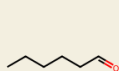
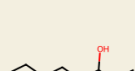


How to Benchmark RaRaMa with Your Own Papers

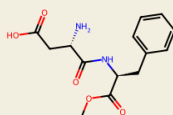
Turn any paper — or your own bench data — into a blind test of the platform, and read the credibility number: how many sigma above chance the prediction sits. Structure in, every number blind-scored.



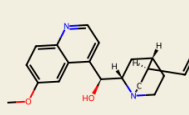
Hexanal



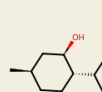
1-Octen-3-ol



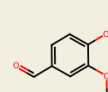
Aspartame



Quinine



Menthol



Vanillin

SECTION 1

The idea in one paragraph

A paper hands you both halves of a test. The structures it studies (a SMILES, a protein sequence, a UniProt accession) plus the conditions (pH, ionic strength, cofactor state) are the input. The values it measured (a threshold, a zeta potential, a melting point, a liking score) are the answer key. You feed the input, the platform predicts, and you compare. The platform never sees the answer key — that is what makes it blind. This is the same loop we use internally to validate every number we publish.

“Structure in, prediction out, then check it against what the paper measured. If the platform cannot anchor a case, it refuses rather than guess.”

The scoring rule — sigma above chance

Any correlation the platform reports counts only the rows where you supplied a measured value, and it reports a **sigma-above-chance** score: how many standard deviations above random the prediction sits pairing. We compute it by shuffling the pairing thousands of times and asking how often chance does as well as the model. Z above ~3 = a real correlation. Z near 0 = no signal, and we say so. When the model exceeds every shuffle, the p-value is reported as an upper bound (for example < 5e-5) — never a canned number. The sigma-above-chance score exists so a weak result can never be dressed up as a strong one.

SECTION 2

Blind by design — you keep the answer

You do not hand us your measured values, and you should not. The platform predicts from structure only (a SMILES, a sequence, an accession) plus the conditions. Your measured numbers stay with you, in a separate held answer key, and you score our predictions against them yourself, after the run.

“Structure in. We never see the measurement. That is the whole point — a prediction is only a benchmark if the model could not have peeked at the answer.”

This matters two ways. First, it protects your data: your proprietary measurements never enter our system. Second, it is what makes the number defensible — every accuracy figure we publish, and every one you compute on your own data, is a true blind test. If you ever do include a measured column (optional), the engine still predicts from structure alone and uses the measured value only to print the comparison; it is never an input to the prediction.

The input format

Every run is one JSON file. Minimum shape:

```
{ "client": "my-project", "options": { "tier": 2 },
  "compounds": [ { "id": "row-1", "smiles": "CCO" } ] }
```

id is yours — echoed back on every row so results join to your sheet. **tier**: 1 = taste, 2 = + aroma / shelf-life / instruments, 3 = + safety / gut / genetics. Give a structure (**smiles**, **sequence**, or **accession**) and the conditions the paper measured at. You normally submit structure only; a measured value is optional and, if included, is used only to print the comparison, never as an input to the prediction (see “blind by design” above). Numbers must be numbers (**"pH": 3.0**, not **"3"**).

```
python -c "import codex_taste_format as fmt; fmt.run('my_panel.json')"
```

SECTION 3

Five worked examples — extract, format, score

1 · Small-molecule threshold paper → odour detection threshold

Example paper: Abraham 2012 / Nagata 2003 (odour thresholds, ppb).

What you pull from the paper: Each compound's SMILES + its measured threshold.

```
{ "client":"odt", "options":{"tier":2}, "compounds":[
  {"id":"hexanal", "smiles":"CCCCCC=O", "measured":4.5},
  {"id":"1-octen-3-ol", "smiles":"CCCCC(O)C=C", "measured":1.0} ] }
```

Native metric returned: Predicted OTV + OAV per volatile (ppb) in results.csv; the usable read is the OTV RANK.

Blind result: OTV read-across RANK: Spearman rho 0.76-0.82, 13.4 sigma above null (held-out 0.755, 88/126). Absolute single-proxy threshold is much harder - Abraham 2012 single structural proxy $r \sim 0.27$ ($z \sim 4.6$), reported as weak, not dressed up. The rank is usable; the absolute value is not. Refuses out-of-domain rather than guess.

