Flavor Forge: Bidirectionele smaakmodellering van Belgisch bier met probabilistische neurale netwerken

Auteurs: *Jens Mertens, Erwin Sarban, Matthias Truyzelaere*

Inleiding

Bier bevat honderden smaakactieve verbindingen. Welke precies waarvoor verantwoordelijk is, en in welke combinatie, weet niemand precies. Dat is al decennialang het kernprobleem in sensorische wetenschap.

Het vertrekpunt voor dit project is de studie van Schreurs et al. (2024), *"Predicting and improving complex beer flavor through machine learning"*, gepubliceerd in *Nature Communications*. Ze analyseerden 250 Belgische commerciële bieren: meer dan 200 chemische metingen per bier, sensorische scores van een getraind expertpanel, en 180.000+ reviews van het online platform RateBeer. Gradient Boosting bleek het beste model: $R^2 = 0.22$ op paneldata, $R^2 = 0.69$ op RateBeer-scores. Door bieren te voorzien van de meest voorspellende verbindingen (ethylacetaat, ethanol, melkzuur) kregen die bieren aantoonbaar hogere beoordelingen.

Wat Schreurs et al. niet deden: de omgekeerde richting modelleren. Als je een bepaald smaakprofiel wilt bereiken, welke chemische samenstelling heb je dan nodig? Dat is het inverse probleem, en dat is lastig: de mapping is niet uniek. Meerdere chemische profielen kunnen dezelfde smaak opleveren. Een puntschatting als antwoord is dan eerder misleidend dan nuttig.

Flavor Forge bouwt verder op de dataset en aanpak van Schreurs et al. en voegt drie dingen toe:

- een probabilistisch inverse model dat per chemische output een kansverdeling geeft in plaats van één getal
- Optuna-hyperparameteroptimalisatie (200 trials, Bayesiaans) in plaats van GridSearchCV
- een interactieve webapplicatie met UMAP-visualisatie en hiërarchische clustering

De drie onderzoeksvragen die we proberen te beantwoorden:

1. Kan machine learning sensorische descriptoren van bier voorspellen op basis van de chemische vingerafdruk?
2. Hoe pak je het inverse probleem (smaak naar chemie) aan als de oplossing niet uniek is?
3. Wat voegen dimensionaliteitsreductie en clustering toe aan de exploratie van bierdata?

De methodesectie beschrijft de dataset en modelkeuzes voor beide richtingen. De resultatenssectie rapporteert de evaluatiemetrieken en visualisaties. Discussie en conclusie plaatsen de bevindingen in de bredere context van het originele onderzoek.

Methode

Dataset

De dataset omvat 250 Belgische bieren, elk beschreven door 227 chemische features (gaschromatografie, spectrofotometrie, biochemische analyses) en 50 sensorische descriptoren gescoord door een getraind expertpanel. De data werden opgesplitst in 175 trainings- en 75 testsamples (70/30, gestratificeerd per bierstijl). De dataset is afkomstig uit Schreurs et al. (2024).

Forward model (chemie naar smaak)

We testten zeven regressiemodellen: vijf tree-based methoden (GradientBoosting, XGBoost, LightGBM, ExtraTrees, RandomForest) en twee lineaire methoden (ElasticNet, PLS). Lineaire modellen faalden door multicollineariteit: de 227 chemische features correleren onderling te sterk. Tree-based modellen zijn robuuster voor dit probleem. GradientBoosting presteerde het best na hyperparameteroptimalisatie via Optuna (200 trials, 5-fold cross-validatie). SHAP-analyse (Lundberg et al., 2020) werd gebruikt om per sensorisch attribuut de meest invloedrijke verbindingen te identificeren.

