TASTEBENCH: Multimodal Benchmark for Sensory Prediction, from Molecules to Sustainable Foods

Anna Thomas [1 2] **Sohum Patnaik** [2] **Caroline Cotto** [3] **Benjamin Sanchez-Lengeling** [4]

Abstract

Sustainable protein discovery lacks the fast computational proxies, analogous to molecular docking or density functional theory, that accelerate drug and materials discovery. Evaluating whether a novel food tastes like its animal-based target requires expensive human sensory panels, bottlenecking the design-build-test loop. We introduce TASTEBENCH, a multimodal benchmark and privacy-preserving competition for sensory prediction, spanning two tasks: a food-level ranking task built on 21K+ human evaluations across 215 plant-based foods in 24 product categories, yielding 935 within-category ranking pairs, and a supporting molecular-level taste classification task over 15K flavor molecules. To enable rigorous interpretation of model performance, we characterize the ground truth: inter-rater agreement among panelists is low (Krippendorff's $\alpha = .077$), and the split-half reliability ceiling of panel-aggregated rankings is .825, establishing the range within which ML systems on this benchmark should be assessed. We evaluate baselines across four input modalities; on the same pairs panelists rated, the best model achieves .661 pairwise accuracy, competitive with the median individual panelist (.650). TASTEBENCH provides the evaluation infrastructure and baselines for measuring progress on computational screening for sustainable protein discovery.

1. Introduction

Drug and materials discovery have been transformed by fast computational proxies - molecular docking, density functional theory - that allow researchers to screen candidates *in silico* before committing to costly experiments (Fan et al., 2019; Jain et al., 2016). Sustainable protein development lacks any such proxy. Today, evaluating whether a plant-based formulation tastes like its animal-based target requires recruiting large panels of human tasters, a process that is slow, expensive, and inherently low-throughput. This bottleneck matters: diversifying protein sources away from animal agriculture is among the highest-leverage interventions for climate change mitigation (IPCC, 2022). Yet 46% of consumers cite poor taste as a barrier to repeat purchase, and large-scale benchmarking has documented persistent taste gaps between sustainable proteins and their animal-based counterparts (NECTAR, 2025).

The formulation space is enormous, as varying protein sources, binding agents, fats, flavors, and processing conditions creates a combinatorial design problem, but the evaluation bottleneck means that only a handful of candidates can be panel-tested per development cycle. Just as molecular docking ranks drug candidates by predicted binding affinity so that only top-scoring compounds advance to cell-based assays, a sensory predictor could rank candidate formulations by predicted similarity to the target animal product, focusing expensive panel testing on the most promising candidates. An AI system that could reliably perform this ranking would accelerate the design-build-test loop. More ambitiously, such a system could eventually participate in closed-loop formulation optimization, in which candidates are iteratively proposed and refined.

Prior work has begun to explore this direction. Thomas et al. (2025) evaluated several LLMs on sensory prediction tasks constructed from an earlier version of the NECTAR dataset, but considered only text-based inputs and did not attempt to improve upon foundation model performance. Separately, recent work on molecular-level taste and olfaction prediction has produced capable models for classifying individual compounds (Zimmermann et al., 2025; Lee et al., 2023), but no prior work has explored whether such molecular representations transfer to food-level sensory prediction. More broadly, no standardized benchmark exists for this task. Progress has been difficult to measure in part because sensory data carries commercial sensitivity: companies that participate in panel testing are reluctant to share results that could identify underperforming products. Building useful

[1]Stanford Computer Science [2]Food Intelligence Lab [3]NECTAR [4]University of Toronto Chemical Engineering. Correspondence to: Anna Thomas <thomasat@stanford.edu>.

Proceedings of the AI for Science Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

evaluation infrastructure for this domain therefore requires not only curating tasks and baselines, but also designing data-sharing mechanisms that preserve brand privacy. Our contributions include:

- **A multimodal, multiresolution benchmark and privacy-preserving competition for sensory prediction.** We introduce TASTEBENCH, spanning two tasks: a food-level ranking task (21K+ evaluations across 215 products, yielding 935 pairs) and a molecular-level taste classification task (15K+ molecules). The benchmark integrates four feature cost tiers - ingredient text, nutrition, molecular composition, and appearance - enabling systematic study of the cost-accuracy tradeoff. Both tasks are released as public Kaggle competitions, using decoy data to prevent brand-identity inference.
- **Ground truth characterization for ML evaluation.** We quantify the reliability of the human judgments against which models are scored. The median panelist achieves .650 pairwise ranking accuracy, inter-rater agreement is low ($\alpha = .077$), and the split-half reliability ceiling is .825. This establishes the interpretive range for model performance: scores near .650 are indistinguishable from a single panelist, while scores approaching .825 hit the noise floor of the ground truth. This confirms that panel-aggregated targets are the appropriate ground truth for ML evaluation in this domain.
- **Computational baselines and empirical findings.** We evaluate unsupervised distance methods, zero-shot foundation models, supervised pairwise models, and ensembles. On within-block pairs, the best model achieves .661 accuracy, competitive with the median panelist; on all within-category pairs, the best model achieves .683. Furthermore, swapping the compound encoder from Zimmermann et al. (2025)'s Flavor Analysis and Recognition Transformer (FART) to a GNN degrades food-level ranking in the best downstream models, suggesting that directional improvements on a component-level ML task do not strictly transfer to the system-level ML task the benchmark is designed to measure.

By establishing this seed benchmark, TASTEBENCH unlocks several critical research questions for the community:

- **Multimodal sensory integration.** How can models effectively fuse visual, nutritional, and structural chemical data to predict noisy, subjective human sensory responses?
- **The chemistry/semantic gap.** Why do text-based LLMs currently outperform dedicated molecular graphs in macroscopic food matrices? How can we build multi-sensory grounded models that capture missing modalities (like texture and olfaction) without merely relying on memorized semantic correlations?
- **Data augmentation in low-data regimes.** How can the community utilize this calibrated seed benchmark to bridge small, high-fidelity psychophysical datasets with massive, unstructured web data (e.g., recipes, photos and reviews) to accelerate discovery?

We further discuss related work in Appendix A.

2. Datasets

TASTEBENCH integrates four data sources to map molecular, text, nutritional, and visual features to human sensory outcomes while preserving privacy.

NECTAR. (NECTAR, 2025; 2026) consists of 271 sustainable protein products and animal-based benchmarks and at least 100 sensory ratings per product, along the dimensions of overall liking, similarity to the target animal product, flavor, texture, appearance, and purchase intent. It is the largest dataset to date on sensory properties of sustainable proteins. Here we use data published in NECTAR's 2025 Taste of the Industry report, consisting of 121 plant-based meats across 14 categories evaluated by 2,684 omnivores, and NECTAR's 2026 Taste of the Industry report, consisting of 94 dairy-free products across 10 categories evaluated by 2,183 omnivores. Available features include ingredient lists, nutritional information, and photos of each food item. The public interactive NECTAR data dashboard is here. To protect the privacy of participating brands, the specific product identities (e.g., brand names) have been de-identified, though the sensory targets, evaluation suite, and underlying product features needed to run the TASTEBENCH benchmark are fully public. Access to the unblinded product identities can be requested here. Our models were evaluated on 215 plant-based and vegetarian meat and dairy analogs across 24 categories. This creates 935 within-category ranking pairs.[1] We limited our selection to products with available animal-product similarity ratings. An additional 32 animal-based products were included as references for feature development. This resulted in a total of 247 products in our dataset. Data collection for 134 additional products will be completed by the end of 2026, and will continue in future years.

Taste Like. (NECTAR, 2024) is a directory of over 1,200 plant-based and fermentation-based products from 147 companies, with ingredient lists, nutrition facts, and allergen metadata. In TASTEBENCH, these products serve a struc-

[1]Three of the 935 pairs have tied panel-mean similarity scores; they are excluded from supervised training and scored as 0.5 in pairwise-accuracy evaluation.

tural role: they anonymize the NECTAR-evaluated products in the public Kaggle competition. Because NECTAR's TASTY Awards publicly identify top performers, releasing features for only NECTAR products would allow for inferring which products scored poorly. Including Taste Like products (which lack sensory ratings) prevents this inference. Taste Like products do not affect scoring.

FoodAtlas. (Youn et al., 2024) is a knowledge graph containing 285,077 triplets of three entity types (*food*, *part*, *chemical*) and four relation types (*contains*, *has part*, *is a*, *has child*) on three evidence quality levels (*high*, *medium*, *low*), with 4,318 of them evaluated by human experts. We use FoodAtlas to expand each ingredient on a NECTAR product label into a list of associated compounds together with their concentrations where reported; each compound's identifier is then mapped to a canonical SMILES which can then be featurized.

FartDB. (Zimmermann et al., 2025) (Flavor Analysis and Recognition Transformer)[2] is a publicly available collection of 15,025 canonicalized small molecules labeled with one of five taste classes: *sweet*, *bitter*, *sour*, *umami*, and *undefined*. The dataset is curated from six public sources:

- **FlavorDB** (Garg et al., 2018): a database of 25,595 flavor molecules aggregated from 936 natural ingredients across 34 food categories, with associated taste and aroma profiles.
- **PlantMolecularTasteDB** (Gradinaru et al., 2022): a database of 1,527 orosensorially active phytochemicals (*sweet*, *sour*, *bitter*, *salty*, *umami*, *pungent*, *astringent*), with their chemical and sensorial profile.
- **ChemTastesDB** (Rojas et al., 2022): a database of 2,944 molecular tastants categorized by the five basic tastes (*sweet*, *bitter*, *umami*, *sour*, and *salty*), or non-basic classes (*tasteless*, *non-sweet*, *multitaste* and *miscellaneous*).
- **Tas2R Agonists DB** (Bayer et al., 2021): a database of 133 food-derived bitter compounds with activity against human bitter taste receptors (TAS2Rs).
- **IUPAC Digitized Dissociation Constants** (Zheng, 2022): a database of "high-confidence" pKa data with 8,843 unique molecules used to identify sour compounds.
- **Umami Compounds** (Suess et al., 2015): a manually curated set of umami molecules extracted from a published review.

[2]We use the GitHub release at https://github.com/fart-lab/fart/tree/main/dataset; a HuggingFace mirror is available at https://huggingface.co/datasets/FartLabs/FartDB.

Molecules with multiple taste labels appear as separate entries. We use the original train/validation/test splits provided by the authors (10,517 / 2,254 / 2,254 molecules). The dataset is heavily skewed toward the sweet class, with umami severely underrepresented (53 training molecules, 6 test molecules). Our models are evaluated on the 2,254-molecule held-out test split for the molecular prediction task.

License terms and access restrictions for all datasets are documented in Appendix C.3. Croissant metadata for the NECTAR and Taste Like datasets, with Responsible AI fields, is available here.

3. Benchmark Design

We evaluate predictive capabilities across two distinct scales: a macroscopic food-level ranking task that simulates human sensory panels, and a microscopic molecular classification task to test the limits of chemical proxies.

Food-Level Similarity Prediction Task While NECTAR contains multiple sensory dimensions, here we focus on predicting similarity to the target animal product, a key driver of overall liking (Giezenaar et al., 2024). Specifically, we frame the problem as a within-category ranking task: given a set of plant-based products in the same category, rank them by their mean panelist-rated similarity to the animal-based counterpart. We use multiple metrics to evaluate model performance: (1) Pairwise ranking accuracy; (2) Spearman ρ and Kendall τ rank correlation (weighted by number of products per category); (3) Recall@k ($k = 1, 2, 3$): fraction of categories where the true best product is in the top k predictions (macro-averaged).

