**Cannabis Strain Characteristics and Consumer Preference: A Multivariate Analysis of Effects, Flavors, and Ratings**

Nesar K S

[ksnesar@outlook.com](mailto:ksnesar@outlook.com)

***Abstract:***

*Abstract—This study presents a comprehensive multivariate analysis of cannabis strain characteristics and their relationship to consumer preferences. Using a dataset of 2,351 cannabis strains, we investigated the distribution of strain types (indica, sativa, hybrid), effects, flavors, and their correlations with consumer ratings. Statistical analysis revealed significant associations between specific effects (Happy, Relaxed, Euphoric) and flavors (Sweet, Earthy, Berry) with higher consumer satisfaction. While strain type did not significantly predict ratings (ANOVA, F = 1.02, p = 0.36), the number of reported effects and flavors strongly correlated with higher ratings (r = 0.61 and r = 0.43, respectively, p < 0.001). Principal Component Analysis and K-means clustering identified distinct strain profiles representing different combinations of effects and flavors. These findings challenge traditional classification paradigms and suggest that effect and flavor profiles are more predictive of consumer satisfaction than conventional strain categorizations. The results provide an evidence-based framework for consumers, producers, and researchers seeking to understand the complex relationships between cannabis characteristics and user experience.*

**I. INTRODUCTION**

**With increasing legalization and acceptance worldwide, cannabis has emerged as both a recreational substance and therapeutic agent across many regions. Understanding consumer preferences for different cannabis strains has become increasingly important for cultivators, retailers, medical practitioners, and consumers themselves [1]. Cannabis strains vary widely in their chemical composition, particularly in their cannabinoid and terpene profiles, which contribute to diverse effects and flavors experienced by users [2].**

**Despite growing interest in cannabis research, there remains limited understanding of how specific strain characteristics correlate with consumer preferences. Traditional classification of cannabis into indica, sativa, and hybrid categories has guided consumer choices for decades, yet emerging research suggests these distinctions may be less meaningful than specific chemical profiles [3]. This research gap presents an opportunity for data-driven analysis to identify patterns and relationships that may better inform both consumers seeking specific experiences and producers developing new strains.**

**This study aimed to analyze a comprehensive dataset of cannabis strains to identify patterns and relationships between strain types, reported effects, flavors, and consumer ratings. The specific research questions guiding this analysis were:**

**1. How are cannabis strains distributed across different types (indica, sativa, hybrid)?**

**2. What are the most common effects and flavors associated with cannabis strains?**

**3. Do strain types differ significantly in their consumer ratings?**

**4. Which effects and flavors most strongly correlate with higher consumer ratings?**

**5. Can cannabis strains be meaningfully clustered based on their effects and flavor profiles?**

**By applying various statistical and machine learning techniques to address these questions, this study contributes to the emerging field of cannabis informatics and provides insight into the factors driving consumer preferences in an increasingly important market.**

**II. RELATED WORK**

Research on cannabis strain characteristics and consumer preferences has expanded considerably in recent years. Several studies have examined the chemical composition of cannabis strains and their relationship to reported effects. Russo [4] argued that the traditional indica-sativa distinction is less meaningful than specific cannabinoid and terpene profiles in determining effects. Similarly, Piomelli and Russo [5] highlighted the limitations of strain-based classifications in predicting user experiences.

The concept of chemovars (chemical varieties) has gained traction as an alternative classification system. Lewis et al. [6] proposed that chemical composition provides a more reliable basis for predicting effects than traditional strain classifications. Their work emphasized the importance of specific cannabinoid ratios and terpene profiles in mediating psychoactive and therapeutic effects.

Baron [7] reviewed the medicinal properties of cannabinoids, terpenes, and flavonoids in cannabis, noting associations between specific compounds and reported effects. However, few studies have directly analyzed large datasets of consumer-reported effects and preferences to identify patterns and relationships.

The current study builds upon this foundation by applying data science techniques to a substantial dataset of consumer-reported experiences, addressing the gap between chemical composition research and real-world consumer preferences.

