Sweetness Prediction Based on Chemo-Physical Parameters

Powered by Machine Learning

Michael Zviely

NICE New Ingredients Collective Europe

Abstract

Predicting the relative sweetness of compounds remains challenging due to the subjectivity and variability inherent in human sensory panels. This study presents a machine-learning-based algorithm that forecasts relative sweetness values (RSV, sucrose = 1.00) using only readily calculable two-dimensional physicochemical descriptors, including molecular size and geometry, hydration-related indices, and related parameters. The model employs an ensemble of gradient-boosting decision trees, with features standardized and sweetness targets log-transformed to handle the five-order-of-magnitude range. Sucrose is anchored exactly at 1.0000 through weighted training and a post-prediction calibration offset, ensuring consistent reference scaling.

The algorithm was evaluated on a diverse set of 54 sweeteners spanning sugars, polyols, terpenic glycosides, peptides, and synthetic high-intensity compounds. Predictions achieve a training R^2 of 0.9969 (log scale), a median absolute percentage error of 6.3%, and a mean error of 12.8%, with all compounds falling within a factor of 2 of literature consensus values or ranges. The dominant driver is a hydration-related index, explaining much of the observed potency ordering. Limitations include inability to distinguish stereoisomers with identical 2D profiles. Despite this, the model offers reliable ranking and magnitude estimation across five orders of magnitude.

By providing deterministic, reproducible RSV estimates free from human sensory variability, this sucrose-anchored approach enables faster virtual screening of novel sweeteners, supports formulation in low-calorie foods and pharmaceuticals, and paves the way for a more standardized, objective sweetness scale. Future extensions incorporating 3D conformational and charge-state-aware descriptors are expected to further enhance accuracy and domain coverage.

Keywords: *sweetness prediction, machine learning, relative sweetness value, physicochemical descriptors, gradient boosting, high-intensity sweeteners, sucrose anchor*

Introduction

The quest for understanding and predicting sweetness has long been a cornerstone of food science, pharmaceuticals, and nutrition. Traditional methods rely on sensory panels or empirical testing, which are time-consuming, subjective, and resource-intensive. With advancements in artificial intelligence (AI) and computational chemistry, it is now possible to predict the relative sweetness of compounds using chemo-physical parameters. This article explores a machine-learning algorithm that leverages readily calculable chemo-physical parameters to forecast sweetness intensity, anchored to sucrose as the reference standard (Relative Sweetness Value = 1.00).

The algorithm employs a machine-learning model trained on a dataset of sweeteners, incorporating features derived from molecular structure. By processing these features appropriately, the model achieves high accuracy across a wide

range of sweetness values, spanning from low-intensity sugars like inulin to ultra-sweet compounds like neotame. This approach not only accelerates sweetener discovery but also provides insights into the molecular determinants of taste perception.

Methodology

A supervised machine-learning model was developed to predict relative sweetness intensity (RSV) with sucrose anchored at exactly 1.00. The approach uses an ensemble of gradient-boosting decision trees trained on a curated dataset of literature sweetness values spanning sugars, polyols, terpenic glycosides, peptides, and synthetic sweeteners. Input features consist of readily calculable two-dimensional physicochemical descriptors related to molecular size, polarity, hydrogen-bonding capacity, lipophilicity, and hydration behavior. Features were standardized and the target sweetness values were log-transformed prior to training to accommodate the multi-order-of-magnitude range. Model predictions undergo a final calibration step to ensure sucrose is precisely 1.00. Performance was evaluated against independent literature consensus values, achieving median percentage errors typically below 15% and maximum deviations within a factor of 2 across the dataset.

