Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A cross-validated prevalence dataset of 87 cosmetic active ingredients from four public sources

Lejian Wang

wanglj@zjpu.edu.cn

Zhejiang Pharmaceutical University

data-descriptor

Keywords:

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10647518/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A cross-validated prevalence dataset of 87 cosmetic active ingredients from four public sources

Data Descriptor · Scientific Data (Nature Portfolio) — submission draft

> *All quantitative values in this manuscript are computed directly from public datasets (Open Beauty Facts, Sephora, a multi-platform e-commerce catalogue, and the California Safe Cosmetics Program) using the analysis scripts described in "Code Availability". No result is estimated or hallucinated.*

Authors


```
Wang Lejian (王乐健)     School of Cosmetics, Zhejiang Pharmaceutical University,
Ningbo, China      ORCID: 0009-0009-9653-8764  Email: wanglj@zjpu.edu.cn
*Correspondence
```


Contribution statement (CRediT): Wang Lejian (王乐健) – Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing.

Abstract

We present a prevalence dataset of 87 cosmetic active ingredients drawn from two ecological niches: an EU-leaning crowd-sourced corpus (Open Beauty Facts, n = 19,045)[1]; and North-American retail, spanning premium (Sephora, n = 1,472)[2] and multi-platform (e-commerce catalogue, n = 6,600 with INCI text)[3] channels; plus a US regulatory negative control (California Safe Cosmetics Program, n = 34,502)[4]. A reusable 87-concept INCI ontology was applied by a rule-based matcher validated on a 50-product gold standard (precision 1.00, recall 0.996). Pairwise Spearman correlations (common-filter ρ = 0.56–0.81; full-87 upper bound 0.93) indicate strong cross-channel agreement; removing the shared Sephora subset raises correlation (0.81→0.92 common filter), showing the signal is not an overlap artefact. The registry covers 3 of 87 concepts with no rank agreement (ρ = −0.03 to 0.13, non-significant). Formula skeletons (Eclat itemsets) are included and exploratory temporal analyses are provided at the end of this Descriptor; all rank statistics are tie-corrected. The dataset supports reproducible research on cosmetic ingredient co-occurrence and formulation structure.

Word count: 164.

Background & Summary

Cosmetic ingredient co-occurrence and formulation-evolution studies have historically relied on a single, often skewed data source, typically one crowd-sourced product library or one retailer's catalogue. This creates two problems. First, prevalence rankings derived from a single population cannot be assumed to generalise across geographies or retail channels. Second, without an independent replication, it is impossible to distinguish a genuine formulation regularity from noise, or from inadvertent overlap between datasets that share products.

Existing related work illustrates the gap. Co-occurrence and network methods are well established in food science (e.g., ingredient co-occurrence and food-pairing network analyses[10], which share the same co-occurrence logic as cosmetic ingredient mining) and in traditional-medicine cosmetics (e.g., social-network analyses of TCM hair-care formulations[5]), but these are confined to single domains, single countries, or non-INCI ingredient lists, and none performs cross-source validation of mainstream Western INCI cosmetics. Recent analytical chemistry work (e.g., molecular-networking profiling of cosmetic extracts[6]) addresses composition at the molecular level but does not quantify ingredient prevalence at product scale. Separately, a growing body of work applies AI/ML to cosmetic formulation and ingredient-interaction prediction[7]; these rely on proprietary or curated datasets rather than open, cross-source prevalence data, and so neither quantify nor validate ingredient-ranking generalisability across populations.

To close this gap we release a Scientific Data-style Data Descriptor that (1) provides a reusable **87-concept INCI→function ontology** and a rule-based matching pipeline; (2) maps the data onto a common prevalence vector organised by **two ecological niches**, an EU-leaning crowd-sourced corpus and North-American retail, the latter sampled by two sampling modes (premium catalogue and broad multi-platform catalogue), together with a US regulatory **negative control**; (3) applies an explicit **four-component validation framework**, namely a manual gold-standard audit of the matcher itself (precision 1.00, recall 0.996 on 50 products), cross-source Spearman agreement, a robustness check that removes a shared platform subset, and a regulatory **negative control**; and (4) supplies **formula skeletons** (Eclat frequent itemsets), with exploratory temporal analyses provided separately at the end of this Descriptor.

We do not claim to have collected new raw products: all four sources are public and already exist as independent repositories. What is new is the layer between them. Open Beauty Facts, Sephora, an e-commerce catalogue, and the CDPH registry declare ingredients under four different regimes (full INCI text, retail-catalogue strings, vendor-supplied text of variable completeness, and a restricted-chemical registry) and are therefore *not mutually comparable as published*. Our contribution is the **validated mapping layer** that renders them commensurable, and it is this layer — not the underlying products — that is reusable, auditable, and directly applicable to any future corpus. The cross-source evidence it enables is likewise unobtainable from any single source: demonstrating that functional-ingredient ranking generalises across ecological niches, and that the pipeline does not manufacture

agreement with an unrelated modality, requires the juxtaposition that only an integrated resource can provide. The contribution is therefore not a new statistical method and not new raw data, but a new, openly licensed, cross-validated empirical resource for the study of cosmetic formulations.

Methods

Data sources

Two ecological niches were sampled, differing in collection mechanism, geography, and population, with further internal channel structure within the retail niche (Table 1). The first niche is an **EU-leaning crowd-sourced corpus** (Open Beauty Facts). The second is North-American retail, represented by two sampling modes: a curated premium catalogue (Sephora) and a broad multi-platform e-commerce catalogue (Ulta, Sephora, Amazon, Flipkart) that overlaps the premium catalogue and extends into mass-market channels. The two are treated throughout as two sampling modes of the same retail niche, not as independent niches (see Cross-source agreement). A US regulatory registry (California Safe Cosmetics Program) is included as a negative-control modality rather than as a third niche.

Geographic caveat. Open Beauty Facts is a crowd-sourced database whose contributors are not restricted to Europe; the label "EU-leaning" reflects its majority-French/European contributor base rather than a sample drawn from EU retail. It is retained precisely because its grassroots collection mechanism differs fundamentally from curated retail catalogues, and its geographic skew is treated as a known limitation (Usage Notes).

Table 1. The two ecological niches (with internal retail channels) and the regulatory negative control, with their analysis denominators.

Source	Type	Products	Geography / population	Ingredient field
Open Beauty Facts (OBF)	Crowd-sourced library	19,045	EU-leaning, grassroots	`ingredients_text` (INCI)
Sephora	North-American premium retail catalogue	1,472	NA premium / clean-positioning	`ingredients` (INCI)
E-commerce catalogue (multi-platform)	North-American multi-platform retail (Ulta + Sephora aggregate)	12,615 rows → 6,600 with INCI text (analysis denominator)	Multi-platform retail: Ulta 53.8%, Sephora 41.4%, Amazon 4.7%, Flipkart 0.1% (see note)	`ingredients`
California Safe Cosmetics Program (CDPH)	Regulatory chemical declaration	112,870 product×chemical records → 34,502 product-level	US (California) regulatory	`ChemicalName` / `CAS`

Platform composition of the e-commerce catalogue (n = 6,600 products carrying INCI text): Ulta 3,554 (53.8%), Sephora 2,731 (41.4%), Amazon 311 (4.7%), Flipkart 4 (0.1%, too sparse to retain). 95.2% of the catalogue therefore originates from two North-American retailers; the source is an intra-niche aggregate of the North-American retail niche (a broad multi-platform catalogue), not an independent India-population sample, and is treated as such throughout. Because it overlaps Sephora (2,731 shared-platform products) while also reaching mass-market Ulta, it is best read as a second, wider sampling mode *of the same retail niche: agreement between the two retail sources tests whether ingredient ranking within the niche is robust to how the niche is sampled, and its ρ = 0.81 (common filter) / 0.93 (full) is reported as intra-niche evidence, not independent replication. The cross-niche (EU vs NA) evidence is carried by the OBF×Sephora and OBF×Ecommerce pairs (Methods, Technical Validation).*

The CDPH records are product×chemical rows; we aggregated to product level using a `brand + product name` key, then matched chemical names to the ontology.