2 · Protein colloid paper → zeta potential

Example paper: Kumar 2025 (pea protein + red beet, zeta at pH 3).

What you pull from the paper: The protein (accession is enough) + the measurement pH.

```
{ "client":"zeta", "options":{"tier":3}, "compounds":[
  {"id":"legumin", "accession":"P15838", "conditions":{"pH":3.0}} ] }
```

Native metric returned: A zeta-vs-pH curve + the value at your submitted pH.

Blind result: Sign and pH-direction correct from sequence alone (positive below the pI, negative above); magnitude labelled a surface-potential upper bound.

3 · Protein thermal-stability paper → denaturation temperature

Example paper: Any DSC study of a food globulin (Td in °C).

What you pull from the paper: Sequence or accession, oligomer state, measured Td.

```
{ "client":"td", "options":{"tier":3}, "compounds":[
  {"id":"glycinin", "accession":"P04776", "oligomer":"11s",
    "n_disulfide":2, "measured":92.0} ] }
```

Native metric returned: Predicted Td with a confidence tier (HIGH on the calibrated core, PROVISIONAL for other folds, REFUSED for a novel/disordered chain).

Blind result: Denaturation MAE 3.2 °C in-family / ~ 6.6 °C cross-family; the dossier reports a wider ± 3 -5 °C confidence band on a single call, not the set-level MAE. Gluten, collagen, and intrinsically-disordered chains are refused with a reason (and still return allergen family + bitter peptides).

4 · Sensory / VOC panel → aroma and off-note load

Example paper: Fan 2021 (strawberry, volatiles vs consumer liking, 148 samples).

What you pull from the paper: Each volatile's SMILES (or name) + its concentration per sample.

```
{ "client":"strawberry", "options":{"tier":2}, "compounds":[
  {"id":"me-butanoate","smiles":"CCCC(=O)OC","concentration_ppb":1800},
  {"id":"hexanal","smiles":"CCCCCC=O","concentration_ppb":900} ] }
```

Native metric returned: Each volatile's aroma family + off-note flag + OAV, and an off-note-load ranking across samples. (You can upload your GC-MS .xlsx directly.)

Blind result: Concentrations unlock the impact ranking; structure alone gives family + intrinsic potency.

5 · Taste / sweetener panel → taste call + a stated scope

Example paper: Any sweetener set or a modified-sugar series.

What you pull from the paper: Each compound's SMILES (name it for the identified potency).

```
{ "client":"sweeteners", "options":{"tier":1}, "compounds":[
  {"id":"aspartame","name":"aspartame",
    "smiles":"COC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CC(=O)O"},
  {"id":"novel-1","smiles":"OC[C@H]1O[C@H](O[C@]2(CC1)O[C@H](CC1)
    [C@@H](O)[C@@H]2O)[C@H](O)[C@@H](O)[C@H]1Cl"} ] }
```

Native metric returned: Taste class, a confidence, and a plain-language mechanism.

Blind result: For a modified sugar where the substitution pattern (not the scaffold) decides the taste, it returns a PROVISIONAL sweet-leaning call and says the pattern is not resolvable from 2-D structure — rather than a confident wrong number.

SECTION 4

Showing the work — one compound, every step

Here is a real run, start to finish, on open data anyone can download. The goal is to show there is no sleight of hand: the measured value is held aside, the model sees only the structure, and the prediction is checked afterward.

Step 1 — the raw material (open data). From the Abraham 2012 odour-threshold set (353 human detection thresholds, distributed openly through PyrFume), I take two compounds. Their measured potency is Log(1/ODT) — higher means a lower threshold, i.e. more potent:

compound	SMILES	measured Log(1/ODT) — the held answer
hexanal	CCCCCC=O	2.71 ← potent
1-hexanol	CCCCCCO	1.59 ← less potent

Source: Abraham et al. 2012, *Chem. Senses* 37(3):207 (doi:10.1093/chemse/bjr094), via the open PyrFume dataset *abraham_2012*.

Step 2 — hold the answer aside. I copy those two Log(1/ODT) numbers into a note and close it. They do not go into the platform. This is the whole discipline.

