Supporting Information for "Jeweler-in-the-Loop: Personalized Alloy Color Optimization via Preference-Based BO"

Chase Katz, Ting-Yu Yang, Parker King, Md Shafiqul Islam, Brent Vela, Raymundo Arróyave

Department of Materials Science and Engineering, Texas A&M University, College Station, TX 77843, USA

1. Hyperparameter Settings

The multi-objective preference model uses the BoTorch PairwiseGP implementation with its default kernel hyperparameter priors. Specifically, the kernel lengthscale follows a Gamma prior with concentration 2.4 and rate 2.7 (constrained to be $>10^{-4}$), and the kernel output scale follows a SmoothedBoxPrior bounded between 10^{-2} and 10^{2}. These defaults are designed for inputs scaled to the unit interval. In this work, alloy compositions are represented as normalized fractions in [0,1], making the default prior assumptions appropriate.

The single-objective model uses a GPyTorch RBF kernel wrapped in a ScaleKernel with positivity constraints but no explicit hyperparameter priors.

2. Optimization behavior

To provide a quantitative assessment of optimization behavior, we added a color-space evaluation based on the gold-likeness utility function. For each candidate alloy, the RGB values predicted by the Thermo-Calc optical model were converted to a scalar gold-likeness score, which measures proximity to a canonical gold color in HSV space. Figure 1 plots this score against each RGB channel value for the full dataset and highlights the alloys selected during the optimization.

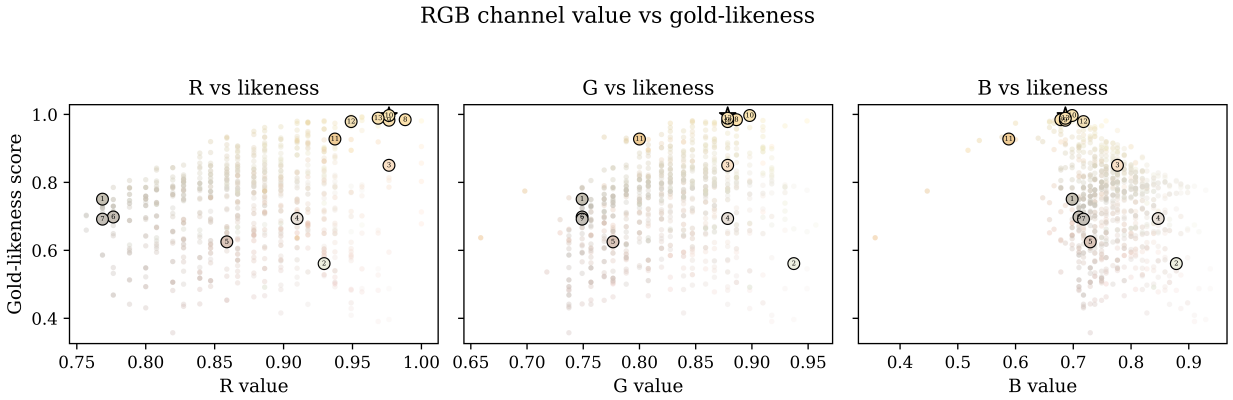

Figure 1: RGB channel value versus gold-likeness score for all candidate alloys. Points show Thermo-Calc–predicted colors mapped to the quantitative gold-likeness utility. Circled markers denote alloys sampled during optimization (labeled by iteration), and the star indicates the target gold color.

Unlike the swatch progression figure (Figure 2), which is qualitative, this plot directly reports the objective used by the simulated preference model. The sampled alloys progressively cluster in the high-utility region, demonstrating that the preference-based Bayesian optimization algorithm moves toward colors closer to the target gold appearance.

