A Flavor Analysis and Recognition Transformer Model

Technical Report

Franz Görlich
Department of Chemistry and Applied Biosciences
Eidgenössische Technische Hochschule Zürich
fgoerlich@ethz.ch

Philipp Pestlin
Department of Chemistry and Applied Biosciences
Eidgenössische Technische Hochschule Zürich
ppestlin@ethz.ch

Henrik Seng
Department of Chemistry and Applied Biosciences
Eidgenössische Technische Hochschule Zürich
heseng@ethz.ch

Leif Sieben
Department of Chemistry and Applied Biosciences
Eidgenössische Technische Hochschule Zürich
lsieben@ethz.ch

Yoel Zimmermann
Department of Chemistry and Applied Biosciences
Eidgenössische Technische Hochschule Zürich
yzimmermann@ethz.ch

Abstract

- Determining the taste of molecules is a time and labor intensive process based on the analysis of taste receptor interactions or human panelists but could potentially be substituted with machine-learning models.
- A dataset was generated based on several publicly available sources, containing SMILES representations and taste labels for 15'032 molecules from which several pretrained ChemBERTa transformer models were fine-tuned for the downstream task of multi-class taste classification.
- The transformer models allowed predictions with high accuracy but did not outperform random forest model based on molecular fingerprints.
- Predictions of rarer categories, such as umami, remain challenging and future work could focus on expanding the scope of these machine learning models to include larger molecules such as peptides.

1 Introduction

Taste emerges from the interaction of molecules with receptors in the mouth. A molecule will only activate a certain taste receptor if such an interaction is both sterically and chemically allowed. Hence, the taste of a molecule should be fully encoded in its molecular structure. Determining the taste of molecules is an important task in food chemistry but remains both time- and cost-intensive [Reineccius and Peterson, 2013]. The ability to predict the taste of molecules *in silico* would not only meaningfully accelerate this process but may also enable the automated screening of substances on an unprecedented scale in the future.

Unlike odor, taste is biologically well defined through specific receptors and five general taste classes are recognized: sweet, sour, bitter, salty and umami [Chaudhari and Roper, 2010]. There has been substantial work in the past to transform taste prediction into a machine learning classification task [Malavolta et al., 2022, Rojas et al., 2023, Kou et al., 2023, Song et al., 2023]. Despite their ubiquity in virtually all sequential learning approaches, transformer models have not yet been utilized for the specific downstream task of taste classification. The objective of this work was therefore to curate a FAIR-compliant dataset specifically developed for molecular taste analysis to then fine-tune a ChemBERTa model [Chithrananda et al., 2020, Ahmad et al., 2022] as a Flavor Analysis and Recognition Transformer (FART).

2 Methods

Figure 1 illustrates the general methodology utilized for the training of the FART models. The following section contains the detailed procedures for the dataset generation, transformer training, additional training of random forest models and the implementation of an interactive web application. The full code can be accessed via GitLab (`https://www.gitlab.ethz.ch/lsieben/digital_chemistry_2024_fart`) and trained models are available on our Hugging Face organization page (https://huggingface.co/FartLabs).

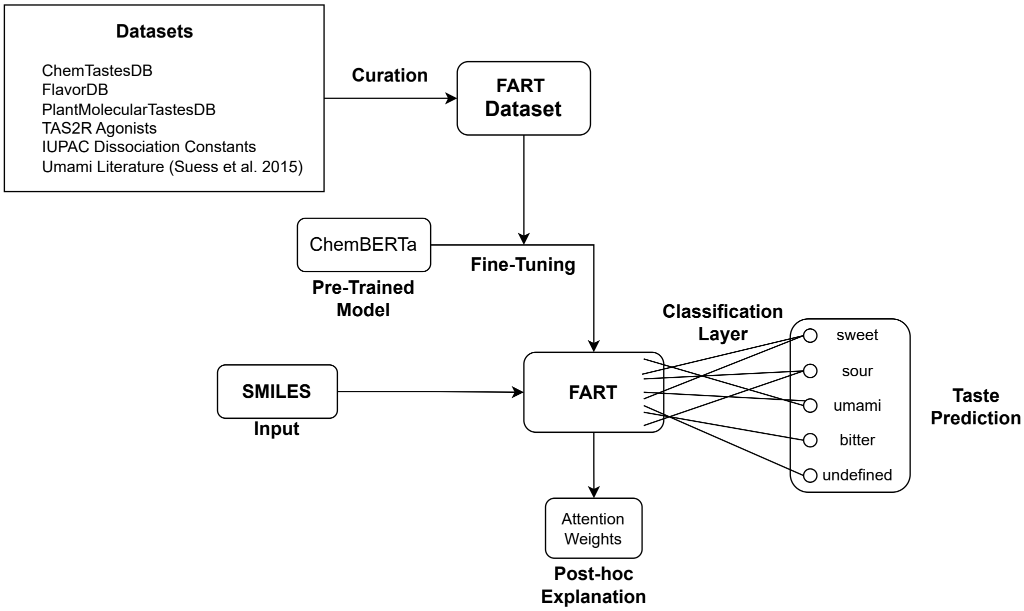


Figure 1: Overview of the methodology used in developing FART.

2.1 Dataset

An important intermediate aim of this work was to curate a large and high quality dataset of molecular tastants. To this end, a large amount of publicly available flavor databases was curated. The dataset utilizes SMILES to represent molecular structure and labels every molecule with one of four taste categories and one category for the remaining molecules: sweet, bitter, sour, umami and undefined. Salty was excluded as a category because only a very small number of molecules actually produce this taste apart from sodium chloride and therefore, it is generally not true that all salts also taste salty [Taruno and Gordon, 2023]. To make this dataset as useful as possible, the data was enriched with information from the PubChem database where available.

