Preference Learning for Color Palette Aesthetics

Lewis Chao
Stanford University
Stanford, CA 94305
lchao9@stanford.edu

1 Introduction & background

Human judgments of color harmony and aesthetic preference play a central role in visual design, branding, and data visualization, yet they are difficult to formalize in a principled way. Designers typically rely on intuitive rules of thumb or traditional color theory (e.g., complementary or analogous schemes), which provide qualitative guidance but lack quantitative validation. At the same time, large-scale online datasets of rated color palettes now make it possible to study color aesthetics empirically. Building on this opportunity, our project uses a five-color MTurk palette corpus to ask whether interpretable machine learning methods can capture the structure of human preferences well enough to support prediction, ranking, and ultimately tool support for non-expert designers.

The central research question is whether engineered color features, combined with classical preference-learning models, can reliably predict which of two palettes a human is more likely to prefer. To address this question, we consider two complementary formulations that mirror the course material. First, we train a sparse linear (LASSO) regression model to predict user-normalized ratings from palette-level features, treating preference as a pointwise prediction problem. Second, we adopt a Bradley–Terry style logistic model over palette pairs, using feature differences to model the probability that one palette is preferred to another. By comparing these models on shared test data, we evaluate both their predictive performance and the extent to which their learned feature weights provide consistent, interpretable explanations of color preference.

This work is grounded in prior literature on preference learning and color aesthetics. Bradley–Terry is widely used to analyze pairwise comparisons, and they form a core example of preference learning in CS329H. In parallel, color harmony research and computational studies such as O'Donovan et al. [2] have shown that relatively simple feature-based models can predict palette ratings from properties like lightness balance, hue range, and contrast. More recent work has explored pairwise color relations [3] and personalized aesthetic prediction [4] using more complex neural models. Our project situates itself at the intersection of these lines of work by combining hand-engineered features in multiple color spaces with interpretable linear and Bradley–Terry-style models, aiming to bridge theoretical insights from color science with practical, reproducible preference-learning techniques.

2 Methods

We use the publicly available dataset introduced by O'Donovan et al. [2], which contains 10,743 five-color palettes sourced from the Adobe Kuler platform. Each palette was evaluated by 40 independent participants on Amazon Mechanical Turk, who rated its aesthetic appeal using a 5-point Likert scale (1 = least pleasing, 5 = most pleasing). We model human preferences over color palettes by first transforming each five-color palette into a comprehensive feature representation that encodes structural properties across multiple color spaces and summarizes hue distribution statistics. Two complementary models are then trained on these features: a sparse linear regression model that predicts user-normalized aesthetic ratings, and a Bradley–Terry–style logistic regression model that learns relative preferences from synthetic pairwise comparisons. Model

performance is evaluated using both pointwise and pairwise metrics, followed by detailed feature-importance and error analyses, and multi-seed experiments with statistical tests to confirm the robustness and reproducibility of the results.

2.1 Dataset and prediction task

We base our study on an MTurk color-palette dataset in which each example is a five-color palette annotated with human preference ratings. For each palette we have:

- A unique identifier (ids) and a textual name ($names$).
- A mean crowdsourced rating $targets$.
- A user-normalized rating $userNormalizedTargets$, where each user's ratings are z-scored and then aggregated; this is our main target variable.
- Five RGB triplets (color1_r, color1_g, color1_b, …), with each component normalized to lie in [0,1].

Our primary supervised learning task is to predict $userNormalizedTargets$ from palette-level features. We also induce a pairwise preference model over palettes and evaluate how well it recovers human ordering on synthetic pairwise comparisons.

2.2 Feature engineering

All feature construction is implemented in a single pipeline that converts raw RGB values into multiple color spaces and derives rich palette-level descriptors.

- **Color-space representations**: For each five-color palette, we transform the RGB values into multiple color spaces to capture complementary perceptual dimensions. In *Lab*, colors are converted from RGB → XYZ → Lab using a D65 reference white, with channels normalized as L/100, a/128, and b/128 for scale consistency. In *HSV*, MATLAB's rgb2hsv provides hue, saturation, and value in [0, 1). To better encode hue's circular nature, we also construct a *CHSV* (circular HSV) representation: hue values are remapped via a cubic-spline fit to empirical hue statistics, and each color is represented as $[\mathrm{S}\cos(2\pi\mathrm{H}), -\mathrm{S}\sin(2\pi\mathrm{H}), \mathrm{V}]$ where H is the remapped hue, S saturation, and V value. Each palette thus yields a 3 × K matrix ($K = 5$) in every color space.

