Our 2020 Spring Colloquium is dedicated in memory of Dr. George Preti.

1944-2020

Monell Chemical Senses Center
2020 Spring Colloquium

BIG DATA AND ARTIFICIAL INTELLIGENCE IN THE CHEMICAL SENSES

April 7, 2020

Cover Image
Photo by Markus Spiske/Unsplash
Cover Design by Alyssa Treff

SCHEDULE OF EVENTS

Event Summary

New methods in the manipulation and analysis of very large and multi-dimensional datasets are transforming how we collect data and conduct chemosensory research. These methods hold the potential to revolutionize how food, flavors and fragrances can be transformed to improve diets, environments and health. But with size comes complexity and challenges which can easily be underestimated and can be difficult to navigate without a guide! Presentations will cover the why, what, how and when of very large, high-dimensional data with examples from many disciplines and application areas in basic and applied chemosensory science, personalized medicine and nutrition. In addition to the educational presentations, experts will be on hand during a panel discussion to answer your questions and give you tips and techniques to help you get started or build on your big data projects.

Tuesday, April 7th

Session I	**The ABC's of AI and Big Data**	
9:00 am	Big Data Basics: An Introduction	J. Mainland
9:30 am	Knowledge Management	J. Ennis*
10:00 am	What is "Artificial Intelligence/Machine Learning" and Why Should I Care?	A. Wiltschko*
10:25 am	**Coffee Break**	
10:30 am	The Pyrfume Project – Organizing Odorant-linked Olfactory Data for Open Research	R. Gerkin*
11:00 am	How to Build a Sensory Dataset for Machine Learning	E. Mayhew
11:30 am	Accessible Artificial Intelligence for Sensory Research	J Moore*
12:00 pm	**Lunch**	
Session II	**Applications**	
1:30 pm	The Machinery of Odor Detection: A Computational Microscope Focused on the Sense of Smell	C.A. de March*
2:00 pm	Using Molecular Profiles to Diagnose Disease	B. Kimball/ N. Beauchamp*
2:30 pm	Revisiting the National Geographic Data as a Training Ground for Big Data Scientists	Monell R-Club

3:00 pm	**Coffee Break**	
3:30 pm	Assessing the Relationship Between Ambient Air Pollution and Olfactory Function: Revisiting the 1986 National Geographic Smell Survey	V. Ramirez*
4:00 pm	Personal Differences in Sensory Experience and Human Health	D. Reed
4:30 pm	Panel Discussion – Getting Started: Questions, Tools, and Considerations	Moderator: C. Thorp*

***Guest Speakers**

Nicholas Beauchamp, PhD
Assistant Professor of Political Science
Northeastern University

Claire A. de March, PhD
Postdoctoral Researcher
Duke University

John Ennis, PhD
President
Aigora

Rick Gerkin, PhD
Associate Research Professor
Arizona State University

Jason Moore, PhD
Director, Institute of Biomedical Informatics
University of Pennsylvania

Vicente Ramirez
PhD Student, Public Health
University of California Merced

Clare Thorp, PhD
Senior Vice President, North America
Creme Global

Alex Wiltschko, PhD
Research Scientist
Google Inc.

PRESENTATION ABSTRACTS

Knowledge Management

John Ennis, PhD

Consumer research is at its best when it responds to business questions. But, too often, we begin our research from scratch when we already own a wealth of knowledge in our historical data. If only we could easily access that knowledge to inform our research design and statistical analysis!

In this presentation, to help practitioners unlock their historical knowledge, we provide an introduction to graph databases, which are a category of databases that emphasize individual units of information and the relationships between them. We explain how this ascendant technology, used extensively in social media and e-commerce, turns many of the traditional limitations of sensory and consumer research - most notably the many diverse data streams we must navigate - into strengths. We then explain how graph databases are ideally suited to connecting diverse data without imposing unneeded structure, thereby facilitating rapid access to historical knowledge across multiple data sources that would otherwise have remained partitioned into silos.

