From Molecules to Perception: A Benchmark Dataset for AI in Sensory Science

Dachuan Zhang*
Department of Food Science and Technology
National University of Singapore
Singapore 117542
dachuan.zhang@nus.edu.sg

Abstract

Sensory perception—how molecules taste, smell, and ultimately feel pleasant or unpleasant—plays a critical role in food formulation, cosmetics, and pharmaceuticals. Yet, while vast datasets exist for molecular bioactivity, large-scale, standardized datasets linking chemical structure to sensory attributes remain scarce. This absence is a major bottleneck: promising molecules are often discarded due to undesirable taste or odor, and current AI efforts rely on fragmented, small-scale data with limited predictive power. We propose the Molecular Sensory Dataset (MSD), an open resource designed to capture molecular taste, odor, intensity, and pleasantness at scale. MSD will integrate high-throughput instrumentation—including electronic nose/tongue arrays, n-noise spectrometry, and n-tonne olfactometry—with standardized sensory descriptors and calibrated human panel ratings of hedonic value. Importantly, the dataset will cover both single molecules and mixtures, reflecting real-world applications where synergistic and masking effects shape perception. By establishing a benchmark for AI-driven prediction, generation, and optimization of sensory properties, MSD promises to accelerate discovery across food, fragrance, and drug development, while advancing fundamental understanding of the chemistry of perception.

1 Task definition

The ability of molecules to evoke sensory experiences, such as taste, smell, and pleasantness, is essential in various fields such as food formulation, fragrance design, and pharmaceutical development. Despite rapid advances in predicting molecular bioactivity, AI models still struggle to accurately forecast molecular sensory attributes. We suggest developing a Molecular Sensory Dataset (MSD): a comprehensive, openly accessible resource that connects molecules with validated sensory descriptors (taste and odor), intensities, and pleasantness ratings, covering both individual molecules and mixtures. This dataset would enable multiple AI applications, including predicting sensory properties from molecular structures, designing and refining new compounds with specific sensory profiles, and clustering molecules and mixtures within combined chemical–perceptual–pleasantness spaces.

2 Dataset rationale

Unlike bioactivity datasets such as binding affinities or toxicity measures, molecular sensory data remains scarce, fragmented, and small in scale. Current resources typically cover only hundreds or thousands of common compounds (*Science*, 381, 999-1006, 2023; *J. Agric. Food Chem.* 71, 18, 6789–6802, 2023), often with inconsistent labeling, limited intensity measurements, and almost no

*ORCID: 0000-0003-2467-6286

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

pleasantness ratings, limiting AI's application in sensory science. This scarcity also represents a critical bottleneck for real-world applications. For example, compounds with ideal pharmacological or functional properties are frequently discarded due to undesirable bitterness, metallic aftertaste, or unpleasant odor (*Chemical Senses*, 50, bjaf003, 2025).

To overcome this gap, the proposed MSD would systematically expand and standardize molecular sensory data through high-throughput instrumentation and human validation. We envision the use of n-noise spectrometry to identify odor-active compounds, n-tonne olfactometry to quantify human-perceivable odor intensity, and electronic nose and tongue arrays to capture reproducible taste and odor fingerprints across large chemical libraries. These instrument-based measures would be coupled with a standardized annotation scheme using established sensory lexicons and normalized scales. Crucially, in addition to recording objective descriptors and intensity values, MSD would incorporate hedonic ratings of pleasantness, gathered through calibrated human sensory panels. This integration ensures that AI models learn not only how molecules are perceived, but also how they are valued by human subjects.

Another key feature of the dataset will be its inclusion of mixtures. Most AI's sensory applications, from flavor design to cosmetics, operate on molecular blends rather than isolated compounds (*Science*, 355, 820-826, 2017). Mixtures often exhibit synergistic or masking effects, altering both perceptual qualities and pleasantness in ways that cannot be inferred from single molecules alone, and are much closer to industrial application in most settings. By systematically characterizing binary and ternary mixtures, the MSD will approximate real-world sensory experiences and provide the foundation for AI-guided formulation design for food, cosmetics, and pharmaceuticals.

