# QuantumScents: Quantum-Mechanical Properties for 3.5k Olfactory Molecules

Jackson W. Burns[†] and David M. Rogers[*,‡]

[†]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*

[‡]*National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831*

E-mail: rogersdm@ornl.gov

**Abstract**

Quantitative Structure-Odor Relationships are critically important for studies related to the function of olfaction. Current literature datasets contain expert-labeled molecules but lack feature data. This paper introduces QuantumScents, a quantum-mechanics augmented derivative of the Leffingwell dataset[1]. QuantumScents contains 3.5k structurally and chemically diverse molecules ranging from 2 to 30 heavy atoms (CNOS) and their corresponding 3D coordinates, total PBE0 energy, molecular dipole moment, and per-atom Hirshfeld charges, dipoles, and ratios. The authors demonstrate that Hirshfeld charges and ratios contain sufficient information to perform molecular classification by training a Message Passing Neural Network with `chemprop`[2] to predict scent labels. The QuantumScents dataset is freely available on Zenodo along with the authors' code, example models, and dataset generation workflow[3].

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the

# Background

Quantitative Structure-Odor Relationships (QSOR) are of immense interest for their potential in describing the mechanism by which humans' olfactory senses function[4]. The DREAM challenge, for example, challenged the scientific community to pursue a mapping of structure to smell[5] and found that reliable prediction was achievable for 8 out of 19 semantic descriptors. So far, the most successful models made use of descriptor sets related to structure as well as 3D coordinates.

Because QSOR modeling is a sub-area of Quantitative Structure-Activity Relationship (QSAR) modeling, new datasets, descriptors, and modeling approaches for QSOR can have wide-ranging impact on diverse problem areas in the agrochemical, pharmaceutical, toxicology, cosmetic, and even chemical synthesis and detection industries[6]. Within QSAR there are two complementary views on special-purpose, interpretive *vs.* general-purpose, predictive models[7]. The first seek to establish trends within a series of similar molecules using few-parameter models – which then point out simple chemical reasons for observed trends. The second seeks to use high-dimensional models to predict properties of a large portion of chemical space – which then internally capture multiple competing effects. Parsimonious selection of input descriptors for these models may hold the key to reconciling these two views, since the remaining descriptors have then proven to be both sufficient and necessary for the prediction task.

In this work, we investigate the hypothesis that Hirshfeld partial atomic charges and

volume ratios[8], element identity, and molecular graph connectivity provide sufficient information for prediction of scent labels. Since early works on molecular structural modeling, it has been established that electrostatic interactions and atomic radii are sufficient for describing the structures of Van der Waals complexes[9]. Recent work convincingly demonstrates that charge and radius also describe a range of specific-anion effects including water binding energy, concentration-dependence of viscosity, ion-mediated reaction rates, enzyme activity, and biopolymer folding[10].

Other recent works on odor prediction from molecular descriptors have utilized 1) random forest models that comb through 4884 physicochemical features of each molecule (including atom types, functional groups, and topological and geometrical properties)[5] and 2) message-passing neural networks with a proprietary design[11] that utilize input features similar to `chemprop`[2]. Random forest models build a decision tree to filter out descriptor combinations important for the prediction task at hand. Message passing networks, also used in this work, combine local information using nonlinear functions to build feature vectors on atoms and bonds that are important to the prediction task at hand. This study is distinguished from the recent work of Lee et. al. by making use of quantum mechanics-derived charge and volume information as input features.

A major product of this work is the QuantumScents dataset, a quantum-mechanics augmented derivative of the established Leffingwell PMP 2001 datatset[1] as prepared by Sanchez-Lengling and coworkers[12]. The original Leffingwell dataset is approximately 3.5k molecules known to the perfume industry that were labeled with freeform text describing their scent. In Ref.[12] the dataset was systematized to 113 different scent descriptors such as "floral", "chocolate", and "sulforous", of which the molecules can be one or more, mapped to the SMILES representation[13] of the molecule. The dataset is unique for its significant structural and chemical diversity with samples ranging from two to thirty heavy atoms and multiple sulfur-containing functional groups.

Figure 1 plots occurrence of scent labels in the dataset with point size representing

number of label occurrences. Although any number of descriptors can occur together, the average number of labels per molecule is 3.7 ± 2.4. It can be seen that some of the labels are very infrequent, making the dataset unbalanced. We will show that this creates a strong dependence of the fit results on the test/train splitting in the section on Dataset Partitioning.

QuantumScents augments this dataset by also providing 3D coordinates for an energy-minimized structure of each molecule as well as quantum-mechanics properties, particularly the per-atom Hirshfeld charges, dipoles, and volume ratios[8]. Note that alternative charge assignment methods exist, and that fitted atomic charges should vary somewhat with the molecule geometry. Briefly, we chose Hirshfeld charges over other available charge partitioning schemes because it depends linearly on the electron density and does not require defining atomic radii. The electron density is a physical observable, and so descriptors computed from it – such as partial atomic charges, multipoles, and "volume ratios"[14] – are more reliable and reproducible than orbital coefficient-based definitions like Mulliken population analysis. Our calculation methodology is described in the Methods section.

In the same study where the Leffingwell data were cleaned[12], and in subsequent work[11], Graph Convolution Networks were successful at predicting the scent labels with a ROC AUC (Area Under Receiver-Operator Curve) value of 0.89. To demonstrate the utility of Hirshfeld parameters in performing this classification task, this work recreates and then extends those results.