Our data used to evaluate sustainable proteins includes multiple levels of cost, visualized in Appendix Figure 3. Specifically, ingredient list text, nutrition facts, and molecular composition features do not require actually making the product. In contrast, generating appearance features does and is thus more costly. Finally, our predictions are compared against ratings from ground truth, large-scale sensory panels, the highest level of cost. A single university-run consumer panel can cost $30 per product per panelist plus administrative fees, with even a modest 20-person evaluation of three samples totaling over $2,100 (ACBS Foodtesting Services, 2023), while typical full-scale outsourced consumer tests range from $18,000 to $85,000 per project depending on panel size and product category (More, 2026).

Molecular-Level Taste Classification Task Given a canonical SMILES string, we predict whether the molecule is sweet, bitter, sour, umami, or undefined in this task. This is the same classification task performed by the chemical language model, FART (Flavor Analysis and Recog-

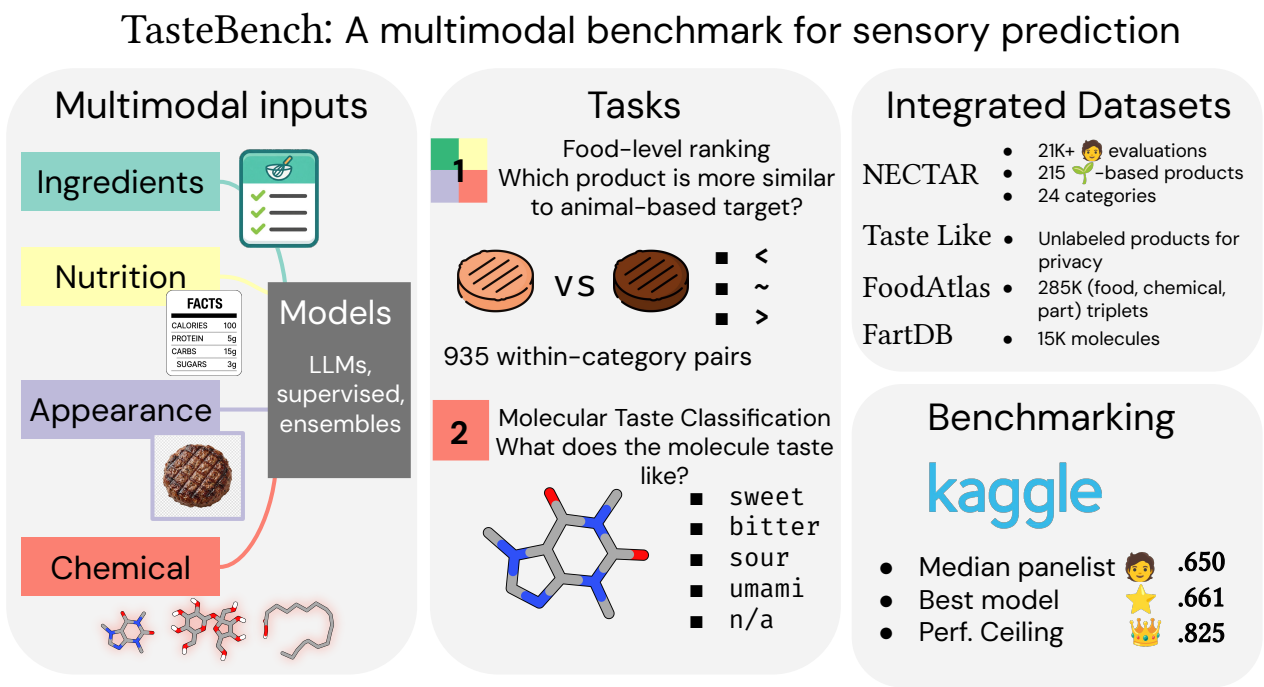

Figure 1. Overview of TASTEBENCH, a multimodal benchmark built on 21k+ human evaluations across 215 plant-based foods in 24 product categories, and 15K flavor molecules.

nition Transformer) (Zimmermann et al., 2025). We extend their contribution by evaluating a different model (D-MPNN; (Heid et al., 2023) we train on the same data splits to understand how embeddings generated from different molecular taste classification models with comparable performance transfer to food-level similarity predictions.

Privacy-Preserving Competition To publicly compare the performance of different approaches on the food-level similarity prediction task, we create a test-only public Kaggle competition. Input features for the sustainable proteins are publicly available, and performance on the hidden test set is computed once a competitor uploads their predictions. A key consideration in this setting is preserving the privacy of companies whose products performed poorly in the NECTAR sensory testing. Thus, solely including features of NECTAR products would, in conjunction with the results of the TASTY Awards (given to the top performing products) enable identification of the brands that didn't win a TASTY Award. We reduce this risk via also incorporating samples from the Taste Like dataset of sustainable proteins (NECTAR, 2024), thus anonymizing the NECTAR products. The Kaggle competition is available here. We discuss defenses against adversarial attacks in Appendix C.2. Similarly, to publicly compare the performance of different approaches on the molecular-level taste classification task, we create a fully public Kaggle competition, available here. This is the FartDB benchmark (Zimmermann et al., 2025), released on Kaggle as a public, transparent leaderboard.

Access Structure and Reproducibility TASTEBENCH separates benchmark use from brand-identifying information. Both Kaggle competitions are public: participants can download input features, submit predictions, and receive scores without access to product identities. We also release the code, trained model weights, and other files needed to reproduce all reported results. Released model weights, out-of-fold predictions, cached features, and reproduction scripts are inventoried in Appendix D. The only controlled-access component is the linkage between de-identified product codes and underlying NECTAR product records, including brand identities. This linkage is not required to use the benchmark, submit to either competition, evaluate predictions, or reproduce results from the released feature files. It is needed only to audit the de-identification procedure or rebuild the benchmark from raw NECTAR records. Because participating companies provided data under anonymity constraints, this linkage is restricted. For reviewer verification, we provide a private access link here to the full data required to reproduce the results in this paper.

4. Baseline Models

To predict food-level similarity, we extract features across four modalities: nutrition, appearance, ingredient text, and molecular composition. Because molecular chemistry is the most complex modality, we evaluate it as a standalone sub-task (Molecular Taste Classification) to determine if better chemical classifiers yield better downstream food-level representations.

4.1. Features & the Molecular Taste Sub-Task

Nutrition We extend the nutrition features of Thomas et al. (2025), who used normalized fat and sodium, to a set of six macronutrients: total fat, sodium, protein, dietary fiber, total carbohydrate, and total sugars. These features were selected based on literature linking the nutrient and sensory properties in each relevant product type (Appendix C.5). All values are normalized to per-100g serving and converted to grams. In our supervised models, all six columns are provided to every product and the model learns which are relevant per category; missing values are imputed with zero. For the unsupervised distance baselines, only the category-specific subset is used. LLM-based models receive the full nutrition facts panel as formatted text in the prompt.

Ingredient Text To generate a text embedding of the ingredients, we embed the full ingredient list as a single string using Qwen3-Embedding-0.6B. We prepend a task-specific instruction ("Given a food ingredient list, identify products with similar taste, texture, and sensory properties") following the model's instruct format. This produces a 1024-dimensional food-level embedding.

Category Subset We group the 24 NECTAR product categories into four subsets that share sensory-relevant macronutrients: *meat*, *nonsweet dairy*, *cheese*, and *sweet dairy* (full mapping in Appendix C.6). Per-category sample sizes are too small for category-specific modeling (as few as 5 products per category), so each product is encoded as one dimension of a 4-dimensional one-hot indicator over these subsets. In our supervised models, this lets the model learn product-type-specific coefficients without per-category overfitting; in our unsupervised distance baselines, the corresponding subset additionally selects which macronutrient features to use (Appendix C.5).

Appearance To generate an appearance embedding for each image, we use the DINOv3 vision transformer model (Siméoni et al., 2025). We extract the CLS token from the last hidden state to generate 1024-dimensional embeddings. 93% of analog products have an image available; for the rest, missing image embeddings are imputed from the k=5 nearest neighbors within the same category, where neighbor distance is computed via cosine similarity over the available (non-image) features.

Chemical Composition & Molecular Taste Classification Task We extract the molecular composition of each ingredient using the FoodAtlas knowledge graph, resolving compounds to canonical SMILES via PubChem or ChEBI. We evaluate two distinct methods for embedding these molecules. First, we default to extracting 768-dimensional embeddings from the CLS token of the FART chemical language model (Zimmermann et al., 2025). These compound-level embeddings are aggregated to the food level via a log-concentration-weighted mean (weights $w_i \propto \log(1 + c_i)$). While complex attention-based aggregators are common (Rajaonson et al., 2025), recent chemosensory benchmarks demonstrate that in high-noise psychophysical datasets (Satarifard et al., 2025), simple log-linear aggregations resist catastrophic overfitting and effectively mimic biological receptor competitive masking, aligning with the Weber-Fechner law (Dehaene, 2003). Ingredient-level embeddings are then averaged across the top three ingredients by FDA label weight.

To determine if improving molecular taste classification yields better downstream food-level rankings, we evaluate a standalone sub-task by training a directed message-passing network (D-MPNN; Heid et al. 2023) from scratch. We parse SMILES into graphs using Chemprop's default atom and bond featurizer, and train the D-MPNN on the same five-class taste classification task (sweet, bitter, sour, umami, undefined) as FART. The model mean-pools atom representations into a 300-dimensional graph embedding. Details are in Appendix C.8. This trained GNN serves two purposes: it establishes our molecular-task baseline (Table 4) and provides alternative 300-dimensional embeddings that we swap with FART to test cross-resolution transfer in our downstream product models (Table 5).

4.2. Food Similarity Models

After generating features, we use the models described below to predict the mean similarity rating for each plant-based protein to its animal-based counterpart.

Multi-modal Rank Fusion (MMRF). For each product category, we first compute the mean feature vector of animal-based reference products per modality (nutrition, compound, image). We then compute each plant-based product's distance to its category's reference centroid, normalize via within-category percentile ranks, and average ranks across modalities to produce a final similarity ranking within each category. We evaluate both cosine and Euclidean distance variants.

LLMs. We tested Gemini 3.1 Pro and Qwen 3.5 397B-A17B, leading closed-source and open-source multimodal foundation models, to understand how they perform with

zero-shot prompting. The prompt is in Appendix C.4. Ingredients, nutrition, and images are used as inputs in the prompts.

Supervised Models. We train supervised models via leave-one-product-out cross-validation (LOOCV) across all 215 plant-based products. All supervised models receive the same concatenated feature vector (product category subset, nutrition proportions, compound aggregations, text embeddings, and image embeddings) with per-modality standardization and dimensions reduced via PCA to 95% explained variance for the high-dimensional modalities (both fit on training data within each LOOCV fold). We evaluate five models: ridge regression; Bradley-Terry, which fits logistic regression on within-category feature-pair differences; Hierarchical Bradley-Terry, which fits per-category-subset models with empirical Bayes shrinkage toward the global coefficients; Kernel RankSVM with an RBF kernel on pair differences; and LightGBM with a pointwise MSE objective.

Ensembles. We combine the best individual model (Gemini 3.1 Pro) with the supervised model (Bradley-Terry) that yields the most improvement in accuracy; Appendix F traces a single product through this full pipeline end-to-end. A comparison of the performance improvement of each supervised model when combined with Gemini can be seen in Appendix E.3.

We use nested leave-one-out cross-validation to prevent label leakage: for each held-out product, inner-loop Bradley-Terry predictions are trained excluding both the held-out product and the evaluation partner. Our results report three combination strategies: non-negative least squares (NNLS), rank averaging, and arithmetic mean.

