Original Research Article

Using Inferred Skin-Type Signals for Personalized Beauty Product Recommendation in a Hybrid System with Multi-Criteria Evaluation

Sherry Chen

Havergal College, 1451 Avenue Rd, Toronto, ON, M5N 2H9, Canada

ABSTRACT

Skin type match is a key consideration of fit in beauty product recommendation, yet most deployed recommendation systems do not have users' skin type information readily available, and standard precision-based metrics may not capture skin compatibility. This paper proposes a methodology to infer customers' skin type and integrates this signal into the hybrid beauty product recommender. Using the Sephora dataset (8,494 products, 294,722 reviews), purpose-built skin-type signals were built—including a skin-type-aware collaborative filtering matrix and a trained skin-type classifier—into various hybrid skincare recommendation configurations. Eight domain-specific metrics are proposed to evaluate the recommenders. A Bayesian weight optimization procedure using a Tree-structured Parzen Estimator (TPE) was applied to find optimal weights for the hybrid system. The results yield four key findings. First, personalization lifts are statistically significant across all four tested skin-type profiles on skin compatibility and routine coherence. Second, no single configuration dominates on all metrics: combining product content with skin profile achieves the highest skin-type compatibility and rank-sensitive precision among profile-aware variants; combining collaborative filtering with skin profile leads on diversity and serendipity; full hybrid provides balanced performance across all metrics. Third, profile weighting produces genuine inter-profile differentiation confirmed by positive adjusted personalization scores across all hybrid variants. Fourth, Bayesian optimization identifies the skin-type classifier as the dominant signal and reveals that enforcing a minimum content weight improves out-of-sample generalization. These results confirm that these inferred skin-type signals and the use of a multi-criteria evaluation framework can significantly improve the quality of the beauty product recommendation.

Keywords: beauty product recommendation; skin-type-aware personalization; skin-type classifier; hybrid recommender systems; multi-criteria evaluation metrics; collaborative filtering; Bayesian weight optimization; catalog coverage

INTRODUCTION

The global online beauty market has grown rapidly over the past decades, with platforms such as Sephora offering tens of thousands of beauty products across categories including skincare, makeup, hair, and fragrance. For consumers, this creates a choice-overload problem: it is challenging to navigate thousands of options to find products suited to their specific needs. For retailers, the challenge is to identify customer needs from browsing history and purchase patterns to deliver recommendations that increase returning customers and satisfaction. An effective recommendation system addresses both problems simultaneously.

Corresponding author: Sherry Chen, E-mail: sherrychensf@gmail.com.

Accepted May 27, 2026
https://doi.org/10.70251/HYJR2348.43261272

In beauty specifically, the most significant dimension of fit is skin type match. A moisturizer that works well for dry skin may overload oily skin; a serum formulated for sensitive skin may not work for normal skin. Recommendation quality in beauty products therefore depends on the system's ability to match products to individual skin profiles, not merely to general preferences or purchase history.

Building such a system faces three challenges. The first is signal inadequacy: general-purpose recommendation signals are not designed to encode skin-type compatibility. The second challenge is evaluation inadequacy: standard precision-based metrics such as Precision@K and NDCG (1) based on product category are not directly applicable, so beauty products need to be assessed through domain-specific metrics. The third challenge is how to assign appropriate weights when integrating the skin-type signal into a hybrid system to balance the sophisticated needs of beauty products customers.

LITERATURE REVIEW

Recommender systems are generally classified into three families. Collaborative filtering (CF) predicts user preferences from those of similar users, with item-based CF computing item-item similarity from co-rating patterns (2). A well-known limitation of CF is that it aggregates preferences across the full reviewer population, biasing similarity scores toward majority demographics and systematically under-representing minority subgroups. Content-based filtering (CBF) recommends items whose feature profiles resemble previously engaged items, typically using TF-IDF representations over item text (3). Hybrid systems combine CF and CBF through weighted scores, mitigating each method's individual weaknesses (4). Burke's foundational survey identifies seven hybrid strategies; this paper adopts weighted fusion, which is interpretable and tunable via the weight assignment procedure described in Methods.

A growing body of content-based research addresses skincare product personalization. Lee *et al*. proposed a content-based system filtering by ingredient compositions (5). Vinutha *et al*. proposed a similar approach analyzing chemical composition with individual skin types (6). Kyaw *et al*. demonstrated the value of ingredient profile vectorization in content-based skincare recommendation (7). Lee *et al*. extended this work by combining a deep neural network ingredient analyzer with a facial skin condition classifier (8). On the collaborative filtering side, Subhan *et al*. applied Neural Network Collaborative Filtering to cosmetics, demonstrating that neural architectures improve recommendation accuracy (9). Across all these systems, a common limitation is that none constructs a purpose-built skin-type signal or discusses how to integrate this signal into a hybrid framework for beauty products.

On the evaluation side, standard precision-based metrics have limited use in beauty recommendation. Beyond accuracy, three families of metrics have been discussed. Diversity metrics address filter-bubble effects by measuring intra-list item variety (10). Ziegler *et al*. demonstrated that diversity-aware re-ranking improves user satisfaction even when doing so reduces accuracy scores. Serendipity metrics capture the degree to which recommendations are simultaneously relevant and surprising (11). Catalog coverage metrics quantify how broadly a system exposes users to the available catalog (12). Celma and Herrera applied coverage and Gini-based distributional equity metrics to music recommendation and found that accuracy-optimized systems concentrate exposure on a small fraction of the catalog (12). These beyond-accuracy metrics have not been incorporated into a domain-specific framework for beauty products.