2.1.1 Data Extraction

The FART dataset combines data from six publicly available sources. ChemTasteDB is one of the largest public databases of tastants and contains 2'944 organic and inorganic tastants from which 2'177 were used to train FART [Rojas et al., 2022]. The database is based on tastes given in the literature. FlavorDB aggregates data on both gustatory and olfactory sensation from a number of sources [Garg et al., 2017]. The FART dabase uses the "flavor profile" given by FlavorDB as most molecules do not have a specific entry for taste. Data from FlavorDB will thus be of somewhat less quality given that some of these flavor profiles will actually be based on smell, not taste. From the 25'595 total molecules, 10'372 could be clearly attributed to one of the four taste categories. FlavorDB is dominated by sweet molecules and is also the source for the data imbalance in the final dataset. PlantMolecularTasteDB contains 1'527 phytochemicals with associated taste of which 906 were used for this dataset [Gradinaru et al., 2022]. The database is based on both literature and other databases, some of which are also listed in other databases used for FART. To obtain more data on bitter compounds, a database of ligands that bind to the human bitter receptor (TAS2) was also considered which yielded 53 previously unseen bitter compounds [Bayer et al., 2021].

The database was further extended with molecules with known pK_A between 2 and 7. Although sour taste is influenced by other factors such as cell permeability, which is the reason why organic acids taste more acidic than inorganic acids such as HCl at the same pH, acidic molecules can be assumed to also taste sour [Roper, 2017]. Acids were collected from an ongoing project based with the International Union of Pure and Applied Chemistry (IUPAC) which digitized three high-quality sources of pK_A values in the literature [Perrin, 1965, 1972, Serjeant and Dempsey, 1979]. A total of 1'513 acids could be obtained in this way although it should be noted that sour taste, as all tastes, is concentration dependent and that some of the weaker acids may not be picked up by humans. The pK_A values should all be measured between 15 and 30 °C in water, excluding any acids which are not water-soluble, and refer to the most acidic proton. Lastly, 19 umami-tasting molecules were collected from the literature [Suess et al., 2015] of which 11 were not given in any other database.

2.1.2 Data Curation

The combined dataset was reduced to the taste label associated with a canonicalized SMILES representation. RDKit [Landrum et al., 2020] was utilized to further curate the dataset. First, all SMILES which did not allow the generation of a valid molecular graph were excluded. To avoid solvent-containing molecules, all entries with multiple uncharged fragments were removed. Charged molecules were additionally excluded to prevent substances with missing counter ions. All SMILES were standardized with the default RDKit standardization procedure. Duplicate could be removed with the help of these standardized SMILES.

While only very few entries with invalid SMILES (21) or charged molecules (342) need to be removed, the amount of entries containing multiple neutral fragments (3783) was more significant. The duplicate removal (14685) reduced the dataset almost by half to a final size of 15'032 entries, see Figure A.1. The large number of duplicates underlines the significant overlap among the databases used. Note that when duplicate entries existed from different sources, which source would be given in the final dataset was arbitrarily determined based on the index. The curated dataset was further enriched by general information (PubChemID, IUPAC name, molecular formula, molecular weight, InChI, InChIKey), accessed through the PubChem API [Kim et al., 2022].

Table 1 summarizes the origin and the taste labels of the entries of the curated FART dataset. The strong data imbalance, where sweet represents over 60% and umami less than 1% of the data, necessitates additional steps during model training to ensure predictions are fair across groups and do not simply greedily focus on the majority class.

Table 1: Overview of the data sources used for FART.

Database	Sweet	Bitter	Sour	Umami	Undefined	Total
ChemTastesDB	787	921	17	47	405	2177
FlavorDB	8665	71	35	0	1601	10372
PlantMolecularTasteDB	91	631	40	0	144	906
TAS2R Agonists	0	53	0	0	0	53
IUPAC Dissociation Constants	0	0	1513	0	0	1513
Suess et al. 2015	0	0	0	11	0	11
Total	9543	1676	1605	58	2150	15032

The dataset, FartDB, was published on Huggingface (`https://huggingface.co/datasets/FartLabs/FartDB`) in agreement with the FAIR principles [Wilkinson et al., 2016]. The data can be accessed through several different interfaces, allowing the utilization in other project

2.2 Transformer Training

In machine learning, particularly in natural language processing and related fields, a common workflow involves pretraining a model on a large, general dataset followed by fine-tuning on a smaller, task-specific dataset. This approach leverages the model's ability to learn general patterns and representations from vast amounts of data during pretraining in an unsupervised fashion, which can then be specialized to specific tasks through fine-tuning.

Transformer models, a type of feed-forward neural network with an attention mechanism, exemplify this approach. They can process sequential data and produce various outputs depending on the application. ChemBERTa models, derived from the RoBERTa architecture [Liu et al., 2019], are a class of transformer models tailored for chemical data.

They are pretrained on the SMILES representation of molecules, using a model-specific tokenizer to handle the input format.

During the pretraining phase, ChemBERTa models are exposed to extensive SMILES data, allowing them to learn general molecular features and patterns. In the subsequent fine-tuning phase, additional layers are added to the pretrained model. This enhanced model is then trained on a smaller, task-specific dataset, optimizing it for particular applications, such as predicting molecular properties or classifications. This workflow ensures that the model benefits from both broad, general knowledge and specialized, task-specific insights.

