savoury or less specific character – such as the smell of roasted vegetables.”</i>	0	0	0	45	45
Fruity	<i>“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.”</i>	0	438	0	628	1066
Gourmand	<i>“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.”</i>	0	0	0	100	100
Green	<i>“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.”</i>	64	205	0	513	782
Herbal	<i>“Herbal notes include culinary herbs (e.g., thyme, rosemary) that often have a green note and impart fresh nuances to a complex fragrance.”</i>	0	0	0	272	272
Honey	<i>“Honey is used to describe materials that have honey characteristics – often sweet and cloying, but sometimes quite harsh and acidic.”</i>	12	38	14	36	100
Marine	<i>“Marine coves smells that you expect to find at the seashore – they tend to be fresh and sometimes ozonic, and often sea water-like.”</i>	0	0	0	21	21
Minty	<i>“These materials impart mint or menthol notes reminiscent of peppermint and spearmint.”</i>	0	57	11	69	137
Musk-like	<i>“These materials belong to an important fragrance note – while they are not obtained from animals, they are created to have an animal-like quality, often powdery and sometimes warm and sweet.”</i>	0	0	0	54	54
Ozonic	<i>“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.”</i>	0	0	0	24	24
Powdery	<i>“Powdery fragrance ingredients are from a larger complex group that impart a warm, sometimes sweet or musky powdery smell.”</i>	0	0	0	72	72
Smoky	<i>“These ingredients have a smoked or phenolic quality, reminding you of the smell from a bonfire or the smell of food burning.”</i>	0	31	15	27	73

Primary IFRA Term	Primary IFRA Term Definition	FN	SA	SS	IFRA	SORD
Spicy	<i>“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.”</i>	0	91	0	115	206
Sulfurous	<i>“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.”</i>	0	67	14	39	120
Tobacco-like	<i>“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.”</i>	0	0	0	8	8
Woody	<i>“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.”</i>	0	108	0	285	393

1.2.4: Generation of Semantic Embeddings for Verbal Scent Descriptors

Semantic embeddings were generated using the Embeddings from Language Models (ELMo) NLP framework developed by Peters and colleagues (Peters et al. 2018). We utilized the implementation provided by Google (<https://tfhub.dev/google/elmo/3>), which was trained on a one-billion-word corpus. ELMo is a word embedding framework which aims to capture the variability in word use (such as syntax) and how many words have contextual meanings based upon surrounding words. The ELMo framework uses long short-term memory (LSTM) neural networks, which take as input a word or another layer of a neural network and output a vector representing the word in semantic space (Peters et al. 2018).

This model provides a 1,024-dimensional embedding vector of floating pointing numbers for every word included in Google’s 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 $422 \times 1,024$ 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. ELMo vectors were also generated for each of the 27 verbal scent descriptor terms featured in the Primary IFRA scent ontology, which enabled the calculations described in the next section.

1.2.5: Odorant Semantic Projection Calculation

“Odorant semantic projections” (OSPs) mean vectors (osp_c) were calculated from the set of all unique 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 this odorant in SORD. If a term occurred in more than one data source for an odorant, the term was only included once in the list of unique terms. We then run each VSD term in the set through ELMo,

producing a respective set of VSD vectors. The mean of each set of vectors is calculated, to yield an average osp_c vector (see **(Eq. 1.1)**);

$$osp_c = \frac{\sum_{t \in C} vsd_t}{\#\{t \in C\}} \quad (\text{Eq. 1.1})$$

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.

1.2.6: 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 (see **(Eq. 1.1)**), and each of the 27 vsd_v vectors representing the terms included in the Primary IFRA scent ontology. The Euclidean ($euc_{matrix}[c, v]$) and cosine ($cos_{matrix}[c, v]$) distances were calculated ($euc_{dist}(A, B), cos_{dist}(A, B)$) (see **(Eq. 1.2)** & **(Eq. 1.3)**) between each odorant embedding and each of the 27 semantic embeddings used to represent the Primary IFRA 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).

$$euc_{dist}(A, B) = \sqrt{\sum_{i=1}^n (A_i - B_i)^2} \quad (\text{Eq. 1.2})$$

$$\text{cos}_{dist}(A, B) = \frac{\sum_{i=1}^n (A_i * B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} * \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (\text{Eq. 1.3})$$

Where in **(Eq. 1.2)** & **(Eq. 1.3)**, A represents an OSP_C and B represents the vector of a targeted VSD term in the Primary IFRA scent ontology.

1.2.7: 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 IFRA VSD term (See **Table 2**). Mono-molecular odorants featured in the SORD with one or more Primary IFRA VSD terms included in their VSD profiles were identified. The distance matrices referenced above (see **(Eq. 1.2)** & **(Eq. 1.3)**) were then used to rank each of the 27 Primary IFRA VSD terms from nearest to farthest to each mono-molecular odorant. For each Primary IFRA 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 IFRA VSD term. This process was iterated for all 27 Primary IFRA VSD terms, using both $\mathbf{euc}_{matrix}[\mathbf{c}, \mathbf{t}]$ and $\mathbf{cos}_{matrix}[\mathbf{c}, \mathbf{t}]$, 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 IFRA VSD terms were removed (*'all-out'*). As a negative control, the rank of Primary IFRA 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'*).

For clarity, VSD terms are removed before averaging corresponding ELMo vectors. This generates an alternate osp_c vector, simulating a scenario where all terms, except each of the terms t in the targeted Primary IFRA scent ontology, are present in the raw data used to annotate mono-molecular odorants. To achieve this scenario, we either remove the single target term (leave-one-out), or all 27 target terms from the raw description of scents in the dataset (all out). For example, taking out the term “balsamic” from odorant raw verbal scent descriptor profiles before assessing the MMR for “balsamic” odorants would be “leave-one-out”. This is done to simulate a scenario where raw data does not include the specific term being assessed; this was repeated once for each term. On the other hand, the “all out” scenario represents one in which all terms in the target ontology are removed.

In order to assess the relationship between semantic distances and odorant scent profile similarity, a secondary set of alternate osp_c vectors were generated from raw verbal scent descriptor profiles for the odorants featured in 2 or more of the sources used to build SORD. If an odorant occurred in n sources, n alternate vectors were generated. For each odorant, average distances between each possible pair of the n alternate (raw) OSP_C vectors were calculated, as well as the average distances from the OSP_C for each odorant and all other odorant vectors in SORD. Then, for each odorant that occurred in 2 or more sources, their average distance to all raw alternate vectors was subtracted from their average distance to all other odorant vectors. In theory, semantic distances between secondary alternate OSP_C vectors for the same odorant should be shorter than average distances between such odorant to the rest of the odorants in SORD.

Further, we assessed the robustness of this protocol for obtaining harmonized categorical VSD profiles (standardized verbal scent descriptor profiles), as opposed to the continuous profiles that result directly from semantic distance calculation (see *(Eq. 1.2) & (Eq. 1.3)*). This task was executed by establishing the relationship between the number of top-ranking (by semantic distance) Primary IFRA 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.

1.2.8: Semantic-space Visualization

Principal Component Analysis (PCA) (Jolliffe and Cadima 2016) and *t*-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton 2008) were employed to visualize the distribution of mono-molecular odorants represented by the osp_c vectors as well as 27 Primary IFRA 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.

1.3: Results and Discussion

1.3.0: Construction of the structure-odor relationship dataset (SORD)

Each raw VSD profile in the SORD 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 sparse. Most unstandardized VSDs present in the dataset describe 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 1.2**. As one can see, the IFRA 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 1.2A**, 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 singular mono-molecular odorants that have replicate records.

In **Figure 1.2B**, 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.

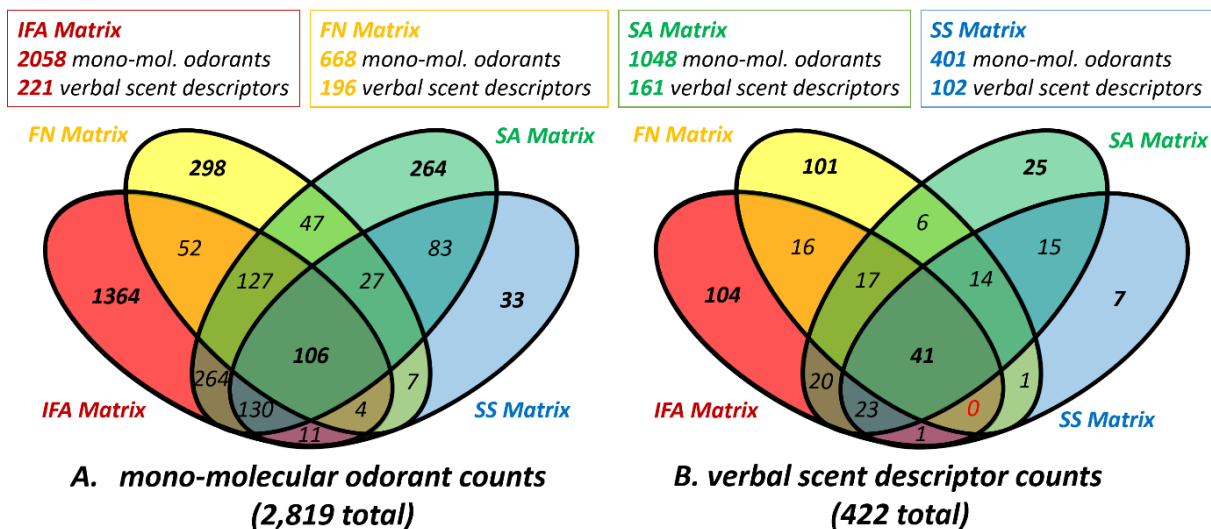


Figure 1.2. Overlap analysis of curated data sources used in this study and their respective contributions to the Dataset. (A) Unique mono-molecular odorants and (B) unique verbal scent descriptors.

1.3.1: Standardized Verbal Scent Descriptor Profile Translation

The specific goal of this study was to harmonize raw VSD profiles included in SORD 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 IFRA scent ontology. The accuracy of semantic distance-based calculations to translate VSD profiles to standardized ontologies was assessed by three approaches: MRR calculation for classification of odorant ‘*all-in*’, ‘*leave-one-out*’, and ‘*all-out*’ OSP_C vectors (see Materials and Methods section). The results of these accuracy assessments are captured in **Table 1.2**. Unsurprisingly, the performance of standardization using “*all-in*” osp_C vectors resulted in the highest MRR values for each Primary IFRA VSD Term, 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 IFRA scent ontology, are employed for the harmonization of VSD profiles (See **Table 1.2**).

Comparison of the average cosine distances between secondary alternate vectors generated for the same odorants (for those odorants occurring in 2 or more of the sources used to build SORD) to average distances from each odorant to all other odorants in the SORD was enabled via plotting of both distributions in a box plot, as well as the difference between each value for each odorant (See **Figure 1.3**). The range observed for average distances to self (distance between alternate odorant vectors) is wider than the range of average distances between all odorant vectors in SORD. While there is a significant overlap between these two distributions, at least 25% of average distances to self are higher than average distances to all other odorants. The distribution of differences between average distances to self and average distances to all for each odorant indicates that there is a symmetrical distribution of distances centered around the mean value of 0.06. Compellingly, 75% of differences between distances are greater than 0, meaning that for the majority of odorants in SORD, semantic distances to self are, in fact, closer than semantic distances to non-self.

Table 1.2. *MRR for cosine and Euclidean distance-based ranking of Primary IFA Terms.*

Primary_IFRA_ vsd_term	cosine distance			euclidean distance			rank_nn- rand
	all_out	leave_1_ out	all_in	all_out	leave_1_ out	all_in	
acidic	0.29	0.29	0.79	0.27	0.26	0.70	0.23
aldehydic	0.15	0.14	0.65	0.18	0.19	0.80	0.18
amber	0.20	0.07	0.58	0.15	0.07	0.51	0.13
animal-like	0.09	0.11	0.62	0.09	0.11	0.64	0.13
anistic	0.12	0.06	0.33	0.40	0.29	0.87	0.18
aromatic	0.19	0.29	0.88	0.11	0.17	0.67	0.13
balsamic	0.16	0.09	0.55	0.05	0.04	0.13	0.15
camphoraceous	0.34	0.19	0.61	0.78	0.60	0.99	0.18
citrus	0.28	0.16	0.80	0.09	0.05	0.36	0.15
earthy	0.20	0.17	0.61	0.17	0.15	0.50	0.14
floral	0.10	0.08	0.70	0.07	0.06	0.59	0.15
food-like	0.07	0.07	0.58	0.08	0.07	0.50	0.14
fruity	0.15	0.11	0.64	0.09	0.07	0.51	0.14
gourmand	0.07	0.07	0.52	0.12	0.11	0.64	0.16
green	0.05	0.04	0.41	0.04	0.04	0.31	0.15
herbal	0.10	0.08	0.63	0.07	0.06	0.49	0.17
honey	0.32	0.16	0.61	0.20	0.12	0.53	0.16
marine	0.05	0.04	0.53	0.04	0.04	0.41	0.14
minty	0.14	0.11	0.59	0.13	0.11	0.59	0.12
musk-like	0.06	0.10	0.57	0.07	0.09	0.63	0.12
ozonic	0.23	0.10	0.65	0.29	0.17	0.87	0.26
powdery	0.10	0.07	0.58	0.06	0.05	0.49	0.14
smoky	0.14	0.10	0.50	0.06	0.07	0.31	0.12
spicy	0.23	0.18	0.71	0.11	0.10	0.54	0.14
sulfurous	0.08	0.07	0.64	0.06	0.07	0.41	0.12
tobacco-like	0.04	0.04	0.42	0.07	0.07	0.58	0.07
woody	0.29	0.23	0.77	0.27	0.24	0.83	0.15

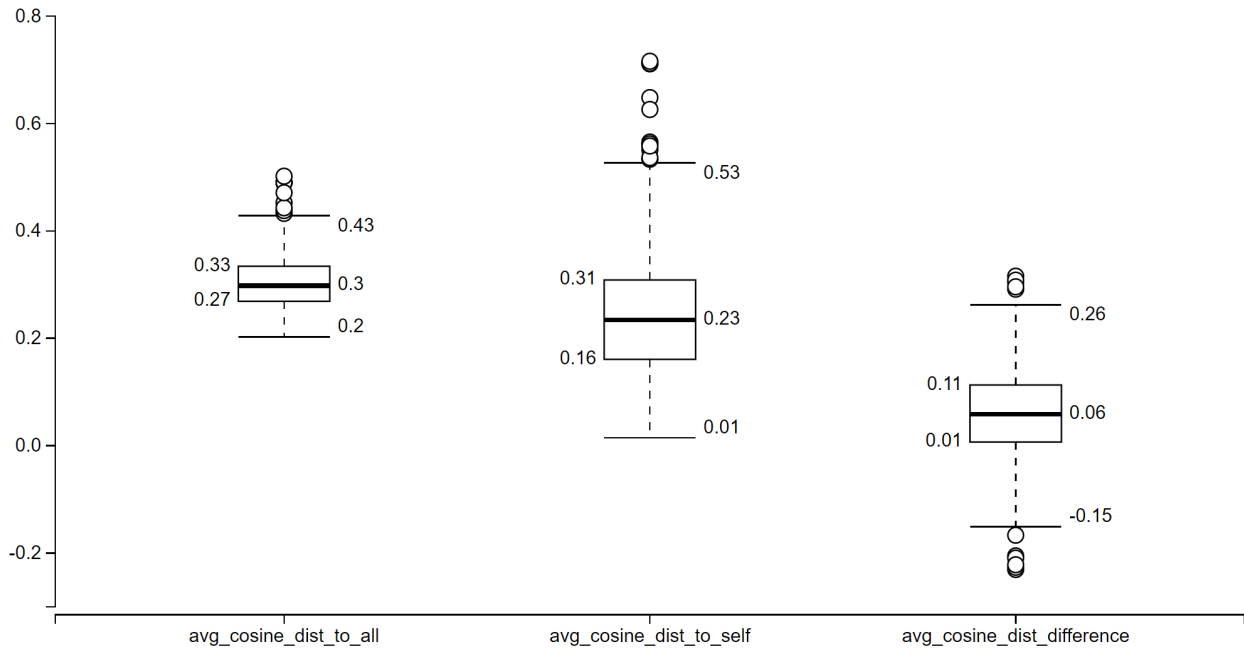


Figure 1.3. Box plot capturing distributions of distances between OSP_C vectors. The label 'avg_cosine_dist_to_all' refers to the average distance from each odorant to each of the other compounds in SORD. 'avg_cosine_dist_to_self' refers to the average distance from each alternate OSP_C vector generated for each odorant that occurred in two or more sources. 'avg_cosine_dist_difference' refers to the difference between 'avg_cosine_dist_to_all' and 'avg_cosine_dist_to_self' for all odorants that occurred in two or more sources.

In **Figure 1.4** we show the translation validation exercises that compared Euclidean and cosine distance based VSD profile harmonization to a random assignment (*'rank randomization'*), in the context of standardized categorical VSD profiles (See Materials and Methods). 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. A 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 1.4** represent the percentage of odorants in the SORD with at least 1 exact match between nearest neighbor (Primary IFRA) and target (original data) verbal scent descriptor lists, as a function of nearest neighbor list length.