Inverse model (smaak naar chemie)

Het inverse probleem is ill-posed: 50 sensorische inputs zijn niet genoeg om 227 chemische outputs uniek te bepalen. Een deterministisch model geeft dan een schijnzekerheid die niet bestaat. We kozen voor een Deep Ensemble van vijf Gaussian MLP-netwerken (Lakshminarayanan et al., 2017). Elk netwerk heeft de volgende architectuur:

`

Input (50) → Linear(50,256) → BatchNorm → ReLU → Dropout(0.3)

→ Linear(256,128) → BatchNorm → ReLU → Dropout(0.3)

→ μ-hoofd: Linear(128, 227)

→ σ-hoofd: Linear(128, 227) + softplus + 1e-3

`

Elk netwerk voorspelt per chemische output een Gaussische verdeling $N(\mu, \sigma^2)$. De verliesfunctie is de negatieve log-likelihood:

$NLL = (1/2) \cdot \log(\sigma^2) + (y - \mu)^2 / (2\sigma^2) + \text{constante}$

σ wordt positief gehouden via softplus-activatie met een vloer van 1e-3 om degeneratie te voorkomen. Training verloopt in twee fasen: eerst MSE-warmup, daarna NLL-optimalisatie, met gradient clipping (max_norm=1.0) en early stopping op validatie-NLL. Het ensemble splitst onzekerheid op in twee componenten: aleatarische onzekerheid (d.w.z. intrinsieke dataruis, uitgedrukt als de gemiddelde σ over de vijf leden) en epistemische onzekerheid (d.w.z. onzekerheid door beperkte data, uitgedrukt als de spreiding van de μ-voorspellingen tussen de vijf leden).

Round-trip validatie

Om te controleren of het systeem intern consistent is, werd een round-trip test uitgevoerd: sensorische testdata → inverse model → geschat chemisch profiel → forward model → gereconstrueerde smaakscores. De resulterende R^2 werd vergeleken met die van het directe forward model.

Visualisatie en clustering

UMAP (McInnes et al., 2018) projecteert alle 250 bieren op een tweedimensionale kaart op basis van hun chemische profiel. Hiërarchische clustering (Ward-methode) op de sensorische data levert een dendrogram dat bieren groepeert op smaakgelijkenis in plaats van chemische vingerafdruk. Het dendrogram werd na de sprint review toegevoegd op suggestie van de product owner; tegelijk werd ook de mogelijkheid om eigen bierrecepten op te slaan ingebouwd.

Webapplicatie

Het systeem is ontsloten via een FastAPI-backend en een Next.js-frontend. De applicatie biedt voorspelling in beide richtingen, een interactieve UMAP Beer Map, het dendrogram en opslag van eigen bieren.

Resultaten

Forward model

Tabel 1: Prestaties forward model (chemie naar smaak), gesorteerd op test R^2

Model	CV R^2	Test R^2	Opmerkingen
GradientBoosting	0.164	0.215	Optuna, 200 trials, 5-fold CV — **beste model**
XGBoost	0.135	0.211	Optuna-geoptimaliseerd
ExtraTrees	0.180	0.198	Default parameters
RandomForest	0.153	0.188	Default parameters
LightGBM	0.124	0.177	Optuna-geoptimaliseerd
ElasticNet	-29.86	-0.016	Faalt door multicollineariteit
PLS	-59.01	-0.165	Faalt door multicollineariteit

Alle tree-based modellen vertonen geen overfitting: test R^2 ligt consistent hoger dan CV R^2, wat duidt op pessimistische cross-validatie door de kleine foldgrootte (n=175). GradientBoosting haalde test R^2 = 0.215, consistent met de benchmark van Schreurs et al. (R^2 = 0.22) en bevestigt dat dit het plafond is voor expertpaneldata bij deze omvang. SHAP-analyse gaf interpreteerbare resultaten per attribuut.

Inverse model

Tabel 2: Prestaties inverse model (smaak naar chemie)

Metriek	Waarde
Overall test R^2	-0.02
NLL (test)	2.58
95%-coverage	94.4%
Aleatarische σ (gem.)	1.82
Epistemische σ (gem.)	0.35
Ratio epistemisch/aleatarisch	0.195
Top feature: kcalperc	R^2 = 0.46
Top feature: color.EBC	R^2 = 0.45
Top feature: lactic_acid	R^2 = 0.43

De overall R^2 van -0.02 klinkt slecht. Maar dat is het verwachte resultaat voor dit type probleem: 50 smaakscores bevatten simpelweg niet genoeg informatie om 227 chemische outputs uniek vast te leggen. De relevante vraag is of de onzekerheidsintervallen kloppen, en die doen dat. Een 95%-coverage van 94.4% bij een nominale waarde van 95% is goed gekalibreerd — dat wil zeggen dat 94.4% van de werkelijke chemische waarden binnen het voorspelde 95%-interval valt, wat aantoont dat de onzekerheidsschattingen statistisch betrouwbaar zijn.