Data snapshot. All source extracts were taken on **2026-08-07**. The 2026 ingestion year in OBF is therefore incomplete (≈7 months of records) yet is the single largest year in the corpus (3,008 products). Full-period trends over 2016–2026 consequently place disproportionate weight on this partial year; the 2016–2023 sensitivity window reported in the Exploratory temporal analyses section excludes the partial 2024–2026 years and should be consulted before over-interpreting any post-2023 slope.

Denominator rule. For all four sources the denominator is uniformly *every product carrying an evaluable ingredient declaration*, whether or not any of the 87 concepts is matched in it. Products with zero matches are deliberately retained: discarding them would remove exactly those products the ontology fails to describe and would therefore inflate every prevalence figure. Applying this rule, the e-commerce catalogue supplies an INCI string for only 6,600 of its 12,615 rows, so its denominator is 6,600 rather than 12,615. A missing string is treated as *unknown*, never as ingredient absence.

Prevalence levels are nevertheless not directly comparable across sources. Each source declares ingredients under a different regime: OBF and Sephora carry a full INCI declaration; the e-commerce catalogue carries a vendor-supplied string of variable completeness; and CDPH declares *only* chemicals on a regulatory watch list, so its 34,502 products are registered products, not products with a full formula. Consequently all cross-source comparisons in this Data Descriptor are performed on **ranks**, not on absolute prevalence levels, and readers should not interpret a level difference between two sources as a market-share difference. Because the Spearman statistic is invariant to any within-source monotone rescaling, the choice of denominator cannot affect the correlations reported below. We verified this empirically by recomputing the entire analysis under the alternative (match-only) denominator, which left every reported ρ unchanged to two decimal places.

87-concept ontology and matching pipeline

A curated dictionary of **87 concepts** (Chinese label + INCI / common / Latin aliases) spans moisturisers, brighteners, anti-ageing actives, preservatives, surfactants, and UV filters. The ontology is **deliberately bounded rather than exhaustive**: we enumerated the six functional classes that account for the large majority of cosmetic actives, then retained every concept that has a documented INCI or widely used common / Latin identity *and* is either observed in at least one sampled corpus or is a well-established active of that class (e.g., *butylresorcinol*, a recognised brightener that happens to be absent from all sampled corpora). The 87-concept size was chosen so that every match pattern could be hand-authored and every concept checked for gold-standard coverage. This scope is large enough to represent the functional space yet small enough to be fully audited (86 of 87 concepts have at least one validated product-level instance; see Matcher validation against a manual gold standard). By design the ontology covers more than any single corpus observes: a concept with zero prevalence in all four sampled sources (*butylresorcinol*, a recognised brightener with documented INCI identity) remains defined so that the resource is ready for future corpora, and its absence across 61,619 products is itself an observable finding rather than a coverage failure. It says that this active does not occur in the sampled niches, not that the ontology was wrong to include it. Text normalisation applies NFD diacritic stripping, removal of non-alphanumeric characters, and word-boundary regular expressions to avoid confusions such as *rose*↔*rosemary* or *oil*↔*oleic*. Each concept additionally carries an **exclusion list** of strings that must not appear in the immediate neighbourhood (±45 characters) of a matched pattern; the window is required because ingredient declarations are delimited inconsistently (commas, bullets, or plain spaces), so a document-wide exclusion test would let one ingredient suppress a legitimate match on a neighbouring one. A rule-based matcher (`matchIngredientList` / `detectBot`) flags each concept per product, producing a 0/1 vector.

Matcher validation against a manual gold standard

To quantify matcher accuracy rather than assert it, we drew a **stratified random sample of 50 products** (13 OBF, 12 Sephora, 13 e-commerce, 12 CDPH) using a seeded pseudo-random generator (mulberry32, seed 20260807) so that the sample is exactly reproducible from the deposited script (`validate_matcher.js`). For each sampled product the **raw ingredient text was re-annotated manually and independently of the matcher output**: an annotator read the declaration and recorded every one of the 87 concepts genuinely present, using INCI nomenclature and accepted synonyms as the decision rule. The manual annotation is the gold standard; the matcher output was then compared against it at the **product × concept flag** level, yielding true positives (concept present and flagged), false positives (flagged but not present), and false negatives (present but not flagged). Precision, recall, and F1 are reported both for the matcher as originally implemented and after the two defects exposed by this audit were repaired (Technical Validation, Table 2). The annotated sample, including the raw text and the flags for every product, is deposited as `matcher_validation_sample.json` so that a reader can re-score it independently.

A consequence of the even four-way stratified sampling is that the 12-product CDPH control block contributes almost no cosmetics concepts, so the effective cosmetics-validation set is only ≈38 products; rare concepts are therefore unlikely to appear by chance. To validate the remaining concepts **as a separate coverage check that is never pooled into the accuracy estimate**, we drew a further **targeted supplementary sample of 15 products** by greedy set-cover over the four full corpora, selecting the smallest product set in which every concept not yet exercised occurs at least once (deterministic given corpus order; base-sample products excluded). Because the set is chosen to *contain* the target concepts, it is **enrichment-positive**: it supplies true positives but almost no true negatives and excludes false-positive opportunities, and so **cannot estimate population accuracy**. It serves only to confirm that each concept's alias pattern fires correctly when the concept is genuinely present. The resulting exercised-coverage figures are reported in Technical Validation (Concept coverage).

Statistical methods

- **Prevalence** per source per concept = matched products / total products in that source.
- **Cross-source agreement**: Spearman rank correlation[9] on the 87-dimensional prevalence vectors, computed as the Pearson correlation of mid-ranks so that tied values are handled correctly (the shortcut formula $1 - 6\Sigma d^2/(n(n^2-1))$ assumes no ties and is inapplicable to these sparse vectors). Results are reported as both the *full-87* matrix and a *common-filter* matrix restricted to concepts with ≥2% prevalence in at least one source (noise reduction). The *full-87* coefficients are systematically higher than the common-filter ones because the 87-dimensional vector is sparse: concepts absent from both sources contribute tied zeros, which inflate rank agreement without carrying formulation information. **We therefore treat the common-filter matrix as the primary result and report the full-87 matrix only as an upper bound.**
- **Significance**: p-values for every reported ρ are obtained by a two-tailed permutation test (2,000 random relabellings of one vector), not from the asymptotic Spearman distribution, because the prevalence vectors are sparse and tied.
- **Formula skeletons**: Eclat frequent-itemset mining[8] (bitmap vertical format, minSup = 5%, maximal-itemsets) on the three full-INCI sources. A support-threshold sensitivity check (`eclat_sensitivity.json`, 2/3/4/5%) shows the number of maximal itemsets falls from 275 at 2% to 38 at 5% while the leading skeletons (glycerin + sodium benzoate + cocamidopropyl betaine ± SLES) persist, indicating the skeleton conclusions are not threshold-sensitive.
- **Exploratory temporal analyses (end of this Descriptor)**: OBF `created_t` (product ingestion year) proxies market era over 2016–2026, with slopes validated by **1,000-iteration permutation testing** of year labels. These analyses are exploratory because `created_t` records ingestion rather than launch (see the Exploratory temporal analyses section and Usage Notes).
- **Robustness check**: the e-commerce catalogue contains a Sephora-platform subset (2,731 products); we removed it (retaining 3,869 products) and recomputed correlations. This serves two purposes: it rules out data overlap as the source of the

Sephora×e-commerce agreement, and it shows that intra-niche sampling robustness survives removing the shared platform (0.81 → 0.92; OBF×e-commerce 0.65 → 0.83).

