Flavor Analysis and Recognition Transformer Model

Franz Goerlich†, Philipp Pestlin†, Henrik Seng†, Leif Sieben† and Yoel Zimmermann†
Digital Chemistry Lecture, Department of Chemistry and Applied Biosciences, ETH Zurich.

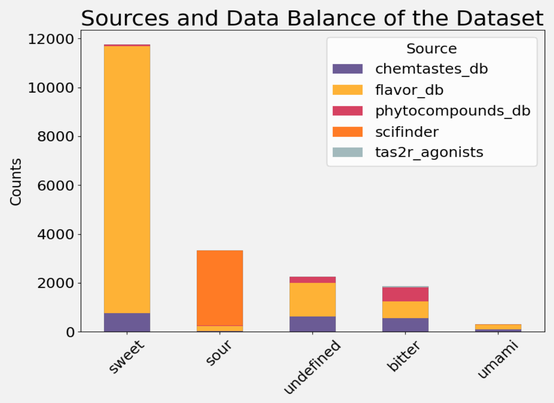
†Equal contribution, names are given in alphabetical order.

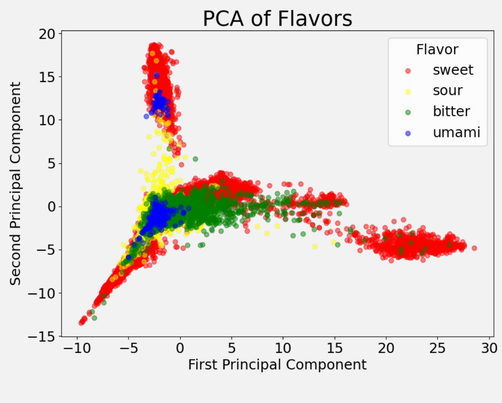
1 Introduction

Determining the taste of molecules is a labor and time intensive process in food chemistry. Various machine learning approaches have been used for taste classification [1]. However, despite their ubiquity in sequential learning tasks, transformer models have not been tested for taste classification so far.

2 Dataset

In total, a dataset of 19'478 molecules was curated from five publicly available datasets [2]:

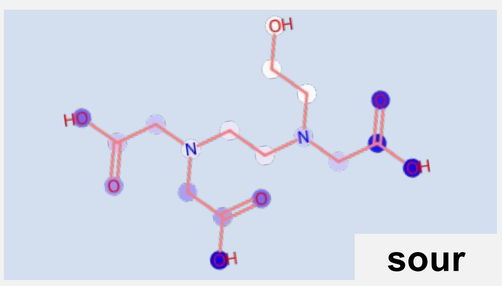


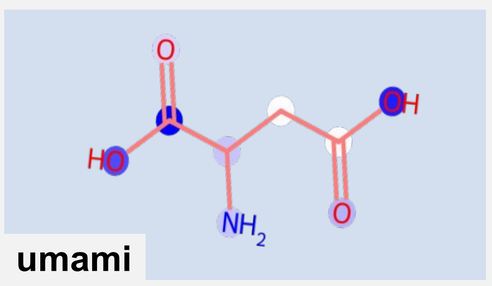


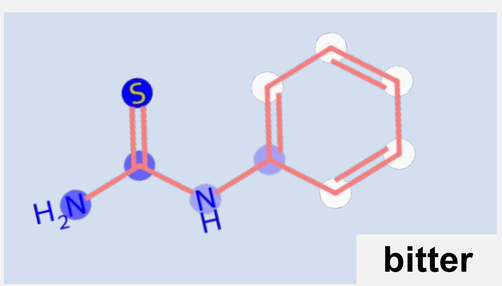
The database contains no duplicate values but when a molecule is associated with two tastes, its SMILES is featured multiple times for each taste. Every entry thus contains one canonicalized SMILES and one taste class. Salty molecules were not considered as there are very few known salty molecules. Molecules that could not be attributed to any of the four categories, mostly molecules with known scent but no taste, were labelled as "undefined".

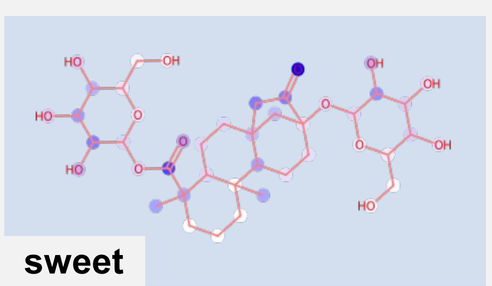
3 Model

The database was used to finetune the large-language model ChemBERTa [3], which is pretrained on 10 million SMILES from PubChem using 6 layers and 12 attention heads. Dark blue areas indicate atoms with high average attention across heads.



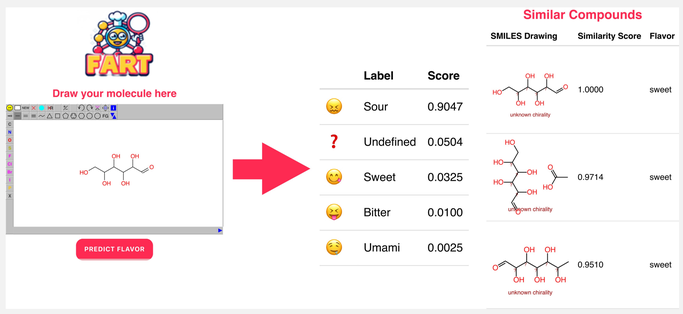






4 Results and Discussion

A given query molecule is checked against the database and the three nearest (Tanimoto distance) molecules are returned. FART outputs the predicted softmax probability for all five categories to classify the query molecule. An interactive web-app with a molecule drawer enables users to easily interact with FART.



Try it yourself:

Approaches	Accuracy	Precision	Recall	F1-Score
FART	**0.8381**	0.8348	**0.8381**	**0.8347**
FART with class weighting	0.8155	**0.8444**	0.8155	0.8248
Random Forest with oversampling	0.8085	0.8316	0.8085	0.8176
Random Forest with weights biases	0.8095	0.8388	0.8095	0.8204

5 Conclusion

- A large dataset of 19'478 molecules, which is close to all publicly known molecule-taste pairs, was collected. This dataset is now accessible online with full FAIR compliance [4].
- Data balance is a major issue as most tested molecules are classified as sweet. Expert generated data was used to collect sour molecules. A loss function penalizing prediction of more common classes was used to reflect data imbalance.
- FART's attention mask matches well with chemical intuition in many cases, e.g. focusing on acid groups for sour compounds.
- FART outperforms two different random forest approaches, which can be considered good baseline models in this case. This suggests that transformers, given sufficient data, can outperform simpler machine learning approaches.
- FART can be easily accessed via an online user interface.

References

1. Y. Song, S. Chang, J. Tian, W. Pan, L. Feng, and H. Ji, “A Comprehensive Comparative Analysis of Deep Learning Based Feature Representations for Molecular Taste Prediction,” Foods, vol. 12, no. 18, p. 3386, (2023), doi: 10.3390/foods12183386.
2. Malavolta, M., Pallante, L., Mavkov, B. et al. "A survey on computational taste predictors". Eur Food Res Technol 248, 2215–2235 (2022), doi: 10.1007/s00217-022-04044-5
3. S. Chithrananda, G. Grand, and B. Ramsundar, “ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction.” arXiv, (2020). Accessed: Feb. 23, 2024. Arxiv.org/abs/2010.09885
4. https://huggingface.co/datasets/FartLabs/FartDB