The weight assignment problem in hybrid recommenders has also attracted increasing attention. Bayesian optimization using Tree-structured Parzen Estimators (TPE) builds a probabilistic model of the objective surface and samples candidate weight combinations from regions of high expected improvement (13). TPE has been shown to find better weight combinations in fewer function evaluations than grid or random search for multi-signal linear systems. However, prior applications of TPE to recommender weight learning have used standard accuracy objectives rather than composite domain-specific objectives.

This paper makes four contributions to beauty product recommendations. First, it introduces a skin-type inference procedure that constructs a purpose-built inferred skin-type signal entirely from product metadata. Unlike existing content-based systems (5, 6, 7) that are directly based on content or ingredient composition, this paper's classifier produces a continuous skin-type compatibility score and is trained on brand-stated product metadata, avoiding circular measurement with the evaluation metrics. Second, it assembles the first domain-specific evaluation framework for beauty recommendation, proposing three new metrics—Skin Compatibility, MAP@5 with skin-type relevance,

and Routine Coherence—together with five beyond-accuracy metrics discussed in the past literature (Intra-List Diversity, Brand Diversity, Serendipity, Catalog Coverage, and Coverage Gini). Third, Bayesian weight optimization applies with a TPE sampler to a composite objective and systematically explores the sensitivity of optimized weights to boundary constraints. Fourth, it reports personalization lift tests against a non-personalized baseline for four distinct skin-type profiles, providing evidence of profile-aware recommendation benefit on a publicly available Sephora beauty dataset.

METHODS AND MATERIALS

Dataset

The dataset used in this study is the publicly available Sephora Products and Skincare Reviews dataset (14). The product information comprises 8,494 products across six primary categories—Skincare, Makeup, Hair, Fragrance, Bath & Body, and Tools & Accessories—and 294,722 customer reviews from 57,912 unique reviewers. Each review record contains a star rating (1–5), free-text review body, and four self-reported demographic attributes: skin_tone (18 values), skin_type (combination, dry, normal, oily, sensitive), hair_color, and eye_color. This analysis is restricted to the 2,283 products that have at least one review. Reviews with fewer than two of the four profile fields populated are excluded.

Recommendation System Components

The recommendation engine fuses six distinct signals, each capturing a different dimension of product–user compatibility.

Signal 1 — Content-Based TF-IDF Similarity (sim_content). Each product is represented as a single text document by combining its primary category, secondary category, product name, and parsed highlight tags, with 'Best for X Skin' labels stripped prior to vectorization. A TF-IDF matrix is computed using max_features = 5,000 and stop-word removal. Cosine similarity between the seed product's vector and all other products generates the signal of sim_content.

Signal 2 — Standard Item-Based Collaborative Filtering (sim_collab). A sparse user-item rating matrix is constructed from all reviews, with each entry being the reviewer's raw star rating (1–5). Item-item cosine similarity is computed on the transpose of this matrix.

Signals 3 to 5 depend on inferring the target skin type for each seed product, performed through the following steps.

Construction of Inferred Skin Type

Step 1: If the product's highlight tags include a brand-stated 'Best for X Skin' label, the corresponding canonical skin type is read directly.

Step 2: If the product name or highlights contain sensitivity-indicating keywords (e.g., 'sensitive', 'fragrance-free'), the type 'sensitive' is added to the set.

Step 3: If no brand tag or keyword is found, the skin-type classifier is called. This classifier is built as a binary Logistic Regression classifier trained separately for each of the five skin types using Sephora's brand-stated 'Best for X Skin' highlight tags as ground truth labels. The feature matrix combines TF-IDF representation (bigrams, max_features = 20,000, sublinear TF scaling) with binary expert keyword features encoding domain-critical ingredient and formulation terms. For each candidate product, the classifier returns a probability in [0, 1] for each of the five canonical skin types with the highest probability selected as inferred skin type. In case of a tie, the dominant skin type among the product's top-rated reviewers is used. Since some products have multi-type tags such as "Best for Dry, Combo, Normal Skin," this can cause products formulated for dry or combination skin to appear as positive training examples for the normal classifier. To correct this, the raw normal-skin probability is blended with a complement term, $0.5 \times \text{normal\_score} + 0.5 \times (1 - \max(\text{oily\_score}, \text{dry\_score}))$, where the second term penalizes any product that scores strongly on the most extreme opposing skin types. The result is that genuinely normal-compatible products, which should score low on both oily and dry classifiers, retain most of their score, while products that only appeared in normal training data due to multi-type tagging are down-weighted.

Step 4: If no result is obtained, the product is assigned the default type (combination).

Signal 3 — Skin-Type-Filtered Collaborative Filtering (sim_collab_st). This corrects the majority-bias in standard CF by restricting the rating matrix to reviewers whose documented skin type matches the inferred target skin type. A minimum of 50 reviews per type is required for statistical reliability.

Signal 4 — Profile-Matched Review Text Similarity (sim_review). Each review's contribution is weighted by its reviewer's fuzzy similarity to the target user profile before aggregation. Reviewer-to-profile similarity is computed as the mean of four fuzzy match scores (skin type, skin tone, hair color, and eye color) using the RapidFuzz library (15).

Signal 5 — Skin-Type Compatibility Classifier

Score (skin_score). This is a clean product-metadata-only personalization signal independent of all reviewer data. The same binary Logistic Regression classifier described in Step 3 is used: for a given user profile with inferred skin type t, the classifier for type t is queried for each candidate product, returning a probability in [0, 1] that reflects how confidently the product's metadata indicates suitability for skin type t. Training on brand-stated metadata rather than reviewer data avoids circular measurement between the personalization signal and the SkinCompat evaluation metric introduced later.

