We thank the reviewers (**R1**, **R2**, **R3**, **R4**, and **R5**) for their thoughtful reviews, and respond to as much as we can given time and space constraints below.

Global vs. pointwise strategyproofness We agree the distinction between pointwise strategyproofness and global strategyproofness is an important one and thank **R4** for pointing out cases where we could further emphasize this. There is some connection between the pointwise regrets we compute and global properties of the mechanism. Duetting et al. have some generalization results that are relevant (see latest arXiv version of their paper). Their Theorem 2.2 gives a generalization bound from pointwise regret estimated on finite samples (what we compute) to true expected regret. Our networks that enforce IR satisfy the assumptions of this theorem. Given true expected regret, Lemma 2.1 allows one to bound the probability of sampling a point where regret is high – not quite a DSIC guarantee but closely related. A crucial point is that Theorem 2.2 is stated in terms of the exact pointwise regret – the true maximum of the difference in utility. It is precisely this quantity which vanilla RegretNet can only approximate but which we can compute. Making explicit reference to these results would definitely be valuable.

Differences in performance from original RegretNet **R1**, **R2**, and **R5** asked about the difference in empirical performance between original RegretNet and our models. Because the main goal of our work was to produce a proof of concept for certifiability, not to get SOTA performance, we made some changes from the original RegretNet architecture and training hyperparameters. Due to RegretNet's sensitivity to hyperparameters, we believe that reproducing optimal results would require a very costly hyperparameter search (for more support of this, see discussion of Rahme et al. under "Additional discussion"). Changes to enable certification include a single trunk architecture rather than separate allocation and payment networks, along with ReLU activations and sparsemax. Additionally, when training, we used different learning rates and much larger batch sizes (and therefore relatively fewer misreport updates) to make training faster. These changes might explain the performance differences. One additional point we want to emphasize is that our modified networks are not necessarily enforcing IC any more strongly than the original RegretNet – they just make it possible to detect with confidence when violations do occur after training is complete.

Previous work in automated mechanism design **R3** points out that more discussion of previous work in learning auctions and automated mechanism design is important. We agree with this and will add such discussion. Our underlying networks are trained using essentially the same RegretNet approach as Duetting et al; our contribution is to use this technique, but modify the network architectures to allow for exact computation of pointwise regret after training is complete. As such, much of the comparison in Duetting et al. to previous work applies to our technique as well. Specifically with regards to the Cai, Daskalakis, Weinberg papers, these are mentioned very briefly in Duetting et al and aim for Bayesian incentive compatibility (BIC), a weaker notion than dominant-strategy incentive compatibility (DSIC). RegretNet, by contrast, aims for an approximate notion of DSIC; this is what we aim for as well, while determining the presence of DSIC violations with greater confidence. We will add discussion briefly in §1 and as a new subsection in §2.

Correctness of certificates **R2** mentions that we do not provide or reference proofs of the correctness of our certificates. The mixed-integer formulation we use gives an exact (up to numerical error) representation of the neural network in the integer program; our certificates just consist of solving this program to find the regret-maximizing misreport. We will explicitly point to the places in the literature where the correctness of these formulations is shown (e.g. Tjeng et al. 2019). The points found as solutions give a lower bound on true regret which is often substantially higher than regret found by gradient ascent; these are also upper bounds certifying maximum true regret, under the assumption that our chosen MIP solver, Gurobi, does correctly solve problems to global optimality when it reports that it has. We will explicitly clarify this assumption as well.

Individual rationality **R4** raises some questions related to IR enforcement. We used both distillation from a teacher (which is perfectly IR by architectural construction) and a Lagrangian penalty to encourage the student network to be IR. We appreciate the feedback and will clarify this. Filtering out IR-violating points refers to testing network output for IR violation and if it occurs, simply charging and allocating nothing – thus player utility is zero, preserving IR, but the auctioneer revenue is also zero.

Additional discussion Subsequent to the NeurIPS submission deadline, a new paper was posted on arXiv, "Auction learning as a two-player game" by Rahme et al. This paper, which we feel is relevant and will discuss briefly if accepted, presents an improved training algorithm and gets better results than RegretNet on the same tasks. A core point of the paper is that RegretNet performance is very sensitive to hyperparameters (discussed throughout, see their Table 1 for specific results), a phenomenon we also observed. The architecture used for their auctioneer network is essentially the same as RegretNet (softmax, IR enforcement via product), so the modifications from our paper could be applied to it. Then the techniques we present could be used to certify networks trained using this improved algorithm, or further improved algorithms in the future.