Benchmarking the Performance of Bayesian Optimization across Multiple Experimental Materials Science Domains

Qiaohao Liang[1,***], **Aldair E. Gongora**[2], **Zekun Ren**[3], **Armi Tiihonen**[1,*], **Zhe Liu**[1,†], **Shijing Sun**[1], **James R. Deneault**[4], **Daniil Bash**[5], **Flore Mekki-Berrada**[6], **Saif A. Khan**[6], **Kedar Hippalgaonkar**[5], **Benji Maruyama**[4], **Keith A. Brown**[2], **John Fisher III**[1], and **Tonio Buonassisi**[1,***]

[1]Massachusetts Institute of Technology, Cambridge, MA, United States
[2]Boston University, Boston, MA, United States
[3]Singapore-MIT Alliance for Research and Technology, Singapore
[4]Air Force Research Laboratory, United States
[5]Agency for Science, Technology and Research (A*STAR), Singapore
[6]National University of Singapore, Singapore
[***]Corresponding author: Qiaohao Liang, hqliang@mit.edu and Tonio Buonassisi, buonassisi@mit.edu

October 9, 2021

Supplementary Notes

This section provides supplementary tables and figures for the study "Benchmarking the Performance of Bayesian Optimization across Multiple Experimental Materials Science Domains."

Supplementary Table 1 - 5 describe the design space of the five experimental datasets in more detail.

Supplementary Figure 1 shows the normalized RMSE values of RF models with different number of decision trees n_{tree} during cross validation across five datasets.

Supplementary Figure 2 shows cross validation results of neural network, GP ARD, and RF ground truth models on Crossed barrel dataset and comparison of the effectiveness of BO algorithms when applied to each of the ground truth models. Supplementary Figure 3 shows cross validation results of neural network, GP ARD, and RF ground truth models on AgNP dataset and comparison of the effectiveness of BO algorithms when applied to each of the ground truth models. All ground truth models have noise $e(\mathbf{x}) = \mathcal{N}(0, 0.01\mu)$ at each point $\mathbf{x}$ in its design space, where μ is the mean of the dataset's objective values. Neural network and RF model represent ground truth models that do not carry strong gaussianity assumption like GP. When ground truth model is GP type, we observe that the relative optimization performance of BO algorithms with GP type surrogate models are improved over those with RF type surrogate models. GP type surrogate models' advantage over RF type ones is reduced when ground truth model is NN or RF, which are free of distributional assumptions. This confirms our concern that specific ground truth models could introduce extra bias into design space that could impact the benchmarking results. Nevertheless, the results in Supplementary Figure 2 - 3 match those in Figure 3 - 4 from manuscript in relative ranking between the three surrogate models. GP with anisotropic kernels (GP ARD) is shown to be better surrogate model across all ground truth models, proving to be a robust model for future optimization campaigns. RF is a close second shown to be comparable to GP ARD at times, and both outclass GP with isotropic kernels (GP).

Supplementary Figure 4 shows the absolute EF_{max} values shown by BO algorithms when guiding materials optimization campaigns across five datasets. It can be viewed together with Figure 4 in the manuscript, where EF_{max} values are normalized for comparison purposes.

[*] This author is now at Aalto University, Espoo, Finland

[†] This author is now at Northwestern Polytechincal University (NPU), Xi'an, Shaanxi, P.R. China

Supplementary Figure 5 - 9 show the performance of BO algorithms across five datasets. Compared to Figure 3 in manuscript, they are more comprehensive and include more surrogate models, acquisition functions, a range of $\overline{\lambda}$ values for LCB and more kernels for GPs.

Supplementary Table 1: Crossed barrel dataset input feature space. The dataset has size 600. It consists of design parameters for the crossed barrel structure. The crossed barrel structures were optimized for max toughness.

Parameter	Kind	Range	Description
n	Discrete	[6, 12] with interval of 2	number of struts
θ	Discrete	[0, 200] with interval of 25	twisting angle [°]
r	Discrete	[1.5, 2.5] with interval of 0.1	thickness [mm]
t	Discrete	[0.7, 1.4] with interval of 0.35	outer radius [mm]

Supplementary Table 2: AgNP dataset input features space. The dataset has size 164. It consists of processing parameters for synthesizing triangular nanoprisms. The synthesized silver nanoparticles were optimized for shape and correspondingly absorbance spectrum. Q_i is the ratio between flow rate of reactant i to total aqueous flow rate.

Parameter	Kind	Range	Description
Q_{seed}	Discrete	[0.5:80] with interval of 5	flow rate ratio of Ag seeds [%]
Q_{AgNO_3}	Discrete	[0.5:80] with interval of 5	flow rate ratio of silver nitrate [%]
Q_{TSC}	Discrete	[0.5:80] with interval of 5	flow rate ratio of trisodium citrate [%]
Q_{PVA}	Discrete	[10:40] with interval of 5	flow rate ratio of polyvinyl alcohol [%]
Q_{total}	Discrete	[200:1000] with interval of 100	total flow rate [μL/min]

Supplementary Table 3: P3HT/CNT dataset input features space. The dataset has size 178. It consists of composition parameters for carbon nanotube polymer blend. The composite blends were optimized for electrical conductivity [S/cm]. A constraint on the parameter space is P3HT + D_1 + D_2 + D_6 + D_8 = 1.