Our key findings are that charge-based input features provide better predictivity than the default descriptors present in `chemprop` (atom hybridization, degree, hydrogen count, chirality, aromaticity, and mass). Nevertheless, the imbalance in the occurrence of scent labels can cause a significant skew in measures of fitness for any descriptor setting. We recommend both that Hirshfeld charges be included in descriptors, and that Szymanski splitting be used to minimize the effects of imbalance in occurrence of target descriptors.

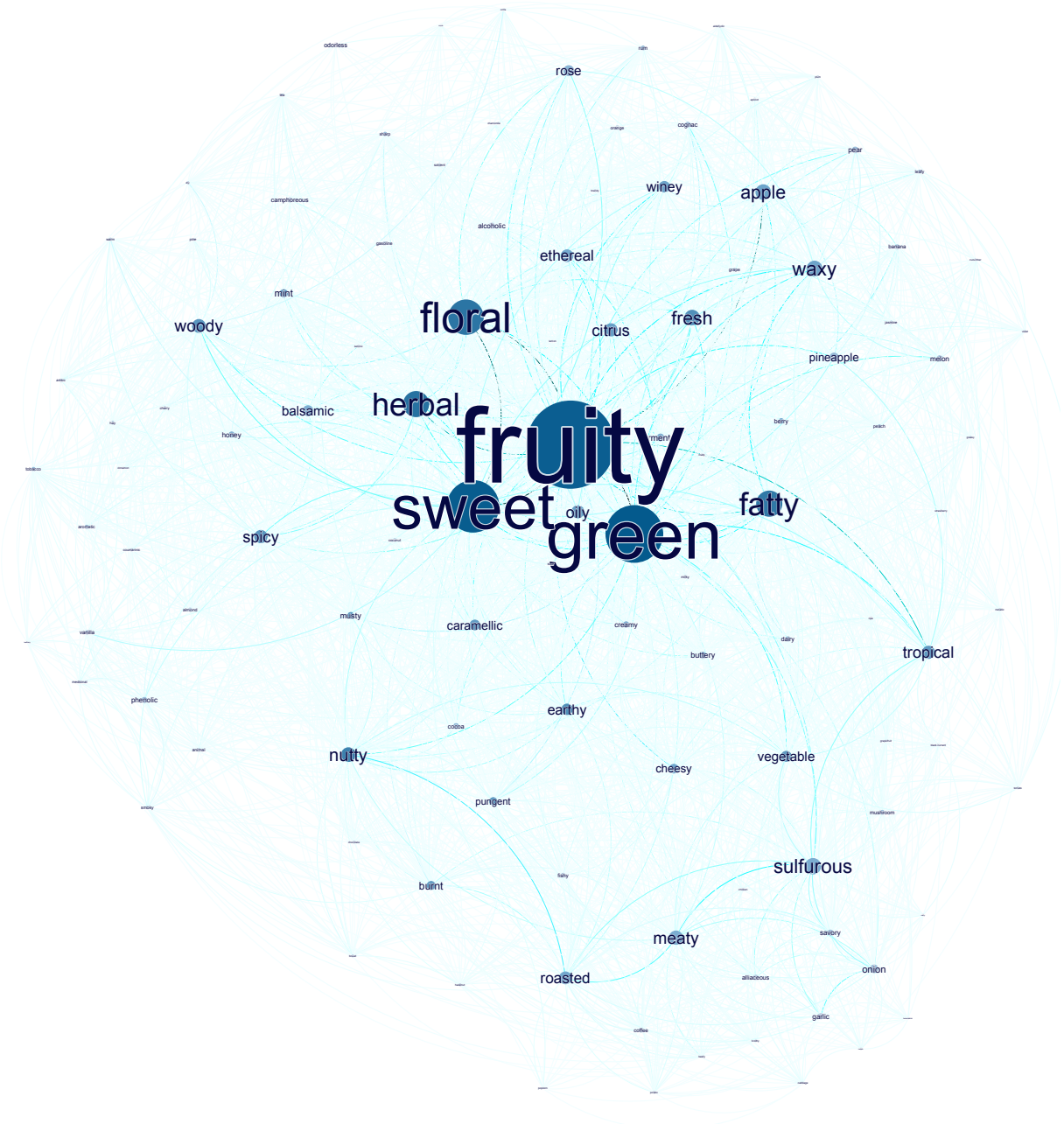


Figure 1: Descriptor Co-Occurrence. Lines connect labels that occur together for the same molecule. The layout minimizes distance between descriptors with strong co-occurrence.

## Methods

Dataset generation, modeling, and packaging were all performed in Python, with the exception of a single line of Bash used for conformer generation The complete workflow is open-source and available on Zenodo[3]. This is an example of the power and flexibility of modern Python for the computational chemistry community. The workflow is demonstrated schematically in Fig. 2.

