



Figure 1: This figure illustrates the model performance under the within-model comparison setting. Specifically, each test objective function corresponds to a GP posterior conditioning on 900 random samples from the objectives provided in the Toy Function, which is detailed in the main paper and Appendix D. For the GP model, we adopted a setting similar to the within-model comparison constructed in Hening & Schuler (2012), where the variance of random noise is set to 0.005^2 , and length scale is 1. The results were averaged over 40 groups of test functions.