Using graph databases, as we illustrate through example, sensory and consumer researchers can conduct what amount to numerous virtual pilot studies before launching new research. Moreover, graph databases support machine learning investigations, which can both answer and suggest new research questions. And, finally, once new research has been conducted, graph databases can help set priors and norms to support analysis and interpretation.

At the conclusion of this presentation, attendees will understand the many benefits offered by graph databases and will be well-positioned to leverage their historical data to new advantage.

For more information about Dr. Ennis visit: https://www.linkedin.com/in/johnmichaelennis/

What Is “Artificial Intelligence/Machine Learning” and Why Should I Care?

Alex Wiltschko, PhD

Predicting the relationship between a molecule's structure and its odor remains a difficult, decades-old task. This problem, termed quantitative structure-odor relationship (QSOR) modeling, is an important challenge in chemistry, impacting human nutrition, manufacture of synthetic fragrance, the environment, and sensory neuroscience. Recent advances in the fields of artificial intelligence (AI) and machine learning (ML) promise to push progress forward in QSOR. In this presentation, we will review exactly what has changed over the last five years that has brought AI/ML to the forefront in industry, and highlight the already large impact it has had on the digitization of sight and sound. Then, we will highlight what AI/ML has to offer in the field of olfaction, and share some early results Google has found in collaboration with Monell.

For more information about Dr. Wiltschko visit: https://research.google/people/105779/

The Pyrfume Project – Organizing Ordorant-Linked Olfactory Data for Open Research

Rick Gerkin, PhD

Progress on theories and predictive models of olfactory perception is limited by obstacles to obtaining and using datasets reported in the academic literature or in industrial settings. We are creating a convenient entrypoint for olfactory psychophysics work by 1) curating and aligning numerous datasets of responses to identified odorants assessed by human perception and by neural responses in both human and animal models; and by 2) creating a framework for testing olfactory prediction models according to their agreement with those datasets, permitting a frank assessment of the state of the field by revealing gaps in current understanding, and enabling model-guided experimental design. These tests will serve as benchmarks for new models to meet and exceed. This project, entitled “Pyrfume”, is being disseminated using Python, web APIs, and a web application.

For more information about Dr. Gerkin visit: https://sols.asu.edu/richard-gerkin

How to Build a Sensory Dataset for Machine Learning

Emily Mayhew, PhD

Machine learning is the implementation of algorithms (series of conditional steps) to build, test, and revise models that results in improved performance on a task with experience. Models generated using machine learning algorithms can learn even complex patterns from data, but researchers need to provide accurate and sufficient data from which to learn. We are interested in using machine learning to find patterns relating molecular structure to odor perception, but must first collect accurate and sufficient data to train a model. Many sensory methods can be used to generate profile data. The ideal method must generate data with (1) high resolution, describing even small sensory differences; (2) high consistency, generating the same profile for a given sample each time it is measured; and (3) adequate speed, such that a large data set can be feasibly collected. We tested 8 trained and 15 novice subjects on 50 odors using six candidate methods of odor characterization. We found that trained subjects were more consistent as a group than novice subjects (trained homogeneity = 36%, novice = 27%), applied more descriptors to each odor (5.2 vs. 4.2 descriptors/odor), and produced higher dimensional profile data (first 2 dimensions of correspondence analysis explained 26% variance vs. 32%). However, trained subjects spent longer on each evaluation (145 s vs. 20 s per odor, $p < 0.001$). By measuring the data quality and the speed of evaluation for different methods and subject types, we can select the approach that yields the highest quality data efficiently. We plan to scale-up the winning approach and generate a large, open-access database on the odor character of molecules that will raise the performance ceiling for structure-odor modeling in olfaction.