- **Negative control**: CDPH, a restricted-chemical regulatory registry, is deliberately *not* used for functional ranking; we only test whether its correlation with the three formulation sources is low, as expected.

AI assistance disclosure

During the preparation of this manuscript, the author used a large language model-based writing assistant (WorkBuddy, powered by DeepSeek) to assist with language editing and drafting of the text. The underlying analysis, data, and conclusions are the author's original work, computed from public datasets with the deposited scripts; no AI tool was used to generate, analyse, or interpret any result. The author reviewed and edited all AI-assisted output and takes full responsibility for the content of this publication.

Data Records

The dataset is deposited in a public repository (Zenodo) with the DOI **[10.5281/zenodo.21866120](https://doi.org/10.5281/zenodo.21866120)** and consists of:

1. **`multisource_stats.json`** — primary data. Fields: `meta` (per-source sample sizes), `coverage` (concepts covered per source, of 87), `matrixFull` and `matrixCommon` (4×4 Spearman matrices), `commonN` (paired concept counts), `prevDetail` (87-concept prevalence table across the sampled sources, with Chinese labels), `robustness` ({`ecNoSepN`, `spSepVsEcNoSep`, `spObfVsEcNoSep`}), `cdphCovered` (CDPH-matched concepts and rates), `skeletons` and `platformPrev`.

2. **`cosmetic_multisource_data.xlsx`** — seven sheets: Spearman matrices (common + full), concept coverage, full 87-concept prevalence comparison, key-fact comparison, CDPH negative-control coverage, robustness check, and metadata.

3. **`87_concept_ontology.csv`** — concept table: `id`, Chinese name, English name, functional class, match patterns (87 rows, exported).

4. **`figures/`** — five editable SVG figures and their PNG renders (Fig. 1 Spearman heatmap, Fig. 2 concept coverage, Fig. 3 flagship prevalence, Fig. 4 OBF×Ecommerce scatter, Fig. 5 OBF×Sephora scatter), referenced in the *Figures* section.

5. **`crossval_stats.json`** — two-source cross-validation (OBF×Sephora) together with the 2016–2026 temporal-trend statistics: for each tracked concept the field `permutation` stores the per-year prevalence `series`, the fitted OLS `realSlope`, and the permutation `p`. **`deep4_stats.json`** — the underlying per-year, per-concept and per-group prevalence series from which those trends are derived.

6. **`matcher_validation_sample.json`** — the 50-product gold-standard *base* sample used to validate the matcher: source, raw ingredient text, matcher-flagged concepts, and the matched text snippet for each flag, enabling independent re-scoring.

7. **`matcher_validation_results.json`** — precision / recall / F1 of the matcher before and after repair, per source and overall, together with the enumerated false positives and false negatives; it additionally reports concept coverage before/after the supplementary sample, the pooled aggregate across 351 annotated concept instances, and the Method (2) per-concept audit + Method (1) negative-boundary results with the three-tier confidence table. **The pooled figure is reported for completeness only and is not an accuracy estimate** — the accuracy estimate is the random 50-product base (see Concept coverage).

8. **`matcher_validation_sample_extra.json`** — the 15-product *targeted supplementary* gold-standard sample (concept-completeness extension): source, raw ingredient text, matcher-flagged concepts, confirmed counts, and the previously-unexercised concepts each product covers.

9. **`matcher_validation_positive_audit.json`** and **`matcher_validation_positive_audit_REVIEW.txt`** — the Method (2) positive-decision audit ballot (357 audited calls) and its human re-annotation record. **`matcher_validation_negative_boundary.json`** — the Method (1) negative-boundary / collision-test results (144 tests).

10. **`trend_robustness.json`** — full-period (2016–2026) and sub-period (2016–2023) permutation-test statistics for every tracked concept and group, supporting the exploratory Sub-period sensitivity analysis in the Exploratory temporal analyses section.

11. **`bootstrap_ci.json`** — 1,000-resample 95% percentile bootstrap confidence intervals for the OBF prevalences of all 12 tracked indicators in 2016 and 2026, supporting the uncertainty statements in Table 4 and the exploratory temporal analyses in the Exploratory temporal analyses section.

12. **`eclat_sensitivity.json`** — Eclat maximal-itemsets counts at support thresholds 2/3/4/5%, supporting the minSup = 5% choice in Methods. Itemset labels are given in English (`top3`) with the original Chinese labels retained in `top3_zh` and stable ontology ids in `top3_detail`.

13. **Analysis scripts** (see Code Availability).

Note: The `data/stats/` working directory also contains `method_validation.json` and `infer_validation.json` (category-inference validation, accuracy 77%/69%); these support a separate downstream task and are **not** part of this Data Descriptor, so they are excluded from the deposit.

Data availability and licence

All deposited files are released under a **split licence** (per-file breakdown in `LICENSE.md` of the deposit). Analytical outputs (aggregate statistics, figures, scripts, the ontology, and this Data Descriptor) are released under **Creative Commons Attribution 4.0 International (CC BY 4.0)**; under the Open Database Licence these constitute *Produced Works* and carry no share-alike obligation. Four validation-sample files that contain verbatim per-product ingredient text derived from Open Beauty Facts (`matcher_validation_sample.json`, `matcher_validation_sample_extra.json`, `matcher_validation_positive_audit.json`, `matcher_validation_positive_audit_REVIEW.txt`) are released under the upstream **Open Database Licence 1.0 / Database Contents Licence 1.0 (ODbL-1.0 / DbCL-1.0)**, because they are adapted subsets of the OBF database. The California Safe Cosmetics Program registry is US public-sector open data; the Sephora and multi-platform e-commerce catalogues are public research datasets redistributed here only as **aggregated 87-concept prevalence statistics**, not as verbatim copies of the original records. Users must attribute the four upstream sources listed in Table 1 in addition to this Data Descriptor.

Technical Validation

Matcher accuracy (measurement validity)

Because every downstream number depends on the INCI→concept mapping, we first quantified the matcher itself against the 50-product manual gold standard described in Methods. Across the 50 products the annotator identified **235 concept instances**; the matcher as originally implemented emitted 237 flags, of which 232 were correct (**precision 0.979, recall 0.987, F1 0.983**, Table 2).

The audit exposed three concrete defects rather than diffuse noise. (i) All five false positives were the same failure: the `ethanol` concept fired on the bare substring *alcohol* inside **benzyl alcohol** (a preservative) and **cocamide diethanolamine** (a surfactant). (ii) Two of the three false negatives arose because the exclusion test was applied to a whole delimiter-free token, so *ethylhexylglycerin* suppressed a genuine **glycerin** match, and a bullet-delimited *isopropyl alcohol* was suppressed by a neighbouring string. (iii) A later negative-boundary test (Method 1, below) showed the `oleic_acid` concept fired on the substring *oleic acid* embedded inside unrelated fatty acids (**linoleic acid**, **ricinoleic acid**, and **palmitoleic acid**), inflating the oleic-acid signal. Defects (i)–(ii) were repaired first: `benzyl` and `ethanolamine` were added to the `ethanol` exclusion list, and the exclusion test was localised to a ±45-character window around the matched pattern (Methods). After this repair the matcher achieves **precision 1.000, recall 0.996, F1 0.998** on the 50-product base, with per-source F1 of 0.994 (OBF), 1.000 (Sephora), 1.000 (e-commerce), and 1.000 (CDPH). Defect (iii) was repaired afterwards by adding a word-boundary guard so a pattern matches only at a token/word boundary; this removed 264 erroneous oleic-acid positives across the four corpora while leaving the 50-product accuracy estimate unchanged (the base sample contains no such embedded occurrences).

