Supplementary Information for A Chemical Language Model for Multi-Class Molecular Taste Prediction

Contents

Supplementary Methods

Figures

The t-SNE plot (perplexity=30) was generated using 1024-Morgan fingerprints (radius=2) based on PCA initialization and the Jaccard distance metric. Heatmap plots for interpretability were generated using custom code utilizing the SimilarityMaps functionality of RDKit.

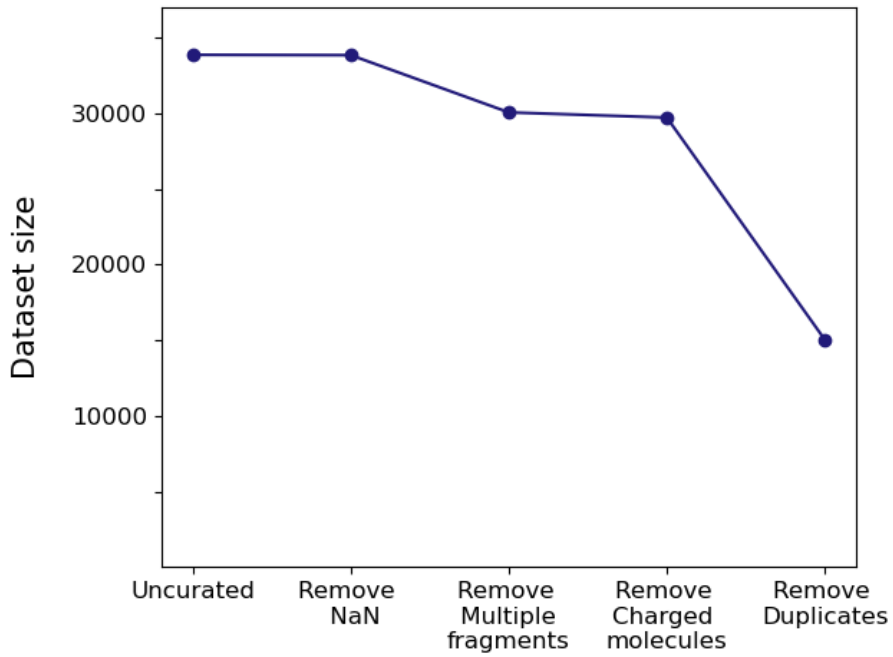
Dataset

An important intermediate aim of this work was to curate a large, high-quality dataset of molecular tastants, combining publicly available data. The dataset utilizes a standard text-based representation of molecules, called SMILES [1]. Every molecule is labeled with a taste: sweet, bitter, sour, umami, and undefined. Where the last category encompasses compounds that had either previously been established as tasteless or compounds that were present in the source databases but for which no clear taste could be associated, this particularly includes many compounds with odor rather than taste labels. Salty was excluded as a taste category because only a very small number of molecules actually produce this taste apart from sodium chloride [2]. Data curation was handled with the cheminformatics package RDKit [3]. The data was enriched with information from the PubChem database.

Supplementary 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	90	631	40	0	144	905
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	9542	1676	1605	58	2150	15031

The FART dataset combines data from six publicly available sources, see Table 1. 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. The database was curated from literature [4]. FlavorDB aggregates data on both gustatory and olfactory sensation from a number of sources [5]. The FART database uses the "flavor profile" given by FlavorDB as most molecules do not have a specific entry for taste. Data from FlavorDB will thus be more heterogeneous given that some of these flavor profiles will actually be based on smell, not taste. Care was taken to only include compounds with unambiguous taste adjectives in the dataset. 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 of the data imbalance in the final dataset. PlantMolecularTasteDB contains 1,527 phytochemicals with associated taste of which 906 were used for this dataset [6]. 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 [7].



Supplementary Figure 1: The dataset size decreases during data curation. Duplicate removal nearly halved the size of the dataset.

Water-soluble, acidic molecules (pK_A between 2 and 7), assumed to taste sour [8], were collected from an ongoing project based with the International Union of Pure and Applied Chemistry (IUPAC) digitizing three high-quality sources of pK_A values in the literature [9–11]. 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. Nonetheless, acidic molecules can be assumed to also taste sour [8]. 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 refer to the most acidic proton and are all measured between 15 and 30 °C in water, i.e. around physiological temperature, excluding any acids that are not water-soluble. Lastly, 19 umami-tasting molecules were collected from the literature [12] of which 11 were not given in any other database.