5. Experiments and Results

5.1. Ground truth reliability and human performance

To interpret model performance, we benchmark against human panelists, following (Lee et al., 2023). NECTAR recruits untrained consumer panelists using a balanced incomplete block design: each panelist rates a subset of products within a category per session, across two sessions on separate days. We compute each panelist's pairwise ranking accuracy against the leave-one-out panel mean, restricted to the products they rated. The median panelist achieves .650 accuracy (IQR .50–.75, $n = 627$), with low inter-rater agreement (Krippendorff's $\alpha = .077$). To compare on the same pair set, we evaluate our best model (described in Section 5.2) on within-block pairs, where it achieves .661, exceeding the median panelist by 1.1 pp; on all within-category pairs (Table 2), it achieves .683, matching approximately $k^* \approx 2$ panelists ($p < .001$). Figure 2 shows accuracy as a function of panel size and Table 1 compares

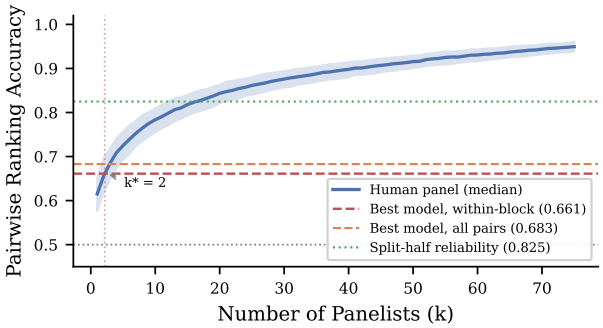

Figure 2. Pairwise accuracy vs. panel size on within-category ranking. The panel-mean curve crosses the within-block best-model line at $k^* \approx 2.2$.

model performance with discrete panel sizes. The split-half reliability of the panel mean ranking (.825) represents an approximate ceiling for model performance.

Table 1. Human panelist baseline for pairwise ranking accuracy. The models perform competitively with small human panels

Method	Pairwise Acc. [95% CI]
Evaluated on within-block pairs	
Individual panelist (median)	.650[†]
Panel of 3	.685 [.640,.725]
Panel of 5	.725 [.686,.760]
Best model (within-block)	.661 [.604,.716]
Evaluated on all within-category pairs	
Best model (all pairs)	.683 [.655,.740]

[†]IQR: 0.50–0.75 across 4,664 panelist-block observations.

5.2. Ensembles exceed median human similarity matching performance

The results for our computational baselines are shown in Table 2. We report 95% BCa bootstrap confidence intervals (10,000 resamples, stratified by category) (Efron, 1987). The best ensemble (NNLS over Bradley-Terry and Gemini 3.1 Pro) achieves .683 pairwise accuracy [.654, .740], exceeding both the strongest individual model (Gemini 3.1 Pro at .654) and the median untrained human panelist (.650); this corresponds to matching approximately $k^* \approx 2$ panelists (Section 5.1).

Among individual models, the leading LLM (Gemini 3.1 Pro, .654) and the strongest supervised model (Kernel RankSVM, .618) are within their CIs of each other, while distance baselines (.571–.574) and LightGBM (.558) lag substantially. NNLS exceeds either component alone (.683 vs. .654 / .610), indicating that the LLM and supervised model capture complementary signal; rank-averaging and arithmetic-mean variants recover most but not all of this gain (.671 and .623 respectively).

Table 2. **Ensembles of supervised models and LLMs deliver the strongest continuous ranking on all three metrics.** Pairwise accuracy and rank-correlation metrics on 215 NECTAR plant-based products (935 within-category pairs; LOOCV) with 95% BCa CIs. **Bold** = best in column; † = CI overlaps the leader (no significant difference at 95%). Top-k retrieval in Table 3.

Model	Pw. Acc.	ρ_S	τ_K
Gemini 3.1 Pro	.654† [.619,.713]	.370† [.250,.509]	.298† [.179,.409]
Qwen 3.5 397B	.630† [.591,.688]	.320† [.203,.459]	.255† [.141,.365]
MMRF (cosine)	.574 [.528,.635]	.212† [.077,.380]	.151† [.023,.278]
MMRF (L2)	.571 [.524,.632]	.202† [.064,.358]	.153† [.031,.277]
Ridge	.579 [.534,.632]	.160 [.028,.302]	.115† [−.005,.226]
Bradley–Terry	.610† [.571,.661]	.266† [.156,.402]	.178† [.064,.287]
Hierarchical BT	.613† [.570,.670]	.228† [.098,.383]	.165† [.043,.283]
Kernel RankSVM	.618† [.579,.674]	.288† [.168,.438]	.206† [.088,.321]
LightGBM	.558 [.509,.615]	.136 [−.003,.293]	.092† [−.032,.219]
NNLS (BT+Gemini)	**.683** [.654,.740]	**.417** [.314,.550]	**.322** [.211,.427]
Rank avg.	.671† [.641,.725]	.398† [.293,.537]	.304† [.192,.409]
Mean	.623† [.587,.675]	.310† [.201,.443]	.216† [.102,.321]
Random (analytical)	.500	.000	.000

Table 3. **Recall metrics indicate the practical utility of these models for recommending the best formulations.** R@k is the fraction of categories whose top-truth product is among the model's top k predictions. Random rates are macro-averaged $\min(k, n_c)/n_c$ over 24 categories. **Bold** = best in column; † = CI overlaps the leader. Continuous ranking metrics in Table 2.

Model	R@1	R@2	R@3
Gemini 3.1 Pro	.271† [.167,.424]	**.514** [.476,.701]	.578† [.491,.786]
Qwen 3.5 397B	.229† [.132,.354]	.354† [.215,.483]	.500† [.375,.656]
MMRF (cosine)	.208† [.104,.323]	.292 [.135,.396]	.333 [.181,.379]
MMRF (L2)	.222† [.104,.335]	.403† [.285,.576]	.542† [.438,.729]
Ridge	.250† [.181,.435]	.292 [.118,.396]	.542† [.493,.749]
Bradley–Terry	.250† [.181,.422]	.292 [.125,.396]	.375 [.187,.458]
Hierarchical BT	**.333** [.271,.526]	.500† [.424,.742]	**.667** [.622,.854]
Kernel RankSVM	.167 [.040,.250]	.292 [.115,.396]	.417 [.196,.521]
LightGBM	.292† [.250,.438]	.333† [.208,.491]	.458† [.333,.646]
NNLS (BT+Gemini)	.292† [.188,.451]	.500† [.424,.693]	.542† [.385,.688]
Rank avg.	.264† [.182,.427]	.361† [.211,.486]	.521† [.375,.677]
Mean	.167 [.035,.250]	.375† [.260,.549]	.417 [.219,.500]
Random (analytical)	.121	.242	.363

Recall metrics indicate the practical utility of these models for recommending the best formulations (Table 3). Hierarchical Bradley-Terry recovers the true best product within the top-3 predictions in 67% of categories (R@3 = .667) and it predicts the true top product in 33% of categories (R@1 = .333, versus .121 for random). Per-modality and per-feature ablations are reported in Appendix E.1.

These results are promising because formal sensory panels are rare for sustainable protein startups. While sensory best practices recommend companies conduct frequent trained panels (8-12 participants) and at least one consumer panel (80-100+ participants) during the product development lifecycle (Nolden et al., 2025), many resource-constrained brands rely only on feedback from 3-5 untrained, internal participants. While this can provide some directional insights for early-stage prototyping, this approach is often insufficient for surfacing robust consumer insights that drive strong commercial performance. Thus, given our best model is competitive with the median untrained panelist, it could provide a low-cost signal where infrastructure for formal panels is unavailable.

5.3. A directional improvement in molecular classification does not guarantee better food ranking

On the held-out FART test split (2,254 molecules), the GNN achieves an accuracy of .904 [.890, .915] compared to FART's .894 [.880, .906]. This 1 pp gap has overlapping confidence intervals and McNemar's test (p=0.07) does not reach statistical significance, suggesting that any accuracy advantage of the GNN is a directional improvement, but it is at best marginal. The AUROC for both models is essentially tied (.978 vs. .974). Macro F1 shows a larger gap (.818 vs. .774), but the umami class alone accounts for Δ=+.167 of this (n=6), leaving only ∼1.3 pp across the remaining four classes. The GNN performs worse than FART on precision (GNN .825 vs. FART .879), but better on recall (GNN .813 vs. FART .739) due to the same umami trade-off: FART predicts umami conservatively (precision 1.000, recall .333) while the GNN predicts it more freely (precision .667, recall .667). Despite comparable or slightly better molecular-task performance, the GNN yields lower pairwise accuracy across the best downstream models.

Table 4. **The trained GNN matches FART's overall accuracy but exhibits a distinct precision-recall tradeoff on umami.** Taste classification on the 2,254-molecule FART test split. The "FART augmented + confidence" test-time augmentation variant is excluded for single-checkpoint comparability. fp = Morgan fingerprints; desc = molecular descriptors. **Bold** = best in column; † = CI overlaps leader. McNemar's test on accuracy gap: $\chi^2 = 3.20, p = 0.07$.

Model	Accuracy	Precision	Recall	F1	AUROC
Reported by Zimmermann et al. (2025)					
XGBoost (fp)	.899	.817	.740	.766	.852
XGBoost (fp+desc)	.896	**.884**	.740	.778	.851
Balanced RF (fp)	.797	.585	.732	.601	.839
FART	.894† [.880,.906]	.879 [.675,.897]	.739† [.668,.869]	.774† [.671,.875]	.974† [.968,.979]
This work					
GNN	**.904** [.890,.915]	.825 [.697,.895]	**.813** [.683,.883]	**.818** [.732,.876]	**.978** [.968,.982]

Table 5. **Swapping FART for GNN embeddings degrades product-level ranking.** Pairwise accuracy on 215 products (LOOCV) replacing FART with GNN compound blocks for non-LLM models. Δ = GNN − FART. † = CI overlaps leader.

Model	C = FART	C = GNN	Δ
Unsupervised			
MMRF (cosine)	.574 [.528,.635]	.545† [.494,.602]	−.029
MMRF (L2)	.571 [.524,.632]	.556† [.509,.612]	−.014
Supervised – linear			
Ridge	.579 [.534,.632]	.586† [.539,.644]	+.006
Supervised – pairwise			
Bradley–Terry	.610† [.571,.661]	.616† [.580,.669]	+.005
Hierarchical BT	.613† [.570,.670]	.624† [.581,.684]	+.011
Kernel RankSVM	.618† [.579,.674]	.597† [.556,.654]	−.020
Supervised – nonlinear			
LightGBM	.558 [.509,.615]	.559† [.509,.617]	+.001
Ensemble (BT + Gemini)			
NNLS	**.683** [.654,.740]	**.635** [.594,.692]	−.048
Rank average	.671† [.641,.725]	.626† [.584,.682]	−.044
Mean	.623† [.587,.675]	.553† [.504,.609]	−.070

Transferring molecular representations to food-level ranking. To determine if comparable molecular-task performance translates to macroscopic sensory predictions, we substitute the FART compound embeddings with our trained GNN embeddings across all non-LLM models. As shown in Table 5, using GNN embeddings directionally reduces pairwise accuracy in the best models. The gaps are largest in the ensemble (Δ=−.048 for NNLS, −.070 for Mean). A better molecular encoder does not guarantee better product-level ranking. This supports our use of FART over the GNN in the main pipeline, but more broadly suggests that molecular taste classification may be a limited proxy for the complex compound interactions that drive product-level preferences.