- **Palette-level descriptors**: From each color-space matrix, we compute several families of palette-level features capturing structure, contrast, and color relationships. We first flatten per-color coordinates (e.g., *lab-D1-C2*, *hsv-D3-C4*) and flatten versions sorted by brightness to encode ordering effects. To model transitions, we calculate adjacent coordinate-wise differences between colors, treating hue circularly in HSV, and then sort these differences to capture the overall distribution of color jumps. For each channel (e.g., L, a, b or H, S, V), we derive summary statistics—mean, standard deviation, median, extrema, and range—to quantify overall lightness, chroma, and contrast. Finally, for RGB, Lab, and CHSV spaces, we fit a plane via SVD to the five color points and record the plane normal, explained variance ratios, and residual distances, describing how linearly or spatially coherent each palette is within its color space.

- **Hue-probability features**: Beyond geometric descriptors, we incorporate data-driven hue statistics, which contains empirical distributions over hue, hue pairs, and hue adjacencies. For each palette, we identify "visible" hues based on saturation and value thresholds and then compute summary statistics (mean, standard deviation, min, max) over: univariate hue probabilities of visible hues and adjacency probabilities between consecutive hues in the palette.

- **Pruning and final feature set**: After feature construction, we remove the original raw RGB columns, retaining only engineered features plus identifiers and ratings. We then prune highly correlated features: we compute the absolute correlation matrix over all non-core features and drop one feature from any pair with $|corr| > 0.95$, while always retaining ids, $names$, $targets$, and $userNormalizedTargets$. The result is a palette-feature matrix with 288 decorrelated features.

2.3 Models

We compare two modeling approaches: a rating-based baseline using LASSO regression, and a Bradley–Terry–style pairwise model implemented as logistic regression on feature differences.

Rating-based baseline (LASSO regression)

Our first model predicts the continuous target $userNormalizedTargets$ directly from the engineered features via LASSO regression. Implementation details are as follows:

- We perform a random train–test split with test size 0.2. For the single-seed analysis, we set random state to be 0; for multi-seed experiments we vary this seed.
- Input features are all non-core columns (excluding *ids, names, targets, userNormalizedTargets*).
- We standardize features using StandardScaler fit on the training data and applied to both train and test.
- We fit a LASSO model with L1 regularization parameter α = 0.001 and maximum iteration = 10000. This is a sparse linear regression model of the form

$$\hat{y} = w^T x + b,$$

  with an L1 penalty $\lambda \, ||w||_1$.

- Finally, we report RMSE on both training and test sets and check for overfitting by comparing the two.

This model directly draws on course material on linear regression and regularization: LASSO encourages sparse solutions, which enhances interpretability through the learned coefficients and mitigates overfitting in high-dimensional feature spaces.

Bradley–Terry–style pairwise model (logistic regression)

Our second model is motivated by the Bradley–Terry framework for pairwise comparisons. Instead of predicting ratings, we predict preferences between pairs of palettes.

We first synthesize pairwise training data:

- From the training split, we repeatedly sample unordered pairs of indices (i, j) with replacement.
- We remove self-pairs and any pairs with $|y_i - y_j| < \delta$, where $\delta = 0.2$, to avoid near-tie comparisons.
- For each remaining pair we define a binary label

$$y_{ij} = 1 \; if \; userNormalizedTargets_i > userNormalizedTargets_j, 0 \; otherwise.$$

We continue sampling until we obtain 10,000 valid training pairs; the function reports how many candidates were sampled and the empirical label distribution.

To approximate the Bradley–Terry model in feature space, we define the pairwise log-odds as a linear function of the difference in latent utilities. We model this by taking feature differences:

$$X_{diff} = X_i - X_j,$$

where X_i and X_j are the standardized feature vectors for palettes i and j. We then fit a logistic regression model

$$P\big(i \succ j \,\big|\, X_i, X_j\big) = \sigma\Big(w^T\big(X_i - X_j\big)\Big),$$

where σ is the logistic sigmoid. In practice we use sklearn.linear_model.LogisticRegression with L2 regularization (penalty="l2", C=1.0, solver="lbfgs", max_iter=1000), and we standardize X_{diff} via StandardScaler fit on the training differences. This formulation is exactly the Bradley–Terry style approach covered in class.

2.4 Experimental protocol and analyses

We evaluate both models using a combination of pointwise rating metrics, pairwise preference metrics, feature-importance analyses, and qualitative visualizations.

Pointwise evaluation of the rating model

On the test split, we evaluate the LASSO model's rating predictions:

- Compute RMSE for training and test sets.
- Generate a scatter plot of true vs predicted *userNormalizedTargets*

Pairwise evaluation on shared test pairs

To compare the BT model and the LASSO baseline fairly, we evaluate them on an identical set of synthetic test pairs:

- From the BT test split, we sample up to 20,000 pairs subject to the same $\delta = 0.2$ filter, yielding test labels y_{ij} as defined above.
- For each pair:
  - The BT model provides a probability $p_{ij} = P(i \succ j)$ from logistic regression on $X_i - X_j$.
  - The LASSO model provides scalar scores $\hat{y}_i$ and $\hat{y}_j$. We interpret $i \succ j$ if $\hat{y}_i > \hat{y}_j$.
- For both models we compute:
  - Pairwise accuracy: the fraction of pairs for which the predicted preference matches the ground truth.
  - ROC AUC:
    - For BT, using p_{ij}.
    - For LASSO, using the score difference $\hat{y}_i - \hat{y}_j$ as a continuous score.

