UNCOVERING THE LIMITS OF IBP-BASED CERTIFIED TRAINING THROUGH SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

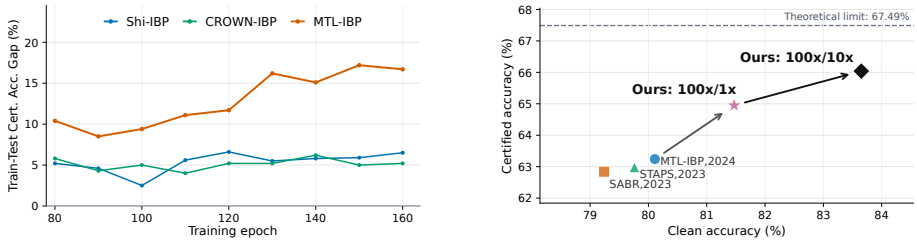
While deep neural networks have enabled remarkable advances across a wide range of supervised learning tasks, their susceptibility to *adversarial attacks* remains a major concern. *Certified training* methods based on *interval bound propagation* (IBP) produce models with mathematically rigorous robustness guarantees, but typically suffer from degraded performance on clean data. While recent methods improved this trade-off by training on both adversarial examples and IBP bounds, progress has stalled, suggesting that IBP-based certified training may be approaching its limits. Here, we show that the progress that has been achieved came at the cost of severe overfitting, which traditional IBP training suffered less from. Thus, to uncover new performance limits of these methods, for the first time, we employ large-scale synthetic datasets and conduct the first systematic scaling study of IBP-based certified training across data and compute. We demonstrate that scaling yields performance improvements comparable to several years of algorithmic progress; for instance, on CIFAR-10 with $\epsilon = \frac{2}{255}$, we obtain a model with **83.65% clean and 65.90% certified accuracy**, substantially surpassing the previous state of the art of 79.87% and 64.54%, respectively. Analysing the resulting models, we find that these improvements are accompanied by changes in the distribution of unstable ReLUs and bound propagation tightness. At the same time, we reveal more fundamental limits of IBP-based certified training: performance quickly saturates with scaling, the gained robustness conflicts with propagation tightness, and stronger models incur increasingly prohibitive verification costs, making verification systems a bottleneck for certifying robust models.

1 INTRODUCTION

Despite the remarkable performance of deep neural networks, their susceptibility to imperceptible, adversarially chosen perturbations that cause misclassification remains a major obstacle to their responsible use in safety-critical settings (Szegedy et al., 2014; Goodfellow et al., 2015). *Adversarial attacks* (see, *e.g.*, Croce and Hein, 2020; Madry et al., 2018) that efficiently identify adversarial examples can be used to assess the robustness of a neural network, but, due to their heuristic nature, may fail to discover adversarial manipulations, even when these exist. Therefore, to obtain rigorous robustness guarantees, *neural network verification* techniques have been developed to provide mathematical certificates on the behaviour of a neural network. However, those incur substantially increased computational costs (see, *e.g.*, Kaulen et al., 2025; König et al., 2024), due to the $\mathcal{NP}$-completeness of the verification problem (Sälzer and Lange, 2021).

To mitigate these computational costs, *certified training* aims to learn networks for which robustness certificates can be obtained more efficiently. These methods minimise an over-approximation of the worst-case loss obtained through *incomplete verification* methods, which provide sound bounds on the reachable output set for a given bounded input region. Perhaps surprisingly, training approaches based on the relatively loose and conceptually simple *interval bound propagation* (IBP) (Mirman et al., 2018; Gowal et al., 2019) achieve state-of-the-art performance (Mao et al., 2025; Kaulen et al., 2026). However, while robustness certificates can be obtained far more efficiently for the resulting networks, even the most sophisticated methods suffer severely from degraded performance on clean inputs (Müller et al., 2023; De Palma et al., 2024b; Kaulen et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: Left: Recent advances in certified training (MTL-IBP, De Palma et al., 2024b) induce worsened overfitting behaviour. Right: Scaling training data by $100\times$ and compute by $10\times$ on CIFAR-10, $\epsilon = \frac{2}{255}$ yields gains comparable to years of algorithmic progress.

Hybrid methods that combine adversarial training with IBP bounds have substantially advanced the clean–certified accuracy trade-off (see, *e.g.*, Müller et al., 2023; De Palma et al., 2024b). However, their performance has begun to saturate (Kaulen et al., 2026), and, to date, it is unclear what prevents further improvements. In this work, we show that these hybrid methods suffer from severe robust overfitting which substantially limits their performance. As shown in Figure 1a, we observed an increasing gap between train and test certified accuracy for the hybrid method MTL-IBP (De Palma et al., 2024b), which is far less pronounced in traditional IBP and CROWN-IBP training. This raises the question of whether the apparent performance ceiling of current methods reflects a fundamental limitation of IBP-based certified training or, instead, insufficient generalisation.

Robust overfitting and its mitigation through additional data are well studied in adversarial training (Rice et al., 2020; Gowal et al., 2021), but it is unclear whether these findings transfer to certified training, where performance additionally depends on learning network parameters that can be tightly bounded. Indeed, previous attempts to scale IBP-based methods did not yield analogous improvements (Altstidl et al., 2024). We therefore investigated, for the first time, whether scaling training data can mitigate overfitting in IBP-based certified training and reveal its performance limits. We demonstrate that for the representative hybrid method MTL-IBP (De Palma et al., 2024b), scaling CIFAR-10 training data reduces overfitting and yields an improvement of 1.64% in clean and 1.47% in certified accuracy under an ℓ_∞ threat model with radius $\epsilon = \frac{2}{255}$. In addition, we show that, once auxiliary data is incorporated, computational effort can be scaled accordingly, yielding even stronger models with up to **83.65% clean and 65.90% certified accuracy**. Thereby, we show that mitigating robust overfitting can close a substantial fraction of the gap to the theoretical certified robustness limit of 67.49% on that benchmark (Zhang and Sun, 2024). Similarly, we establish a new state of the art on CIFAR-10 at $\epsilon = \frac{8}{255}$ and on TinyImageNet at $\epsilon = \frac{1}{255}$. Overall, this shows that resource scaling yields gains comparable to several years of algorithmic progress (see Figure 1b).

To understand the source of these gains beyond mitigated overfitting, we analysed the resulting models and found that scaling changes the structure and verifiability of the learned networks, leading to systematic shifts in ReLU instability and bound-propagation tightness. Thus, the strongest resulting models become increasingly difficult to certify using complete verification, making verification itself a bottleneck for establishing guarantees that the models may in fact satisfy. Together, these results show that current methods had not yet reached their performance limits, while also exposing more fundamental limitations of both the training objective and the downstream verification pipeline.