6. Discussion

TASTEBENCH provides evaluation infrastructure for predicting how sustainable proteins will perform in sensory evaluation. Our baselines establish that current models can exceed the median untrained panelist on pairwise ranking, while the reliability ceiling (.825) provides an upper bound for future progress. Even at current accuracy, these models are practically useful as screening tools. Under a simplified selection model (Appendix B), using the NNLS ensemble to select the top 5% of burger candidates yields an estimated 32% improvement in mean sensory similarity over random selection. The best models also recover the true top product within their top-3 predictions in a majority of categories, suggesting utility for triaging candidates before expensive panel testing. In our molecular prediction experiments, a GNN encoder matching FART's classification accuracy degrades food-level ranking, suggesting that improving molecular classifiers alone will not close the performance gap.

The modality gap and data contamination. Our baselines reveal that zero-shot multi-modal LLMs can outperform dedicated molecular representations. However, this advantage might stem from "data contamination": LLMs having ingested public CPG databases and recipe co-occurrences during pre-training. Because text and images cannot fully capture physical texture, taste or olfaction, advancing chemical mixture models remains essential to overcome this data scarcity and discover truly out-of-distribution formulations.

A seed benchmark for the community. Given the cost of gathering human sensory panels, TASTEBENCH is designed as a foundational "seed benchmark." By rigorously establishing the human reliability ceiling, it provides a calibrated framework to measure algorithmic progress. This setup explicitly invites the community to augment these high-fidelity tasks with large-scale, unstructured web data (e.g., consumer reviews, scraped formulations) to further close the gap to human perception.

Limitations. The benchmark covers meat and dairy only, though will be updated as NECTAR expands to eggs and seafood. Per-category sample sizes are small (median 36 pairs, but as few as 10), producing wide category-level confidence intervals; primary metrics are computed across all 935 within-category pairs. Our evaluation focuses on the similarity dimension, and other sensory dimensions may require different modeling approaches. We have not yet established an expert food scientist baseline or explored prompt engineering for the foundation models.

Impact Statement

Sensory prediction could accelerate sustainable protein development, with downstream benefits for emissions, land use, and water use. However, biased models could narrow the formulation design space, and overreliance on screening before models are sufficiently reliable could lead to premature elimination of promising candidates. Models trained on American omnivore panels may systematically under-rank products formulated for non-Western palates; additional data collection on diverse populations is planned.

References

ACBS Foodtesting Services. Sensory evaluation, 2023. URL https://foodtesting.arizona.edu/sensory-eval. Department of Animal and Comparative Biomedical Sciences. Accessed May 2026.

Bakhsh, A., Lee, S.-J., Lee, E.-Y., Sabikun, N., Hwang, Y.-H., and Joo, S.-T. A novel approach for tuning the physicochemical, textural, and sensory characteristics of plant-based meat analogs with different levels of methylcellulose concentration. *Foods*, 10(3):560, 2021. doi: 10.3390/foods10030560.

Bayer, S., Mayer, A. I., Borgonovo, G., Morini, G., Di Pizio, A., and Bassoli, A. Chemoinformatics view on bitter taste receptor agonists in food. *Journal of Agricultural and Food Chemistry*, 69(46):13916–13924, 2021. doi: 10.1021/acs.jafc.1c05057. URL https://doi.org/10.1021/acs.jafc.1c05057.

D'Andrea, A. E., Kinchla, A. J., and Nolden, A. A. A comparison of the nutritional profile and nutrient density of commercially available plant-based and dairy yogurts in the United States. *Frontiers in Nutrition*, 10:1195045, 2023. doi: 10.3389/fnut.2023.1195045.

Dehaene, S. The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, 7(4):145–147, 2003. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(03)00055-X. URL https://www.sciencedirect.com/science/article/pii/S136466130300055X.

Efron, B. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

Fan, J., Fu, A., and Zhang, L. Progress in molecular docking. *Quantitative Biology*, 7(2):83–89, 2019.

Feng, D., Dai, W., Li, C., Pernigo, A., and Liang, P. P. Smellnet: A large-scale dataset for real-world smell recognition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aUnheo6zFD.

Ganesan, B., Brown, K., Irish, D. A., Brothersen, C., and McMahon, D. J. Manufacture and sensory analysis of reduced- and low-sodium Cheddar and Mozzarella cheeses. *Journal of Dairy Science*, 97(4):1970–1982, 2014. doi: 10.3168/jds.2013-7443.

Garg, N., Sethupathy, A., Tuwani, R., NK, R., Dokania, S., Iyer, A., Gupta, A., Agrawal, S., Singh, N., Shukla, S., Kathuria, K., Badhwar, R., Kanji, R., Jain, A., Kaur, A., Nagpal, R., and Bagler, G. FlavorDB: A database of flavor molecules. *Nucleic Acids Research*, 46(D1): D1210–D1216, 2018. doi: 10.1093/nar/gkx957. URL https://doi.org/10.1093/nar/gkx957.

Giezenaar, C., Orr, R. E., Godfrey, A. J. R., Maggs, R., Foster, M., and Hort, J. Profiling the novel plant-based meat alternative category: Consumer affective and sensory response in the context of perceived similarity to meat. *Food Research International*, 188:114465, 2024.

Godschalk-Broers, L., Sala, G., and Scholten, E. Meat analogues: Relating structure to texture and sensory perception. *Foods*, 11(15):2227, 2022. doi: 10.3390/foods11152227.

Gradinaru, T.-C., Petran, M., Dragos, D., and Gilca, M. PlantMolecularTasteDB: A database of taste active phytochemicals. *Frontiers in Pharmacology*, 12:751712, 2022. doi: 10.3389/fphar.2021.751712. URL https://doi.org/10.3389/fphar.2021.751712.

Grasso, N., Roos, Y. H., Crowley, S. V., Arendt, E. K., and O'Mahony, J. A. Composition and physicochemical properties of commercial plant-based block-style products as alternatives to cheese. *Future Foods*, 4:100048, 2021. doi: 10.1016/j.fufo.2021.100048.

Guinard, J.-X., Zoumas-Morse, C., Mori, L., Uatoni, B., Panyam, D., and Kilara, A. Sugar and fat effects on sensory properties of ice cream. *Journal of Food Science*, 62(5):1087–1094, 1997. doi: https://doi.org/10.1111/j.1365-2621.1997.tb15044.x. URL https://ift.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2621.1997.tb15044.x.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. Chemprop: a machine learning package for chemical property prediction. *Journal of chemical information and modeling*, 64(1):9–17, 2023.

IPCC. Climate change and land: IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. Technical report, Intergovernmental Panel on Climate Change, 2022.

Jain, A., Shin, Y., and Persson, K. A. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials*, 1(1):1–13, 2016.

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., et al. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381 (6661):999–1006, 2023.

Liem, D. G., Miremadi, F., and Keast, R. S. J. Reducing sodium in foods: The effect on flavor. *Nutrients*, 3(6): 694–711, 2011. doi: 10.3390/nu3060694.

Maddala, G. S. *Limited-dependent and qualitative variables in econometrics*. Econometric Society Monographs. Cambridge University Press, 1983.

Magwere, A. A., Keast, R., Gamlath, S., Nandorfy, D. E., Pematilleke, N., and Gambetta, J. M. A comparative study of the sensory and physicochemical properties of cow milk and plant-based milk alternatives. *Journal of Food Science*, 90(7):e70370, 2025. doi: 10.1111/1750-3841.70370.

More, A. B. Food sensory analysis service market research report 2034: Segments by service type, application, end-user, and region – global industry analysis, growth, share, size, trends, and forecast 2025–2034. Market Research Report FB-584161, Dataintelo, April 2026. URL https://dataintelo.com/report/food-sensory-analysis-service-market. 272 pages.

NECTAR. NECTAR Acquires Taste Like's Comprehensive Data on Alternative Protein Products. https://www.nectar.org/news/taste-like-data, 2024.

NECTAR. Taste of the industry 2025. https://www.nectar.org/sensory-research/2025-taste-of-the-industry, 2025.

NECTAR. Taste of the industry 2026. https://www.nectar.org/sensory-research/2026-taste-of-the-industry, 2026.

Nolden, A., Bomkamp, C., Kirchner, J., and Kogar, N. M. Sensory evaluation of alternative proteins: A best practices guide. Technical report, Good Food Institute, Washington, D.C., 2025. URL https://doi.org/10.62468/novg4853.

Rajaonson, E. M., Rajabi Kochi, M., Mejia Mendoza, L. M., Moosavi, M., and Sanchez-Lengeling, B. Chemixhub: Datasets and benchmarks for chemical mixture property prediction. In Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., and Chen, N. (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2f803abdcad9de35b45d5a656dade45c-Paper-Datasets_and_Benchmarks_Track.pdf.

Rojas, C., Ballabio, D., Pacheco Sarmiento, K., Pacheco Jaramillo, E., Mendoza, M., and García, F. Chemtastesdb: A curated database of molecular tastants. *Food Chemistry: Molecular Sciences*, 4:100090, 2022. ISSN 2666-5662. doi: https://doi.org/10.1016/j.fochms.2022.100090. URL https://www.sciencedirect.com/science/article/pii/S2666566222000181.

Rune, C. J. B., Clausen, M. P., and Giacalone, D. Sensory evaluation of plant-based cheese: a systematic review with a focus on texture and mouthfeel. *Critical Reviews in Food Science and Nutrition*, 66(4):754–779, 2026. doi: 10.1080/10408398.2025.2531220.

Satarifard, V., Sisson, L., Han, Y., Ilídio, P., Hladiš, M., Lalis, M., Song, X., Yin, W., Ravia, A., Zheng, C. X., Andreoletti, G., Albrecht, J., Pellegrino, R., Wang, Z., Yang, S., D'hondt, R., Ghinis, A., de Boer, J., Nakano, F. K., Gharahighehi, A., DREAM Olfactory Mixtures Prediction Consortium, Sanchez-Lengeling, B., Keller, A., Vosshall, L. B., Fiorucci, S., Tewari, A., Topin, J., Vens, C., Björkman, M., Kragic, D., Sobel, N., Christakis, N. A., Mainland, J. D., and Meyer, P. High-fidelity tuning of olfactory mixture distances in the perceptual space of smell through a community effort. *bioRxiv*, December 2025. doi: https://doi.org/10.64898/2025.12.13.694160.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

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

Thomas, A., Yee, A., Mayne, A., Mathur, M. B., Jurafsky, D., and Gligorić, K. What can large language models do for sustainable food? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=f6SFHNfuMu.

Tom, G., Ser, C. T., Rajaonson, E. M., Lo, S., Park, H. S., Lee, B. K., and Sanchez-Lengeling, B. From molecules to mixtures: Learning representations of olfactory mixture similarity using inductive biases. *arXiv preprint arXiv:2501.16271*, 2025.

Youn, J., Li, F., Simmons, G., Kim, S., and Tagkopoulos, I. Foodatlas: Automated knowledge extraction of food and chemicals from literature. *Computers in Biology and Medicine*, 181:109072, 2024. doi: 10.1016/j.compbiomed.2024.109072.

Zheng, J. IUPAC/Dissociation-Constants: v1.0 (v1-0_initial-release). Zenodo, 2022. URL https://doi.org/10.5281/zenodo.7236453.

Zimmermann, Y., Sieben, L., Seng, H., Pestlin, P., and Görlich, F. A chemical language model for molecular taste prediction. *npj Science of Food*, 9(1):122, 2025.