This evaluation protocol directly measures how well each model recovers pairwise preferences, not just absolute ratings. Qualitatively, these analyses reveal which aspects of palette structure (e.g., Lab lightness statistics, hue adjacency, RGB contrast) drive rating prediction versus pairwise preference.

Feature importance and visualization

We analyze feature importances and visualize palettes to better understand the learned models.

- For LASSO, we treat the absolute value of each coefficient as a measure of importance and sort features by $|\beta|$. We print the top features and their magnitudes.
- For BT, we analogously sort by the absolute logistic regression coefficients applied to the feature differences. We print the top coefficients and compare the resulting feature ranking to LASSO's.
- We create horizontal bar plots of the top ~20 features for each model
- We compare the top feature sets:
  - Overlap between the two top-lists.
  - Features that appear only in the LASSO top-list.
  - Features that appear only in the BT top-list.

Qualitatively, these analyses reveal which aspects of palette structure (e.g., Lab lightness statistics, hue adjacency, RGB contrast) drive rating prediction versus pairwise preference.

Error analysis

We perform several error analyses on the shared test pairs:

- BT worst mistakes (high-confidence errors). We define a certainty measure for BT as $max(p_{ij}, 1 - p_{ij})$. Among misclassified pairs, we select those with highest certainty and visualize up to three via 2×5 color swatches, reporting their indices, p_{ij}, and true labels. These highlight cases where the model is confidently wrong.
- BT borderline correct cases. Among correctly classified pairs, we sort by $|p_{ij} - 0.5|$

and report the cases closest to 0.5. These are the most ambiguous comparisons where the model is almost indifferent yet correct, illustrating the decision boundary in practice.

- BT vs LASSO disagreements. We find pairs where BT and LASSO disagree and define a simple "strength of disagreement" score as BT certainty multiplied by $|\hat{y}_i - \hat{y}_j|$. We rank these and inspect the top few, logging indices, BT probabilities, LASSO score differences, each model's predicted label, and the true label. These pairs reveal where the two modeling paradigms—pairwise BT vs pointwise LASSO—make qualitatively different judgments.

3 Results & discussion

3.1 Overall predictive performance

We first assess how well the rating-based LASSO baseline predicts user-normalized ratings and how well both models recover pairwise preferences. Recall that the standard deviation of *userNormalizedTargets* in the dataset is approximately 0.33, so RMSEs around 0.22–0.23 correspond to explaining a substantial fraction of the variance.

For a representative single train–test split (seed 0), the LASSO model achieved a test RMSE of approximately 0.224 on *userNormalizedTargets*, with very similar performance on the training set. This indicates that, given the engineered features, a sparse linear model can capture much of the structure in the ratings without strong evidence of overfitting. The scatter plot of true vs. predicted ratings shows points tightly concentrated around the diagonal, with somewhat increased spread near the extremes, consistent with slightly higher difficulty in predicting very low- or high-rated palettes.

On the same test split, we compared the Bradley–Terry (BT) logistic model and the LASSO baseline on an identical set of synthetic test pairs. The BT model achieved a pairwise accuracy of roughly 0.858 and a ROC AUC of about 0.937, while the LASSO-induced pairwise predictor achieved a similar accuracy of about 0.857 and AUC of roughly 0.935. Thus, both models recover human preferences for most comparisons and rank palettes in a broadly consistent way.

To assess robustness, we repeated the entire pipeline across five random seeds. Averaged over these runs, LASSO's test RMSE was 0.2245 with a standard deviation of 0.0041, indicating stable pointwise performance. The BT model's mean pairwise accuracy across seeds was 0.8646 (std 0.0063), while LASSO's was 0.8613 (std 0.0075). A paired t-test on per-seed pairwise accuracies yielded a t-statistic of 2.24 and $p = 0.0884$, suggesting that BT tends to outperform LASSO slightly on pairwise accuracy, but that this advantage is not statistically significant at the conventional 0.05 level. Overall, both models perform strongly and comparably, with a weak trend in favor of the explicitly pairwise BT formulation.

3.2 Feature importance and learned representations

We next analyze which features are most influential for each model. Since both LASSO and BT are linear in the engineered feature space (for ratings and differences, respectively), their coefficients admit a straightforward interpretation as feature importances.