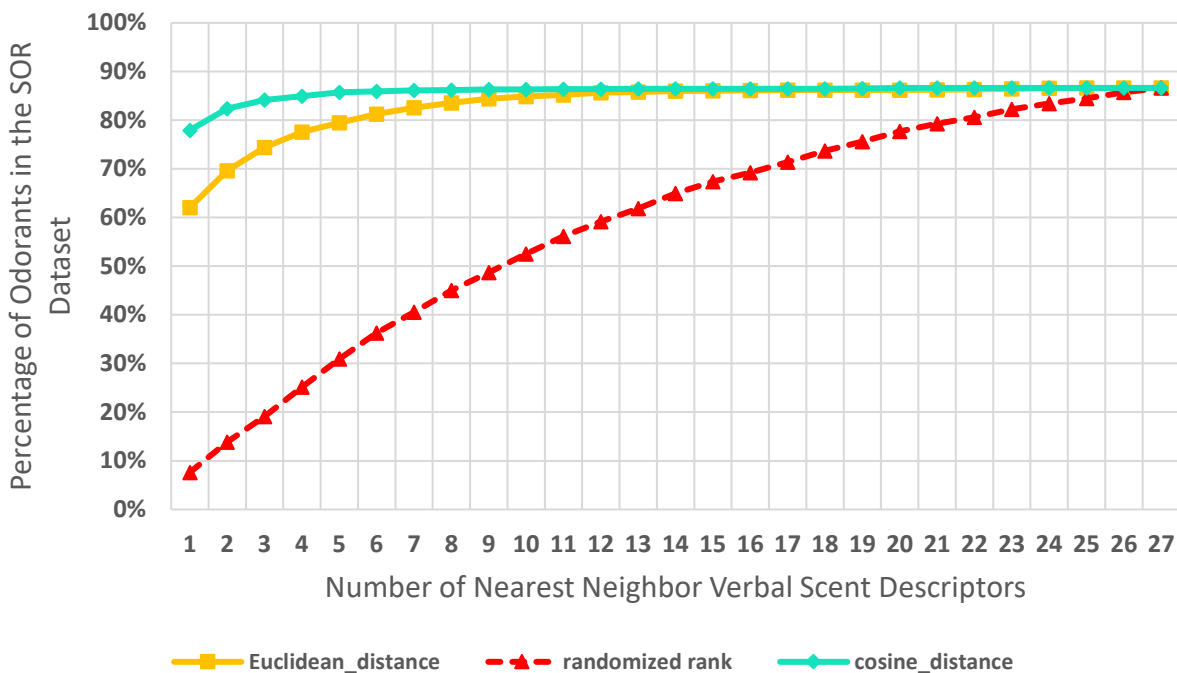


Figure 1.4. Percentage of odorants in the SORD Dataset compared to sets of nearest VSD terms, ranked by semantic distance.

At 3 nearest neighbors, at least 75% of the odorant verbal scent descriptor profile translations were validated by the observed exact matches between the 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 IFRA 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 IFRA labels. The overall high prevalence of odorants containing Primary IFRA 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 a 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**).

As the evaluation of **Figure 1.4** 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 1.5**. 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.

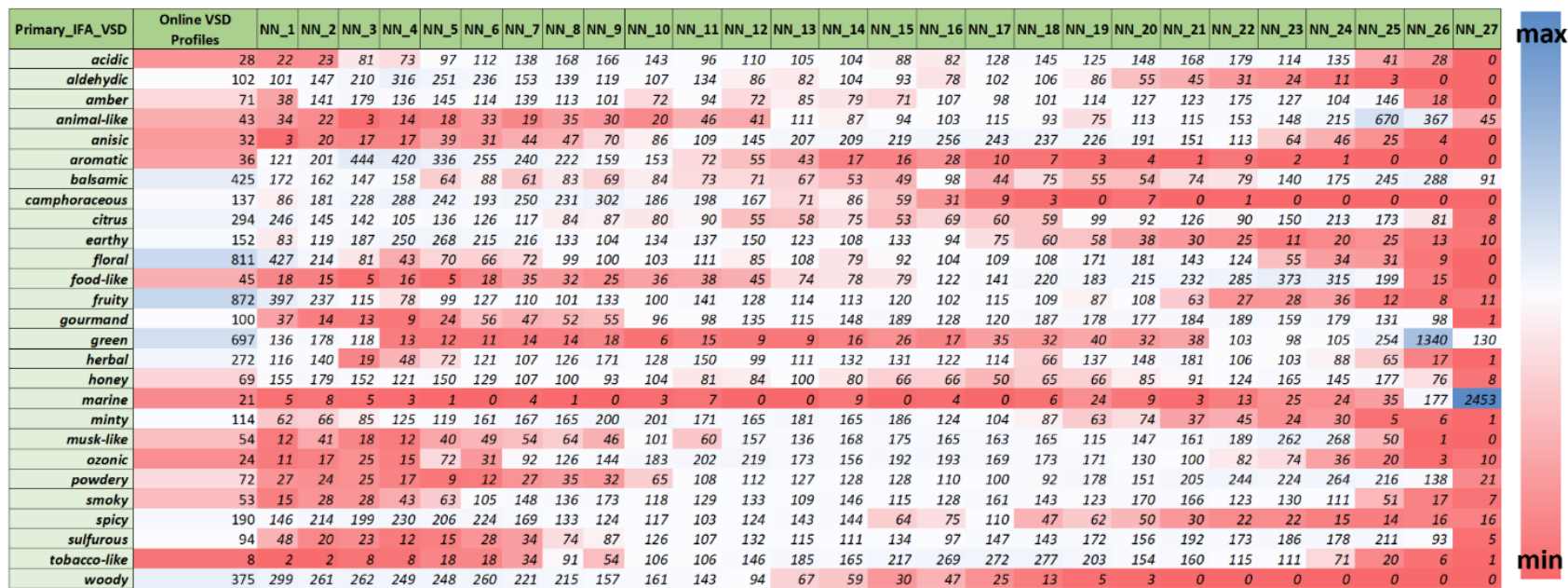


Figure 1.5. Heatmap summarizing VSD profile categorical harmonization results for SORD. Counts in the cells of each row represent the number of odorants with each Primary IFRA VSD as its nearest neighbor VSD. The column “Online VSD Profiles” contains the frequency of each term in the online data used to build the SORD. Every other column represents the frequency of each term in standardized profiles defined by sets of nearest neighbors, ranging from one to 27 nearest neighbor terms. The distribution of primary IFRA VSDs with high values in cells within columns NN_1 through NN_5 bear resemblance to the original distribution of Primary IFRA VSDs found in online sources. Red color indicates cells with low counts, and blue color indicates cells with high counts.

In **Figure 1.6**, 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 the 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. **Figure 1.7** offers an alternative view of the same space afforded by the t-SNE analysis performed to yield **Figure 1.6**. Importantly, both figures show how harmonizing online scent perception-based data into standardized VSD profiles with natural language descriptors enables the discrete clustering of odorants according to their multi-dimensional scent profiles.

Inspection of PCA plots used to visualize the region of semantic space occupied by the 422 unique VSD terms in raw VSD profiles within the SORD (See **Figure 1.8**), 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. This outcome is similar to that which has recently been achieved by (Hörberg et al. 2020), who were similarly able to achieve a coherent mapping of the semantic space occupied by verbal scent descriptor terms via NLP techniques. 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'*) (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 IFRA scent ontology. Overall, we observe that each term has a distinct set of discrete relationships to other terms.

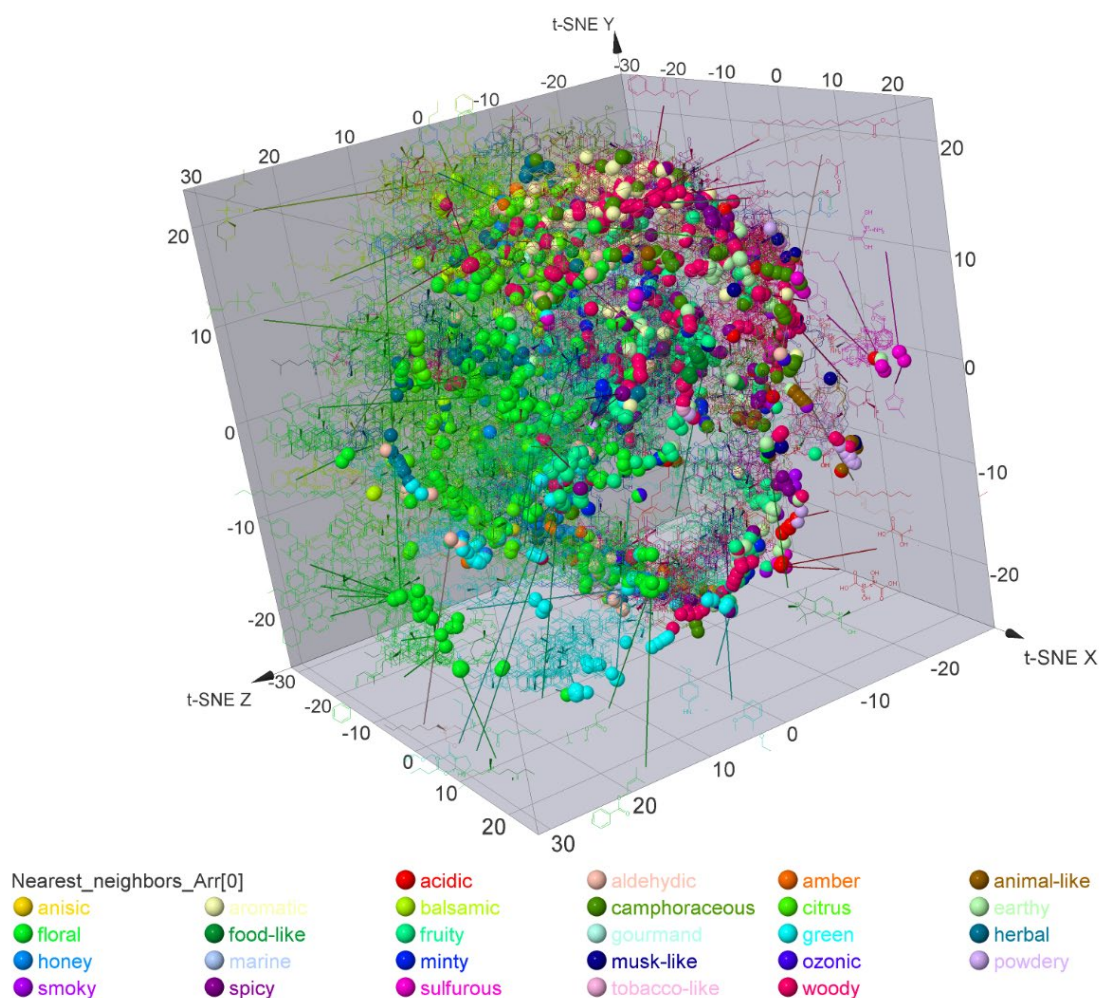


Figure 1.6. Colors of points and chemical structures included in the above plot correspond to the nearest neighbor Primary IFRA VSD term to the OSP_C vectors representative of VSD profiles, calculated via ranking of cosine distances. This categorical labeling scheme shows how the harmonization of online 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.

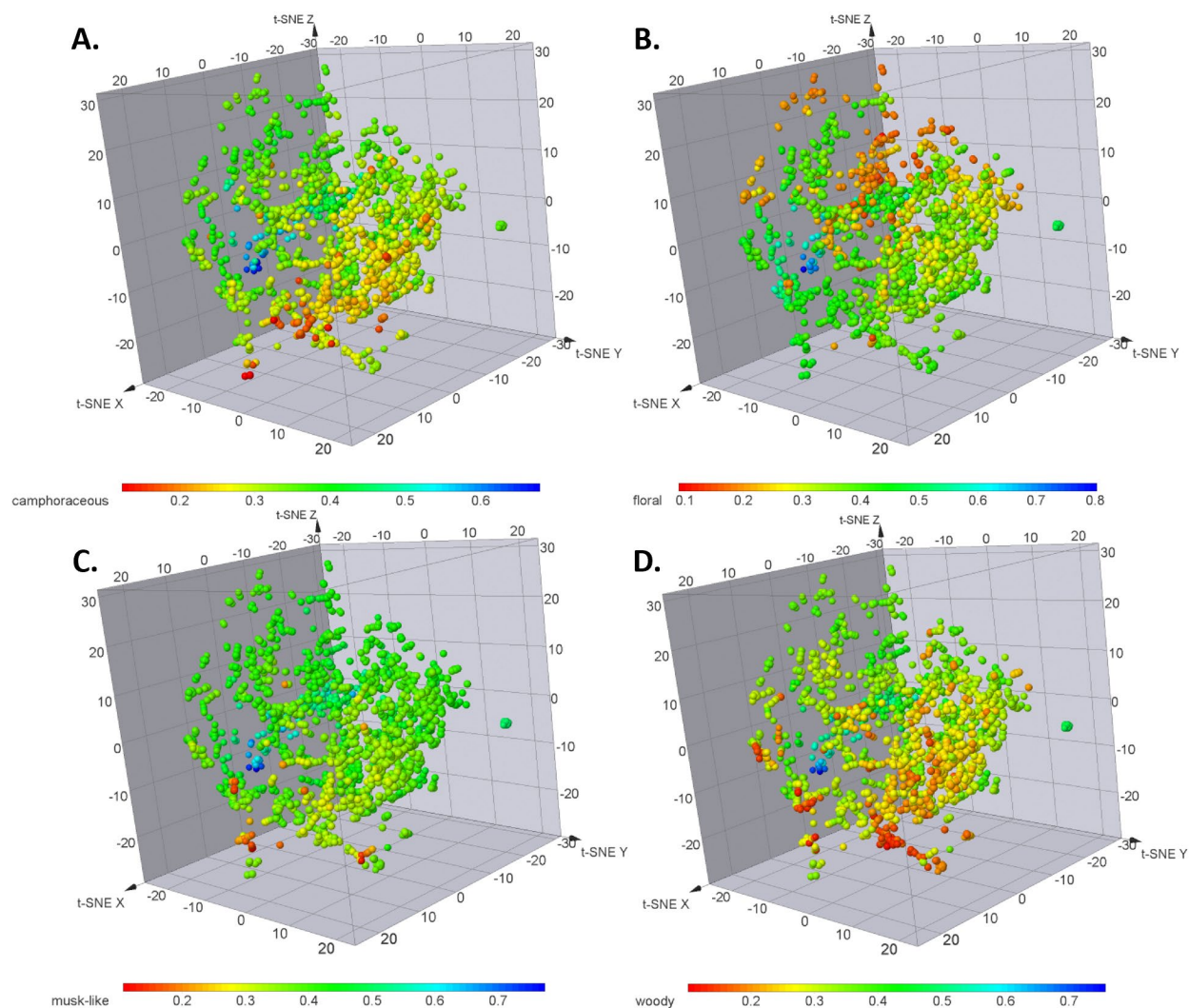


Figure 1.7. Colors of points included in the above plot correspond to the cosine distances between selected Primary IFRA VSD terms (A. ‘camphoraceous’, B. ‘floral’, C. ‘musk-like’, D. ‘woody’) to OSP_C vectors representative of VSD profiles. This visualization scheme further demonstrates how the harmonization 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. Lower distances (more red) indicate higher semantic similarity, and higher distances (more blue) indicate lower semantic similarity.

1.3.2: 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). Because unique VSD terms such as ‘maple’ exist in specific locations within the semantic space defined by ELMo vectors, their effect on the average of vectors from the other terms used to describe an odorant can vary. In essence, the harmonization process captures the singular presence or absence of unique VSD terms in the known record of an odorant, in order to capture as much of the variability in the ways a given molecule has been labeled as possible. For this reason, we do not factor in the relative prevalence of a VSD term, because we were not looking for the most popular way to describe an odorant, but rather to identify all the different ways an odorant has been described.

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 (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 profiles (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).

As discussed in the introduction of this chapter, studies have also been devoted strictly to the statistical analysis of the semantic space occupied by VSD terms. 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 (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.

We have put forward a method for mapping arbitrary descriptions used by perfumers and other professions to create a universal map, which can be used in the context of alternative collections of target terms and alternative approaches to NLP. The scope of this study does not

allow for a comprehensive analysis of similar approaches, but instead provides a methodology for our team and others to provide a path forward. It is natural to contemplate the similarity of scent. In our work, we employed concepts of semantic similarity; we did not compare the distance between scents directly. The standardized scent profiles generated using our method can still be used to cluster similar odorants based on observations that odorants share nearest neighbor target terms in semantic space.

Modern NLP approaches can revolutionize scent research. Recently, as mentioned in the introductions section of this chapter, 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. Therefore, the 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 VSD profile

obtained from different sources, regardless of the idiosyncrasy of VSD terms included in both categorical and continuous classifications of odorant scent profiles.