In this work, multiple transformer models were trained using the dataset described in section 2.1 based on pretrained ChemBERTa models with an additional classification layer with 5 output neurons. Training was performed over five epochs and validated every 500 training steps where the current state of the model was cached. After training, the model state with the best validation loss was selected. Besides training on the dataset using a cross entropy loss, training was also performed using a class-weighted cross entropy loss as well as using an augmented dataset to account for under and over represented classes.

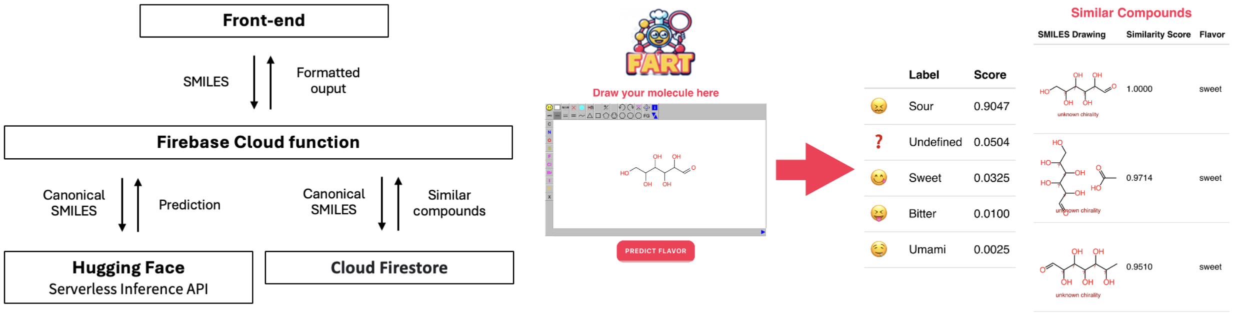
2.3 Random Forest Models for Baseline Comparison

Random forest (RF) models are often considered good baseline models in classification tasks for more complex model architectures, such as transformers, given their robust performance, efficient training and low model complexity. A RF model is an ensemble learning method that trains multiple, in this case 150, decision trees during the training phase, combining their predictions to improve accuracy and reduce overfitting. Gradient boosting methods, such as XGBoost [Chen and Guestrin, 2016], are typically used to improve performance.

In this work, the same RF classifier was trained under three different conditions: (i.) standard RF model, (ii.) Synthetic Minority Over-sampling Technique (SMOTE), (iii.) weighted sampling across classes. The latter two approaches are meant to adjust for data imbalance, i.e. the dataset being dominated by sweet molecules. The RF models were evaluated with a multiclass logarithmic loss function, results are given in Table 2. For more details on the training process see Appendix B.

2.4 Web Application Development

A web application was built to allow for quick testing. Firebase was chosen as the hosting and database infrastructure. The front-end was built with NextJS, the back-end using Flask (Python). An overview of the back and front-end can be seen in Figure 2.



Label	Score
Sour	0.9047
Undefined	0.0504
Sweet	0.0325
Bitter	0.0100
Umami	0.0025

Similar Compounds

SMILES Drawing	Similarity Score	Flavor
unknown chirality	1.0000	sweet
unknown chirality	0.9714	sweet
unknown chirality	0.9510	sweet

Figure 2: Overview of the web application: data handling (left) and front-end (right)

The front-end features the JSME molecule editor, which allows users to easily draw structures of interest [Bienfait and Ertl, 2013]. The SMILES encoding is then canonicalized and the database of known flavors is searched, returning the three most similar compounds by RDKit Fingerprint similarity. At the same time, an API call to the Hugging Face Serverless Infererence API is made, returning the prediction from the FART model. The data is finally formatted and displayed in two tables [Hugging Face].

3 Results and Discussion

Transformer models have a number of disadvantages compared to RF models in terms of computational complexity, robustness and interpretability. FART however is explicitly not a black-box model, given its attention weights can be used to determine which molecular features were particularly important during taste classification. For a RF model, a similar level of interpretability would be achieved by extracting its feature importance scores. A disadvantage in terms of computational resources remains as the fine-tuning of the transformers on two NVIDIA T4 GPUs takes about 1 hour compared to around 10 minutes for a RF model.

Table 2: Performance overview between the trained transformers and RF models. Scores are given as averages across tastes either weighted or unweighted by the size of the taste category in the test set. Scores for RF models were obtained through five-fold cross-validation.

		Weighted average			Unweighted average		
Model	Accuracy	Precision	Recall	F1 Score	Precision	Recall	F1 Score
RF, no adjustments	**0.86**	0.86	**0.86**	**0.86**	0.69	0.78	0.72
RF, weighted sampling	0.85	**0.87**	0.85	**0.86**	0.68	0.76	0.71
RF, SMOTE	0.84	0.86	0.84	0.85	0.65	0.72	0.67
ChemBERTa 1[1], not weighted	**0.86**	0.86	**0.86**	**0.86**	0.72	0.72	0.72
ChemBERTa 1, weighted	**0.86**	**0.87**	**0.86**	**0.86**	**0.83**	0.77	**0.79**
ChemBERTa 2[2], augmented	0.79	0.72	0.80	0.80	0.72	**0.80**	**0.79**

Unless transformers can offer significantly better performance, RF models generally remain a better choice for machine learning classification tasks. For the dataset used here, the two transformers (the smaller ChemBERTa 2 and the larger ChemBERTa 1 version) performed similar to the RF models, see Table 2. The transformer model with weighted cross-entropy loss narrowly beating out all other models. The performance of the RF models is similar to other models from the literature such as BitterSweet if class-specific scores are compared [Banerjee and Preissner, 2018]. These results are also consistent with other model architectures such as graph or convolutional neural networks [Song et al., 2023].