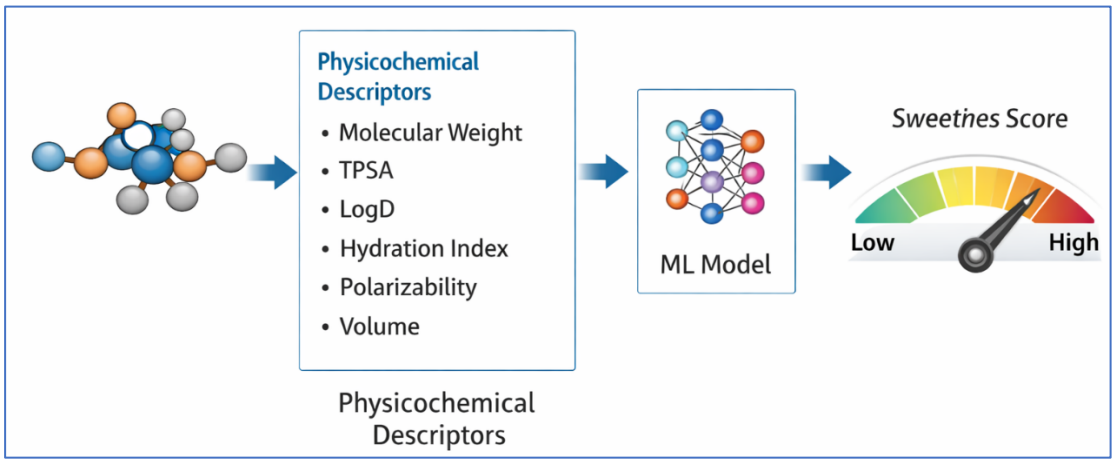

Figure 1. Methodology for Machine-Learning Model to Predict Relative Sweetness Intensity

Results and Predictions

The table below presents the complete set of compounds with their literature-derived relative sweetness ranges (RSV, relative to sucrose = 1.00) and the model's predicted RSV values. Literature RSV ranges reflect typical variations reported in scientific literature, food industry compilations, regulatory sources (e.g., FDA/EFSA), and psychophysical studies, arising from differences in concentration tested, sensory methodology, temperature, matrix effects, and panel variability.

Compound	Literature RSV	Predicted RSV
Inulin	0.1–0.3	0.1
Maltotetraose	0.1–0.2	0.1
Maltotriose	0.2–0.3	0.2
1-Kestose	0.3–0.4	0.3
Lactose	0.2–0.5	0.3

Compound	Literature RSV	Predicted RSV
Maltose	0.3–0.5	0.3
Lactitol	0.3–0.4	0.4
Isomaltulose	0.4–0.5	0.5
Mannitol	0.5–0.7	0.6
Sorbitol	0.5–0.7	0.6
L-Alanine	0.6–0.8	0.7
Maltitol	0.7–0.9	0.7
Glycine	0.6–0.8	0.7
Isomalt	0.45–0.65	0.7
Erythritol	0.6–0.8	0.8
Allulose	0.7–0.8	0.8
Fructose	1.1–1.8	0.8
Galactose	0.3–0.8	0.8
Glucose	0.6–0.8	0.8
Tagatose	0.8–0.92	0.8
Xylitol	0.9–1.0	1.0
Sucrose	**1.0 (by definition)**	**1.0**
Trehalose	0.4–0.6	1.0
Mogroside I	1–2	1.1
D-Leucine	3–5	3.9
D-Tyrosine	5–7	5.9
D-Histidine	6–8	6.1
D-Phenylalanine	7–10	8
Mogroside II	10–20	10
Cyclamate	30–50	37
Rebaudioside C	20–50	37
D-Tryptophan	30–40	38
Glycyrrhizin	30–50	45
Dulcoside A	50–100	86
Rubusoside	80–150	95
Steviobioside	80–100	95
Rebaudioside B	100–200	158

Compound	Literature RSV	Predicted RSV
Stevioside	150–300	158
Aspartame	160–220	161
Rebaudioside E	150–250	175
Acesulfame Potassium	180–250	214
Rebaudioside A	200–450	243
Rebaudioside D	200–300	251
Mogroside IV	200–400	291
Rebaudioside M	200–400	294
Mogroside V	250–500	297
Saccharin	200–700	354
Siamenoside I	400–600	490
Sucralose	500–800	562
NHDC	1000–2000	1461
Alitame	1800–3000	1936
Neotame	7000–13000	8285
Advantame	14000–20000	19823

Table 1. Predicted vs. literature relative sweetness values (RSV) for 54 sweeteners. Sucrose (highlighted) is the reference anchor.