References

Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., … Meyer, P. (2017). Predicting human olfactory perception from chemical features of odor molecules. Science, 355(6327), 820–826. https://doi.org/10.1126/SCIENCE.AAL2014

Sanchez-Lengeling, B., Wei, J. N., Lee, B. K., Gerkin, R. C., Aspuru-Guzik, A., & Wiltschko, A. B. (2019). arXiv. Machine Learning for Scent: Learning Generalizable Perceptual. http://arxiv.org/abs/1910.10685

For more information about Dr. Mayhew visit: https://www.monell.org/faculty/people/emily_mayhew

Accessible Artificial Intelligence for Sensory Research

Jason Moore, PhD

Machine learning is a common computational approach for identifying patterns in big data. However, machine learning can be complex, time consuming, and intimidating for inexperience users. One of the challenges is knowing which machine learning algorithms to choose and how to tune their parameters. Automated machine learning (AutoML) has emerged as a solution to the problem. With AutoML, algorithm selection and parameter tuning are done automatically thus making machine learning more accessible. We have taken this a step further and built an AutoML system that can learn from experience to automate this process. Our PennAI method and software consists of a machine learning code base, a controller for launching analyses, a database of results that serves as a memory for the system, an artificial intelligence algorithm that can learn what algorithm is right for a given data set, and a user-friendly interface that allows users to do machine learning with a single press of a button. We have evaluated PennAI on a collection of more than 150 real and simulated data sets and demonstrated that it learns from experience and performs as well or better than the top AutoML methods to date. We discuss how this software can transform sensory research by making machine learning accessible and easy regardless of computational experience and training.

For more information about Dr. Moore visit:
https://www.med.upenn.edu/apps/faculty/index.php/g275/p8803452

The Machinery of Odor Detection: A Computational Microscope Focused on the Sense of Smell

Dr. Claire A. de March, PhD

The complexity of the odor chemical space and the large number of odorant receptors (ORs) associated with their combinatorial activation make understanding odor coding an enormous challenge. More specifically, being able to predict the behavior of an OR bound to an agonist, an antagonist or a non-agonist is still challenging.

Using a joint approach combining molecular modeling and experimental data on several ORs, we have built a model that can capture the active or inactive state of these proteins when bound to odorants with different potencies. This approach also allowed us to identify the receptor stability as a crucial feature for ORs expression in non-olfactory cells.

The methodology is illustrated on challenging cases. By the aim of computational tools combined with site directed mutagenesis, we predict the activation of human and mouse ORs by their strong agonists, and their inactivation form by non-agonists. These models allow us to investigate the activation mechanism of all mammalian olfactory receptors from the binding cavity to the G protein binding site, revealing the network of conserved amino acids involved in OR activation.

We also demonstrate that the divergences from the conserved residues in this family are cryptic mutations which increase OR repertoire diversity at the expense of functionality. The divergence from conserved residues destabilizes the structure of the OR which in turn decrease its cell surface expression in heterologous cells.

Such powerful approaches will help unravel odor-coding in the nervous system and facilitate the understanding the mechanism of neuronal activation induced by an odor.

References

de March, C. A.; Yu, Y.; Ni, M. J.; Adipietro, K. A.; Matsunami, H.; Ma, M.; Golebiowski, J., Conserved residues control Activation of mammalian G protein-coupled odorant receptors. Journal of the American Chemical Society 2015, 137 (26), 8611-8616.

Ikegami, K.; de March, C. A.; Nagai, M. H.; Ghosh, S.; Do, M.; Sharma, R.; Bruguera, E. S.; Lu, Y. E.; Fukutani, Y.; Vaidehi, N.; Yohda, M.; Matsunami, H., Structural instability and divergence from conserved residues underlie intracellular retention of mammalian odorant receptors. Proc Natl Acad Sci U S A 2020, 117 (6), 2957-2967.