A. Related Work

Taste Prediction. Zimmermann et al. (2025) introduced FART, a chemical language model fine-tuned on 15K molecular tastants for multi-class taste prediction across sweet, bitter, sour, and umami categories, achieving over 91% accuracy. Prior to FART, molecular taste prediction relied on binary classifiers specialized to individual taste classes, using representations such as Morgan fingerprints or molecular descriptors paired with tree-based models. Thomas et al. (2025) evaluated several LLMs on sensory prediction tasks constructed from an earlier version of the NECTAR dataset, finding that LLMs can exceed human food scientist performance on experimental design for plant-based meats as judged by expert evaluators. However, that work considered only text-based inputs and did not attempt to improve upon foundation model performance with learned representations. TASTEBENCH extends both lines of work: it uses FART's molecular representations as one feature modality while introducing a food-level ranking task that connects molecular-level predictions to product-level sensory outcomes, a link that no prior work has established.

Olfactory Prediction. Computational olfaction provides the closest methodological precedent for our work. Lee et al. (2023) trained a graph neural network to construct a principal odor map (POM) that predicted odor qualities of novel odorants from molecular structure; on a prospective validation set of 400 molecules, the model matched the panel mean more closely than the median human panelist, establishing the benchmarking paradigm we adopt for our human performance baseline (Section 5.1). Tom et al. (2025) extended POM to molecular mixtures with POMMix, using attention mechanisms to aggregate single-molecule GNN embeddings into mixture representations. This molecule-to-mixture aggregation problem is structurally analogous to our compound-to-food aggregation pipeline (Section 4.1), though we operate in the taste rather than olfactory domain and aggregate across ingredients rather than mixture components. Feng et al. (2026) introduced SmellNet, a large-scale sensor-based benchmark for real-world odor classification with 180K time steps across 50 substances, providing a benchmarking template for chemosensory perception that complements our label-based approach.

B. Theoretical Analysis

A natural question is: how good must the sensory prediction model be in order to be useful? In a simplified model, we can use properties of the truncated bivariate normal to begin to understand the relationship between performance of the predictive model and downstream product improvements. Let S be the property of interest, e.g. sensory similarity to the target animal product, and let P be the predicted value, with $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$, and $P \sim \mathcal{N}(\mu_P, \sigma_P^2)$.

The quantity we want to understand is $\frac{E[S|P>T]-E[S]}{E[S]} = \frac{E[S|P>T]-\mu_S}{\mu_S}$: the percent improvement in the target attribute when using our predictive model to select a subset of candidate formulations.

Proposition B.1. $\frac{E[S|P>T]-\mu_S}{\mu_S} = \frac{\sigma_S \rho}{\mu_S} \frac{\phi(\frac{T-\mu_P}{\sigma_P})}{1-\Phi(\frac{T-\mu_P}{\sigma_P})}$, *where $\Phi(x)$ is the standard normal CDF, $\phi(x)$ is the standard normal PDF, and ρ is the correlation between S and P.*

Proof. The proposition follows from standardizing S and P and applying a result from (Maddala, 1983) that $E[X|Y > T] = \rho \frac{\phi(T)}{1-\Phi(T)}$ for standard bivariate normals X and Y with correlation ρ. □

For example, in the burgers category the $\frac{\sigma_S}{\mu_S}$ ratio was 0.19, though this is the variance of products on the market rather than an internal candidate pool. With a correlation between the predictor and the true property of 0.81 (both values from the NNLS ensemble; see Table 12 for per-category inputs), and taking the top 5% of candidates, this translates to a percent improvement of 32% in the target property when using the predictive model. This increases linearly as the variance of the pool of candidates and the correlation between predictive model and true property increase. This number is illustrative; the realized improvement on a real internal pool depends on its variance, which we cannot observe.

C. Experimental Details

C.1. Cost Hierarchy

Figure 3 visualizes the hierarchy of cost in sustainable protein evaluation. We study whether lower-cost features can predict the gold standard evaluation of a large-scale, representative sensory panel. By benchmarking lower-cost computational proxies (e.g., zero-shot LLMs) against the highest-cost human ground truth, TASTEBENCH assesses the viability of these novel evaluation methodologies.

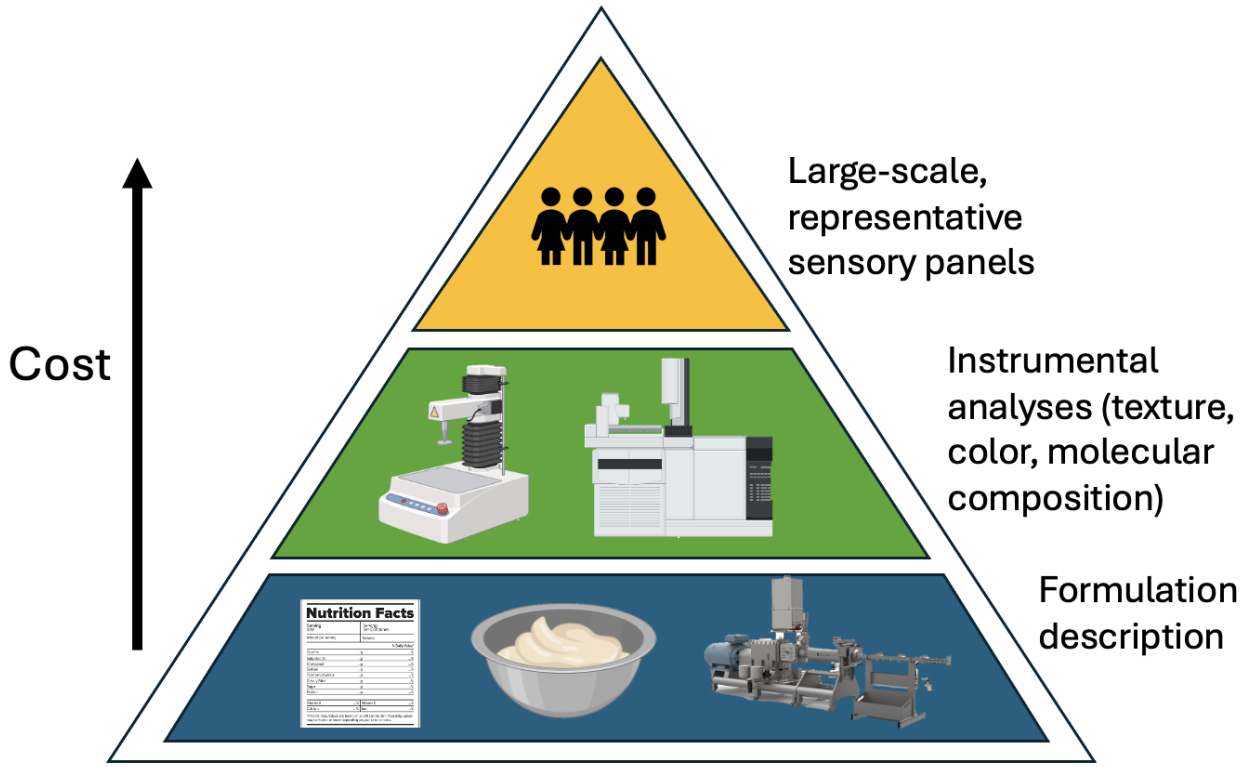

Figure 3. Hierarchy of cost in sustainable protein evaluation.

C.2. Privacy-Preserving Kaggle Competition

The most harmful attack is fine-grained pairwise inference: an attacker submits rankings that swap two specific products and reads the panel's ordering from the resulting leaderboard delta of $\pm\frac{1}{935} \sim 0.001$. Two layers of defense raise the cost of this attack. First, leaderboard scores are rounded to two decimal places, an order of magnitude coarser than the single-pair signal, and participants are restricted to five submissions per day. Second, Kaggle's public/private leaderboard split places 11% of pairs in a private holdout that is not probeable during the competition and is frozen afterward, so an attacker cannot adaptively target a specific brand's product.

C.3. Dataset Licenses and Terms of Use

- **NECTAR sensory ratings, ingredient/nutrition CSVs, and product images.** Licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) by NECTAR, with access gated under NECTAR's NDA. We provide a private access link here for reviewer verification. Other researchers can request access via the Google Form linked in Section 2.
- **Taste Like CPG product directory.** This dataset has been acquired by NECTAR and can be accessed by the same private access link here and requested via the Google Form linked in Section 2.
- **FoodAtlas v**4.0**.** Licensed under the Apache License 2.0.
- **FartDB (FART train/val/test splits).** Licensed under the MIT License. Pinned to commit `bde90e6562ce5d248e76af791fab29ffc9ae901b` of `https://github.com/fart-lab/fart`; vendored at `molecular/data/splits/`. The upstream project requests citation of Zimmermann et al. (2025) for any work using the dataset.

C.4. LLM prompt

System prompt

```
<role>
You are an expert food scientist.
Your task is to predict which of two
plant-based products would be perceived
as more similar to its animal-based
counterpart in a blind taste test.
</role>

<output_format>
Respond with ONLY the number 1 or 2.
No other text.
</output_format>
```

User prompt (all modalities)

```
<context>
Two plant-based {category} products
are compared for similarity to their
animal-based {category} counterpart.
Consider ingredients, nutrition facts,
and visual appearance.
</context>

<data>
<product id="1">
<ingredients order="descending_by_weight">
{ingredient list 1}
</ingredients>
<nutrition unit="per_100g">
{nutrition facts 1}
</nutrition>
</product>

<product id="2">
(same structure)
</product>
</data>

<images>
The first image shows Product 1 and
the second image shows Product 2.
</images>

<task>
In a blind taste test, which product
would omnivores rank as more similar
to the animal-based {category}?
</task>
```

Figure 4. Prompt template for LLM-based models. Feature ablation removes the corresponding XML blocks.

C.5. Nutrition feature selection

We include only nutrition label features where nutritional distance is a valid proxy for sensory distance, selecting subsets based on the dominant sensory drivers in each product type (Table 6).

Table 6. Nutrition features by category group. All values normalized to per-100g serving.

Group	Features
Meat (14 cat.)	Total Fat, Sodium, Protein, Dietary Fiber
Non-Sweet Dairy (6 cat.)	Total Fat, Sodium
Cheese (2 cat.)	Total Fat, Sodium, Total Carbs
Sweet Dairy (2 cat.)	Total Fat, Total Sugars

Feature selections are supported by: total fat, sodium, protein, and dietary fiber in meat analogs (Thomas et al., 2025; Godschalk-Broers et al., 2022; Bakhsh et al., 2021); fat and sodium in non-sweet dairy (Magwere et al., 2025; Liem et al., 2011); total fat, sodium, and total carbohydrate in cheese (Rune et al., 2026; Ganesan et al., 2014; Grasso et al., 2021); and total fat and total sugars in sweet dairy (Guinard et al., 1997; D'Andrea et al., 2023).

C.6. Category Subsets

Per-category sample sizes in NECTAR are too small for category-specific modeling (as few as 5 products per category, mean ≈ 9). We therefore group the 24 categories into the same four subsets used for nutrition feature selection (Appendix C.5):

Meat **(n=121):** bacon, bratwurst, breakfast sausages, burgers, breaded chicken filet, chicken strips, deli ham, deli turkey, hot dogs, meatballs, nuggets, pulled pork, steak, unbreaded chicken breast.

Nonsweet dairy **(n=62):** barista milk, butter, cream cheese, creamer, milk, sour cream.

Cheese **(n=16):** cheddar cheese, mozzarella.

Sweet dairy **(n=16):** ice cream (hard serve), yogurt.

Each product is encoded as one dimension of a 4-dimensional one-hot indicator over these subsets for the supervised models. Total: 215 plant-based products.