The combined dataset was reduced to the taste label associated with a canonicalized SMILES representation. The open-source cheminformatics package RDKit [3] was used to further curate the dataset. First, all SMILES that 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. Duplicates could be removed with the help of these standardized SMILES.

While only very few entries with invalid SMILES (21) or charged molecules (342) needed to be removed, the number of entries containing multiple neutral fragments (3783) was more significant. The duplicate removal (14685) reduced the dataset by almost half to a final size of 15,031 entries, see Figure 1. The large number of duplicates underlines the significant overlap among the databases used. When duplicate entries existed from different sources, which source would be given in the final dataset was arbitrarily determined based on the

index. The final dataset exhibits a strong data imbalance, where sweet represents over 60% and umami less than 1% of the data.

The curated dataset was further enriched by general information (PubChemID, IUPAC name, molecular formula, molecular weight, InChI, InChIKey), accessed through the PubChem API [13]. The dataset, FartDB, was published in agreement with the FAIR principles [14] and can be accessed through several different interfaces to encourage its use by other research projects.

Model Training

All transformer models were trained on multiple NVIDIA T4 GPUs in Google Cloud using the HuggingFace Transformers library [15]. For all experiments, the ChemBERTa checkpoint `seyonec/SMILES_tokenized_PubChem_shard00_160k` on HuggingFace was used, consisting of 6 layers and a total of 83.5 million parameters. Training on the unaugmented dataset was run for 20 epochs, while training on the augmented dataset was run for 2 epochs. A weight decay of 0.01 was applied, and a batch size of 16 was used. For all other parameters, the default values for fine-tuning were used. 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. The following weighted loss function was used for the weighted model:

$$\mathcal{L} = -\sum_{i=1}^{N} w_{y_i} \cdot \log\left(\frac{e^{z_{i,y_i}}}{\sum_{j=1}^{C} e^{z_{i,j}}}\right), \tag{1}$$

where N is the number of samples in the batch, C is the number of classes, $z_{i,j}$ is the logit (raw output) for the jth class of the ith sample, y_i is the true class label for the ith sample, and w_{y_i} is the weight associated with the true class y_i.

Supplementary Table 2: Overview of FART model variations.

Model	Link
FART	FartLabs/Stable_A
FART augmented	FartLabs/Stable_B
FART augmented + weighted loss function	FartLabs/Stable_C

Model Evaluation

Multi-Class Averages

To evaluate multi-class classification performance, macro and weighted averages are commonly used to summarize metrics across all classes. The macro (unweighted) average is computed with

$$\text{Macro} = \frac{1}{C}\sum_{i=1}^{C} M_i, \tag{2}$$

where C is the number of classes and M_i is the metric (e.g., precision, recall, F1-score) for the ith class. The weighted average is given by

$$\text{Weighted} = \sum_{i=1}^{C} \frac{n_i}{N} \cdot M_i, \tag{3}$$

where n_i is the number of instances in class i, N is the total number of instances across all classes, and M_i is the metric for the ith class.

Multi-Label Data

Molecules that could be associated with multiple tastes during data curation are given as duplicates in the dataset with the same canonicalized SMILES but different taste labels. To test how FART (augmented, no weighted loss function) evaluates on these multi-label molecules, we considered all labels above a probability of 0.2, i.e. higher than a uniform distribution across the classes, as relevant rather than considering the highest probability as done for normal evaluation. Of the 213 molecules, which were nearly all seen during training, FART only correctly associates 23 of the molecules with only and all of the training labels. In 18 cases FART associates too many labels, in 15 of these the additional label is "undefined". Overwhelmingly, however, FART collapses the multiple labels into a single one. In 101 of the 213 cases, only a single but correct label was predicted. Ultimately, during learning as well as inference, FART is tasked with producing a single output label and hence it is unsurprising that it struggles with this parallel multi-class prediction task. More work is needed to develop models that more accurately reflect the nature of multi-class tastants.