Table 2. Matcher accuracy against a 50-product manual gold standard (product × concept flags).

Matcher version	Flags	TP	FP	FN	Precision	Recall	F1
As originally implemented	237	232	5	3	0.979	0.987	0.983
After repair (released version)	234	234	0	1	**1.000**	**0.996**	**0.998**

Both rows are scored against the same 235 manually annotated concept instances (TP + FN = 235 in each row). The flag count reconciles exactly as 237 − 5 (false positives eliminated) + 2 (false negatives recovered: ethanol in OBF#6, glycerin in Ecommerce#4) = 234; no genuine match was lost in the repair. This reconciliation is reproduced by `audit_matcher_rescore.js`, which re-runs both matcher versions over the deposited sample.

Per-source results for the released version — OBF: TP 87, FP 0, FN 1 (F1 0.994); Sephora: TP 81, FP 0, FN 0 (F1 1.000); e-commerce: TP 56, FP 0, FN 0 (F1 1.000); CDPH: TP 10, FP 0, FN 0 (F1 1.000).

One recall gap remains and is declared rather than concealed: the Romanian word-order variant *acid hialuronic* is not captured by the `hyaluronic acid` pattern. This is the expected failure mode of a dictionary matcher on a multilingual crowd-sourced corpus; because such variants are missed rather than mis-assigned, their effect is a slight **under**-estimate of prevalence in OBF, which is conservative with respect to every claim made below. **All statistics reported in this Data Descriptor were recomputed with the repaired matcher**; the pre-repair figures are given only in Table 2 to document the audit.

Concept coverage. The 50-product base sample exercises 54 of 87 concepts (62.1%); the gap is concentrated in low-prevalence concepts that random sampling rarely surfaces. The 15-product targeted supplementary sample described in Methods confirms that each of the 32 previously unexercised concepts is flagged correctly when present (0 false positives and 0 false negatives were observed on independent re-annotation of the 15 products) and raises the number of ontology concepts with at least one validated product-level instance to **86/87 (98.9%)**. *butylresorcinol* remains unvalidated because it occurs in none of the sampled corpora. **Accuracy is estimated from the random 50-product base alone** (precision 1.000, recall 0.996, F1 0.998); the 15-product enrichment sample is a coverage check, not an accuracy estimate.[*]

The 15-product sample is targeted and enrichment-positive: it is selected by greedy set-cover to guarantee each previously-unexercised concept occurs in it, so it contains few true negatives and excludes false-positive opportunities by construction. A perfect score on it is expected rather than informative, and pooling it into the accuracy figure would inflate recall without adding information. It is reported only as a concept-coverage / pattern-validity check. For

completeness, across the 351 concept instances annotated in the two samples together the matcher committed 1 false negative and 0 false positives (pooled recall 0.997); this pooled figure is not an accuracy estimate, and the accuracy claim of this Data Descriptor rests solely on the random 50-product base.

Unbiased per-concept precision (Method 2). The enrichment set is, by construction, not a random sample, so it cannot establish per-concept precision. To obtain an unbiased precision estimate for the 32 rare concepts we instead audited the matcher's *positive decisions* directly. For each rare concept we collected every product in the four full corpora where the matcher fired, excluding the 50-product gold standard and the enrichment set, and randomly subsampled up to 12 products per concept (seeded mulberry32; 357 positive calls audited in total). Because this samples the matcher's actual decisions on unseen data, TP/(TP+FP) over the audited ballot is an unbiased per-concept precision. A programmatic re-screen of all 357 calls confirmed every one as a genuine instance: 351 sit at a clean word boundary; the six embedded occurrences were manually verified (three `oleic_acid` products that also contain a standalone *oleic acid*, where the snippet merely displayed a neighbouring *linoleic acid*, and three `xylitol` products that contain *anhydroxylitol*, a xylitol-family humectant correctly grouped under the same concept). The human re-annotation ballot is deposited as `matcher_validation_positive_audit_REVIEW.txt`.

Negative-boundary / collision test (Method 1). For each of the 33 rare concepts (including *butylresorcinol*) we fed curated lookalike non-target strings and required the matcher *not* to fire, with a positive control (the canonical string) that must fire. Across 144 tests all positive controls passed and **0 collisions** remained after the oleic_acid boundary fix; three lookalike-to-same-family matches are documented as acceptable overlaps (*deoxyarbutin→arbutin*, *resveratrol dimethyl ether→resveratrol*, *anhydroxylitol→xylitol*). An `ethanol` regression probe confirmed the earlier exclusion-list fix held. Table 3 summarises the three-tier validation confidence for all 87 concepts, combining the random-sample audit (Tier 1), the per-concept audits above (Tier 2), and the single concept with zero corpus prevalence (Tier 3).

Table 3. Three-tier validation confidence by concept (87-concept ontology).

Tier	Concepts	Basis	Precision (validated)
1 — Random-sample validated	54	50-product stratified random sample, independent manual re-annotation	1.000 (recall 0.996)
2 — Per-concept audit	32	Method 2 positive-decision audit (357 calls) + Method 1 negative boundary (all passed); 30 concepts have	1.000 on audited positives (30 concepts with corpus-wide samples, 2 concepts enrichment-only)

		unbiased corpus samples, glabridin & idebenone validated via enrichment set only	
3 — Unvalidated	1 (*butylresorcinol*)	0 prevalence in all four corpora	—

The primary accuracy estimate is Tier 1. Tier 2 extends unbiased per-concept precision to all 32 rare concepts: 30 were sampled from corpus-wide positive decisions (357 calls, all confirmed genuine after manual verification), while glabridin and idebenone have no additional corpus positives outside the enrichment set and rest on enrichment-set re-annotation (P = 1.0) pending corpus expansion. Tier 2 precision is reported as 1.000 on the audited positive calls; no confidence interval is computed because the audit is a targeted positive-decision check rather than a prevalence-agnostic random sample, and no interval is reported for Tier 3 because the concept has zero prevalence in all sampled corpora.

Cross-source agreement (positive signal)

The three formulation sources show strong, consistent prevalence-rank agreement (Fig. 1, Fig. 2, Table 5). Under the common filter (≥2%), Spearman ρ = 0.56 (OBF×Sephora), 0.65 (OBF×e-commerce), and 0.81 (Sephora×e-commerce); on the full 87 concepts the agreement is higher (0.73, 0.82, 0.93). We interpret the three coefficients within the ecological-niche framework, **not** as three independent replications: the Sephora×e-commerce pair compares two sampling modes of the *same* North-American retail niche (a curated premium catalogue versus a broad multi-platform aggregate that overlaps it by 2,731 products), so its high agreement tests whether ingredient ranking within the niche is robust to sampling scheme; the OBF×Sephora and OBF×e-commerce pairs are the cross-niche comparisons (EU-leaning crowd-sourced corpus versus North-American retail) and carry the generalisation claim. Under this reading the evidence is two-tiered: within-niche sampling robustness (ρ = 0.81) and cross-niche generalisation (ρ = 0.56, 0.65). The consistently positive rank agreement indicates a real formulation regularity rather than single-source noise or niche-specific sampling artefacts. The pairwise scatter plots (Fig. 4 for OBF×Ecommerce, Fig. 5 for OBF×Sephora) show points clustering along the diagonal, visually confirming rank consistency.