In summary, our contributions are as follows:

1. We identify severe robust overfitting in recent hybrid IBP-based methods and mitigate it with large-scale diffusion-generated data.
2. We show that once overfitting is reduced, compute and model capacity can also be increased and thereby conduct the first systematic scaling study of IBP-based certified training, achieving new state-of-the-art results.
3. We analyse how scaling changes the learned models and find systematic shifts in ReLU instability and bound-propagation tightness. At the same time, we uncover more fundamental limitations: performance quickly saturates with further scaling, while stronger models do not lead to uniformly tighter bounds and incur increasingly prohibitive verification costs.

2 Background and Related Work

In the following, we present the relevant background for our work, including neural network verification, certified training and augmenting training data with synthetic images generated using diffusion.

2.1 Neural Network Verification

The neural network verification task refers to formally proving whether an input–output property *holds* or is *violated* for a given neural network. We focus on local robustness for classifiers $f_\theta : \mathbb{R}^d \to \mathbb{R}^c$: given an input $\mathbf{x}_0 \in \mathbb{R}^d$ with scalar label $y_0 \in \mathbb{N}$, the prediction should remain constant for all perturbations in $\mathcal{B}^\epsilon_\infty(\mathbf{x}_0) = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_0\|_\infty \leq \epsilon\}$, *i.e.*,

$$\forall \mathbf{x}' \in \mathcal{B}^\epsilon_\infty(x_0) : \arg\max_j f_\theta(\mathbf{x}')_j = y_0 \tag{1}$$

Equivalently, the problem can be formulated as computing the sign of the following problem, where $\mathbf{z}(\mathbf{x}, y) \in \mathbb{R}^c$ is defined as the vector of logit differences, *i.e.*, $\mathbf{z}(\mathbf{x}, y) := f_\theta(\mathbf{x})_y \cdot \mathbf{1} - f_\theta(\mathbf{x})$:

$$\min_{\mathbf{x}' \in \mathcal{B}^\epsilon_\infty(\mathbf{x})} \min_{i \neq y} \mathbf{z}(\mathbf{x}', y)_i \tag{2}$$

Obtaining a solution to this problem is $\mathcal{NP}$-complete (Sälzer and Lange, 2021). Thus, current state-of-the-art solvers employ cheap incomplete verification techniques (see, *e.g.*, Zhang et al., 2018; Singh et al., 2019; Xu et al., 2021) that yield sound lower bounds to the logit margins $\underline{\mathbf{z}}(\mathbf{x}, y)_i \leq \mathbf{z}(\mathbf{x}, y)_i, i \in \{1, \ldots, c\}$ that may be tight enough to decide the problem. These methods can be combined with the branch and bound paradigm (Bunel et al., 2020) to recursively tighten bounds on the outputs (see, *e.g.*, Wang et al., 2021; Ferrari et al., 2022; De Palma et al., 2024a; Duong et al., 2025); the procedure thus obtained is provably *complete*.

2.2 Certified Training

Madry et al. (2018) introduced the problem of training robust neural networks as a min-max optimisation problem concerned with minimising the expected worst-case loss within a given threat model. For the local robustness property, given a loss function $\mathcal{L} : \mathbb{R}^c \times \mathbb{N} \to \mathbb{R}$ and a data distribution $(\mathbf{x}, y) \sim D$, this refers to finding a set of parameters

$$\theta \in \arg\min_{\theta'} \mathbb{E}_D[\max_{\mathbf{x}' \in \mathcal{B}^\epsilon_\infty(\mathbf{x})} \mathcal{L}(f_{\theta'}(\mathbf{x}'), y)]. \tag{3}$$

An exact solution to the inner maximisation would require solving Equation 2 precisely and is therefore $\mathcal{NP}$-complete. Adversarial training replaces the inner maximisation by an under-approximation based on adversarial examples $\mathbf{x}_{\text{adv}}$, yielding the adversarial loss $\mathcal{L}_{\text{adv}} := \mathcal{L}(f_\theta(\mathbf{x}_{\text{adv}}), y)$ (Madry et al., 2018). In *certified training*, an over-approximation of the worst-case loss obtained through incomplete verification methods is employed instead. Using the previously defined lower bound to the logit differences of f_θ, the verified loss $\mathcal{L}_{\text{ver}}$ is defined as (Wong and Kolter, 2018):

$$\max_{\mathbf{x}' \in \mathcal{B}^\epsilon_\infty(\mathbf{x})} \mathcal{L}(f_\theta(\mathbf{x}'), y) \leq \mathcal{L}_{\text{ver}} := \mathcal{L}(-\underline{\mathbf{z}}(\mathbf{x}, y), y) \tag{4}$$

Best-performing methods are based on *interval bound propagation* (IBP) (Mirman et al., 2018; Gowal et al., 2019), which, using interval arithmetic, propagates a hyper-box encompassing the input set through the network to obtain sound, yet usually loose, bounds on the reachable output set. The interval arithmetic underlying IBP is summarised in Appendix A.1. Competitive IBP training typically requires ϵ ramp-up schedules, together with specialised initialisation and regularisation techniques that stabilise bound propagation (Shi et al., 2021). More recent methods improved the clean–certified accuracy trade-off by combining adversarial and certified training objectives (see, *e.g.*, Müller et al., 2023; Mao et al., 2023), yielding substantial performance gains. Since then, however, further improvements have been largely incremental (see, *e.g.*, Mao et al., 2025; Kaulen et al., 2026). De Palma et al. (2024b) unified these approaches through the notion of *expressive losses*, which interpolate between adversarial and certified training objectives, and identify expressiveness as the decisive factor underlying state-of-the-art performance. Among several representative losses, they introduce MTL-IBP, $\mathcal{L}_{\text{MTL}} = \alpha \cdot \mathcal{L}_{\text{ver}} + (1 - \alpha) \cdot \mathcal{L}_{\text{adv}}$, which achieves state-of-the-art results across commonly used benchmark settings (Kaulen et al., 2026). This performance is generally assessed in terms of *clean* and *certified accuracy* where the former denotes the proportion of correctly classified inputs and the latter denotes the proportion of inputs that are certifiably robust.