For the LASSO model, the largest-magnitude coefficient by a wide margin is associated with *labMean-D1*, the mean Lab lightness across the palette. This confirms the intuitive idea that overall lightness is a primary determinant of perceived palette quality. Other highly ranked LASSO features include global RGB means (*rgbMean-D1*, *rgbMean-D2*, *rgbMean-D3*), the mean value channel (*hsvMean-D3*), and various contrast-related statistics such as *rgbMaxMinDiff-D2*, *hsvMaxMinDiff-D1*, and higher-order brightness descriptors like *hsvMedian-D3* and sorted difference features. Together, these patterns suggest that the rating model pays attention to both the average brightness/chroma of the palette and the spread of colors within it.

For the BT model, the most influential features are also dominated by lightness but exhibit a more localized, color-by-color emphasis. The top features include *labMean-D1*, *labMax-D1*, and several

individual lightness coordinates and their sorted variants (e.g., *lab-D1-C2, lab-D1-C3, labSorted-D1-C1–C4*). In addition, hue-probability features such as *hsvHueAdjProbLogMin* and *hsvHueAdjProbLogMax* appear high in the ranking, indicating that the pairwise model exploits information about how "typical" the palette's hue transitions are. The feature importance bar plots (Fig. 1) make these differences visually apparent: LASSO allocates weight more diffusely across global summary statistics, while BT concentrates weight on a smaller set of specific lightness and hue-adjacency cues.

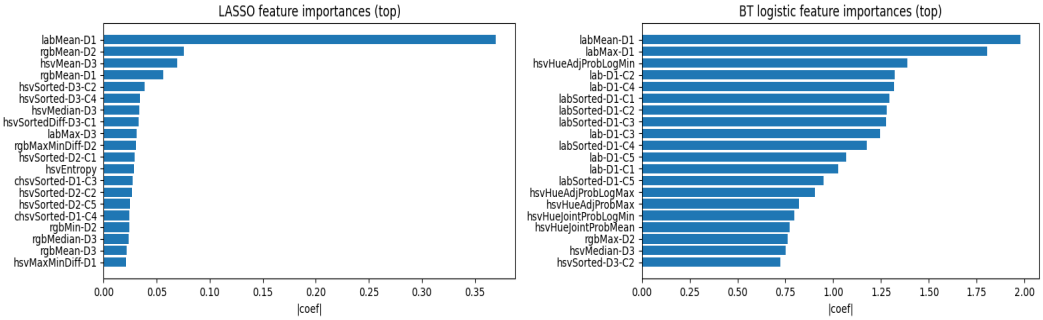

Figure 1: Feature importance scores from LASSO and Bradley–Terry models.

Comparing the two models' top feature lists reveals both commonalities and divergences. Both models rank *labMean-D1* their most important features, underscoring the centrality of palette lightness in human judgments. They also agree that brightness-related HSV summaries like *hsvMedian-D3* and sorted value features (*hsvSorted-D3-C2*) are informative. On the other hand, LASSO relies more heavily on global RGB statistics and max–min contrasts, while BT places more emphasis on individual lightness coordinates and hue-adjacency probability features, consistent with its focus on pairwise discrimination rather than absolute ratings.

3.3 Preference behavior and qualitative palette examples

Quantitatively, BT's and LASSO's pairwise accuracies are very similar, but their qualitative behavior sheds additional light on what they have learned. We begin by examining the test pairs on which the BT model is most certain. Among pairs with non-degenerate probabilities, BT often outputs probabilities around 0.995 for the preferred palette, indicating extreme confidence. Visual inspection of these "most certain" pairs (Fig. 2) reveals that BT tends to confidently favor palettes whose colors exhibit coherent lightness structure and harmonious hue transitions over palettes with harsher contrasts or more unusual hue combinations.

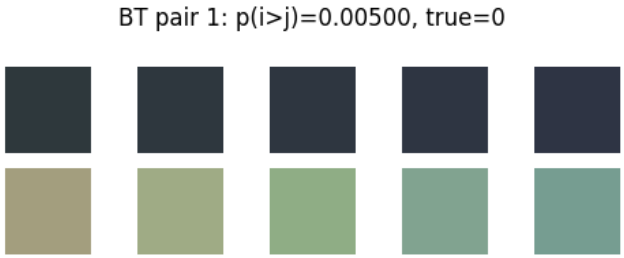

Figure 2: Most certain pairs of BT model with correct prediction (truth: top $\prec$ bottom).

We then turn to error analysis. The BT worst mistakes—pairs where BT predicts the wrong winner with very high certainty—highlight the limits of our feature set and model. For example, some mistakes involve palettes that are both light but differ in subtle aspects of color semantics (e.g., muted vs. vibrant shades) that are not fully captured by low-level statistics, leading BT to strongly prefer the "wrong" palette. An example of the worst cases (Fig. 3) suggests that incorporating higher-level attributes (e.g., semantic color labels or context) might be necessary to

resolve such subtleties.