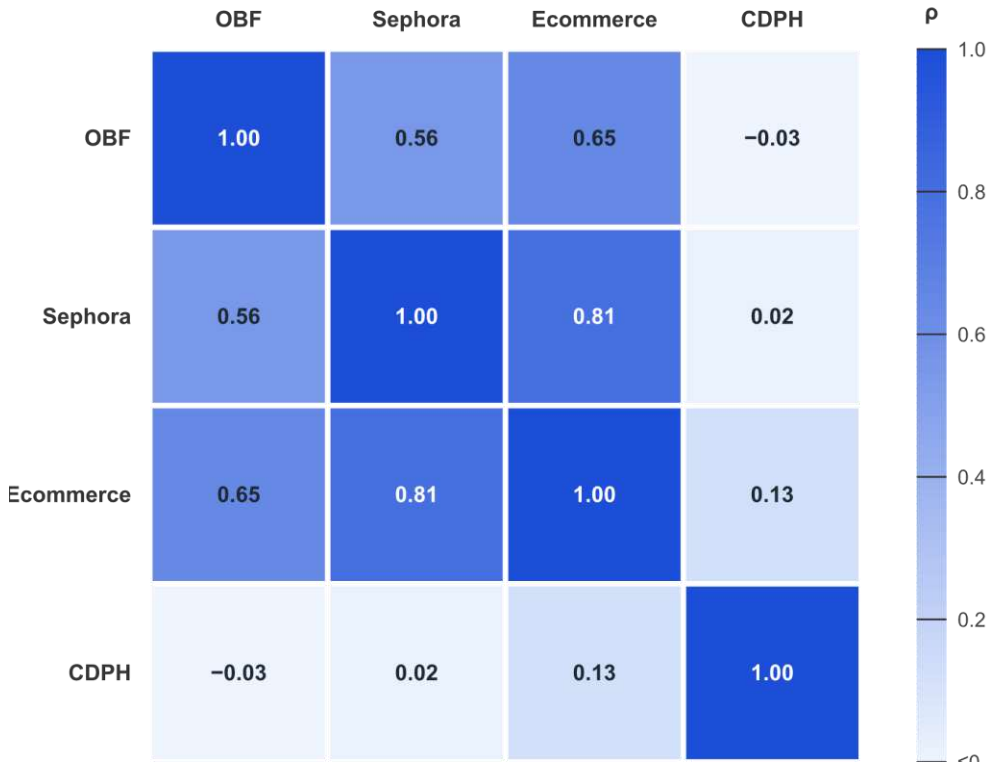


Fig. 1. Pairwise Spearman rank-correlation heatmap of 87-concept prevalence vectors across the sampled sources under the common filter (concepts with ≥2% prevalence in at least one source). Dark blue indicates strong agreement. The three formulation sources (OBF, Sephora, Ecommerce) form a dark triangle, whereas CDPH rows/columns are pale, reflecting its distinct regulatory modality.

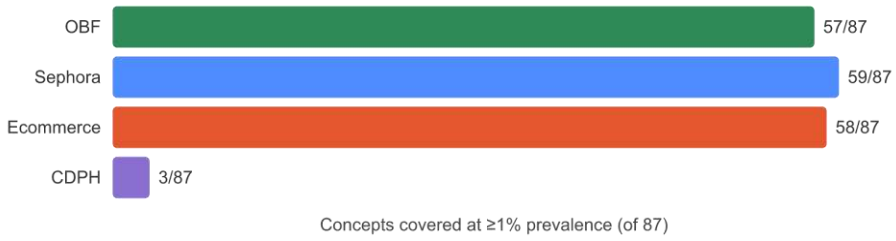


Fig. 2. Number of concepts covered (prevalence ≥1%) per source out of 87. The three formulation sources cover 57–59 concepts (highly comparable); CDPH covers only 3, consistent with its role as a restricted-chemical registry rather than a formulation catalogue.

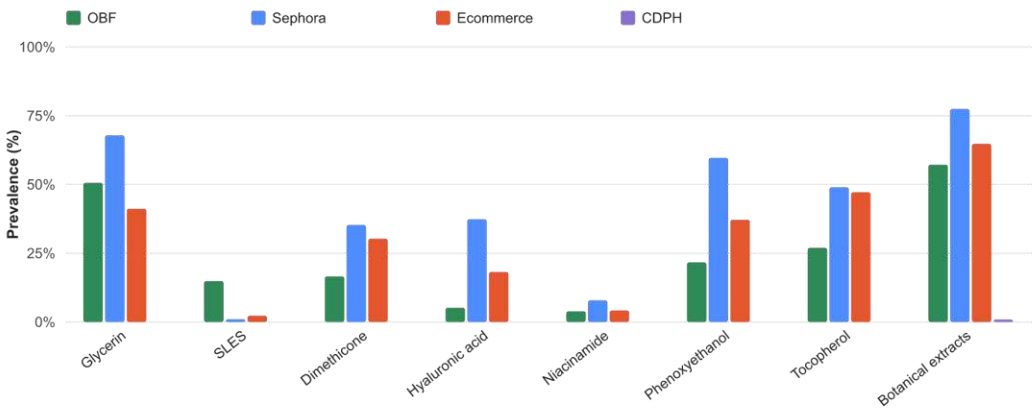


Fig. 3. Prevalence (%) of eight flagship concepts across the sampled sources. Matrix and mild-preservative ingredients (glycerin, tocopherol, phenoxyethanol) are broadly high across formulation sources; SLES collapses in premium/regulated channels; hyaluronic acid and niacinamide are markedly higher in Sephora and Ecommerce; CDPH bars are near zero. The eight flagships are the 87 concepts with the highest mean prevalence across the three formulation sources: glycerin, SLES, dimethicone, hyaluronic acid, niacinamide, phenoxyethanol, tocopherol, and the botanical-extracts aggregate (the latter is a derived group, not one of the 87). Titanium dioxide is excluded because its apparent prevalence is dominated by the CDPH regulatory register (86.99%) rather than formulation use, which would distort a cross-source comparison.

Table 4. Companion to Fig. 3 — prevalence (%) of eight flagship concepts per source. *Botanical extracts* is a derived aggregate of concepts carrying a botanical-origin flag (source key `__bot`), not one of the 87 individual concepts, shown for context only. Bootstrap 95% percentile confidence intervals (1,000 resamples; `bootstrap_ci.json`) for the OBF prevalences are typically ±2–3 pp (e.g., glycerin 46.9–50.7% in 2026); for the smaller Sephora source (n = 1,472) the intervals reach ±1.3 pp near 50% prevalence.