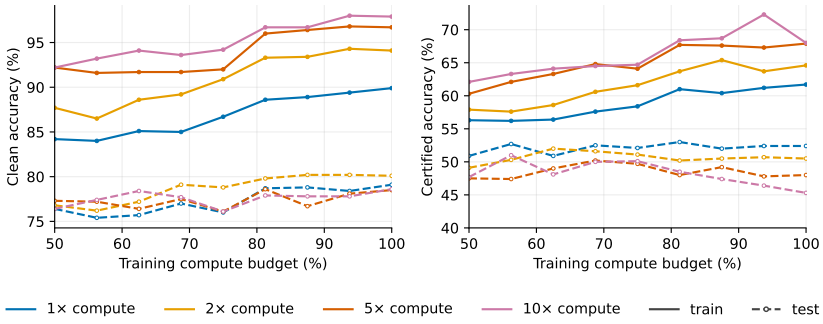


Figure 2: Robust overfitting of the hybrid certified training method MTL-IBP (De Palma et al., 2024b). Increasing the number of training epochs leads to higher clean and certified accuracy on the training data. However, while performance on the unperturbed test set remains stable, certified accuracy decreases severely with increasing compute, indicating robust overfitting (Rice et al., 2020).

2.3 Synthetic Data for Robustness

Robust overfitting is a known limitation of adversarial training, where increased robustness on the training set leads to degrading generalisation (Rice et al., 2020). Recently, diffusion models (Ho et al., 2020; Song et al., 2021) have enabled high-quality synthetic image generation, with *elucidating diffusion models* (EDMs) achieving state-of-the-art sample quality (Karras et al., 2022; Kim et al., 2023). Integrating additional training data into adversarial training has proven as a particularly effective mitigation strategy against robust overfitting, with recent state-of-the-art approaches relying heavily on synthetic images generated through diffusion models (see, *e.g.*, Hendrycks et al., 2019; Rebuffi et al., 2021; Gowal et al., 2021). Higher data quality enabled additional performance improvements (Wang et al., 2023), with recent work showing that these gains follow scaling-law dynamics (Bartoldson et al., 2024).

For certified robustness, synthetic data has received considerably less attention. Altstidl et al. (2024) investigated the scaling behaviour of specialised architectures with low Lipschitz constants (Hu et al., 2023; Xu et al., 2022; Zhang et al., 2022a;b) that are state of the art for deterministic certified robustness against ℓ_2 and large ℓ_∞ perturbations. They showed that diffusion-generated data benefits performance and that further scaling compute and model size yields new state-of-the-art results. However, their preliminary experiments with MTL-IBP reported *catastrophic* performance when scaling training data and compute, suggesting that these benefits do not trivially transfer to IBP-based certified training. Nevertheless, the state-of-the-art performance of IBP-based certified training in low-perturbation regimes on CIFAR-10, as well as on high-dimensional datasets, such as TinyImageNet, motivates a thorough investigation of its overfitting and scaling behaviour.

3 Scaling IBP-based Certified Training

In the following, we investigate to which extent state-of-the-art certified training methods suffer from overfitting, before conducting a systematic scaling study to uncover whether synthetic training data can mitigate this limitation and, in turn, can enable further scaling of compute and model capacity.

3.1 Robust Overfitting in IBP-based Certified Training

While recent hybrid methods remarkably improved performance of IBP-based certified training (Kaulen et al., 2026), subsequent progress has been mostly incremental. To uncover whether these performance limits are inherent to IBP-based training objectives or caused by insufficient generalisation, we investigated the overfitting behaviour of the representative state-of-the-art method MTL-IBP (De Palma et al., 2024b) using the experimental setup of De Palma et al. (2024b) on CIFAR-10 with $\epsilon = \frac{2}{255}$. For this, we evaluated clean and certified accuracy on 1000 randomly selected inputs from the training and test sets for increasing compute budgets using α-CROWN (Xu et al., 2021) for verification. We adopted the commonly used training schedule of 160 epochs as

baseline setting (see, *e.g.*, Kaulen et al., 2026; De Palma et al., 2024b; Müller et al., 2023) and scaled the number of training epochs by factors $\{2, 5, 10\}$ while proportionally extending the warm- and ramp-up phases. In Figure 2, we show the resulting development of training and testing set performance. While both certified and clean accuracy continue to improve on the training set with increased compute, clean testing performance remains relatively stable. In contrast, certified accuracy on the testing set strongly deteriorates as the compute budget is increased.

Interestingly, this established phenomenon known as *robust overfitting* in adversarial training (Rice et al., 2020) was not known to be a major bottleneck of certified training, since it is less pronounced in traditional IBP-based methods (see Appendix C.1). Thereby, our results suggest that, with the introduction of adversarial objectives into certified training, robust overfitting has also become a limitation of state-of-the-art certified training methods. Consequently, the performance of state-of-the-art certified training appears to be limited less by its ability to fit robust predictors than by its ability to generalise the robustness thus achieved to previously unseen data.

3.2 Mitigating Overfitting through Scaling

Having identified robust overfitting as a potential limitation of IBP-based certified training, we investigated whether robust generalisation can be improved by using additional training data. Similar in spirit to recent scaling studies in the robustness literature (Wang et al., 2023; Bartoldson et al., 2024; Altstidl et al., 2024) and the broader machine learning community (Hestness et al., 2017; Kaplan et al., 2020; Zhai et al., 2022), we further investigated whether mitigating overfitting through auxiliary data enables additional gains from scaling compute and model capacity. To this end, we systematically studied the scaling behaviour of the representative hybrid method MTL-IBP (De Palma et al., 2024b) across data, compute, and capacity.

Scaling data. We employed class-conditional images generated by EDMs using models trained exclusively on the respective training sets, following the procedure of Wang et al. (2023). To avoid biases induced through data selection, we naïvely subsampled from the diffusion data uniformly at random. In each training batch, we mixed real data from the training set with generated data according to a predefined ratio. In our experiments, given a dataset of length d, we analysed behaviour with regard to a scaling factor c defined as training on $c \cdot d$ images, out of which d are real and $d \cdot (c-1)$ are generated.

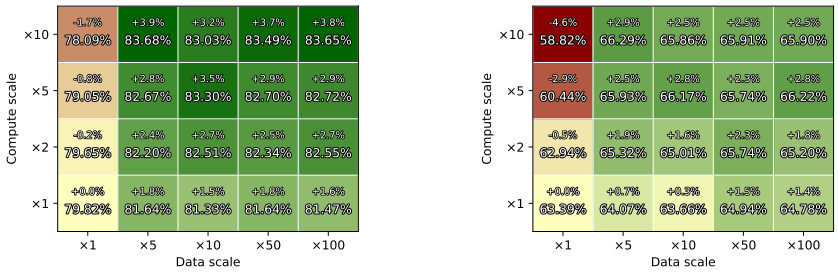
Scaling compute. We define one epoch as $\lceil d/b \rceil$ optimisation steps, where d is the number of real training images and b the batch size, even when generated data is used. This enables a clear separation of the effects of scaling and compute. 'Scaling compute by a factor k' refers to training for $k \cdot e$ epochs, where e denotes the length of the baseline training schedule. Importantly, we scaled warm- and ramp-up phases proportionally to the overall training budget to ensure consistent optimisation dynamics.

