We thank the reviewers for their feedback. Below we respond to some of the main concerns.

Clarification of the experimental goals R1 and R3 were dissatisfied with the small scale of our experiments. We are happy to run any additional experiments that are deemed crucial for better understanding of our method. In fact, we are happy to leave the choice of additional data and ensemble methods up to the reviewers. However, we would like to first clarify what we were hoping to illustrate with the current setup, and further discuss what constitutes meaningful comparisons for a wrapper method such as our J+aB that comes with a model-free guarantee.

The primary goal of our experiments is to demonstrate that our method achieves near $1-\alpha$ coverage numerically (according to the theory, $1-2\alpha$ is guaranteed). The secondary goal is to verify that although J+aB runs faster than J+ ensemble, $\hat{C}^{\text{J+aB}}_{\alpha,n,B} \approx \hat{C}^{\text{J+ ensemble}}_{\alpha,n,B'}$. The final, lesser goal is to relate known stabilizing properties of bagging by comparing J+aB (or J+ ensemble) vs J+ non-ensemble. These goals are either stated or implied in Lines 251-6, 294-303, but we promise to make them more explicit in the camera-ready version.

There were two main reasons for running experiments on a small scale with a couple of data sets. The first is the page limit. The second is the cost of running J+ ensemble; we needed the experimental parameters to be quite small to be certain of obtaining results by the deadline. This is remarked in Lines 284-6. However, given our experimental goals, we did not see the lack of scale as a significant defect. The advantage of model-free framework is that our coverage guarantee is impossible to break irrespective of the data and the choice of ensemble. We can always choose to look at more data sets and more ensemble architectures, but this will only produce more plots that all look very similar. Meanwhile, the secondary goal amounts to a sanity check, and we have said that the final goal is of lesser significance.

The issue of width is certainly of interest, but here, we would argue that the only meaningful comparisons are those with other wrapper methods. For example, a split conformal variant is competitive with J+aB in terms of computational cost, but is expected to lose in terms of statistical efficiency. Although this is rather obvious by construction, it may be interesting to investigate whether this would translate to meaningful differences in performance. This is a comparison of efficiency that we are happy to add to our current results. Otherwise, the precision of the intervals would be most heavily affected by the fit of the chosen ensemble with the data. However, as we have ceded this choice to the user, opting to develop a fully flexible method that works irrespective of the quality of this choice, we believe that additional comparisons involving more particular instances of $\mathcal{R}$ or φ are not as useful and tangential to the topic.

Breakdown of computational complexity R3 requested a summary of computational complexity. Here, we provide the total number of occurrences for three different types of operations, which can be used to derive the final cost. We focus on bootstrapping, and match the number of models as in Supplement, Lines 166-9. The table below demonstrates that if the model-fitting cost dominates, the cost of J+aB is roughly that of obtaining a single ensemble prediction. We do not claim any advantage for our method when the cost is dominated by aggregation or evaluation. See Line 170.

	#calls to $\mathcal{R}$	#evaluations
Jackknife	$n+1$	$n+1$
Jackknife+	n	$2n$

	#calls to $\mathcal{R}$	#calls to φ	#evaluations
J+ Ensemble	$B'n$	n	$2B'n$
J+aB (on average)	$B'/(1-\frac{1}{n})^m$	n	$2B'/(1-\frac{1}{n})^m$
Ensemble	B'	1	B'

Tradeoff between computational and statistical efficiency R5, as well as R3, expressed concerns about a tradeoff between computational and statistical efficiency for the J+aB vs J+ ensemble comparison. The short answer is that one method does not always win over the other. See Figure S2. First, since $\tilde{B}$ is a user-specified parameter, it can be picked so that the numbers of models in $\hat{\mu}_{\varphi\setminus i}$ are matched on average. See Supplement, Lines 166-9. Second, the more important difference is the *correlation* among $\{\hat{\mu}_{\varphi\setminus i}\}_{i=1}^{n}$. Conditional on the observed data, $\{\hat{\mu}_{\varphi\setminus i}\}_{i=1}^{n}$ are dependent in the case of J+aB and independent in the case of J+ ensemble. (Note that unconditionally they are always highly correlated for both.) What this means for the precision is expected to depend on the data and the choice of ensemble. In any case, this difference is expected to be much smaller than, say, that for the J+aB vs split conformal comparison.

R1 2) The takeaway of Table 1 is in Lines 289-91 (as well as Lines 251-2). It is not our goal to see which among the nine (all *instances* of J+aB) performs best. 3) The results for RF vs Ridge are completely expected given the known results on bagging. The point we are trying to illustrate is in Lines 295-7. 4) Figure 1 shall be amended.

R3 3) The number of trials was *doubled* for Supplement C, which reduced the standard errors. 4) J+aB can be applied to any ensemble algorithm in the form of Algorithm 1 as long as it is agnostic to the ordering of the input data. $\mathcal{R}$ or φ may be arbitrarily complicated, e.g., $\mathcal{R}$ may involve built-in hyperparameter tuning; φ may have adaptive weights. 1) No further condition on $\tilde{B}$ is necessary. The validity comes from construction of an exchangeable array of residuals. 2) For fully distribution-free guarantee, a random B is necessary, as the array is not exchangeable with B fixed. Figure S3 assumes *bagging*, so fixes a particular resampling and a particular φ. Also, the title and the abstract will be amended.