Step 3 — format the structure only.

```
{ "client":"odt-walkthrough", "options":{"tier":2}, "compounds":[
  {"id":"hexanal", "smiles":"CCCCCC=O"},
  {"id":"1-hexanol", "smiles":"CCCCCCO"} ] }
```

Step 4 — run it.

```
python -c "import codex_taste_format as fmt; fmt.run('odt_walkthrough.json')"
```

Step 5 — what comes back. The platform predicts an odour threshold for each, from structure alone:

compound	predicted ODT	
hexanal	4.5 ppb	potent
1-hexanol	500 ppb	less potent

"I did not tell it hexanal was the potent one. It read the aldehyde group, put it on the chain-length curve, and returned the lower threshold on its own."

Step 6 — compare and score. Now I reopen my note. Measured: hexanal 2.71 > 1-hexanol 1.59 (hexanal more potent). Predicted: hexanal 4.5 ppb << 1-hexanol 500 ppb (hexanal more potent). The order agrees. Do this over the whole set and you get the number in the benchmark table below: Spearman 0.76-0.82, 13.4 sigma above chance, with the compounds it cannot place refused, not guessed.

SECTION 5
What actually comes back (real output)

These are unedited screenshots of a real RaRaMa dossier — the same file a client downloads. First, the per-compound taste table: every call carries its confidence and the mechanism behind it.

RaRaMa · TASTE READOUTHans-Made Research Inc.

FTIR — functional-group fingerprint · Naringin

Transmittance (%)

Wavenumber (cm⁻¹)

Labels: C-H aliphatic (2825), C=O ketone (1670), C-O aromatic (1250), C-O ether (1100), C-O alcohol (1050)

Scans from the functional-group combination table: 4000-1500 cm⁻¹ = Functional group zone, 1500-600 cm⁻¹ = Fingerprint, depth = intensity, width = H band broadening

Submission manifest — every compound accounted for

Completeness check (no compound is silently dropped): 9 submitted · 9 scored · 5 off-note liabilities · 0 abstained (out of standard tastant domain) · 0 could not be read. Every valid structure is computed from the framework dimension and appears in the per-compound table and the fingerprint heatmap below.

Per-compound read — the call and the mechanism behind it

Bottom line: 9 compound(s) scored — 3 Sweet · 2 Bitter · 2 Astringent · 2 aroma-only. 5 carry an off-note liability (flagged below).

#	Compound	Input	Taste call	Off-note	Mechanism (why it fired)	Food safety & stability
1	Sucralose	SMILES	SWEET	no	a chlorodeoxy-sugar sweetener (sucralose class) — chlorine substituted onto the sugar backbone gives a high-potency sweet that the body does not metabolise	PASS
2	Acesulfame potassium	SMILES	SWEET	YES	Cyclic sulfonyl-N sweetener — a 1,2,3-oxathiazin-4(3H)-one 2,2-dioxide (the acesulfame class; the -O-S(=O)2-N- is IN the ring). NOT the cyclamate sulfamate; the ring seats the sulfonyl-N in the sweet pocket, and the motif cross-activates bitter receptors at functional dose → a concentration-dependent bitter/metallic tail	PASS
3	Aspartame	SMILES	SWEET	no	Artificial Sweet	PASS
4	Naringin	SMILES	BITTER	YES	a flavanone neohesperidoside — the bitter glycoside motif of naringin and neohesperidin	PASS
5	Quercetin	SMILES	ASTRINGENT	YES	Astringent	PASS
6	Chlorogenic acid	SMILES	ASTRINGENT	YES	Caffeoyl catechol-polyphenol ester (chlorogenic/rosmarinic class) — precipitates salivary proteins, astringent	PASS

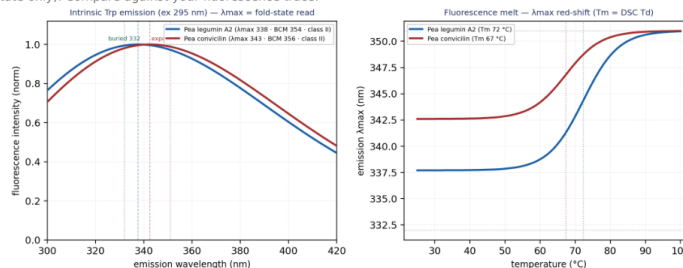
Confidential — prepared for RaRaMa Sample Dossier V2.
Integrity (SHA-256 manifest): 93ade973ace5507c...