Figure 3: Most certain pairs of BT model with incorrect prediction (truth: top $\prec$ bottom).

Borderline correct cases, where BT's probability is very close to 0.5 but the prediction happens to be correct, tend to involve palettes with very similar mean ratings. In these examples, the model is nearly indifferent, and small perturbations in the data or model parameters could easily flip the prediction. This behavior is reassuring: the model's uncertainty correctly reflects the difficulty of these comparisons.

Finally, we analyze strong disagreements between BT and LASSO, focusing on pairs where the two models give opposite predictions while both being reasonably confident. In several such cases, BT appears to rely on fine-grained lightness and hue-adjacency patterns, favoring palettes with more consistent lightness progression even when LASSO's predicted ratings are higher for the alternative. Conversely, LASSO occasionally prefers palettes with somewhat higher global brightness or contrast, even when their internal hue transitions are less typical. These disagreements, which we log along with palette indices and scores, highlight how pairwise models and rating models can prioritize different aspects of palette structure, even when their aggregate performance is similar.

3.4 Comparison to baselines and related approaches

Our study focuses on two relatively simple linear models—a rating-based LASSO model and a BT-style pairwise logistic model. Despite their simplicity, these models achieve strong performance: test RMSE of approximately 0.22–0.23 on standardized ratings and pairwise accuracies above 0.85 with ROC AUC around 0.94. These numbers compare favorably to what one might expect from trivial baselines, such as predicting the global mean rating or random preference, and demonstrate that our hand-engineered feature set is informative.

While we do not directly compare to complex non-linear models (e.g., deep neural networks over raw RGB sequences), the structure of our models aligns with classic approaches in the literature: linear regression and LASSO for interpretable prediction, and Bradley–Terry or Thurstone-type models for pairwise comparison. The fact that relatively low-capacity models already capture a large fraction of the structure suggests that much of the signal in this dataset is accessible through low-level color statistics and simple geometric relationships. More sophisticated models might yield incremental gains, but our results already provide a strong, interpretable baseline.

3.5 Limitations and lessons learned

Several limitations of our approach deserve discussion. First, our pairwise training data are synthetic: we derive pair labels from continuous ratings rather than collecting true pairwise judgments. This assumes that differences in *userNormalizedTargets* translate monotonically into preferences, and that individual-level noise averages out, which may not always hold. True pairwise MTurk data would provide a more direct test of Bradley–Terry assumptions.

Second, our models are linear in a hand-engineered feature space. While this makes them easy to interpret and connect to course material, it may limit their ability to capture complex non-linear relationships or interactions between colors. For example, semantic associations (e.g., "autumn" vs. "neon") or higher-order compositional structure are not explicitly modeled. Exploring kernel

methods or neural architectures over palette representations could be a fruitful direction.

Despite these limitations, the project yielded several useful insights. We found that relatively simple models, when paired with rich color-space features and hue statistics, can predict human palette ratings and pairwise preferences with high accuracy. Lightness-related features—particularly Lab mean lightness—consistently emerge as the most important predictors across models, underscoring the central role of brightness structure in perceived palette quality. Hue-adjacency and hue-probability features also matter, especially for the pairwise BT model, suggesting that how colors are arranged around the color wheel influences preference beyond mere hue counts.

We also learned that rating-based models and BT models, while achieving similar aggregate performance, can disagree substantially on specific cases. These disagreements expose complementary perspectives: LASSO emphasizes global summary statistics, whereas BT emphasizes fine-grained contrasts and transitions. This reinforces a key theme from the course: modeling choices (pointwise vs. pairwise, absolute scores vs. relative comparisons) shape what structure the model can exploit and how its errors manifest. Overall, the combination of interpretable feature engineering, classic linear models, and careful evaluation provides a solid foundation for future work on more expressive models and richer data.

4 Conclusion & future work

This project set out to model human preferences over five-color palettes using interpretable feature engineering and preference-learning methods inspired by research on color harmony and Bradley–Terry–style models of choice. Methodologically, we contributed (i) a comprehensive palette representation spanning multiple color spaces (RGB, Lab, HSV, CHSV) augmented with hue-probability and geometric descriptors, (ii) a rating-based LASSO baseline that predicts user-normalized aesthetic scores, and (iii) a Bradley–Terry logistic model trained on synthetic pairwise comparisons. Empirically, we found that both models achieve strong and consistent performance: the LASSO model attains a test RMSE of roughly 0.22–0.23 on standardized ratings, while both BT and LASSO recover pairwise preferences with accuracies above 0.85 and ROC AUC near 0.94 on shared test pairs, with a small but not statistically significant advantage for BT. Multi-seed experiments confirm the stability of these findings, and feature-importance and error analyses reveal that lightness structure and hue-adjacency statistics are central to predicting palette preference.