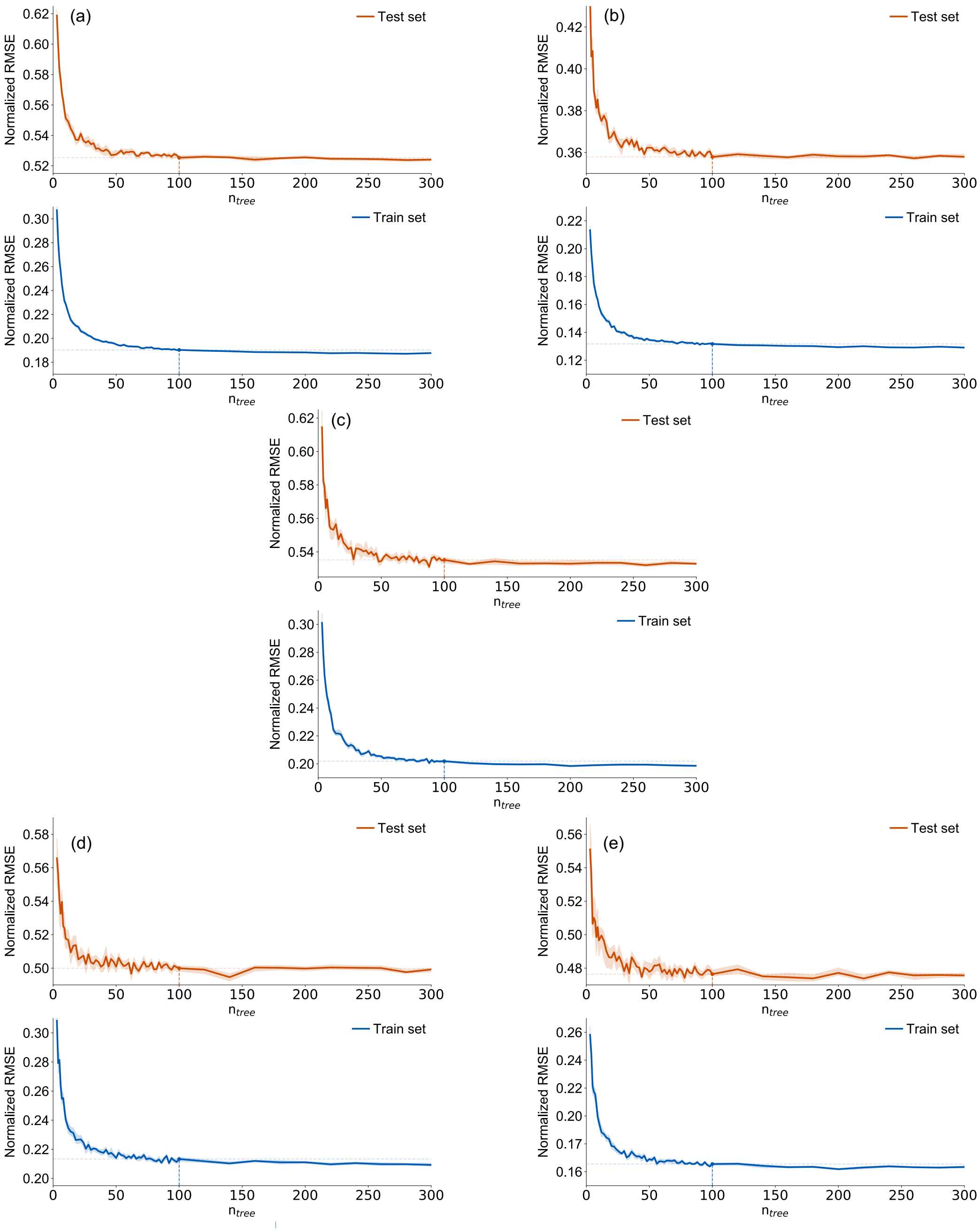
Parameter	Kind	Range	Description
P3HT	Continuous	[0, 100]	composition ratio of P3HT [%]
D_1	Continuous	[0, 100]	composition ratio of D1 CNT sample [%]
D_2	Continuous	[0, 100]	composition ratio of D2 CNT sample [%]
D_6	Continuous	[0, 100]	composition ratio of D6 CNT sample [%]
D_8	Continuous	[0, 100]	composition ratio of D8 CNT sample [%]

Supplementary Table 4: Perovskite dataset input features space. The dataset has size 94. It consists of composition parameters for halide perovskites $Cs_xMA_yFA_{1-x-y}PbI_3$ thin films. The perovskite films were optimized for environmental stability. A constraint on the parameter space is CsPbI + FAPbI + MAPbI = 1.

