A Appendix

A.1 Considerations for Sampling Around the Offline Dataset

In this subsection, we explore an alternative sampling strategy for the pseudo-labeling process. Instead of generating new samples around the current optimization point, this strategy generates samples directly around the offline dataset $\mathcal{D}$. To ascertain the effectiveness of our chosen strategy against this alternative, we perform experiments on two tasks: D'Kitty (continuous) and TF8 (discrete).

Table 4 showcases the results. For both tasks, our strategy consistently yields higher scores, affirming its superior performance over the alternative. The advantage of our chosen strategy can be attributed to its dynamic nature. By sampling around the current optimization point, we gather more insightful information for the local fine-tuning of the proxy. This strategy allows the co-teaching process to adapt and evolve according to the optimization trajectory, leading to improved performances.

Table 4: Comparison of Sampling Strategies.

Method	Sampling along Gradient Path **(Ours)**	Sampling from $\mathcal{D}$
TF8	$\mathbf{0.958 \pm 0.008}$	0.871 ± 0.067
D'Kitty	$\mathbf{0.968 \pm 0.020}$	0.955 ± 0.006

A.2 Comparative Performance Analysis using Median Scores

In addition to the maximum scores discussed in the main paper, we also present the median (50^{th} percentile) scores across all seven tasks. The best design in the offline dataset, denoted as $\mathcal{D}$(**best**), along with the mean and median rankings are provided for comprehensive comparison.

Performance in Continuous Tasks. Table 5 illustrates the performances of ICT compared with other methods in continuous tasks. It is noteworthy that ICT exhibits performance on par with the best-performing methods. Compared with the vanilla gradient ascent (Grad), ICT demonstrates superior performance, thus affirming its effectiveness in addressing out-of-distribution issues. Moreover, ICT is generally better than the mean ensemble (Mean), which demonstrates the effectiveness of our strategy. These results support the use of ICT as a robust baseline for offline MBO.

Table 5: Experimental results on continuous tasks for comparison (median).

Method	Superconductor	Ant Morphology	D'Kitty Morphology	Hopper Controller
$\mathcal{D}$(**best**)	0.399	0.565	0.884	1.0
BO-qEI	0.300 ± 0.015	0.567 ± 0.000	$\mathbf{0.883 \pm 0.000}$	0.343 ± 0.010
CMA-ES	0.379 ± 0.003	-0.045 ± 0.004	0.684 ± 0.016	-0.033 ± 0.005
REINFORCE	$\mathbf{0.463 \pm 0.016}$	0.138 ± 0.032	0.356 ± 0.131	-0.064 ± 0.003
CbAS	0.111 ± 0.017	0.384 ± 0.016	0.753 ± 0.008	0.015 ± 0.002
Auto.CbAS	0.131 ± 0.010	0.364 ± 0.014	0.736 ± 0.025	0.019 ± 0.008
MIN	0.336 ± 0.016	$\mathbf{0.618 \pm 0.040}$	$\mathbf{0.887 \pm 0.004}$	0.352 ± 0.058
Grad	0.321 ± 0.010	0.559 ± 0.032	0.856 ± 0.009	0.354 ± 0.010
Mean	0.334 ± 0.003	0.569 ± 0.010	0.876 ± 0.003	0.386 ± 0.003
Min	0.354 ± 0.026	0.571 ± 0.011	$\mathbf{0.883 \pm 0.000}$	0.359 ± 0.004
COMs	0.316 ± 0.026	0.560 ± 0.002	0.879 ± 0.002	0.341 ± 0.009
ROMA	0.372 ± 0.019	0.479 ± 0.041	0.853 ± 0.007	0.389 ± 0.005
NEMO	0.318 ± 0.008	$\mathbf{0.592 \pm 0.000}$	0.880 ± 0.000	0.355 ± 0.002
BDI	0.412 ± 0.000	0.474 ± 0.000	0.855 ± 0.000	$\mathbf{0.408 \pm 0.000}$
IOM	0.352 ± 0.021	0.509 ± 0.033	0.876 ± 0.006	0.370 ± 0.009
ICT$_{\text{(ours)}}$	0.399 ± 0.012	$\mathbf{0.592 \pm 0.025}$	0.874 ± 0.005	0.362 ± 0.004

Performance in Discrete Tasks. The median scores for discrete tasks are reported in Table 6. ICT consistently demonstrates high performance for both TF Bind 8 and TF Bind 10. However, for the NAS task, which has a higher dimensionality than the two tasks, the optimization process becomes notably more complex. Further, the simplistic encoding-decoding strategy employed in the design bench may not accurately capture the intricacies of the neural network's accuracy, potentially contributing to ICT's suboptimal performance on the NAS task.

