BLOG ›

Learning to Smell: Using Deep Learning to Predict the Olfactory Properties of Molecules

THURSDAY, OCTOBER 24, 2019

Posted by Alexander B Wiltschko, Senior Research Scientist, Google Research

Smell is a sense shared by an incredible range of living organisms, and plays a critical role in how they analyze and react to the world. For humans, our sense of smell is tied to our ability to enjoy food and can also **trigger vivid memories**. Smell allows us to appreciate all of the fragrances that abound in our everyday lives, be they the proverbial roses, a batch of freshly baked cookies, or a favorite perfume. Yet despite its importance, smell has not received the same level of attention from machine learning researchers as have vision and hearing.

Odor perception in humans is the result of the activation of 400 different types of **olfactory receptors** (ORs), expressed in 1 million **olfactory sensory neurons** (OSNs), in a small patch of tissue called the **olfactory epithelium**. These OSNs send signals to the **olfactory bulb**, and then to further structures in the brain. Based on analogous advances in deep learning for sight and sound, it should be possible to directly predict the end sensory result of an input molecule, even without knowing the intricate details of all the systems involved. Solving the odor prediction problem would aid in discovering new synthetic odorants, thereby reducing the ecological impact of harvesting natural products. Inspection of the resulting olfactory models may even lead to new insights into the biology of smell.

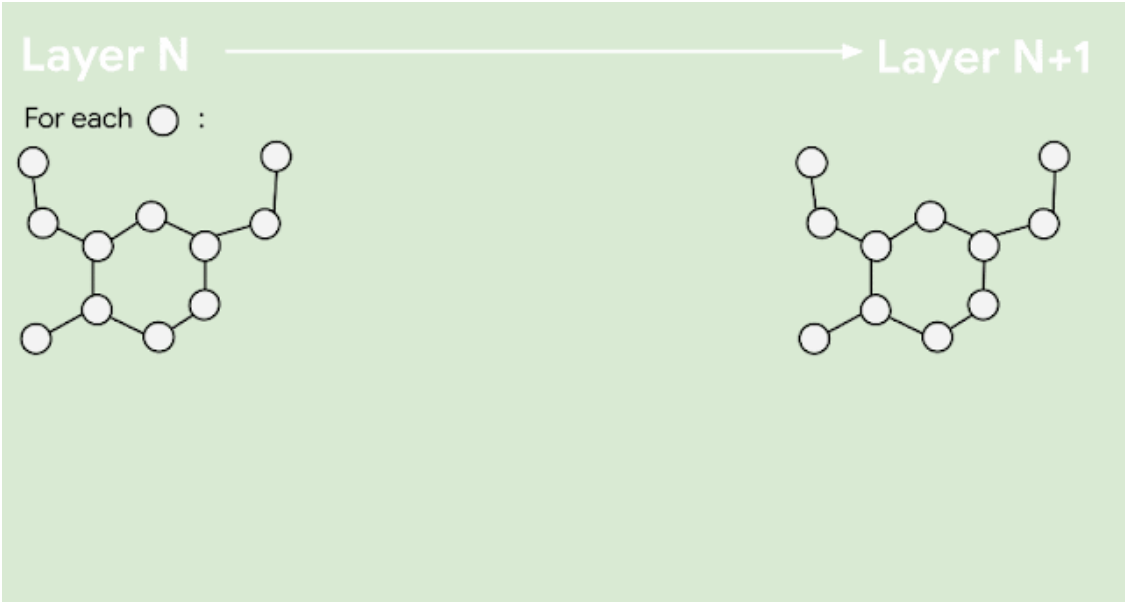
Small odorant molecules are the most basic building blocks of flavors and fragrances, and therefore represent the simplest version of the odor prediction problem. Yet each molecule can have multiple odor descriptors. ***Vanillin***, for example, has descriptors such as *sweet, vanilla, creamy,* and *chocolate*, with some notes being more apparent than others. So odor prediction is also a ***multi-label classification*** problem.