**III. METHODOLOGY**

**A. Dataset Description**

The analysis was conducted on a dataset containing information about 2,351 cannabis strains. Each entry included:

- Strain name

- Type (indica, sativa, or hybrid)

- Consumer rating (scale: 0-5)

- Effects (comma-separated list)

- Flavors (comma-separated list)

- Description

The data was relatively complete, with only 46 missing flavor entries (1.96%) and 33 missing descriptions (1.40%). There was one duplicate strain entry ("B-Witched") which was retained for the analysis as it may represent different phenotypes of the same strain.

**B. Analytical Approach**

The exploratory data analysis followed a structured approach:

1) Dataset Overview: Examining basic statistics, data types, missing values, and distributions of key variables.

2) Distribution Analysis: Analyzing the distribution of strain types, ratings, effects, and flavors using histograms, boxplots, violin plots, and word clouds.

3) Relationship Analysis: Investigating correlations between effects/flavors and ratings, as well as creating network graphs to visualize co-occurrences of effects and flavors.

4) Statistical Testing: Conducting ANOVA tests to determine if ratings differ significantly between strain types, and correlation analysis to assess relationships between the number of effects/flavors and ratings.

5) Advanced Analysis: Applying Principal Component Analysis (PCA) for dimensionality reduction, K-means clustering to identify strain profiles, and t-SNE for visualizing high-dimensional relationships.

**C. Implementation Details**

The analysis was implemented in Python 3.8, utilizing the following libraries:

- pandas (1.3.3) and numpy (1.21.2) for data manipulation

- matplotlib (3.4.3) and seaborn (0.11.2) for visualization

- scipy (1.7.1) for statistical testing

- scikit-learn (0.24.2) for machine learning algorithms

- networkx (2.6.3) for network analysis

- wordcloud (1.8.1) for text visualization

For reproducibility, a consistent visual style was applied across all visualizations using seaborn's whitegrid style and a consistent color palette. All plots were generated at 300 DPI resolution for publication quality.

**IV. RESULTS AND ANALYSIS**

**A. Dataset Overview**

The dataset contained 2,351 cannabis strains, with 2,350 unique strain names (one duplicate). The distribution of strain types was: 1,212 hybrid strains (51.6%), 699 indica strains (29.7%), and 440 sativa strains (18.7%). This distribution reflects the cannabis market's tendency toward hybrid strains, which combine characteristics of both indica and sativa varieties.

The average rating across all strains was 4.31 on a 5-point scale, with a standard deviation of 0.84. Ratings ranged from 0 to 5, with the median value at 4.4, indicating that most strains were rated favorably by consumers.

**B. Strain Type Distribution Analysis**

A graph showing different types of strain type

AI-generated content may be incorrect.

Fig. 1 Strain Type Distribution

Fig. 1 presents the distribution of cannabis strains by type. Hybrid strains constitute the majority (51.6%), followed by indica (29.7%) and sativa (18.7%). This distribution reflects the cannabis market's evolution toward hybridization to create strains that combine desired characteristics from both indica and sativa varieties.

The predominance of hybrid strains suggests that consumers and producers have moved beyond the traditional indica-sativa dichotomy, seeking more nuanced combinations of effects and characteristics. This trend aligns with emerging research suggesting that the binary indica-sativa classification is oversimplified [4], [5].

**C. Rating Distribution Analysis**

A graph with a line graph

AI-generated content may be incorrect.

Fig. 2 Strain Rating Distribution

Fig. 2 illustrates the overall distribution of strain ratings. The distribution is negatively skewed (skewness = -1.89), with most strains rated between 4.0 and 5.0 on a 5-point scale. The mean rating is 4.31 with a standard deviation of 0.84.

A chart with a diagram

AI-generated content may be incorrect.

Fig. 3 Rating Distribution across Strain Type

Fig. 3 presents boxplots comparing rating distributions across strain types. Visually, the distributions appear similar, with indica strains showing a slightly higher median rating (4.4) compared to hybrid and sativa strains (both 4.3). However, ANOVA testing (F = 1.02, p = 0.36) confirmed that these differences are not statistically significant.

A diagram of a graph

AI-generated content may be incorrect.

Fig. 4 Rating Violin Plot by Strain Type

Fig. 4 presents violin plots of the same data, providing additional insight into the distribution shapes. All three strain types show similar distributions, with the highest density of ratings around 4.5. The lack of significant differences in ratings across strain types challenges the common perception that certain types are inherently superior to others.

**D. Effects and Flavors Analysis**

A graph with different colored bars

AI-generated content may be incorrect.