Unlike previous work however, the scope of the dataset prepared allowed FART and the RF classifiers to learn taste prediction across four of the five human taste categories as well categorizing undefined tastes. The undefined category can be seen as a proxy for model uncertainty. The transformer models generally performed very similar to the RF models but using a weighted loss function to penalize greedy prediction of the majority class helped in improving predictions for all taste categories rather than just the majority class (sweet) as can be seen in the unweighted averages. Using an augmented dataset (see Appendix D for more details) did not improve overall performance but provided more comparable performance across categories which again is best reflected in the unweighted averages.

Given the dataset of 15’032 molecule-taste pairs, it is clear that ML based classification of taste from molecular structure alone can be achieved with high accuracy and precision. While FART does not outperform RF models trained on the same dataset in terms of overall accuracy, it does achieve reliable taste prediction across all five categories with results being interpretable both through the attention weights and through comparison with the training data, which is displayed along the taste prediction on the FART web application. In conclusion, FART offers a reliable, easily accessible and transparent way to use transformer models for taste prediction for four of the five human taste categories.

4 Outlook

The reliable prediction of taste was achieved with different machine-learning approaches based on SMILES representations of the molecules. The models presented here are competitive with other state-of-the-art taste-prediction models while offering a broader scope of predictions including four of the five human tastes. These models also include a class for undefined molecules, which should enable them to generalize better to unknown chemical spaces compared to models which are trained on a known subspace with only labeled data points, e.g. sweet and bitter molecules. For example, some of the transformers trained in this work predict table salt as "undefined" rather than erroneously forcing it into one of the four other categories. Furthermore, these taste predictions by FART can be made interpretable by analyzing attention weights.

[1] HF ref.: seyonec/SMILES_tokenized_PubChem_shard00_160k

[2] HF ref.: DeepChem/ChemBERTa-77M-MLM

The FAIR compliant dataset created in the course of this project can also serve as a useful starting point for future ML models that wish to train on publicly available data on molecular tastants. Major challenges remain in the prediction of rare tastants such as umami and in particular salty. While ChemBERTa is limited by its token window, future ML models may also allow peptides or other large molecules as input.

FART would likely be most useful in the discovery of new structural classes of tastants where a large number of molecular scaffolds could be screened for a particular taste. ML tools have been used to identify new scaffolds for artificial sweeteners for example Bouysset et al. [2020]. Such an approach could be combined in a broader pipeline screening for other desirable properties such as synthesizability, adsorption, toxicity etc. to discover novel scaffolds and open up new chemical space for food chemistry.

References

G. Reineccius and D. Peterson. Principles of food flavor analysis. In David Kilcast, editor, *Instrumental Assessment of Food Sensory Quality*, Woodhead Publishing Series in Food Science, Technology and Nutrition, pages 53–102. Woodhead Publishing, 2013. ISBN 978-0-85709-439-1. doi:10.1533/9780857098856.1.53.

Nirupa Chaudhari and Stephen D. Roper. The cell biology of taste. *Journal of Cell Biology*, 190(3):285–296, 08 2010. ISSN 0021-9525. doi:10.1083/jcb.201003144.

Marta Malavolta, Lorenzo Pallante, Bojan Mavkov, Filip Stojceski, Gianvito Grasso, Aigli Korfiati, Seferina Mavroudi, Athanasios Kalogeras, Christos Alexakos, Vanessa Martos, Daria Amoroso, Giacomo Di Benedetto, Dario Piga, Konstantinos Theofilatos, and Marco Agostino Deriu. A survey on computational taste predictors. *European Food Research and Technology*, 248(9):2215–2235, 2022. ISSN 1438-2385. doi:10.1007/s00217-022-04044-5.

Cristian Rojas, Davide Ballabio, Viviana Consonni, Diego Suárez-Estrella, and Roberto Todeschini. Classification-based machine learning approaches to predict the taste of molecules: A review. *Food Research International*, 171:113036, 2023. ISSN 0963-9969. doi:10.1016/j.foodres.2023.113036.

Xingran Kou, Peiqin Shi, Chukun Gao, Peihua Ma, Huadong Xing, Qinfei Ke, and Dachuan Zhang. Data-driven elucidation of flavor chemistry. *Journal of Agricultural and Food Chemistry*, 71(18):6789–6802, 2023. doi:10.1021/acs.jafc.3c00909.

Yu Song, Sihao Chang, Jing Tian, Weihua Pan, Lu Feng, and Hongchao Ji. A comprehensive comparative analysis of deep learning based feature representations for molecular taste prediction. *Foods*, 12(18), 2023. ISSN 2304-8158. doi:10.3390/foods12183386.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv Preprint*, 2020. doi:10.48550/arXiv.2010.09885.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv Preprint*, 2022. doi:10.48550/arXiv.2209.01712.

Akiyuki Taruno and Michael D. Gordon. Molecular and cellular mechanisms of salt taste. *Annual Review of Physiology*, 85(Volume 85, 2023):25–45, 2023. ISSN 1545-1585. doi:10.1146/annurev-physiol-031522-075853.

Cristian Rojas, Davide Ballabio, Karen Pacheco Sarmiento, Elisa Pacheco Jaramillo, Mateo Mendoza, and Fernando García. Chemtastesdb: A curated database of molecular tastants. *Food Chemistry: Molecular Sciences*, 4:100090, 2022. ISSN 2666-5662. doi:10.1016/j.fochms.2022.100090.

Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, Rakhi NK, Shubham Dokania, Arvind Iyer, Ayushi Gupta, Shubhra Agrawal, Navjot Singh, Shubham Shukla, Kriti Kathuria, Rahul Badhwar, Rakesh Kanji, Anupam Jain, Avneet Kaur, Rashmi Nagpal, and Ganesh Bagler. FlavorDB: a database of flavor molecules. *Nucleic Acids Research*, 46(D1):D1210–D1216, 10 2017. ISSN 0305-1048. doi:10.1093/nar/gkx957.

Teodora-Cristiana Gradinaru, Madalina Petran, Dorin Dragos, and Marilena Gilca. Plantmoleculartastedb: A database of taste active phytochemicals. *Frontiers in Pharmacology*, 12, 2022. ISSN 1663-9812. doi:10.3389/fphar.2021.751712.

Sebastian Bayer, Ariane Isabell Mayer, Gigliola Borgonovo, Gabriella Morini, Antonella Di Pizio, and Angela Bassoli. Chemoinformatics view on bitter taste receptor agonists in food. *Journal of Agricultural and Food Chemistry*, 69 (46):13916–13924, 2021.

Stephen D. Roper. Taste: Mammalian taste bud physiology. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier, 2017. ISBN 978-0-12-809324-5. doi:10.1016/B978-0-12-809324-5.02908-4.

Douglas Dalzell Perrin. *Dissociation Constants of Organic Bases in Aqueous Solution, Supplement*. IUPAC, Butterworths, 1965.

Douglas Dalzell Perrin. *Dissociation Constants of Organic Bases in Aqueous Solution*. IUPAC, Butterworths, 1972.

E. P. Serjeant and Boyd Dempsey. *Ionisation Constants of Organic Acids in Aqueous Solution*. Oxford IUPAC Chemical Data Series, Oxford/Pergamon, 1979.

B. Suess, D. Festring, and T. Hofmann. 15 - umami compounds and taste enhancers. In J.K. Parker, J.S. Elmore, and L. Methven, editors, *Flavour Development, Analysis and Perception in Food and Beverages*, Woodhead Publishing Series in Food Science, Technology and Nutrition, pages 331–351. Woodhead Publishing, 2015. ISBN 978-1-78242-103-0. doi:10.1016/B978-1-78242-103-0.00015-1.

Greg Landrum, Paolo Tosco, Brian Kelley, sriniker, gedeck, NadineSchneider, Riccardo Vianello, Ric, Andrew Dalke, Brian Cole, AlexanderSavelyev, Matt Swain, Samo Turk, Dan N, Alain Vaucher, Eisuke Kawashima, Maciej Wójcikowski, Daniel Probst, guillaume godin, David Cosgrove, Axel Pahl, JP, Francois Berenger, strets123, JLVarjo, Noel O'Boyle, Patrick Fuller, Jan Holst Jensen, Gianluca Sforna, and DoliathGavid. rdkit/rdkit: 2020_03_1 (q1 2020) release, March 2020. URL `https://doi.org/10.5281/zenodo.3732262`.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022. ISSN 0305-1048. doi:10.1093/nar/gkac956.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 2016. ISSN 2052-4463. doi:10.1038/sdata.2016.18. URL `http://dx.doi.org/10.1038/sdata.2016.18`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL `https://arxiv.org/abs/1907.11692`.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, 8 2016. doi:10.1145/2939672.2939785. URL `http://dx.doi.org/10.1145/2939672.2939785`.

B. Bienfait and P. Ertl. JSME: a free molecule editor in JavaScript. *Journal of Cheminformatics*, 5:24, 2013.

Hugging Face. Hugging face inference api documentation. `https://huggingface.co/docs/api-inference/index`. Accessed: 2024-06-24.

Priyanka Banerjee and Robert Preissner. Bittersweetforest: A random forest based binary classifier to predict bitterness and sweetness of chemical compounds. *Frontiers in Chemistry*, 6, 2018. ISSN 2296-2646. doi:10.3389/fchem.2018.00093.

Cédric Bouysset, Christine Belloir, Serge Antonczak, Loïc Briand, and Sébastien Fiorucci. Novel scaffold of natural compound eliciting sweet taste revealed by machine learning. *Food Chemistry*, 324:126864, 2020. ISSN 0308-8146. doi:https://doi.org/10.1016/j.foodchem.2020.126864. URL `https://www.sciencedirect.com/science/article/pii/S0308814620307263`.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL `http://jmlr.org/papers/v18/16-365.html`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011. URL `http://arxiv.org/abs/1106.1813`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized smiles strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11:1–13, 2019. doi:10.1186/s13321-019-0393-0.

Appendix

A Dataset

Figure A.1 shows the effect on the size of the dataset for each step of the curation pipeline, as outlined in section 2.1. Removing duplicates roughly halved the dataset size underlining how similar these dataset are to each other.