Signal 6 — Popularity Score (pop). A static quality signal computed as pop = 0.6 × (avg_rating / 5.0) + 0.4 × log(review_count + 1) / log(max_review_count + 1). The 0.6/0.4 split gives greater weight to rating quality than evidence volume; these values were manually chosen and could be tuned to reflect different deployment priorities.

Configuration Design

The final score for each candidate product can be a weighted linear combination: score = w_content · sim_content + w_collab_st · sim_collab_st + w_collab · sim_collab + w_review · sim_review + w_skin · skin_score + w_pop · pop, where all weights sum to 1.0. Two post-processing mechanisms — Maximal Marginal Relevance (MMR) re-ranking (16) and ε-greedy (Stochastic) are added to the full hybrid configuration to evaluate whether post-processing diversity mechanisms can recover variety that the scoring function alone does not produce. All twelve configurations are summarized in Table 1.

Evaluation Metrics

Eight domain-specific metrics are proposed. All are within [0, 1]. Metrics are computed over 549 deduplicated stratified seed queries (up to 150 per skin type).

Metric 1 — Skin Compatibility (SkinCompat). This is a newly proposed metric measuring the mean skin-type lift across all recommendations, where lift = proportion of a product's top-rated reviewers whose skin type matches the inferred type, divided by that skin type's base rate, rescaled as max (lift − 0.5, 0) × 2.

Metric 2 — Mean Average Precision with Skin-Type Relevance (MAP@5). This metric incorporates skin-type relevance into Average Precision at K = 5, with a product classified as relevant if its normalized lift for the user's skin type is ≥ 1.2× the dataset baseline.

Metric 3 — Routine Coherence (RoutineCoh). This is a newly proposed metric reflecting the fraction of unique secondary sub-categories among the ten recommended products. A skincare recommendation is most useful when it spans complementary product types, e.g., cleanser, serum, moisturizer, SPF, etc. rather than returning five variants of the same step.

Metric 4 — Intra-List Diversity. Mean pairwise TF-IDF cosine dissimilarity across all pairs of recommended products.

Metric 5 — Brand Diversity (BrandDiv). The fraction

Table 1. *Twelve experimental configurations with component signal weights and post-processing parameters (9 baseline configurations and three variants of the full hybrid configuration).*

Configuration	w_content	w_collab_st	w_collab	w_review	w_skin	w_pop	Skin-Filter	MMR λ	Stoch. ε
Random	---	---	---	---	---	---	No	---	---
Content Only	1.00	0.00	0.00	0.00	0.00	0.00	No	1.0	0.0
Content+Collab	0.65	0.00	0.20	0.00	0.00	0.15	No	1.0	0.0
Content+Reviews	0.65	0.00	0.00	0.20	0.00	0.15	No	1.0	0.0
Popularity	0.20	0.00	0.00	0.00	0.00	0.80	No	1.0	0.0
Cont+Skin Profile	0.35	0.00	0.00	0.00	0.65	0.00	Yes	1.0	0.0
Collab+Skin Profile	0.00	0.35	0.00	0.00	0.65	0.00	Yes	1.0	0.0
Rev+Skin Profile	0.00	0.00	0.00	0.35	0.65	0.00	Yes	1.0	0.0
Full Hybrid	0.20	0.25	0.15	0.00	0.40	0.00	Yes	1.0	0.0
Full Hybrid+MMR	0.20	0.25	0.15	0.00	0.40	0.00	Yes	0.6	0.0
Full Hyb Stochastic	0.20	0.25	0.15	0.00	0.40	0.00	Yes	1.0	0.30
Full Hyb+MMR+Stoch.	0.20	0.25	0.15	0.00	0.40	0.00	Yes	0.6	0.30

of unique brand names among the ten recommended products.

Metric 6 — Serendipity. Mean (1 − popularity_percentile) × (1 − sim_to_seed) × (avg_rating / 5.0) per recommendation (11).

Metric 7 — Catalog Coverage. The fraction of the 2,283-product catalog surfaced at least once across all 549 seed queries (12).

Metric 8 — Coverage Gini (CovGini). Measures of how evenly recommendations are distributed across the catalog. 0 = perfectly uniform; 1 = one product monopolizes all recommendations.

Supplementary Tests

Three supplementary tests were designed to answer questions about architectural differences, genuine personalization, and statistical significance. Test 1 (Ranking Disagreement) computes Kendall's τ and Rank-Biased Overlap (RBO) (17) between configuration pairs across 549 seeds, with a two-sided significance test (H_0: $\tau = 0$). This test is to check if these configurations actually generate different ranked lists. Test 2 (Cross-Profile Recommendation Overlap) fixes the configuration (Full Hybrid) and varies the user profile across four simulated users—Dry+Fair, Oily+Light, Combination+Deep, and Normal+Tan to see if the profile signal genuinely move recommendation lists apart. This is done by running each profile against the same 549 seeds and computing an adjusted personalization score that isolates the contribution of explicit profile weighting from seed-driven diversity. Test 3 (Personalization Lift) asks if adding profile information to the full hybrid improves recommendations over a system that uses no profile at all. It applies a one-sided Wilcoxon signed-rank test (H_1: Full Hybrid > Content Only) on paired per-seed scores across three metrics and four profiles to test the significance of personalization lift. The Wilcoxon test is chosen because per-seed metric distributions are bounded and non-Gaussian.