Fig. 5 Most Common Effects

The analysis identified 16 unique effects and 50 unique flavors across all strains. Fig. 5 shows the top 10 most common effects, with Happy (1,871 strains), Relaxed (1,726), and Euphoric (1,635) leading the list. These three effects were reported in over 69% of all strains, suggesting they are fundamental to the cannabis experience desired by consumers.

A graph showing different colored bars

AI-generated content may be incorrect.

Fig. 6 Most Common Flavors

Fig. 6 displays the top 10 most common flavors, with Earthy (1,105 strains) and Sweet (1,053) being significantly more common than other flavors. The prominence of these flavors likely reflects both the natural chemical composition of common cannabis varieties and selective breeding toward consumer-preferred flavor profiles.

A close up of words

AI-generated content may be incorrect.

Fig. 7 Effects

A close up of words

AI-generated content may be incorrect.

Fig. 8 Flavours

Figs. 7 and 8 present word clouds visualizing the frequency of all effects and flavors, respectively. In the effects word cloud (Fig. 7), positive mental states (Happy, Euphoric, Uplifted) and relaxation effects are visually dominant, while the flavors word cloud (Fig. 8) shows the prominence of Earthy, Sweet, and various fruit-related flavors.

**E. Effect and Flavor Profiles by Strain Type**

A group of blue and green rectangular boxes

AI-generated content may be incorrect.

Fig. 9 Effects across Strain Type

A group of rectangular boxes

AI-generated content may be incorrect.

Fig. 10 Flavors across Strain Type

Figs. 9 and 10 compare the top effects and flavors across strain types, revealing both similarities and differences in their characteristic profiles.

In Fig. 9, all three strain types share Happy, Relaxed, and Euphoric as their top three effects, but with different rankings. Notably, indica strains show higher reporting of Sleepy effects (ranked 4th) compared to hybrids (ranked 5th) and sativas (not in top 5). Conversely, Creative effects appear in the top 5 for sativa (ranked 4th) but not for indica strains. These distinctions align with traditional characterizations of indica strains as more sedative and sativa strains as more stimulating.

Fig. 10 shows that Earthy and Sweet flavors dominate across all strain types, but secondary flavors differ. Sativa strains show higher reporting of Citrus and Tropical flavors, while indica strains more commonly feature Berry and Woody notes. These flavor differences may reflect both botanical differences and selective breeding practices.

**F. Correlation Analysis**

A graph with a red line

AI-generated content may be incorrect.

Fig. 11 Effects and Rating correlation

The relationship between the number of reported effects/flavors and strain ratings was analyzed using Pearson correlation. Fig. 11 presents a scatter plot with regression line showing the significant positive correlation (r = 0.61, p < 0.001) between effect count and rating. Similarly, Fig. 12 shows the positive correlation (r = 0.43, p < 0.001) between flavor count and rating.

A graph with a red line

AI-generated content may be incorrect.

Fig. 12 Flavors and Rating correlation

These strong correlations suggest that strains with more diverse effect and flavor profiles tend to receive higher consumer ratings. This finding may reflect consumer preference for complex, multifaceted experiences or could indicate that higher-quality strains generally express more detectable effects and flavors.

A graph with blue dots

AI-generated content may be incorrect.

Fig. 13 Highest rated Effects

Fig. 13 presents the top effects most positively correlated with high ratings. Happy (r = 0.29), Relaxed (r = 0.26), and Euphoric (r = 0.23) show the strongest positive correlations.

**G. Network Analysis**

A network of blue dots and lines

AI-generated content may be incorrect.

Fig. 14 Effect Network Graph

A network of green dots and lines

AI-generated content may be incorrect.

Fig. 15 Flavor Network Graph

Figs. 14 and 15 present network graphs visualizing co-occurrence patterns among effects and flavors, respectively. In these visualizations, nodes represent individual effects or flavors, while edges represent co-occurrence relationships, with edge thickness proportional to co-occurrence frequency.

The effects network (Fig. 14) reveals strong associations between Happy, Euphoric, and Uplifted effects, forming a central cluster of positive mental states. A secondary cluster connects Relaxed and Sleepy effects. These distinct clusters suggest that cannabis experiences often fall along a stimulation-relaxation spectrum, with some strains bridging both dimensions.