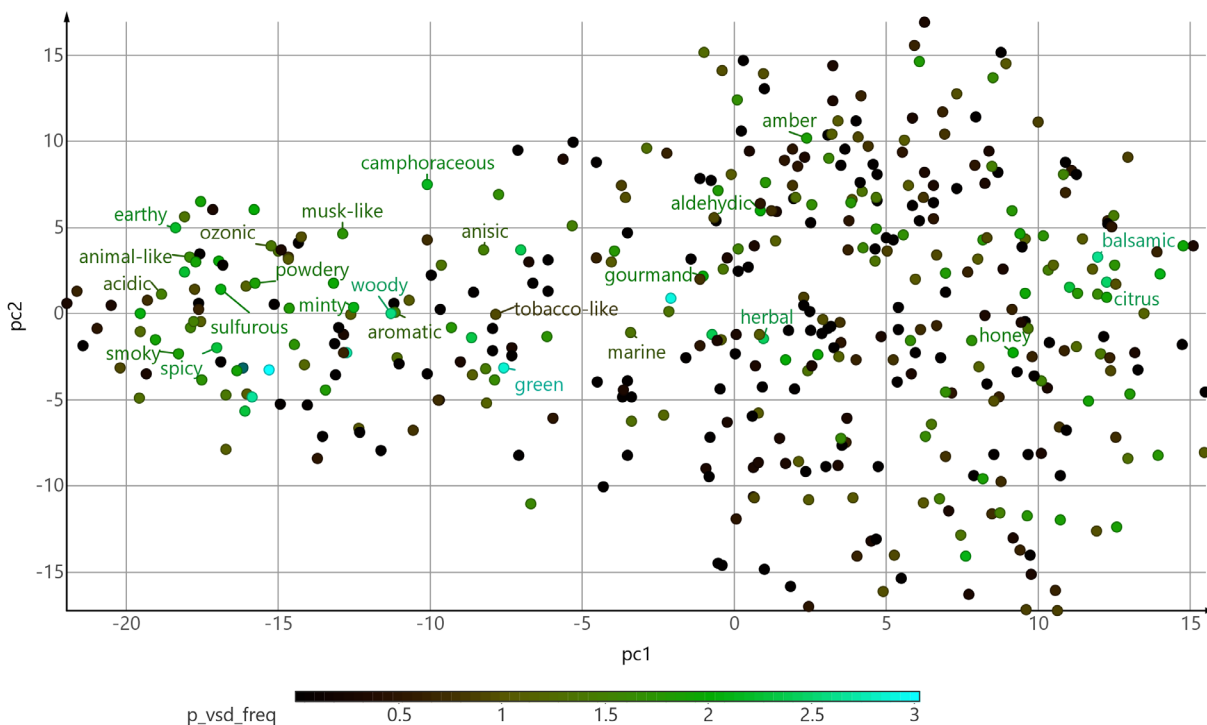


Figure 1.8. *Semantic space analysis of 422 unique VSD terms observed in this study using PCA on semantic vectors 1,024-dimensional space representative of ELMO embeddings. For visualization purposes, we limited labels for points corresponding to the 27 terms included in the Primary IFRA scent ontology. Point color corresponds to the log frequency at which each term occurs in the SORD.*

1.4: 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; and 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. For example, other distance or similarity metrics, such as Pearson correlations, could be calculated with in place of cosine distances; or the semantic embedding algorithm could be a more traditional one, such as Word2Vec (Mikolov et al. 2013). As a cautionary note, results can vary as a function of the type of semantic embedding, target scent ontology, and distance algorithm selection.

In the next chapter, we demonstrate further utility of the SORD, as it is used to develop a quantitative structure-odor relationship (QSOR) model for the discovery of odorants with targeted scent properties. We have provided 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 KNIME workflow used to harmonize online verbal scent descriptor profiles of odorants to user-defined scent ontologies has been provided for public use online at https://figshare.com/projects/AJT_Dissertation_UNC_CH_ESOP_CBMC_2022/137364, as a tool to researchers interested in performing their own scent perception-based data harmonization.

CHAPTER 2: SCENT-INFORMATICS: DEVELOPMENT AND VALIDATION OF QUANTITATIVE STRUCTURE-ODOR RELATIONSHIP MODELS TO PREDICT STANDARDIZED SCENT PROFILES OF ODORANT MOLECULES

2.1: Introduction

Computational modeling techniques have become common practice in drug discovery and chemical safety assessment. Only recently, Quantitative Structure-*Activity* Relationship (QSAR) modeling has been adapted to yield Quantitative Structure-*Odor* Relationship (QSOR) models to predict scent qualities from chemical structures. However, using these methods for the rational discovery of novel scents has remained incipient. Herein, we describe the development of predictive QSOR models employing a modified version of the Multi-Descriptor Read Across (MuDRA) method to achieve statistically validated and interpretable models. To enable model development, we have digitized and quantified standardized verbal scent qualities using natural language processing approaches, and built statistical models to predict these qualities from chemical descriptors of scents. More specifically, we **(i)** employed Primary International Fragrance Association (IFA) scent ontology comprising 27 standard terms to describe any scent in the form of semantic distances between any discretionary verbal scent descriptors and each of these 27 IFA terms; **(ii)** built multi-objective MuDRA models predicting 27-positional verbal scent profile from chemical descriptors of the scents and employed 5-fold *external* cross-validation and an independent test set for model validation; and **(iii)** used the validated models for virtual screening of the *SuperNatural II* database of natural products to identify mono-

molecular odorants with specific scent qualities and evaluated selected predictions with independent assessment. All curated data, Python scripts, and KNIME workflows used in this study are publicly available at

https://figshare.com/projects/AJT_Dissertation_UNC_CH_ESOP_CBMC_2022/137364.

In both the pharmaceutical and fragrance industry, research progress and commercial success depend on the discovery of new molecular entities with targeted properties. Since the seminal development of mathematical models that describe Quantitative Structure-Activity Relationships (QSAR) (Hammett 1935; Hansch et al. 1962), QSAR modeling and related computational techniques have evolved and become integral to drug discovery both in the academic and industrial settings. QSAR modeling has enabled the enhancement of drug discovery pipelines via virtual screening, a process where hit molecules are identified from large libraries of chemical structures to be tested experimentally (Cherkasov et al. 2014).

In pharmaceutical studies, the targeted drug properties are typically objective (*i.e.*, receptor binding). Conversely, the evaluation and comparison between odorant chemical structures have been traditionally performed heuristically by trained chemists and perfumers to establish discrete sets of rules defining specific structure-odor relationships (Rossiter 1996). Nonetheless, in recent years, computational methods of scent quality assessment have been introduced. One early example of this transition from heuristic to computational techniques described the use of least squares regression algorithms to predict perceived odorant '*pleasantness*' from chemical descriptors (Khan et al. 2007). In addition, multiple publications have reported the recent development of Quantitative Structure-*Odor* Relationship (QSOR) models (Gutiérrez et al. 2018; Keller et al. 2017; Khan et al. 2007; Kowalewski and Ray 2020; Sanchez-Lengeling et al. 2019; Sharma et al. 2021). Fundamentally, these studies have

consolidated QSOR modeling as a new branch of cheminformatics that we can define as *scentinformatics*, an area of research primarily focused on identifying relationships between the chemical structures and scent qualities of odorant materials.

QSOR techniques have emerged naturally at the interface between cheminformatics and scent perception research via translation of traditional QSAR techniques used for virtual screening during drug discovery campaigns. However, while the representation of chemical structures as computed chemical descriptors is common in both cases, formal representation of the target biological endpoint is naturally more challenging for scents, as scent description and categorization are based on verbal perception. Notably, a publication by Keller et al. presented a series of models designed to predict numeric descriptors that capture the relative intensity of odorant scent profiles, using 21 unique Verbal Scent Descriptor (VSD) terms (Keller et al. 2017).

A recent observation (Rugard et al. 2021) that 3-5 VSD terms describe the majority of odorants encountered in scent perception databases implies that predicting a scent quality defined by singular VSD terms will only partially describe the scent profiles of mono-molecular odorants. While prediction of singular scent qualities such as *'pleasantness'* is undoubtedly useful both from scientific and commercial viewpoints, it is even more so for predictions of comprehensive scent profiles, which include a set of multiple VSD terms commonly used to classify odorants. The use of machine learning techniques to predict scent profiles capturing a range of widely recognizable scent qualities seen in models described in the literature (Keller et al. 2017; Rugard et al. 2021) serve as critical examples of recent advancements in QSOR modeling techniques.

While many of the QSOR models featured in the studies indicated above reported acceptable metrics of predictive accuracy for some VSD terms, there still exist many terms

which are not as readily predicted with high accuracy. Part of this issue may stem from the fact that many odorant datasets are imbalanced. The lack of consistency between VSD terms used to label odorants across different studies and databases led our (Thieme et al. 2022) and other groups (Gutiérrez et al. 2018) to employ natural language processing techniques to standardize VSD terms and enable the development of QSOR models. Another issue with earlier QSOR models is the presence of activity cliffs, where minor differences in chemical structure cause significant changes in reported VSD terms. To minimize the possible effects of activity cliffs in any particular chemical descriptor space, we employed a modified version of the multi-descriptor read-across (MuDRA) method (Alves et al. 2018a) that averages the assessment of target properties of chemicals made in different descriptor spaces.

Herein, we describe the training and validation of a novel QSOR model that predicts standardized VSD profiles where the standard terms proposed by the International Fragrance Association (IFRA) have been employed. Furthermore, we employ the developed QSOR models for virtual screening to characterize chemicals lacking experimental scent properties. These predictions can contribute to the discovery of mono-molecular odorants with targeted scent properties. In summary, as a result of this study, we have employed a recently developed method to standardize the verbal description of any scent (Thieme et al. 2022) and devised innovative QSOR models to predict the standardized scent profile of any mono-molecular odorant from its chemical structure. These two approaches are integrated into a single workflow, publicly available at

https://figshare.com/projects/AJT_Dissertation_UNC_CH_ESOP_CBMC_2022/137364.

2.2: Materials and Methods

The general workflow employed in this work is depicted in **Figure 2.1**.

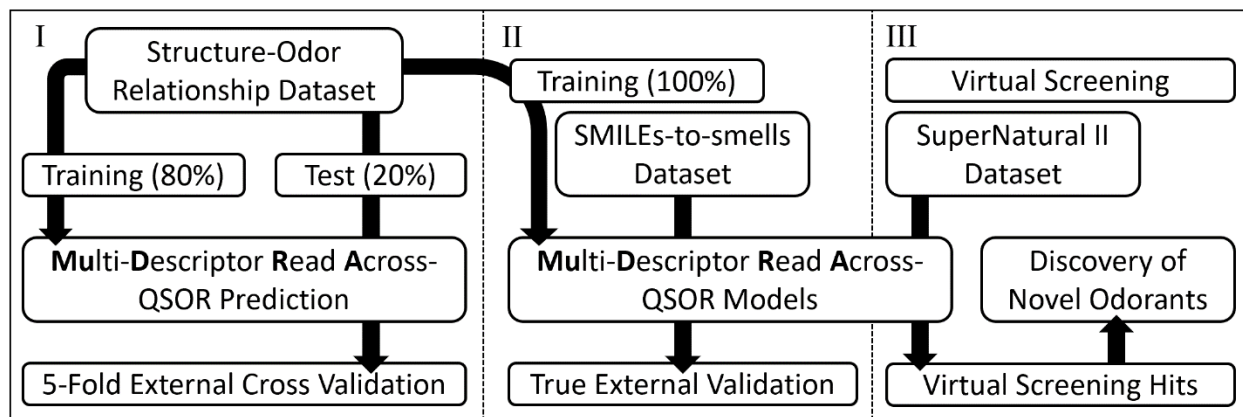


Figure 2.1. General workflow depicting the three major steps of the study design starting from the Structure-Odor Relationship Dataset (SORD): (I) 5-fold cross-validation of MuDRA model trained with SORD. (II) “True” external validation of MuDRA model trained using all of SORD to make predictions for compounds in the ‘SMILES-to-smell’ dataset. (III) Virtual screening of the SuperNatural II dataset to discover novel odorants with targeted scent properties.

2.2.0: Datasets

2.2.0.0: Structure-Odor Relationship Dataset (SORD)

As described in Chapter 1, we recently collected, curated, and integrated a large dataset of mono-molecular odorants and raw VSD profiles that we named the “Structure-Odor Relationship Dataset” (SORD) (Thieme et al. 2022). Raw VSD profiles are presented as lists of all unique VSD terms used as labels to indicate the scent quality of odorants in the sources accessed to build SORD (Thieme et al. 2022). The original SORD contained chemical structures of 2,819 unique mono-molecular odorants and a count of 422 unique VSD terms occurring in at least 1 or more raw VSD profiles. As mentioned above, a study (Rugard et al. 2021) recently observed that the majority of odorants encountered in online databases are described using 3-5 distinct VSD terms selected from the Primary International Fragrance Association (IFRA) scent ontology (International Fragrance Association 2020). This ontology consists of 27 operationally

defined VSD terms deemed to be well suited for odorant classification by a worldwide coalition of trained perfumers and scent researchers. The set of VSD terms included in the Primary IFA scent ontology (vsd_{IFA}^t) are composed by 'acidic', 'aldehydic', 'amber', 'animal-like', 'anisic', 'aromatic', 'balsamic', 'camphoraceous', 'citrus', 'earthy', 'floral', 'food-like', 'fruity', 'gourmand', 'green', 'herbal', 'honey', 'marine', 'minty', 'musk-like', 'ozonic', 'powdery', 'smoky', 'spicy', 'sulfurous', 'tobacco-like', and 'woody'.

To describe each odorant in the SORD, 3 primary VSD terms were selected to enable the efficiency and practicality of QSOR model development. Using a natural language processing approach named Embeddings from Language Models (ELMo) (AllenNLP 2022; Joshi et al. 2018), we harmonized raw VSD profiles in the SORD by embedding VSD terms as vectors in a high-dimensional semantic space using semantic cosine distance calculation (Thieme et al. 2022). As a result of this data transformation, the VSD profiles for molecules in the SORD dataset were represented as a 2,819 x 27 matrix, where each row corresponds to a unique odorant and respective values in each column correspond to 1 of 27 cosine distance values (cdv). This set of values forms the '*quantitative*' scent profile.

This representation allows for the systematic and uniform comparison between non-standard and standard terms featured in odorant profiles. The top 3 shortest semantic distance values between the verbal description of each odorant in the original dataset and 27 standard VSD term profiles were then used to systematically identify 3 VSD terms to be used as categorical labels for each odorant. Thus, each odorant was described by a series of 27 integers corresponding to each distinct targeted VSD term, referred to herein as '*binary activity*' values. Together, these independent '*binary activity*' values comprise the binary VSD profiles (vsd_{bin}), where a value of 1 indicates the use of a VSD term for categorical labeling of an

odorant, and a value of -1 implies the inverse. This transformed version of the SORD dataset was employed as a training set to develop QSOR MuDRA models. Both the dataset and protocols for its creation are fully described and available in the Supplemental Materials of a previous study (Thieme et al. 2022).

2.2.0.1: SMILES-to-smells Dataset

To create an additional *test* set for the QSOR model, we collected another odorant-scent dataset from the supplementary information section of (Sharma et al. 2021). This dataset is a compilation of scent-perception data from multiple sources. The “*SMILES-to-smells*” dataset contains records for 4,040 compounds with a curated list of VSD terms and referenced sources. Chemical names were used to retrieve the chemical structures using the PubChem API (PubChemPy 2022). We retrieved valid chemical structures for 3,831 out of 4,040 names. Structures were standardized using the ChemAxon Standardizer (ChemAxon 2021). We identified 709 replicate records for odorants with identical structures within the “*SMILES-to-smells*” dataset, corresponding to 277 unique odorant structures. The remaining 3,122 records each corresponded to singular odorant structures. In addition, there were 2,269 records for compounds also present in the training set, which were removed. The remaining 1,560 records were then used as an additional *test* set.

The curated VSD profiles in the “*SMILES-to-smells*” dataset were then used to generate “*binary activity*” VSD profiles. Here, ‘*binary activity*’ values for odorants in the ‘*SMILES-to-smells*’ dataset are determined based on the presence or absence of each of the 27 Primary IFRA terms within curated VSD profiles (see the previous section). This activity determination was readily achieved for most targeted VSD terms, which were used consistently between the Primary IFRA scent ontology and according to a curation protocol reported elsewhere (Sharma

et al. 2021). However, certain terms were not used consistently, such as *'herbal'* versus *'herbaceous'*, and *'ozonic'* versus *'ozone'*, and were recognized as inconsistent and modified thereafter to perform this operation for all 27 terms included in the Primary IFRA scent ontology. Three hundred eighty one records that did not have any assigned VSD terms out of the 27 terms were removed from the dataset, to ultimately yield a set of 1,179 odorants, each tagged with *"binary activity"* VSD profiles for use as an additional *test* set.

2.2.0.2: SuperNatural II Dataset

The *'SuperNatural II'* dataset (Banerjee et al. 2015) was kindly provided upon request and used for virtual screening. The dataset had 325,273 natural product chemical structures, which were then filtered to keep only compounds with calculated molecular weight and LogP values between the ranges of 125 to 325 amu and -0.5 to 6.0, respectively. These limits were selected based on the observation that molecular weight and LogP values for most odorants in the SORD training set are within the ranges stated above. This filtering yielded a subset of the *'SuperNatural II'* database (80,396 molecules), which was further reduced to 79,780 molecules after removing 616 molecules also present in the SORD. After curation, *SuperNatural II* contained 79,780 unique natural products that (a) do not occur (or have shared stereoisomers) in the SORD, which is used as the training set in this study, and (b) have chemical structures that have odorant-like calculated physicochemical properties.

2.2.1: Scent Multi-Descriptor Read Across Modeling

A variant of the Multi-Descriptor Read Across (MuDRA) methodology (Alves et al. 2018a) was devised to develop QSOR models that predict verbal scent descriptor profiles from chemical descriptors. QSOR models described herein were constructed and executed using the KNIME Analytics Platform (KNIME 2020). For both *training* and *test* set molecules, 4

distinct sets of chemical descriptors were calculated, including (1) MACCS fingerprints (Anderson 1984); (2) Avalon fingerprints (RDKit 2022); (3) Morgan fingerprints (Morgan 1965); (4) RDKit Descriptors (RDKit 2020).

