The Chemistry of Smell: Learning Generalizable Perceptual Representations of Small Molecules

Benjamin Sanchez-Lengeling[1,*], Jennifer N Wei[1,*], Brian K Lee[1], Richard C Gerkin[2], Alán Aspuru-Guzik[3], and Alexander B Wiltschko[1]

[1]Google Research, Brain Team
[2]Arizona State University
[3]Department of Chemistry, University of Toronto
[3]Department of Computer Science, University of Toronto
[3]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[3] Canadian Institute for Advanced Research, Toronto, Ontario, Canada
[*]Contributed equally

Abstract

In machine learning for domains in physics and chemistry, an important problem is learning meaningful and useful representations of physical phenomena that are both predictive and generalize to new tasks. One such problem is predicting quantitative structure-odor relationships (QSOR). A solution would impact human nutrition, manufacture of synthetic fragrance, the environment, and sensory neuroscience. In this paper, we propose the use of graph neural networks for this decades-old task, and show results that significantly out-perform prior methods on a novel data set labeled by olfactory experts. Additional analysis shows that the learned embeddings from graph neural networks capture a meaningful representation of the underlying relationship between structure and odor, as demonstrated by strong performance on two challenging transfer learning tasks. Based on these early results with graph neural networks for molecular properties, we hope the field can build towards doing for olfaction what machine learning has already done for vision.

1 Introduction

Predicting analytical properties of molecules is an area of growing research in machine learning, particularly as models for learning from graph-valued inputs improve in sophistication and robustness. A molecular property prediction problem that has received comparatively little attention during this surge in research activity is the prediction of Quantitative Structure-Odor Relationships (QSOR)[1]. This is a 70+ year old problem straddling chemistry, physics, sensory neuroscience and machine learning. This has remained an open problem for so long due to its difficulty—very small changes in a molecule's structure can have dramatic effects on its odor, a phenomenon known in medicinal chemistry as an "activity cliff" (1; 2). A classic example is *Lyral*, which is a commercially successful *muguet* molecule (a floral scent often used in dryer sheets). Its structural neighbors are not always perceptual neighbors, and vice versa (Figure 1).

Advances in deep learning for vision and audition suggest that we might be able to directly predict the end sensory result of an input stimulus, even without detailed knowledge of the systems and circuits linking the physical world and our internal sensations. Advances in deep learning for olfaction would aid in the discovery of new synthetic odorants, thereby reducing ecological impact of harvesting natural products, and would advance our understanding of sensory perception in

[1] As opposed to Quantitative Structure-Activity Relationships (QSAR), a term from medicinal chemistry

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

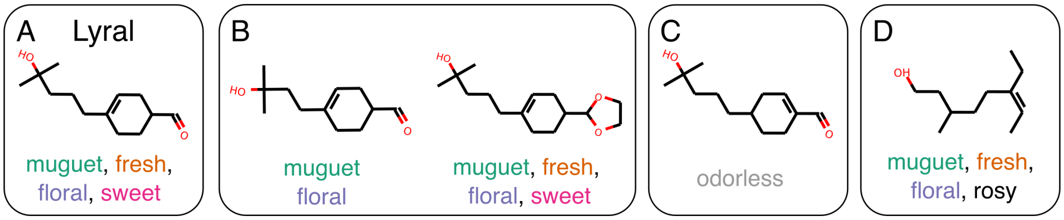

Figure 1: **Structurally similar molecules do not necessarily have similar odor descriptors. A.** Lyral, the reference molecule. **B.** Molecules with similar structure can share similar odor descriptors. **C.** However, a small structural change can render the molecule odorless. **D.** Further, large structural changes can leave the odor of the molecule largely unchanged. Example from Ohloff, Pickenhagen and Kraft (5).

the brain by offering new ways of representing and processing olfactory data. Here, we curated a dataset of molecules associated with expert-labeled odor descriptors[2], and train Graph Neural Networks (GNN) to predict these odor descriptors using a molecule's graph structure alone. Further, we show that the embeddings learned to solve this task generalize well to downstream tasks, a rare occurrence in machine learning applications in chemistry (3; 4). This indicates we have captured a general-purpose representation of a molecule's odor properties, which might be useful in the future for rational molecular design or screening.