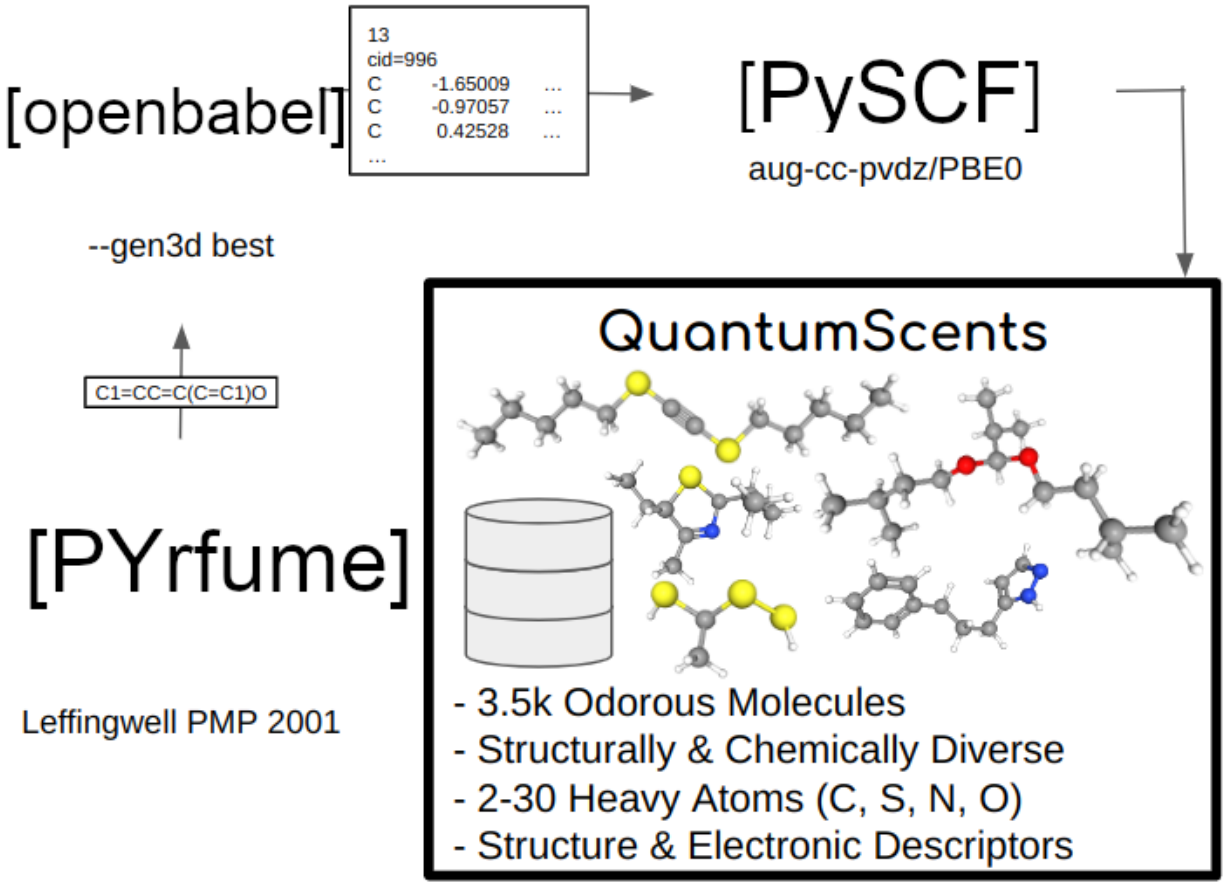


Figure 2: Generation of QuantumScents Dataset. Openbabel conformer generation is followed by energy minimization and property calculation in pyscf.

To begin, the cleaned Leffingwell dataset was retrieved from the `pyrfume`[15] project repository. Each molecule in the dataset was then converted from its SMILES embedding to the corresponding 3D coordinates of a minimum-energy structure using `openbabel`[16]. All

molecules were protonated to charge-neutral states in this step, although approximately 5% of the dataset would likely be charged at physiologically relevant pH levels (7-8) as determined by Dimorphite-DL[17]. 3D conversion leveraged the `--gen3d best` option to generate coordinates, scan torsion rotations, and energy minimize using MMFF94[18]. Torsion rotation scanning was skipped for 17 of the molecules (`--gen3d fast`) and/or pubchem known structures[19] were used because `--gen3d best` failed. Using these structures DFT simulations were performed with PySCF[20–22] using Restricted Kohn-Sham[23] at the PBE0/aug-cc-pvdz level of theory[24–26]. The Hirshfeld extension module[27] was used for calculation of Hirshfeld charges within PySCF. These simulations were parallelized using mpi4py[28,29] and run on the Andes cluster at the Oak Ridge Leadership Computing Facility.

## Data and Software Availability

The QuantumScents dataset contains atomic coordinates, total PBE0 energy, molecular dipole moment, and per-atom Hirshfeld charges, dipoles, and ratios for each of the molecules. Molecules range in size from two to thirty heavy atoms including carbon, sulfur, nitrogen, and oxygen. QuantumScents diverges from the most cited quantum property datasets in this regard. QuantumMachine9 (QM9) and QuantumMachine7-X (QM7-X) are both far larger and explore a complete combinatorial space but are consequently limited in size to nine and seven heavy atoms, respectively[30,31]. Quantum property datasets like AlexandriaLib[32] are more similar to QuantumScents in that they focus on smaller sets of structurally diverse molecules aggregated for their field-specific applications: thermodynamic data aggregation and olfaction research.