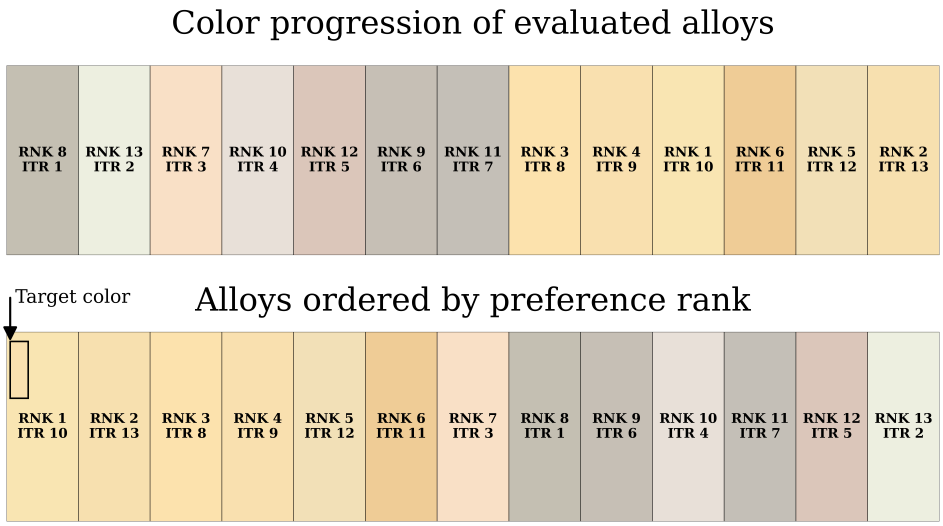

Figure 2: Progress of evaluated alloy colors along with their associated preference ranking.

3. Alternative Methods

Figure 3 shows the evolution of the best-observed utility as a function of optimization iteration. All Bayesian optimization variants rapidly improve the objective within the first few iterations, reaching near-optimal utility after only a small number of evaluations. In contrast, the random search baseline progresses gradually and exhibits larger variability between runs. The three Gaussian process kernels produce nearly indistinguishable performance, indicating that the observed acceleration arises from surrogate-guided search rather than kernel choice. The primary advantage of Bayesian optimization therefore occurs in the early stages of the campaign, where it identifies high-utility compositions substantially faster than uninformed sampling.

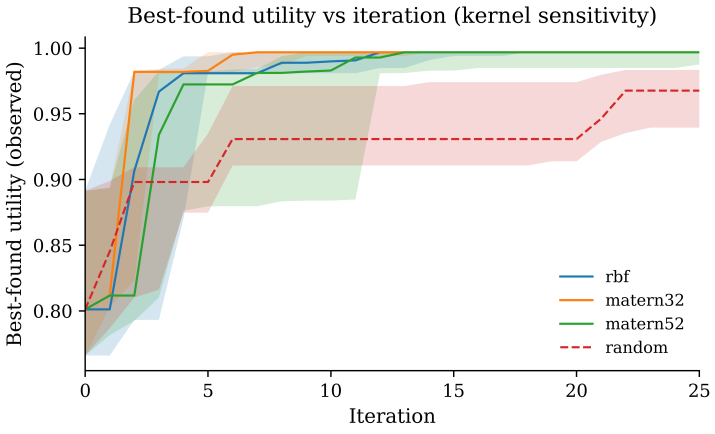

Figure 3: Best-found utility versus iteration for single-objective optimization under different Gaussian process kernels (RBF, Matérn-3/2, Matérn-5/2) and a random search baseline. Solid lines represent the median across 10 random seeds, and shaded regions indicate the interquartile range (25–75%). All runs used 25 optimization iterations and identical initialization size. The Bayesian optimization method clearly outperforms random sampling, particularly in early iterations, showing the benefit of surrogate-guided search.

4. Hyperparameter Sensitivity

The preference-based Bayesian optimization framework involves several classes of hyperparameters. To assess robustness, we performed a sensitivity analysis around the baseline configuration, varying kernel structure and acquisition optimization settings while keeping the iteration fixed at 25. Each configuration was evaluated across 10 random seeds. Performance was measured using the best utility for the single-objective study and hypervolume for the bi-objective study.

For the single-objective study, we evaluated sensitivity to the Gaussian process kernel by comparing an RBF kernel with Matérn-3/2 and Matérn-5/2 kernels. Figure 4 shows the best-found utility as a function of iteration. Solid lines represent the median across 10 random seeds, and shaded regions indicate the interquartile range (25–75%). All kernel choices show rapid convergence and achieve nearly identical final performance, indicating that optimization behavior is not strongly sensitive to kernel choices. We also evaluated sensitivity to the number of randomly initialized alloys (3, 5, and 8 initial points). As shown in Figure 5, larger initialization sizes yield earlier improvement. However, all configurations converge to similar final utility within 25 iterations, indicating that the optimization is robust to reasonable variations in initialization size.