The MuDRA approach infers the target property of the queried compound from those of their structural analogs identified within each of the multiple chemical spaces defined by the respective descriptor sets. Likewise, the Scent MuDRA algorithm infers a standard scent profile of the queried compound from those of nearest neighbor analogs identified in defined chemical spaces, where scent properties are represented as a series of quantified VSD terms. This implementation of MuDRA was devised as follows:

(i) Pairwise Tanimoto similarity S_i values between the query odorant of interest and its i^{th} neighbor (B) is calculated from the Jaccard distance d_{Jac} (Willett et al. 1998), where there are D descriptor spaces with the p_1, p_2, \dots, p_D descriptors $x_1^j, \dots, x_{p_j}^j$ and $j=1, \dots, D$. For each i^{th} compound of a dataset, the similarity $S_{i,B}^j$ with compound B in space j is calculated (See (Eq. 2.1)).

$$S_{i,B}^j = 1 - d_{Jac} = \frac{\sum_{j=1}^{p_j} x_i^j x_{i,B}^j}{\sum_{j=1}^{p_j} (x_i^j)^2 + \sum_{j=1}^{p_j} (x_{i,B}^j)^2 - \sum_{j=1}^{p_j} x_{j=1}^{p_j} x_{i,B}^j} \quad (\text{Eq. 2.1})$$

(ii) Calculated similarity scores, $S_{i,B}^j$, are normalized from 0 to 1, where a value of 1 represents identical pairs.

(iii) The VSD prediction ($vsd_i^{\text{pred}, \text{MuDRA}}$) for a given IFA term (this is, the binary scent activity predicted by the Scent MuDRA), is defined by dividing the sum of the mean calculated similarity scores, $S_{i,B}^j$, by the product of the respective cosine distance value of its i^{th} neighbor in

with compound B in space j (cdv_B^j , see the previous section) by its VSD ‘binary activity’ value ($vsd_bin_B^j$, see the previous section) (see (Eq. 2.2)).

$$vsd_i^{pred, MuDRA} = \frac{\sum_{j=1}^D \sum_{B_j=1}^{n_j} S_{i,B_j}^j}{\sum_{j=1}^D \sum_{B_j=1}^{n_j} cdv_{i,B_j}^j vsd_bin_{i,B_j}^j} \quad (\text{Eq. 2.2})$$

(iv) the previous steps are repeated for each of the 27 IFRA terms to fill the entire predicted VSD profile of an odorant.

2.2.2: External Cross-Validation

A 5-fold external cross-validation procedure was used to assess the predictive accuracy of QSOR Scent MuDRA models trained using the SORD. The SORD was randomly split into 5 subsets, and each subset was then iteratively used as a *test* set while the remaining 4 subsets were combined to be used as *training* sets. The predicted VSD profile of each odorant in the SORD (when this compound is in the test fold) was then used to rank odorants according to the predicted likelihood their scent qualities can be described by each IFRA term corresponding to these values. Additionally, 10 rounds of y -randomization were performed to guarantee the predictivity of the models were not due by chance.

2.2.3: Additional External Validation

An additional round of external validation was performed to further assess the predictive accuracy of QSOR models trained using SORD. This time, in contrast to the 5-fold external cross-validation described above, SORD was used in its entirety as the *training* set; and the *test* set was the ‘SMILES-to-smells’ dataset. Ten rounds of Y -randomization per predicted VSD term (27) were performed, as in the above section.

2.2.4: Statistical metrics

The accuracy of models was estimated by the area-under-the-curve (AUC) values calculated for receiver operator characteristic (ROC) for each IFA term (AUC^t , See (Eq. 2.3)); capturing the relationships between the true positive rates (TPR) and false positive rates (FPR) as a function of odorant ranking by the predicted VSD values, where true activity x is defined by the odorant ‘binary activity’ VSD profiles described above, and t corresponds to any 1 VSD term out of the 27 VSD terms in vsd_{IFA}^t .

$$AUC^t = \int_{x=0}^1 TPR(FPR(x))dx \quad (\text{Eq. 2.3})$$

2.2.5 Virtual Screening

The MuDRA models described above were employed to predict scent qualities for compounds in the ‘SuperNatural II’ dataset. Predicted scores were used to rank the compounds by their predicted likelihood of being described by each of the 27 IFRA terms. For each structure in the ‘SuperNatural II’ dataset, 27 ‘prioritization scores’ ($vsd_i^{pred, MuDRA}$) were output as a series of numerical values; where larger values correspond to the higher likelihood (estimated via semantic and chemical distance calculations) that the screened compounds possess scent qualities that can be described by the respective 27 targeted VSD terms.

Lists of the top 1,000 ranked virtual screening hits were submitted as queries to the PubChem API (PubChemPy 2022), in order to identify the compounds in SuperNatural II that have distinct chemical structures in PubChem. We reasoned that structures listed in PubChem are likely to be listed in other online databases, and vice versa. These hit compounds were prioritized in order of ascending rank for each targeted VSD term, in search of screened molecules with

predictions supported by data external to the training set used in this study. The rank number of the first identified odorant with external support for prediction validity was recorded.

2.3: Results

2.3.0: Training Set Curation from SORD

As detailed above, SORD contains both raw and standardized VSD profiles for 2,819 unique mono-molecular odorants. Although raw VSD profiles were not used in the procedures described in this manuscript, they were used to generate **Figure 2.2** as a way to visualize the relative prominence of each term, especially those predicted by our models, in SORD.

The calculation of ‘*binary activity*’ values from standardized profiles enabled the assignment of each odorant included in SORD to 3 of 27 possible scent classes defined by vsd_{IFA}^t . This curation step was a pre-requisite both for modeling and validation of model predictions. To evaluate the overlap between activity class labels used to annotate odorants in SORD, odorants were split into 351 groups, one group per unique pairs of the 27 possible activity classes (for example, ‘acidic’, ‘aldehydic’; ‘acidic’, ‘amber’, and so on...). For each activity class pair, the number of unique odorants in which fall into their respective group were summed. **Figure 2.3** is a heatmap that demonstrates the degree of overlap between activity classes in the SORD, via count of unique odorants labeled with different binary combinations of the 27 terms in vsd_{IFA}^t .

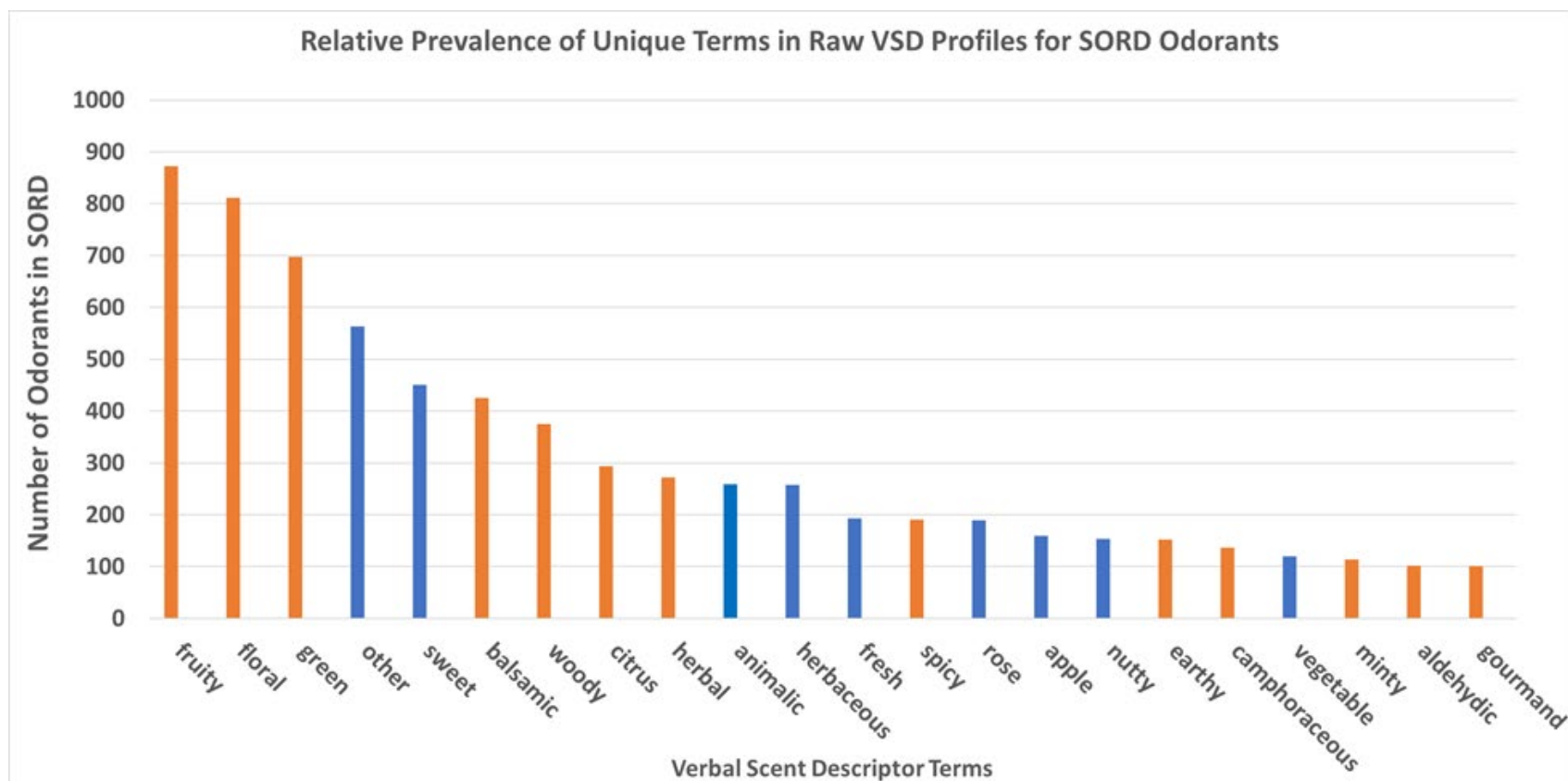


Figure 2.2. Bar chart with counts of frequency of each unique term in the raw VSD profiles of SORD odorants. Bars corresponding to terms included in $\text{vsd}_{\text{IFA}}^t$ are colored in orange, those not included are colored in blue. VSD terms with less than 100 incidences in SORD are not shown but are included in the data table (see Supplementary Information Table S2) used to generate the above chart.

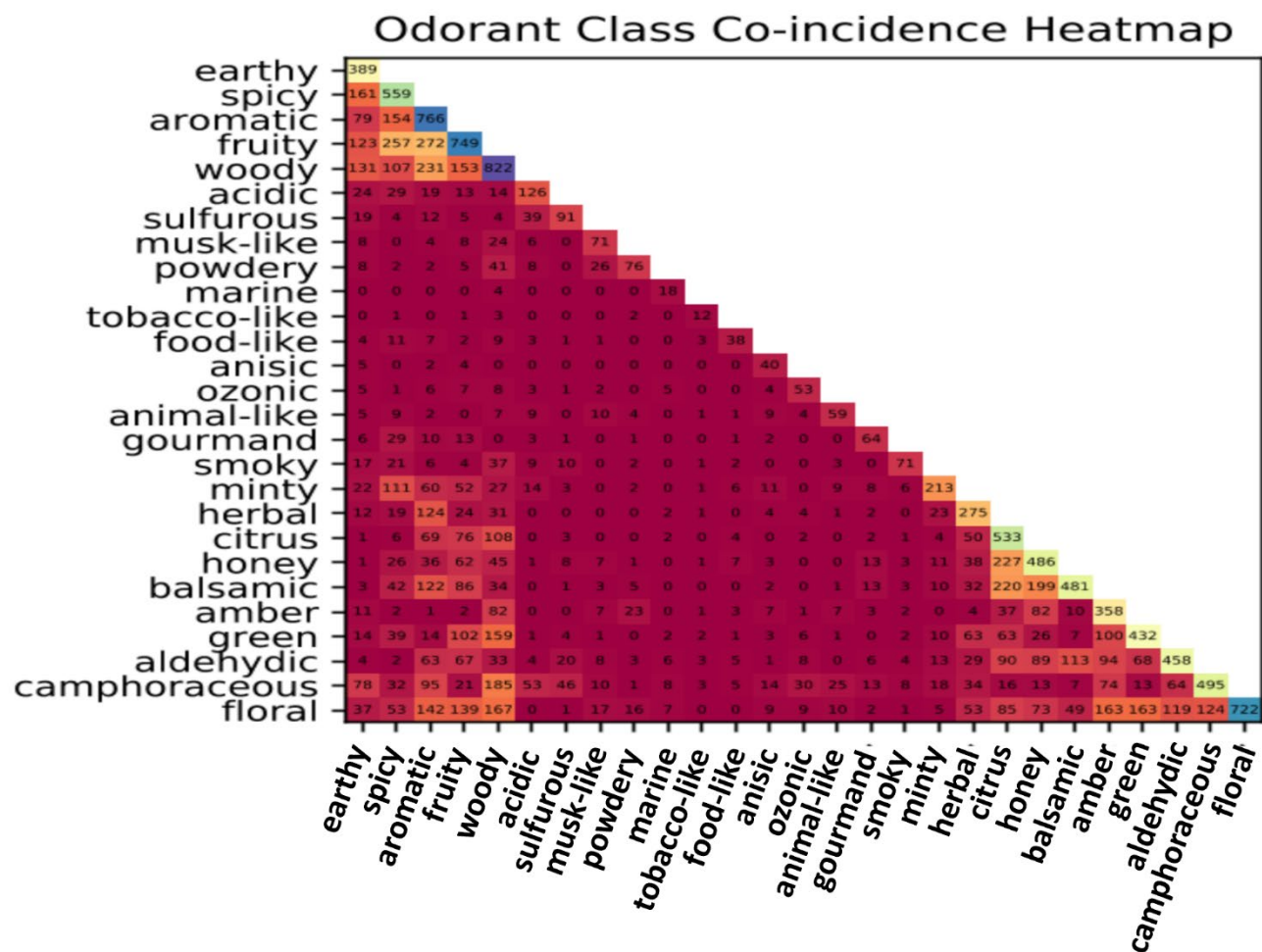


Figure 2.3. Heatmap showing the overlap between activity classes used to label mono-molecular odorants. Red cells have low counts, and blue cells have high counts of odorants belonging to each activity class pair. The diagonal cells represent the number of total number of odorants in each singular activity class, with values identical to the number of odorants that falls under each activity class in SORD; and the cells at intersections of singular activity classes represent the total number of odorants in the SORD that simultaneously belong to two independent activity classes. This diagram helps one to understand the relative frequency at which these labels overlap on mono-molecular odorants in the SORD.

2.3.1: External Cross-Validation of QSOR MuDRA Models

The predictive accuracy measures obtained from 5-fold external cross-validation are summarized in **Table 2.1**. Calculated ROC AUC values range between 0.55 (*'anistic'*, *'earthy'*) to 0.87 (*'sulfurous'*), and 12 out of 27 targeted VSD terms (*'acidic'*, *'animal-like'*, *'camphoraceous'*, *'floral'*, *'fruity'*, *'green'*, *'herbal'*, *'musk-like'*, *'ozonic'*, *'powdery'*, *'sulfurous'*, *'tobacco-like'*) were predicted within the measure of accuracy we considered acceptable (ROC AUC \geq 0.65, ROC AUC- ROC AUC for y-randomized training data \geq 0.1). The relatively high predictive accuracy for the VSD term *'sulfurous'* is likely derivative of the fact that the presence or absence of sulfur atoms in a given compound factors heavily into whether or not their scent is described as sulfurous. Conversely, the poorly predicted VSD terms *'anistic'* and *'earthy'* are not as readily explained by the presence or absence of singular chemical moieties.

Percentages of odorants in the SORD with *'binary activity'* values of 1 are reported along with ROC AUC values for each of the 27 Primary IFRA VSD terms reported in **Table 2.1**. ROC AUC values greater than or equal to 0.65 were established as the threshold to indicate acceptable predictive accuracy, although ideally we were seeking values of 0.7 and higher (Zach 2021). Additionally, calculation of the differences between ROC AUC values from 5-fold external cross-validation using SORD and a y-randomized version of SORD can be used to guarantee that the predictive performance of QSOR MuDRA models trained were not due to chance alone. ROC AUC values (A) that are greater than y-randomized ROC AUC values (B) by 0.1 or more, were considered by our group to predict significantly better than y-randomized counterparts (See **Table 2.1**).