Scaling network capacity. We utilised the CNN-7 architectures introduced by Shi et al. (2021), which have become the *de facto* standard for evaluating IBP-based certified training (see, *e.g.*, Kaulen et al., 2026; De Palma et al., 2024b; Mao et al., 2025). Following Mao et al. (2024), who demonstrated that increased network width can improve performance, we additionally trained a wider variant of CNN-7 with doubled layer widths under the best-performing scaling settings.

3.3 Setup of Experiments

We built upon the certified training implementations and evaluation pipelines provided by the CTRAIN library (Kaulen and Hoos, 2025) and evaluated scaling on CIFAR-10 (Krizhevsky et al., 2009) with $\epsilon \in \{\frac{2}{255}, \frac{8}{255}\}$, and on TinyImageNet (Le and Yang, 2015) with $\epsilon = \frac{1}{255}$. Unless stated otherwise, we used the hyperparameter settings of De Palma et al. (2024b) and defined the unscaled baseline as training for 160 epochs with a batch size of 128.

We evaluated the resulting models with regards to clean and certified accuracy on the original test sets. To determine whether an input is provably robust, we employed the state-of-the-art (Kaulen et al., 2025; König et al., 2024) neural network verification system α,β-CROWN (Wang et al., 2021; Xu et al., 2021; Zhang et al., 2018) in its standard configuration, with a per-instance timeout of 300 s in wall-clock time. To ensure that our findings are not dependent on the chosen timeout, we provide additional experiments with increased verification budgets in Appendix C.5.



(a) Clean Acc., $\epsilon = \frac{2}{255}$

(b) Certified Acc., $\epsilon = \frac{2}{255}$

(c) Clean Acc., $\epsilon = \frac{8}{255}$

(d) Certified Acc., $\epsilon = \frac{8}{255}$

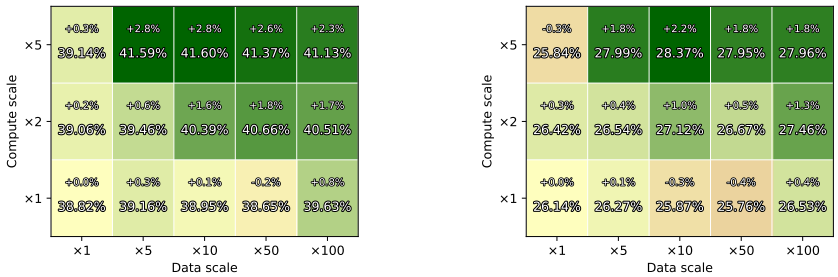
Figure 3: Clean and certified accuracies together with improvements relative to the unscaled baseline for CIFAR-10, $\epsilon \in \{\frac{2}{255}, \frac{8}{255}\}$ on the CNN7 architecture when scaling compute by factors $\{2, 5, 10\}$ and scaling data by factors $\{5, 10, 50, 100\}$.

For CIFAR-10, we evaluated over a grid of data scaling factors $\{1, 5, 10, 50, 100\}$ and compute scaling factors $\{1, 2, 5, 10\}$. For TinyImageNet, we considered a reduced grid, due to the higher costs associated with model training, using compute scaling factors $\{1, 2, 5\}$. In preliminary experiments, we investigated the optimal ratio of real to generated data by evaluating a grid in increments of 10% using the baseline training schedule. We identified approximately optimal fractions of 30% real data for CIFAR-10, which is consistent with prior work (Wang et al., 2023; Altstidl et al., 2024), and 60% for TinyImageNet.

3.4 Results

In Figure 3, we present the performance of models trained with MTL-IBP on CIFAR-10 for $\epsilon \in \{\frac{2}{255}, \frac{8}{255}\}$ with increased data and compute scaling. For CIFAR-10 at $\epsilon = \frac{2}{255}$, additional data already improved performance under the standard compute budget, with gains immediately appearing at a data scaling factor of 5 and saturating around a factor of 50. The best data-only scaled model reached 81.64% clean and 64.94% certified accuracy, surpassing previously known best results by 1.77% and 0.40%, respectively (Kaulen et al., 2026). Scaling data and compute jointly yielded strongest results: When using $100\times$ data and $10\times$ compute, MTL-IBP achieved 83.65% clean and 65.90% certified accuracy, improving over the unscaled baseline by 3.83% and 2.51%, respectively, thereby establishing a new state of the art in certified training. Notably, these gains are comparable to years of algorithmic progress, *e.g.*, the improvement from SABR (Müller et al., 2023) to MTL-IBP (De Palma et al., 2024b) was 0.87% and 0.40% in clean and certified accuracy.

For $\epsilon = \frac{8}{255}$, scaling yields smaller but still meaningful gains. Data scaling alone mainly improved clean accuracy, with modest changes in certified accuracy. The best configuration used $50\times$ data and $10\times$ compute, reaching 57.27% clean and 35.58% certified accuracy, a new state of the art in clean accuracy. Notably, certified performance saturated near levels achieved by prior work that utilised longer schedules for training at the full ϵ value (see, *e.g.*, Mao et al., 2025; De Palma et al., 2024b). Importantly, these gains cannot be traced back to seed variance (see Appendix C.3) and transfer to SABR and, to a lesser extent, to traditional IBP training (see Appendix C.2). In addition, we found that quality of the synthetic data only has a limited effect on performance (see Appendix C.4).



(a) Clean Acc., $\epsilon = \frac{1}{255}$

(b) Certified Acc., $\epsilon = \frac{1}{255}$

Figure 4: Clean and certified accuracies together with improvements relative to the unscaled real-only baseline for TinyImageNet, $\epsilon = \frac{1}{255}$ on the CNN7 architecture when scaling compute by factors $\{2, 5\}$ and scaling data by factors $\{5, 10, 50, 100\}$.

Figure 4 illustrates the effects of scaling certified training on TinyImageNet with $\epsilon = \frac{1}{255}$. Here, the same qualitative pattern can be observed; scaling data in isolation yielded moderate improvements only, while the best performance required increasing compute as well. The strongest model used $10\times$ data and $5\times$ compute, achieving 41.60% clean and 28.37% certified accuracy, thereby establishing a new state of the art. However, performance largely saturated beyond data scaling factors of $5\times$, indicating that factors other than data availability limit performance on this dataset.