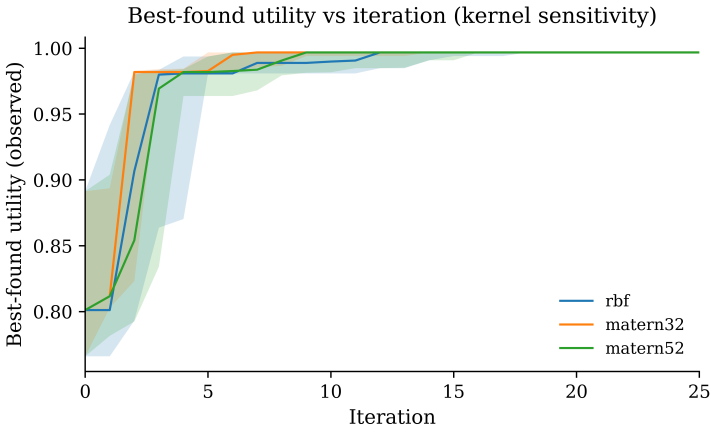

Figure 4: Best-found utility versus iteration for different Gaussian process kernels (RBF, Matérn-3/2, and Matérn-5/2). Solid lines show the median across 10 random seeds and shaded regions indicate the interquartile range (25–75%). All runs used 25 optimization iterations and identical initialization size.

In the multi-objective setting, default weakly informative priors provided by the BoTorch PairwiseGP were used. These priors are designed for inputs normalized to the range [0,1], consistent with the representation of alloy compositions as normalized fractions. We evaluated sensitivity to the kernel by comparing RBF, Matérn-3/2, and Matérn-5/2 kernels. Hypervolume was computed in a fixed objective space with a reference point chosen slightly worse than the worst normalized objective values. Figure 6 shows hypervolume as a function of iteration. Solid lines represent the median across 10 random seeds, and shaded regions indicate the interquartile range (25–75%). Similar convergence trends across kernels indicate that the multi-objective optimization behavior is not strongly dependent on the kernel choice.

Surrogate model hyperparameters, including kernel lengthscales and output scales, were not manually tuned. These quantities were learned automatically at each iteration by maximizing the Laplace-approximated marginal likelihood of the Pairwise Gaussian process. Positivity constraints were imposed on all kernel hyperparameters.

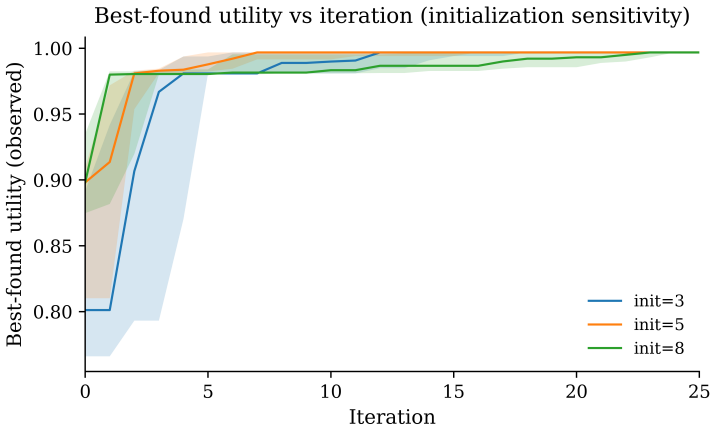

Figure 5: Best-found utility versus iteration for different numbers of randomly initialized alloys (3, 5, and 8). Solid lines show the median across 10 random seeds and shaded regions indicate the interquartile range (25–75%). All runs used the RBF kernel and 25 optimization iterations.

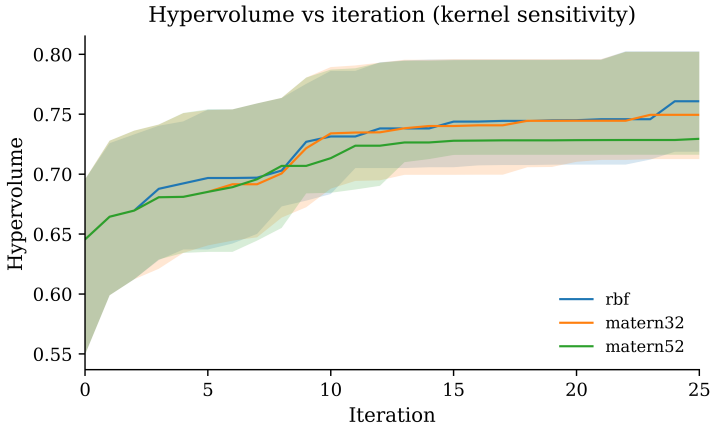

Figure 6: Hypervolume versus iteration for multi-objective optimization under different color-surrogate kernels (RBF, Matérn-3/2, Matérn-5/2). Solid lines show the median across 10 random seeds and shaded regions indicate the interquartile range (25–75%). All runs used 25 iterations and identical initialization size.