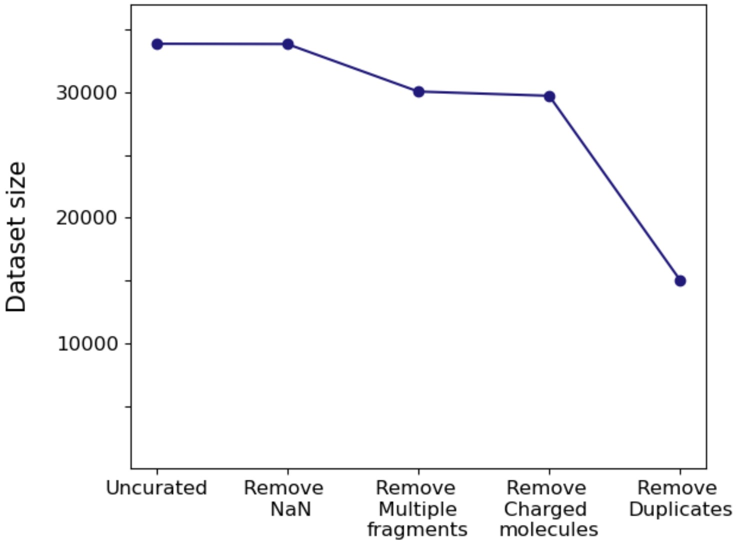


Figure A.1: Dataset size decrease during data curation steps.

B Random Forest Models

All transformer models were trained on 2 NVIDIA T4 GPUs in Google Cloud using the SMOTE algorithm as implemented in the imblearn [Lemaître et al., 2017] and the weighted sampling as implemented in the scikit-learn [Pedregosa et al., 2011] libraries. Both SMOTE and weighted sampling are non-parametric methods but whereas the latter only applies higher weight to real data points, SMOTE actually produces synthetic data points by interpolating across the five nearest neighbors for each minority datapoint Bowyer et al. [2011].

RF models were based on the XGBoost method [Chen and Guestrin, 2016] and RDKit [Landrum et al., 2020] was used to transform SMILES into fingerprints. Input features consisted of a list of Morgan fingerprints generated from the standardized SMILES obtained during data curation. Taste classes were encoded as integers. The classifier was trained with a subsampling rate of 80% and a learning rate of 0.01. A multiclass softmax probability objective function was used and the models were evaluted based on the multiclass logarithmic loss.

Table 3: Classification report for the RF model without any adjustments for data imbalance.

Class	Precision	Recall	F1-score	Support
Bitter	0.59	0.81	0.68	1216
Sour	0.89	0.82	0.85	1740
Sweet	0.93	0.93	0.93	9513
Umami	0.33	0.73	0.45	26
Undefined	0.75	0.63	0.68	2537
Accuracy			0.86	15032
Macro avg	0.69	0.78	0.72	15032
Weighted avg	0.86	0.86	0.86	15032

Table 4: Classification report for the RF model with weighted classes.

Class	Precision	Recall	F1-score	Support
Bitter	0.71	0.67	0.69	1676
Sour	0.79	0.90	0.85	1605
Sweet	0.97	0.88	0.93	9543
Umami	0.32	0.55	0.40	58
Undefined	0.61	0.80	0.69	2150
Accuracy			0.85	15032
Macro avg	0.68	0.76	0.71	15032
Weighted avg	0.87	0.85	0.86	15032

Table 5: Classification report for the RF model with oversampling of minority classes (SMOTE).

Class	Precision	Recall	F1-score	Support
Bitter	0.63	0.67	0.65	1676
Sour	0.81	0.85	0.83	1605
Sweet	0.96	0.89	0.93	9543
Umami	0.22	0.45	0.30	58
Undefined	0.61	0.74	0.67	2150
Accuracy			0.84	15032
Macro avg	0.65	0.72	0.67	15032
Weighted avg	0.86	0.84	0.85	15032

Oversampling and weighted sampling show very similar classification performance as expected. Compared to a RF model without adjustments for data imbalance, both bitter and sweet molecules are detected more accurately whereas umami and undefined have lower scores. Seemingly, oversampling did little to improve classification for the rarest taste, umami.

C Transformer Models

All transformer models were trained on 2 NVIDIA T4 GPUs in Google Cloud using the HuggingFace Transformers library [Wolf et al., 2020]. The training parameters are specified in Table 6. For all other parameters the default values were used.

Table 6: Training parameters used in our experiments.

Parameter	Value
Maximum Sequence Length	512
Learning Rate	5×10^{-5}
Weight Decay	0.01
Warmup Steps	500

Training was continued until overfitting was observed, as indicated by the loss function on the evaluation dataset, or until the loss had saturated. At this point, the best model checkpoint, corresponding to the lowest evaluation loss, was selected for further analysis. An example training run is shown in Figure C.1.

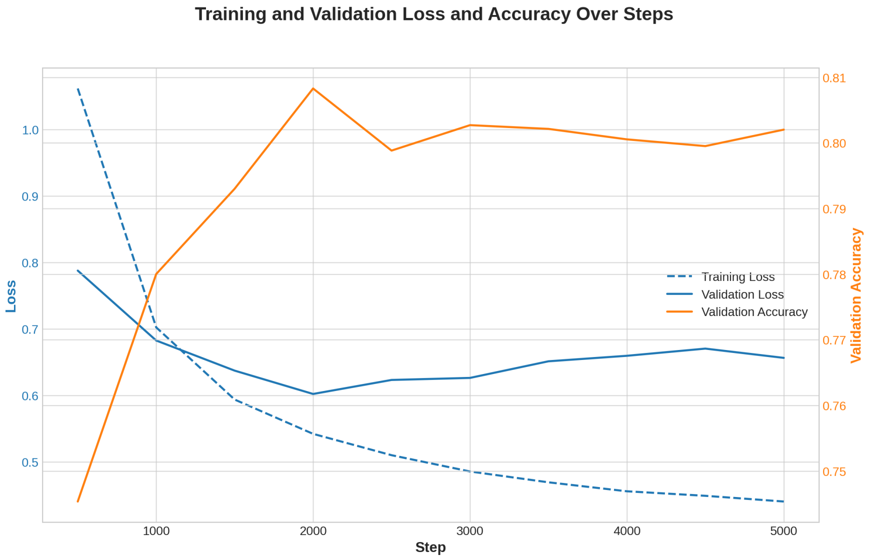


Figure C.1: Training and evaluation loss and accuracy over training steps for one example fine-tuning run. Every step corresponds to 16 data points.