Table 6: Experimental results on discrete tasks & ranking on all tasks for comparison (median).

Method	TF Bind 8	TF Bind 10	NAS	Rank Mean	Rank Median
$\mathcal{D}$(**best**)	0.439	0.467	0.436		
BO-qEI	0.439 ± 0.000	0.467 ± 0.000	0.544 ± 0.099	$7.7/15$	$8/15$
CMA-ES	0.537 ± 0.014	0.484 ± 0.014	$\mathbf{0.591 \pm 0.102}$	$8.4/15$	$6/15$
REINFORCE	0.462 ± 0.021	0.475 ± 0.008	-1.895 ± 0.000	$10.9/15$	$14/15$
CbAS	0.428 ± 0.010	0.463 ± 0.007	0.292 ± 0.027	$12.9/15$	$13/15$
Auto.CbAS	0.419 ± 0.007	0.461 ± 0.007	0.217 ± 0.005	$13.4/15$	$13/15$
MIN	0.421 ± 0.015	0.468 ± 0.006	0.433 ± 0.000	$7.7/15$	$9/15$
Grad	0.528 ± 0.021	0.519 ± 0.017	0.438 ± 0.110	$7.7/15$	$8/15$
Mean	0.539 ± 0.030	$\mathbf{0.539 \pm 0.010}$	0.494 ± 0.077	$5.3/15$	$5/15$
Min	$\mathbf{0.569 \pm 0.050}$	0.485 ± 0.021	$\mathbf{0.567 \pm 0.006}$	**3.7/15**	$4/15$
COMs	0.439 ± 0.000	0.467 ± 0.002	0.525 ± 0.003	$8.4/15$	$8/15$
ROMA	$\mathbf{0.555 \pm 0.020}$	0.512 ± 0.020	0.525 ± 0.003	$5.6/15$	$5/15$
NEMO	0.438 ± 0.001	0.454 ± 0.001	$\mathbf{0.564 \pm 0.016}$	$7.7/15$	$7/15$
BDI	0.439 ± 0.000	0.476 ± 0.000	0.517 ± 0.000	$6.7/15$	$8/15$
IOM	0.439 ± 0.000	0.477 ± 0.010	-0.050 ± 0.011	$7.9/15$	$7/15$
ICT$_{\text{(ours)}}$	$\mathbf{0.551 \pm 0.013}$	$\mathbf{0.541 \pm 0.004}$	0.494 ± 0.091	$4.3/15$	**3/15**

Summary. ICT excels by achieving the best median ranking and a top-two mean ranking. These rankings consolidate ICT's standing as a strong method for both continuous and discrete tasks.

A.3 Hyperparameter Setting

We report the details of hyperparameters in our experiments. The number of iterations, T, is set to 200 for continuous tasks and 100 for discrete tasks. For most continuous and discrete tasks, we employ the Adam optimizer [32] to fine-tune the proxies. The learning rates are set at $1e-3$ and $1e-1$ for continuous tasks and discrete tasks, respectively. In the case of the Hopper Controller task, the input dimension is significantly larger, at 5126, and we adopt a smaller learning rate $1e-4$ for fine-tuning to ensure stability of the optimization process. Regarding the learning rate for the meta-learning framework, we use the Adam optimizer [32] with a learning rate $2e-1$ for continuous tasks and $3e-1$ for discrete tasks, respectively.

A.4 Analysis of Co-teaching and Sample Reweighting Efficacy

In our analysis, we focus on two key steps of our method: (1) pseudo-label-driven co-teaching and (2) meta-learning-based sample reweighting. We evaluate the efficacy of these steps by comparing generated samples with their corresponding ground truth. It's important to note that during the training phase, ground-truth scores are inaccessible to all algorithms and are used strictly for evaluation. Our method incorporates three proxies $f_{\boldsymbol{\theta}_1}(\cdot)$, $f_{\boldsymbol{\theta}_2}(\cdot)$, and $f_{\boldsymbol{\theta}_3}(\cdot)$. We employ $f_{\boldsymbol{\theta}_1}(\cdot)$ for pseudo-labeling and $f_{\boldsymbol{\theta}_2}(\cdot)$, $f_{\boldsymbol{\theta}_3}(\cdot)$ for co-teaching. We run ICT over 50 time steps for both D'Kitty (continuous) and TF8 (discrete) tasks.