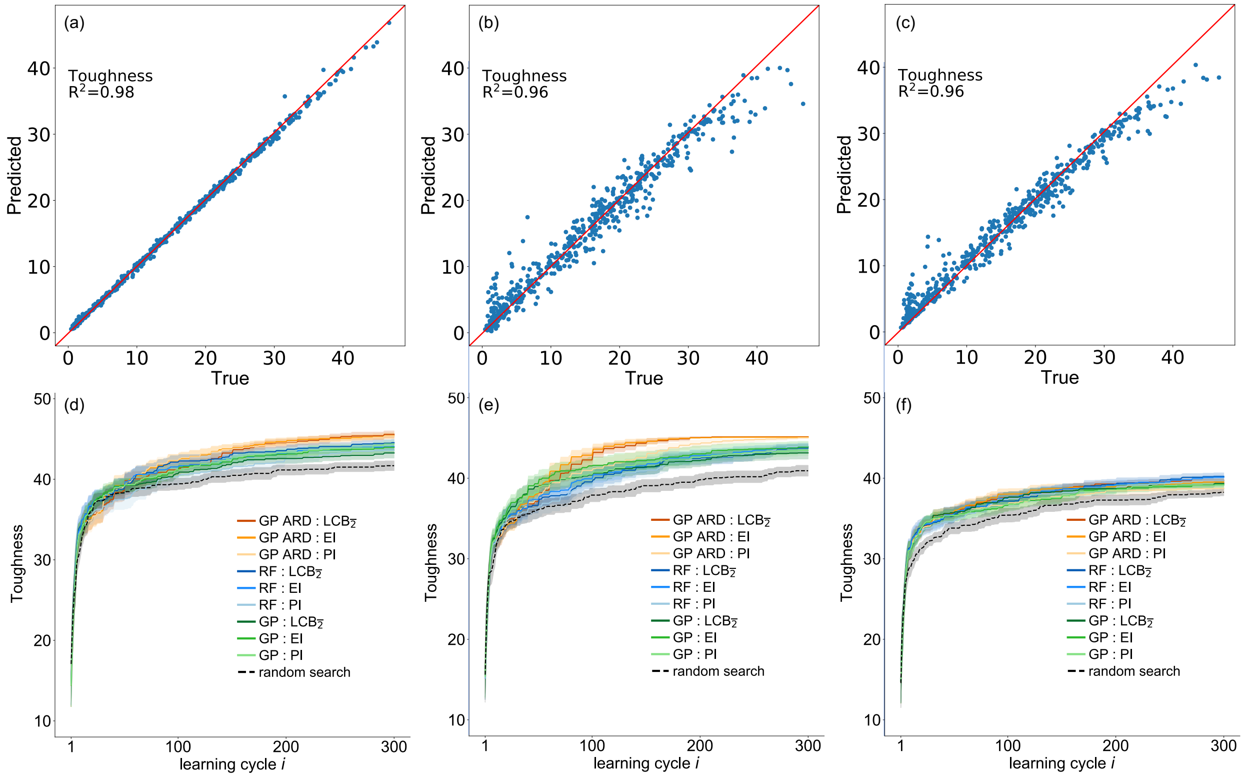
Parameter	Kind	Range	Description
CsPbI	Discrete	[0, 100] with interval of 1	composition ratio of CsPbI [%]
FAPbI	Discrete	[0, 100] with interval of 1	composition ratio of FAPbI [%]
MAPbI	Discrete	[0, 100] with interval of 1	composition ratio of MAPbI [%]

Supplementary Table 5: AutoAM dataset input features space. The dataset has size 100. It consists of additive manufacturing control parameters for printing specific shapes. The additive manufacturing system was optimized for printing shapes with best shape score..

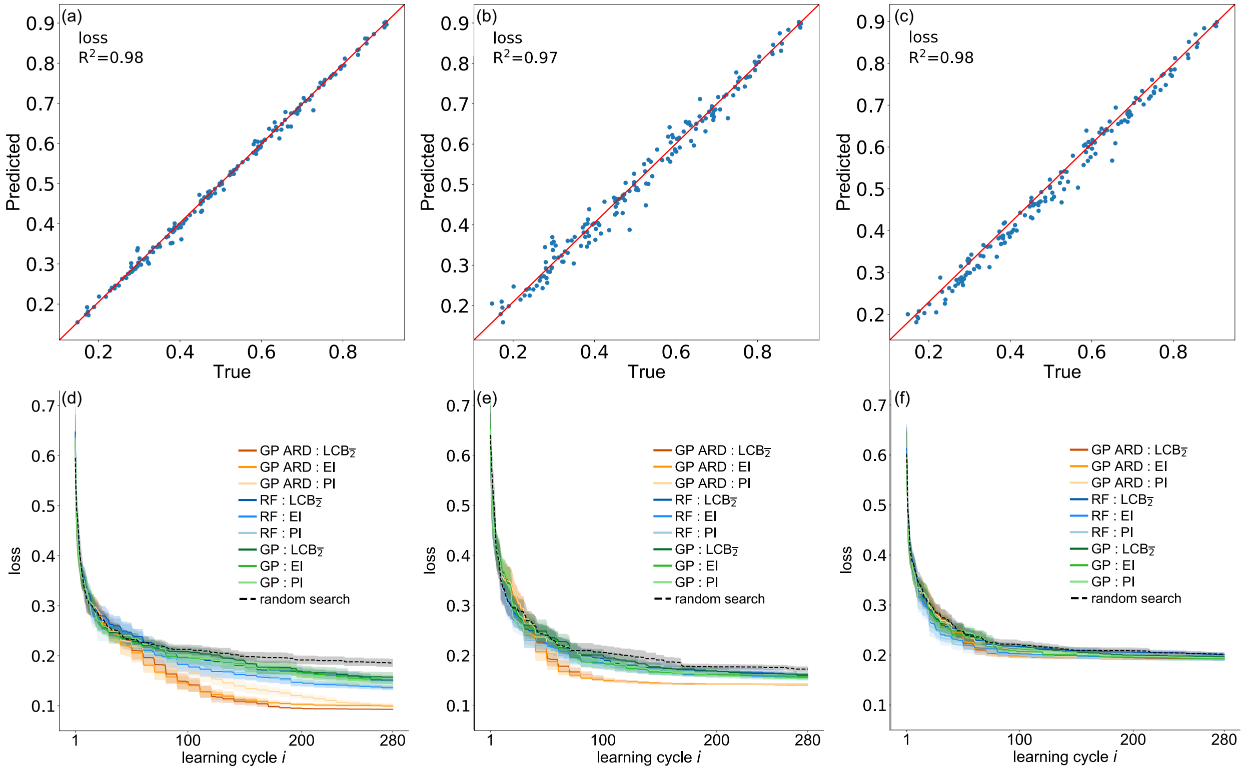
Parameter	Kind	Range	Description
Prime Delay	Continuous	[0, 5]	delay in extrusion before deposition [s]
Print Speed	Continuous	[0.1, 10]	speed of printing movement [mm/s]
X Offset Correction	Continuous	[-1, 1]	printing head offset in X direction [mm]
Y Offset Correction	Continuous	[-1, 1]	printing head offset in Y direction [mm]



