We thank the reviewers for the careful comments, questions, and recommendations that surely will improve the quality of our paper. Below we highlight some common themes within the reviews and finish with some general clarifications.

1) Motivation. After reevaluating our work considering the reviews, we agree that our work neglects to adequately motivate the importance of the study of guaranteed adversarial robustness for NPCs and the role these models play in Machine Learning (ML). Unfortunately, the newly introduced broader impact section enticed us to shift most of the motivation from the introduction to this new section. We will revert this shift in the camera-ready version and also add some additional background information on NPCs here. As discussed in the broader impact section, NPCs are considered to be one of the most interpretable ML models. This makes NPCs a preferred choice in medicine, where models are required to be interpretable for use in clinical trials. This is highlighted in the recent study *Urine steroid metabolomics for the differential diagnosis of adrenal incidentalomas in the EURINE-ACT study: a prospective test validation study,* published in the renowned journal THE LANCET. In this work, GMLVQ is used to construct an extraction analysis method for detecting adrenal tumors. The resulting model is evaluated in one of the largest studies of ML in medicine to date, consisting of over 2000 participants and spanning almost 10 years. In addition to this, as R#4 rightfully pointed out, NPC-like models are often used in the few-shot and meta-learning fields, resulting in several cross-breeds between deep learning models and NPCs (e.g., *Prototypical Networks for Few-shot Learning).*

2) Related work. As R#1 and R#2 mentioned, there is a strong relation between NPCs and 1-NN. This relation is discussed in the last related work paragraph, but it will be extended to include work mentioned by the reviewers. *Evaluating the Robustness of Nearest Neighbor Classifiers: A Primal-Dual Perspective* contains a formulation similar to the hypothesis margin, but is limited to L^1-, L^2-, and L^∞-norms for the attack and L^2-norms for the classifier metric. Additionally, their methods are order of magnitudes slower and do not achieve the same robustness as GLVQ. *Towards Certificated Model Robustness Against Weight Perturbations* discusses weight perturbations without the relation to adversarial changes to the input and does not use the margin concept. However, the approaches could be complementary. *Robustness for Non-Parametric Classification: A Generic Attack and Defense* is related to the adversarial robustness of NPCs because of the relation to 1-NN. However, their method is more time consuming and is not studied with respect to seminorm-based classification rules like LMNN. *On lp-norm Robustness of Ensemble Decision Stumps and Trees* and *Robustness Verification of Tree-based Models* are partly covered by the cited papers [46, 47], but they will be included and discussed in the camera-ready version. *Interpretable machine learning with reject option* studies adversarial rejection in a similar manner as our work, but focuses on GMLVQ models trained with reject option.

3) Additional experiment. As highlighted by R#1, NPCs are often applied to tabular data. During the experimental design, we performed an additional experiment using tabular datasets more closely related to the usual NPC domain:

		Robust Stumps		Robust Trees		GLVQ with L^∞-norm	
Dataset	ϵ of L^∞-attack	CTE	URTE	CTE	URTE	CTE	URTE
breast-cancer	0.3	5.1%	10.9%	0.7%	**6.6%**	**0.0%**	8.7%
diabetes	0.05	27.3%	31.8%	27.3%	35.7%	**22.0%**	**33.7%**
cod-rna	0.025	11.2%	22.6%	**6.9%**	**21.4%**	7.6%	**21.4%**

In the experiment, GLVQ is compared to *Provably Robust Boosted Decision Stumps and Trees against Adversarial Attacks* [46], using a similar setup. The comparison shows that an NPC trained using a triplet loss has a guaranteed robustness similar to the robustified stumps and trees. We agree with the reviewer that by selecting the presented experiments to make the comparison to state-of-the-art methods, we, unfortunately, ignore the power of NPCs for tabular data. We will include the additional experiment in the supplementary to counterbalance this.

4) Clarifications. First, RSLVQ serves the role of highlighting that an NPC trained without a triplet loss of seminorms does not necessarily become robust. Hence, the atrocious robustness of RSLVQ, as observed by R#1, is therefore the expected/desired behavior—how RSLVQ can be robustified is not clear yet, as questioned by R#2. We will move the overview found in Sec. C.1 from the supplementary to the main paper to clarify the model selection intentions. Similarly, the importance of the seminorm independence will also be amplified. In short, this allows for using task-specific or adaptive metrics. NPCs such as LMNN and GMLVQ are therefore still provably robust against adversarial examples under the seminorm they use for classification—according to the provided theorems. Second, as R#2 correctly observed, all compared methods—except for RSLVQ—are also robustification approaches. With the classification of the used models as either verification or certification, we refer to the method used for obtaining the URTE. Hence, the combination of verification/certification and robustification is possible. Similar, as R#3 mentioned, the robustification of IBP is indeed based on a certification method, as it does not optimize the exact robust loss during training. The URTE of IBP is however obtained using a MIP solver, and therefore the provided URTE falls within the verification class. This nuance will be made more concrete in the final version. Lastly, some short notes: We will add a dedicated section to the supplementary to discuss LVQ and specifically its training process. We will refer to this section in Sec. 3 to answer questions regarding the training of LVQ models. Computational cost/time complexity of the models will be included in the tables. A figure will be added to Sec. 4 to explain the geometrical intuition behind the formulas and proofs. We will clarify and rephrase the sentences mentioned by R#3 and clarify that "–" in Table 1 means not reported or calculable.