2 Olfactory Dataset

We assembled an expert-labeled set of $n = 5030$ molecules from two separate sources: GoodScents perfume materials database, $n = 3786$ (6) and Leffingwell PMP 2001 database, $n = 3561$ (7). The datasets share 2317 overlapping molecules. Molecules are labeled with one or more odor descriptors by olfactory experts (usually a practicing perfumer), creating a multi-label prediction problem. GoodScents describes a list of 1–15 odor descriptors for each molecule (Figure 2A), whereas Leffingwell uses free-form text. Odor descriptors were canonicalized using GoodScents' ontology, and overlapping molecules inherited the union of both datasets' odor descriptors. After filtering for odor descriptors with at least 30 representative molecules, 138 odor descriptors remained, including an *odorless* descriptor (Figure 2B). Some odor descriptors are extremely common, like *fruity* or *green*, while others are rare, like *radish* or *bready*. This bias may be due to our dataset being a collection of materials for perfumery and on the degree of specificity of the odor descriptors. Note that there is an extremely strong co-occurrence structure among odor descriptors that reflects a common-sense intuition of which odor descriptors are similar and dissimilar (Figure 2C). For example, there is a *dairy* cluster that includes the *dairy*, *yogurt*, *milk*, and *cheese* descriptors. There is also a *fruity* cluster with *apple*, *pear*, *pineapple*, *pear*, etc., and a *bakery* cluster that includes *toasted*, *nutty*, and *cocoa*. Previous approaches in QSOR often train one model per odor descriptor. Here, we apply GNNs to all odor descriptor tasks at once, allowing us to take advantage of this correlation structure.

3 Results

3.1 Classification Performance Comparison

We consider two types of GNNs: Message Passing Neural Networks (MPNN) (8) and Graph Convolution Networks (GCN) (9), and compare against baselines of random forests (RF) and k-nearest neighbors (kNN). For baseline featurizations, we used bit-based RDKitFingerprints (bFP), count-based Morgan fingerprints (cFP), and Mordred features (10). We report several metrics (Table 1), as each metric can highlight different performance characteristics, however we primarily compare models on mean AUROC (averaged across odor descriptors). We found that MPNNs and GCNs perform similarly. Both MPNNs and GCNs significantly outperform all baseline models.

[2] In QSAR terminology, *descriptors* are used for input features in a model. The term *odor descriptors* is used in QSOR for odorant properties to predict.