The dataset itself is freely available on Zenodo[3]. It includes workflow scripts and example models discussed below. The example models demonstrate how to open the compressed data as a Python dictionary using `hdfdict`[33] and converting to `RDKit`[34] molecules using `xyz2mol`[35].

# Example Model

To demonstrate the loading and usage of the QuantumScents dataset as well as the utility of Hirshfeld charges in performing molecular property assignment, the modeling performed by Sanchez-Lengling and co-workers alongside the cleaning of the Leffingwell Dataset [12] is reproduced and expanded upon. First the dataset is partitioned into training, validation, and test subsets using multiple approaches for comparison. Next a Message Passing Neural Network (MPNN) is trained with `PyTorch`[36] via `chemprop`[2,37]. The partitioning schemes, model architecture, and results are explored below.

## Directed-Message Passing Neural Network

Directed-Message Passing Neural Networks (D-MPNNs) are a class of artificial neural network that operate on graphs to aggregate a learned representation via passing of 'messages' (features) between nodes and edges. This featurization is then used as the input layer to an archetypal Feedforward Neural Network that translates the embeddings into the desired feature space. `chemprop` is a cheminformatics software package that implements the D-MPNN + FNN architecture for molecular property prediction[37] and is used here to re-implement the architecture described in the reference work. For further details on the implementation of the described architecture, see Heid et al.[2].

By default `chemprop` uses a pre-determined set of atomic and bond features to initialize the representation and supports loading molecules from the SMILES format only. To enable loading of structures from coordinates and to change the default feature layout some minor modifications were made to `chemprop` and are included in the `QuantumScents` repository as a submodule.

## Dataset Partitioning

The highly unbalanced nature of the target space in this dataset requires more intentional sampling approaches than simple random partitioning which is likely to leave target labels out of small subsets. Many methods exist that attempt to account for this imbalance, two of which are explored here: Kennard-Stone algorithm based on Mahalanobis distance (MDKS)[38] as implemented in the `astartes` dataset partitioning package[39] and Szymanski splitting[40] as used in the reference work[1].

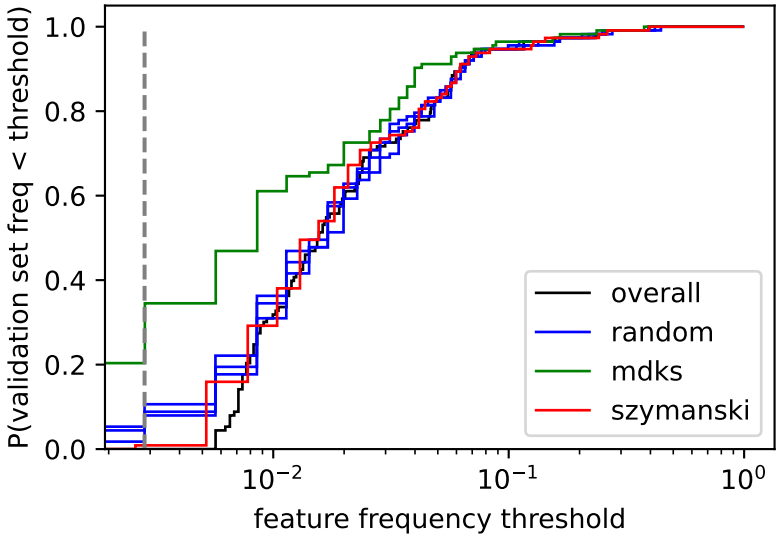


Figure 3: Comparison of label frequency in the validation set generated by various splitting algorithms (cumulative distribution function). Nonzero initial value indicates the percent of labels absent in the validation set. Vertical line marks the frequency corresponding to one occurrence in the validation set (for N=352). Szymanski splitting used N=385 validation samples.

Figure 3 shows another way to visualize the unbalanced target label space. The cumulative distribution function compares the fraction of each scent label in the validation split for random, MDKS, and Szymanski partitioning approaches. The blue (random) splits usually pick molecules that do not exhibit any examples for 2-5% of the known labels – showing the

[1]The exact split assignments were made available in the reference paper but have not been used here because the tool to generate them (`scikit-multilearn`[41]) has been abandoned. Instead the original code by Szymanski was used[40] with minor modifications (included in the Zenodo[3] repository) to enable partitioning into an 80/10/10 training/validation/test split.

difficulty of including all the scent labels in the validation set of 352 molecules. In addition, the "random" line is above the "overall" line until we reach features at the 1% level. This shows that rare features are systematically under-represented in random and MDKS splits.

Random and MDKS sampling result in multiple validation labels infrequently being excluded from the validation data, whereas Szymanski splitting ensures each label is at least represented if not close to its fraction in the overall dataset. This guarantees the model's prediction task is interpolation (the desired application) rather than extrapolation. This increases performance because interpolation is less challenging. Due to the huge deficiency in the MDKS sampling approach, only Szymanski and Random splitting are considered further.

Beyond simply changing the distribution of data in the various sets, the sampling approach also impacts the model performance. Shown in the plots below is the change in Area Under Receiver-Operator Curve (ROC AUC) for the validation set across epochs averaged across ten repetitions of Szymanski and random sampling. ROC AUC ranges from a perfectly predictive at 1 to a completely incorrect at 0. Each model was trained for only 90 epochs using the architecture described in the Results section.