Interpretability Framework

Integrated Gradients [16] is a method for attributing a deep neural network's prediction to its input features. The core idea is to integrate the gradients of the output taken along a linear path from a baseline input to the input at hand. Mathematically, for a neural network $F(x)$, an input x and baseline input x' (e.g. the zero input), the attribution for the ith feature is:

$$\text{IntegratedGrads}_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial F\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha. \tag{4}$$

The method satisfies important axioms like sensitivity (if inputs differ in one feature but have different predictions, that feature should receive attribution) and implementation invariance (attributions are identical for functionally equivalent networks). The method is readily available for Hugging Face Transformers models through the transformers-interpret package [17].

Supplementary Table 1: Performance Overview with Weighted Averages

Supplementary Table 3: Performance comparison between the trained transformers and baseline classifiers. Scores are given as weighted averages across taste classes which penalizes wrong but rare predictions on a minority class less compared to an unweighted average. Scores for Random Forest and XGBoost were obtained through five-fold cross-validation. Area under the receiver operating characteristic (AUROC) values are calculated as one-vs-rest for each taste class and then combined into a weighted average.

		Weighted average				
Model	Accuracy	Precision	Recall	F1 Score	AUROC	Support
XGBoost: fingerprints (fp)	0.8572	0.8616	0.8572	0.8564	0.8821	100%
XGBoost: fp+descriptors	0.8526	0.8522	0.8526	0.8506	0.8716	100%
Balanced Random Forest: fp	0.7375	0.8105	0.7375	0.7580	0.8296	100%
FART	0.8621	0.8650	0.8621	0.8607	**0.9617**	100%
FART augmented	0.8670	0.8610	0.8670	0.8626	**0.9643**	100%
FART augmented + weighted	0.8532	0.8721	0.8532	0.8592	0.9576	100%
FART augmented + confidence	**0.8837**	**0.8986**	**0.8837**	**0.8887**	**0.9686**	93%

Supplementary Tables 2-4: Class-Resolved Performance Data for XGBoost and Random Forest

Supplementary Table 4: Class-resolved performance data for the XGBoost model trained on Morgan fingerprints. The entire data set is considered through five-fold cross validation.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.8104	0.5943	0.6857		1676
Sour		0.8242	0.8941	0.8577		1605
Sweet		0.9288	0.9257	0.9273		9542
Umami		0.7600	0.3276	0.4578		58
Undefined		0.6338	0.7447	0.6848		2150
Weighted Average		0.8616	0.8572	0.8564	0.8821	
Unweighted Average		0.7915	0.6973	0.7227	0.8250	
Overall	0.8572					15031

Supplementary Table 5: Class-resolved performance data for the XGBoost model trained on Morgan fingerprints in addition to 15 Mordred descriptors. The entire data set is considered through five-fold cross validation.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.7842	0.6116	0.6872		1676
Sour		0.8292	0.8773	0.8526		1605
Sweet		0.9146	0.9299	0.9222		9542
Umami		0.7273	0.2759	0.4000		58
Undefined		0.6490	0.6949	0.6712		2150
Weighted Average		0.8522	0.8526	0.8506	0.8716	
Unweighted Average		0.7809	0.6779	0.7066	0.8133	
Overall	0.8526					15031

Supplementary Table 6: Class-resolved performance data for the Balanced Random Forest using Morgan fingerprints. The entire data set is considered through five-fold cross validation.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.6845	0.4039	0.5081		1676
Sour		0.5583	0.8766	0.6822		1605
Sweet		0.9436	0.7785	0.8531		9542
Umami		0.0567	0.7241	0.1051		58
Undefined		0.5263	0.7121	0.6053		2150
Weighted Average		0.8105	0.7375	0.7580	0.8296	
Unweighted Average		0.5539	0.6991	0.5507	0.8154	
Overall	0.7375					15031