Table 2.1. *Summary of results from 5-fold external cross validation.*

Primary IFA VSD Term	% of Odorants in SORD	ROC AUC (A)	Y-Randomized ROC AUC (B)	(A)– (B)
Acidic	4.5%	0.66	0.50	0.16
Aldehydic	16%	0.60	0.50	0.10
Amber	13%	0.64	0.50	0.14
Animal-like	2.1%	0.68	0.50	0.18
Anisic	1.4%	0.55	0.50	0.05
Aromatic	27%	0.56	0.50	0.06
Balsamic	17%	0.57	0.50	0.07
Camphoraceous	18%	0.65	0.50	0.15
Citrus	19%	0.64	0.51	0.13
Earthy	14%	0.55	0.50	0.05
Floral	26%	0.71	0.50	0.21
Food-like	1.3%	0.56	0.49	0.07
Fruity	27%	0.71	0.50	0.21
Gourmand	2.3%	0.58	0.49	0.09
Green	15%	0.71	0.50	0.21
Herbal	9.8%	0.65	0.50	0.15
Honey	17%	0.61	0.50	0.11
Marine	0.6%	0.66	0.48	0.18
Minty	7.6%	0.64	0.50	0.14
Musk-like	2.5%	0.79	0.49	0.30
Ozonic	1.9%	0.67	0.49	0.18
Powdery	2.7%	0.69	0.50	0.19
Smoky	2.5%	0.62	0.49	0.13
Spicy	20%	0.60	0.49	0.11
Sulfurous	3.2%	0.87	0.50	0.37
Tobacco-like	0.4%	0.68	0.49	0.19
Woody	29%	0.64	0.49	0.15

Some terms, for instance, ‘*anistic*’ ‘*aromatic*’ and ‘*marine*’, were not predicted better or predicted insignificantly better than with y-randomized training data, by the QSOR MuDRA model. At the onset of this experiment, expectations were modest for the predictive performance of multiple VSD terms at once, using a singular QSOR model. We acknowledge the high level of complexity of the task; multiple variables of this method can be tuned, such as (i) sensitivity to the language models, (ii) choice of target terms, and (iii) strategies for converting raw qualitative scent descriptions to numeric descriptions. Overall, the total performance of the model across the whole set of predicted terms did exceed these modest expectations.

2.3.2: Independent External Test Set Curation

The curation of the ‘*SMILES-to-smell*’ dataset detailed in the previous section yielded an independent $test_{ind}$ set for external validation of the QSOR MuDRA model detailed above. Key steps of this process were (1) removal of $test_{ind}$ set odorants (and structural isomers thereof) also present in the $training_{SOR}$ set, as a prediction of these structures would artificially enhance accuracy metrics; and (2) assignment of binary activity class labels for each of the 27 terms t in vsd_{IFA}^t to $test_{ind}$ set odorants, where a value of 1 is assigned to label an odorant as active in each of the classes defined by vsd_{IFA}^t identified in their raw VSD profiles (See Materials and Methods).

2.3.3: Independent External Validation of QSOR MuDRA Models

The predictive accuracy measures obtained from external cross-validation using the ‘SMILES-to-smell’ dataset are summarized in **Table 2.2**. Calculated ROC AUC values range between 0.48 (*acidic*) to 0.79 (*amber*), and 10 out of 27 targeted VSD terms (*amber*, *animal-like*, *balsamic*, *camphoraceous*, *food-like*, *fruity*, *minty*, *ozonic*, *sulfurous*, *woody*) were recognized as predicted within an acceptable measure of accuracy (ROC AUC \geq 0.65, ROC AUC - ROC AUC for y-randomized training data \geq 0.1). With metrics just below acceptable, *smoky* was reported at ROC AUC = 0.64; *citrus*, *floral*, and *herbal* were reported at ROC AUC = 0.63. Percentages of odorants in the ‘SMILES-to-smell’ dataset with ‘binary activity’ values of 1 are reported along with ROC AUC values for each of the 27 Primary IFRA VSD terms reported in **Table 2.2**. Y-randomization was also performed during independent test set validation.

The low predictive performance of *acidic* compared to the high predictive performance of *sulfurous* may reflect the broad chemical meaning of acidity compared to the narrow meaning implied by sulfurous. The performance of *aromatic* was the lowest out of all 27 targeted VSD terms in terms of absolute ROC AUC value, and the ROC AUC value generated was 0.1 lower than the y-randomized ROC AUC for *aromatic*. For some terms, predictions made by our model were poor; however, the model performed better than random guessing simulated by Y-randomization for the majority of the terms. The exceptions were *acidic*, *aromatic*; and *gourmand* although *aldehydic*, *anisic*, *earthy*, *green*, *honey*, *marine*, *musk-like*, and *powdery*, and *smoky* have differences that are greater than 0, the difference between the two values was less than or equal to 0.1.

Table 2.2. Summary of results from external test set validation using the ‘SMILES-to-smell’ dataset.

Primary IFA VSD Term	% of Odorants in SMILES-to-Smells	ROC AUC (A)	Y-Randomized ROC AUC (B)	(A) -- (B)
Acidic	0.4%	0.48	0.52	-0.04
Aldehydic	3.0%	0.61	0.51	0.10
Amber	3.1%	0.79	0.47	0.32
Animal-like	3.4%	0.65	0.50	0.15
Anisic	2.9%	0.55	0.51	0.04
Aromatic	0.6%	0.44	0.54	-0.10
Balsamic	8.3%	0.68	0.50	0.18
Camphoraceous	3.3%	0.66	0.51	0.15
Citrus	10%	0.63	0.50	0.13
Earthy	8.3%	0.57	0.51	0.06
Floral	32%	0.63	0.50	0.13
Food-like	4.1%	0.67	0.52	0.15
Fruity	38%	0.70	0.50	0.20
Gourmand	1.9%	0.49	0.50	-0.01
Green	29%	0.53	0.50	0.03
Herbal	0.5%	0.63	0.50	0.13
Honey	6.2%	0.62	0.54	0.08
Marine	1.8%	0.50	0.42	0.08
Minty	1.4%	0.70	0.49	0.21
Musk-like	6.8%	0.62	0.52	0.10
Ozonic	1.7%	0.75	0.45	0.30
Powdery	10%	0.56	0.51	0.05
Smoky	9.1%	0.64	0.47	0.17
Spicy	3.5%	0.59	0.51	0.08
Sulfurous	21%	0.78	0.50	0.28
Tobacco-like	34%	0.61	0.50	0.11
Woody	18%	0.65	0.50	0.15

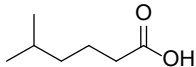
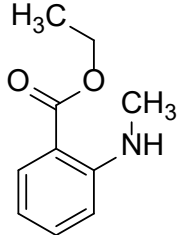
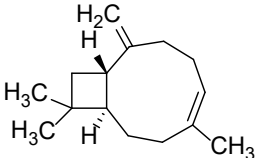
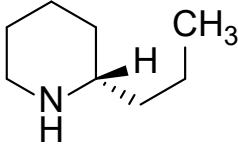
2.3.4: Virtual Screening of SuperNatural II

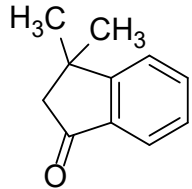
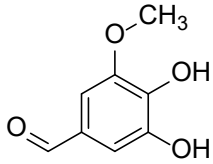
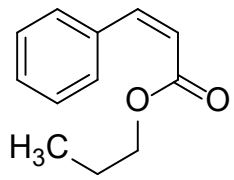
Virtual screening of *SuperNatural II* was performed in both an effort to (a) assess the predictive capability of the QSOR MuDRA model built for this study using a larger test set, and (b) identify a prioritized set of natural products that may serve as novel odorants with targeted properties. Curation (elaborated in the Materials and Methods section above) was performed to remove compounds also present in the *training* set (SORD), as well as to remove compounds

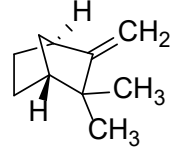
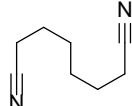
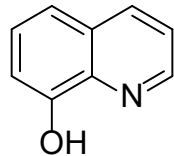
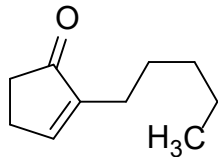
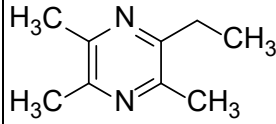
outside of the ranges of calculated molecular weight and LogP values derived for mono-molecular odorant structures in SORD. For each scent term t in vsd_{IFA}^t , the corresponding scores in predicted profiles of $vsd_i^{pred, MuDRA}$ values for odorant-like compounds in *SuperNatural II* (cf. Eq. 1-3) were ranked to prioritize virtual screening molecules for inspection. Then, 27 sets of the top 1,000 ranked compounds were identified, one for each t in vsd_{IFA}^t . The compound names from these sets were used as literature mining queries via the PubChem API, which enabled the separation of those compounds that do not appear in the PubChem database.

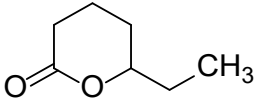
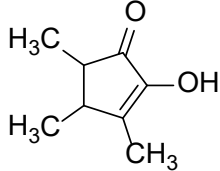
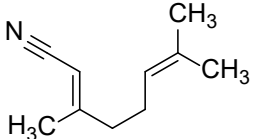
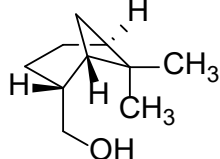
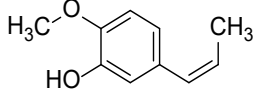
One compound was identified as a hit example for each activity class, meaning we could identify a confirmed hit in the top 1000 ranked compounds out of ~70,000 screened compounds. In order to identify hit compounds, compound names were input as internet search engine queries along with the keywords “scent” and “odor”. The rank number of the first compound in each ranked set that was identified as a true positive according to an external source was recorded. Compounds were used in queries in order of their rank for each predicted activity class, and this manual search approach was taken for ranked sets of 1,000 compounds for each of the 27 classes. While this process was time consuming, we were able to identify a hit compound for each activity class. Hit compounds and their predicted ranks from virtual screening are summarized in **Table 2.3**. Definitions are provided along with Primary IFA VSD terms to help guide the reader's interpretation of the prediction meaning. In order to provide readers with examples of virtual screening hits from *SuperNatural II*, the name and chemical structure of hit odorants identified via a manual review of the top 1,000 odorants ranked by $vsd_i^{pred, MuDRA}$ values are reported (See Materials and Methods Section), along with the rank number for each hit odorant identified. Additionally, the rationale for each selected hit is provided along with external references providing support to the accuracy of targeted scent property prediction.

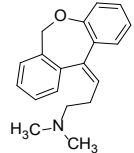
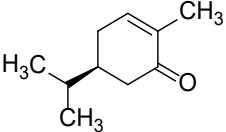
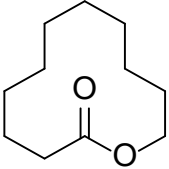
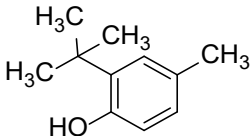
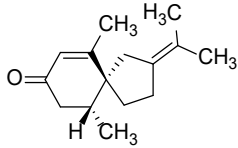
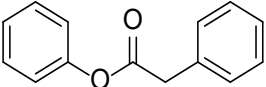
Table 2.3. Summary of virtual screening results, 1 'hit' compound selected per VSD term in the Primary IFRA scent ontology.

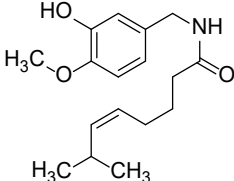
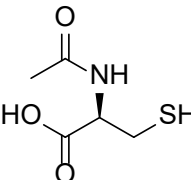
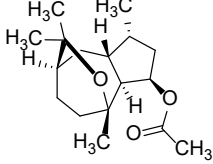
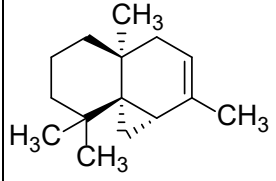
Primary IFRA VSD Term	Primary IFRA VSD Term Definition	SNII Hit Rank	SNII Hit Name	SNII Hit Structure	SNII Hit Rationale	External Support for Hit Rationale
Acidic	“Acidic means a fragrance note that smells sharp and somewhat pungent. Acidic notes may help boost a citrus note or impart natural qualities.”	49	5-methyl-hexanoic acid		Sour, fatty-cheese, oily in fruit dilution	(GSC 2022)
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”	22	ethyl methyl anthranilate		soft sweet mandarin petitgrain	(GSC 2022)
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.”	248	beta-caryophyllene		sweet woody spicy clove dry	(GSC 2022)
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.”	199	coniine		Coniine is eliminated from the body through the lungs and kidneys and the peculiar mousy odor of the urine and exhaled air is diagnostic.	(Hotti and Rischer 2017)

Primary IFRA VSD Term	Primary IFRA VSD Term Definition	SNII Hit Rank	SNII Hit Name	SNII Hit Structure	SNII Hit Rationale	External Support for Hit Rationale
Anisic	“Anisic materials are those that smell similar to natural aniseed materials like tarragon or fennel.”	34	3,3-dimethyl-indanone		“...found that 3,3-dimethyl-1-indanone exhibits apart from the spicy, safranal and leathery odor and taste, myrrh, ionone and fruity aspects which is a highly desirable odor and taste combination. Furthermore, it was found that 3,3-dimethyl-1-indanone can ideally substitute safranal in perfumery and flavor applications, and therefore offers an outstanding alternative for the use-limited safranal.”	(Bajgrowicz and Gygax 2000)
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.”	41	5-hydroxy-vanillin		“In case you were wondering, the product smells much like vanillin, but perhaps a bit more like caramel, and not as strong.”	(Myristicinaldehyde 2018)
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.”	703	n-propyl cinnamate		balsamic musty vine amber cortex	(GSC 2022)

Primary IFRA VSD Term	Primary IFRA VSD Term Definition	SNII Hit Rank	SNII Hit Name	SNII Hit Structure	SNII Hit Rationale	External Support for Hit Rationale
Camphoraceous	“A fresh, strong and diffusive smell that is characterized by natural camphor and other herbs such as rosemary or marjoram.”	48	(+)-camphene		fresh herbal woody fir needle camphoraceous	(GSC 2022)
Citrus	“Citrus notes are given by the smell of fruit from the citrus family – such as orange, lemon or grapefruit.”	165	1,8-octadinitrile		fresh, sweet, waxy, floral, citrus, mandarin	(GSC 2022)
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.”	126	8-hydroxyquinoline		"characteristic" [GSC]. Phenolic odor.	(PubChem 2022)
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.”	72	amyl cyclopentenone		woody floral jasmin tuberose	(GSC 2022)
Food-like	“Food-like describes food substances of a savoury or less specific character - such as the smell of roasted vegetables”	873	2-ethyl-3,5,6-trimethylpyrazine		“Sensory investigation indicated that pyrazines have a synergistic effect on the perception of roasted aroma”	(Yan et al. 2021)

Primary IFRA VSD Term	Primary IFRA VSD Term Definition	SNII Hit Rank	SNII Hit Name	SNII Hit Structure	SNII Hit Rationale	External Support for Hit Rationale
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.”	13	dela-heptalactone		coconut oily green earthy	(GSC 2022)
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.”	2	2-hydroxy-3,4,5-trimethylcyclopent-2-en-1-one		sweet burnt spicy caramellic maple coffee bready licorice	(GSC 2022)
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.”	253	geranyl nitrile		citrus, green, oily	(GSC 2022)
Herbal	“Herbal notes include culinary herbs (e.g., thyme, rosemary) that often have a green note and impart fresh nuances to a complex fragrance.”	9	myrtenol		naturally occurs at 1% in sage oil (saliva officinalis)	(GSC 2022)
Honey	“Honey is used to describe materials that have honey characteristics – often sweet and cloying, but sometimes quite harsh and acidic.”	40	meta-eugenol		spicy carnation	(GSC 2022)

Primary IFRA VSD Term	Primary IFRA VSD Term Definition	SNII Hit Rank	SNII Hit Name	SNII Hit Structure	SNII Hit Rationale	External Support for Hit Rationale
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.”	244	doxepin		“slight, amine-like odor”	(Cunha 2021)
Minty	“These materials impart mint or menthol notes reminiscent of peppermint and spearmint.”	38	carvotanacetone		minty	(GSC 2022)
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.”	116	oxacyclododecan-2-one		woody, amber, dry, musky	(GSC 2022)
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.”	174	2-tert-butyl-para-cresol		cresol medicinal leather	(GSC 2022)
Powdery	“Powdery fragrance ingredients are from a larger complex group that impart a warm, sometimes sweet or musky powdery smell.”	113	beta-vetivone		quinoline-like, fruity (cassis, grapefruit) aroma with a woody by-note	(Leffingwell 2002)
Smoky	“These ingredients have a smoked or phenolic quality, reminding you of the smell from a bonfire or the smell of food burning.”	2	phenyl benzoate		phenolic coal tar	(GSC 2022)

Primary IFRA VSD Term	Primary IFRA VSD Term Definition	SNII Hit Rank	SNII Hit Name	SNII Hit Structure	SNII Hit Rationale	External Support for Hit Rationale
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.”	11	capsaicin		mild warm herbal	(GSC 2022)
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.”	12	N-acetyl-L-cysteine		“like rotten eggs”	(AHC 2006)
Tobacco-like	“These ingredients are created to give a smell of tobacco before it has been lit or smoked. They tend to be sweet and warm notes, sometimes with the smell of dried fruit.”	359	alpha-kessyl acetate		occurs naturally in valeriana officianalis rhizome oil	(GSC 2022)
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.”	203	thujopsene		cedarwood oil constituent	(Eybna Technologies 2022)

2.4: Discussion

There are two natural concerns that arise in QSAR modeling for odorant discovery. The first concern is elementary in principle: the molecules selected for virtual screening should likely function as odorants. The only accurate way to confirm whether or not a molecule functions as an odorant is through the firsthand experiences of individuals participating in scent perception-based surveys. In the absence of this information, physicochemical descriptors of chemical structures, such as LogP and molecular weight, can be used to select compounds that are likely to be volatile and lipophilic enough to enter the nose via inhalation to reach the olfactory epithelium, where they bind to and activate the olfactory receptors (Ache and Young 2005).