The speed with which the network converges is impacted by sampling as is the ROC AUC. Szymanski sampling achieves higher average performance, less variability, and faster convergence between repetitions than random sampling. Repeat training with different random seeds did not have an impact on model score on the testing set after convergence was reached. After convergence with longer training times the various Szymanski splits achieve similar performance on holdout data. Going forward reported metrics are averaged across three folds the sake of well-defined metrics (see Discussion).

## Results

The architecture and training parameters were first adapted from the reference paper: 400 epochs, 512 batch size, depth 3 message passing, 0.1 dropout, hidden size 300, and a 5 layer FNN. With this configuration, the Szymanksi splits discussed above, and the default features

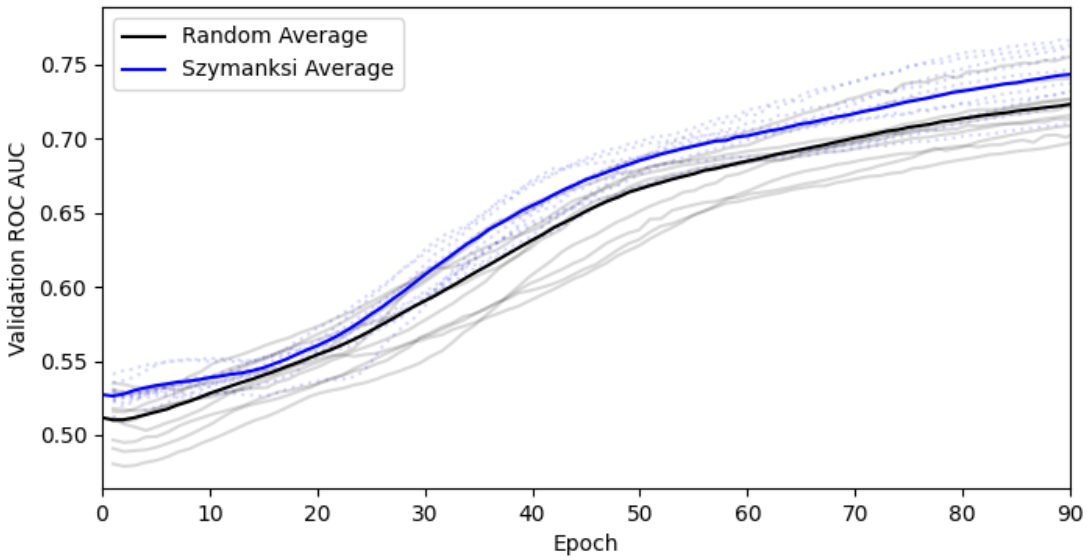


Figure 4: Impact of sampling algorithm on validation score. Validation ROC AUC is plotted over the first 90 epochs of training for ten different realizations of the random (grey lines with average in black) and Szymanski (blue dotted lines with average in solid blue) splitting strategies.

in `chemprop` the testing ROC AUC was 0.845, in line with the validation ROC AUC of 0.853. These values serve as our baseline for comparison to alternative features and architectures.

Next the Hirshfeld charges (provided in QuantumScents) were appended to the default atomic features of `chemprop` leaving the bond features unchanged. This increased the validation and testing ROC AUC slightly to 0.874 and 0.865, respectively.

Subsequently, we modified the atom encoding to eliminate all features except the (one-hot encoded) atomic number, Hirshfeld charge, and Hirshfeld ratio. Going forward, this collection is refered to as the Hirshfeld features. We left bond features at their defaults, a one-hot encoding of bond type (unknown, single, double, triple, aromatic), and 3 boolean indicators (for conjugated, isInRing, and stereospecific). We then repeated the model training. Impressively, the ROC AUC is nearly identical at 0.876 for the validation data and 0.867 for the holdout set. The above results are summarized in the table below.

| Feature Set | Validation ROC AUC | Testing ROC AUC |
|---|---|---|
| `chemprop` Default | 0.845 | 0.853 |
| Default + Hirshfeld | 0.874 | 0.865 |
| Hirshfeld Only | 0.876 | 0.867 |

Also striking is the impact of the Hirshfeld features on the training itself. Shown below is the validation ROC AUC across epochs during training.