Voor een subset van chemische features werkt het ensemble wel degelijk. Kcalperc, color.EBC en lactic_acid halen R^2 boven 0.40: features die sterk samenhangen met waarneembare smaak. Features zonder duidelijk sensorisch signaal blijven onvoorspelbaar.

De epistemisch/aleatarisch ratio van 0.195 geeft aan dat de meeste onzekerheid intrinsiek is (dataruis, de fundamentele niet-uniekheid van het probleem) en niet het gevolg is van een tekort aan

modelcapaciteit. Meer trainingsdata helpt, maar lost het onderliggende probleem niet op.

Round-trip validatie

Smaak → chemie → smaak geeft $R^2 = 0.138$, tegenover 0.215 bij directe forward-voorspelling. Een verlies van 0.087 R^2-punten voor een omweg via het chemische domein valt mee, zeker gezien de negatieve overall R^2 van het inverse model zelf.

Beer Map en clustering

UMAP bracht duidelijke structuur aan het licht. Kriek ($\sigma = 0.31$) en Lambic clusteren compact en goed gescheiden. Die stijlen hebben een heel specifiek chemisch profiel. Blond, Tripel en Strong Ale overlappen grotendeels, wat klopt: ze zijn chemisch nauwer verwant dan hun stijllabels suggereren. De between/within-variantieverhouding bedraagt 0.56, waarbij een hogere waarde betere scheiding tussen stijlen aangeeft. Een ratio van 0.56 wijst op betekenisvolle maar niet-perfecte clustering, wat verwacht wordt voor chemisch verwante bierstijlen.

Het dendrogram op sensorische data geeft een andere groepering dan de UMAP op chemische data. Sommige bieren die chemisch op elkaar lijken, smaken toch anders, en andersom. Die twee perspectieven tegelijk zichtbaar maken was het doel van de visualisatie.

Discussie en beperkingen

De forward resultaten zijn vergelijkbaar met Schreurs et al. Er was ruimte voor verbetering, maar die bleek er niet te zijn: $R^2 \approx 0.22$ lijkt het plafond voor expertpaneldata bij 175 trainingssamples. Meer data zou helpen; betere modellen niet per se.

Het inverse model is eerlijk over zijn beperkingen. Overall $R^2 = -0.02$ is geen mislukking: het model zegt "ik weet het niet zeker" op de juiste momenten, en de coverage-resultaten bevestigen dat. Wat het niet kan, is een unieke chemische samenstelling geven voor een gewenst smaakprofiel. Dat kan principieel ook niet.

De voornaamste bottleneck is de dataset: 175 trainingssamples voor 227 outputs is krap. Uitbreiden met meer bieren (bij voorkeur ook niet-Belgische stijlen) zou de epistemische onzekerheid verkleinen en de bruikbaarheid van het inverse model verbeteren.

Na de sprint review zijn drie uitbreidingen toegevoegd: een verfijnde Beer Map, het dendrogram en de mogelijkheid om eigen bieren op te slaan. Die kwamen voort uit concrete feedback bij gebruik van de applicatie, wat betekent dat het systeem in elk geval bruikbaar genoeg was om echt te testen.

Conclusie

Machine learning kan biersmaak gedeeltelijk voorspellen vanuit chemie. GradientBoosting haalt $R^2 = 0.215$, consistent met de literatuur. Beter wordt het niet met de huidige datagrootte.

Het inverse probleem is aanpakbaar, maar niet oplosbaar. Een Deep Ensemble van Gaussian MLPs geeft geen uniek antwoord (want dat bestaat niet), maar wel een gekalibreerde kansverdeling over de mogelijkheden. De 95%-coverage van 94.4% toont dat die intervallen statistisch te vertrouwen zijn.