Modules

This analysis was conducted using Python 3.10. Key libraries included NumPy, Pandas, Matplotlib, and scikit-learn (18) for data processing and machine learning. The TF-IDF vectorizer and Logistic Regression classifier were sourced from scikit-learn. Collaborative filtering matrices were built using scipy sparse matrices. Fuzzy string matching used the RapidFuzz library. The Wilcoxon signed-rank test and Kendall's τ were computed using scipy.stats. Bayesian weight optimization used the Optuna library with the TPE sampler. All simulations use random_seed = 42 for reproducibility. The code used to generate the analyses and results in this study is publicly available at https://github.com/sherrychen29/Recommendation-System-for-Beauty-Products.git

RESULTS

Main Evaluation

Table 2 presents evaluation results for all twelve configurations using eight metrics over 549 stratified seed queries. The upper section covers the nine baseline configurations; the lower section shows the three Full Hybrid post-processing variants.

Random sampling defines practical upper bounds for variety-oriented metrics. Popularity achieves the worst CovGini (0.825) and lowest Serendipity (0.013), and its SkinCompat of 0.079 falls below Random (0.153), confirming that bestsellers earn broad appeal by being more generic rather than skin-type-specific.

Content Only achieves SkinCompat = 0.181 and MAP@5 = 0.195, confirming that TF-IDF product similarity provides a weak but genuine skin-type signal. Content Only also achieves the highest Coverage (0.733) and lowest CovGini (0.525) among non-random configurations. Adding collaborative filtering (Content+Collab) yields a drop in SkinCompat (0.181 → 0.164) and MAP@5 (0.195 → 0.181). Content+Reviews achieves the highest BrandDiv (0.703) among content-based configurations but the lowest RoutineCoh (0.262) of any configuration.

Skin-profile-aware configurations generally improve skin compatibility. Content+Skin Profile achieves the highest SkinCompat and MAP@5 among personalized configurations, with lifts of +0.045 and +0.076 over Content Only. Collab+Skin Profile leads to routine coherence, diversity, and serendipity but its Coverage of 0.123 falls below even the Popularity baseline (0.181). Diversity and Serendipity are intra-list metrics measured per seed, while Coverage is an aggregate metric across all seeds, so the two sets of metrics offer different insights. Full Hybrid achieves near-top SkinCompat and MAP@5 with moderate scores across all remaining metrics, reflecting a balanced signal mix.

Within the post-processing variants, Full Hybrid + MMR improves within-list variety and personalization at the cost of coverage; stochastic selection broadens catalog exposure at minimal personalization cost; and FH Stochastic and FH+MMR+Stochastic are operationally equivalent, as explained in Table 3b.

Table 2. *Mean evaluation scores across all configurations (n = 549 seeds). Bold = best value (non-random) per metric in each section. For all metrics, higher values reflect better performance; CovGini is an exception, where lower values indicate more equitable catalog distribution.*

Configuration	SkinCompat	MAP@5	RoutCoh	Diversity	BrandDiv	Serendip	Coverage	CovGini
Random	0.153	0.137	0.601	0.922	0.940	0.446	0.907	0.346
Content Only	0.181	0.195	0.296	0.624	0.688	0.251	0.733	0.525
Content+Collab	0.164	0.181	0.295	0.626	0.681	0.209	0.727	0.527
Content+Reviews	0.160	0.178	0.262	0.631	0.703	0.185	0.694	0.536
Popularity	0.079	0.059	0.439	0.787	0.763	0.013	0.181	0.825
Cont+Skin Prof	0.226	0.271	0.354	0.689	0.742	0.280	0.493	0.688
Collab+Skin Prof	0.201	0.220	0.528	0.807	0.740	0.311	0.123	0.825
Rev+Skin Prof	0.196	0.250	0.298	0.748	0.777	0.215	0.374	0.706
Full Hybrid	0.223	0.264	0.368	0.704	0.733	0.274	0.517	0.687
Full Hybrid Post-Processing Variants								
Full Hybrid+MMR	0.232	0.270	0.447	0.785	0.848	0.294	0.483	0.692
Full Hyb Stochastic	0.227	0.256	0.409	0.744	0.778	0.296	0.546	0.641
FH+MMR+Stochastic	0.227	0.256	0.409	0.744	0.778	0.296	0.546	0.641

Note: Random and Popularity serve as boundary setters rather than deployed systems.

Ranking Disagreement

Table 3a covers the results from cross-configuration pairs and Table 3b covers the results from Full Hybrid family pairs.

In Tier 1, Content+Skin Profile vs Full Hybrid (τ = 0.808, overlap = 87.0%) is highest: the skin classifier acts as a shared signal determining roughly seven-eighths of product selection. Content Only vs Content+Collab (τ = 0.753) shows CF adds only modest reshuffling because it amplifies the same popularity gradient already in TF-IDF. In Tier 2, all four pairs involve a skin-type signal on at least one side. In Tier 3, eight pairs show negligible or statistically insignificant rank correlation.

All pairs sit in Tier 1 ($\tau \geq 0.50$), confirming that every variant preserves the base scoring function's dominant ordering. When ε-greedy is active, it fills each slot independently by drawing from remaining candidates, so MMR has nothing to act on and the two configurations produce identical outputs. The more architecturally distinct pair is Full Hybrid vs Full Hybrid+MMR (τ = 0.649), and its closest neighbour FH+MMR vs FH Stochastic (τ = 0.627), because MMR substitutes items and reshuffles the retained ones, which creates genuine list-level differences even though the underlying scores are identical.