Supplementary Tables 5-8: Class-Resolved Performance Data for the FART models

Supplementary Table 7: Class-resolved performance data for the FART model trained on the unaugmented dataset with an unweighted loss function.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.7955	0.5882	0.6763		238
Sour		0.8078	0.9075	0.8548		227
Sweet		0.9286	0.9267	0.9276		1459
Umami		1.0000	0.3750	0.5455		8
Undefined		0.6658	0.7523	0.7064		323
Weighted Average		0.8650	0.8621	0.8607	0.9617	
Unweighted Average		0.8395	0.7099	0.7421	0.9644	
Overall	0.8621					2255

Supplementary Table 8: Class-resolved performance data for the FART model trained on the augmented dataset with an unweighted loss function.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.7594	0.6765	0.7156		238
Sour		0.8571	0.8987	0.8774		227
Sweet		0.9096	0.9520	0.9303		1459
Umami		0.5000	0.3750	0.4286		8
Undefined		0.7279	0.6130	0.6655		323
Weighted Average		0.8610	0.8670	0.8626	0.9643	
Unweighted Average		0.7508	0.7030	0.7235	0.9649	
Overall	0.8670					2255

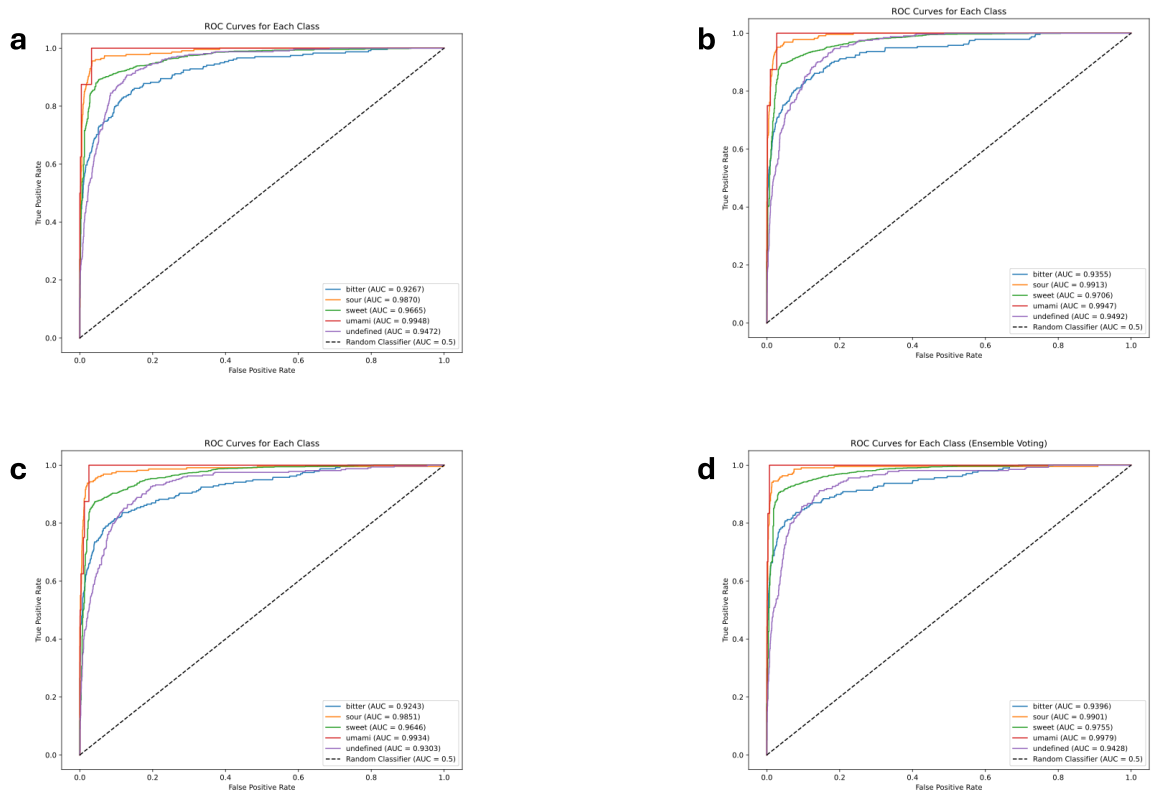
Supplementary Table 9: Class-resolved performance data for the FART model trained on the augmented dataset with a weighted loss function.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.6680	0.7269	0.6962		238
Sour		0.8480	0.9339	0.8889		227
Sweet		0.9690	0.8780	0.9213		1459
Umami		0.5714	0.5000	0.5333		8
Undefined		0.6091	0.7864	0.6865		323
Weighted Average		0.8721	0.8532	0.8592	0.9576	
Unweighted Average		0.7331	0.7650	0.7452	0.9595	
Overall	0.8532					2255