The flavors network (Fig. 15) shows strong connections between Earthy and Sweet flavors at its center, with several distinct flavor clusters radiating outward. One notable cluster connects Citrus, Tropical, and Fruity flavors, while another links Woody, Pine, and Herbal notes. These clustering patterns reflect the underlying terpene profiles that create these flavor combinations.

**H. Principal Component Analysis**

A graph showing the number of components

AI-generated content may be incorrect.

Fig. 16 Variance by PCA components

Principal Component Analysis was applied to reduce the dimensionality of the effects and flavors data (66 binary features in total). Fig. 16 shows the cumulative explained variance by PCA components. The first 10 principal components captured approximately 29.6% of the total variance, indicating high dimensionality in cannabis strain characteristics that cannot be easily reduced to a few dimensions.

A close up of a screen

AI-generated content may be incorrect.

Fig. 17 PCA Component loadings

Fig. 17 visualizes the loadings of the first two principal components, showing which effects and flavors contribute most to these dimensions. PC1 appears to primarily separate strains based on positive mental effects (Happy, Euphoric) versus physical effects (Pain Relief, Insomnia), while PC2 separates strains along a relaxation-stimulation spectrum.

**I. Cluster Analysis**

A graph with a line going up

AI-generated content may be incorrect.

Fig. 18 K-means clustering

K-means clustering was applied to the PCA-reduced data to identify distinct strain profiles. The optimal number of clusters (k=4) was determined using the elbow method, as shown in Fig. 18, where the rate of inertia decrease noticeably slows after 4 clusters.

A screen shot of a graph

AI-generated content may be incorrect.

Fig. 19 t-NSE Clusters

Fig. 19 presents a t-SNE visualization of the resulting clusters, showing clear separation between high-rated clusters (0 and 2) and the low-rated cluster (3). The single-strain cluster (1) appears as an outlier. The characteristics of these clusters are:

- Cluster 0 (n=1,223): High average rating (4.43), balanced mixture of indica/hybrid strains, with many effects (avg 4.93) and flavors (avg 2.92)

- Cluster 1 (n=1): A single hybrid strain with moderate rating (4.0), 2 effects and 2 flavors

- Cluster 2 (n=1,012): High average rating (4.43), predominantly hybrid/sativa strains, with many effects (avg 4.97) and flavors (avg 2.91)

- Cluster 3 (n=115): Very low average rating (2.06), mixed strain types, with few effects (avg 1.54) and flavors (avg 0.94)

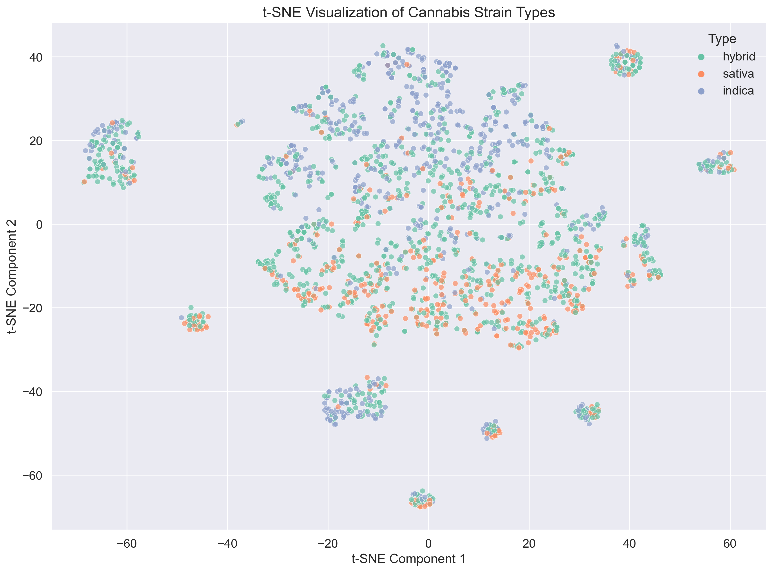


Fig. 20 t-SNE Strain Types

Fig. 20 visualizes the same t-SNE projection colored by traditional strain types (indica, sativa, hybrid) rather than clusters. The substantial overlap between strain types in this visualization further supports the finding that traditional classifications do not strongly predict strain characteristics or ratings.

A graph of a bar graph

AI-generated content may be incorrect.