Cross-Profile Overlap

Table 4a reports pairwise profile overlaps. All four hybrid variants show significant inter-profile differentiation ($p < .001$). Dry+Fair vs Oily+Light is the most differentiated pair (overlap 0.016–0.017), reflecting the large compositional distance between dry-skin and oily-skin formulations.

Personalization scores are also constructed to capture the differentiation in the recommended list, which are defined as 1-mean (profile pairwise overlap). These are subtracted by the baseline scores (1-mean of profile overlap with no-profile) to arrive at the adjusted personalization score.

The personalization scores are positive for all four hybrid configurations which shows the system is differentiating the recommendations based on profiles.

Personalization Lift Test

Table 5 reports Full Hybrid versus Content Only personalization lifts for four simulated user profiles over 549 seeds with Wilcoxon signed-rank tests p-value and significance reported.

Dry+Fair and Oily+Light show statistically significant lifts in all three metrics across all variants. Combination+Deep sees significant lifts in SkinCompat

Table 3a. *Pairwise ranking disagreement across configuration families (n = 549 seeds). Kendall's τ measures the rank correlation on shared items with higher score indicating more similar ordering. Rank-Biased Overlap (RBO) measures top-position-weighted similarity not requiring shared items, with higher scores meaning more similar ordering. Overlap is the mean fraction of shared items with higher scores indicating more similarity. Pairs are grouped by τ threshold: Tier 1 = τ ≥ 0.50; Tier 2 = 0.10 ≤ τ < 0.50 and p < .05; Tier 3 = τ < 0.10 or p ≥ .05. Significance is calculated for the two-sided test of H0: τ=0 with * indicating p < .05, ** indicating p < .01, and *** indicating p < .001 and n.s. standing for "not significant," meaning the p-value exceeded the 0.05 threshold.*

Configuration Pair	Mean τ	Mean RBO	Overlap	p-value	sig
Tier 1 — Strong agreement (τ ≥ 0.50)					
Cont+Skin Profile vs Full Hybrid	0.808	0.530	87.0%	< .001	***
Content Only vs Content+Collab	0.753	0.562	86.2%	< .001	***
Content+Collab vs Content+Reviews	0.643	0.503	76.2%	< .001	***
Content Only vs Content+Reviews	0.595	0.473	70.3%	< .001	***
Tier 2 — Moderate agreement (0.10 ≤ τ < 0.50, p < .05)					
Collab+Skin Profile vs Full Hybrid	0.223	0.158	30.9%	< .001	***
Content Only vs Full Hybrid	0.204	0.140	19.2%	< .001	***
Content Only vs Cont+Skin Profile	0.167	0.134	19.8%	< .001	***
Cont+Skin Profile vs Collab+Skin Profile	0.159	0.123	26.9%	< .001	***
Tier 3 — Weak / independent (τ < 0.10 or p ≥ .05)					
Popularity vs Full Hybrid	0.059	0.060	7.3%	< .001	***
Content Only vs Rev+Skin Profile	0.017	0.031	5.3%	0.106	n.s.
Content Only vs Collab+Skin Profile	0.009	0.028	2.0%	0.110	n.s.
Rev+Skin Profile vs Full Hybrid	0.004	0.135	30.0%	0.446	n.s.
Collab+Skin Profile vs Rev+Skin Profile	−0.070	0.148	24.7%	0.989	n.s.
Popularity vs Collab+Skin Profile	0.006	0.020	1.3%	0.183	n.s.
Random vs Content Only	0.000	0.001	0.6%	---	---
Random vs Full Hybrid	0.000	0.001	0.3%	---	---

Table 3b. *Pairwise ranking disagreement within the Full Hybrid family (n = 549 seeds). Column definitions are identical to Table 3a. Results are sorted in descending order by τ. FH Stochastic and FH+MMR+Stochastic are operationally identical (τ = 1.000).*

Full Hybrid Variant Pair	Mean τ	Mean RBO	Overlap	p-value	sig
FH Stochastic vs FH+MMR+Stochastic	1.000	0.651	99.9%	---	---
Full Hybrid vs FH Stochastic	0.954	0.455	71.7%	< .001	***
Full Hybrid vs FH+MMR+Stochastic	0.954	0.455	71.7%	< .001	***
Full Hybrid vs Full Hybrid+MMR	0.649	0.494	69.3%	< .001	***
FH+MMR vs FH Stochastic	0.627	0.378	55.9%	< .001	***
FH+MMR vs FH+MMR+Stochastic	0.627	0.378	55.9%	< .001	***

Table 4a. *Mean pairwise recommendation overlap (proportion of shared products) across four hybrid variants (n = 549 seeds, top-10). Lower = more differentiated. ***p < .001 (one-sample t-test, H_0: overlap = 1.0).*

Profile Pair	Full Hybrid	FH+MMR	FH Stochastic	FH+MMR+Stoch.	sig
Dry+Fair vs Oily+Light	0.017	0.017	0.016	0.016	***
Dry+Fair vs Combination+Deep	0.091	0.060	0.074	0.074	***
Dry+Fair vs Normal+Tan	0.052	0.045	0.048	0.048	***
Dry+Fair vs No Profile	0.169	0.137	0.136	0.136	***
Oily+Light vs Combination+Deep	0.241	0.183	0.187	0.187	***
Oily+Light vs Normal+Tan	0.040	0.034	0.038	0.038	***
Oily+Light vs No Profile	0.148	0.116	0.119	0.119	***
Combo+Deep vs Normal+Tan	0.142	0.130	0.125	0.125	***
Combo+Deep vs No Profile	0.194	0.150	0.162	0.162	***
Normal+Tan vs No Profile	0.266	0.211	0.216	0.216	***