In "**Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules**", we leverage graph neural networks (GNNs), a kind of deep neural network designed to operate on **graphs** as input, to directly predict the odor descriptors for individual molecules, without using any handcrafted rules. We demonstrate that this approach yields significantly improved performance in odor prediction compared to

Graph Neural Networks for Odor Prediction

Since molecules are analogous to graphs, with atoms forming the vertices and bonds forming the edges, GNNs are the **natural model of choice** for their understanding. But how does one translate the structure of a molecule into a graph representation? Initially, every node in the graph is represented as a vector, using any preferred featurization — atom identity, atom charge, etc. Then, in a series of **message passing** steps, every node broadcasts its current vector value to each of its neighbors. An update function then takes the collection of vectors sent to it, and generates an updated vector value. This process can be repeated many times, until finally all of the nodes in the graph are summarized into a single vector via summing or averaging. That single vector, representing the entire molecule, can then be passed into a fully connected network as a learned molecular featurization. This network outputs a prediction for odor descriptors, as provided by perfume experts.

Each node is represented as a vector, and each entry in the vector initially encodes some atomic-level information.



For each node we look at adjacent nodes and collect their information, which is then transformed with a neural network into new information for the centered node. This procedure is performed iteratively. Other variants of GNNs utilize edge and graph-level information.

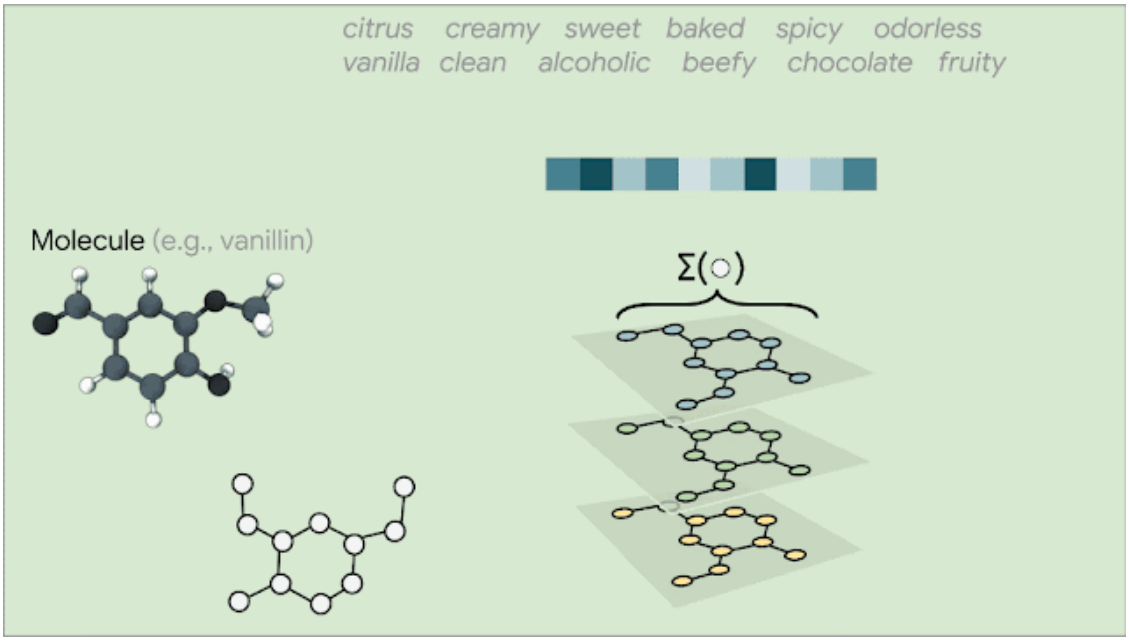
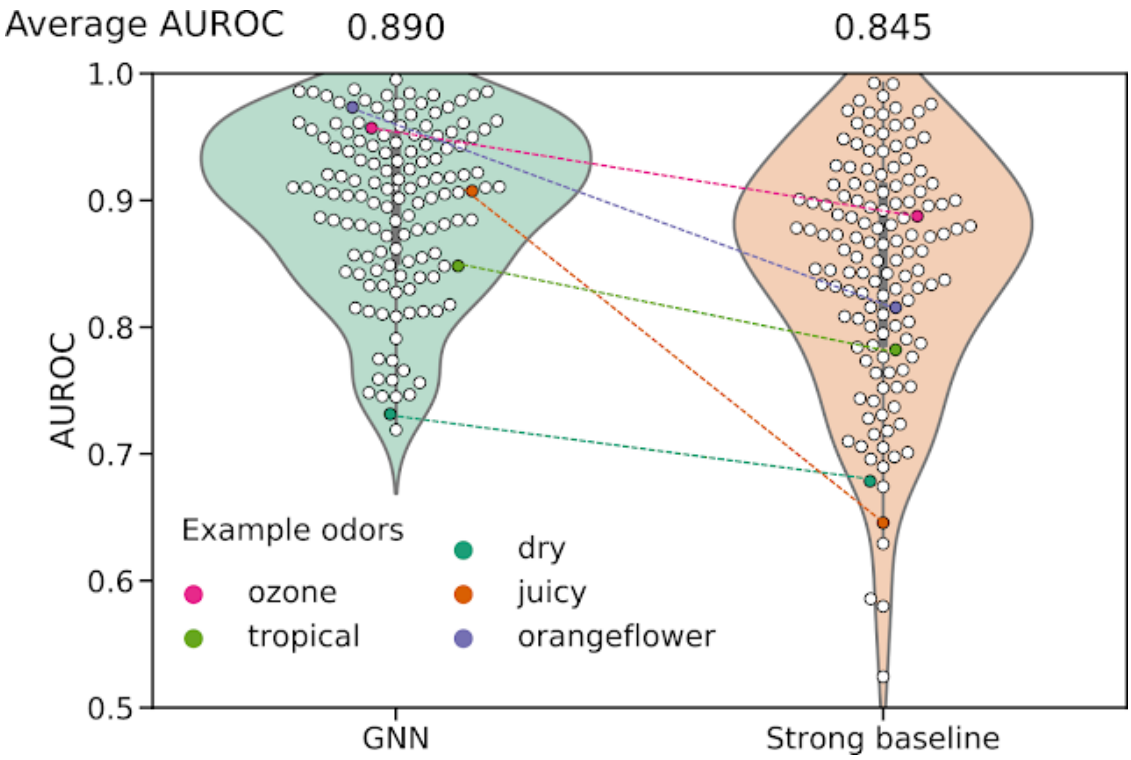