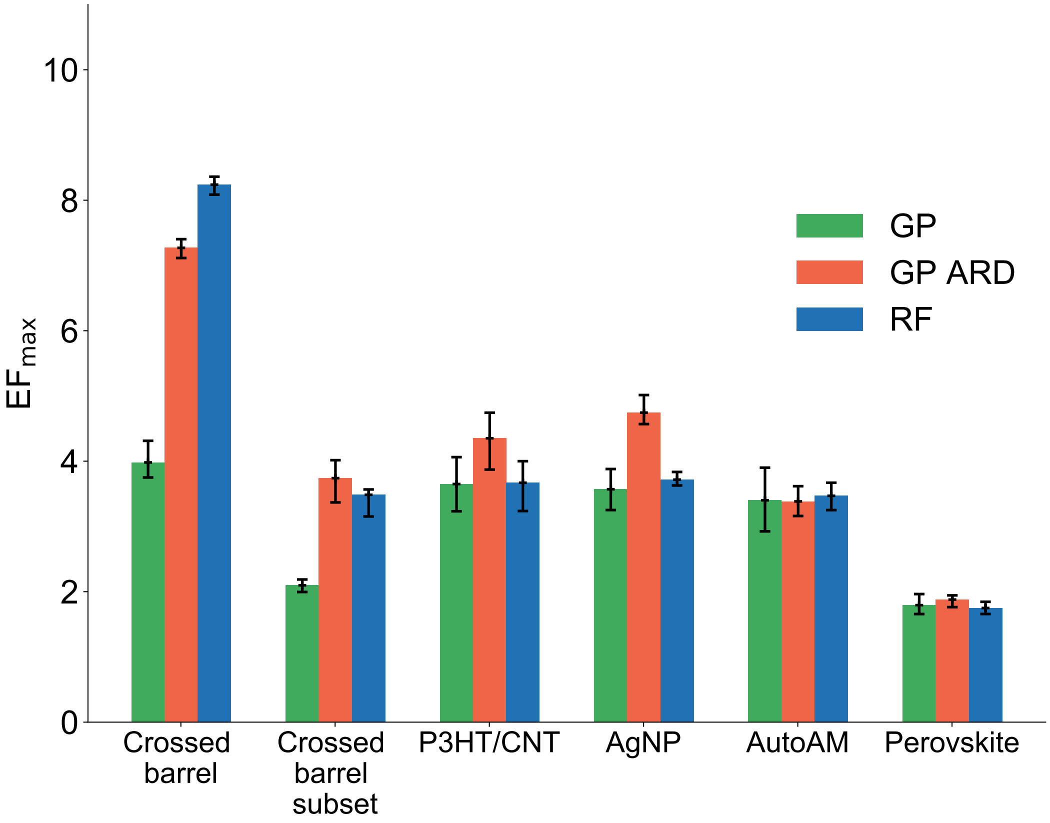
Supplementary Figure 1: The normalized RMSE values of RF models with different number of decision trees n_{tree} during cross validation. **(a)(b)(c)(d)(e)** are results for Crossed barrel, AgNP, P3HT, Perovskite, and AutoAM dataset respectively. The RF models had n_{tree} between 5 to 300. For each model, 50 different random test and train split were conducted, followed by 5-fold cross validation and evaluation of model on the respective training set and test set. Variation of normalized RMSE at each n_{tree} is visualized by plotting the median as well as shaded regions representing the 5th to 95th percentile of the 50 different evaluations. The figure shows that RF with $n_{\text{tree}} = 100$ is a suitable initial hyperparameter for RF surrogate models because it not only achieves similar prediction accuracy as RF models with larger n_{tree} but also avoids the risk of overfitting from larger n_{tree}.