Table 4b. *Personalization score summary for all four hybrid variants. PersonalScore measures how different the two profiles' recommendation lists are from each other; Baseline measures how different each profile's list is from what you would get with no profile at all. Adjusted = PersonalScore − Baseline. Positive adjusted scores confirm that profile weighting genuinely moves recommendation lists apart beyond seed-product differences alone as measured in baseline score.*

Configuration	PersonalScore	Baseline	Adjusted
Full Hybrid	0.903	0.806	+0.097
Full Hybrid + MMR	0.922	0.846	+0.076
Full Hybrid Stochastic	0.919	0.842	+0.077
FH+MMR+Stochastic	0.919	0.842	+0.077

Table 5. *Personalization lift of all hybrid variants vs Content Only per profile (n = 549 seeds). Δ = mean (hybrid variant) − mean (Content Only). Wilcoxon signed-rank, one-sided tests (H_1: hybrid > Content Only). * Indicates p < .05, ** indicates p < .01, and *** indicates p < .001, where p is the probability of observing a lift at least as large as the one measured if the null hypothesis were true. A result marked *** means there is less than a 0.1% chance the observed lift arose by chance, making it the strongest evidence of a genuine personalization effect. n.s. = not significant.*

Profile × Variant	ΔSkinCompat	p	sig	ΔMAP@5	p	sig	ΔRoutCoh	p	sig
Dry + Fair									
Full Hybrid	+0.086	< .001	***	+0.210	< .001	***	+0.086	< .001	***
Full Hybrid+MMR	+0.100	< .001	***	+0.232	< .001	***	+0.164	< .001	***
Full Hyb Stochastic	+0.086	< .001	***	+0.187	< .001	***	+0.131	< .001	***
Oily + Light									
Full Hybrid	+0.206	< .001	***	+0.266	< .001	***	+0.087	< .001	***
Full Hybrid+MMR	+0.214	< .001	***	+0.271	< .001	***	+0.122	< .001	***
Full Hyb Stochastic	+0.210	< .001	***	+0.269	< .001	***	+0.109	< .001	***

Continued Table 5. *Personalization lift of all hybrid variants vs Content Only per profile (n = 549 seeds). Δ = mean (hybrid variant) − mean (Content Only). Wilcoxon signed-rank, one-sided tests (H_1: hybrid > Content Only). * Indicates p < .05, ** indicates p < .01, and *** indicates p < .001, where p is the probability of observing a lift at least as large as the one measured if the null hypothesis were true. A result marked *** means there is less than a 0.1% chance the observed lift arose by chance, making it the strongest evidence of a genuine personalization effect. n.s. = not significant.*

Profile × Variant	ΔSkinCompat	p	sig	ΔMAP@5	p	sig	ΔRoutCoh	p	sig
Combination + Deep									
Full Hybrid	+0.012	< .001	***	+0.003	0.226	n.s.	+0.068	< .001	***
Full Hybrid+MMR	+0.021	< .001	***	+0.007	0.008	**	+0.159	< .001	***
Full Hyb Stochastic	+0.016	< .001	***	+0.002	0.166	n.s.	+0.123	< .001	***
Normal + Tan									
Full Hybrid	+0.012	0.008	**	+0.001	0.489	n.s.	+0.087	< .001	***
Full Hybrid+MMR	+0.008	0.059	n.s.	−0.014	0.934	n.s.	+0.166	< .001	***
Full Hyb Stochastic	+0.015	< .001	***	+0.003	0.354	n.s.	+0.121	< .001	***

and RoutineCoh. Normal+Tan shows significant positive lifts in SkinCompat and RoutineCoh, but MAP@5 lifts are not statistically significant, as normal skin is the most prevalent demographic and the population-level signal already implicitly favors normal-compatible products.

Bayesian Weight Optimization

This paper also explores how to use Bayesian weight optimization to identify the best weight mix in a hybrid system. The composite objective is defined as 0.40 × SkinCompat + 0.30 × MAP@5 + 0.20 × RoutineCoh + 0.10 × Serendipity where the objective weights reflect the relative importance of each metric. The initial search bounds of weights are defined in the baseline scenario in Table 6. A TPE sampler ran 60 trials on a 70% held-out validation split to find the optimal weight mix. The TPE sampler converged to the following optimal weights: w_content = 0.055, w_collab = 0.228, w_skinrating = 0.611, w_collab_st = 0.106. Compared to hand-designed Full Hybrid, the optimizer increased the skin-type classifier weight (0.40 → 0.611) and substantially reduced w_collab_st (0.25 → 0.106).

To test the sensitivity to the imposed bounds, four boundary scenarios were created in addition to the baseline scenario. Table 6 defines the scenarios, reports their composite scores, and lists optimal weights.

Three findings stand out. First, a content floor prevents overfitting: Content Floor (w_content ≥ 0.15) achieves the best test score (0.3188) despite the worst validation score (0.2608). Second, the skin signal has a ceiling at approximately 0.61: pushing to 0.751 (Skin Dominant) reduces test performance. Third, the baseline bounds are not constraining: the Wide-Open scenario found comparable performance to Baseline, confirming the original search space was well-calibrated.