Finally, we tested whether additional model capacity helps in light of scaled data and compute, by doubling CNN-7 layer widths for the best-performing scaling settings. On CIFAR-10, this improved clean accuracy by 1.72% at $\epsilon = \frac{2}{255}$ and 0.47% at $\epsilon = \frac{8}{255}$, but reduced complete certified accuracy by 5.95% and 0.98%, respectively. We found that this drop is caused by harder verification problems, *e.g.*, for $\epsilon = \frac{2}{255}$, the fraction of unsolved instances increased from 6.72% to 14.68%.

4 Uncovering Fundamental Limits of IBP-based Training

In the following, we analyse generalisation, ReLU instability, bound tightness, and verification cost and uncover more fundamental limitations that remain beyond scaling.

4.1 Generalisation

First, we investigated whether the observed performance gains through scaling can indeed be traced back to the mitigation of robust overfitting. For this, similarly to the experiments from Section 3.1, we evaluated the training–testing gap in certified accuracy of the scaled models on 1000 randomly selected inputs per split and used α-CROWN for verification. We observed that the gap between training and testing certified accuracy decreased with increasing amounts of auxiliary data and did not widen again when additional compute was introduced, suggesting improved robust generalisation. For CIFAR-10 with $\epsilon = \frac{2}{255}$, the gap in certified accuracy approached zero at data scaling factors of 50 and above, whereas at $\epsilon = \frac{8}{255}$, the training–testing gap reached its minimum already at a data scaling factor of 5. One possible explanation lies in the different MTL-IBP loss weightings used for the two perturbation radii, following De Palma et al. (2024b). At $\epsilon = \frac{8}{255}$, the verification loss receives substantially more weight than at $\epsilon = \frac{2}{255}$ (50% compared to 0.04%). Consequently, training relies less heavily on the adversarial loss, which may contribute to the reduced robust overfitting observed. For TinyImageNet, overfitting similarly decreased with increasing amounts of auxiliary data, with the train–test gap approaching zero at $10\times$ data.

Overall, these results support the hypothesis that mitigating robust overfitting is crucial for improving IBP-based training, while also showing that the potential gains are largest in settings that rely more heavily on adversarial training and consequently exhibit more severe robust overfitting. Additional details on our analysis are available in Appendix D.1.

4.2 ReLU Activation Instability

A ReLU is unstable for a given input region if its pre-activation can take both positive and negative values (Li et al., 2023), which is particularly relevant for certified robustness, as these neurons need to be relaxed during verification. To investigate how scaling changes the internal structure of the learned networks, we analysed upper bounds on ReLU instability obtained with α-CROWN (Xu et al., 2021) on 1 000 randomly selected CIFAR-10 test inputs.

While the overall fraction of unstable ReLUs changed only moderately across scaling settings, we observed a shift in *where* instability occurs within the network. For CIFAR-10 with $\epsilon = \frac{2}{255}$, the unscaled baseline exhibited its highest instability in the activations preceding the classification layer, where 16.6% of ReLUs were unstable on average. With $100\times$ data, this fraction decreased to 12.6%, and further to 8.2% when compute was additionally scaled by $10\times$. At the same time, instability increased within the feature extractor; for instance, in the second convolutional layer that contains $128\times$ more ReLUs than the classification layer, the fraction of unstable activations increased by 0.8%. We observed the same behaviour for $\epsilon = \frac{8}{255}$, where final-layer instability decreased from 8.3% to 4.3% in the best-performing model, while instability in the first layer increased by approximately 1%. Thus, scaling does not simply make the network more or less stable overall, but shifts non-linear behaviour from the classifier towards the feature extractor. This suggests that stronger models learn more expressive feature representations while making the final classification more stable within the perturbation region. Layer-wise results are provided in Appendix D.2.

4.3 Verification Hardness

Finally, we observed that scaling certified training increased certified accuracy while simultaneously rendering the resulting models more difficult to verify. For CIFAR-10 with $\epsilon = \frac{2}{255}$, the PAR-10 score increased from 168 for the unscaled model to 204 for the best-performing scaled CNN-7, while the fraction of unsolved instances increased from 5.51% to 6.72%. This effect became substantially stronger when model capacity was increased: for the wide CNN-7 with maximal scaling of data and compute, PAR-10 reached 445, while 14.68% of the instances remained unsolved. As reducing timeouts is essential for further improving performance, we investigated the source of these cases.

Towards this end, we define the relaxation deficit as the difference between the α-CROWN lower bound on the worst-case margin and the certification threshold of zero. A larger deficit therefore means that the initial relaxation is further from establishing robustness and requires more tightening through branch and bound, making verification more challenging. Looking into provably robust and unsolved instances, on CIFAR-10 with $\epsilon = \frac{2}{255}$, we found that for the scaled CNN-7, the median deficit is 0.51 for timeouts compared to 0.16 for solved instances. The effect becomes particularly pronounced for the wider model, where the median deficit is 0.89 for unsolved instances compared to 0.24 for solved instances. Accordingly, the relaxation deficit discriminates well between solved instances and timeouts: varying a threshold on the deficit yields ROC-AUC values between 0.90 and 0.94 across the investigated models, indicating that a higher deficit signals timeouts. However, when examining the distribution over all correctly classified inputs, the relaxation deficits improved with scaling: the mean deficit decreased from 0.23 for the baseline to 0.18 with joint data and compute scaling, while the 90th percentile, reflecting the upper tail of the distribution, decreased from 0.93 to 0.78. Thus, the increased verification cost is not explained by a broad deterioration of the relaxation, but may arise from changes in the subset of inputs that become robust as performance improves.

To investigate this further, we considered the $6,276$ inputs that were correctly classified and either verified robust or unresolved by the verifier in all three configurations. On this common set, the 90th percentile of the relaxation deficit decreased from 0.165 for the baseline to 0.013 with data scaling and to 0 with additional compute. Hence, scaling does not make these common inputs harder to bound. In contrast, among inputs outside the common set that were correctly classified and either verified robust or unresolved in a given configuration, the relaxation deficits were substantially larger. Their 90th-percentile deficit was 0.868 for the baseline, 0.628 with data scaling, and 0.744 with joint data and compute scaling. At the same time, the number of such inputs increased from 614 to 829 and 986, respectively. These results indicate that scaling increases the number of inputs that are correctly classified and either verified robust or unresolved, and these additional inputs are substantially harder to bound than those in the common set. This creates a small subset of substantially harder verification instances, even though the initial relaxation improves for most inputs.

4.4 Discussion

Our study revealed that the apparent performance limits of IBP-based certified training were largely caused by robust overfitting, which can be effectively mitigated using diffusion-generated data. However, through the conducted scaling study, we uncovered several more fundamental limitations that prohibit meaningful progress in certified training.