These results extend prior work highlighting the roles of lightness balance, limited hue range, and smooth progression in color harmony. The prominence of Lab lightness and hue-adjacency features as top predictors quantitatively supports classic design principles, while the Bradley–Terry formulation links our findings to well-established probabilistic choice frameworks. Overall, a simple, interpretable model using structured color features captures much of the variance in human palette preferences, providing a strong baseline for comparison with future deep or generative approaches.

Future research should collect true pairwise judgments (rather than synthetic ones from ratings) to test Bradley–Terry and related models more faithfully, and explore richer nonlinear formulations—such as kernelized BT, gradient-boosted trees, or neural architectures—that capture higher-order interactions and semantic color relationships. Incorporating user and context variables (e.g., cultural background, design task) could further explain heterogeneity in taste. Finally, human-in-the-loop evaluation, where model predictions guide interactive palette-recommendation tools, offers a practical next step to assess how well these models enhance user satisfaction and design quality.

References

[1] Schloss, K. B., & Palmer, S. E. (2011). *Aesthetic response to color combinations: Preference, harmony, and similarity. Attention, Perception, & Psychophysics, 73*(2), 551–571.

[2] O'Donovan, P., Agarwala, A., & Hertzmann, A. (2011). *Color compatibility from large datasets. ACM Transactions on Graphics (TOG),* 30(4), 63:1–63:12.

[3] Yang, S., Wang, M., Lin, Y., & Chen, W. (2019). *A color-pair based approach for accurate color harmony estimation*. *Journal of Visual Communication and Image Representation*, 59, 443–450.

[4] Yang, X., Zhou, F., Wang, W., & Wang, Y. (2024). *Personalized aesthetic prediction for color themes via probabilistic modeling*. *ACM Transactions on Applied Perception (TAP),* 21(2), Article 12.

GitHub Repository

https://github.com/lewiseng/preference-learning-for-color-palette-aesthetics

Appendix I: Pre-analysis Plan

Preference Learning for Color Palette Aesthetics: Pre-Analysis Plan

Lewis Chao
Stanford University
Stanford, CA 94305
lchao9@stanford.edu

1 Overview & motivation

This project investigates modeling human aesthetic preferences for color palettes using interpretable machine learning models trained on pairwise comparison data. The guiding research question is: *Can an interpretable model learn to predict which of two color palettes is more aesthetically pleasing to a human observer?* This builds on CS329H topics including preference learning via Bradley–Terry model, human-in-the-loop learning, and alignment with human judgment.

Effective color choice greatly impacts user experience and communication in design [1]. However, designers often rely on intuition or traditional color theory rules (e.g., complementary or triadic schemes), which lack quantitative validation. Data-driven preference learning can verify or refine these theories and provide personalized recommendations. In practice, an aesthetic preference model could power tools to recommend or adjust palettes to suit user tastes, improving efficiency for non-experts.

This plan is designed for a solo researcher working within a short timeframe. The scope is realistic: we will reuse an existing dataset of color palettes and their preference ratings (or pairwise comparisons). The methods will emphasize simplicity and interpretability – for example, using linear or small models and easily computable color features – to keep implementation manageable and to allow insight into why a palette is preferred. This focused scope ensures we can complete the work and obtain meaningful results in the available time.

2 Literature review

Color palette aesthetics have received significant attention in both design research and computational modeling. Several studies have attempted to quantify what makes a color palette visually appealing, typically through predictive modeling or data-driven heuristics. Our project builds on this literature but brings a novel focus on pairwise preference learning. Below, we review key related works.

- **O'Donovan et al. (2011)** – "Color Compatibility from Large Datasets"
  This foundational study collected over 10,000 color palettes from Adobe Kuler and COLOURLovers and trained a LASSO regression model to predict aesthetic scores based on features like hue distribution, lightness, and contrast. They found that palettes with moderate contrast and high lightness tended to receive higher ratings. The model was interpretable and effective, showing that even simple linear models could predict human ratings [1]. However, their method relied on absolute aesthetic scores, which limited its ability to capture nuanced or relative preferences. Our project draws directly from their feature design but instead applies a pairwise preference model (Bradley–Terry) to better reflect comparative judgments.

- **Yang et al. (2019)** – "A Color-Pair Based Approach for Accurate Color Harmony Estimation"
  This study proposed a two-stage model: first estimating harmony between individual

color pairs, then aggregating pairwise scores to assess the overall aesthetic of a palette [3]. The approach improved upon prior models by capturing local interactions between colors rather than treating palettes as flat feature sets. Though highly relevant, this model also used regression-style outputs and did not incorporate preference learning directly. Our project shares their attention to color relationships (e.g., pairwise hue distance) but instead emphasizes learning from pairwise comparisons and directly modeling which palette is preferred over another.