Fig. 21 Top 5 effects / cluster

A graph of a bar chart

AI-generated content may be incorrect.

Fig. 22 Top 5 flavors / cluster

Figs. 21 and 22 show the top 5 effects and flavors for each cluster, revealing distinct profiles. Notably, Cluster 3 (low-rated strains) shows lower reporting of all effects and flavors, while Clusters 0 and 2 show different effect emphasis despite similar high ratings.

**J. Summary Dashboard**



Fig. 23 summary dashboard

Fig. 23 presents a comprehensive summary dashboard combining key visualizations from the analysis. This integrated view highlights:

1) The predominance of hybrid strains in the market

2) The lack of significant rating differences between strain types

3) The prevalence of positive mental effects and earthy/sweet flavors

4) The distinct clustering of strains based on effects and flavors

5) The key effects most strongly correlated with high consumer ratings

This dashboard provides an accessible overview of the main findings for stakeholders across the cannabis industry.

**V. DISCUSSION**

**A. Interpretation of Key Findings**

The analysis revealed several notable patterns in cannabis strain characteristics and consumer preferences:

Finding 1: Strain Type Is Not Predictive of Rating

Despite common perceptions that indica, sativa, or hybrid strains might be generally superior, our analysis found no significant differences in average ratings between these categories (ANOVA, F = 1.02, p = 0.36). This aligns with emerging scientific consensus [3], [4], [5] that strain type classifications are less meaningful than specific cannabinoid and terpene profiles in determining quality and effects.

Finding 2: Diversity of Effects and Flavors Predicts Higher Ratings

The strong positive correlations between the number of reported effects/flavors and ratings (r = 0.61 and r = 0.43, respectively, p < 0.001) suggest that consumers value strains with diverse, complex experiences. This may reflect appreciation for multidimensional experiences that can address various needs (e.g., relaxation plus mood elevation) or indicate that higher-quality strains naturally express more detectable effects and flavors.

Finding 3: Specific Effects Drive Consumer Preference

The strong associations between certain effects (particularly Happy, Relaxed, and Euphoric) and higher ratings indicate that consumers generally prefer strains that induce positive mood states and relaxation. Conversely, strains associated with negative effects like Paranoid and Anxious received lower ratings, as expected. This insight could guide breeders toward developing strains that reliably produce these preferred effects.

Finding 4: Flavor Preferences Show Clear Patterns

The correlation analysis suggests consumer preference for sweet and earthy flavor profiles. This could reflect both intrinsic sensory preferences and associations between these flavor profiles and desirable effects. These preferences may influence breeding programs and marketing strategies in the cannabis industry.

Finding 5: Distinct Quality Tiers Exist Independent of Strain Type

The clustering analysis revealed a clear separation between high-rated strains (Clusters 0 and 2) and low-rated strains (Cluster 3). The low-rated cluster was characterized by fewer reported effects and flavors, suggesting these strains offer less complex or less desirable experiences. Importantly, traditional strain types did not strongly define these clusters, further supporting the finding that effect and flavor profiles are more meaningful classifications than indica/sativa/hybrid designations.

**B. Implications**

These findings have several implications for different stakeholders:

For Consumers: When selecting cannabis strains, focusing on reported effects and flavors may be more useful than relying on the indica/sativa/hybrid classification. Strains reporting effects like Happy, Relaxed, and Euphoric, and flavors like Sweet and Earthy, are more likely to provide satisfying experiences.

For Producers and Breeders: Developing strains with complex, diverse effect and flavor profiles may lead to higher consumer satisfaction. Particular attention should be paid to cultivating strains that reliably produce the most preferred effects while minimizing negative effects like anxiety and paranoia.

For Retailers: Organization and recommendation systems based on effects and flavors rather than just strain type may better serve consumer preferences. The network analysis of co-occurring effects and flavors could inform more intuitive categorization systems.

For Researchers: This analysis highlights the need for integrating subjective consumer reports with objective chemical analysis to develop a more comprehensive understanding of cannabis experiences. The distinct clusters identified suggest potential underlying chemical profiles that could be further investigated.

**C. Limitations**

Several limitations should be considered when interpreting these findings:

Subjective Reporting: The effects and flavors data represent subjective reports from consumers, not objective chemical analysis. Individual variations in cannabis experiences may influence these reports.