Once a set of accessible odorant-like molecules are compiled in a virtual screening library, the second concern emerges: the predictions of the model should be trustworthy. This concern is of relevance to all cases of QSAR modeling. In theory, for a given test set molecule \mathbf{M}_{test} , proximity to training set molecules $\mathbf{M}_{\text{training}}$ in a high-dimensional chemical space $\mathbf{C}_{\text{space}}$, tends to correspond to the similarity in biochemical activity. By extension of this theory, the proportion of nearest neighbors $\mathbf{M}_{\text{training}}$ that are ‘*active*’ out of a set of nearest neighbor compounds to \mathbf{M}_{test} , can serve as a means of activity prediction for \mathbf{M}_{test} .

In reality, nearest neighbors can be quite far apart, and therefore it is unwise to blindly apply the theory outlined above. In other words, any given \mathbf{M}_{test} always has at least 1 nearest neighbor $\mathbf{M}_{\text{training}}$ compound, but the actual distance between them may be so far that there is a significant difference in terms of their biochemical activity, because ‘nearest’ is a relative term. The phrase ‘*activity cliff*’ is used to describe a near neighbor pair within a dataset where there is a disagreement between the activities of the two compounds in the pair. For training sets to be of

acceptable quality (modelability) to build predictive QSOR models, the number of activity cliffs present in them should be minimized (Golbraikh et al. 2014).

2.4.0: Scent-MuDRA versus MuDRA

Multi-Descriptor Read Across was selected as the method to make predictive QSOR models in this study. This selection was made primarily on the widely recognized observation that many structurally similar odorants have dissimilar VSD profiles (Rossiter 1996), and vice versa. Because MuDRA relies on chemical similarity measured in more than one chemical descriptor space, the likelihood that predictions will be corrupted by a single nearest neighbor pair is reduced.

Some modifications were made to the general MuDRA methodology previously described by our group (Alves et al. 2018a). In a previous study (Keller et al. 2017), authors reported the development of a series of individual models, each trained to predict 1 of 21 activity classes, out of a scent profile comprised of 21 VSD terms. While this approach is viable, we wanted to assess the feasibility of a technique that could make predictions of the entire scent profile with a single model. This is the first key difference between general and Scent MuDRA: the original MuDRA has been used to predict singular activities, whereas Scent MuDRA predicts multi-objective VSD profiles. In this study, we predict 27 VSD terms, but the number of terms is variable, and is ultimately defined by the number of terms in the target scent ontology.

The second key difference is in the incorporation of semantic information into Scent MuDRA predictions. While general MuDRA relies on chemical similarity to make predictions by averaging the target bioactivity of nearest chemical neighbors defined in diverse chemical spaces, Scent MuDRA averages semantic profiles of chemical neighbors of the test compounds. Through integration of these two distinct high-dimensional spaces, one used to capture the

variability of odorant chemical structure, and the other to capture the relative similarity between semantic entities such as VSD terms used to describe odorants. Scent MuDRA serves to quantitate subjective experience and enables prediction of quantitative scent profiles for novel odorant molecules from their chemical descriptors.

2.4.1: How to Use Scent MuDRA

Scent MuDRA is intended to be time efficient to employ, while at the same time still be customizable according to the needs of individual researchers. Prospectively, this technique can be implemented using any odorant training set that has VSD profiles standardized via relative semantic distance to a standard set of VSD terms defining scent profile activity classes. Formally, there are no limits to the sets of chemicals and VSD terms that can be used for this purpose. Practically, such limits do exist.

One limit is applicability domain, or the region of chemical space covered by the training set used to build models. If test set molecules fall too far outside the applicability domain of a predictive model, predictive capability of even the most modellable training set will still fall short. By using a singular training set that contains 27 activity classes for each odorant, instead of curating 27 training sets to build 27 individual models; the applicability domain of Scent MuDRA models can be extended to the limit defined by the number of unique odorant structures that can be integrated into one training set. The number 27 in this case reflects the number of VSD terms in the Primary IFRA scent ontology, used as the target ontology in this study.

Of course, this number is arbitrary. As stated above, Scent MuDRA is intended to be customizable. This means that the number, as well as specificity, of predicted VSD terms is virtually limitless. Furthermore, $vsd_i^{pred, MuDRA}$ scores can be used as numeric profiles of

odorant scent quality, according to targeted scent ontology (vsd_{IFA}^t). This use case may be relevant in the case of groups studying relatively small (<100) sets of structurally related odorants in terms of the discrete differences between their complete scent profiles. Alternatively, the predictive scores generated by Scent MuDRA for each independent activity class included in multi-objective profiles can be used for ranking and prioritization of odorant-like molecules during virtual screening.

2.4.2: Validation of QSOR MuDRA Models via Virtual Screening of SuperNatural II

Virtual screening of the *SuperNatural II* database was performed as detailed in the Materials and Methods Section, with selected ‘hit’ compounds reported in **Table 2.3**. While the measures of predictive accuracy obtained via 5-fold cross-validation and test set prediction reported in the Results section above help quantify model performance on a statistical basis, this exercise serves to validate the performance of the Scent MuDRA QSOR model produced in a context that simulates real-world applications of predictive models. The rank of selected hit compounds was recorded, where a lower value indicates better prioritization by rank.

In the case of ‘*gourmand*’, which did not perform very well in terms of ROC AUC values during external validation rounds, the 2nd ranked compound (**2-hydroxy-3,4,5-trimethyl-cyclopent-2-en-1-one**) was identified as a hit. This is interesting given that the ROC AUC value for ‘*gourmand*’ was lower than the threshold of acceptability for test set validation and 5-fold external cross-validation. The rationale for selection of **2-hydroxy-3,4,5-trimethyl-cyclopent-2-en-1-one** as a hit for the VSD term ‘*gourmand*’ is provided in the columns **SNII Hit Rationale** and **External Support for Hit Rationale**. In this case, the hit compound was found in an external odorant database (good scents company), where it was described with the VSD terms ‘*sweet*’, ‘*burnt*’, ‘*spicy*’, ‘*caramellic*’, ‘*maple*’, ‘*coffee*’, ‘*bready*’, and ‘*licorice*’.

The hit selection for *'animal-like'*, **coniine**, is an alkaloid with a “peculiar mousy odor” which is derived from hemlock; the infamous toxic plant notorious for the death of Socrates (Hotti and Rischer 2017). The hit selection for *'spicy'* was **capsaicin**, the active ingredient in spicy peppers responsible for creating sensations of heat. It is unsurprising that capsaicin was not included in our training set, since it is not typically considered to be an odorant; it was still found to have a record in the good scents company database, along with the VSD terms *'mild'*, *'warm'*, and *'herbal'*. In many cases, this database served as the external source to verify that the ranked prioritization of compounds according to Scent MuDRA predictions, did in fact prioritize odorants with targeted scent properties (GSC 2022).

In a few other cases, unique references were used to identify that ranked prioritizations were reflected by external subjective data. **N-Acetyl Cysteine** was selected as the hit for *'sulfurous'*, and this selection was supported by a statement recorded from a symposium presentation on the subject of **N-Acetyl Cysteine** in the context of its use as a supplement. Barbara Insley Crouch, PharmD described the scent of the supplement as a characteristic “rotten egg” smell (AHC 2006). **5-hydroxyvanillin** was selected as the hit for *'aromatic'*, and verified by an anonymous chemist on the “Sciencemadness Discussion Board” web forum under the username *'Myristicinaldehyde'*, who claims to have synthesized the compound; “In case you were wondering, the product smells much like vanillin, but perhaps a bit more like caramel, and not as strong.” (Myristicinaldehyde 2018).

A recent review on the occurrence of marine ingredients in fragrance by (Riad et al. 2021) presents a series of canonical with *'marine'* scents, which broadly means that they are reminiscent of a marine environment. The prescription medication **doxepin**, was selected for *'marine'*, and is noted to have a “slight, amine-like odor” according to the RxList website

(Cunha 2021). Although amines have a wide diversity of VSD terms that have been associated with them, fishy scents are one of the most commonly associated with simple aliphatic amines. Interestingly, the structure of doxepin appears to overlap with the olfactophore (arrangement of functional groups in 3D space required to elicit a '*marine*' scent) defined by the standard '*marine*' odorant "Calone 1951" and analogs thereof that are also '*marine*' odorants; which is also presented in the review article referenced above (Riad et al. 2021). While the scent description of doxepin is not highly specific, this drug was selected as a hit for marine to exemplify a chemical that can be perceived as an odorant, and also be used as a pharmaceutical; as it is not uncommon that a molecule falling under one industrial classification, such as cosmetics, also fall in other categories such as pharmaceutical excipients and active ingredients (Alves et al. 2018b). The example of **coniine**, mentioned above as the selection for '*animal-like*', represents an analogous scenario as it falls under the category of toxin, as opposed to medicine or cosmetic. Finally, the selection of **3,3-dimethyl-inandone** as the hit for '*anisic*', was supported by its description in a patent that identified its scent qualities. **3,3-dimethyl-inandone** has been identified as a safer replacement for the odorant safranal, which has the "characteristic warm, spicy odor of saffron" (Bajgrowicz and Gygax 2000); and while this is not an exact match with terms such as the examples given in the Primary IFRA definition for '*anisic*', both fennel and tarragon are annotated in the good scents company website as being spicy and herbal.

2.5: Conclusions

As the body of scent perception-based data available for collection, curation, and integration into training sets for QSOR models increases, Scent MuDRA serves as an innovation in QSOR modeling methodology. By predicting multi-objective scent profiles using both chemical and semantic similarity, Scent MuDRA relies on the same fundamental principles that guide natural scent perception and description of scent percepts via natural language. The validity of this method has been assessed via 5-fold cross-validation, test set validation, and virtual screening. It is the expressed intent of our group that this technique be made available to other groups seeking to identify new odorants with targeted scent properties. The scent MuDRA workflow detailed in this chapter has been made available online at

https://figshare.com/projects/AJT_Dissertation_UNC_CH_ESOP_CBMC_2022/137364.

CHAPTER 3: SCENT-KOP: SUBJECTIVE SCENT-PERCEPTION AND BIOMEDICAL DATA KNOWLEDGE GRAPH

3.1: Introduction

Biomedical knowledge graphs have expanded our ability to find connections between previously disparate areas of biomedical research. Building upon our recent development of the ROBOKOP biomedical knowledge graph, we have developed a specialized knowledge graph focusing on odorant small molecules and scent perception, that we have named SCENT-KOP. SCENT-KOP captures scent qualities described in specialized respective data sources that include both hand curated expert assessments and natural language processing driven methodologies. SCENT-KOP is the first knowledge graph to integrate scent descriptors with existing biomedical knowledge graph. In the final chapter of this dissertation, we illustrate this tool and demonstrate its utility in supporting the discovery of medicinal odorants, with a case study.

3.1.0: How We Organize Biomedical Knowledge in Knowledge Graphs.

The field of biomedical knowledge graphs is a quickly advancing area of bioinformatics. These graphs aim to capture existing biomedical information in a form which is flexible and interdisciplinary. This is accomplished by expressing core terms and concepts as nodes, or vertices, in a graph (some examples of nodes could be **aspirin**, **pancreatic cancer**, **fever**, etc.). The relationships between these concepts are captured as edges linking nodes in a graph (an example of a relationship could be **aspirin treats fever**). A benefit of a graphical approach over a classical approach to data arrangement is these graph databases can easily interconnect data

from various sources thereby capturing functional relationships between biomedical concepts (Ji et al. 2020).

The ROBOKOP knowledge graph is a general purpose biomedical knowledge graph which has been under development for the last 5 years (Bizon et al. 2019; Morton et al. 2019). This knowledge graph aims to represent the current state of biomedical knowledge by aggregating over thirty-eight biomedical databases. While the ROBOKOP database has shown great utility for general biomedical issues (Fecho et al. 2021); it often lacks the niche knowledge which would allow for very specific disciplines and experts of these disciplines to leverage the graph. The recent COVID-19 pandemic was one such instance of a knowledge gap; in response to this gap in capability of the ROBOKOP graph, our team produced COVID-KOP (Korn et al. 2021), an extension which utilized the same underlying database, but was supplemented with specific COVID-19 information. The COVID-KOP database was used to find real world candidate drugs that can potentially treat COVID-19 (Bobrowski et al. 2021).

SCENT-KOP aims to merge the general information provided by the ROBOKOP graph with more focused scent data. In this study we document our process for gathering and normalizing this data, the technical process for integrating the datasets, and provide some case studies to explore the utility of our new database.

3.1.1: What are Medicinal Odorants?

The earliest evidence of fragrance industry dates back over 4,000 years, to an ancient factory unearthed in Cyprus, which is thought to have been the site of manufacture for both cosmetic and medicinal products (Belgiorno 2016). Historically, many botanical oils, animal musks, natural residues, and man-made concoctions used in cosmetic fragrances have also seen

use as active ingredients in medicinal preparations (Ali et al. 2015; Angelucci et al. 2014). In holding with the historical trend, many odorants used as fragrance ingredients today, are used as active ingredients in pharmaceuticals. Notable examples include camphor (DrugBank 2022a), eucalyptol (DrugBank 2022b), eugenol (DrugBank 2022c), menthol (DrugBank 2022d), and turpentine (DrugBank 2022e).

Many molecules used in fragrance also have biochemical effects that can be mechanistically linked to indications common drugs are used to treat. Odorants have also been reported in the literature to elicit clinical outcomes directly, either via the psychophysiological effect of their odor (Sowndhararajan and Kim 2016), or by ingestion (Donelli et al. 2019; McKay and Blumberg 2006). For example, the scent of isopropyl alcohol has been reported to reduce nausea in chemotherapy patients (Lindblad et al. 2018). On the other hand, (-)-Linalool is an odorant that naturally occurs in lavender essential oil, oral ingestion of linalool produces a state of anxiolysis via modulation of GABA, glutamate, and serotonin activity in the CNS (Donelli et al. 2019). What this means is that the medicinal effects of odorants could be mediated by scent perception itself, a non-scent related biological process, or a combination of both mechanisms.

It has been demonstrated that drugs, cosmetics, and pesticides occupy overlapping regions of chemical space (Alves et al. 2018b). Odorant molecules should not be thought of as only capable of eliciting scent percepts; the fact that they are capable of binding olfactory receptors to elicit specific psychological responses makes them drug-like in nature, and they are capable of binding ectopic olfactory receptors (olfactory receptors expressed outside of the olfactory epithelium), and other biochemical targets such as the NMDA ion channel and SERT (Donelli et al. 2019). In Chapter 1, we demonstrated the use of natural language processing techniques to generate standardized verbal scent descriptor profiles for odorant molecules

(Thieme et al. 2022). The dataset that emerged from this process was used to train Quantitative Structure-Odor Relationship (QSOR) models (See Chapter 2). Herein, we utilize this dataset to generate the SCENT-KOP, a novel knowledge graph focusing on biological effects of scent molecules. By accessing biomedical data linked to mono-molecular odorants, it becomes feasible to search in the odorant chemical space for new uses of odorants as drugs or biochemical probes, on the basis of observations made using SCENT-KOP.

3.2: Materials and Methods

3.2.0: Mono-molecular odorants and Verbal Scent Descriptors

Mono-molecular odorants are volatile small molecules that elicit scent percepts via activation of olfactory receptors. Oftentimes, odorant scent quality is characterized by verbal scent descriptor (VSD) terms, i.e., words or phrases that communicate the sensory qualities of odorant scent profiles (Kaepler and Mueller 2013), as discussed in detail in Chapter 1. In the Structure Odor Relationship Dataset (SORD) used in this study (Thieme et al. 2022), there were 2,819 unique mono-molecular odorants, along with 422 unique VSD terms that have been manually applied as labels. As described previously (Thieme et al. 2022), these arbitrary scent descriptors were transformed for each scent molecule into standardized VSD profiles containing semantic (cosine) distance values to 27 VSD terms featured in the standard Primary IFRA scent ontology (International Fragrance Association 2020). The collection, curation, and integration of the SORD is described in detail in Chapter 1 and elsewhere (Thieme et al. 2022).

3.2.1: Integrating Scent Datasets.

We used the Automat graph builder to extend the existing ROBOKOP Knowledge Graph (https://github.com/RENCI-AUTOMAT/Data_services). This service enables users to add novel datasets to enrich the existing knowledge graph by instantiating a Python class representing a

particular biomedical source. Once all sources have been developed, a build process may be run. In this build, the user provides a configuration file written in YAML specifying a list of data sources, the service collects data from each of them, transforms the data into a series of graph nodes and edges, and normalizes all of the labels utilizing the node normalization SRI (Standards and Reference Implementations Component) (<https://nodenormalization-sri.renci.org/docs>).

The source files utilized in the building of the scent knowledge base can be found on the RENCI Stars server (https://stars.renci.org/var/data_services/scent_data/). We catalogued five distinct groups of inputs each given their own file. (1) `primary_ifa_vsd_list.txt`: A list of the 27 verbal scent descriptors produced by the IFRA (International Fragrance Association 2020) was used as our gold standard for all other terms. (2) `sor_dataset_human_generated_vsd_list.txt`: A list of the 422 verbal scent descriptors found in the Scent Odorant Dataset; this is a list of every term used by human experts to describe the scents of various chemicals. (3) `sor_dataset_robokop_id_list.txt`: A list of the 2,684 chemical odorants catalogued in the Scent Odorant Dataset; each chemical is identified by the appropriate PubChem Identifier. (4) `sor_dataset_mmod_sor_dataset_vsd_edges.csv`: A list of the 11,906 tuples of the form (**chemical_odorant**, **verbal_scent_descriptor**, *relationship*); here we catalogue the cases from the SORD where human experts described the scent of a specific chemical, with each scent label given its own tuple. (5) `sor_dataset_mmod_primary_ifa_vsd_cos_dist_weighted_edges.csv`: A list of the 72,468 ($27 * 2,684$) tuples of the form (**chemical_odorant**, **verbal_scent_descriptor**, *relationship*, *cosine distance*); each of these tuples describes the cosine distance calculated between each chemical odorant and all twenty-seven IFRA VSDs. Each of these cosine distances are calculated using the methodology detailed in our previous work (Thieme et al. 2022).