Limitation 1: Mitigating overfitting reveals new performance limits. By adding generated data, we can largely close the gap between training and test performance, thus showing that robust overfitting can be substantially mitigated. However, there are limits to the improvement in certified accuracy that can be achieved in this way. In fact, both training and test performance stagnate, once the model has learned the available robust training signal; an effect that was particularly pronounced in our experiments for the larger perturbation radius of $\frac{8}{255}$, where only small improvements over the state of the art could be realised. This suggests that further improvements will require changes to the training objective or model architecture rather than simply additional training data.

Limitation 2: Robustness gains conflict with propagation tightness. As scaling improved generalisation, additional challenging inputs became certifiably robust. However, accommodating these additional robust instances appears to require increasingly complex network behaviour that is harder to bound tightly. This tension became particularly apparent when model capacity was increased: for the wide CNN-7, relaxation quality deteriorated even on inputs that remained robust across configurations. Thus, current IBP-based objectives might not be capable of producing models that can be tightly bounded for all robust instances at the same time.

Limitation 3: Better models become harder to verify. As shown previously, scaling substantially increases verification cost, which became especially problematic when model capacity was scaled. This is particularly unfortunate, because such models can still achieve very strong certified accuracy, *e.g.* 59.95%, suggesting that further improvements in the state of the art may be within reach and may even be close to recently uncovered theoretical limits of 67.49% on CIFAR-10, $\epsilon = \frac{2}{255}$ (Zhang and Sun, 2024). However, to reliably measure these advancements, substantial improvements in verification systems will be required.

5 Conclusions and Future Work

In this work, we conducted the first systematic study of the scaling behaviour of IBP-based certified training. Our results identify robust overfitting as an important limitation of recent state-of-the-art methods and show that mitigating it through data scaling substantially improves both clean and certified accuracies, yielding new state-of-the-art results across commonly studied benchmark settings. At the same time, the saturation of these gains demonstrates that insufficient generalisation explains only part of the apparent performance ceiling and exposes more fundamental limitations of current IBP-based methods.

Our analysis reveals how these limitations emerge as certified training becomes stronger. Initially, scaling improves worst-case margins and extends robustness to additional inputs. While relaxation quality improves for most inputs, these newly covered inputs form a small but substantially harder tail that current state-of-the-art complete verifiers frequently cannot solve. With increased model capacity, this effect becomes more pronounced: further robustness gains are accompanied by deteriorating propagation tightness even on inputs that were already robust under smaller models, further increasing verification difficulty. As a result, the bottleneck of certified training progressively shifts from learning robust models towards formally proving the robustness exhibited by these models.

Our findings suggest several directions for future work. Firstly, certified training objectives and architectures could be designed not only to increase worst-case margins individually, but also to preserve tight bound propagation as robustness extends to more challenging inputs. Secondly, scaling behaviour should become an explicit criterion when evaluating certified training methods, since the ability to exploit additional data, compute, and capacity is itself an important property that current methods do not exhibit. Lastly, advances in complete verification will be necessary to keep pace with increasingly strong certifiably trained models. Alternatively, our results could provide motivation to revisit the original motivation of IBP-based training and to focus on learning models whose robustness can be certified efficiently using inexpensive incomplete verifiers.

6 AI USE STATEMENT

We used elucidating diffusion models (EDMs) to generate the class-conditional synthetic CIFAR-10 and TinyImageNet datasets central to this work, and discriminator-guided EDMs to generate data for an additional CIFAR-10 experiment. In addition, we used LLMs to assist with the implementation of experimental code, to proofread and improve manuscript wording and to assist with the generation of figures. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, propose hypotheses, design experiments, translate material, clean or reformat datasets, or interpret results; tasks involving proofs and qualitative or thematic analysis were not applicable to this work. The authors reviewed all AI-assisted text, code, and figures, checked them against the underlying experiments and source materials, and verified that the manuscript accurately reports the methods and results. We take full responsibility for the final content of this work.

7 ETHICS STATEMENT

Our work reveals an important limitation of state-of-the-art certified training methods, *i.e.*, robust overfitting, and demonstrates that scaling data effectively mitigates this issue. In addition, we show how additional compute scaling enables further performance improvements and identify more fundamental limits of IBP-based certified training beyond overfitting. With this, we provide important guidance on how to obtain best-performing classifiers with provable guarantees that may prove invaluable in safety-critical applications, and additionally, provide important guidance to the robustness and neural network verification communities. Nevertheless, our results only apply to the evaluated model architectures, datasets, and bounded ℓ_∞ perturbations; they do not establish safety under other attacks, distribution shifts, or deployment conditions. Therefore, it is crucial that practitioners conduct a scaling study similar to the one presented here on their specific threats and datasets before deploying certifiably robust models. In addition, the generation of synthetic data, extended training schedules and complete verification consume substantial computing resources.

8 REPRODUCIBILITY STATEMENT

Our experiments relied on the open-source datasets CIFAR-10 and TinyImageNet and built on the open-source software CTRAIN 0.4.2 and PyTorch 2.11.0. The experiments were run on NVIDIA H100 GPUs; complete verification used α,β-CROWN (commit `6b8bbc`) with a 300 s per-instance cutoff. We give additional details in Appendix B. All code needed to reproduce the experiments and detailed configurations are available at `https://anonymous.4open.science/r/scaling_certified_training_uncovers_limits-B207/`.

REFERENCES

Altstidl, T., Dobre, D., Kosmala, A., Eskofier, B. M., Gidel, G., and Schwinn, L. (2024). On the scalability of certified adversarial robustness with generated data. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages 1–24.

Bartoldson, B. R., Diffenderfer, J., Parasyris, K., and Kailkhura, B. (2024). Adversarial robustness limits via scaling-law and human-alignment studies. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, pages 3046–3072.

Bunel, R., Lu, J., Turkaslan, I., Torr, P. H., Kohli, P., and Kumar, M. P. (2020). Branch and bound for piecewise linear neural network verification. *Journal of Machine Learning Research*, 21(42):1–39.

Croce, F. and Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119, pages 2206–2216.

De Palma, A., Behl, H. S., Bunel, R., Torr, P. H. S., and Kumar, M. P. (2024a). Scaling the convex barrier with sparse dual algorithms. *Journal of Machine Learning Research*, 25(61):1–51.

De Palma, A., Bunel, R., Dvijotham, K. D., Kumar, M. P., Stanforth, R., and Lomuscio, A. (2024b). Expressive losses for verified robustness via convex combinations. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, pages 1–28.