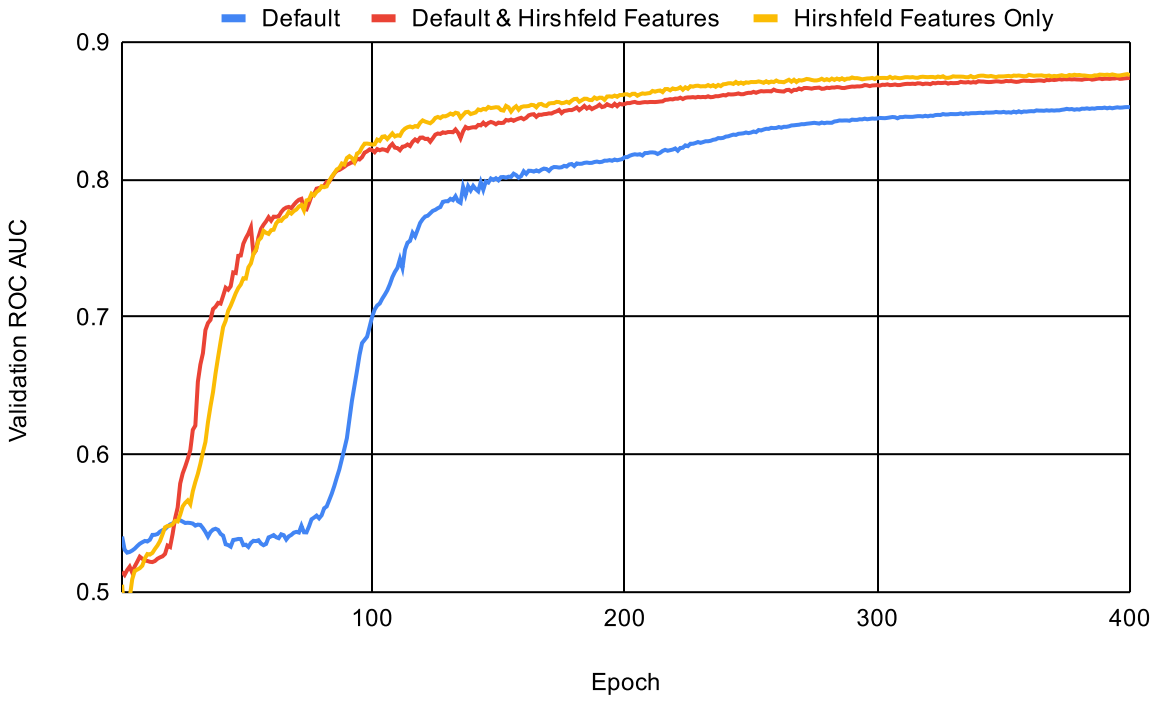


Figure 5: Featurization Impact on Validation ROC AUC. Presence of Hirshfeld charges and volume ratios dramatically improves predictivity compared to the default atomic descriptors.

The network was able to converge far more rapidly with Hirshfeld features regardless of the inclusion of other features. While the time savings at the scale of this dataset size are less pronounced for larger datasets this savings could be substantial. This feature ablation resulted in nearly identical performance indicating the the Hirshfeld features alone are sufficient to inform the network and contain all cogent information to describe the molecule.

Modest hyperparameter optimization yielded negligible improvements in validation or testing ROC AUC, though we did find that the FNN size could be reduced to only 2 layers. Figure 6 shows the per-task performance for this architecture sorted from best to worst.

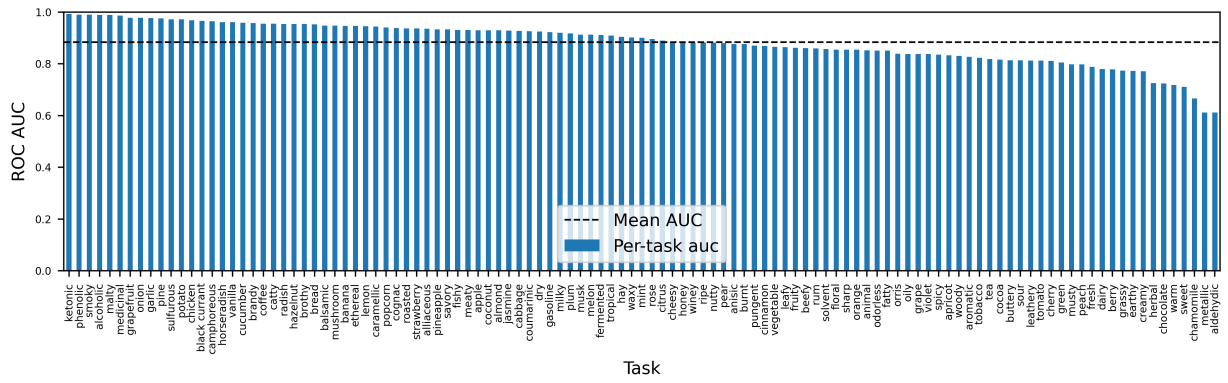


Figure 6: Results: Predictivity for individual scent labels.

The distribution of performance across labels is skewed, but is not correlated to the label frequency in the dataset ($r = 0.023$, $p = 0.81$). Compared to the reference paper, the performance is less consistent across labels. Again, this can be attributed to size and content of the holdout data. The per-task performance when using the default `chemprop` features is provided in the SI.

# Discussion

The prediction score (ROC AUC) of the initial chemprop network (0.845 test, 0.853 validation) does not exactly match the results of the Sanchez-Lengling study (0.894 [12]), which used similar input descriptors. Nevertheless, we view them as comparable given the differences in data partitioning and test/validation set selection methodology between these studies.

Including electronic-structure derived charges and atomic volume ratios resulted in a marked improvement of model performance. This can be seen in Fig. 5 both as a faster increase in validation score with respect to training epochs as well as a higher score at convergence. This improved performance demonstrates that the Hirshfeld charges do contain

cogent information for this classification task.

More critical though is the impact of *replacing* the default features with Hirshfeld charges. Because the gains in predictivity persisted in the absence of the other default atom descriptors (atom hybridization, degree, hydrogen count, chirality, aromaticity, and mass), we can conclude that charge-based input features provide better predictivity than the default atom descriptors present in `chemprop`. This conclusion provides support to the argument that atomic charge and molecular topology are key determinants of odor recognition.

It should be noted that the molecules present in this dataset have varying degrees of structural flexibility. For any given molecule, multiple low-energy structures may be populated in a physical sample. Each of those structures will, in turn, have a different electron density, and correspondingly different atomic partial charges and volumes.