C.7. Food-Level Model Training Details

Featurization. See Section 4.1 for the five feature blocks (**S** category, **N** nutrition, **C** compound, **T** ingredient text, **I** image), the FDA-rank top-3 ingredient aggregation, and the log-concentration weighting motivated by the Weber–Fechner law (ablation in Table 13).

Preprocessing. Per-modality `StandardScaler` fit on the training fold; high-dimensional modalities (C, T, I) additionally PCA-reduced to 95% explained variance, fit within each LOOCV fold to prevent train-time leakage. Missing image embeddings are imputed by the k=5 nearest neighbors within the same category over the available non-image features; missing nutrition values are zero-imputed.

LOOCV protocol. 215 folds, one plant-based product held out per fold. For pairwise models (Bradley–Terry, hierarchical BT, Kernel RankSVM), training pairs are constructed from within-category preference comparisons of training-fold products only; the held-out product never appears on either side of any training pair. For pointwise models (Ridge, LightGBM), targets are mean panelist similarity ratings. Predictions across folds are concatenated into a single per-product OOF prediction file at `food_similarity/results/oof_predictions/{model}_S{subset}.csv`, which is the input to all paper render scripts.

Hyperparameters. All models use random seed 42 where applicable. Defaults below are committed in `food_similarity/models/` and were not swept; this is intentional to keep the model comparison apples-to-apples and avoid the appearance of per-model tuning advantage.

- **Ridge.** `sklearn.linear_model.Ridge`, $\alpha = 1.0$, `fit_intercept=True`.
- **Bradley–Terry.** `sklearn.linear_model.LogisticRegression` on within-category feature-pair differences. $C = 1.0$, `max_iter`= 1000.
- **Hierarchical Bradley–Terry.** Per-subset BT for *meat*, *nonsweet dairy*, and *other dairy* (the latter merging cheese and sweet dairy, which have 2 categories each — too few for stable group-level estimation); empirical Bayes shrinkage toward the global coefficients. Shrinkage strength $\lambda_s = \alpha/(\alpha + n_{\text{pairs},s})$ with α set to the median of $\{n_{\text{pairs},s}\}$, so a median-sized subset receives 50% shrinkage.
- **Kernel RankSVM.** `sklearn.svm.SVC` on pairwise difference features with RBF kernel, $C = 1.0$, $\gamma =$ `"scale"`, `max_iter`= 5000.

- **LightGBM.** Regression objective (RMSE), `n_estimators=` 100, `learning_rate=` 0.1, `num_leaves=` 31, `min_child_samples=` 5. `deterministic=True`, `force_row_wise=True`, `num_threads=1` for byte-identical reproduction.

Distance baselines (MMRF). For each category, compute the mean feature vector of animal-based reference products per modality. Score each plant-based product by its distance to that centroid (cosine or Euclidean), rank-normalize within category per modality, and average ranks across enabled modalities. No training, no fold structure.

Zero-shot LLMs. Prompt template, model versions (Gemini 3.1 Pro, Qwen 3.5 397B-A17B), and modality combinations are documented in Appendix C.4. Pair ordering is deterministic via MD5 hash to neutralize position bias; pairwise judgments are aggregated to a per-product score by win rate (number of pairs won divided by number of pairs participated in — equivalent to Copeland's method), and within-category ranks are computed from these scores.

Ensembles. Nested LOOCV combines Bradley–Terry (over Category Subset, Nutrition, Compound, Ingredient Text, Image features, abbreviated SNCTI) with Gemini 3.1 Pro. For each outer-fold held-out product i, the inner-loop BT predictions for the evaluation partners j are produced from a BT trained on $\{k : k \neq i,\ k \neq j\}$ (213 products), preventing partner-level label leakage. We report three combination strategies fit per fold: non-negative least squares (`scipy.optimize.nnls`), arithmetic mean, and rank averaging.

Confidence intervals. BCa bootstrap (Efron, 1987), $n_{\text{boot}} = 10{,}000$, stratified by category, seed 42. CIs are computed once per (model, metric, scope) and cached at `food_similarity/results/cis_*.csv` so table re-rendering does not re-run the bootstrap.

C.8. GNN Training Details

Featurization. Each molecule's canonicalized SMILES is parsed with RDKit and featurized using Chemprop's default `SimpleMoleculeMolGraphFeaturizer`. Atom features (72 dimensions): atomic number, total degree, formal charge, chirality tag, hydrogen count, hybridization, aromaticity, atomic mass. Bond features (14 dimensions): bond type, conjugation, ring membership, stereochemistry.

Architecture. Directed bond-level message passing (`BondMessagePassing`) with hidden dimension 300 and batch normalization, mean aggregation over atoms, and a single-hidden-layer classification feed-forward network (also hidden dimension 300).

Training. Adam optimizer with a Noam-like learning rate schedule (initial 10^{-4}, peak 10^{-3}, final 10^{-4}). Batch size 50, maximum 50 epochs, cross-entropy loss (optionally class-weighted, see hyperparameter grid below), early stopping on validation loss with patience 10. Within-run checkpoint selection: minimum validation loss. Random seed: 42.

Hyperparameter grid. The 12 configurations are the Cartesian product of depth $\in \{3, 5\}$, dropout $\in \{0, 0.2\}$, and class weighting $\in$ {none, inverse-frequency, sqrt-inverse-frequency}. Across-grid selection: maximum validation macro-F1, committed before evaluating on the FART test set or any NECTAR data. Validation macro-F1 for all 12 configurations is reported in Table 7; the selected configuration leads the runner-up by 0.8 pp.

Table 7. Validation macro-F1 for all 12 configurations in the GNN hyperparameter grid (depth × dropout × class weighting). The selected configuration (**bold**) is the highest val macro-F1; this checkpoint is used in Tables 4 and 5. Selection committed before evaluating on the FART test set or any NECTAR data.

Depth	Dropout	Class weighting	Val macro-F1
5	**0.2**	**none**	**.825**
5	0	none	.817
5	0.2	sqrt-inverse-frequency	.804
3	0	none	.800
3	0.2	none	.789
3	0	sqrt-inverse-frequency	.776
5	0	sqrt-inverse-frequency	.771
3	0.2	sqrt-inverse-frequency	.764
3	0	inverse-frequency	.759
5	0.2	inverse-frequency	.713
5	0	inverse-frequency	.711
3	0.2	inverse-frequency	.696

Embedding extraction. The 300-dimensional penultimate-layer representation used as the compound encoder in Table 5 is the activation after the classifier FFN's input projection, before the final classification head.

C.9. Compute Resources

All supervised models, distance-based baselines, and the D-MPNN graph neural network were run on a single MacBook Pro (Apple M4 Pro, 14 cores, 48 GB RAM) running macOS, using CPU only. No GPU acceleration was used. Bootstrap confidence intervals (10,000 resamples per evaluation) were computed in parallel across cores. End-to-end reproducibility takes under a few hours on this workstation, plus the OpenRouter API calls for LLM evaluation.

LLM evaluations (Gemini 3.1 Pro and Qwen 3.5 397B-A17B) were performed via the OpenRouter API using the prompt template in Appendix C.4.

D. Reproducibility

This section outlines the files we release to reproduce the results in the paper. The high-level access structure is summarized in Section 3; this appendix lists each artifact, its location, and what it backs in the paper. All file paths below are relative to the anonymous repository at `https://anonymous.4open.science/r/tastebench-4C66/`.

Food-level task — supervised baselines. The supervised baselines (Ridge, Bradley–Terry, hierarchical BT, Kernel RankSVM, LightGBM) are fit per fold under leave-one-out cross-validation, producing 215 per-fold fits per model. Rather than serialize these intermediate fits, which would inflate the artifact without adding reproducibility value, we ship the deterministic per-product out-of-fold predictions at `food_similarity/results/oof_predictions/`, one CSV per (model, feature-subset) combination. Each CSV is the canonical input to its corresponding paper render script. Hyperparameters, training protocol, and the random seed (42) are documented in Appendix C.7; given the released code, cached feature matrices, and fixed seed, any reviewer can regenerate the OOFs byte-identically by running `food_similarity/reproduce.sh`.

Food-level task — LLM baselines. Gemini 3.1 Pro and Qwen 3.5 397B-A17B are accessed via the OpenRouter API; we do not redistribute their weights. The deterministic outputs (prompt completions and the resulting per-product win-rate scores) ship at `food_similarity/results/oof_predictions/llm_*.csv`. Each is a deterministic function of the prompt template (Appendix C.4), the MD5-hashed pair-ordering, and the provider's serving snapshot at the time of the run.

Molecular task — GNN checkpoints. All twelve trained D-MPNN checkpoints from the hyperparameter grid (Appendix C.8) ship at `molecular/results/grid/run_*/ckpt.pt`, each accompanied by a `config.yaml` (the hyperparameters that produced it) and a `val_metrics.json` (per-class validation F1; the aggregated table across all runs is at `molecular/results/grid/grid_summary.csv`). The validation-best run is symlinked at `molecular/results/grid/best/`, and is the checkpoint whose accuracy is reported in the GNN row of Table 4 and whose penultimate-layer activations form the compound encoder evaluated in Table 5. Total checkpoint footprint is approximately 15 MB. Held-out FART test predictions for each run are committed at `molecular/results/grid/run_*/fart_test_eval/predictions.parquet`.

Pretrained external encoders. Three external models are used for feature extraction and are not redistributed:

- **DINOv3 ViT-L/16** (Siméoni et al., 2025), fetched from the official Hugging Face release (food-level image encoder).
- **Qwen3-Embedding-0.6B**, fetched from the official Hugging Face release (food-level ingredient-text encoder).
- **FART** (Zimmermann et al., 2025), fetched from the authors' GitHub release (compound encoder used by both tasks).

Cached embeddings (one matrix per modality, keyed by product code or canonical SMILES) ship at `shared/data/caches/` and are byte-checksummed against `food_similarity/results/input_cache_sha256.txt`. Regeneration scripts that produce these caches from the upstream model releases live at `shared/scripts/prepare_caches/`.

Confidence-interval caches. 95% BCa bootstrap intervals (Efron, 1987) ($n_{\text{boot}} = 10{,}000$, stratified by category, seed 42) are cached alongside each table at `food_similarity/results/cis_*.csv` and `molecular/results/cis_*.csv`. The cached intervals make table re-rendering fast (no bootstrap re-run); the seeds and n_{boot} are baked into both the cache filenames and each render script's header comment so any drift is visible in a diff.

End-to-end reproduction scripts. Two top-level scripts cover the two reviewer paths. `verify_paper.sh` re-renders every paper table and figure from the committed out-of-fold predictions and parquet predictions, with no external data and no model training; this is the lightweight reviewer verification path. `food_similarity/reproduce.sh` regenerates the food-level OOFs from the gated NECTAR ratings (requires NECTAR access; see `data/GATED.md`); this is the heavyweight retrain path. The molecular GNN grid is regenerated via `molecular/scripts/submit_grid.sh dmpnn_grid` followed by `python -m molecular.src.train.select_best_and_evaluate`; the D-MPNN trains on CPU and no GPU is required.

E. Additional Results

E.1. Model Ablation

We report two complementary ablations. Table 8 sweeps the supervised baselines and unsupervised distance models (MMRF) over all 15 non-empty subsets of the $\{\mathrm{N}, \mathrm{C}, \mathrm{T}, \mathrm{I}\}$ feature blocks defined in Section 4.1; Table 9 sweeps Gemini 3.1 Pro and Qwen 3.5 397B-A17B across all 7 non-empty subsets of $\{\text{ingredients}, \text{nutrition}, \text{image}\}$. Pairwise accuracy with 95% BCa CIs is reported in both. The headline result - the SNCTI / ingredients+image configuration - is the one carried into the main table; the ablation documents that no smaller subset matches it within its CI for either the supervised or LLM family.