Table 6. *Bayesian optimization boundary exploration: search bounds, composite scores, and optimal weights across five scenarios. Val Composite is the composite score achieved on the 70% validation split; Test Composite is the score computed on the remaining 30% test set. Val->Test Lift measures the change in score when moving from validation set to test set which measures generalization behavior.*

Scenario	Trials	w_content bounds	w_skinrating bounds	Hypothesis
Baseline	60	[0.05, 0.60]	[0.10, 0.65]	Existing bounds
Skin Dominant	60	[0.01, 0.40]	[0.20, 0.85]	Test if skin dominates more
Wide Open	80	[0.01, 0.90]	[0.01, 0.90]	Remove all boundary bias
Content Floor	60	[0.15, 0.50]	[0.10, 0.70]	Enforce minimum 15% content weight
Collab Boost	60	[0.05, 0.40]	[0.10, 0.65]	Test if collab signals are under-utilized

Continued Table 6. *Bayesian optimization boundary exploration: search bounds, composite scores, and optimal weights across five scenarios. Val Composite is the composite score achieved on the 70% validation split; Test Composite is the score computed on the remaining 30% test set. Val->Test Lift measures the change in score when moving from validation set to test set which measures generalization behavior.*

Scenario	Trials	w_content bounds	w_skinrating bounds	Hypothesis
Composite Scores				
Scenario	**Val Composite**	**Test Composite**	**Val→Test Lift**	
Baseline	0.2696	0.3113	+0.0417	
Skin Dominant	0.2729	0.3075	+0.0346	
Wide Open	0.2727	0.3081	+0.0354	
Content Floor	0.2608	0.3188 (best)	+0.0580	
Collab Boost	0.2682	0.3114	+0.0432	
Optimal Weights				
Scenario	**w_content**	**w_collab**	**w_skinrating**	**w_collab_st**
Baseline	0.055	0.228	0.611	0.106
Skin Dominant	0.047	0.044	0.751	0.158
Wide Open	0.033	0.159	0.677	0.131
Content Floor	0.173	0.078	0.611	0.138
Collab Boost	0.043	0.342	0.517	0.099

DISCUSSION

Signal Importance and Configuration Comparison

The eight-metric evaluation reveals distinct and complementary strengths across configuration families. The prior content-based systems (5, 6, 7) are most directly comparable to the Content Only configuration here, which achieves SkinCompat = 0.181 and MAP@5 = 0.195 using TF-IDF similarity alone. The gain from adding the purpose-built skin-type classifier is immediate: Content+Skin Profile achieves SkinCompat = 0.226 and MAP@5 = 0.271, a lift of +0.045 and +0.076 respectively, demonstrating that the trained classifier contributes signal that ingredient-list similarity alone cannot replicate.

On the other hand, adding collaborative filtering to content yields a drop in SkinCompat and MAP@5, likely because the full-pool rating matrix is implicitly popularity-biased due to the minimum number of review requirement, diluting the incidental skin-type clustering in TF-IDF space.

Adding reviews to content achieves the highest BrandDiv (0.703) among content-based configurations but the lowest RoutineCoh (0.262) of any configuration, revealing that review-text similarity clusters products by the vocabulary reviewers use so the signal broadens brand exposure at the cost of concentrating recommendations within a single product type.

Furthermore, adding skin profile into collaborative filtering (Collab+Skin Profile) leads to higher routine coherence, diversity, and serendipity but its Coverage of 0.123 falls below the Popularity baseline which indicates a severe filter bubble: restricting the rating matrix to skin-type-specific reviewers concentrates similarity scores on a small highly-reviewed subset.

Overall, a structural tension between personalization depth and catalog coverage runs through every comparison which calls for the needs of a hybrid model to achieve a balanced recommendation reflecting the customized objective of the system. The two post-processing mechanisms can be used to address different deployment needs: MMR for within-list routine-building variety, stochastic perturbation for cross-session catalog exposure.

Personalization in the Recommendations

The cross-profile overlap results confirm genuine inter-profile personalization. All four hybrid variants

show positive adjusted personalization scores, meaning profile weighting moves recommendation lists apart beyond what seed-product differences alone produce. The Wilcoxon personalization lift tests show some evidence that profile-aware recommendation benefit is both statistically significant and practically large for Oily+Light and Dry+Fair profile. Normal+Tan profile shows significant positive lifts in SkinCompat and RoutineCoh, but MAP@5 lifts are not statistically significant.

Hybrid Weight Optimization

The Bayesian optimization result provides a data-driven calibration of the signal space. The TPE optimizer assigns w_skinrating = 0.611 and reduces w_content to 0.055, confirming that TF-IDF content similarity contribution is negligible once a strong skin-type classifier is present. The boundary exploration analysis further shows that enforcing a content floor (w_content ≥ 0.15) improves test generalization, suggesting that practitioners should retain a minimum content similarity weight to prevent classifier overfitting.

Limitation

The criteria used to evaluate the recommenders in this paper are based on the eight theoretical metrics instead of real customer feedback through questionnaires or online experience through A/B tests; therefore, it may not be able to capture actual user satisfaction or purchase behavior. In addition, the test is solely conducted on the Sephora dataset, so conclusion generalizations to another dataset or catalogue is yet to be tested. Lastly, the ground truth of skin type used for classifier training in this research is based on the "Best for X Skin" labels provided by the brand, so the accuracy and reliability of this manufacturer label are critical for the success of the recommender.