Supplementary Table 10: Performance overview for the confidence model trained on the augmented dataset with a weighted loss function. A prediction is only made when all 10 augmented SMILES for a given molecule result in the same label. Molecules for which no consensus is reached are not predicted. The percentages in the support column indicate how many molecules the model predicted on out of the unaugmented test set.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.7260	0.7644	0.7447		208 (87%)
Sour		0.8879	0.9406	0.9135		219 (96%)
Sweet		0.9800	0.9086	0.9430		1401 (96%)
Umami		0.6000	0.5000	0.5455		6 (75%)
Undefined		0.6278	0.8095	0.7072		273 (85%)
Weighted Average		0.8986	0.8837	0.8887	0.9686	
Unweighted Average		0.7643	0.7846	0.7708	0.9692	
Overall	0.8837					2107 (93%)

Supplementary Figure 2: Receiver Operating Characteristics (ROC) for FART Models



Supplementary Figure 2: Receiver operating characteristics (ROC) for all FART models. (**a**) Unaugmented training data with unweighted loss function. (**b**) Augmented training data with unweighted loss function. (**c**) Augmented training data with weighted loss function. (**d**) Confidence model (10 models in agreement) based on augmented training data and weighted loss function.

Supplementary Tables 9-10: Comparison Between Augmented and Unaugmented FART Models on Non-Canonical SMILES

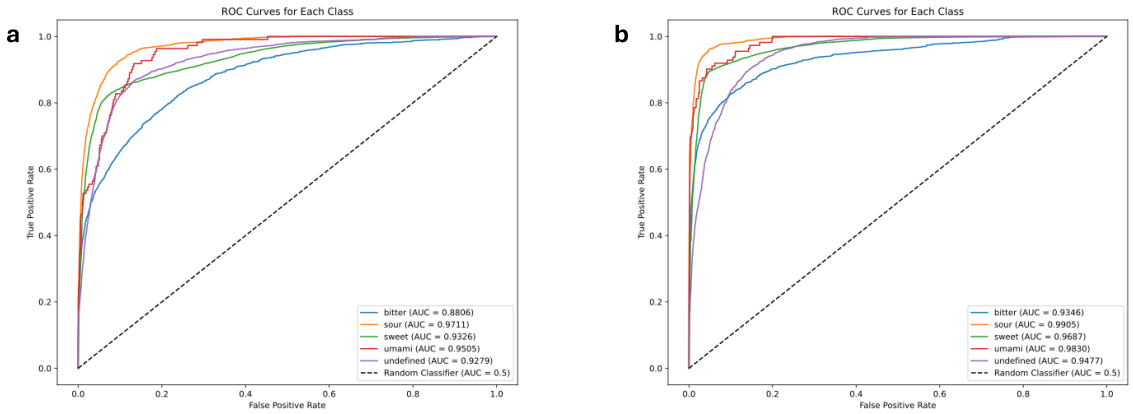
Supplementary Table 11: Performance of the unaugmented FART model with unweighted loss function on an augmented test set including non-canonical SMILES. The performance drops markedly compared to an evaluation on only canonical SMILES suggesting that the unaugmented FART has not robustly learned a mapping from structure to taste.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.5167	0.5841	0.5483		3258
Sour		0.6734	0.8236	0.7410		3039
Sweet		0.9358	0.8503	0.8910		20270
Umami		0.8095	0.1545	0.2595		110
Undefined		0.6105	0.7335	0.6664		4161
Weighted Average		0.8213	0.8013	0.8075	0.9304	
Unweighted Average		0.7092	0.6292	0.6213	0.9325	
Overall Accuracy	0.8013					30838

Supplementary Table 12: Performance of the augmented FART model with unweighted loss function on an augmented test set including non-canonical SMILES. The performance remains essentially unchanged compared to evaluating on canonical SMILES as predictions are now robust towards non-canonical input.