Table 7: Classification report for transformer model without weighted loss function.

Class	Precision	Recall	F1-score	Support
Bitter	0.7082	0.6707	0.6889	246
Sour	0.9055	0.8311	0.8667	219
Sweet	0.9288	0.9185	0.9236	1448
Umami	0.4000	0.4000	0.4000	5
Undefined	0.6615	0.7537	0.7046	337
Accuracy			0.8572	2255
Macro avg	0.7208	0.7148	0.7168	2255
Weighted avg	0.8613	0.8572	0.8586	2255

Table 8: Classification report for the transformer model with weighted loss function.

Class	Precision	Recall	F1-score	Support
Bitter	0.7100	0.6667	0.6876	246
Sour	0.8682	0.8721	0.8702	219
Sweet	0.9538	0.8978	0.9249	1448
Umami	1.0000	0.6000	0.7500	5
Undefined	0.6324	0.8220	0.7148	337
Accuracy			0.8581	2255
Macro avg	0.8328	0.7717	0.7895	2255
Weighted avg	0.8709	0.8581	0.8619	2255

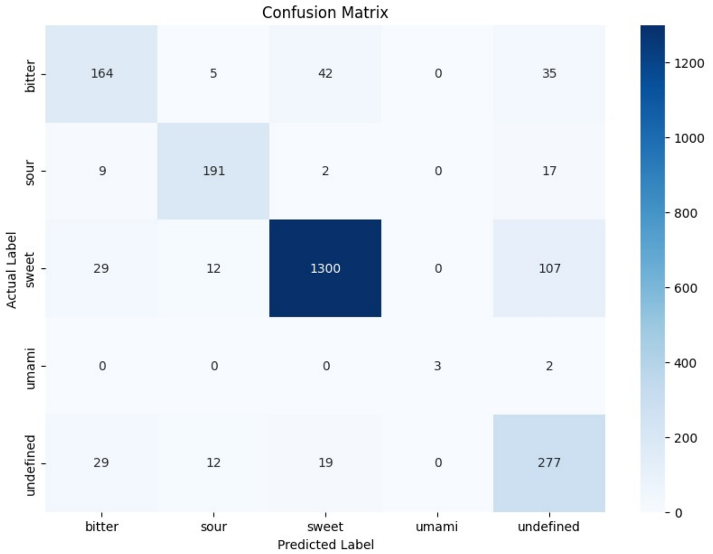


Figure C.2: Confusion matrix for the transformer model with weighted loss function

Table 9: Classification report for a transformer model that was only fine-tuned on an augmented dataset with exclusively non-canonical SMILES strings.

Class	Precision	Recall	F1 Score	Support
Bitter	0.7660	0.6591	0.7086	1634
Sour	0.8705	0.8440	0.8570	1442
Sweet	0.7142	0.8446	0.7739	1441
Umami	0.8484	0.9184	0.8820	1152
Undefined	0.7716	0.7280	0.7492	1522
Accuracy			0.7895	7191
Macro avg	0.7194	0.7988	0.7941	7191
Weighted avg	0.7910	0.7895	0.7878	7191

D Augmentation

SMILES strings are not a unique representation of molecular graphs. To enhance the core dataset of canonical SMILES, we incorporated randomized SMILES and trained several additional models. This approach aimed not only to address class imbalance but also to leverage the reported benefits of data augmentation in improving the generalization capabilities of molecular language models, such as those based on recurrent neural networks [Arús-Pous et al., 2019]. However, it is important to note that our efforts were constrained by the training scheme of the pretrained model, as the ChemBERTa models were trained exclusively on canonical SMILES Chithrananda et al. [2020]. Table 10 shows the augmentation factors used for our experiments.