For each chemical in SORD, we mapped each odorant chemical to a corresponding PubChem identifier. We were able to map 2674 of the 2684 odorants in the dataset. Any odorant with an unmappable identifier was discarded. If a molecule already existed as a node within the ROBOKOP database, all previous information on the node was kept, and an *odorant* label was added to the node. This integration of existing nodes with the scent information is one of the key contributions of the SCENT-KOP, as we discuss in further detail in the Case Study (Section 1.4). Our ability to integrate into the existing ROBOKOP knowledge base enables us to find patterns linking scent descriptors to other biomedical entities which we would otherwise have difficulty in uncovering.

To help facilitate the discovery of the described scent relationships, we have introduced a new label for chemical entity nodes to identify them as mono-molecular odorants. The new label *odorant* is applied to all of the 2674 chemicals we found in the SORD database. This label allows us to exclude nodes which have not yet been identified.

941 of the 2,674 *odorants* molecules were already found in ROBOKOP database (previously existing with the label *chemical entities*). Then, we introduced the 1,733 remaining compounds into the database as new nodes with the label *odorants*. All VSD terms were added as unique *verbal scent descriptor* nodes belonging to one of two distinct sub-classes (manually labeled terms, or standard cosine distance terms). Edges were inserted between mono-molecular odorant nodes and verbal scent descriptor nodes, representing the entire set of VSD terms used in online data sources to annotate odorant scent profiles. Additionally, 27 new edges were added between each mono-molecular odorant and the 27 standard Primary IFRA VSD terms. These standardized edges are weighted by calculated semantic distance between online odorant VSD terms and each of the 27 standard Primary IFRA VSD terms.

3.3: Results and Discussion

3.3.1: Description of Scent-KOP.

SCENT-KOP is constructed as a combination of the data described above, which has been normalized and integrated into the existing ontological spaces of the ROBOKOP database. This ontological normalization process is done by the NCATS Translator NodeNormalization service (<https://github.com/TranslatorSRI/NodeNormalization>). Once all nodes have been integrated into the same namespace, we integrate them into a singular Neo4J database. The SCENTKOP database contains 418 *verbal_scent_descriptor* nodes, 2,674 *odorant* nodes and 38,258 *disease* nodes (incorporated in ROBOKOP). Additionally, these nodes are interconnected through edges; we have introduced 6,957 *human_labeled* relationships connecting *odorants* and *verbal_scent_descriptors*. Additionally, we have 72,198 computationally generated relationships connecting *odorants* and *verbal_scent_descriptors* using the standardized verbal scent descriptor methodology referenced above (Thieme et al. 2022). Each of these computed edges contains a cosine distance property, which ranks how closely the descriptor is associated to the odorant. These computed edges represent a distinct set of edges human labeled relationships, there is no overlap between these two sets of edges.

3.3.2: Applications/Case Studies.

Recently, our group has defined the concept of clinical outcomes pathways (COPs). COPs represent a novel method to map drug action, as a discrete series of biochemical events of increasing complexity starting with a molecular initiating event (i.e. biochemical target binding) and ending with events that can be classified as clinical outcomes (Korn et al. 2022). SCENT-KOP can be used to identify COPs for known medicinal odorants, such as peppermint essential oil. Indeed, a cursory web search for medicinal uses of peppermint essential oil indicates it is sometimes used topically to treat headaches (WebMD 2022). Therefore, a query was input to SCENT-KOP in search of paths linking the verbal scent descriptor term “peppermint” to the disease node “migraine disorder”. The subgraph that was returned from this query is shown below in **Figure 3.1**.

In **Figure 3.1** multiple mechanistic clinical outcomes pathways converge on single disease to illustrate the multiplicity of drug action in the case of peppermint oil used to treat migraine headaches. The mono-molecular odorants **methyl salicylate** and **(1)-menthol** are shown to interact with different genes, such as cell surface receptors and cytokines, which can then each act on different downstream processes to influence outcomes related to pain and inflammation. Interestingly, the genes that were returned by the query pictured below were not olfactory receptors; but each of the biochemical targets IL-6, TRPA1, and OPRK1, are all known to be involved in the mediation of pain signals (Beck and Dix 2019; Sebba 2021; Souza Monteiro de Araujo et al. 2020).

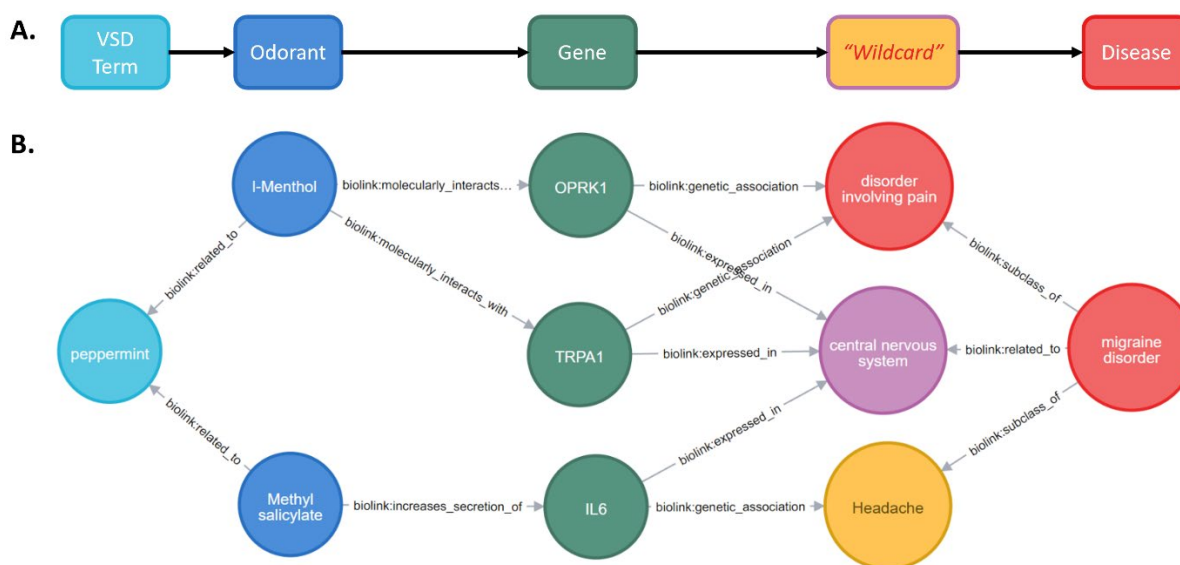


Figure 3.1. Resultant SCENT-KOP subgraph is shown in panel “B”, from a query linking the verbal scent descriptor term “peppermint” to the disease node “migraine disorder”; through pathways consisting of one odorant, one gene node, and one ‘wildcard’ node, as depicted in panel “A”. The full query used in this instance was `MATCH (vsd: verbal_scent_descriptor)-[r0]-(od:odorant)-[r1]-(gn: `biolink:Gene`)-[r2]-(x1)-[r3]-(ds: `biolink:Disease`)` WHERE `vsd.name="peppermint" AND ds.name="migraine disorder" RETURN *`. The x1 segment in the above query represents a wildcard node, a node that can belong to any node class, whereas the rest of the nodes in pathways are specified in the query. Node classes shown above include ‘verbal scent descriptors’ (light blue), ‘odorants’ (blue), ‘genes’ (green), ‘anatomical entities’ (magenta) ‘phenotypic features’ (yellow), and ‘disease states’ (red).

The above “peppermint” odorant case study shows how SCENT-KOP can be used in an exploratory manner to search for the 941 odorants that have been included in ROBOKOP and investigate the COPs that they can participate in. For the remaining 1,733 molecules that are not present in ROBOKOP, the challenge remains to obtain data linking them to biomedical entities currently existing in ROBOKOP. This process is iterative, and we intend to continue to add new edges connecting known odorants to biomedical entities found in odorant-linked databases, such as those that include toxicological data and olfactory receptor binding data.

In the interim, we wanted to see if we can use SCENT-KOP to examine indirect relationships between the scent terms used in standardized VSD profiles, and diseases. We hypothesize that just as chemical similarity can often be used to explain and predict similarity of

action between drugs, similarity in the perceptual profile of odorant molecules can be used to explain and predict similarity between biochemical activity profiles of medicinal odorants. Briefly, this analysis involved querying the SCENT-KOP for the set of all unique pathways linking the 27 standard VSD terms (each represented by a single node) included in the IFRA Primary scent ontology to disease nodes in SCENT-KOP. This query returned a matrix with 90,585 rows (each row represents a unique path linking verbal scent descriptor terms to disease nodes) and 5 columns ((1) verbal scent descriptor term, (2) cosine distance, (3) odorant, (4) odorant relationship to disease, and (5) disease) that was downloaded from SCENT-KOP as a .csv file. In total, this matrix contained all paths linking cosine distances to standard VSD terms for a total of 303 unique odorants, which were linked as a whole to 717 unique diseases.

This matrix was then transformed to yield a 717×27 matrix containing the average cosine distance values from each path linking VSD terms to specific diseases. For each odorant, there were 27 unique cosine distance values. Therefore, for each odorant connected to diseases, there were 27 cosine distance values in the total set of 27 paths linking standard scent terms to disease states. These averaged values were used to calculate Z-scores, that were then used to rank the scent term-disease pairs that have a significant correlation to each other in SCENT-KOP in terms of the cosine distances featured in disease-linked odorant scent profiles. In this way, quantitative measures of indirect relationships between scent percepts and diseases within SCENT-KOP can be assessed, as opposed to direct relationships between specific odorants and diseases.

The statistical measures obtained using that above procedure were then used to rank diseases that are linked to '*minty*' scented odorants, the fourth condition ranked out of 717 was "diabetes mellitus". In other words, this analysis indicated that there is a statistically significant

link in SCENT-KOP between the verbal scent descriptor term “minty”, and “diabetes”. To test the legitimacy of this statistical pairing, the hit ‘*minty*’ compound **carvotanacetone**, from a recent virtual screening exercise in search of new odorants conducted by our group (See Chapter 2; **Table 2.3**), was assessed in terms of its link to diabetes. Compellingly, there is a study demonstrating that **carvotanacetone** is anti-diabetic in rodents (Alsanea and Liu 2017).

Further, searches for genes identified in **Figure 3.1** and diabetes on PubMed reveals that each of the 3 biochemical targets have been independently associated with diabetes in the literature (Hiyama et al. 2018; Kristiansen and Mandrup-Poulsen 2005; Shang et al. 2015). In effect, these observations link both **(l)-menthol** and **methyl salicylate** to diabetes as well. One study describes **(l)-menthol** as an enhancer of glucose homeostasis and attenuator of pancreatic β -cell apoptosis in a rodent model of type II diabetes (Muruganathan et al. 2017). Whether or not there exist true mechanistic relationships between scent perception, and the treatment or pathophysiology of disease states, is a question of great interest as a future area of research to be explored. This secondary case study serves as a clear example of how SCENT-KOP can also be used to examine indirect, yet meaningful, relationships between nodes and edges that are of relevance to the treatment of disease, and how observation of such relationships can support the elucidation of new hypotheses.

3.4: Conclusion

Built as a derivative of ROBOKOP, by augmentation of additional nodes and edges corresponding to molecules, their properties, and functional relationships between nodes, SCENT-KOP allows for more specific investigation at the interface of biomedical and perceptual research. Because chemo-sensation is a natural mode of subjective bioactivity prediction (if something smells like it will make you sick, don’t eat it), subjective scent perception-based data

may allow for discovery of new links between odorant compounds, olfaction, psychophysiological and pharmacological processes in the context of human disease.

Opportunities can be created through the merging of the disciplines of scent perception research and the emerging field of biomedical knowledge graph studies. Identification of new drugs, new combinations of existing drugs, new targets for the treatment of specific indications are all potential outcomes of using SCENT-KOP. Ideally, identification of new medicinal odorants can support drug repurposing efforts, as many of these materials are already available and generally recognized as safe. SCENT-KOP enables us to take a bird's-eye view on problems that have historically been approached at the ground level. We hope to enable such progress through the creation of the SCENT-KOP. The SCENT-KOP tool is freely accessible at

<http://scentkop.apps.renci.org/>.

CONCLUSIONS

At the outset of this dissertation, we hypothesized that we could adapt cheminformatics approaches initially oriented toward pharmaceutical science applications toward the study of mono-molecular odorants and structure-odor relationships. In Chapter 1, we demonstrate the use of NLP techniques in the harmonization of verbal scent descriptor profiles for mono-molecular odorants from idiosyncratic and raw to coherent and standardized formats. In Chapter 2, we exemplify the utility of standardized verbal scent descriptor profiles as training sets for externally validated QSOR models designed to predict mono-molecular odorant profiles from chemical structures. In Chapter 3, we explore the biomedical data linked to mono-molecular odorants in the SORD, through the implementation of SCENT-KOP.

There appears to be no such thing as a biologically neutral odorant. Just as chemical similarity has been used to prioritize bioassays for drug-like molecules, similarity-based techniques can also be applied to odorant molecules. As discussed in Chapter 3, medicinal odorants represent a specific opportunity to test the utility of such similarity-based techniques. By virtue of the mechanism in which scents are perceived, where chemosensory impulses are initiated by the binding of odorants to olfactory receptors, all odorants share a key drug-like property: they are capable of binding specific biochemical targets in order to elicit their effects. SCENT-KOP operates by mapping '*systems of systems*' in the conceptual space around mono-molecular odorants. Many links have been made to odorants and other biomedical entities, now the challenge that remains is to organize these links in a way that enables reasoning in regard to the study of olfaction, as well as to other areas of research, such as clinical research and drug

discovery. The case studies highlighted in Chapter 3 represent an interesting lead for future research projects.

Elucidation of clinical outcomes pathways based on knowledge from SCENT-KOP can support scent repurposing efforts. In order to map relationships between scent percepts, odorants, disease, and other biomedical entities, it was crucial to first establish a means for standardizing scent perceptual data. Thereafter, it was possible to build training sets for predictive QSOR models, designed to predict user-defined, standardized scent profiles. After scent perception-based data for online odorants had been collected, curated, integrated, and harmonized for QSOR tasks, implementation of the SCENT KOP knowledge graph was straightforward. In total, these three elements come together to form a platform for comprehensive characterization of scents and their biomedical properties and prediction of scent properties from their chemical structure.

Overall, we feel that we have achieved the goals outlined in the in the Introduction of this dissertation, and that the work herein serves as an example of successful translation of cheminformatic technique from medicinal chemistry, chemical biology, and pharmaceutical science to the domain of perceptual scent research. Such a framework, where computational methods have been used to quantify qualitative aspects of scent perceptual data, potentially enables rational odorant discovery. As described in the introduction, there exists a vast landscape of discrete applications for scented products. So long as scent percepts can be tied to consumer applications, either heuristically or on some systematic basis, automated systems can (and will) then be employed to support the discovery of odorants with targeted properties.

REFERENCES

- Ache BW, Young JM. 2005. Olfaction: Diverse Species, Conserved Principles. *Neuron* 48:417–430; doi:10.1016/j.neuron.2005.10.022.
- AHC M. 2006. Pearls of wisdom about NAC administration - Interview with Barbara Insley Crouch, PharmD, MSPH. *Contin Med Educ Publ*. Available: <https://www.reliasmedia.com/articles/124892-8216-pearls-8217-of-wisdom-about-nac-administration> [accessed 7 February 2022].
- Ahmed L, Zhang Y, Block E, Buehl M, Corr MJ, Cormanich RA, et al. 2018. Molecular mechanism of activation of human musk receptors OR5AN1 and OR1A1 by (R)-muscone and diverse other musk-smelling compounds. *Proc Natl Acad Sci U S A* 115:E3950–E3958; doi:10.1073/pnas.1713026115.
- Ali B, Al-Wabel NA, Shams S, Ahamad A, Khan SA, Anwar F. 2015. Essential oils used in aromatherapy: A systemic review. *Asian Pac J Trop Biomed* 5:601–611; doi:10.1016/j.apjtb.2015.05.007.
- AllenNLP. 2022. AllenNLP - ELMo — Allen Institute for AI. Available: <https://allennlp.org/allennlp/software/elmo> [accessed 24 February 2022].
- Alsanea S, Liu D. 2017. BITC and S-Carvone Restrain High-Fat Diet-Induced Obesity and Ameliorate Hepatic Steatosis and Insulin Resistance. *Pharm Res* 34:2241–2249; doi:10.1007/s11095-017-2230-3.
- Alves VM, Golbraikh A, Capuzzi SJ, Liu K, Lam WI, Korn DR, et al. 2018a. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure–Activity Relationship Models. *J Chem Inf Model* 58:1214–1223; doi:10.1021/acs.jcim.8b00124.
- Alves VM, Muratov EN, Zakharov A, Muratov NN, Andrade CH, Tropsha A. 2018b. Chemical toxicity prediction for major classes of industrial chemicals: Is it possible to develop universal models covering cosmetics, drugs, and pesticides? *Food Chem Toxicol* 112:526–534; doi:10.1016/j.fct.2017.04.008.
- Anderson S. 1984. Graphical representation of molecules and substructure-search queries in MACCS. *J Mol Graph* 2:83–90; doi:10.1016/0263-7855(84)80060-0.
- Angelucci FL, Silva V V., Dal Pizzol C, Spir LG, Praes CEO, Maibach H. 2014. Physiological effect of olfactory stimuli inhalation in humans: an overview. *Int J Cosmet Sci* 36:117–123; doi:10.1111/ics.12096.
- Arn H, Acree TE. 1998. Flavornet: A database of aroma compounds based on odor potency in natural products. In: *Developments in Food Science*. Vol. 40 of Elsevier. 27.
- Bajgrowicz JA, Gygax P. 2000. EP1184447A1 - Composition having organoleptic characteristics of Safranal. Google Patents. Available: <https://patents.google.com/patent/EP1184447A1/en> [accessed 8 February 2022].
- Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, Dunkel M. 2015. Super Natural II--a database of natural products. *Nucleic Acids Res* 43:D935-9; doi:10.1093/nar/gku886.
- Beck TC, Dix TA. 2019. Targeting peripheral κ -opioid receptors for the non-addictive treatment of pain. *Futur Drug Discov* 1; doi:10.4155/fdd-2019-0022.
- Belgiorno MR. 2016. *The Perfumes of Cyprus. From Pyrgos to Francois Coty the route of a millenary charm*.