Taste Class	Accuracy	Precision	Recall	F1 Score	AUROC	Support
Bitter		0.7582	0.6805	0.7173		3271
Sour		0.8524	0.8810	0.8665		3060
Sweet		0.9081	0.9522	0.9296		20257
Umami		0.5610	0.4107	0.4742		112
Undefined		0.7163	0.5921	0.6483		4158
Weighted Average		0.8596	0.8658	0.8613	0.9643	
Unweighted Average		0.7592	0.7033	0.7272	0.9649	
Overall Accuracy	0.8658					30858

Supplementary Figure 3: Receiver Operating Characteristics for Evaluation on Non-Canonical SMILES



Supplementary Figure 3: The performance of the FART model trained only canonical SMILES, i.e. on the unaugmented train set, (**a**) drops markedly when evaluating on non-canonical SMILES. The performance of the FART model trained on augmented SMILES (**b**) is robust towards non-canonical SMILES.

References

1. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. en. *Journal of Chemical Information and Computer Sciences* **28,** 31–36. ISSN: 0095-2338. http://dx.doi.org/10.1021/ci00057a005 (Feb. 1, 1988).
2. Taruno, A. & Gordon, M. D. Molecular and Cellular Mechanisms of Salt Taste. *Annual Review of Physiology* **85,** 25–45. ISSN: 1545-1585 (2023).
3. Landrum, G. *et al. rdkit/rdkit: 2020_03_1 (Q1 2020) Release* version Release_2020_03_1. Mar. 2020. https://doi.org/10.5281/zenodo.3732262.
4. Rojas, C. *et al.* ChemTastesDB: A curated database of molecular tastants. *Food Chemistry: Molecular Sciences* **4,** 100090. ISSN: 2666-5662 (2022).
5. Garg, N. *et al.* FlavorDB: a database of flavor molecules. *Nucleic Acids Research* **46,** D1210–D1216. ISSN: 0305-1048 (Oct. 2017).
6. Gradinaru, T.-C., Petran, M., Dragos, D. & Gilca, M. PlantMolecularTasteDB: A Database of Taste Active Phytochemicals. *Frontiers in Pharmacology* **12.** ISSN: 1663-9812 (2022).
7. Bayer, S. *et al.* Chemoinformatics View on Bitter Taste Receptor Agonists in Food. *Journal of Agricultural and Food Chemistry* **69,** 13916–13924 (2021).
8. Roper, S. D. *Taste: Mammalian Taste Bud Physiology* 2017. https://www.sciencedirect.com/science/article/pii/B9780128093245029084.
9. Perrin, D. D. *Dissociation Constants of Organic Bases in Aqueous Solution, Supplement* (IUPAC, Butterworths, 1965).
10. Perrin, D. D. *Dissociation Constants of Organic Bases in Aqueous Solution* (IUPAC, Butterworths, 1972).
11. Serjeant, E. P. & Dempsey, B. *Ionisation Constants of Organic Acids in Aqueous Solution* (Oxford IUPAC Chemical Data Series, Oxford/Pergamon, 1979).
12. Suess, B., Festring, D. & Hofmann, T. in *Flavour Development, Analysis and Perception in Food and Beverages* (eds Parker, J., Elmore, J. & Methven, L.) 331–351 (Woodhead Publishing, 2015). ISBN: 978-1-78242-103-0.
13. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Research* **51,** D1373–D1380. ISSN: 0305-1048 (Oct. 2022).
14. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018. ISSN: 2052-4463. https://doi.org/10.1038/sdata.2016.18 (2016).
15. Wolf, T. *et al. Transformers: State-of-the-Art Natural Language Processing* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) (Association for Computational Linguistics, Online, Oct. 2020), 38–45. https://aclanthology.org/2020.emnlp-demos.6.
16. Sundararajan, M., Taly, A. & Yan, Q. *Axiomatic Attribution for Deep Networks* 2017. arXiv: 1703.01365 [cs.LG]. https://arxiv.org/abs/1703.01365.
17. Pierse, C. *Transformers Interpret* version 0.5.2. 2023. https://github.com/cdpierse/transformers-interpret.