Table 10: Augmentation factors used for training

Class	Augmentation Factor
Sweet	0
Bitter	7
Sour	7
Umami	250
Undefined	5

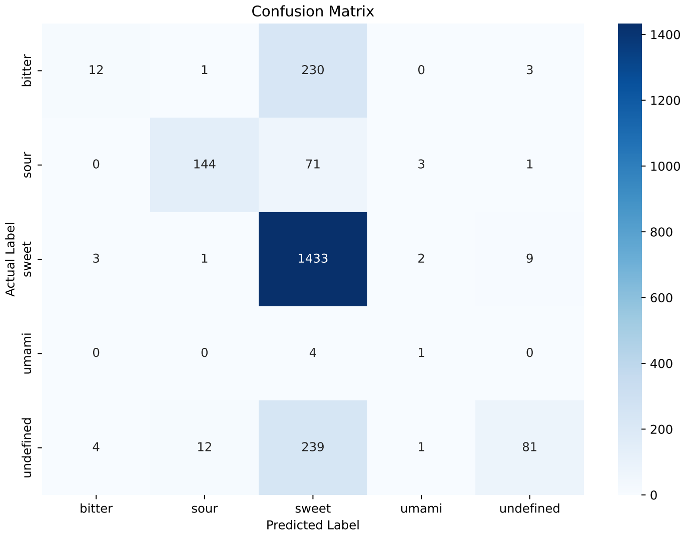


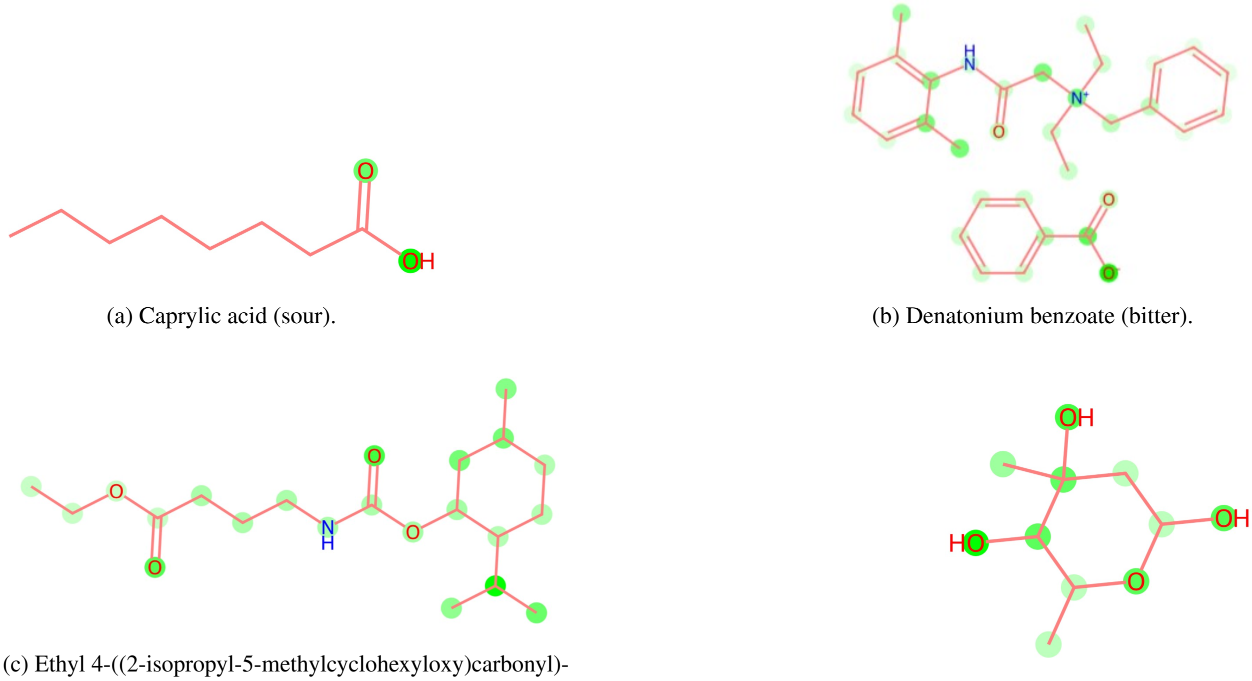
Figure D.1: Confusion matrix for the model trained on augmented SMILES but evaluated on the non-augmented test-set.

When evaluated on the non-augmented test set, the confusion matrix shown in Figure D.1 reveals that the model learns to detect canonical SMILES, as the class of sweet molecules had not been augmented. This indicates a potential bias in the model's learning process towards the canonical representation of SMILES strings. To address this issue, we trained an additional model where the canonical SMILES were replaced by non-canonical SMILES. In a production setting, it would be sensible to use an ensemble approach for prediction. This would involve generating multiple SMILES representations for each molecule, running each representation through the model, and taking the consensus of these predictions.

After data augmentation, the worst performing taste is actually the majority class of sweet but all classes have an F1-score above 0.70 compared to F1-scores for umami between 0.30 and 0.45 for the RF models, see Table 9.

E Attention Weights

In an attempt to make FART interpretable, the “soft” attention weights were extracted from the last layer and averaged over all attention heads. The resultant score was superimposed onto the chemical structure where each atom is highlighted according to its attention weight, see Figure E.1. In general, FART manages to pick up on some expected chemical patterns such as detecting carboxylic acids for sour compounds, see Figure E.1a. Some non-obvious patterns were also present, as Figure E.1c shows, the umami compound ethyl 4-((2-isopropyl-5-methylcyclohexyloxy)carbonyl)butanoate, where a surprising amount of attention is localized at the central carbon of the isopropyl group. This is actually not so surprising as the isopropyl group is a common pattern among many synthetic umami compounds and four of the nineteen molecules collected from the literature [Suess et al., 2015] display this motif.



(a) Caprylic acid (sour).

(b) Denatonium benzoate (bitter).

(c) Ethyl 4-((2-isopropyl-5-methylcyclohexyloxy)carbonyl)-butanoate (umami).

(d) Pyranose (sweet)

Figure E.1: Examples for attention weights for all four taste categories. A darker green corresponds to an atom receiving more average attention. In all cases, FART predicted the correct label for these molecules.

Denatonium benzoate, see Figure E.1b, is an example for a SMILES which FART was not fine-tuned for as it includes an anion (i.e benzoate). The attention weight does focus primarily on the quaternary amine as could be expected but some attention is also diverted to the anion. Removing the benzoate interestingly makes FART more confident in its prediction (from 0.95 to 0.99) which could be expected given it did not train on predicting tastes across multiple substructures.

In general, FART is relatively robust towards small transformations of the SMILES such as changing stereochemistry. Both D- and L-Glucose for example are detected as sweet and adding more terminal carbons to a fatty acid will not change the predicted label. Interestingly, adding even a single carbon in a side-chain does change FART’s prediction from sweet to undefined. While FART has learned to generalize some molecular patterns, its attention is not perfect and is typically more spread out than what most chemists would consider in predicting the taste from a molecular structure.