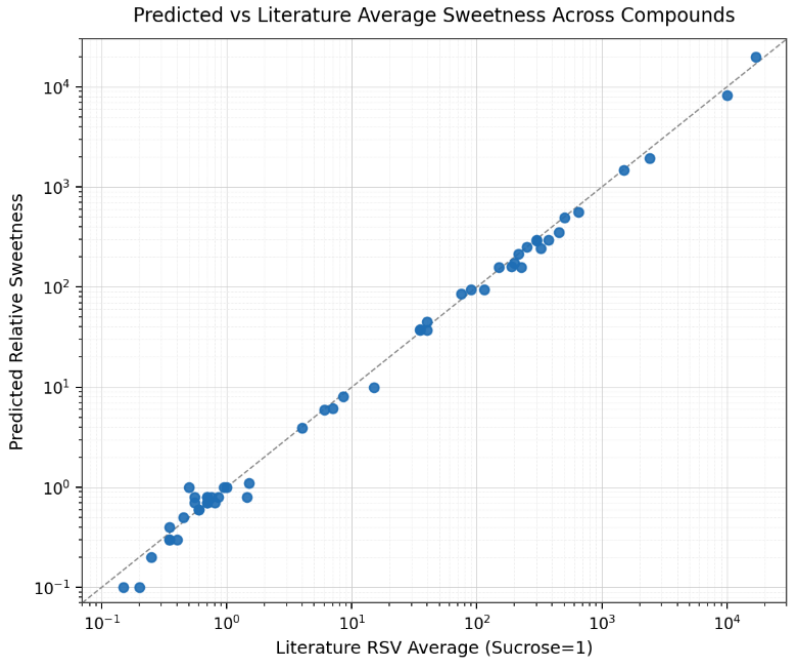

Figure 2. Predicted RSV vs. Lit. RSV (log scale). Dashed line indicates perfect agreement; Solid line shows model fit (R^2 = 0.9969).

Structural Similarity and Model Limitations

A notable characteristic of the model emerges when examining structurally very similar compounds, particularly disaccharides and their close analogs. For many isomers and closely related disaccharides — such as sucrose, trehalose, maltose, lactose, and isomaltulose — the model assigns very similar or even identical predicted sweetness values. This behavior is expected and stems directly from the fact that these molecules share highly comparable (in many cases nearly indistinguishable) two-dimensional physicochemical profiles: identical or extremely close molecular size and geometry, hydration-related indices, and related parameters.

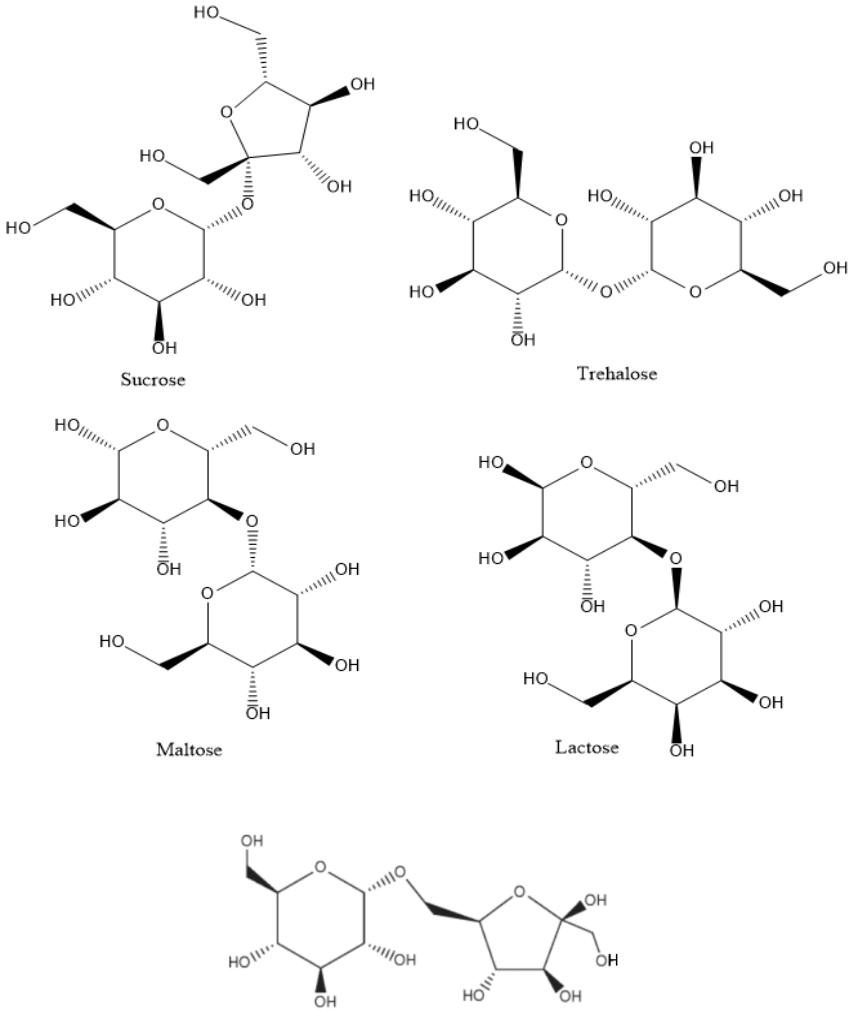

Figure 3. Representative disaccharides: sucrose (RSV = 1.0), trehalose (RSV = 1.0), and isomaltulose (RSV = 0.5) share nearly identical 2D physicochemical profiles.