- **Yang et al. (2024)** – "Personalized Aesthetic Prediction for Color Themes"
  This more recent work acknowledged preference heterogeneity by modeling user-specific aesthetics. The authors used neural networks with user embeddings to capture individual differences in taste, achieving better predictive accuracy on personalized ratings [4]. While promising, this method is computationally intensive and requires large user datasets, making it less accessible or interpretable. Our project opts for model simplicity and interpretability over personalization, but we recognize the value of individual modeling for future extensions (e.g., active learning or user-specific preference elicitation).

While past work has made substantial progress in predicting aesthetic quality using rating-based models and complex neural networks, pairwise preference modeling remains underutilized in color theory applications. Our approach provides a structured and interpretable way to model relative preferences, aligning with human comparative behavior and the theoretical foundations of preference learning.

3 Proposed methods & analysis

This project aims to model human preferences for color palettes using interpretable preference learning methods, primarily the Bradley–Terry model. The modeling process will involve transforming color palettes into structured feature representations, deriving pairwise preference data from existing sources, fitting a model to these comparisons, and evaluating its predictive performance on held-out comparisons.

3.1 Dataset and preprocessing

We will use the publicly available dataset introduced by O'Donovan et al. (2011), which contains 10,743 five-color palettes sourced from the Adobe Kuler platform. Each palette was evaluated by 40 independent participants on Amazon Mechanical Turk, who rated its aesthetic appeal using a 5-point Likert scale (1 = least pleasing, 5 = most pleasing). The final dataset provides the average rating for each palette, offering a consistent and controlled benchmark for modeling human aesthetic judgments. Compared to organically collected community data (e.g., from Kuler or COLOURLovers), this dataset minimizes popularity and exposure bias. Although the data consists of absolute ratings, our project focuses on pairwise preference learning. To align with this framework, we generate synthetic comparisons by randomly sampling palette pairs and assigning the preference label to the palette with the higher mean rating. Ties will be excluded or resolved randomly. While this approach does not replicate explicit head-to-head evaluations, it is an approximation in preference learning and enables the application of Bradley–Terry to derive meaningful insights from rating-based data.

3.2 Feature representation

Each palette will be encoded as a vector of interpretable features rooted in color theory. These features are designed to capture the visual properties that might influence aesthetic appeal and can be grouped into four broad categories:

- **Color Statistics**: Measures of central tendency and dispersion of color values, particularly in perceptual color spaces such as CIE Lab or HSV. For example:
  - Mean and standard deviation of lightness (L*).
  - Mean saturation and hue variance.
  - Minimum and maximum brightness.

- **Color Harmony Metrics**: Classical color theory suggests that certain relationships—like complementary or analogous colors—are more harmonious. We will compute:
  - Average pairwise hue distance (in degrees on the color wheel).
  - Presence of complementary (180° hue separation) or triadic (120° separation) color arrangements.
  - Hue balance or imbalance scores.
- **Color Diversity and Contrast**:
  - Number of unique hues.
  - Entropy of hue distribution.
  - Contrast across brightness or saturation.
- **Palette Structure Features**:
  - Monotonicity of lightness or saturation across the ordered palette.
  - Whether the palette forms a perceptual gradient.
  - Spatial or ordering features (e.g., "dark to light" sequences).

3.3 Preference model

We will fit a Bradley–Terry regression model, where the preference between two palettes is modeled as a function of their respective features. Let x_i and x_j be the feature vectors of two palettes. The probability that palette i_i is preferred over j_i is given by:

$$P(i \succ j) = \frac{\exp(\beta^T x_i)}{\exp(\beta^T x_i) + \exp(\beta^T x_j)}$$

Here, β is a vector of learned coefficients, representing the contribution of each feature to the overall aesthetic "worth" of a palette. The model is equivalent to a logistic regression over the difference in feature vectors between two items. Training will proceed via maximum likelihood estimation over the set of comparison outcomes.

This model is ideal for our setting for several reasons:

- It directly reflects the pairwise structure of the data.
- It's interpretable: each weight β_k tells us whether and how much a specific feature (e.g., lightness) contributes to preference.
- It generalizes to unseen palettes via features, unlike item-level BT models which require repeated items.

As a baseline, we will also implement a simple regression model (e.g., linear regression or LASSO) to predict numeric ratings directly. This allows us to compare the performance of models trained on ratings versus those trained on comparisons.