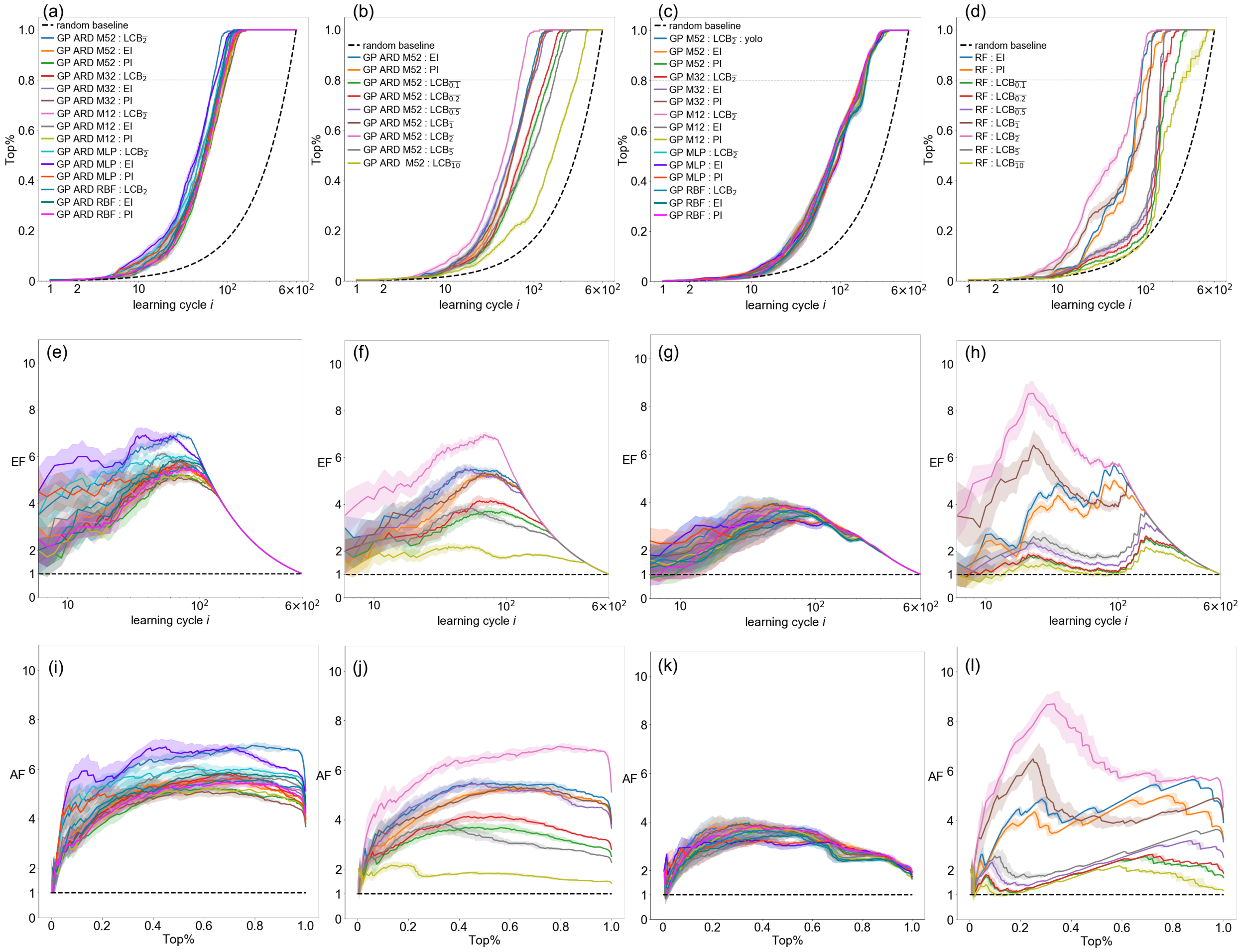
Supplementary Figure 2: The Leave-one-out cross validation results shown in parity plots for neural network, GP, and RF models fitted on the Crossed barrel dataset and the performance of different BO algorithms on these emulated design spaces. In **(a)(d)**, the neural network model has structure (4, 120, 240, 480, 960, 480, 240, 120, 1). It is fully connected between layers and has activation function Leaky ReLU. In **(b)(e)**, the GP model is anisotropic and uses Mátern52 kernel. In **(c)(f)**, the RF model has an ensemble of 500 decision trees and is trained with bootstrapping. BO algorithms are trained for 30 times as they start from different initial experiment locations. Variation at each learning cycle i is visualized by plotting the median as well as shaded regions representing the 5th to 95th percentile of the aggregated 30-run ensembles. For consistency and comparison purposes, in each of the BO campaigns, both the randomly obtained initial experiments and subsequent batches have the size of 10.



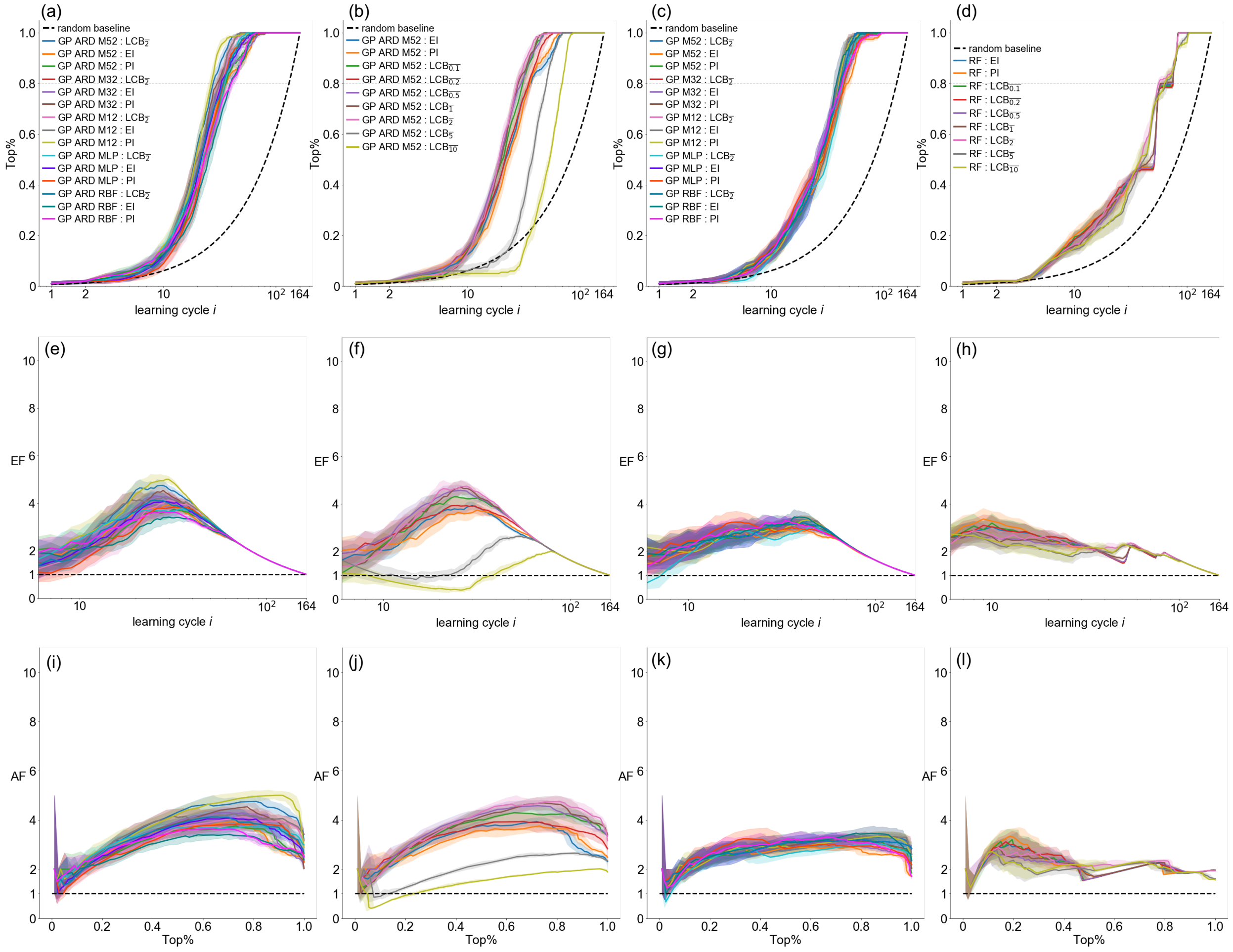
Supplementary Figure 3: The Leave-one-out cross validation results shown in parity plots for neural network, GP, and RF models fitted on the AgNP dataset and the performance of different BO algorithms on these emulated design spaces. In **(a)(d)**, the neural network model has structure (5, 60, 120, 240, 480, 240, 120, 60, 1). It is fully connected between layers and has activation function Leaky ReLU. In **(b)(e)**, the GP model is anisotropic and uses Mátern52 kernel. In **(c)(f)**, the RF model has an ensemble of 500 decision trees and is trained with bootstrapping. BO algorithms are trained for 30 times as they start from different initial experiment locations. Variation at each learning cycle i is visualized by plotting the median as well as shaded regions representing the 5th to 95th percentile of the aggregated 30-run ensembles. For consistency and comparison purposes, in each of the BO campaigns, both the randomly obtained initial experiments and subsequent batches have the size of 10.