Our dataset includes only the single structure found using OpenBabel's `--gen-3d` option, which translates the SMILES embedding into coordinates (see Methods), followed by energy minimization at the DFT-PBE0 level of theory. This process is *likely* to find the lowest energy, structure in ~94% of cases[18]. Spot checking some conformations and performing further geometry optimization during simulation with PySCF may improve the data. However, a thorough conformational search in the style of QM7-x[31] or AlexandriaLib[32] may be more informative for the biological applications of QuantumScents.

If this model were to be extended to molecules larger than 30 atoms, we expect two new, competing effects to come into play. First, because the MPNN captures local structural information, label prediction will become more dependent on the set of substructures present. Local patterns present in atomic charges and volumes extracted from the electronic structure will likely remain in place, despite changes to conformation and environment. If scent labels are associated with clusters of atoms, this would tend to make prediction more robust. Second, environment-dependence of the whole molecule geometry (including intramolecular hydrogen-bonding or rearrangement due to solvation in aerosols) would create an additional layer of uncertainty in whole-molecule properties such as total charge or solvent accessibility.

This uncertainty would tend to make prediction of macromolecular properties less robust.

The imbalance of labels in this dataset (shown in Fig. 3 presents additional challenges to creating a useful metric for validation. Common classification metrics like recall and precision become ill-defined when positive or negative labels are missing from the validation dataset or model predictions. The "accuracy" metric by itself also unhelpful, since a 'model' that predicts all negative labels has an accuracy of ~95% – due to data imbalance. We chose to use ROC AUC as our measure of predicticity because it is always defined and was used in the reference study[12]. Nevertheless, a metric with more physical intuition (e.g. perceptual similarity[11]) may be more useful for the olfaction applications of these data.

Related to the imbalance of labels in our data, we see a dependence on the details of the method used to create test/train/validation splits. Figure 4 shows that Szymanski splitting gives better validation results than Random sampling. Comparing to Fig. 3, we can see the importance of the Szymanski strategy, ensuring each label appears with a frequency close to its fraction in the dataset. We conclude that when there is imbalance in the occurrence of labels, careful attention to splitting methods should be used to avoid random variation in measures of fitness.

# Conclusions

The QuantumScents dataset is a quantum-mechanics augmented version of the established Leffingwell PMP 2001[1] dataset providing 3D coordinates for an energy minimized structure of each molecule and relevant quantum-mechanics properties. Using only the atomic numbers, bond types, and Hirshfeld charges and ratios an MPNN was able to predict the scent labels with a ROC AUC on par with literature precedent. The dataset is unique for its significant structural and chemical diversity with samples ranging from two to thirty heavy atoms and multiple sulfur-containing functional groups.

For the computational chemistry community, QuantumScents demonstrates the power

and flexibility of modern Python to facilitate research and for olfaction research introduces chemical and electronic structure data which can form the basis for further quantitative investigations. Future work should look to explore the relationship of conformation to the capacity to predict scent and potentially development of a novel metric that is more useful for such highly unbalanced label spaces.

Supporting Information. A plot of the per-task prediction performance when using the default `chemprop` features is provided in the SI. This material is available free of charge via the Internet at https://pubs.acs.org.

## Acknowledgement

The authors gratefully acknowledge financial support from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0023112. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

**Disclaimer**

reflect those of the United States Government or any agency thereof.

# References


(1) Leffingwell, J. C. Leffingwell and Associates. 2005; https://github.com/pyrfume/pyrfume-data/blob/main/leffingwell/leffingwell_readme.pdf.

(2) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: Machine Learning Package for Chemical Property Prediction. *ChemRxiv* **2023**,

(3) Burns, J.; Rogers, D. QuantumScents: Quantum-Mechanical Properties for 3.5k Olfactory Molecules. 2023; https://doi.org/10.5281/zenodo.8239854.

(4) Hoehn, R. D.; Nichols, D. E.; Neven, H.; Kais, S. Status of the Vibrational Theory of Olfaction. *Frontiers in Physics* **2018**, *6*.

(5) Keller, A.; Gerkin, R. C.; Guan, Y.; Dhurandhar, A.; Turu, G.; Szalai, B.; Mainland, J. D.; Ihara, Y.; Yu, C. W.; Wolfinger, R.; Vens, C.; Schietgat, L.; De Grave, K.; Norel, R.; DREAM Olfaction Prediction Consortium,; Stolovitzky, G.; Cecchi, G. A.; Vosshall, L. B.; Meyer, P. Predicting human olfactory perception from chemical features of odor molecules. *Science* **2017**, *355*, 820–826.

(6) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.

(7) Fujita, T.; Winkler, D. A. Understanding the Roles of the "Two QSARs". *J. Chem. Inf. Model.* **2016**, *56*, 269–274.

(8) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta* **1977**, *44*, 129–138.

(9) Buckingham, A. D.; Fowler, P. W. A model for the geometries of Van der Waals complexes. *Canadian J. Chem.* **1985**, *63*, 2018–2025.

(10) Gregory, K. P.; Wanless, E. J.; Webber, G. B.; Craig, V. S. J.; Page, A. J. The electrostatic origins of specific ion effects: quantifying the Hofmeister series for anions. *Chem. Sci.* **2021**, *12*, 15007–15015.