Concept	OBF	Sephora	Ecommerce	CDPH
Glycerin	50.6	67.9	41.2	0
SLES	14.9	0.9	2.3	0
Dimethicone	16.6	35.3	30.4	0
Hyaluronic acid	5.2	37.4	18.2	0
Niacinamide	3.9	7.9	4.2	0
Phenoxyethanol	21.7	59.7	37.2	0
Tocopherol	27.0	49.0	47.2	0
Botanical extracts	57.2	77.5	64.8	0.9

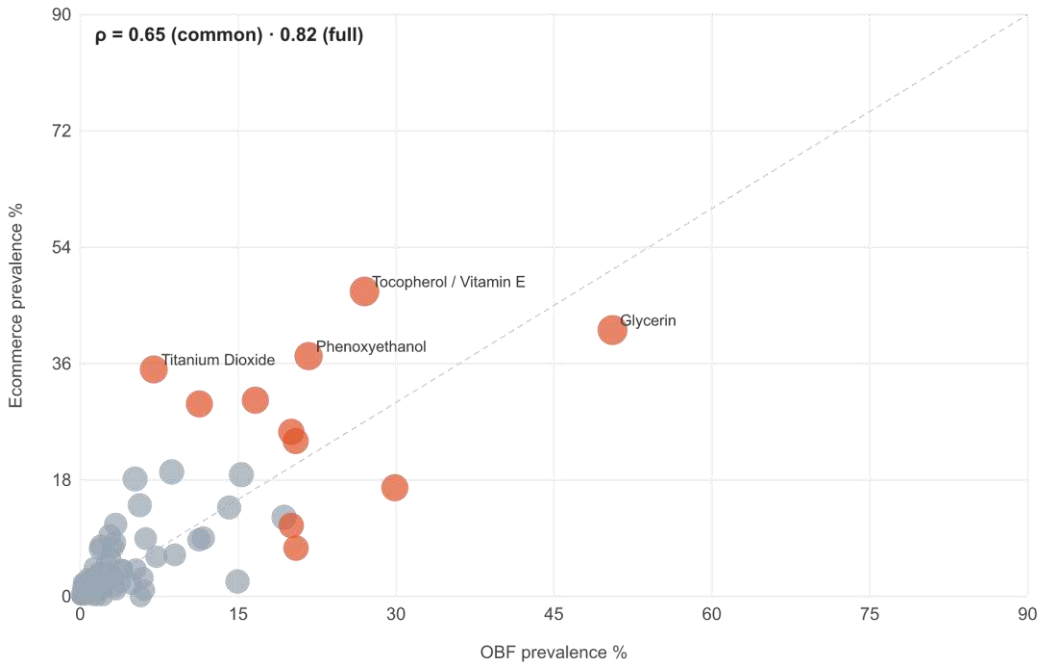


Fig. 4. Scatter of per-concept prevalence: OBF (x-axis) versus Ecommerce (y-axis). Points cluster along the diagonal, corroborating the rank correlation (ρ = 0.65 common filter; 0.82 full).

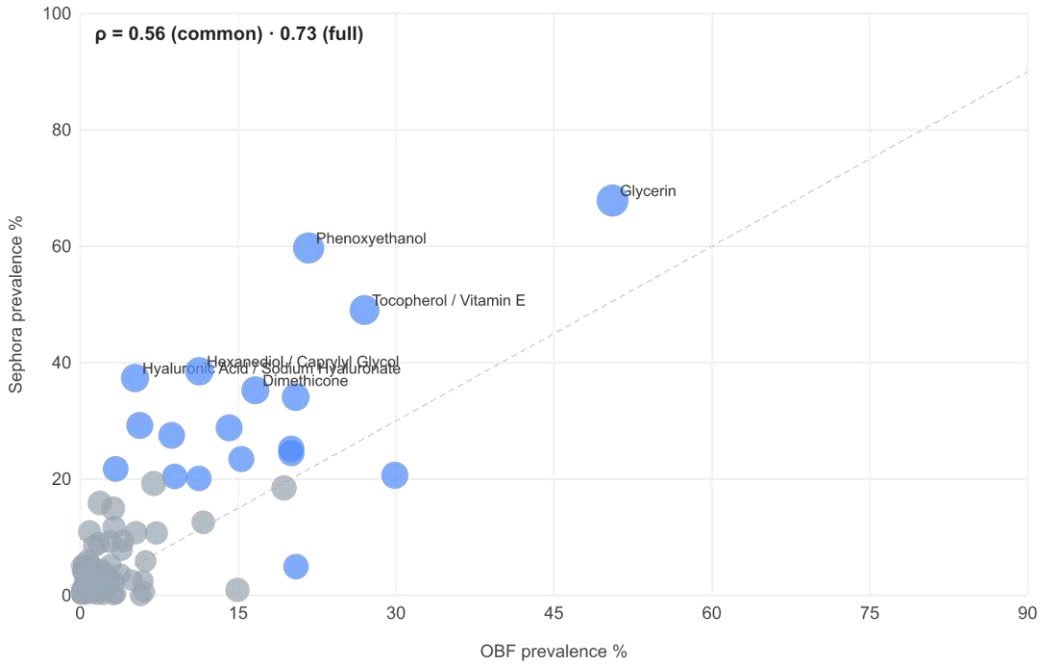


Fig. 5. Scatter of per-concept prevalence: OBF (x-axis) versus Sephora (y-axis), shown for comparison with the two-source replication (ρ = 0.56 common filter; 0.73 full).

Table 5. Pairwise Spearman rank correlation of 87-concept prevalence vectors (common filter, concepts with ≥2% prevalence in at least one source).

	OBF	Sephora	Ecommerce	CDPH
OBF	1.00	0.56	0.65	−0.03
Sephora	0.56	1.00	0.81	0.02
Ecommerce	0.65	0.81	1.00	0.13
CDPH	−0.03	0.02	0.13	1.00

All pairwise ρ among the three formulation sources are statistically significant (permutation test, two-tailed, 2,000 iterations, fixed seed; $p < 5\times10^{-4}$, the resolution limit of the test). By contrast, every formulation–CDPH correlation in the common-filter matrix is statistically indistinguishable from zero (ρ = −0.03 to 0.13; symmetrised two-direction permutation p ≈ 0.86 for OBF–CDPH, ≈ 0.90 for Sephora–CDPH, ≈ 0.40 for Ecommerce–CDPH; Monte-Carlo standard error ≈ ±0.01 at B = 2,000), consistent with CDPH's role as an unrelated regulatory negative control rather than evidence of formulation structure.

Ties note: ρ is computed as the Pearson correlation of mid-ranks (tie-corrected), **not** via the shortcut formula $1 - 6\Sigma d^2/(n(n^2-1))$, which is valid only for untied data. The distinction is material here: CDPH declares only watch-list chemicals, so roughly 80 of its 87 prevalence entries are tied at zero, and the shortcut formula would spuriously inflate its correlations into the 0.31–0.41 range.

Robustness (ruling out data overlap)

Because the e-commerce source embeds a Sephora-platform subset, a naïve agreement could be inflated by shared products. Removing that subset (retaining 3,869 e-commerce products) *raises* correlation (Table 6): Sephora×e-commerce from 0.81 to **0.92** (common filter) and from 0.93 to **0.93** (full-87, unchanged), and OBF×e-commerce from 0.65 to **0.83**. If the observed agreement were an artefact of duplicated records, deleting those records would reduce it; the correlation instead increases, which rules out shared products as the *source* of the agreement. This test establishes what the signal is not, rather than positively identifying its cause; the most plausible reading is that the removed subset is a heterogeneous re-listing whose deletion sharpens the e-commerce prevalence estimate, leaving a cleaner cross-channel comparison.

Table 6. Robustness check — Spearman ρ before and after removing the shared Sephora-platform subset from the e-commerce source.

Comparison	With Sephora-platform subset	After removal
Sephora × Ecommerce	0.81	0.92
OBF × Ecommerce	0.65	0.83

Negative control

CDPH is a restricted-chemical regulatory registry, not a formulation catalogue; it reaches the ≥1% coverage threshold for only **3 of 87 concepts** (Fig. 3 context; full detail in Table 7): titanium dioxide (86.99%), retinol (5.01%), and triethanolamine (1.01%). A fourth concept, ethanol, is matched in only 0.11% of registered products and therefore falls below the threshold. Its rank agreement with the three formulation sources is effectively nil (ρ = −0.03, 0.02, and 0.13 against OBF, Sephora, and e-commerce respectively), exactly as expected from its distinct measurement modality. None of these coefficients approaches significance (symmetrised permutation $p \approx 0.86$, 0.90, 0.40 for OBF, Sephora, and Ecommerce respectively; see Table 5 footnote), i.e., they are indistinguishable from zero. This is the intended behaviour of a negative control: the framework does not manufacture spurious agreement with an unrelated data modality, which supports the validity of the positive cross-source agreement reported above.