Page 5

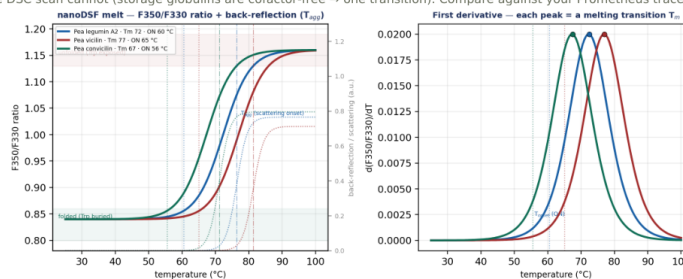
Real dossier — per-compound taste calls, confidence, and mechanism (sample dossier).

And the protein-functionality readouts: the fold-state instrument reads and the zeta-potential-versus-pH curve the platform computes from sequence, to set against your Zetasizer trace. The sign flips at the isoelectric point — positive below it, negative above.

melt, its thermal midpoint tracks the CD and DSC estimates (a prolamin / disordered chain has none — the emission reports the fold state only). Compare against your fluorescence trace.



Predicted nanoDSF (F350/F330) thermal shift — the modern label-free way labs read Tm: the intrinsic-fluorescence ratio rises as the core unfolds, and each peak of its first derivative is a melting temperature (the same midpoint as the DSC/CD estimates). A single-domain protein shows one peak; WHERE a bilobal cofactor protein is present, it melts in two transitions — an early cofactor-free lobe (S1) and a late cofactor-bound lobe (S2) — so the S1/S2 areas report the fractional cofactor saturation a single DSC scan cannot (storage globulins are cofactor-free → one transition). Compare against your Prometheus trace.



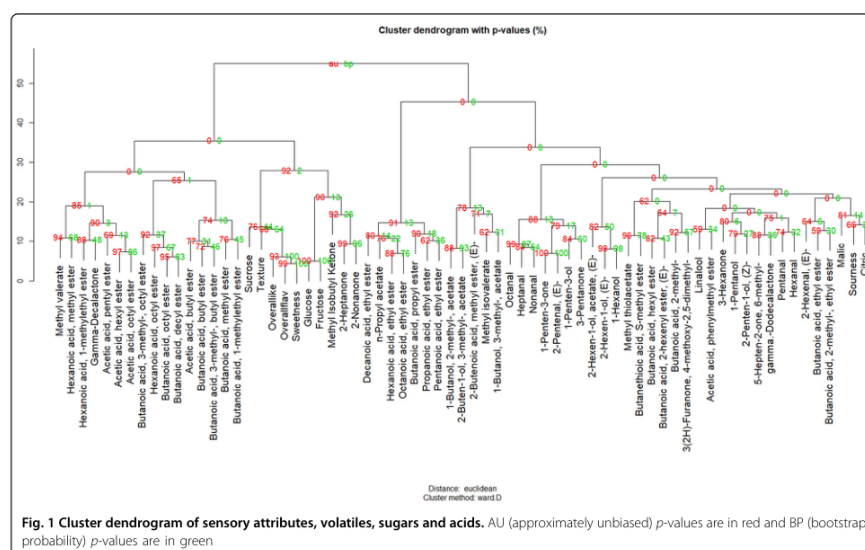
Predicted zeta potential vs pH — the colloidal-stability map: the isoelectric point ($\zeta = 0$) is the pH of least stability (flocculation/precipitation), and $|\zeta| > 30$ mV is the stable-dispersion window. Structure-derived from the charge profile; compare against your Zetasizer trace.

Real dossier — intrinsic-fluorescence fold read, nanoDSF melt, and zeta-vs-pH (sample dossier).