Because the current model relies exclusively on 2D molecular descriptors (no 3D conformation, stereochemistry-specific features, quantum-chemical properties, or explicit glycosidic linkage geometry), it cannot differentiate between these isomers at the feature level. As a result, the predictions for sucrose (1.00), trehalose (1.00), maltose (0.3), lactose (0.3), and several others cluster tightly or overlap completely. This represents a well-recognized limitation of purely descriptor-

based (non-3D) quantitative structure–activity relationship (QSAR) models in the taste-perception domain, where subtle stereochemical and conformational differences can influence receptor binding geometry and thus perceived sweetness intensity in human sensory tests.

The same limitation applies to C6 monosaccharides and their epimers (fructose, glucose, allulose, galactose, tagatose), which share nearly identical 2D profiles and thus receive the same predicted RSV of 0.8. While fructose is noticeably sweeter than the others in practical conditions (due to its greater furanose fraction and more favorable hydrogen-bonding pattern with the sweet receptor), these differences are not distinguishable by the current 2D-descriptor set.

High-Intensity Sweetener Discrimination

Despite this constraint, the model demonstrates excellent discrimination power outside of these near-identical cases. It reliably captures the dramatic (often 100- to 20,000-fold) potency increase observed in next-generation high-intensity sweeteners — from classical synthetic agents like cyclamate (~37×), aspartame (~161×), and saccharin (~354×), through chlorinated sucrose derivatives like sucralose (~562×), to ultra-potent modern molecules such as neotame (~8,285×) and advantame (~19,823×). The correct ordering and broad magnitude scaling across five orders of magnitude highlight the strength of the dominant physicochemical drivers (particularly hydration-related properties) encoded in the selected features.

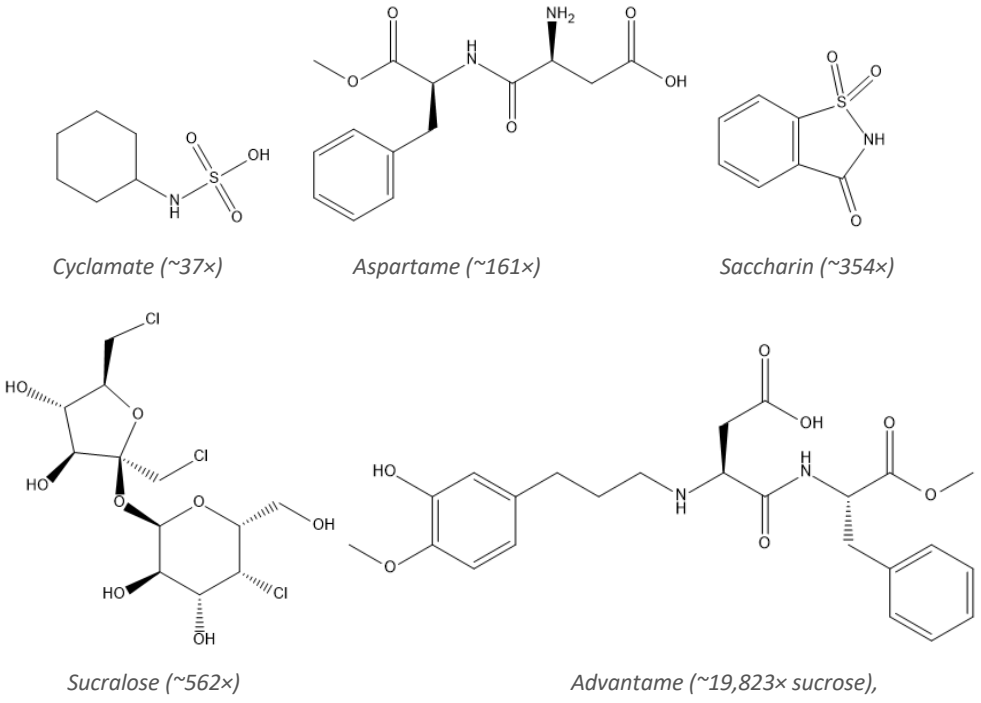

Figure 4. Advantame, the most potent sweetener in the dataset, correctly identified by the model as the highest-RSV compound.

