We thank the reviewers for their helpful comments. Below, we address some of the points made regarding our work's contributions and relation to prior work.

**On automated attacks (Reviewer 1 and 3).** Reviewer 1 and Reviewer 3 argue that "AutoAttack" (Croce & Hein, 2020) is a counter-example to our claim that adaptive attacks are inherently necessary and important to study. However, we believe that the results from that work exactly *support* our claim:

1. **AutoAttack fails to attack many defenses we break completely**. AutoAttack does not fully break the challenging "k-winners take all" defense (19% accuracy), whereas we reduce it to 0% accuracy. Our adaptive attack also outperforms AutoAttack on "ME-Net" and on at least two other defenses we independently evaluated ("Asymmetrical Adversarial Training" and "Are Generative Classifiers More Robust").
2. **AutoAttack only applies to standard feed-forward classifiers.** Many defenses depart from the standard defense template that AutoAttack supports. Of the 13 defenses we study, 5 aim at detecting adversarial examples. For these, formulating an appropriate loss function to optimize is precisely the challenge in developing a strong attack, and it is unclear how this process could be automated by AutoAttack.

   AutoAttack also cannot be directly applied to "Temporal Dependency" (a speech-to-text model) and "Robust Sparse Fourier Transform" (which is aimed at perturbations of small $\ell_0$ norm).

   In fact, a majority of defenses evaluated by AutoAttack use adversarial training, for which there is usually no need for an "adaptive" attack as the inference phase is unchanged. This addresses a question of Reviewer 1, on why we refrained from evaluating such defenses in our paper.
3. **It is easy to build an ineffective defense that AutoAttack fails to break**. First, pick any of the defenses evaluated on AutoAttack for which the query-based attacks (FAB and Square) fail. Then, add a component that masks gradients (e.g., a non-differentiable quantization layer) to defeat PGD. Now AutoAttack fails, even though the defense is easy to circumvent with an adaptive attack.

   (This example also addresses another question of Reviewer 1: since query-based attacks routinely fail, e.g., for randomized defenses, we cannot use them to reliably evaluate all defenses.)

We believe AutoAttack is a strong, *non-adaptive* baseline. However, it is orthogonal to our paper, which argues that static automated attack strategies are not sufficient. The above points illustrate why.

**On adversarial models (Reviewer 2).** We apologize for not clarifying this in the paper. All 13 defenses assume a white-box adversary with full access to the defense parameters. The defenses differ slightly in the perturbation norms and bounds that they consider, and these are mostly incomparable. As a result, all our attacks assume white-box model access and operate with the same perturbation bounds as in the original evaluation. While some of our attacks use black-box optimization techniques, this is solely to side-step gradient-masking. We still view these as white-box attacks.

**On related work & technical novelty (Reviewer 3).** We view the fact that "defenses are broken by existing techniques" as a strength rather than a weakness. Introducing new attack techniques would give the incorrect impression that we currently lack the appropriate technical tools to evaluate defenses properly. But for all of the defenses we studied, a better attack can be built using only tools that are well-known in the literature.

Thus, **the issue with current defense evaluations is methodological rather than technical.** Proposing new attack techniques is not going to fix this. Instead, our paper clearly documents how to make use of existing techniques to build strong adaptive attacks.

This is what differentiates our work from prior work that proposed and argued for adaptive attacks (e.g., Carlini & Wagner 2017, Athalye et al. 2018, Carlini et al. 2019). When those papers were written, defense authors indeed lacked the motivation and tools to conduct strong adaptive evaluations. But of the 13 defenses we study, nearly all claim to perform an adaptive attack evaluation following current best practices. The main contribution of our paper is thus to highlight that there still is a systemic lack of strong adaptive evaluations in this field, and to show how to remedy this.