UMAP en hiërarchische clustering brengen structuur aan het licht die in de ruwe data niet direct zichtbaar is. De chemische en sensorische clustering geven twee complementaire perspectieven op dezelfde 250 bieren.

Flavor Forge is een proof-of-concept, en dat is precies wat het moet zijn. De beperkingen zijn reëel en gedocumenteerd. Het systeem werkt binnen die grenzen.

Referenties

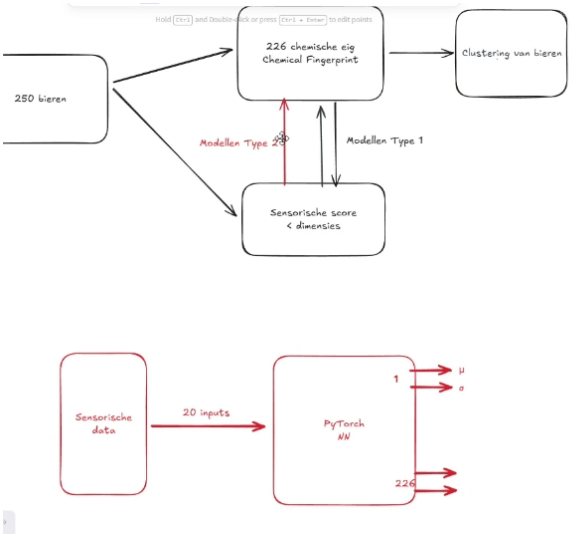
Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, *30*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*, 56–67.

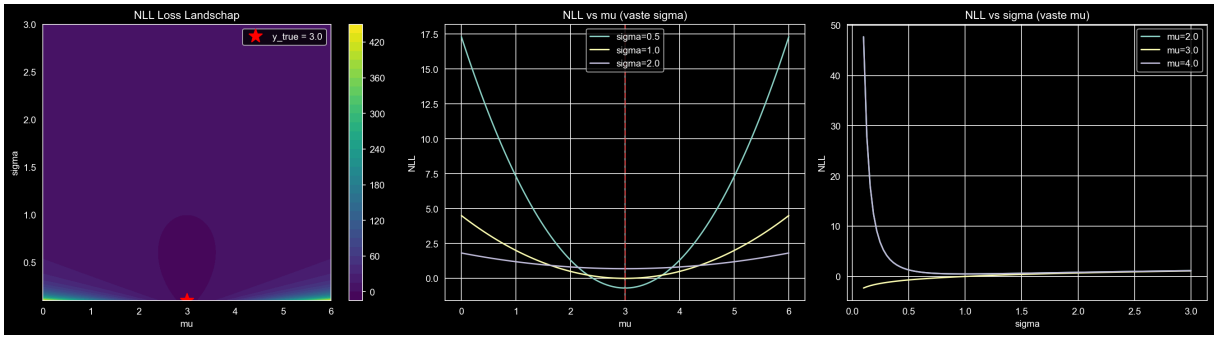
McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Schreurs, M., Piampongsant, S., Roncoroni, M., Cool, L., Herrera-Malaver, B., Vanderaa, C., Theßeling, F. A., Kreft, L., Botzki, A., Malcorps, P., Daenen, L., Wenseleers, T., & Verstrepen, K. J. (2024). Predicting and improving complex beer flavor through machine learning. *Nature Communications*, *15*, 2368.