For more information about Dr. de March visit: https://mgm.duke.edu/faculty-and-research/primary-faculty/hiroaki-matsunami-phd/hiroaki-matsunami-phd-lab-members/claire-de-march-phd/

Using Molecular Profiles to Diagnose Disease

Bruce Kimball, PhD and Nick Beauchamp, PhD

The recent enthusiasm for exploiting chemical signals for disease diagnosis arises from decades of anecdotal and empirical evidence and recent advances in chemical analysis. In addition, dogs and rodents have been taught, via associative learning, to discriminate among many health-related conditions on the basis of olfaction. However, there are many reasons to prefer using chemometric approaches -- instrumental chemical analysis and statistical modelling -- to replicate the sensitivity and selectivity of trained animals. Among diagnosticians, there is almost universal preference for instrumental approaches versus trained animals. Importantly, these instrumental techniques rarely report the appearance of unique odorants in response to disease. Disease and healthy states are generally represented by different patterns among the same collection of odorants. Since these patterns may be quite complex and non-linear, relatively new statistical methods and large training datasets may be necessary in order to replicate the performance of animal discrimination models.

We examine here the important considerations for exploration of disease biomarkers and a variety of machine learning methods for uncovering patterns of volatile metabolites underlying disease states. The challenges are two-fold: signal patterns are complex, and the training data are relatively small. We test a variety of methods suited for small, complex data for out-of-sample predictive accuracy (such as neural networks, random forests, and support vector machines) and further apply a variety of methods for boosting small-sample accuracy (such as synthetic oversampling and ensemble methods). We find that the best combination of methods produces an out-of-sample accuracy comparable to that of animal discrimination. While singular chemical signatures remain elusive, complex patterns of signals are both detectable, and may shed some light on the processes employed by olfaction in animal discrimination.

For more information about Dr. Kimball visit: https://www.monell.org/faculty/people/bruce_kimball

For more information about Dr. Beauchamp visit:
https://cssh.northeastern.edu/people/faculty/nicholas-beauchamp/

Revisiting the National Geographic Smell Survey as a Training Ground for Big Data Scientists

Joel Mainland, PhD and Danielle Reed, PhD

Big data requires tools that are not often taught to students in food science, sensory biology, or chemistry—all fields relevant to taste and smell. Data science skills are only slowly being added to companies that make the products that rely on flavors and fragrances. Thus, there is a need for data scientist training programs for people that can combine expertise in taste and smell biology with the skills to analyze and interpret millions or hundreds of millions of observations. We meet this training need at Monell, in part, through R-Club, which meets weekly to train newcomers and experienced programmers in the use of the statistical language R. This software is freely available and is the conduit for the newest methods of statistical analysis, e.g., machine learning, natural language processing, and image analysis. To hone our skills we held a brief hack-a-thon to analyze the 1986 National Geographic Smell Survey which surveyed 1.42 million readers and remains the largest such study to understand the human sense of smell. Here we share our training process for data scientists and display how analysis of this classic dataset can yield potentially actionable insights.

References

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Wysocki, C. J., & Gilbert, A. N. (1989). National Geographic Smell Survey. Effects of age are heterogenous. *Ann N Y Acad Sci, 561*, 12-28. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2735671

For more information about Dr. Mainland visit: https://www.monell.org/faculty/people/joel_mainland

For more information about Dr. Reed visit: https://www.monell.org/faculty/people/danielle_reed

Assessing the Relationship Between Ambient Air Pollution and Olfactory Function: Revisiting the 1986 National Geographic Smell Survey

Vicente Ramirez, PhD

Olfactory function, which is of interest to health science researchers, is an important sensory mechanism used to help understand our environment. Previous literature shows that exposure to airborne chemicals, such as those found in air pollution, has been associated with decrements in olfactory function. However, the effect of pollutants in the ambient environment on olfactory sensitivity has seldom been explored, and there is little known about the types of air pollutants that affect our ability to smell. Further, many of these studies have focused on comparing high and low exposure groups, and thus do not address important aspects of this relationship. To address these gaps, we examined the relationship between chronic exposure to ambient air pollution and olfactory function. We merged the 1986 National Geographic Smell Survey with historic air quality data from the USA EPA data repository. Our findings reveal small, but statistically significant, association between average air pollutant concentration and the perceived intensity of our panel of odorant cues across multiple air pollutants, including NOx, lead, and suspended particulate matter. Further, exposure to air pollution was associated with the correct classification of odors. Particularly, a standard deviation increase in NOx was associated with increased odds of incorrectly identifying amyl acetate (OR=1.11, 95%CI=1.06-1.17) and eugenol (OR 1.14, 95%CI=1.06-1.22). These findings are consistent with previous epidemiological evidence that exposure to air pollution is associated with impact on chemosensory function. This analysis, which uses historic data, suggests that modern large-scale studies assessing individual pollutant exposures and response to different odorants may help to understand the complex nature of these phenomena.