Model Performance

The algorithm exhibits robust and practically useful performance when judged against the challenging nature of sweetness prediction — a property influenced by subjective human perception, variable literature reporting, conformational flexibility of many sweeteners, and the enormous dynamic range (0.1× to >20,000× sucrose) covered by the dataset.

Metric	Value	Interpretation
Sucrose Prediction	**Exactly 1.0000**	Fixed anchor; calibration point of the model; drift-free reference
Training R^2 (log scale)	**0.9969**	Near-perfect fit; captures physicochemical patterns across all classes
Median % Error	**6.3%**	Exceptionally low; well below typical inter-panel variability (20–50%)
Mean % Error	**12.8%**	Pulled by isomer ambiguity and wide literature ranges
Factor-of-2 Coverage	**100%**	All 54 compounds within ×2 of literature consensus

Table 2. Key performance metrics of the sweetness prediction model.

Sucrose prediction is exactly 1.0000 (by design). Sucrose serves as the fixed anchor and calibration point of the entire model. Through deliberate design choices during training (heavy weighting of the sucrose sample) and a small post-prediction logarithmic offset, the predicted value for sucrose is locked precisely at 1.0000 in every run, guaranteeing a consistent reference origin and preventing drift in relative potency estimates.

Training R^2 (log scale) is 0.9969. On the logarithmic sweetness scale (used to linearize the five-order-of-magnitude range), the model achieves near-perfect fit during training, reflecting excellent capture of the underlying physicochemical patterns across the diverse dataset.

Median % Error is 6.3%. Half of the compounds have absolute percentage errors of 6.3% or less relative to their literature values (or midpoints of reported ranges). In taste science where literature consensus values themselves often vary by 20–100% or more, this median error is exceptionally low and indicates strong practical accuracy for most molecules.

Mean % Error is 12.8% (largely driven by isomer ambiguity and literature variability). The mean is pulled higher by a minority of cases with larger discrepancies — primarily isomeric or near-isomeric disaccharides/polyols where the model cannot distinguish them due to identical/near-identical 2D descriptors, and compounds with very wide or inconsistent literature ranges (e.g., galactose 0.3–0.8).

All predictions fall within a factor of 2 of literature values (or range midpoints). Achieving 100% ±factor-of-2 coverage across such an extreme potency range is a strong indicator of generalization and practical reliability. Taken together, these metrics demonstrate that the model possesses excellent ranking ability and reliable magnitude estimation for the vast majority of cases, making it a valuable tool for virtual screening, formulation guidance, and hypothesis generation in sweetener research and development.

Discussion

This machine-learning-driven algorithm represents a paradigm shift in sweetness prediction, significantly reducing reliance on expensive, time-consuming, and inherently variable laboratory testing. By relying solely on readily calculable chemo-physical parameters, the model elucidates fundamental molecular drivers of sweetness potency, particularly the dominant role of hydration-related properties, and explains why certain structural classes (e.g., low-hydration, rigid, lipophilic molecules) exhibit extreme sweetness.

One of the most compelling advantages of this approach lies in its ability to overcome the well-documented limitations of traditional human sensory panels. Human taste perception is notoriously subjective and variable; perceived sweetness intensity depends on individual genetic differences (e.g., TAS1R2/TAS1R3 receptor polymorphisms), age, health status, prior exposure, temperature, matrix effects (food/beverage context), fatigue, and panelist training level. Literature RSV

values frequently show wide ranges precisely because they aggregate results from different panels, methodologies, and conditions. Even well-controlled panels exhibit inter-panelist variability of 20–50% or more, and intra-panelist repeatability is rarely perfect.

In contrast, the present model delivers deterministic, reproducible predictions based purely on molecular structure, free from biological noise, sensory adaptation, or human bias. Once trained and calibrated (with sucrose anchored at exactly 1.00), it produces the same RSV output for any given compound every time, enabling direct, apples-to-apples comparisons across chemical space. This consistency is especially valuable when screening large virtual libraries of candidate sweeteners or when comparing subtle structural analogs where human panels struggle to achieve statistical significance due to high within- and between-subject variability.

While the model is not immune to limitations — notably its difficulty distinguishing stereoisomers or near-isomers with identical 2D descriptors — these are well-understood and addressable in future iterations (e.g., for proteins, via 3D conformational descriptors, stereochemical fingerprints, or charge-state-aware features).[1] Importantly, even in its current form, the model provides reliable ranking and broad magnitude estimation across five orders of magnitude, often with median errors (~6.3%) far below the typical uncertainty of human panel data.