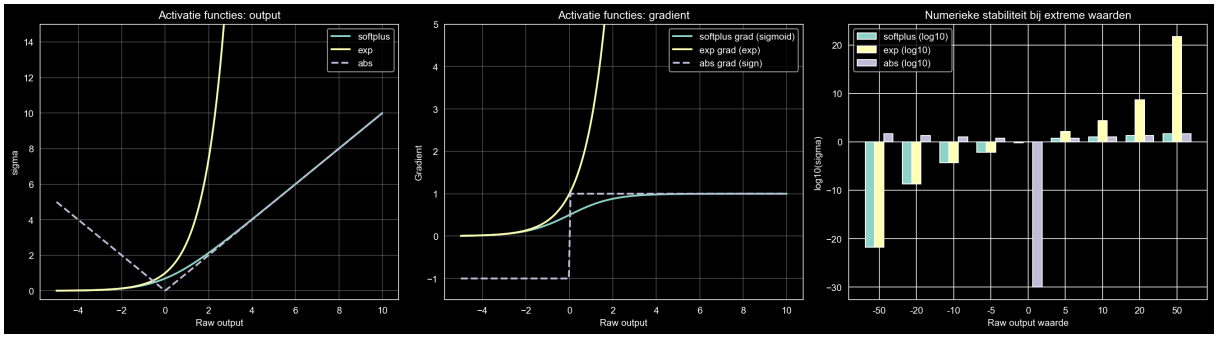
Figurenlijst



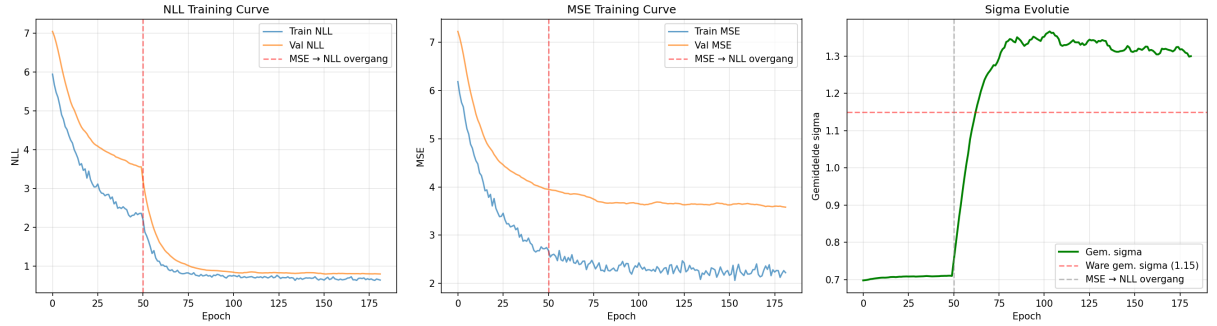
Figuur 1: Architectuurdiagram van het bidirectionele modelleringssysteem (chemie ↔ smaak).



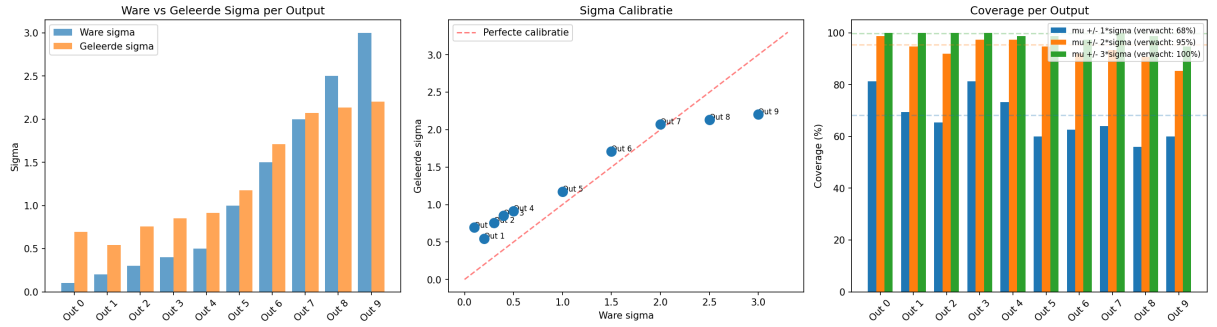
Figuur 2: NLL loss landschap als functie van μ en σ. Links: 2D landschap; midden: NLL vs μ bij vaste σ; rechts: NLL vs σ bij vaste μ.



Figuur 3: Vergelijking van σ-activatiefuncties (softplus, exp, abs): output, gradient en numerieke stabiliteit bij extreme waarden.



Figuur 4: NLL- en MSE-trainingscurves met σ-evolutie. De stippellijn markeert de overgang van MSE- naar NLL-fase.



Figuur 5: Evaluatie van het Gaussian ensemble. Links: ware vs geleerde σ; midden: kalibratiediagram; rechts: coverage per output.