Pseudo-label-driven co-teaching. The step involves selecting 64 samples with smaller losses for fine-tuning the proxies while ignoring the remaining 64 samples. To assess the effectiveness of this strategy, we calculate $\mathcal{L}^{Sel}$, the mean squared error (MSE) between the pseudo-labeled and ground truth scores of the selected 64 samples, and $\mathcal{L}^{Ign}$, the MSE for the ignored samples. These calculations are averaged over 50 steps. We find that for D'Kitty, $\mathcal{L}^{Sel}$ is 0.124 lower than $\mathcal{L}^{Ign}$ and for TF8, it's 0.006 less than $\mathcal{L}^{Ign}$. These results validate the efficacy of this step, as the selected samples more closely align with the ground truth.

Meta-learning-based sample reweighting. In this step, we aim to assign larger weights to cleaner samples and smaller weights to noisier ones among the total of 64 samples. We measure the efficacy of this step by calculating $\mathcal{L}^{Large}$, the MSE between the pseudo-labeled and ground-truth scores of the 32 samples with larger weights, and $\mathcal{L}^{Small}$, the MSE for the 32 samples with smaller weights. These calculations are averaged over 50 steps. We observe that for D'Kitty, $\mathcal{L}^{Large}$ is 0.010 lower than $\mathcal{L}^{Ign}$. For TF8, $\mathcal{L}^{Large}$ is 0.005 less than $\mathcal{L}^{Small}$. These findings indicate that the samples with larger weights are indeed closer to the ground truth, substantiating the effectiveness of this step.

A.5 Examining Hyperparameter Sensitivity Further

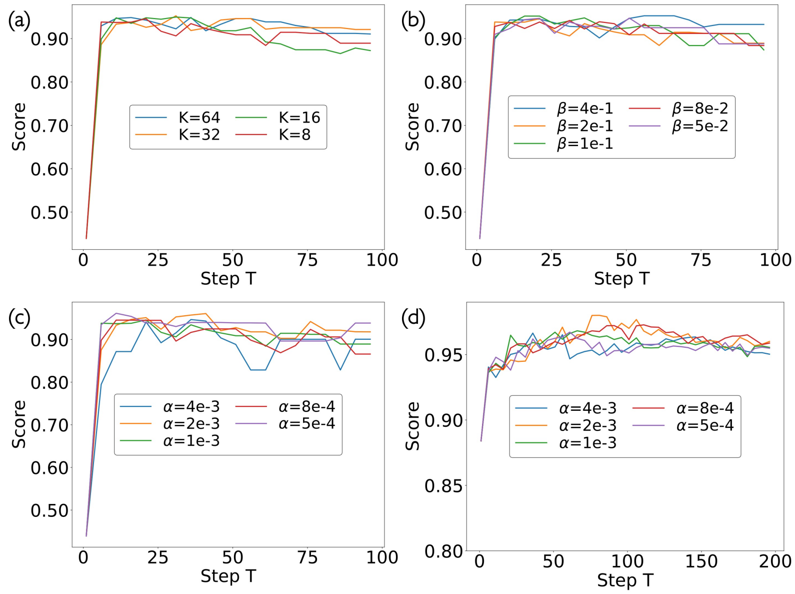

Figure 5: Extended Analysis on Hyperparameter Sensitivity.

Building on the analysis from Sec 4.6, we delve deeper into hyperparameter sensitivity, focusing on the TF8 task. Specifically, we investigate the influence of the number of selected samples (K) in the first step, and the learning rate (β) in the second step.

- Figure 5 (a) displays the 100^{th} percentile normalized ground-truth score as a function of the time step T for different K values (8, 16, 32, 64). ICT demonstrates stability over a specific range for varying K values, showcasing its robustness. Notably, ICT reaches optimal designs around $t = 20$ and maintains this level, further validating its resilience against different optimization steps T.
- Figure 5 (b) plots the 100^{th} percentile normalized ground-truth score as a function of the learning rate (β) in TF8. ICT maintains a consistent performance across diverse β values, corroborating its robustness concerning the hyperparameter β in TF8.

Furthermore, we evaluate the effect of the fine-tuning learning rate α in both TF8 and D'Kitty tasks. Figures 5 (c) and 5 (d) reveal a consistent performance across varied α values for both tasks, highlighting ICT's robustness towards the fine-tuning learning rate.

A.6 Limitation

We validate the effectiveness of ICT across a broad spectrum of tasks. Nevertheless, certain evaluation methodologies do not completely represent authentic situations. For instance, in the superconductor task [5], we adhere to the established convention of utilizing a random forest regression model as the oracle, in line with previous studies [1]. Regrettably, this model may not perfectly mirror the complexities of real-world cases, resulting in discrepancies between our oracle and the ground-truth. Future collaborations with domain experts can potentially refine these evaluation methods. Overall, given the straightforward formulation of ICT, combined with empirical proof of its robustness and effectiveness across diverse tasks in the design-bench [1], we maintain confidence in its capability to effectively generalize to other scenarios.