For more information about Vicente Ramirez visit: https://publichealth.ucmerced.edu/content/vicente-ramirez

Personal Differences in Sensory Experience and Human Health

Danielle Reed, PhD

The UK Biobank project collected medical records, DNA genotype and a food intake questionnaire from nearly a half a million people living in the United Kingdom. This Big Data initiative allowed us to examine how genetic variation predicts differences in food choice. This question is interesting for several reasons. First it can tell us something about the mechanism of why people differ in their choice of foods and second, it can point to rational ways to personalize food distribution to different market segments. One of the observations we made was that people reported eating more fruit if they had a particular variant of an olfactory receptor (*OR6B1*; p= $3.59X10^{-49)}$. This observation was supported in a general way by a second big data initiative from the direct-to-consumer testing program 23andMe. They found that the preference for a fruit-flavored dessert item (strawberry vs other flavors of ice cream) was also predicted by the same olfactory receptor variant. These data suggest that some people may be able to smell some desirable aspects of fruitiness and like fruit and fruit flavors more than people with the alternative allele. Personalizing foods is an important growth area and understanding genetic effects gives a rationale and evidence-based method to understand why people choose the foods they do.

References

Cole, J. B., Florez, J. C., & Hirschhorn, J. N. (2019). Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic loci and establishes causal relationships between educational attainment and healthy eating. *bioRxiv*, 662239. doi:10.1101/662239

Reed, D. R., Tanaka, T., & McDaniel, A. H. (2006). Diverse tastes: Genetics of sweet and bitter perception. *Physiol Behav, 88*(3), 215-226. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16782140

https://blog.23andme.com/23andme-research/you-scream-i-scream-our-genes-scream-for-ice-cream/.

For more information about Dr. Reed visit: https://www.monell.org/faculty/people/danielle_reed

PANEL DISCUSSION – Getting Started: Questions, Tools, and Considerations

Clare Thorp, Ph.D.

As scientists, we typically structure our experimental research with the intention of disproving our null hypothesis. We do this by focusing on the variable in question and designing all other variables to be, in essence, *ceteris paribus*. In other words, we evaluate the effect of one variable on another by ensuring all other potentially influential variables remain static.

This works very well for certain scientific questions, for example evaluating a dose-response, or ruling-out / ruling-in critical variables for future studies. However, this approach is much less effective when evaluating a dynamic system where multiple variables interact with each other to create a net overall effect. Such dynamic systems are everywhere in nature. They exist in the microbiome, for example, or in the complex range of responses we have to chemical stimuli such as flavors and fragrances.

Systematic reviews based on a weight of evidence approach across multiple studies is one way to evaluate these types of systems, but it takes time, multiple experiments and considerable expert knowledge to achieve. Hand in hand with expert knowledge also comes internal heuristics and bias. - We are always human beings first and scientists second, and we cannot help but to introduce these subconscious decision-making paradigms into our decision-making process.

This session will examine the opportunities that big data and artificial intelligence offer to evaluate complex systems in a way that is data agnostic. In order to do this however we need to first ‘get started’. The moderator will kick-start the session with some examples of where this approach has been implemented for a complex biological system, and where probability can be used to address gaps, uncertainties and variability in existing data sets. The panel will weigh in with their examples and insights, and answer questions posed by the audience. Listeners will be left with a practical understanding of what ‘getting started’ looks like when taking a data-driven, data-agnostic approach to conducting research on complex biological systems.

For more information about Dr. Thorp visit: https://www.linkedin.com/in/clare-thorp-ph-d-bbb52930/