We thank the reviewers for their comments. We address individual concerns below. If you think we address your concerns, please consider raising the score.

Reviewer 1: *Simple method and limited contribution.* We believe the simplicity of our method is a strength, not a weakness. In addition, it's not enough to just "plug and play" Cohen et al.'s method, as our No-Denoiser baseline shows (Tables 1&2 and Figure 2). Thus our algorithmic contributions, though simple, are important.

Relationship with prior work. All prior work which apply a preprocessing step are empirical defenses that hope to remove malicious perturbations by doing a preprocessing step. In our work, we apply a denoiser not to remove the malicious noise, but to make the pre-trained classifier accurate under Gaussian perturbation of its input, therefore making randomized smoothing effective when applied to this pre-trained classifier.

The white-box scenario is misleading. We believe Reviewer 1 misunderstood the meaning and context of "white-box" in our paper. Whether the attacker has access to the denoiser or not doesn't affect our method as we aren't empirically removing adversarial noise. Our guarantees are provable due to the use of randomized smoothing.

Practicality: Performance gap w.r.t to Cohen et al's method. One should expect our method, without retraining the classifier, performs at most as well as Cohen et al. which trains the classifiers with smoothing in the loop. While the gap between these methods point to important future work direction, the value of our contribution is clear by comparing against the No-Denoiser baseline, which is the *real "plug-and-play"* referred to by the reviewer in point 1. See Tables 1&2 and Figures 2,3,&4.

What threat models can the algorithm handle? Since our method uses randomized smoothing, it can in theory handle all threat models that randomized smoothing can handle (including l1, l2, linf, and Wasserstein). The only change in our method would be the way our denoisers are trained. For instance, instead of training a denoiser to remove Gaussian noise (for L2 certification), the denoiser shall be trained to remove noise sampled from other distributions (e.g. Laplace distribution for L1 threat models). We agree this might be confusing in our paper so we will make sure to clarify it.

No proper theoretical justification. We are not sure what theoretical justification the reviewer is pointing to here. Our method applies randomized smoothing to a composition of a classifier and denoiser, instead of only applying it to a classifier as all prior works on randomized smoothing do. Therefore, all the theoretical guarantees of randomized smoothing hold for us.

The paper is tough to follow and read. We will reorganize the paper to be more self-contained in the main text.

Reproducibility We provide detailed experimental details in Appendix B, along with a detailed code + pre-trained models replicating all the experiments. So we are confused why the reviewer thinks the paper is not reproducible.

Reviewer 2: *The authors break the promise of avoiding any re-training that is given in the paper.* We stress that we never re-train the base classifier neither in the white-box nor black-box settings. In the white-box setting, we assume we know the base classifier, and we backpropagate through it, but we only update the denoiser. The whole purpose of our paper is to get non-trivial certification results without re-training the classifier itself. We hope this clarifies the confusion; we will update the paper accordingly.

Comparison to Madry's adversarial training. Madry's defense is empirical, whereas we are interested in certified defenses. But in any case, Madry's defense requires adversarially training the classifier, whereas in our setting, the classifier isn't allowed to be re-trained/modified at all. It is interesting to study whether PGD-like defenses can be applied to pre-trained classifiers without modifying the latter, but this is outside the scope of this paper.

Limited novelty. We agree that our approach looks similar to the use of denoising auto-encoder from Lecuyer et al. However, our approach is distinguished from Lecuyer et al in several ways:
1. The use of denoising auto-encoder in Lecuyer et al is solely aimed at speeding up training for certifying large models. In our case, our motivation is to effectively apply randomized smoothing to pretrained models.
2. More importantly, we comprehensively investigate various training strategies (MSE/stability/classification objectives) and application settings (white-box/black-box), which are not investigated in Lecuyer et al.

Practicality issues. Our approach utilizes randomized smoothing, thus any practicality issue of randomized smoothing passes on to our approach. We agree with the reviewer on this point and we will make it clearer in the next revision.

Gap between ours and Cohen et al. increases as the model becomes more complex. This is indeed an interesting observation that needs further investigation. It might be the case that larger architectures require different denoiser architectures/training schemes. We don't claim we train our denoisers in the best way possible, and we believe with improved training of the denoiser, the gap can be further reduced.

Reconstruction artifacts of STAB+MSE compared to MSE We think the reason for this is that STAB+MSE makes the denoiser more customized to the base classifier, resulting in more corrupted reconstruction. MSE loss only considers removing Gaussian noise, thus leading to visually better output. This requires further investigation and is left for future work.

Reviewer 3: Thanks for your comments!