Duong, H., Nguyen, T., and Dwyer, M. B. (2025). Neuralsat: A high-performance verification tool for deep neural networks. In *Proceedings of the 37th International Conference on Computer Aided Verification (CAV 2025)*, pages 409–423.

Ferrari, C., Mueller, M. N., Jovanović, N., and Vechev, M. (2022). Complete verification via multi-neuron relaxation guided branch-and-bound. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, pages 1–15.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations, (ICLR 2015)*, pages 1–11.

Gowal, S., Dvijotham, K. D., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. (2019). Scalable verified training for provably robust image classification. In *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pages 4842–4851.

Gowal, S., Rebuffi, S., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. (2021). Improving robustness using generated data. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 4218–4233.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. (2019). Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 15637–15648.

Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *CoRR*, abs/1712.00409.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1–12.

Hu, K., Zou, A., Wang, Z., Leino, K., and Fredrikson, M. (2023). Unlocking deterministic robustness certification on imagenet. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 1–19.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, volume 37, pages 448–456.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 1–13.

Kaulen, K. and Hoos, H. (2025). CTRAIN - a training library for certifiably robust neural networks (extended abstract). In *Proceedings of the 2nd International Symposium on AI Verification (SAIV 2025)*, pages 238–249.

Kaulen, K., Ladner, T., Bak, S., Brix, C., Duong, H., Flinkow, T., Johnson, T. T., Koller, L., Manino, E., Nguyen, T. H., et al. (2025). The 6th international verification of neural networks competition (VNN-COMP 2025): Summary and results. *arXiv preprint arXiv:2512.19007*.

Kaulen, K., Shavit, H., and Hoos, H. H. (2026). Rethinking evaluation paradigms in ibp-based certified training. In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*, pages 1–38.

Kim, D., Kim, Y., Kwon, S. J., Kang, W., and Moon, I. (2023). Refining generative process with discriminator guidance in score-based diffusion models. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, pages 16567–16598.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*.

König, M., Bosman, A. W., Hoos, H. H., and van Rijn, J. N. (2024). Critically assessing the state of the art in neural network verification. *Journal of Machine Learning Research*, 25(12):1–53.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Le, Y. and Yang, X. S. (2015). Tiny imagenet visual recognition challenge.

Li, L., Xie, T., and Li, B. (2023). SoK: Certified robustness for deep neural networks. In *Proceedings of the 44th IEEE Symposium on Security and Privacy (S&P 2023)*, pages 1289–1310. IEEE.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*, pages 1–23.

Mao, Y., Balauca, S., and Vechev, M. (2025). CTBENCH: A library and benchmark for certified training. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, pages 1–12.

Mao, Y., Müller, M. N., Fischer, M., and Vechev, M. T. (2023). Connecting certified and adversarial training. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 1–19.

Mao, Y., Müller, M. N., Fischer, M., and Vechev, M. T. (2024). Understanding certified training with interval bound propagation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*, pages 1–23.

Mirman, M., Gehr, T., and Vechev, M. T. (2018). Differentiable abstract interpretation for provably robust neural networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pages 3575–3583. PMLR.

Müller, M. N., Eckert, F., Fischer, M., and Vechev, M. T. (2023). Certified training: Small boxes are all you need. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, pages 1–21.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 1–12.

Rebuffi, S., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. (2021). Data augmentation can improve robustness. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 29935–29948.

Rice, L., Wong, E., and Kolter, J. Z. (2020). Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 8093–8104.

Sälzer, M. and Lange, M. (2021). Reachability is NP-complete even for the simplest neural networks. In *Proceedings of the 15th International Conference on Reachability Problems (RP 2021)*, pages 149–164.

Shi, Z., Wang, Y., Zhang, H., Yi, J., and Hsieh, C. (2021). Fast certified robust training with short warmup. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 18335–18349.

Singh, G., Gehr, T., Püschel, M., and Vechev, M. (2019). An abstract domain for certifying neural networks. In *Proceedings of the 3rd ACM on Programming Languages (POPL 2019)*, pages 1–30.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, pages 1–36.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*, pages 1–10.

Wang, S., Zhang, H., Xu, K., Lin, X., Jana, S., Hsieh, C.-J., and Kolter, J. Z. (2021). Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 29909–29921.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. (2023). Better diffusion models further improve adversarial training. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, pages 36246–36263.

Wong, E. and Kolter, Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pages 5286–5295. PMLR.

Xu, K., Shi, Z., Zhang, H., Wang, Y., Chang, K., Huang, M., Kailkhura, B., Lin, X., and Hsieh, C. (2020). Automatic perturbation analysis for scalable certified robustness and beyond. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1–13.

Xu, K., Zhang, H., Wang, S., Wang, Y., Jana, S., Lin, X., and Hsieh, C.-J. (2021). Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, pages 1–15.

Xu, X., Li, L., and Li, B. (2022). LOT: layer-wise orthogonal training on improving l2 certified robustness. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 1–12.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2022). Scaling vision transformers. In *Proceedings of the 35th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 12104–12113.

Zhang, B., Jiang, D., He, D., and Wang, L. (2022a). Boosting the certified robustness of l-infinity distance nets. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, pages 1–21.

Zhang, B., Jiang, D., He, D., and Wang, L. (2022b). Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 1–16.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. (2018). Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 4944–4953.

Zhang, R. and Sun, J. (2024). Certified robust accuracy of neural networks are bounded due to bayes errors. In *Proceedings of the 36th International Conference on Computer Aided Verification (CAV 2024)*, pages 352–376.