SECTION 6

Reading a real open-access paper

You do not need a curated dataset. Any open-access paper with structures and a measured value works. Below is a figure from an open-access strawberry-flavour study; the kind of material you would pull volatiles and a sensory outcome from, hold the outcome aside, and predict from structure.



compound to consumer liking in the first two datasets. Fructose and glucose had lesser but important influence on liking in all three datasets. Important volatiles obtained for consumer liking included 10 esters, 3 aldehydes, 1 lactone, 3 ketones, and 1 alcohol. Eleven of these compounds were significant for the *t*-test (Table 1 and Supplementary Table S5), providing evidence that they had positive impact on consumer liking that was independent of total sugars. Butanoic acid, ethyl ester (V19), known to be one of the most abundant and important contributors to strawberry aroma, showed significant positive correlations ($r = 0.31$) with liking but did not pass the PLS threshold for two of the three datasets. Other abundant esters like butanoic acid, methyl ester (V9); hexanoic acid, methyl ester (V40) and hexanoic acid, ethyl ester (V44) all showed positive but low to moderate correlations ($r = 0.29, 0.17$, and 0.13 , respectively) with liking and did not reach PLS thresholds. Fifteen compounds were identified via PLS as important drivers of both sweetness and liking models, reflecting the high correlation between sweetness and liking (Table 1).

Predicting sweetness and liking with sugars, acids, and volatiles

Since sugars and acids are known to be important factors influencing sweetness perception and consumer liking, we first built a model with sugars and acids.

This model produced an R^2 of 0.41 in testing CV (independent test dataset) for sweetness, indicating that 41% of variation in sweetness could be accounted for by sugars and acids. To further investigate the influences of volatiles on sweetness and liking, we included them as variables and tested prediction accuracies with a broad spectrum of algorithms (Fig. 5a), including decision-tree based Random Forest (RF), bayes GLM, GLM, GLMBOOST, and LASSO. With the inclusion of volatiles, LASSO and GLMBOOST models explained the most variation for sweetness in testing CV ($R^2 = 0.65$ and 0.69), outperforming the model with sugars and acids by 28% for the test dataset. In addition, LASSO and GLMBOOST generated the lowest standard deviation of R^2 among all models (Fig. 5a). Similar results were produced in prediction of liking but with a slightly lower total variance explained. LASSO model accounted for 55% in testing CV, compared to only 30% in models with sugars and acids alone (Fig. 5b). These results provide further strong evidence of volatile compound enhancement of strawberry sweetness perception and consumer liking independently of sugars.

Chemical networks in strawberry fruit

Broad variations in quantity of chemical compounds were measured for the 148 strawberry samples. For example, the most abundant ester (butanoic acid, methyl

Reproduced under CC-BY 4.0 from Fan et al. 2021, *Horticulture Research* 8:66 (doi:10.1038/s41438-021-00502-5). Used here as an example of source material; no endorsement implied.

The recipe is always the same: pull the structures, note the measured outcome somewhere the platform never sees, run the structures, then check. If the sigma-above-chance is high, you have a real result. If it is low, we tell you so.

Downloaded from academic.oup.com/hr/article/doi/10.1038/s41438-021-00502-5/6446614 by guest on 23 August 2023

SECTION 7

Where the platform does very well (blind / zero-shot)

Every number below is a blind or held-out benchmark scored against a public answer key the model never saw. Sources are traceable; the limits follow on the next panel.

Lane	What is predicted	Result	N	Source / basis
Taste	E-tongue bitterness (external lab)	R^2 0.95 · r 0.97 · MAE 1.10	20	RCFTR ASTREE, U Manitoba — contracted correlation, NOT a blind report
Taste	Umami bioactive peptides	call validated, held (no bare MCC — grader emits no single clean value)	88	iUmami / UMP442 — trained iUmami-SCM ~0.61 is a different one-vs-rest task, NOT a beat
Taste	6-class taste vs held key	72.7% (null 32.3%, 45.9% majority)	3,517	ChemTastesDB full public set, zero-train
Taste	Sweet detection (e-Sweet)	75.8% · MCC 0.559	1,600	Zheng 2019, zero-train
Taste	Sweet/bitter cross-dataset (held-out)	MCC 0.53 sweet / 0.45 bitter (transfer)	161/171	BitterSweet, disjoint from ChemTastesDB
Aroma	Odour detection threshold (read-across)	ρ 0.76–0.82 · 13.4 sigma	88 / 126	Nagata 2003 + Abraham 2012 (held-out ρ 0.755)
Aroma	Aroma family recall (structure→family)	51–68% recall · 6–12% null	2,505–4,297	Leffingwell + GoodScents + 3 lexicons, zero-train
Aroma	VOC → tomato flavour (external)	Pearson 0.807 · Spearman 0.825	15	held tomato panel, zero-shot
Protein	Denaturation temperature (DSC)	MAE 3.2 °C in-family · ~6.6 °C cross-family	6–24	commercial soy/pea/canola/faba, sequence-only
Protein	Allergen family (IUIS) · celiac split	88% exact (251/285) · +8 super (91% exact-or-super) · 2 wrong · 24 abstain · 7/7 grains	285	WHO/IUIS 120 + 165 expansion, sequence-only
Protein	Intact mass / ϵ 280 (QC-exact)	ρ 0.92 / exact to the unit	26	NIST-SRM Δ 0.11 Da, sequence-only
Protein	Gel storage modulus G' (rank)	ρ 0.926 · 11/11 concordant	11	Ge 2023, sequence-only
Colloid	Zeta-potential sign & pH-direction	correct blind	3	Kumar 2025, sequence-only