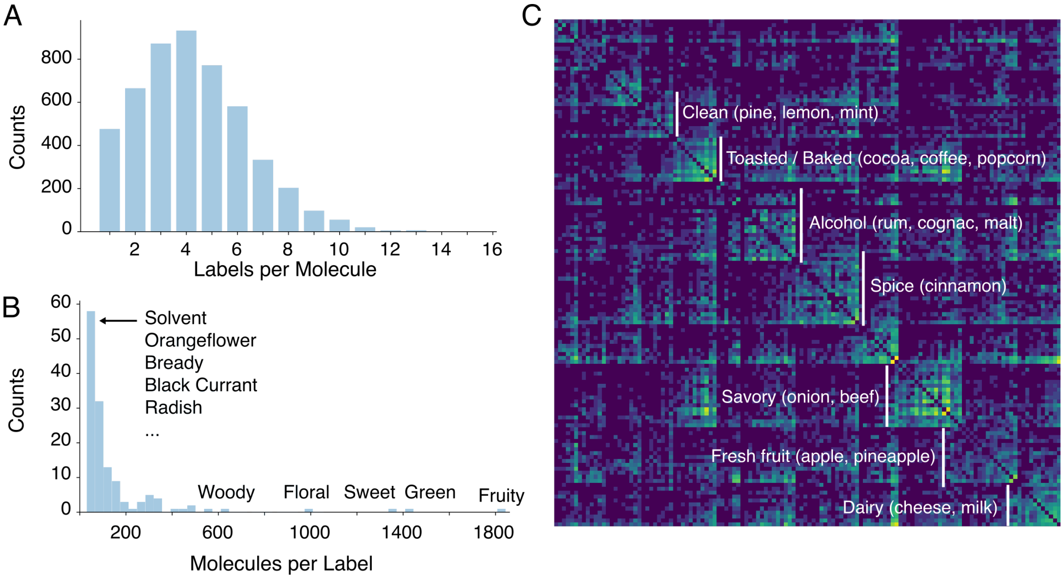

Figure 2: **Dataset overview.** **A.** Distribution of odor descriptor frequencies. **B.** Distribution of label density. **C.** Co-ocurrence matrix for odor descriptors. The 10 most frequent descriptors are removed for visual clarity, and remaining descriptors re-ordered using spectral clustering. Main odor groups with examples are highlighted.

	AUROC	Precision	Recall	F1
MPNN	**0.890 [0.882, 0.898]**	**0.379 [0.352, 0.399]**	0.387 [0.366, 0.408]	**0.362 [0.335, 0.375]**
GCN	**0.894 [0.888, 0.902]**	**0.379 [0.351, 0.398]**	0.390 [0.365, 0.412]	**0.360 [0.337, 0.372]**
RF-Mordred	0.850 [0.838, 0.860]	0.311 [0.288, 0.333]	0.393 [0.372, 0.417]	0.306 [0.283, 0.319]
RF-bFP	0.832 [0.821, 0.842]	0.321 [0.293, 0.339]	0.371 [0.350, 0.390]	0.295 [0.272, 0.308]
RF-cFP	0.845 [0.835, 0.854]	0.315 [0.280, 0.332]	0.375 [0.354, 0.398]	0.295 [0.272, 0.311]
KNN-bFP	0.791 [0.778, 0.803]	0.328 [0.305, 0.347]	0.390 [0.366, 0.411]	0.323 [0.299, 0.335]
KNN-cFP	0.796 [0.785, 0.809]	0.333 [0.307, 0.351]	0.365 [0.342, 0.389]	0.316 [0.292, 0.327]

Table 1: **Odor descriptor prediction results.** mean, 95% CI [lower, upper] bootstrap bounds reported. Numbers reported are an unweighted mean across all 138 odor descriptors. Precision/recall decision thresholds are optimized for F1 score. Best values for each metric are in bold; recall had no clear winner.

3.2 Evaluating Embedding Performance in a Lookup Task

We wished to evaluate whether embeddings extracted from our trained GCN space reflected odor perceptual space. Specifically, we compare whether molecules with small cosine distances in our GCN embeddings were perceptually similar, as compared to using Tanimoto distances on bFP features. While the latter strategy is commonly used in cheminformatics for molecule retrieval, molecules with similar structural features do not always smell the same (Figure 1), so we anticipated that it would not perform as well at retrieving similar-smelling molecules.

To test this idea, we trained a k-nearest neighbors classifier ($k = 20$) using cosine distance on GCN embeddings, and using Tanimoto distance on bit-based Morgan fingerprints. GCN embeddings (AUROC = 0.818, 95% CI [0.806, 0.830]) outperformed bFP (AUROC = 0.782, 95% CI [0.773, 0.797]). By inspecting the nearest neighbors found by each method (Figure 3), we can see that both methods yield molecules with similar structural features, but retrieval using GCN embeddings produces molecules that are more perceptually similar to the source molecule. We conclude that relative to bit-based fingerprints, GCN embeddings emphasize similarity of smell over similarity in structure.