Table 8. Feature ablation: pairwise accuracy (point on top, 95% BCa CI from 10,000 resamples below) for supervised models and unsupervised distance predictors. S (category subset) is always included for supervised models. N = nutrition, C = compound, T = text, I = image. BT = Bradley–Terry, HBT = Hierarchical BT, KSVM = Kernel RankSVM, LGBM = LightGBM. **Bold** = best subset per model; † = CI overlaps the column leader (no significant difference at 95%). Values below .500 indicate worse-than-random.

	Supervised (S+subset)					MMRF	
Subset	Ridge	BT	HBT	KSVM	LGBM	Cos	L2
N	.447 [.389,.498]	.498 [.441,.555]	.493 [.437,.544]	.509 [.459,.565]	.511† [.454,.565]	.566† [.518,.625]	.549† [.497,.610]
C	.547† [.497,.603]	.572† [.523,.630]	.565† [.516,.623]	.563† [.516,.618]	.548† [.500,.602]	.498† [.440,.554]	.504† [.444,.559]
T	.594† [.548,.658]	.614† [.574,.672]	.567† [.517,.622]	.597† [.552,.654]	.546† [.491,.602]	.499† [.442,.549]	.520† [.468,.572]
I	.411 [.350,.453]	.514 [.458,.570]	.540† [.487,.596]	.512 [.458,.566]	.473 [.417,.525]	.559† [.504,.622]	.569† [.514,.633]
NC	.548† [.499,.605]	.570† [.521,.630]	.563† [.514,.616]	.559† [.511,.616]	.543† [.489,.602]	.545† [.495,.601]	.541† [.490,.599]
NT	.589† [.543,.649]	**.641** [.602,.700]	.548† [.495,.603]	.595† [.550,.652]	.560† [.512,.617]	.558† [.513,.610]	.561† [.515,.618]
NI	.440 [.383,.487]	.529 [.478,.583]	.539† [.487,.595]	.510 [.455,.564]	.511† [.457,.566]	**.579** [.535,.635]	.581† [.537,.640]
CT	.589† [.544,.645]	.611† [.571,.668]	.539† [.485,.598]	**.619** [.581,.678]	.557† [.505,.616]	.487 [.435,.532]	.497 [.443,.546]
CI	.471 [.410,.522]	.537 [.483,.594]	.580† [.537,.636]	.576† [.528,.635]	.543† [.487,.602]	.550† [.497,.608]	.555† [.503,.616]
TI	.595† [.551,.654]	.608† [.570,.662]	.623† [.582,.680]	.613† [.575,.672]	.500† [.444,.553]	.544† [.493,.598]	.577† [.529,.633]
NCT	**.606** [.566,.665]	.618† [.580,.677]	.534† [.479,.593]	.617† [.579,.675]	**.587** [.537,.647]	.530† [.483,.578]	.550† [.499,.606]
NCI	.504 [.448,.558]	.559† [.506,.618]	.591† [.548,.649]	.573† [.524,.631]	.521† [.468,.575]	.574† [.528,.635]	.571† [.524,.632]
NTI	.586† [.541,.646]	.606† [.567,.658]	**.630** [.590,.689]	.612† [.573,.671]	.539† [.487,.595]	.568† [.521,.623]	.587† [.539,.645]
CTI	.575† [.529,.628]	.621† [.585,.673]	.622† [.581,.680]	**.619** [.580,.675]	.539† [.484,.596]	.530† [.477,.587]	.551† [.501,.609]
NCTI	.579† [.534,.632]	.610† [.571,.661]	.613† [.570,.670]	.618† [.579,.674]	.558† [.509,.615]	.566† [.519,.622]	**.591** [.547,.650]

Table 9. Input ablation: pairwise accuracy with 95% BCa CIs (10,000 resamples) for zero-shot LLMs across modality subsets. Ingr. = ingredient list text; Nutr. = nutrition facts; Img. = product image. **Bold** = best per model; † = CI overlaps the column leader (no significant difference at 95%).

Input	Gemini	Qwen
Ingr.	.607† [.566,.664]	.588† [.540,.648]
Nutr.	.592† [.543,.655]	.587† [.537,.646]
Img.	.589† [.544,.648]	.572† [.526,.627]
Ingr.+Nutr.	.622† [.579,.682]	.604† [.558,.665]
Ingr.+Img.	**.654** [.619,.713]	**.630** [.591,.688]
Nutr.+Img.	.610† [.566,.671]	.573† [.526,.632]
All	.625† [.582,.687]	.603† [.556,.664]

E.2. Per-category model comparison

Table 10 breaks down pairwise accuracy by NECTAR product category for every model in the main results table. The breakdown shows substantial heterogeneity across the 24 categories, with some categories notably easier than others for all models. Per-category sample sizes are small (as few as 5 products, median 9), so per-cell CIs are wide and individual category leaders should not be over-interpreted.

Table 10. Per-category pairwise accuracy (point on top, 95% BCa CI from 10,000 resamples below; bootstrap restricted to within-category products) for one model per family. **Bold** = best model per category; † = CI overlaps the row leader (no significant difference at 95%). n = number of products in category; Pairs = $\binom{n}{2}$ within-category ranking pairs. Wide CIs on small-n categories reflect the small effective sample size and are reported faithfully.

Category	n	Pairs	Gemini	BT	BT+Gemini
Bacon	10	45	**.556** [.322,.789]	.511† [.222,.789]	.533† [.289,.778]
Barista Milk	10	45	.767† [.633,.922]	.700† [.521,.889]	**.811** [.656,.933]
Bratwurst	9	36	**.694** [.403,.903]	.639† [.375,.875]	.667† [.375,.875]
Breaded Chicken Filet	5	10	.300† [.100,.700]	**.800** [.300,.950]	.300† [.100,.700]
Breakfast Sausages	10	45	.589† [.311,.844]	.578† [.378,.711]	**.644** [.333,.900]
Burgers	10	45	.756† [.544,.911]	**.800** [.622,.922]	**.800** [.600,.933]
Butter	10	45	.689† [.444,.833]	.689† [.456,.878]	**.711** [.478,.833]
Cheddar Cheese	8	28	**.768** [.607,.929]	.661† [.304,.893]	.732† [.589,.911]
Chicken Strips	10	45	.522† [.367,.744]	.533† [.244,.822]	**.578** [.389,.778]
Cream Cheese	10	45	.667† [.378,.856]	.600† [.322,.811]	**.778** [.478,.922]
Creamer	9	36	**.875** [.778,.958]	.722† [.361,.889]	.833† [.625,.944]
Deli Ham	9	36	.458† [.181,.764]	**.708** [.458,.875]	.542† [.250,.819]
Deli Turkey	7	21	.667† [.333,.857]	.714† [.381,.905]	**.810** [.571,.952]
Hot Dogs	10	45	.667† [.356,.856]	**.756** [.589,.922]	.733† [.567,.889]
Ice Cream Hard Serve	10	45	.467† [.256,.756]	**.711** [.478,.889]	.533† [.333,.789]
Meatballs	5	10	**.600** [.100,.900]	.100† [.050,.250]	**.600** [.100,.900]
Milk	18	153	.699† [.536,.833]	.693† [.595,.801]	**.830** [.716,.928]
Mozzarella	8	28	**.589** [.268,.911]	.214† [.089,.429]	.464† [.161,.786]
Nuggets	10	45	**.533** [.322,.756]	.356† [.156,.556]	.333† [.156,.611]
Pulled Pork	7	21	**.667** [.357,.905]	.333† [.095,.619]	.571† [.333,.833]
Sour Cream	5	10	**.600** [.250,.900]	.500† [.100,.900]	**.600** [.250,.900]
Steak	9	36	**.833** [.722,.931]	.556† [.222,.806]	.778† [.542,.903]
Unbreaded Chicken Breast	10	45	.756† [.522,.911]	.600† [.367,.800]	**.800** [.600,.911]
Yogurt	6	15	**.467** [.133,.733]	.133† [.067,.300]	.400† [.133,.700]

E.3. Gemini ensemble comparison

Table 11 shows the pairwise accuracy of combining Gemini 3.1 Pro with each supervised model using a non-negative least squares meta-learner.

Table 11. Effect of swapping the supervised base in the BT+Gemini NNLS ensemble. Each row reports the pairwise accuracy of a supervised model alone (Standalone) and combined with Gemini 3.1 Pro via nested LOOCV NNLS (+ Gemini NNLS), on the SNCTI feature set ($n = 215$ NECTAR plant-based products, 935 within-category pairs). Point estimate above, 95% BCa CI (10,000 resamples) below. **Bold** = best in column; † = CI overlaps the leader (no significant difference at 95%).

Supervised model	Standalone	+ Gemini NNLS
Gemini 3.1 Pro (alone)	**.654** [.619,.714]	—
Bradley–Terry	.610† [.571,.662]	**.683** [.655,.740]
Hierarchical BT	.613† [.571,.671]	.665† [.634,.720]
Ridge	.579† [.535,.633]	.641† [.604,.699]
Kernel RankSVM	.618† [.579,.674]	.659† [.628,.713]
LightGBM	.558 [.509,.616]	.619† [.579,.679]

E.4. Per-category NNLS metrics

Table 12 supplies the per-category inputs (σ_S/μ_S, ρ_S, r) used in the screening-utility worked example in Section B (Proposition B.1), together with per-category sample sizes, NNLS pairwise accuracy with 95% BCa CIs, and the rank of the true top-rated product. Expected uplift from a model-driven shortlist scales with $(\sigma_S/\mu_S) \cdot \rho$, so the categories with the largest screening gains are those ranking high on both columns.

Table 12. Per-category metrics for the NNLS ensemble (BT + Gemini), covering both the per-category evaluation metrics and the inputs used in the worked example (Section B). n = number of products in category; Pairs = $\binom{n}{2}$ within-category ranking pairs. σ_S/μ_S is the coefficient of variation of the panel-mean similarity within the category. ρ_S is Spearman rank correlation; r is Pearson correlation between predicted and panel-mean scores. "True-best rank" is the position of the highest-rated product in the model's ranking (1 = top). Pairwise accuracy CIs are 95% BCa (10,000 resamples). Aggregated across these 24 categories, NNLS achieves R@1 = .292, R@2 = .500, R@3 = .542.