The refusal is part of the result: on the odour-threshold sets the platform declines ~40–50% of compounds as out-of-domain rather than emit a false number. The accuracy above is on what it accepts.

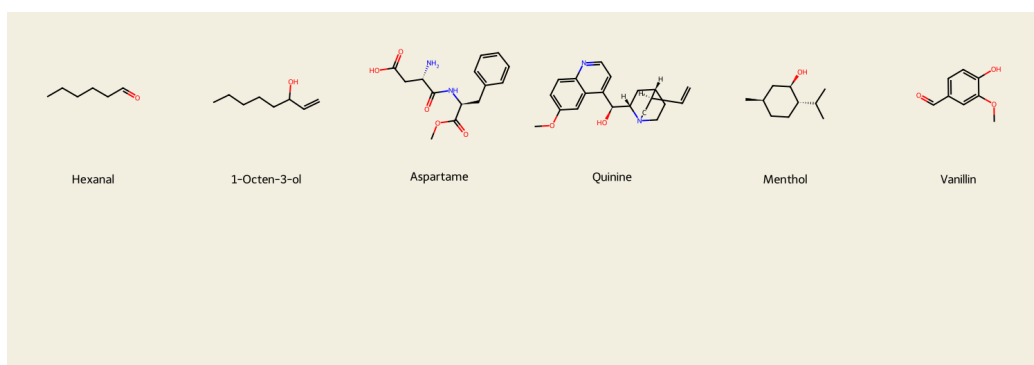
SECTION 8

Known limits & nulls (so you trust the rest)

Case	What we see	Why we report it
Odour intensity	$r \sim 0.10$ — no signal (trained SOTA ~ 0.78)	DREAM / Keller; a null, and we say so, not a result
Bitter-peptide recall	MCC 0.05–0.10 (absolute call)	the defensible read is RANK-ORDERING of fragments, not a bitter/not verdict; umami peptides work, absolute bitter/sour/salty do not
Cross-lane melt (MS-Tm)	$\rho -0.17$ — does not track	a whole-proteome melt is a different observable from our DSC axis; 90.6% abstained
Relative-sweetness magnitude	$ r 0.41$ — bounded	we rank sweeteners well; the exact x-sucrose number is not derivable from 2-D structure
Receptor-activation (full repertoire)	AUROC 0.62 (trained 0.96+)	zero-shot vs a model trained on the receptor set; portability within a class is the claim
Zeta / modified-sugar isomers	upper bound / PROVISIONAL	sign + direction are the read; the exact magnitude / substitution pattern is disclosed, not faked

Our positioning: we do NOT beat a model trained on its home dataset at its home task. The moat is zero-shot breadth, leakage-immunity, and portability — giving a defensible answer on a molecule or construct nobody has measured yet, where the alternative is a bench run and a week in a queue.

This is the point of the sigma-above-chance score: score your own data the same way, and if the Z is small, the platform is telling you the truth about its own limits. That candour is the product.



Ready to benchmark RaRaMa on your own paper?

Format one JSON panel — structure only — run it, then score the predictions against the answer key you kept. That is the whole loop. If the sigma-above-chance is high, you have a real result; if it is low, we have told you so. Want us to run the first one with you?

Dustin Hansley · dustin-rarama@hansley.ca · 204-333-0234 · www.hansley.ca