3 Acceleration potential

The availability of such a dataset has the potential to accelerate multiple scientific and industrial domains. In food science, MSD would enable AI-guided flavor formulation that balances nutritional and sustainability requirements with sensory appeal and consumer liking. In cosmetics, it would support rational fragrance design and the creation of excipients that are not only functional but also pleasant to the senses. In pharmaceuticals, early-stage screening of drug candidates could incorporate taste and odor pleasantness, reducing costly reformulations and improving patient compliance. Beyond applications, the dataset would advance fundamental science by providing a bridge between molecular structure, receptor interactions, perceptual descriptors, and hedonic valence, deepening our understanding of how chemistry maps to human sensory experience. MSD could unlock new advances in molecular generative models, multimodal learning that integrates structure and perception, and AI-driven product design pipelines, enabling the broad scope of AI-driven material discovery to be much closer to practical application and commerce.

4 Data-creation pathway

The pathway for data creation would combine multiple sources and techniques. Public flavor and odor registries will be integrated with libraries of novel molecules, while high-throughput robotic assays and gas chromatography–olfactometry would ensure reproducibility and calibration. Human sensory panels would be used to align instrument-based measurements with perceptual reality, particularly for pleasantness ratings, which remain inherently subjective. Mixtures would be prioritized through factorial design strategies to ensure coverage of the most relevant compound combinations in food, cosmetic, and pharmaceutical contexts. All data would be standardized, anonymized, and linked to open chemical identifiers such as InChI and SMILES, enabling easy integration with molecular modeling frameworks.

5 Cost and scalability

In terms of cost and scalability, we estimate that a high-quality dataset of approximately 10,000 diverse molecules and 1,000 mixtures could be generated with an investment of USD 1–2 million, covering instrumentation, assays, sensory panels, and personnel. With the automation of sensor-based measurements, the marginal cost of data collection would decrease significantly, making it feasible to

scale toward hundreds of thousands of measurements in collaboration with academic and industrial partners worldwide.

6 Conclusion

In conclusion, MSD addresses a fundamental but underexplored bottleneck in AI for molecular discovery: the absence of large-scale, standardized data linking chemical structure to sensory perception and pleasantness. By combining high-throughput instrumentation, mixture-level characterization, and human panel evaluation, MSD promises to establish a new foundation for AI-accelerated innovation across food, cosmetics, and pharmaceuticals. Much like ImageNet for vision or the Protein Data Bank for structural biology, MSD has the potential to redefine this field, enabling AI not only to predict and classify molecules by taste and odor, but also to understand and optimize what human consumers actually find pleasant.

Funding disclosures

This research is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (A-8003718-00-00) and the Start-Up Grant of the National University of Singapore (A-0010237-00-00).

Acknowledgements

We would also like to thank our collaborators and colleagues, including Prof. Shaoquan Liu and Prof. Jianshe Chen, for their valuable input to this project.

References

Kou, X., Shi, P., Gao, C., Ma, P., Xing, H., Ke, Q., & Zhang, D. (2023). Data-Driven Elucidation of Flavor Chemistry. *Journal of Agricultural and Food Chemistry*, 71(18), 6789-6802. https://doi.org/10.1021/acs.jafc.3c00909.

Nguyen, H., Lin, C., Bell, K., Huang, A., Hannum, M., Ramirez, V., . . . Reed, D. R. (2025). Worldwide study of the taste of bitter medicines and their modifiers. *Chemical Senses*, 50. https://doi.org/10.1093/chemse/bjaf003.

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.

Lee, B. K., Mayhew, E. J., Sanchez-Lengeling, B., Wei, J. N., Qian, W. W., Little, K. A., . . . Wiltschko, A. B. (2023). A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661), 999-1006. https://doi.org/doi:10.1126/science.ade4401.