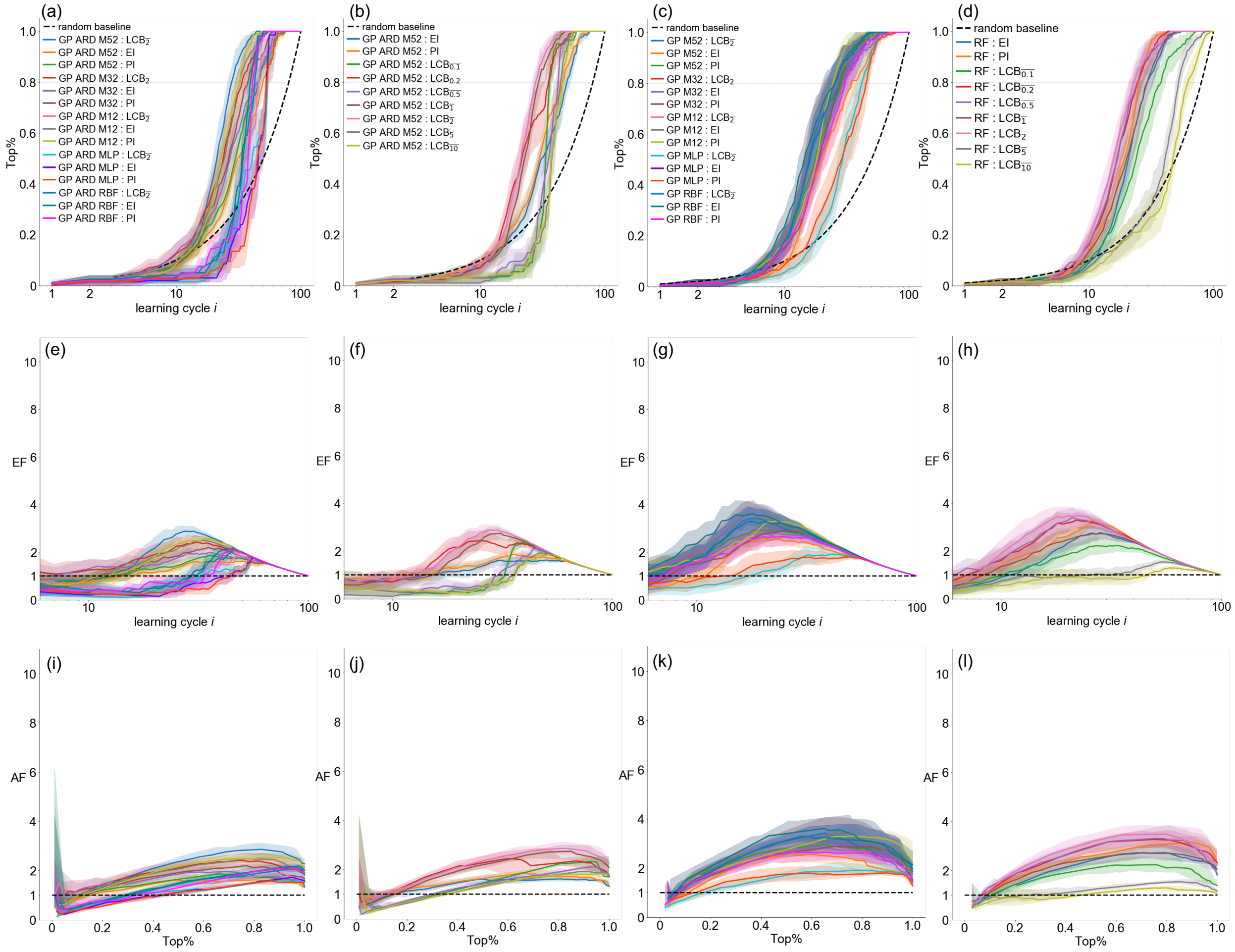
Supplementary Figure 4: Absolute EF$_{max}$ demonstrated by BO algorithms equipping GP without ARD, GP with ARD, and RF as surrogate models and all using LCB$_{\overline{2}}$ as acquisition function. For each algorithm applied across datasets, the median of EF$_{max}$ is shown by the barplots, and its 5th to 95th percentile are are shown by respective floating bars. The crossed barrel subset is collected by running BO algorithm GP : EI until all candidates with top 5% toughness are found, representing an "easier" path towards optimums.