CONCLUSION

This paper constructed inferred skin-type signals and built them into the design of various beauty product recommendation configurations using the Sephora dataset. Eight evaluation metrics specific to beauty products were proposed and supplemented with statistical tests. Bayesian weight optimization was utilized to explore optimal weights for the hybrid system. Four findings stand out. First, no single profile-aware configuration dominates on all personalization metrics: among the baseline configurations, Content+Skin Profile achieves the highest SkinCompat and MAP@5 and Collab+Skin Profile leads Diversity and Serendipity; Hybrid offers a way to assign priorities to different metrics with some post-processing techniques potentially worth exploring: adding re-ranking mechanism MMR can improve within-list routine-building variety, while adding stochastic perturbation on top could see improvement of the serendipity and coverage. Second, all four hybrid variants produce positive adjusted personalization scores, confirming genuine inter-profile differentiation beyond seed-driven diversity. Third, Bayesian optimization confirms the skin-type classifier as the dominant signal (w_skinrating = 0.611), and shows that a content floor (w_content ≥ 0.15) improves test generalization. Fourth, personalization lifts are statistically significant across all four profiles on SkinCompat and RoutineCoh, with Oily+Light and Dry+Fair additionally showing significant MAP@5 gains.

These results confirm that the inferred skin type is an important signal to consider in the beauty product recommendation. It can be integrated into other filtering systems and further be used in the hybrid system. Evaluating beauty product recommendations requires a multi-criteria framework to cover different dimensions of customers' needs. The methodological contributions—construction of skin-type signals, the proposed eight domain-specific metrics, and Bayesian weight learning for a composite-objective—provide a replicable framework applicable to similar beauty product datasets.

Future work could pursue two directions. First, the Sephora dataset is skewed toward combination and normal skin types, with very few products carrying explicit sensitive-skin labels; future research could explore analysis on a dataset with more balanced skin-type representation. Second, A/B testing the full hybrid configuration against a content-only baseline in a live retail setting would establish whether the statistically significant personalization lifts measured here translate into measurable improvements in purchase conversion or satisfaction ratings.

ACKNOWLEDGEMENTS

The Sephora Products and Skincare Reviews dataset was sourced from Kaggle under a public domain license (14). The author acknowledges the open-source communities behind scikit-learn, pandas, scipy, RapidFuzz, and joblib, whose tools formed the computational foundation of this research.

FUNDING SOURCES

No external funding was received for this research.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. Shani G, Gunawardana A. Evaluating recommendation systems. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender Systems Handbook. *Boston: Springer*. 2011; p.257–97. doi:10.1007/978-0-387-85820-3_8
2. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. *Computer*. 2009; 42 (8): 30–7. doi:10.1109/MC.2009.263
3. Lops P, de Gemmis M, Semeraro G. Content-based recommender systems: State of the art and trends. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender Systems Handbook. *Boston: Springer*. 2011; p.73–105. doi:10.1007/978-0-387-85820-3_3
4. Burke R. Hybrid recommender systems: Survey and experiments. *User Model User-Adapt Interact*. 2002; 12 (4): 331–70. doi:10.1023/A:1021240730564
5. Lee G, Jiang X, Parde N. A content-based skincare product recommendation system. In: Proceedings of the 22nd IEEE International Conference on Machine Learning and Applications (ICMLA). *IEEE*. 2023; p.2039–43. doi:10.1109/ICMLA58977.2023.00308
6. Vinutha M, Dayananda RB, Kamath A. Personalized skincare product recommendation system using content-based machine learning. In: Proceedings of the 4th International Conference on Intelligent Technologies (CONIT). *IEEE*. 2024; p.1–6. doi:10.1109/CONIT61985.2024.10626458
7. Kyaw TZL, Uttama S, Panwong P. Leveraging ingredient profiles in content-based skincare product recommendation. In: Proceedings of the 8th International Conference on Information Technology (InCIT). *IEEE*. 2024; p.319–24. doi:10.1109/incit63192.2024.10810620
8. Lee J, Yoon H, Kim S, Lee C, Lee J, Yoo S. Deep learning-based skin care product recommendation: a focus on cosmetic ingredient analysis and facial skin conditions. *J Cosmet Dermatol*. 2024; 23 (6): 2066–77. doi:10.1111/jocd.16218
9. Subhan S, Syarif DL, Widhihastuti E, Rakainsa SK, Sam'an M, Ifriza YN. Improved recommender system using neural network collaborative filtering (NNCF) for e-commerce cosmetic product. *SINERGI*. 2025; 29 (1): 155–62. doi:10.22441/SINERGI.2025.1.014
10. Ziegler CN, McNee SM, Konstan JA, Lausen G. Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web. New York: ACM. 2005; p.22–32. doi:10.1145/1060745.1060754
11. Adamopoulos P, Tuzhilin A. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Trans Intell Syst Technol*. 2014; 5 (4): 54. doi:10.1145/2559952
12. Celma Ò, Herrera P. A new approach to evaluating novel recommendations. In: Proceedings of the 2008 ACM Conference on Recommender Systems. New York: ACM. 2008; p.179–86. doi:10.1145/1454008.1454038
13. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. *Adv Neural Inf Process Syst*. 2011; 24: 2546–54.
14. Nour MA. Sephora products and skincare reviews [Dataset]. Kaggle. 2023. Available from: https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews (accessed on 2025-03-19)
15. Bachmann M. RapidFuzz Version 3.13.0. GitHub; 2025. Available from: https://github.com/rapidfuzz/RapidFuzz doi: 10.5281/zenodo.15133267
16. Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. *New York: ACM*. 1998; p.335–6. doi:10.1145/290941.291025
17. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. *ACM Trans Inf Syst*. 2010; 28 (4): 1–38. doi:10.1145/1852102.1852106
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011; 12: 2825–30.