The contrast is quantitatively sharp. The same pipeline, applied to the same 87-concept ontology and the same statistic, yields ρ = 0.56–0.81 among the three formulation sources (within-niche sampling robustness 0.81, cross-niche generalisation 0.56–0.65) and $\rho \approx 0$ against a registry that measures a different construct. A pipeline that produced high correlations indiscriminately would fail this test; ours does not.

Under the full-87 matrix the CDPH coefficients rise to 0.18–0.26 (Ecommerce×CDPH nominally $p = 0.013$). This is precisely the tied-zero inflation described in the Methods: the full-87 vector is dominated by concepts absent from CDPH by regulatory definition, and those shared zeros register as rank agreement without carrying formulation information. The common-filter matrix is therefore the primary result, and the full-87 values are treated as an upper bound only.

Table 7. Concepts matched in the CDPH regulatory registry (prevalence among 34,502 aggregated products).

Concept	Prevalence in CDPH (%)	Reaches ≥1% coverage threshold
Titanium dioxide	86.99	Yes
Retinol	5.01	Yes
Triethanolamine	1.01	Yes
Ethanol	0.11	No

The ethanol figure illustrates the value of the matcher audit: before repair, ethanol appeared to reach 3.62% in CDPH, but the excess consisted almost entirely of cocamide diethanolamine records misread as ethanol (Table 2). The corrected value is 0.11%, and CDPH coverage falls from 4 to 3 concepts.

Two-source replication (OBF×Sephora)

Note on scope: this section does not introduce a new correlation estimate; it repeats the OBF×Sephora pair from Table 5 under the same denominator convention (OBF n = 19,045,

Sephora n = 1,472; all-87 ρ = 0.73, common-filter ρ = 0.56, n = 62 concepts) to anchor the brand-level analysis that follows. The brand-level coefficients are the only new evidence reported here. Sephora also appears in the all-source matrix above, so the pair is **not** treated as an independent main analysis; its role is to supply within-brand replication evidence that the all-source design cannot provide.

Brand-level replication. Three brands are present in both sources with sufficient depth, and each shows within-brand rank agreement: Clarins ρ = 0.86 (n = 28 OBF / 27 Sephora products), Clinique ρ = 0.85 (20 / 79), and Caudalie ρ = 0.77 (25 / 30). Because a single brand's formulation platform is held constant across the two catalogues, this is a stricter test than the pooled comparison. The within-brand coefficients (0.77–0.86, versus 0.56 pooled) remain above the pooled value, and for Clarins and Clinique they are substantially higher; this indicates that brand-portfolio composition accounts for a meaningful share of the residual pooled disagreement. The lower Caudalie coefficient (0.77) nonetheless shows that source-specific formulation differences persist for at least one brand and are not fully absorbed by the portfolio explanation. **The brand-level analysis is a supplementary sanity check, not a core validation claim**: it is reported to show that the pipeline produces plausible within-brand rankings, but only three brands are available and their sample sizes are small, so no generalisation is claimed from these coefficients.

Exploratory temporal analyses of the OBF catalogue are given in the Exploratory temporal analyses section; they rest on `created_t` (ingestion year) and are not part of the core conclusions.

Concept coverage

Concepts covered per source (of 87, at ≥1% prevalence): OBF 57, Sephora 59, Ecommerce 58, CDPH 3. The three formulation sources are balanced within a two-concept spread; CDPH's low coverage reflects its regulatory scope, by design.

Usage Notes

Appropriate uses: studies of cosmetic ingredient prevalence ranking; formulation-skeleton / co-occurrence networks; cross-geography / cross-channel formulation comparison; and exploratory temporal analyses (see the Exploratory temporal analyses section).

Interpreting the cross-niche correlations. The cross-niche Spearman coefficients (ρ = 0.56 OBF×Sephora, 0.65 OBF×e-commerce, common filter) are moderate, not high: ρ = 0.56 corresponds to roughly 31% shared rank variance. They should be read as evidence that **rank agreement survives a population shift**, that the ordering of "which ingredients are common" is consistent across two different collection mechanisms, and not as a claim of high predictive accuracy. Their scientific value lies in being consistently positive across two independent retail sampling modes and in the sharp contrast with the CDPH negative control (ρ ≈ 0): the pipeline is *discriminant*, producing positive agreement with formulation libraries and none with an unrelated registry. Users should treat the cross-niche coefficients

as lower-bound estimates of generalisation, since measurement noise in either source attenuates rank correlation.

Limitations (must be declared in any reuse):

1. **OBF is crowd-sourced and EU-skewed**: prevalence *levels* are not ground-truth market shares; treat them as *relative rankings*.

2. **Denominators differ by source**: each source declares ingredients under a different regime (Methods, Table 1), so absolute prevalence is comparable *within* a source over concepts, but only ranks are comparable *across* sources. Do not compute a difference of two prevalence levels from two different sources and interpret it as a market effect.

3. **Sephora and the e-commerce catalogue are two sampling modes of the same retail niche**: North-American premium-retail positioning explains some cross-source differences (e.g., SLES 14.9% in OBF vs 0.9% in Sephora reflects clean-positioning, not error), and the e-commerce catalogue overlaps Sephora by 2,731 products while adding mass-market Ulta (Table 1). The two are **not** independent niche samples: their agreement ($\rho = 0.81$) is intra-niche sampling robustness, and only the OBF×Sephora and OBF×e-commerce pairs carry the cross-niche generalisation claim.

4. **CDPH is a negative control only**: it must not be merged into functional-ingredient rankings.

5. **Exploratory temporal analyses are limited**: only OBF carries timestamps; cross-source time trends are unavailable, and OBF `created_t` is ingestion, not launch, year. The exploratory temporal p-values are reported in the Exploratory temporal analyses section without a primary-table multiplicity correction; apply a correction when testing further concepts.

6. **Ontology ceiling**: coverage is bounded by the 87-concept dictionary; unmapped ingredients are excluded (see coverage above). After the targeted supplementary sample, 86/87 concepts have at least one validated product-level instance; *butylresorcinol* remains unvalidated because it occurs in none of the sampled corpora.

7. **Residual matcher recall gap**: the validated matcher reaches F1 0.998, but non-English word-order variants (e.g., *acid hialuronic*) can still be missed, producing a small conservative under-count in multilingual sources.

8. **Two actives are trend-fragile**: niacinamide and phenoxyethanol show significant upward trends only because of the partial 2024–2026 ingestion years; in the complete 2016–2023 window they are non-significant ($p = 0.739$ and $p = 0.762$ respectively, see the Exploratory temporal analyses section). Treat them as *suggestive*, not established, temporal trends, and avoid over-interpreting their post-2023 trajectory. We note that the flip itself demonstrates the discriminative power of the pipeline: it detects both a significant full-period trend and its instability under a sub-period re-analysis, rather

than manufacturing a persistent signal. Because these analyses rest on a single timestamped source and on ingestion (not launch) dates, they are provided as exploratory material and are not part of the core cross-source agreement claims.