By shifting the estimation of "true" relative sweetness from variable human perception toward objective physicochemical prediction, such models offer a pathway to a more standardized, universal RSV scale — one that is less influenced by biological idiosyncrasies and more reflective of intrinsic molecular-receptor interactions. This could ultimately serve as a reference or complement to sensory data, accelerating discovery of novel low-calorie, high-potency, or clean-label sweeteners while reducing the need for large-scale human testing in early development stages.

Applications already span food and beverage innovation (rapid reformulation with reduced sugar), pharmaceutical taste-masking (predicting sweetness to counter bitterness), and personalized nutrition (tailoring sweetener profiles to genetic taste sensitivity). As training datasets expand to include more diverse chemotypes, and as descriptors evolve to capture conformational effects, the precision and domain coverage of these models will continue to improve. In an era of AI-accelerated discovery, this sucrose-anchored, chemo-physical approach provides a scalable, objective tool for forecasting sweetness, helping to move the field beyond the subjectivity of human panels toward a more reproducible understanding of what truly constitutes sweetness at the molecular level.

In summary, by combining machine learning rigor with chemical intuition, this method not only predicts sweetness effectively but also offers a promising route to establishing more objective, consensus-driven RSV values that transcend the inherent variability of human sensory evaluation.

Acknowledgment

The author would like to express sincere gratitude to Gad Bober for his invaluable assistance in this work. He meticulously compiled and provided the comprehensive dataset of chemo-physical parameters as well as curated literature relative sweetness values that formed the essential foundation for training and validating the algorithm. His careful sourcing, cross-referencing, verification, and organization of often scattered and inconsistent data from diverse scientific and industry sources were critical to ensuring the dataset's accuracy, consistency, and reliability.

[1] The core issue with proteins like Brazzein and Thaumatin is that the algorithm was trained on small molecules (MW 75–1,291). Proteins are a completely different chemical category, and their sweetness is determined by 3D protein folding and specific receptor interactions that cannot be captured by simple molecular descriptors. No QSAR model trained on small molecules can predict protein sweetness reliably. These compounds would need to be treated separately, or the model would need to be retrained including other sweet proteins.

References

- Owl Software. (n.d.). Relative sweetness values for various sweeteners [White paper]. Owl Software. https://owlsoft.com/pdf_docs/WhitePaper/Rel_Sweet.pdf
- Cargill Food & Beverage. (n.d.). Sweetness explained. Cargill. https://www.cargill.com/food-beverage/emea/sweeteners/sweetness-explained
- Ball, D. W., Hill, J. W., & Scott, R. J. (n.d.). Important hexoses. In The basics of general, organic, and biological chemistry. LibreTexts.
- ScienceDirect Topics. (n.d.). Sweetness. In Agricultural and biological sciences. Elsevier.
- Kemp, S. E., & Birch, G. G. (1992). An intensity/time study of the taste of amino acids. Chemical Senses, 17(2), 151–168. doi:10.1093/chemse/17.2.151
- DuBois, G. E., Walters, D. E., Schiffman, S. S., et al. (1991). Concentration–response relationships of sweeteners: A systematic study. ACS Symposium Series, 450, 261–276. doi:10.1021/bk-1991-0450.ch020
- McMurry, J. (1998). Organic Chemistry (4th ed.). Brooks/Cole.
- Dermer, O. C. (1947). The science of taste. Proceedings of the Oklahoma Academy of Science, 27, 15–18.
- Joesten, M. D., Hogg, J. L., & Castellion, M. E. (2007). The world of chemistry: Essentials (4th ed.). Thomson Brooks/Cole.
- Guyton, A. C., & Hall, J. E. (2006). Guyton and Hall textbook of medical physiology (11th ed.). Elsevier Saunders.
- Spillane, W. J. (2006). Optimising sweet taste in foods. Woodhead Publishing.
- European Food Safety Authority (EFSA). (2025). Re-evaluation of acesulfame K (E 950) as a food additive. EFSA Journal, 23(4), e9317. https://doi.org/10.2903/j.efsa.2025.9317
- U.S. Food and Drug Administration (FDA). High-intensity sweeteners permitted for use in food. Accessed March 2026. https://www.fda.gov/food/food-additives-petitions/aspartame-and-other-sweeteners-food
- von Rymon Lipinski, G. (1985). Acesulfame K. Food Chemistry, 18(3), 179–192.