- Bizon C, Cox S, Balhoff J, Kebede Y, Wang P, Morton K, et al. 2019. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *J Chem Inf Model* 59:4968–4973; doi:10.1021/acs.jcim.9b00683.
- Bobrowski T, Chen L, Eastman RT, Itkin Z, Shinn P, Chen CZ, et al. 2021. Synergistic and Antagonistic Drug Combinations against SARS-CoV-2. *Mol Ther* 29:873–885; doi:10.1016/j.ymthe.2020.12.016.
- Bushdid C, Magnasco MO, Vosshall LB, Keller A. 2014. Humans can discriminate more than 1 trillion olfactory stimuli. *Science* 343:1370–2; doi:10.1126/science.1249168.
- ChemAxon. 2021. ChemAxon - Software Solutions and Services for Chemistry & Biology. Available: <https://chemaxon.com/> [accessed 7 December 2021].
- Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. 2014. QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57:4977–5010; doi:10.1021/jm4004285.
- Cunha JP. 2021. Silenor (Doxepin Tablets): Uses, Dosage, Side Effects, Interactions, Warning. RxList. Available: <https://www.rxlist.com/silenor-drug.htm> [accessed 7 February 2022].
- Donelli D, Antonelli M, Bellinazzi C, Gensini GF, Firenzuoli F. 2019. Effects of lavender on anxiety: A systematic review and meta-analysis. *Phytomedicine* 65:153099; doi:10.1016/j.phymed.2019.153099.
- Dravnieks A. 1985. *Atlas of odor character profiles*.
- DrugBank. 2022a. Camphor: Uses, Interactions, Mechanism of Action | DrugBank Online. Available: <https://go.drugbank.com/drugs/DB01744> [accessed 16 February 2022].
- DrugBank. 2022b. Eucalyptol: Uses, Interactions, Mechanism of Action | DrugBank Online. Available: <https://go.drugbank.com/drugs/DB03852> [accessed 16 February 2022].
- DrugBank. 2022c. Eugenol: Uses, Interactions, Mechanism of Action | DrugBank Online. Available: <https://go.drugbank.com/drugs/DB09086> [accessed 16 February 2022].
- DrugBank. 2022d. Menthol: Uses, Interactions, Mechanism of Action | DrugBank Online. Available: <https://go.drugbank.com/drugs/DB14123> [accessed 16 February 2022].
- DrugBank. 2022e. Turpentine: Uses, Interactions, Mechanism of Action | DrugBank Online. Available: <https://go.drugbank.com/drugs/DB11120> [accessed 16 February 2022].
- Dunkel M, Schmidt U, Struck S, Berger L, Gruening B, Hossbach J, et al. 2009. SuperScent--a database of flavors and scents. *Nucleic Acids Res* 37:D291–D294; doi:10.1093/nar/gkn695.
- Eybna Technologies. 2022. Thujopsene Profile | Sesquiterpenes. Available: <https://www.eybna.com/terpene/thujopsene-terpene-profile/> [accessed 22 February 2022].
- Fecho K, Balhoff J, Bizon C, Byrd WE, Hang S, Koslicki D, et al. 2021. Application of MCAT questions as a testing tool and evaluation metric for knowledge graph-based reasoning systems. *Clin Transl Sci* 14:1719–1724; doi:10.1111/cts.13021.
- Fourches D, Muratov E, Tropsha A. 2016. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model* 56:1243–1252; doi:10.1021/acs.jcim.6b00129.
- Golbraikh A, Muratov E, Fourches D, Tropsha A. 2014. Data Set Modelability by QSAR. *J Chem Inf Model* 54:1–4; doi:10.1021/ci400572x.

- GSC. 2022. The Good Scents Company Information System. Available: <http://www.thegoodscentscompany.com/> [accessed 7 February 2022].
- Gutiérrez ED, Dhurandhar A, Keller A, Meyer P, Cecchi GA. 2018. Predicting natural language descriptions of mono-molecular odorants. *Nat Commun* 9:4979; doi:10.1038/s41467-018-07439-9.
- Haddad R, Medhanie A, Roth Y, Harel D, Sobel N. 2010. Predicting Odor Pleasantness with an Electronic Nose. L.J. Graham, ed *PLoS Comput Biol* 6:e1000740; doi:10.1371/journal.pcbi.1000740.
- Hammett LP. 1935. Some Relations between Reaction Rates and Equilibrium Constants. *Chem Rev* 17:125–136; doi:10.1021/cr60056a010.
- Hansch C, Maloney PP, Fujita T, Muir RM. 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194:178–180; doi:10.1038/194178b0.
- Hiyama H, Yano Y, So K, Imai S, Nagayasu K, Shirakawa H, et al. 2018. TRPA1 sensitization during diabetic vascular impairment contributes to cold hypersensitivity in a mouse model of painful diabetic peripheral neuropathy. *Mol Pain* 14; doi:10.1177/1744806918789812.
- Hörberg T, Larsson M, Olofsson J. 2020. Mapping the semantic organization of the English odor vocabulary using natural language data. *PsyArXiv*; doi:https://doi.org/10.31234/osf.io/hm8av.
- Hotti H, Rischer H. 2017. The killer of Socrates: Coniine and Related Alkaloids in the Plant Kingdom. *Molecules* 22:1962; doi:10.3390/molecules22111962.
- Iatropoulos G, Herman P, Lansner A, Karlgren J, Larsson M, Olofsson JK. 2018. The language of smell: Connecting linguistic and psychophysical properties of odor descriptors. *Cognition* 178:37–49; doi:10.1016/j.cognition.2018.05.007.
- International Fragrance Association. 2020. IFRA Fragrance Ingredient Glossary. IFRA Fragr Ingred Gloss. Available: <https://ifrafragrance.org/priorities/ingredients/glossary> [accessed 15 November 2021].
- Ji S, Pan S, Cambria E, Marttinen P, Yu PS. 2020. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *IEEE Trans Neural Networks Learn Syst* 33:494–514; doi:10.1109/TNNLS.2021.3070843.
- Jolliffe IT, Cadima J. 2016. Principal component analysis: a review and recent developments. *Philos Trans R Soc A Math Phys Eng Sci* 374:20150202; doi:10.1098/rsta.2015.0202.
- Joshi V, Peters M, Hopkins M. 2018. Extending a Parser to Distant Domains Using a Few Dozen Partially Annotated Examples. *ACL 2018 - 56th Annu Meet Assoc Comput Linguist Proc Conf (Long Pap 1: 1190–1199)*.
- Kaeppeler K, Mueller F. 2013. Odor Classification: A Review of Factors Influencing Perception-Based Odor Arrangements. *Chem Senses* 38:189–209; doi:10.1093/chemse/bjs141.
- Keller A, Gerkin RC, Guan Y, Dhurandhar A, Turu G, Szalai B, et al. 2017. Predicting human olfactory perception from chemical features of odor molecules. *Science* (80-) 355:820–826; doi:10.1126/science.aal2014.
- Khan RM, Luk C-H, Flinker A, Aggarwal A, Lapid H, Haddad R, et al. 2007. Predicting Odor Pleasantness from Odorant Structure: Pleasantness as a Reflection of the Physical World. *J Neurosci* 27:10015–10023; doi:10.1523/JNEUROSCI.1158-07.2007.

- KNIME. 2020. KNIME | Open for Innovation. Available: <https://www.knime.com/> [accessed 17 February 2020].
- Korn D, Bobrowski T, Li M, Kebede Y, Wang P, Owen P, et al. 2021. COVID-KOP: Integrating emerging COVID-19 data with the ROBOKOP database. *Bioinformatics* 37:586–587; doi:10.1093/bioinformatics/btaa718.
- Korn D, Thieme AJ, Alves VM, Yeakey M, Borba JVV, Capuzzi SJ, et al. 2022. Defining clinical outcome pathways. *Drug Discov Today*; doi:10.1016/j.drudis.2022.02.008.
- Kowalewski J, Ray A. 2020. Predicting Human Olfactory Perception from Activities of Odorant Receptors. *iScience* 23:101361; doi:10.1016/j.isci.2020.101361.
- Kristiansen OP, Mandrup-Poulsen T. 2005. Interleukin-6 and diabetes: The good, the bad, or the indifferent? *Diabetes* 54; doi:10.2337/diabetes.54.suppl_2.S114.
- Leffingwell JC. 2002. The beta-Vetivones. Chirality and Odour. Available: <http://www.leffingwell.com/chirality/vetivone.htm> [accessed 7 February 2022].
- Lehrer A. 2009. Aromas and Wine Wheels. In: *Wine and Conversation*. Oxford University Press. 42–50.
- Lindblad AJ, Ting R, Harris K. 2018. Inhaled isopropyl alcohol for nausea and vomiting in the emergency department. *Can Fam Physician* 64: 580.
- McKay DL, Blumberg JB. 2006. A review of the bioactivity and potential health benefits of peppermint tea (*Mentha piperita* L.). *Phyther Res* 20:619–633; doi:10.1002/ptr.1936.
- Merck KGaA, Darmstadt G and/or its affiliates. 2019. Flavors & Fragrances Catalog | Sigma-Aldrich. Sigma Aldrich Website. Available: <https://www.sigmaaldrich.com/industries/flavors-and-fragrances/learning-center/catalog-request.html> [accessed 17 December 2019].
- Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient Estimation of Word Representations in Vector Space.; doi:10.48550/arxiv.1301.3781.
- Morgan HL. 1965. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J Chem Doc* 5:107–113; doi:10.1021/c160017a018.
- Morton K, Wang P, Bizon C, Cox S, Balhoff J, Kebede Y, et al. 2019. ROBOKOP: an abstraction layer and user interface for knowledge graphs to support question answering. *Bioinformatics* 35:5382–5384; doi:10.1093/bioinformatics/btz604.
- Muruganathan U, Srinivasan S, Vinothkumar V. 2017. Antidiabetogenic efficiency of menthol, improves glucose homeostasis and attenuates pancreatic β -cell apoptosis in streptozotocin–nicotinamide induced experimental rats through ameliorating glucose metabolic enzymes. *Biomed Pharmacother* 92:229–239; doi:10.1016/j.biopha.2017.05.068.
- Myristicinaldehyde. 2018. Preparation of 5-hydroxyvanillin. *Sci Discuss Board*. Available: <https://www.sciencemadness.org/whisper/viewthread.php?tid=83371> [accessed 7 February 2022].
- Noble A. 2022. Discover Ann Noble’s Aroma Wheel. Available: <https://www.winearomawheel.com/ann-noble-aroma-wheel.html> [accessed 13 January 2022].
- Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. 2018. Deep Contextualized Word Representations. *Proc 2018 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Vol 1* (Long Pap 1:2227–2237; doi:10.18653/v1/N18-1202).

- Pistollato F, Madia F, Corvi R, Munn S, Grignard E, Pains A, et al. 2021. Current EU regulatory requirements for the assessment of chemicals and cosmetic products: challenges and opportunities for introducing new approach methodologies. *Arch Toxicol* 95:1867–1897; doi:10.1007/s00204-021-03034-y.
- PubChem. 2022. 8-Hydroxyquinoline. Available: <https://pubchem.ncbi.nlm.nih.gov/compound/8-Hydroxyquinoline> [accessed 5 April 2022].
- PubChemPy. 2022. PubChemPy documentation — PubChemPy 1.0.4 documentation. Available: <https://pubchempy.readthedocs.io/en/latest/index.html> [accessed 13 January 2022].
- RDKit. 2022. Avalon package — The RDKit 2019.09.1 documentation. Available: <https://www.rdkit.org/docs/source/rdkit.Avalon.html> [accessed 17 February 2020].
- RDKit. 2020. RDKit. Available: <https://www.rdkit.org/> [accessed 17 February 2020].
- Riad N, Zahi MR, Bouzidi N, Daghbouche Y, Touafek O, El Hattab M. 2021. Occurrence of Marine Ingredients in Fragrance: Update on the State of Knowledge. *Chemistry (Easton)* 3:1437–1463; doi:10.3390/chemistry3040103.
- Rossiter KJ. 1996. Structure–Odor Relationships. *Chem Rev* 96:3201–3240; doi: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. *J. Baudry, ed PLoS One* 16:e0252486; doi:10.1371/journal.pone.0252486.
- Sanchez-Lengeling B, Wei JN, Lee BK, Gerkin RC, Aspuru-Guzik A, Wiltschko AB. 2019. Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules.
- Sander T, Freyss J, Von Korff M, Rufener C. 2015. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *J Chem Inf Model* 55:460–473; doi:10.1021/ci500588j.
- Sebba A. 2021. Pain: A Review of Interleukin-6 and Its Roles in the Pain of Rheumatoid Arthritis. *Open Access Rheumatol Res Rev Volume* 13:31–43; doi:10.2147/OARRR.S291388.
- Shang Y, Guo F, Li J, Fan R, Ma X, Wang Y, et al. 2015. Activation of κ -opioid receptor exerts the glucose-homeostatic effect in streptozotocin-induced diabetic mice. *J Cell Biochem* 116:252–259; doi:10.1002/jcb.24962.
- Sharma A, Kumar R, Ranjta S, Varadwaj PK. 2021. SMILES to Smell: Decoding the Structure-Odor Relationship of Chemical Compounds Using the Deep Neural Network Approach. *J Chem Inf Model* 61:676–688; doi:10.1021/acs.jcim.0c01288.
- Souza Monteiro de Araujo D, Nassini R, Geppetti P, De Logu F. 2020. TRPA1 as a therapeutic target for nociceptive pain. *Expert Opin Ther Targets* 24:997–1008; doi:10.1080/14728222.2020.1815191.
- Sowndhararajan K, Kim S. 2016. Influence of Fragrances on Human Psychophysiological Activity: With Special Reference to Human Electroencephalographic Response. *Sci Pharm* 84:724–751; doi:10.3390/scipharm84040724.
- Spence C. 2020. Using Ambient Scent to Enhance Well-Being in the Multisensory Built Environment. *Front Psychol* 11; doi:10.3389/fpsyg.2020.598859.
- Statista. 2021. Global: fragrance care market revenue 2012-2025 | Statista. Available: <https://www.statista.com/forecasts/1268484/worldwide-revenue-fragrance-care-market> [accessed 9 December 2021].

- Thieme A, Korn D, Alves V, Muratov E, Tropsha A. 2022. Novel Classification of Mono-Molecular Odorants using Standardized Semantic Profiles.; doi:10.26434/CHEMRXIV-2022-H64SB.
- van der Maaten L, Hinton G. 2008. Visualizing Data using t-SNE. *J Mach Learn Res* 9: 2579–2606.
- WebMD. 2022. Peppermint Oil Uses, Benefits, Effects, and More. Available: <https://www.webmd.com/a-to-z-guides/peppermint-oil-uses-benefits-effects> [accessed 17 February 2022].
- Willett P, Barnard JM, Downs GM. 1998. Chemical Similarity Searching. *J Chem Inf Comput Sci* 38:983–996; doi:10.1021/ci9800211.
- Wise PM, Olsson MJ, Cain WS. 2000. Quantification of Odor Quality. *Chem Senses* 25:429–443; doi:10.1093/chemse/25.4.429.
- Yan Y, Chen S, Nie Y, Xu Y. 2021. Quantitative Analysis of Pyrazines and Their Perceptual Interactions in Soy Sauce Aroma Type Baijiu. *Foods* 10:441; doi:10.3390/foods10020441.
- Zach. 2021. What is Considered a Good AUC Score? *Statology*. Available: <https://www.statology.org/what-is-a-good-auc-score/> [accessed 23 April 2022].
- Zarzo M. 2015. A Sensory 3D Map of the Odor Description Space Derived from a Comparison of Numeric Odor Profile Databases. *Chem Senses* 40:305–313; doi:10.1093/chemse/bjv012.
- Zarzo M. 2008. Relevant psychological dimensions in the perceptual space of perfumery odors. *Food Qual Prefer* 19:315–322; doi:10.1016/j.foodqual.2007.10.007.
- Zarzo M. 2012. What is a fresh scent in perfumery? Perceptual freshness is correlated with substantivity. *Sensors (Basel)* 13:463–83; doi:10.3390/s130100463.