Illustration of a GNN for odor prediction. We translate the structure of molecules into graphs that are fed into GNN layers to learn a better representation of the nodes. These nodes are reduced into a single vector and passed into a neural network that is used to predict multiple odor descriptors.

This representation doesn't know anything about spatial positions of atoms, and so it can't distinguish **stereoisomers**, molecules made of the same atoms but in slightly different configurations that can smell different, such as **(R)- and (S)-carvone**. Nevertheless, we have found that even without distinguishing stereoisomers, in practice it is still possible to predict odor quite well.

For odor prediction, GNNs consistently demonstrate improved performance compared

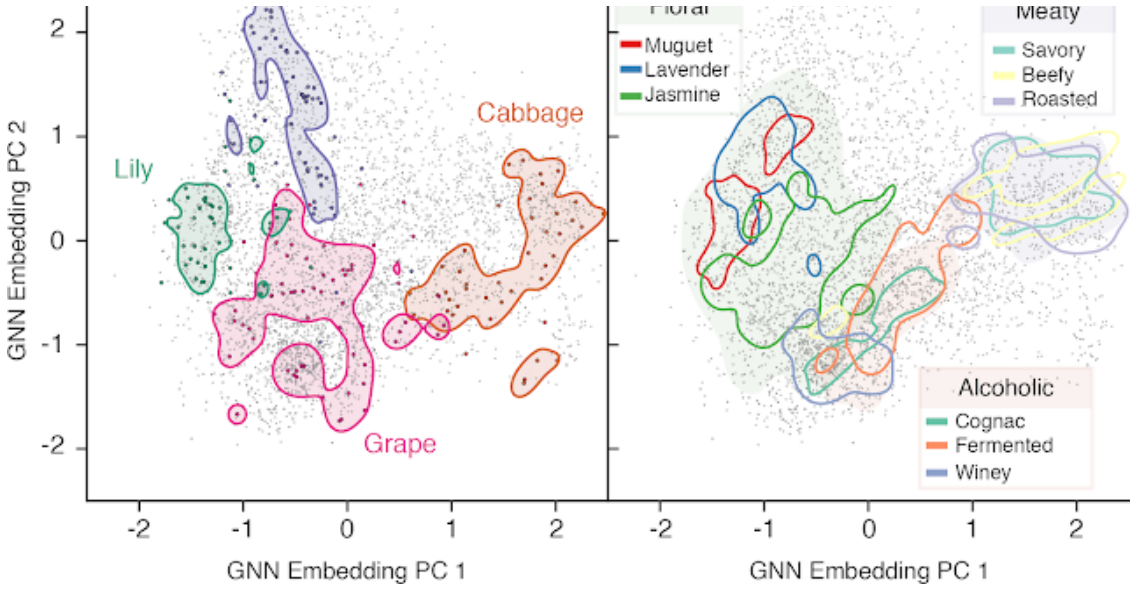
tries to predict.



*Example of the performance of a GNN on odor descriptors against a **strong baseline**, as measured by the **AUROC** score. Example odor descriptors are picked randomly. Closer to 1.0 means better. In the majority of cases GNNs outperform the field-standard baseline substantially, with similar performance seen against other metrics (e.g., **AUPRC, recall, precision**).*

Learning from the Model, and Extending It to Other Tasks

In addition to predicting odor descriptors, GNNs can be applied to other olfaction tasks. For example, take the case of classifying new or refined odor descriptors using only limited data. For each molecule, we extract a learned representation from an intermediate layer of the model that is optimized for our odor descriptors, which we call an "odor embedding". One can think of this as an olfaction version of a **color space**, like RGB or CMYK. To see if this odor embedding is useful for predicting related but different tasks, we designed experiments that test our learned embedding on related tasks for which it was not originally designed. We then compared the performance of our odor embedding representation to a **common chemoinformatic representation** that encodes structural information of a molecule, but is agnostic to odor and found that the odor embedding generalized to several challenging new tasks, even matching state-of-the-art on some.



2D snapshot of our embedding space with some example odors highlighted. ***Left:*** *Each odor is clustered in its own space.* ***Right:*** *The hierarchical nature of the odor descriptor. Shaded and contoured areas are computed with a* ***kernel-density estimate*** *of the embeddings.*

Future Work

Within the realm of machine learning, smell remains the most elusive of the senses, and we're excited to continue doing a small part to shed light on it through further fundamental research. The possibilities for future research are numerous, and touch on everything from designing new olfactory molecules that are cheaper and more sustainably produced, to digitizing scent, or even one day giving those without a sense of smell access to roses (and, unfortunately, also rotten eggs). We hope to also bring this problem to the attention of more of the machine learning world through the eventual creation and sharing of high-quality, open datasets.

Acknowledgements

This early research is the result of the work and advisement of a team of talented researchers and engineers in Google Brain — Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Carey Radebaugh, Max Bileschi, Yoni Halpern, and D. Sculley. We are delighted to have collaborated on this work with Richard Gerkin at ASU and Alán Aspuru-Guzik at the University of Toronto. We are of course building on an enormous amount of prior work, and have benefitted particularly from work by Justin Gilmer, George Dahl and others on ***fundamental methodology in GNNs****, among many other works in neuroscience, statistics and chemistry. We are also grateful to helpful comments from Steven Kearnes, David Belanger, Joel Mainland, and Emily Mayhew.*

Labels: Chemistry Deep Learning Supervised Learning