3.4 Evaluation plan

We will evaluate model performance using a combination of metrics suited to preference prediction:

- **Pairwise Accuracy:** The percentage of held-out comparisons where the model correctly predicts which palette is preferred.
- **Spearman's Rank Correlation:** Measures agreement between predicted aesthetic scores and ground-truth rankings.
- **AUC (Area Under the ROC Curve):** For binary classification of preferred vs. non-preferred palettes, this metric summarizes the model's ability to distinguish aesthetically superior palettes.
- **Interpretability Review:** We will analyze the learned coefficients $\beta$to identify which features most strongly influence preferences. This will also serve as a sanity check against known findings (e.g., gradient palettes tend to be preferred).

We will split the data into training and test sets, ensuring no data leakage across comparisons. We may perform k-fold cross-validation for robustness.

4 Timeline & responsibilities

Week 1 (Nov 5–11): Finalize literature review, obtain dataset, define and extract palette features.

Week 2 (Nov 12–18): Implement pairwise comparison model (Bradley–Terry with feature vector), train using synthetic comparisons, evaluate on held-out data, and conduct baseline regression comparison.

Week 3 (Nov 19–25): Perform simulation-based active learning experiments, analyze model weights, and interpret learned aesthetic indicators in terms of color theory.

Week 4 (Nov 26–Dec 3): Write final report and prepare figures and supplementary results. All coding, writing, and analysis will be performed by the solo researcher.

Risks include limited dataset quality or low model performance. These will be mitigated by fallback to open datasets and simpler logistic or ranking models. The project emphasizes simplicity, interpretability, and clear connection to course concepts.

References

[1] Schloss, K. B., & Palmer, S. E. (2011). *Aesthetic response to color combinations: Preference, harmony, and similarity. Attention, Perception, & Psychophysics, 73*(2), 551–571.

[2] O'Donovan, P., Agarwala, A., & Hertzmann, A. (2011). *Color compatibility from large datasets. ACM Transactions on Graphics (TOG),* 30(4), 63:1–63:12.

[3] Yang, S., Wang, M., Lin, Y., & Chen, W. (2019). *A color-pair based approach for accurate color harmony estimation*. *Journal of Visual Communication and Image Representation*, 59, 443–450.

[4] Yang, X., Zhou, F., Wang, W., & Wang, Y. (2024). *Personalized aesthetic prediction for color themes via probabilistic modeling*. *ACM Transactions on Applied Perception (TAP),* 21(2), Article 12.

Appendix II: Disclosures

Impact Statement

While primarily focused on color palettes, this framework for learning human preferences has broader implications for visual curation, potentially extending to image and video ranking. A significant ethical risk involves data bias; if the pairwise training data does not reflect diverse cultural perspectives, the model may enforce a homogenized standard of aesthetics. Furthermore, there is a societal risk of over-reliance, where algorithmic predictions stifle human agency—such as dictating personal choices—thereby reducing individual creativity.

To mitigate these risks, I position this technology as a collaborative support tool rather than a final arbiter, ensuring humans remain in the loop to correct prediction errors. Regarding compliance, I maintained strict adherence to Stanford's standards of academic integrity by utilizing exclusively anonymized data, responsibly attributing foundational methodologies, and transparently documenting the model's limitations to prevent the interpretation of predicted preferences as objective truths.

Research Statement

To ensure the integrity of this research, I employed a rigorous methodology designed to identify and mitigate plagiarism, bias, and data inaccuracies. Regarding plagiarism and the attribution of ideas, I strictly adhered to citation protocols, ensuring that both text and conceptual frameworks were credited to their original scholars. I went beyond standard literature reviews by engaging deeply with primary sources; this included viewing lectures and oral presentations by authors to verify that I fully understood their original intent and nuance before synthesizing their findings.

To address potential inaccuracies and bias, I prioritized the use of direct source data rather than relying on secondary interpretations, which reduces the risk of transmission errors. I also maintained a reflexive approach throughout the analysis, critically assessing the limitations of existing studies alongside the constraints of my own methodology. This process allowed me to distinguish between supported evidence and conjecture, ensuring that the final conclusions remain objective and accurate.

Artificial Intelligence Reflection Statement

I utilized AI tools primarily as an accelerator for literature discovery and a generator for technical implementation. While the original research trajectory was strictly self-generated, AI proved invaluable in surfacing relevant studies and drafting complex code sequences that often exceeded standard efficiency. However, this process significantly impacted my learning by reinforcing the necessity of critical oversight; because AI outputs can contain subtle logic errors or hallucinations, I found it mandatory to scrutinize every line of generated code and verify source accuracy manually.

Looking forward, I envision AI functioning as a high-velocity collaborator that requires active human direction. My experience demonstrates that while AI can dramatically increase research speed, the scholar must remain the architect of the workflow. Future use relies on the principle that we must possess a deep understanding of the subject matter to effectively guide the AI, ensuring it serves as a verified engine for execution rather than an unchecked source of truth.