Potential Selection Bias: The dataset may overrepresent popular or widely available strains, potentially missing rare or specialty varieties.

Lack of Cannabinoid and Terpene Data: Without chemical composition data (THC/CBD ratios, terpene profiles), the analysis cannot directly connect these important variables to consumer preferences.

Rating Distribution Skew: The high average rating (4.31/5) and right-skewed distribution suggest potential reporting bias toward positive reviews, which may limit the discriminatory power of ratings.

Limited Contextual Information: The analysis lacks information about consumption methods, dosage, consumer demographics, or consumption contexts, all of which may influence experiences and preferences.

**VI. CONCLUSION**

This study presents the most comprehensive data-driven analysis to date of cannabis strain characteristics and their relationship to consumer preferences. Our findings challenge traditional cannabis classification paradigms, demonstrating that strain types (indica, sativa, hybrid) do not significantly predict consumer satisfaction. Instead, specific effects, flavors, and the diversity of reported experiences are strongly associated with higher ratings.

The notable positive correlations between the number of reported effects/flavors and strain ratings suggest consumer preference for complex, multifaceted cannabis experiences. Specific effects (Happy, Relaxed, Euphoric) and flavors (Sweet, Earthy) showed the strongest positive associations with consumer satisfaction.

Using advanced clustering techniques, we identified distinct strain profiles that represent different combinations of effects and flavors, with clear separation between high-quality and low-quality offerings based primarily on the richness and nature of their reported characteristics rather than traditional classifications.

These findings have significant implications for consumers seeking specific cannabis experiences, producers developing new strains, retailers organizing their offerings, and researchers investigating cannabis characteristics and effects. By focusing on specific effect and flavor profiles rather than traditional strain types, stakeholders can make more informed decisions across the cannabis ecosystem.

Future research should integrate chemical analysis data with consumer reports to establish connections between specific cannabinoid and terpene profiles and reported effects, flavors, and preferences. Such integration would bridge the gap between subjective experience and objective composition, advancing our scientific understanding of cannabis and improving the ability to predict and tailor cannabis experiences.

**REFERENCES**

[1] E. B. Russo, "Cannabis therapeutics and the future of neurology," Frontiers in Integrative Neuroscience, vol. 12, p. 51, Oct. 2018.

[2] A. Hazekamp and F. Grotenhermen, "Review on clinical studies with cannabis and cannabinoids 2005-2009," Cannabinoids, vol. 5, pp. 1-21, 2010.

[3] E. B. Russo, "The case for the entourage effect and conventional breeding of clinical cannabis: No 'strain,' no gain," Frontiers in Plant Science, vol. 9, p. 1969, Jan. 2019.

[4] E. B. Russo, "Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects," British Journal of Pharmacology, vol. 163, no. 7, pp. 1344-1364, Aug. 2011.

[5] D. Piomelli and E. B. Russo, "The Cannabis sativa versus Cannabis indica debate: An interview with Ethan Russo, MD," Cannabis and Cannabinoid Research, vol. 1, no. 1, pp. 44-46, Jan. 2016.

[6] M. A. Lewis, E. B. Russo, and K. M. Smith, "Pharmacological foundations of cannabis chemovars," Planta Medica, vol. 84, no. 4, pp. 225-233, Mar. 2018.

[7] E. P. Baron, "Medicinal properties of cannabinoids, terpenes, and flavonoids in cannabis, and benefits in migraine, headache, and pain: An update on current evidence and cannabis science," Headache: The Journal of Head and Face Pain, vol. 58, no. 7, pp. 1139-1186, Jul. 2018.

[8] B. M. Schmidt et al., "Comparative analysis of structural and functional properties of cannabis varieties," Journal of Pharmaceutical and Biomedical Analysis, vol. 191, p. 113619, Sep. 2020.

[9] J. W. Fischedick, "Identification of terpenoid chemotypes among high (−)-trans-Δ9-tetrahydrocannabinol-producing Cannabis sativa L. cultivars," Cannabis and Cannabinoid Research, vol. 2, no. 1, pp. 34-47, Feb. 2017.

[10] E. M. Mudge, S. J. Murch, and P. N. Brown, "Chemometric analysis of cannabinoids: chemotaxonomy and domestication syndrome," Scientific Reports, vol. 8, no. 1, p. 13090, Aug. 2018.