(11) Lee, B. K.; Mayhew, E. J.; Sanchez-Lengeling, B.; Wei, J. N.; Qian, W. W.; Little, K. A.; Andres, M.; Nguyen, B. B.; Moloy, T.; Yasonik, J.; Parker, J. K.; Gerkin, R. C.; Mainland, J. D.; Wiltschko, A. B. A principal odor map unifies diverse tasks in olfactory perception. *Science* **2023**, *381*, 999–1006.

(12) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. 2019.

(13) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(14) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.

(15) Castro, J. B.; Gould, T. J.; Pellegrino, R.; Liang, Z.; Coleman, L. A.; Patel, F.; Wallace, D. S.; Bhatnagar, T.; Mainland, J. D.; Gerkin, R. C. Pyrfume: A Window to the World's Olfactory Data. *bioRxiv* **2022**,

(16) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

(17) Ropp, P. J.; Kaminsky, J. C.; Yablonski, S.; Durrant, J. D. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics* **2019**, *11*, 14.

(18) Yoshikawa, N.; Hutchison, G. R. Fast, efficient fragment-based coordinate generation for Open Babel. *Journal of Cheminformatics* **2019**, *11*, 49.

(19) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Research* **2022**, *51*, D1373–D1380.

(20) Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K.-L. PySCF: the Python-based simulations of chemistry framework. *WIREs Computational Molecular Science* **2018**, *8*, e1340.

(21) Sun, Q.; Zhang, X.; Banerjee, S.; Bao, P.; Barbry, M.; Blunt, N. S.; Bogdanov, N. A.; Booth, G. H.; Chen, J.; Cui, Z.-H.; Eriksen, J. J.; Gao, Y.; Guo, S.; Hermann, J.; Hermes, M. R.; Koh, K.; Koval, P.; Lehtola, S.; Li, Z.; Liu, J.; Mardirossian, N.; McClain, J. D.; Motta, M.; Mussard, B.; Pham, H. Q.; Pulkin, A.; Purwanto, W.; Robinson, P. J.; Ronca, E.; Sayfutyarova, E. R.; Scheurer, M.; Schurkus, H. F.; Smith, J. E. T.; Sun, C.; Sun, S.-N.; Upadhyay, S.; Wagner, L. K.; Wang, X.; White, A.; Whitfield, J. D.; Williamson, M. J.; Wouters, S.; Yang, J.; Yu, J. M.; Zhu, T.; Berkelbach, T. C.; Sharma, S.; Sokolov, A. Y.; Chan, G. K.-L. Recent developments in the PySCF program package. *The Journal of Chemical Physics* **2020**, *153*, 024109.

(22) Sun, Q. Libcint: An efficient general integral library for Gaussian basis functions. *Journal of Computational Chemistry* **2015**, *36*, 1664–1671.

(23) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(24) Kendall, R. A.; Dunning, J., Thom H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of Chemical Physics* **1992**, *96*, 6796–6806.

(25) Woon, D. E.; Dunning, J., Thom H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *The Journal of Chemical Physics* **1993**, *98*, 1358–1371.

(26) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **1999**, *110*, 6158–6170.

(27) Rogers, D. M. hirshfeld. 2023; https://github.com/frobnitzem/hirshfeld.

(28) Dalcin, L.; Fang, Y.-L. L. mpi4py: Status Update After 12 Years of Development. *Computing in Science & Engineering* **2021**, *23*, 47–54.

(29) Rogowski, M.; Aseeri, S.; Keyes, D.; Dalcin, L. mpi4py.futures: MPI-Based Asynchronous Task Execution for Python. *IEEE Transactions on Parallel and Distributed Systems* **2023**, *34*, 611–622.

(30) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, 140022.

(31) Hoja, J.; Medrano Sandonas, L.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio Jr., R. A.; Tkatchenko, A. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific Data* **2021**, *8*, 43.

(32) Ghahremanpour, M. M.; van Maaren, P. J.; van der Spoel, D. The Alexandria library, a quantum-chemical database of molecular properties for force field development. *Scientific Data* **2018**, *5*, 180062.

(33) Gündert, S. hdfdict. 2023; https://github.com/SiggiGue/hdfdict.

(34) Landrum, G. RDKit: Open-Source cheminformatics. 2023; https://www.rdkit.org/.

(35) Kim, Y.; Kim, W. Y. Universal Structure Conversion Method for Organic Molecules: From Atomic Connectivity to Three-Dimensional Geometry. *Bulletin of the Korean Chemical Society* **2015**, *36*, 1769–1777.

(36) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32. 2019; pp 8024–8035.

(37) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388, PMID: 31361484.

(38) Saptoro, A.; Tadé, M. O.; Vuthaluru, H. A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models. *Chemical Product and Process Modeling* **2012**, *7*.

(39) Burns, J.; Spiekermann, K.; Bhattacharjee, H.; Vlachos, D.; Green, W. Machine Learning Validation via Rational Dataset Sampling with astartes. 2023; https://doi.org/10.5281/zenodo.8147205.

(40) Szymański, P.; Kajdanowicz, T. A Network Perspective on Stratification of Multi-Label Data. Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications. ECML-PKDD, Skopje, Macedonia, 2017; pp 22–35.

(41) Szymański, P.; Kajdanowicz, T. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints* **2017**,