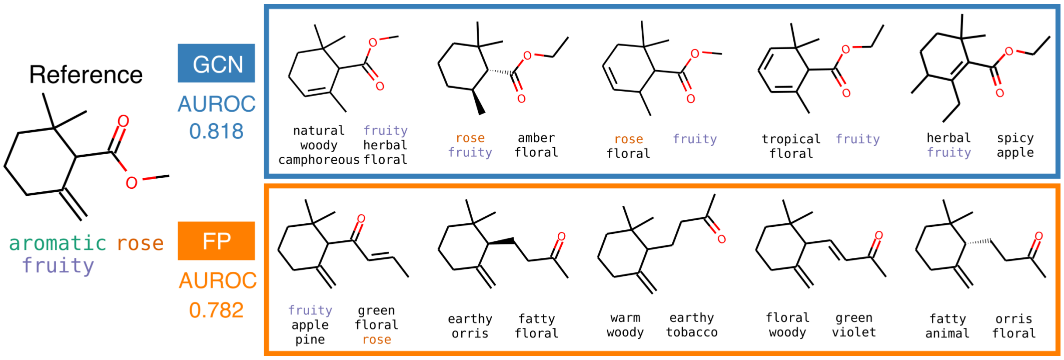

Figure 3: **Nearest neighbor retrieval.** k=5 nearest neighbors are shown for cosine similarity on GCN embeddings and for Tanimoto distance on bit-based Morgan fingerprints. Reported AUROCs are averaged over all odor descriptors, with $k = 20$.

4 Evaluating Generalization Performance of Odor Embeddings

We have shown that GNNs significantly outperform existing methods on predicting odor descriptors, and that our embedding space has useful local structure that can be used to search for similarly-smelling molecules. We now explore whether these odor embeddings generalize to other odor prediction tasks.

4.1 Transfer Learning to Previously-Unseen Odor Descriptors

An odor descriptor may be newly invented or refined (e.g., molecules with the *pear* descriptor might be later attributed a more specific *pear skin, pear stem, pear flesh, pear core* descriptor). A useful odor embedding would be able to perform transfer learning (11) to this new descriptor, using only limited data. To approximate this scenario, we ablated one odor descriptor at a time from our dataset. Using the embeddings trained from $(N - 1)$ odor descriptors as a featurization, we trained a random forest to predict the previously held-out odor descriptor. We used cFP and Mordred features as a baseline for comparison. The results are shown in Figure 4. GNN embeddings significantly outperform Morgan fingerprints and Mordred features on this task, but as expected, still perform slightly worse than a GNN trained on the target odor. This indicates that GNN-based embeddings may generalize to predict new, but related, odors.

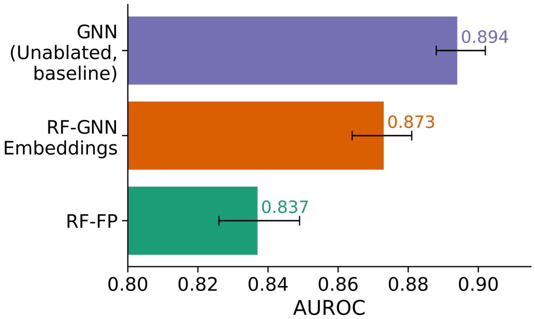

Figure 4: **Mean AUROC on a previously held-out odor.** Average AUROC scores across all labels on the single label ablation task. The error bars denote 95% confidence intervals. The top bar denotes the performance of the model trained on all of the labels. The middle bar denotes the performance of a random forest model trained using the GNN embeddings from a model trained on $(N - 1)$ odor labels. The bottom bar denotes a random forest trained on counting Morgan fingerprints.

