We thank all three reviewers for unanimously recognizing the significance and merits of our work. We have addressed all their raised concerns below. And we promise to release all codes and pre-trained models upon acceptance.

Structure and Clarity (R2). We thank **R2** for pointing out the important issue. Despite that R2 considers our contributions as significant, we agree with R2 that "*it needs a clearer explanation...*", and further "*the paper could be restructured so all of it fits in the 8 pages limit without compromising readability*".

Our concrete action plan to re-organize the existing materials is as follows:

- First, we will merge Fig. 1 (two attention mechanisms) from Supporting Information (**SI**) into Fig. 1 in the main text. Accordingly, we will elaborate on the details of model architectures, including the matrices $\boldsymbol{Q}$ and $\boldsymbol{M}$, in Section 3.2.2 of the main text rather than Section 1 of **SI**. In addition, we will annotate important notations in the new Fig. 1. In this way we will make model architectures and our first contribution (population-based meta learning) more organized and more clarified as suggested by R2.
- Second, we will describe the posterior distribution, $P(\boldsymbol{x}^*|\boldsymbol{D}_t)$, in Section 3.2.3 of the main text. This could make our second contribution (differential entropy in meta loss) more clear to the readers.
- Third, we will move Fig. 2 from **SI** into the main text as Fig. 2(d)–(f) to better explain results in Section 4.1.
- Fourth, we will move Fig. 3 from **SI** into the main text to clearly explain transferability results in Section 4.4.

We will make space in the main text for the above, by moving the pseudo code of Algorithm 1 and some details about protein docking experiments (Section 4.5) into **SI**. We have already prepared a preliminary version with those revisions.

Ablation Study (R2). We deeply acknowledge the valuable suggestion. To elucidate "which parts have significant effects", we performed an ablation study to progressively show each part's contribution. Starting from the DM_LSTM baseline ($\mathbf{B}_0$), we incrementally crafted four models: running DM_LSTM for k times under different initializations and choosing the best solution ($\mathbf{B}_1$); using k independent particles, each with the two point-based features, the intra-particle attention module, and the hidden state ($\mathbf{B}_2$); adding the two population-based features and the inter-particle attention module to $\mathbf{B}_2$ so as to convert k independent particles into a swarm ($\mathbf{B}_3$); and eventually, adding a differential entropy term in meta loss to $\mathbf{B}_3$, resulting in our **Proposed** model.

We tested the five methods ($\mathbf{B}_0$–$\mathbf{B}_3$ and **Proposed**) on 10D and 20D Rastrigin functions with the same settings as in Section 4.2. We compare the function minimum values returned by these methods in the table below (mean $\pm$ standard deviation over 100 runs, each using 1000 function evaluations).

Dimension	$\mathbf{B}_0$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	**Proposed**
10	55.4$\pm$13.5	48.4$\pm$10.5	40.1$\pm$9.4	20.4$\pm$6.6	12.3$\pm$5.4
20	140.4$\pm$10.2	137.4$\pm$12.7	108.4$\pm$13.4	48.5$\pm$7.1	43.0 $\pm$9.2

Our key observations are as follows. i) $\mathbf{B}_1$ v.s. $\mathbf{B}_0$: their performance gap is marginal. As suggested by R2, this proves that our performance gain is not "just from having k independent runs"; ii) $\mathbf{B}_2$ v.s. $\mathbf{B}_1$ and $\mathbf{B}_3$ v.s. $\mathbf{B}_2$: Whereas including intra-particle attention in $\mathbf{B}_2$ already notably improves the performance compared to $\mathbf{B}_1$, including population-based features and inter-particle attention in $\mathbf{B}_3$ presents the largest performance boost. This confirms that our method to majorly "benefit from the attention mechanisms"; iii) **Proposed** v.s. $\mathbf{B}_3$: adding entropy from the posterior gains further, thanks to its balance of exploration and exploitation. **We hope that the ablation study adds to a "thorough experimental evaluation" and convinces R2 better.**

Contribution to the ML field (R1). We respectively disagree that our work was " a fairly straightforward combination of optimizer learning and population-based optimization". Our work, for the first time, tackles a **novel and important ML topic** (meta learning for population-based optimization) that leads to solving very rugged non-convex optimization problems. Moreover, we believe **our methodology to be highly innovative and have broad implications** to other topics in optimization and learning. First, an important complicacy in population-based optimization lies in the collaboration among particles, which also presents a bottleneck when extending current point-based meta-optimizers. We pioneered to address the bottleneck via the novel inter-particle attention mechanisms across LSTMs. Second, an entropy term in meta loss, based on the posterior directly over the optimum, was designed to balance exploration and exploitation, which is also " *a problem in other state-of-the-art (meta learning) approaches*" (Quote R2). Each of those components contribute substantially, as shown in our ablation study above.

We hope the above clarification has convinced R1 of our notable ML methodology innovations. In fact, we notice the other two reviewers agreed on our work's significance and novelty. Quoting R2: *"I consider that the contributions of their work are novel, especially the proposed architecture" "the contributions are significant"*, and R3: *"Their work opens the door to solving more sophisticated optimization using L2L"*.

Comparison to (Chen et al, 2017) and References (R3). Although no official codes are available, we re-implemented (Chen et al, 2017) and found its performance comparable to $\mathbf{B}_0$ (Andrychowicz et al., 2016), possibly due to their similar model architectures. We will add references on attention mechanisms from the computer vision community.