Category	n	Pairs	σ_S/μ_S	Pw. Acc.	ρ_S	r	True-best rank
Bacon	10	45	.191	.533 [.289,.778]	+.091	+.235	5
Barista Milk	10	45	.155	.811 [.656,.933]	+.821	+.753	1
Bratwurst	9	36	.169	.667 [.375,.875]	+.450	+.433	1
Breaded Chicken Filet	5	10	.064	.300 [.100,.700]	−.600	−.674	4
Breakfast Sausages	10	45	.125	.644 [.333,.900]	+.382	+.124	1
Burgers	10	45	.190	.800 [.600,.933]	+.758	+.812	4
Butter	10	45	.135	.711 [.478,.833]	+.636	+.602	2
Cheddar Cheese	8	28	.104	.732 [.589,.911]	+.695	+.746	2
Chicken Strips	10	45	.113	.578 [.389,.778]	+.164	+.136	4
Cream Cheese	10	45	.231	.778 [.478,.922]	+.685	+.814	7
Creamer	9	36	.226	.833 [.625,.944]	+.817	+.921	4
Deli Ham	9	36	.160	.542 [.250,.819]	+.117	−.074	7
Deli Turkey	7	21	.183	.810 [.571,.952]	+.750	+.720	1
Hot Dogs	10	45	.191	.733 [.567,.889]	+.697	+.624	2
Ice Cream Hard Serve	10	45	.174	.533 [.333,.789]	+.091	+.432	6
Meatballs	5	10	.256	.600 [.100,.900]	+.300	+.271	1
Milk	18	153	.190	.830 [.716,.928]	+.798	+.794	9
Mozzarella	8	28	.137	.464 [.161,.786]	−.119	−.336	8
Nuggets	10	45	.155	.333 [.156,.611]	−.455	−.275	7
Pulled Pork	7	21	.164	.571 [.333,.833]	+.214	+.197	2
Sour Cream	5	10	.160	.600 [.250,.900]	+.200	+.405	2
Steak	9	36	.214	.778 [.542,.903]	+.733	+.696	1
Unbreaded Chicken Breast	10	45	.137	.800 [.600,.911]	+.782	+.812	1
Yogurt	6	15	.134	.400 [.133,.700]	−.200	−.386	3

E.5. Ingredient Aggregation Ablation

Section 4.1 aggregates compound embeddings to ingredients via a log-concentration-weighted mean, motivated by the Weber–Fechner law. Table 13 compares five candidate aggregation rules (linear-concentration, log-concentration, max, arithmetic mean, and top-3-by-concentration) under both Bradley–Terry alone and the BT+Gemini NNLS ensemble. Within-block and all-pairs pairwise accuracy are reported with 95% BCa CIs (within-block via cluster bootstrap on blocks). Differences between rules lie within 95% CI overlap on both scopes; we adopt log-concentration on theoretical grounds (Weber–Fechner) rather than empirical advantage.

Table 13. Compound-to-ingredient aggregation ablation on the food-similarity task (LOOCV, $n = 215$ NECTAR products, 935 within-category pairs). For ingredient I with compounds $i \in I$ at concentrations c_i and compound embeddings z_i, the per-ingredient embedding z_I follows one of the rules below. *All-pairs* covers all within-category pairs; *within-block* restricts to the BIBD blocks panelists saw (apples-to-apples with humans). 95% BCa CIs in brackets (10,000 resamples; cluster bootstrap on blocks for within-block). **Bold** = best per (model, scope); † = CI overlaps the leader (no significant difference at 95%).

Aggregation	Embedding z_I	BT		BT + Gemini NNLS	
		All-pairs	Within-block	All-pairs	Within-block
Uniform mean	$\frac{1}{n}\sum_i z_i$	.592† [.551,.645]	.555† [.487,.616]	.673† [.642,.730]	.656† [.595,.713]
Element-wise max	$\max_i z_i$ (per coord.)	.605† [.566,.657]	.572† [.509,.631]	.679† [.649,.740]	**.661** [.600,.717]
Top-3 by conc.	$\frac{1}{3}\sum_{i\in\mathrm{top}_3(c)} z_i$	**.651** [.621,.702]	.608† [.534,.670]	**.685** [.657,.737]	.648† [.585,.705]
Linear conc.	$\sum_i \frac{c_i}{\sum_j c_j} z_i$	.622† [.586,.676]	**.613** [.554,.670]	.676† [.647,.730]	.651† [.588,.707]
Log conc.	$\sum_i \frac{\log(1+c_i)}{\sum_j \log(1+c_j)} z_i$	.610† [.571,.662]	.577† [.504,.640]	.683† [.655,.740]	**.661** [.599,.714]

E.6. Per-class breakdown for molecular prediction

Table 14 reports per-class precision, recall, F1, and AUROC for FART and our val-best D-MPNN on the 2,254-molecule FART test set. The per-class F1 CIs overlap between the two models on every class: gaps are within $+2.3$ pp on the four non-umami classes (largest on bitter, where the GNN nominally leads on both precision and recall). On umami, where the macro-F1 gap is concentrated, FART achieves higher precision (1.000) at the cost of lower recall (0.333) while the GNN achieves balanced precision and recall (0.667 each); however, with only n=6 test molecules a single additional correct prediction shifts F1 by $\approx$0.17, and the FART–GNN F1 gap is itself within CI overlap.

Table 14. Per-class breakdown for Table 4: Precision, Recall, F1 and one-vs-rest AUROC of FART and GNN on each FART test class, with 95% BCa CIs (10,000 resamples; point on top, CI in brackets below). Support row gives the number of test molecules per class. Umami CIs span much of $[0, 1]$ because $n = 6$. **Bold** = winner of FART vs GNN per (class, metric); † = CI overlaps the leader (no significant difference at 95%).

	Sweet	Bitter	Sour	Umami	Undefined
Support	1473	233	238	6	304
FART					
Precision	.944† [.931,.955]	.856† [.801,.901]	.897† [.853,.931]	**1.000** [.000,1.000]	.698† [.648,.746]
Recall	.943† [.930,.954]	.712† [.651,.767]	.916† [.874,.947]	.333† [.000,.800]	**.789** [.742,.834]
F1	.944† [.935,.952]	.778† [.731,.818]	.906† [.877,.931]	.500† [.000,.909]	.741† [.702,.778]
AUROC	.979† [.973,.983]	.951† [.932,.965]	.995† [.992,.997]	**.989** [.975,.999]	.958† [.946,.967]
GNN					
Precision	**.945** [.932,.956]	**.869** [.817,.912]	**.906** [.864,.937]	.667† [.000,1.000]	**.736** [.686,.783]
Recall	**.953** [.941,.963]	**.742** [.683,.796]	**.933** [.894,.960]	**.667** [.000,1.000]	.770† [.720,.816]
F1	**.949** [.940,.957]	**.801** [.757,.840]	**.919** [.891,.942]	**.667** [.000,.889]	**.752** [.713,.790]
AUROC	**.981** [.975,.985]	**.958** [.939,.970]	**.996** [.993,.997]	.988† [.927,1.000]	**.966** [.958,.973]

F. Architecture with Worked Example

Figure 5 traces a single product, a plant-based burger in the Burgers category ("Product A"), through the full TasteBench ensemble pipeline. The same pipeline is applied independently to every product, and we show one instance to make the data flow concrete. All four modalities (ingredient text, nutrition, product image, and molecular composition) are encoded into per-modality vectors, which are standardized and, for the high-dimensional text, image, and compound blocks, reduced by per-modality PCA to 95% explained variance (fit within each LOOCV fold). The low-dimensional nutrition and category-subset blocks are not PCA-reduced. The standardized, concatenated SNCTI vector is the input to the supervised component (Bradley–Terry, the supervised model carried into the final ensemble) fit by logistic regression on within-category feature-pair differences and scored at inference as a product's expected number of within-category tournament wins, $\sum_k \sigma\big(w \cdot (x - x_k)\big)$ over the same-category opponents x_k. The non-molecular modalities are additionally read in raw form by the zero-shot LLM scorer (Gemini 3.1 Pro), which does not consume the encoded or fused features; the LLM score is a product's win rate over its within-category pairwise comparisons. The two scores, s_{BT} and s_{LLM}, are combined by non-negative least squares ($w_1, w_2 \geq 0$, fit by nested leave-one-out cross-validation) to produce the

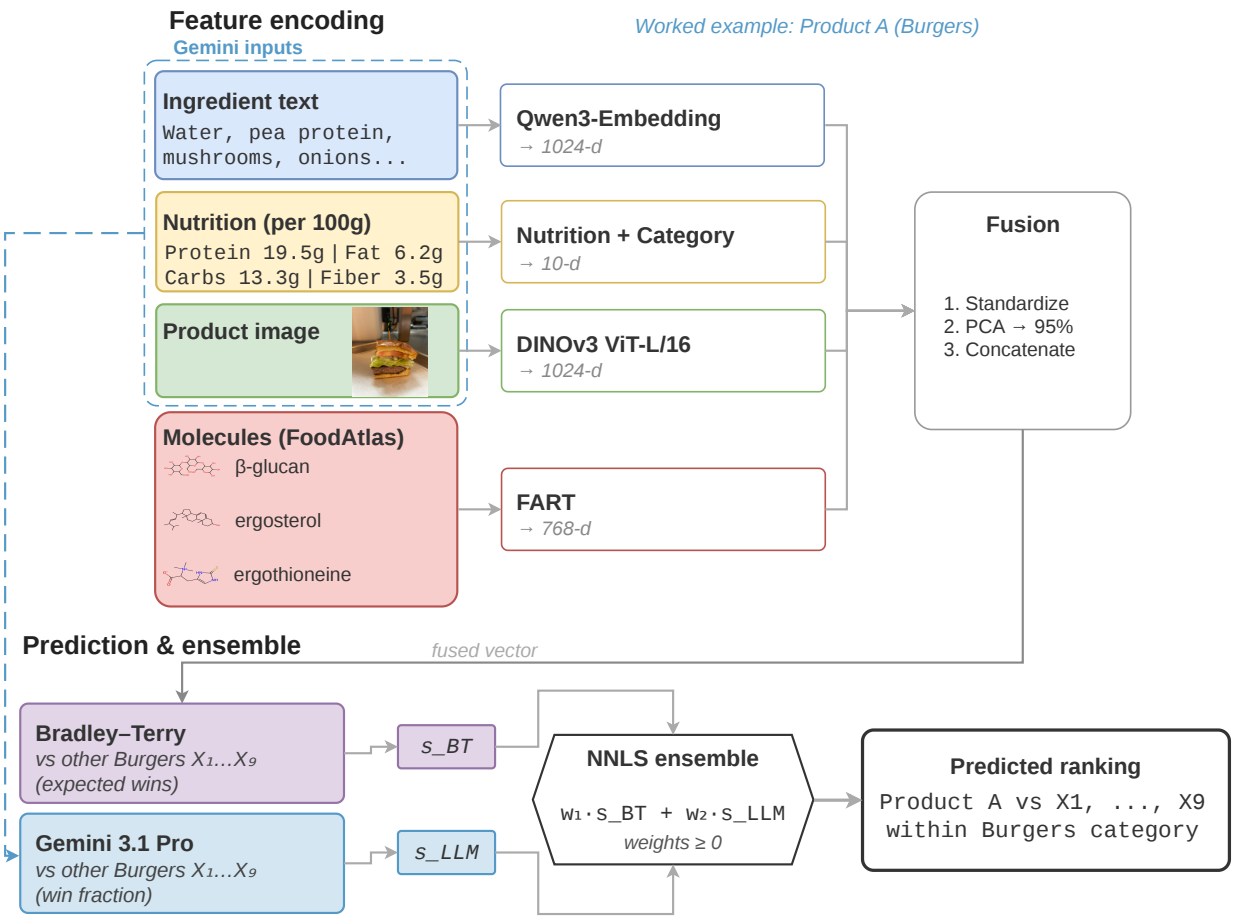

Figure 5. Worked example of the ensemble pipeline for one plant-based burger (Product A) in the Burgers category ($n = 10$). All four modalities feed the supervised Bradley–Terry scorer through encoding and fusion; the non-molecular modalities are also read in raw form by the zero-shot LLM scorer (Gemini 3.1 Pro). The two scores are combined via non-negative least squares to produce the predicted within-category ranking.

predicted within-category ranking. The figure's "Gemini inputs" grouping marks the modalities the LLM can read directly from the prompt template (Appendix C.4). The deployed ensemble uses the ingredients-plus-image configuration, with the ingredients-plus-nutrition-plus-image variant evaluated as an ablation (Table E.2). Ingredient and nutrition values are abbreviated for readability, and the nine opponents shown ($X_1, \ldots, X_9$) are specific to the ten-product Burgers category, the sizes of other categories vary.