4.2 Transfer Learning to the DREAM Olfactory Prediction Challenge

The DREAM Olfaction Prediction Challenge (12) was an open competition to build QSOR models on a dataset collected from untrained panelists. The DREAM dataset has several differences from our own. First, it was a regression problem – panelists rated the amount that a molecule smelled of a particular odor descriptor on a scale from 1 to 100. Second, it had 476 molecules compared to our $\sim 5\text{k}$ (although our dataset contains nearly all of the DREAM molecules). Third, the ratings were provided by a large panel of untrained individuals over a short period of time, whereas ours were gleaned from a small set of experts over many years. The DREAM challenge measured model performance as the Pearson's r correlation of model predictions with the mean reported intensity of each odor descriptor.

The winning DREAM model used random forest models with a combination of several sources of features, primarily Dragon and Morgan fingerprints, among other sources of information (12). Using only our embedding with a random forest model, we achieve a mean Pearson's $r = 0.55$, and the state-of-the-art model described above achieved a mean Pearson's $r = 0.54$. While we can have better average performance in 13 tasks, after taking into account confidence intervals, we find the performance is indistinguishable between the two models for both r and R^2 regression scores (data not shown).

Overall, this indicates that our QSOR modeling approach can generalize to adjacent perceptual tasks, and captures meaningful and useful structure about human olfactory perception, even when measured in different contexts, with different methodologies.

5 Conclusion

We assembled a novel and large dataset of expertly-labeled single-molecule odorants, and trained a graph neural network to predict the relationship between a molecules structure and its smell. We demonstrated state-of-the-art results on this QSOR task with respect to field-recognized baselines. Further, we showed that the embeddings capture meaningful structure on both a local and global scale. Finally, we showed that the embeddings learned by our model are useful in downstream tasks, which is currently a rare property of modern machine learning models and data in chemistry. Thus, we believe our model and its learned embeddings might be generally useful in the rational design of new odorants. For an expanded version of this preprint, please see (13).

6 Acknowledgments

We thank D. Sculley, Steven Kearnes, Jasper Snoek, Emily Reif, Carey Radebaugh, David Belanger, Joel Mainland and Emily Mayhew for support, suggestions and useful discussions on the manuscript. We thank Max Bileschi and Yoni Halpern for their technical guidance. We thank Aniket Zinzuwadia for discussion and who did preliminary work on the DREAM and GoodScents datasets for his senior thesis. We thank Bill Luebke of GoodScents and John Leffingwell of Leffingwell and Associates for their generosity in sharing their data for research use. We thank all of our colleagues in the Google Brain Cambridge office for creating and maintaining such a supportive and stimulating environment.

References

[1] C S Sell. On the unpredictability of odor. *Angewandte Chemie International Edition*, 45(38):6254–6261, 2006.

[2] Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.*, 55(7):2932–2942, April 2012.

[3] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with One-Shot learning. *ACS Cent Sci*, 3(4):283–293, April 2017.

[4] Clyde Fare, Lukas Turcani, and Edward O Pyzer-Knapp. Powerful, transferable representations for molecules through intelligent task selection in deep multitask networks. September 2018, doi:1809.06334.

[5] Günther Ohloff, Wilhelm Pickenhagen, and Philip Kraft. *Scent and Chemistry*. Wiley, January 2012.

[6] The good scents company - flavor, fragrance, food and cosmetics ingredients information. http://www.thegoodscentscompany.com/. Accessed: 2019-9-4.

[7] John C Leffingwell. Leffingwell & associates, 2005.

[8] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272, International Convention Centre, Sydney, Australia, 2017. PMLR.

[9] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Gómez-Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems*, pages 2215–2223, 2015.

[10] Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *J. Cheminform.*, 10(1):4, February 2018.

[11] S J Pan and Q Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, October 2010.

[12] Andreas Keller, Richard C Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, Celine Vens, Leander Schietgat, Kurt De Grave, Raquel Norel, DREAM Olfaction Prediction Consortium, Gustavo Stolovitzky, Guillermo A Cecchi, Leslie B Vosshall, and Pablo Meyer. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, February 2017.

[13] Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. October 2019, doi:1910.10685.