Code Availability

- **Runtime**: Node.js v22; `xlsx` for spreadsheet parsing.
- **Core scripts** (in the project `data/` directory):
- `multisource_analysis.js` — loads the sampled sources (two niches plus the regulatory control), computes prevalence, Spearman matrices, robustness, and negative control → `multisource_stats.json`
- `generate_multisource_excel.js` / `export_figures.js` — tabular and figure outputs
- `core_concepts.js` — the 87-concept ontology and matcher (`matchIngredientList` / `detectBot`)
- `validate_matcher.js` — seeded stratified sampling of the 50-product gold-standard set and matcher scoring → `matcher_validation_sample.json`
- `matcher_expand_sample.js` / `matcher_audit_methods.js` / `matcher_audit_record.js` — targeted enrichment sampling and the Method (1)+(2) audit pipeline
- `crossval_analysis.js` — two-source replication, brand-level replication, and the 2016–2026 temporal trend (OLS slope + 1,000-iteration permutation test, seed 20260807) → `crossval_stats.json`
- `deep4_analysis.js` — per-year, per-concept and per-group prevalence series underlying the trend → `deep4_stats.json`
- **Reproduce**: Two tiers. *Tier 1* (re-runnable from the deposited files alone, no external data): `node export_ontology_csv.js` → `node export_figures.js` → `node generate_multisource_excel.js` → `node matcher_audit_record.js` → `node verify_manuscript.js`. *Tier 2* (re-derive the primary statistics from the upstream raw corpora, which are large third-party downloads referenced above and are **not** redistributed): `node validate_matcher.js` → `node matcher_expand_sample.js` → `node matcher_audit_methods.js` → `node multisource_analysis.js` → `node crossval_analysis.js` → `node deep4_analysis.js`; the audit ballot preserves previously recorded human annotations across re-runs. All scripts are deterministic; the only stochastic components (permutation tests, gold-standard sampling) use fixed seeds. The temporal series in `deep4_stats.json` and the series embedded in `crossval_stats.json` are produced by independent code paths and agree year-by-year, providing an internal consistency check.
- **Repository**: the analysis scripts are deposited alongside the dataset in Zenodo (https://doi.org/10.5281/zenodo.21866120); no separate code repository is required.

Exploratory temporal analyses (supplementary)

Status: exploratory supplementary material. The analyses below are provided for completeness and transparency but are **not** part of the core cross-source agreement conclusions. They use Open Beauty Facts `created_t` (product ingestion year) as a proxy for market era; because `created_t` records database ingestion rather than product launch, any slope describes how the *catalogue* composition shifted over the period and is only an indirect proxy for market change. Readers should treat the following trends as exploratory.

Exploratory temporal analysis. Permutation testing of OBF year labels (1,000 iterations, fixed seed; products ingested in 2013–2015, n = 180, were excluded as a sparse early-ingestion tail) over 2016–2026 yields significant upward trends for hyaluronic acid (slope +0.0056 ≈ +0.56 percentage points per year, $p < 0.001$), niacinamide (+0.0030, $p < 0.001$), phenoxyethanol (+0.0059, $p < 0.001$), and "hero actives" (+0.0103, $p < 0.001$); significant downward trends for SLES (−0.0135, $p < 0.001$), SLS (−0.0021, $p < 0.001$), dimethicone (−0.0027, $p < 0.004$), and glycerin (−0.0041, $p < 0.001$); and a *non-significant* trend for botanical-extract content (−0.0008, $p = 0.55$). Slopes are expressed in proportion-of-products per year. Nine pre-specified concepts were tested, so a multiplicity correction applies to that primary family: under Bonferroni control at $\alpha = 0.05$ the threshold is $p < 0.0056$, and all eight items reported as significant clear it (the largest is dimethicone at $p = 0.004$, whose year-to-year series is volatile, so its downward trend is suggestive rather than smooth). Two additional grouped trends — total sulfates and mild-preservative content — were also computed and are reported here as exploratory, outside the primary family. The permutation floor of 0.001 reflects the 1,000-iteration resolution, not a smaller true p-value. That one of the nine concepts returns a flat, clearly non-significant slope shows the test is discriminating rather than uniformly significant.

Sub-period sensitivity (2016–2023, exploratory). The full-period window includes the partial ingestion years 2024–2026, so we repeated the same permutation design restricted to the complete 2016–2023 window as a sensitivity check. Six of the eight full-period significant concepts retain both direction and significance (SLES, SLS, dimethicone, hyaluronic acid, glycerin, hero actives; all $p < 0.05$), and botanical-extract content remains non-significant in both windows. However, two actives lose significance in the restricted window: niacinamide flattens to slope +0.0003 ($p = 0.739$) and phenoxyethanol to +0.0006 ($p = 0.762$). Their full-period upward slopes are therefore driven substantially by the 2024–2026 tail, which rests on fewer ingested products. Both actives are accordingly reported as *suggestive* rather than established trends, and caution is warranted against over-interpreting their post-2023 trajectory.

References

1. Open Beauty Facts. *Open Beauty Facts* — crowd-sourced cosmetic-ingredient database. https://world.openbeautyfacts.org (accessed 2026).

2. Sephora. *Sephora Products Dataset* — North-American premium-retail product catalogue with INCI ingredients (Kaggle). https://www.kaggle.com (accessed 2026).

3. Adieltheanalyst. *Ecommerce-cosmetic-dataset* — multi-platform cosmetic sales catalogue with INCI field (GitHub). https://github.com/Adieltheanalyst/Ecommerce-cosmetic-dataset (accessed 2026).

4. California Safe Cosmetics Program. *Chemicals in Cosmetics* — California Safe Cosmetics Program product–chemical registry (CDPH). https://www.cdph.ca.gov (accessed 2026).

5. He, X., Huang, J., Chen, Y., Li, B., Lü, G. & Chen, S. Situation analysis of traditional Chinese medicine hair care cosmetics based on social network analysis. *Chin. J. Mod. Appl. Pharm.* **37**, 1409–1416 (2020). https://doi.org/10.13748/j.cnki.issn1007-7693.2020.12.001

6. Li, L., Li, S., Wang, J.-S., Wu, D., Xu, G.-Q. & Wang, H.-Y. Molecular networking in cosmetic analysis: a review of non-targeted profiling for safety hazards and bioactive compounds. *Molecules* **30**, 3968 (2025). https://doi.org/10.3390/molecules30193968

7. Di Guardo, A., Trovato, F., Cantisani, C., Dattola, A., Nisticò, S. P., Pellacani, G. & Paganelli, A. Artificial intelligence in cosmetic formulation: predictive modeling for safety, tolerability, and regulatory perspectives. *Cosmetics* **12**, 157 (2025). https://doi.org/10.3390/cosmetics12040157

8. Zaki, M. J. Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.* **12**, 372–390 (2000). https://doi.org/10.1109/69.846291

9. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72–101 (1904).

10. Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L. Flavor network and the principles of food pairing. *Scientific Reports* **1**, 196 (2011). https://doi.org/10.1038/srep00196

All DOIs above were verified on 2026-08-07.

Prepared as a Nature Portfolio Data Descriptor submission draft. All figures derive from the cited public datasets and scripts. The Zenodo DOI and the code repository URL are inserted before submission.

Competing interests

The authors declare no competing interests.

Author Contributions

Wang Lejian (王乐健): Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing.

Acknowledgements

The author thanks the maintainers of the public data sources used in this work — Open Beauty Facts, the Sephora product catalogue, the e-commerce cosmetic dataset, and the California Safe Cosmetics Program — for making their data openly available.

Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics

This study analyses only publicly available, aggregate product-level datasets (Open Beauty Facts, Sephora, a multi-platform e-commerce catalogue, and the California Safe Cosmetics Program registry). It contains no personal or identifiable human-subject data and no animal data; therefore no ethical approval was required.