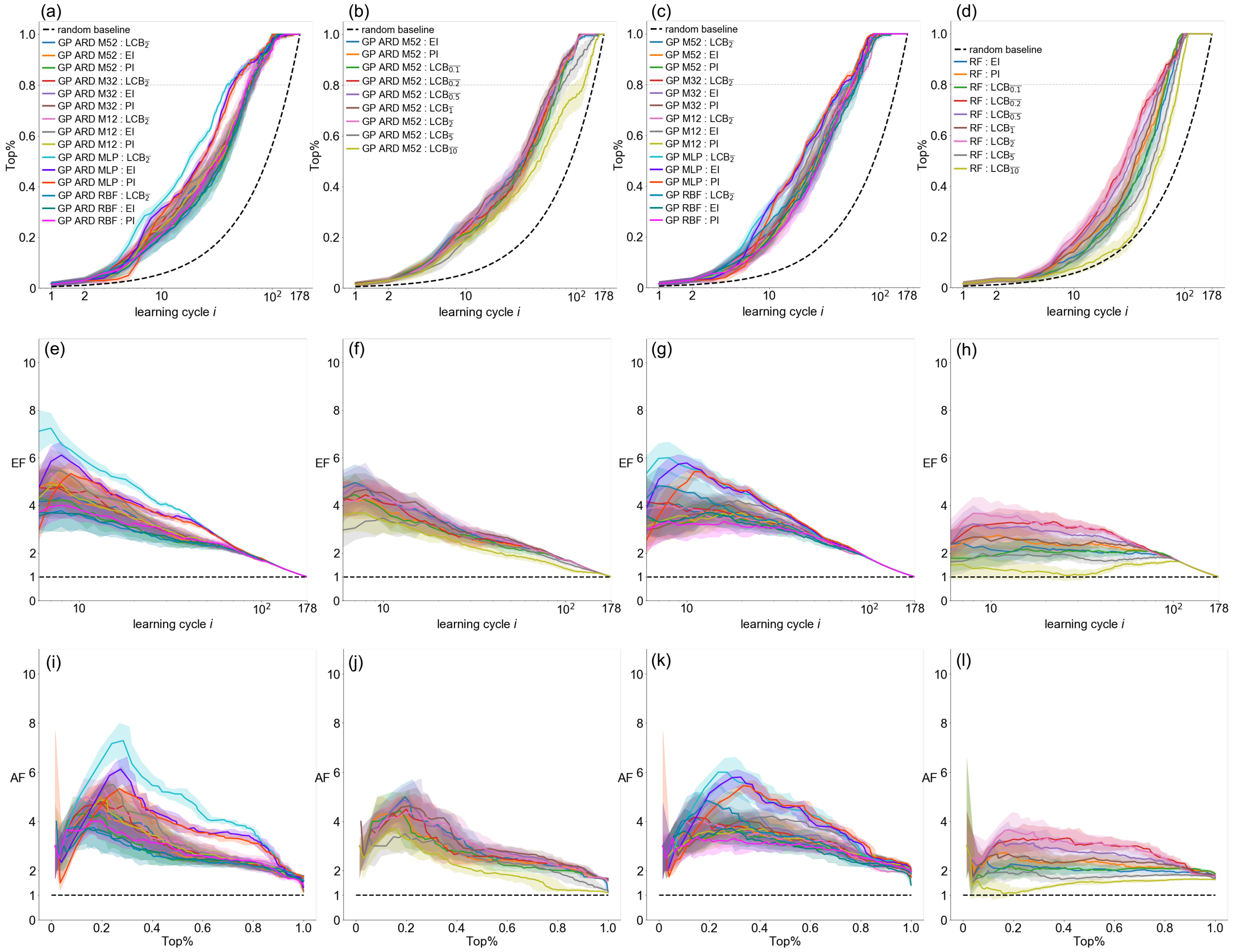
Supplementary Figure 5: The aggregated performance of BO algorithms on the Crossed barrel dataset. The comprehensive benchmark involves multiple surrogate models, kernel type for GP, and acquisition functions. The performance of BO algorithms with GP ARD surrogate model and various kernels and acquisition functions can be observed in **(a)(e)(i)** and **(b)(f)(j)**. The performance of BO algorithms with GP surrogate model and various kernels and acquisition functions can be observed in **(c)(g)(k)**. The performance of BO algorithms with RF surrogate model and various kernels and acquisition functions can be observed in **(d)(h)(l)**. Variation at each learning cycle is visualized by plotting the median as well as shaded regions representing the 5th to 95th percentile of the aggregated 50-run ensembles.



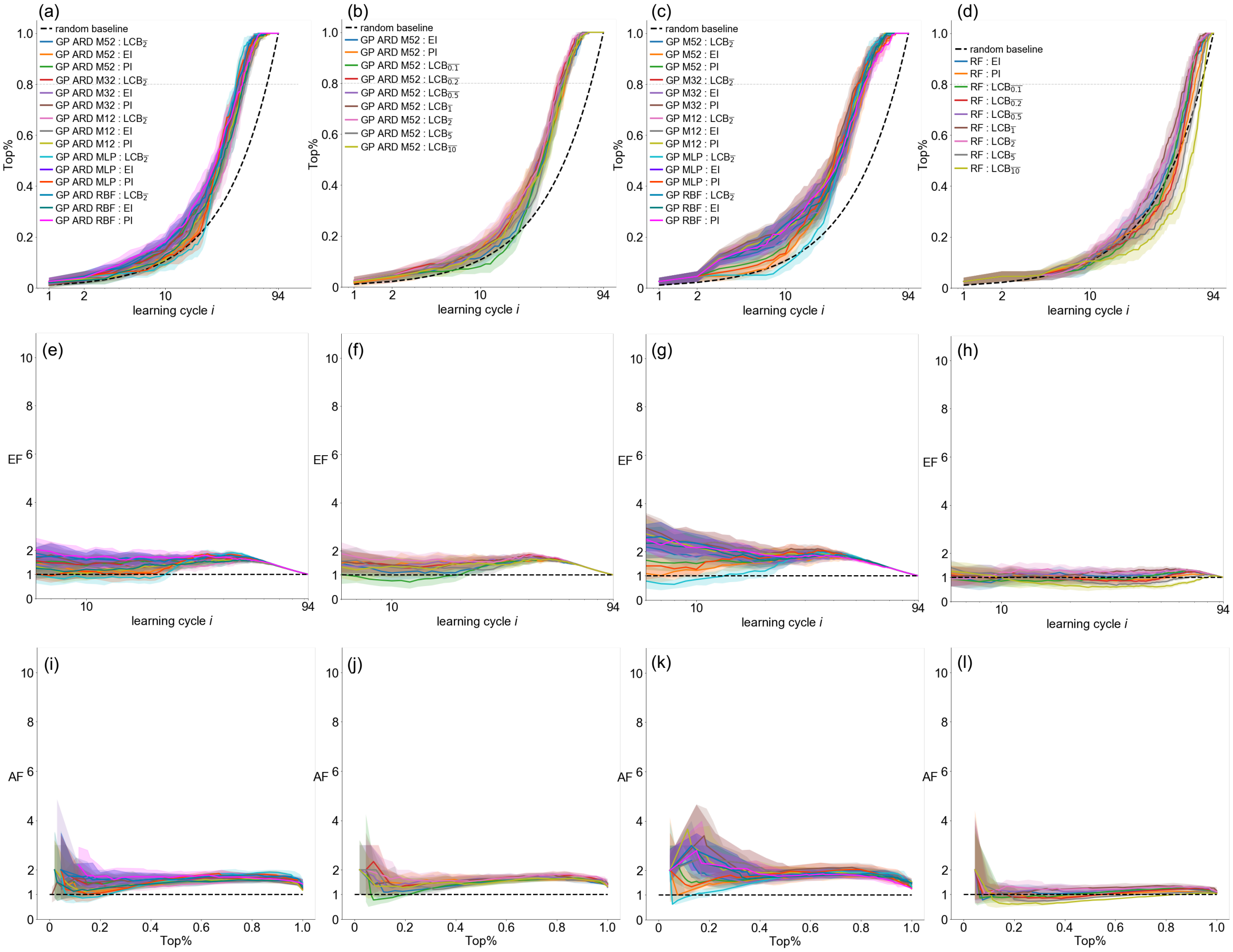
Supplementary Figure 6: The aggregated performance of BO algorithms on the AgNP dataset. The comprehensive benchmark involves multiple surrogate models, kernel type for GP, and acquisition functions. The performance of BO algorithms with GP ARD surrogate model and various kernels and acquisition functions can be observed in **(a)(e)(i)** and **(b)(f)(j)**. The performance of BO algorithms with GP surrogate model and various kernels and acquisition functions can be observed in **(c)(g)(k)**. The performance of BO algorithms with RF surrogate model and various kernels and acquisition functions can be observed in **(d)(h)(l)**. Variation at each learning cycle is visualized by plotting the median as well as shaded regions representing the 5^{th} to 95^{th} percentile of the aggregated 50-run ensembles.



Supplementary Figure 7: The aggregated performance of BO algorithms on the AutoAM dataset. The comprehensive benchmark involves multiple surrogate models, kernel type for GP, and acquisition functions. The performance of BO algorithms with GP ARD surrogate model and various kernels and acquisition functions can be observed in **(a)(e)(i)** and **(b)(f)(j)**. The performance of BO algorithms with GP surrogate model and various kernels and acquisition functions can be observed in **(c)(g)(k)**. The performance of BO algorithms with RF surrogate model and various kernels and acquisition functions can be observed in **(d)(h)(l)**. Variation at each learning cycle is visualized by plotting the median as well as shaded regions representing the 5th to 95th percentile of the aggregated 50-run ensembles.



Supplementary Figure 8: The aggregated performance of BO algorithms on the P3HT/CNT dataset. The comprehensive benchmark involves multiple surrogate models, kernel type for GP, and acquisition functions. The performance of BO algorithms with GP ARD surrogate model and various kernels and acquisition functions can be observed in **(a)(e)(i)** and **(b)(f)(j)**. The performance of BO algorithms with GP surrogate model and various kernels and acquisition functions can be observed in **(c)(g)(k)**. The performance of BO algorithms with RF surrogate model and various kernels and acquisition functions can be observed in **(d)(h)(l)**. Variation at each learning cycle is visualized by plotting the median as well as shaded regions representing the 5th to 95th percentile of the aggregated 50-run ensembles.



Supplementary Figure 9: The aggregated performance of BO algorithms on the Perovskite dataset. The comprehensive benchmark involves multiple surrogate models, kernel type for GP, and acquisition functions. The performance of BO algorithms with GP ARD surrogate model and various kernels and acquisition functions can be observed in **(a)(e)(i)** and **(b)(f)(j)**. The performance of BO algorithms with GP surrogate model and various kernels and acquisition functions can be observed in **(c)(g)(k)**. The performance of BO algorithms with RF surrogate model and various kernels and acquisition functions can be observed in **(d)(h)(l)**. Variation at each learning cycle is visualized by plotting the median as well as shaded regions representing the 5th to 95th percentile of the aggregated 50-run ensembles.