A Additional Background & Related Work

A.1 Interval Bound Propagation

Given a neural network $f_\theta : \mathbb{R}^d \mapsto \mathbb{R}^c$, $c, d \in \mathbb{N}$ that maps inputs $\mathbf{x} \in \mathbb{R}^d$ to outputs $f_\theta(\mathbf{x}) \in \mathbb{R}^c$, interval bound propagation (IBP) (Gowal et al., 2019) obtains sound lower bounds on the vector of logit differences $\mathbf{z}(\mathbf{x}, y) \in \mathbb{R}^c$ defined as $\mathbf{z}(\mathbf{x}, y) := f_\theta(\mathbf{x})[y] \cdot \mathbf{1} - f_\theta(\mathbf{x})$. Only coordinates corresponding to classes other than y enter the robustness condition; the y-coordinate is identically zero. For this, consider f_θ as the composition of L linear layers $h_{1,\dots,L}$ with $h_i(\mathbf{x}^{i-1}) = \mathbf{W}_i \cdot \mathbf{x}^{i-1} + \mathbf{b}_i$ and the ReLU activation $\sigma(\mathbf{x}) := \max(0, \mathbf{x})$, *i.e.*, $f_\theta = h_L \circ \sigma \circ h_{L-1} \circ \cdots \circ \sigma \circ h_1$. Using interval arithmetic, the axis-aligned box $\mathcal{B}_1$ that encompasses $h_1(\mathcal{B}^\epsilon_\infty)$, with $\mathcal{B}^\epsilon_\infty = \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}_0\|_\infty \leq \epsilon\}$, has centre $\overline{\mathbf{x}}_1 = \mathbf{W}_1\mathbf{x}_0 + \mathbf{b}_1$ and half-width $\boldsymbol{\delta}_1 = |\mathbf{W}_1|(\epsilon\mathbf{1})$. To approximate the reachable outputs of $\sigma(\mathcal{B}_1)$, due to the non-linearity of the ReLU function, lower and upper bounds have to be propagated separately, *i.e.*, $\mathbf{l}_2 = \sigma(\overline{\mathbf{x}}_1 - \boldsymbol{\delta}_1)$ and $\mathbf{u}_2 = \sigma(\overline{\mathbf{x}}_1 + \boldsymbol{\delta}_1)$. The resulting hyper-box $\mathcal{B}_2$ has centre $\overline{\mathbf{x}}_2 = \frac{\mathbf{u}_2 + \mathbf{l}_2}{2}$ and half-width $\boldsymbol{\delta}_2 = \frac{\mathbf{u}_2 - \mathbf{l}_2}{2}$. By continuing this process, we can compute a hyper-box that encompasses the reachable output set of f_θ, thereby allowing for the calculation of $\underline{\mathbf{z}}(\mathbf{x}, y)$.

A.2 Comparison to Altstidl et al. (2024)

Altstidl et al. (2024) also investigated scaling MTL-IBP on CIFAR-10 with $\epsilon = \frac{8}{255}$ using additional diffusion-generated data and increased compute. However, they reported decreases in both clean and certified accuracy relative to their unscaled baseline. In contrast, using the same synthetic data, we observe substantial improvements from scaling. To understand this discrepancy, we compared the experimental setups and identified two important differences in the training schedules. First, Altstidl et al. (2024) adopted the ImageNet-64 MTL-IBP configuration from De Palma et al. (2024b) for CIFAR-10, whereas we use the dataset-specific CIFAR-10 configuration. In particular, the CIFAR-10 configuration employs a longer ϵ ramp-up and performs the learning-rate decay later. Second, when increasing the compute budget, we proportionally extend the ramp-up schedule together with the overall training duration. A longer ramp-up phase can make the transition to the full perturbation radius more gradual, potentially improving optimization stability and final certified performance. These differences suggest that an appropriately scaled training schedule is important for obtaining gains from additional data and compute.

B Additional Details on the Setup of Experiments

In the following, we give additional details on the setup of our empirical scaling study. The code employed to conduct all presented experiments can be found at `https://anonymous.4open.science/r/scaling_certified_training_uncovers_limits-B207/`.

B.1 Hardware Details

All experiments were run on a compute cluster where each node is equipped with two Intel Xeon Platinum 8480+ CPUs with 210MB of L3 cache, four NVIDIA H100 SXM GPUs and 2TB of RAM running Rocky Linux 9. Each training and verification experiment utilised 14 CPU cores, 220GB of RAM and one GPU.

B.2 Datasets

In our scaling study, we employed commonly used vision datasets that are regularly used to evaluate certified training methods (see, *e.g.*, Shi et al., 2021; De Palma et al., 2024b; Müller et al., 2023; Mao et al., 2025). CIFAR-10 (Krizhevsky et al., 2009) consists of RGB images with three channels of size 32×32 pixels. Those are associated to 10 classes of simple objects such as airplanes or birds. Here, the training split includes $50\,000$ images while the test set includes $10\,000$ images. In addition, we utilised the TinyImageNet (Le and Yang, 2015) dataset, which is a subset of ImageNet including $100\,000$ train and $10\,000$ test images belonging to 200 classes. The images have three RGB colour channels and a resolution of 64×64 pixels. In line with previous research, we

normalised the datasets before training and used data augmentations (see, *e.g.*, Mao et al., 2025; De Palma et al., 2024b; Müller et al., 2023; Shi et al., 2021; Xu et al., 2020). More precisely, we augmented the datasets with random horizontal flips and random cropping to 32×32 pixels after padding by 2 pixels on CIFAR-10, and to 64×64 pixels after padding by 4 pixels on TinyImageNet. Importantly, these augmentations were also applied to the additional images generated by means of diffusion incorporated into the training process. Finally, we trained on the corresponding training sets, potentially augmented with additional data, and reported clean and certified accuracies on the respective test splits.

B.3 Software Versions

For training and certifying the scaled models, we built upon the open-source CTRAIN library (Kaulen and Hoos, 2025), which is available under a MIT license. We utilised version `0.4.2` of CTRAIN and adapted it to support the conducted scaling experiments. For verification, we utilised α,β-CROWN (Wang et al., 2021; Xu et al., 2021; Zhang et al., 2018) at commit `6b8bbc`, which is available under a permissive open-source license. Lastly, the underlying core library for training and verification was `torch` (Paszke et al., 2019) in version `2.11.0`.

B.4 Architectures

For our experiments, we used the CNN-7 architecture of Shi et al. (2021), who showed that the incorporation of BatchNorm layers Ioffe and Szegedy (2015) before every ReLU activation helps performance by reducing imbalance between active and inactive neurons. The model used for CIFAR-10 consists of two convolutional layers with 64 filters of size 3×3 each with a stride and padding of 1, followed by one convolutional layer with 128 filters of identical size and padding, but with a stride of 2. Thereafter follow two convolutional layers with, again, 128 filters and stride and padding of 1. Each of these layers is followed by a Batch normalisation layer and a ReLU activation. The classification head consists of a linear layer with 512 neurons, followed by Batch normalisation and a ReLU activation, before the final layer, which is a linear layer with size corresponding to the number of classes in the dataset. For TinyImageNet, we followed (De Palma et al., 2024b) and modified the last convolutional layer to also have stride 2, resulting in a reduced input to the classification head; in the experiments of De Palma et al. (2024b), this resulted in improved performance. Lastly, the wider variants of the two architectures used in our experiments were obtained by doubling the number of filters in all convolutional layers, *i.e.*, using 128 and 256 filters instead of 64 and 128. The number of neurons in the classification layer remained unchanged for these models.

B.5 Training Details

Generally, we followed the training setup of De Palma et al. (2024b) for our experiments, of which we give a detailed description in the following.

Initialisation. At the start of training, the network is initialised with the specialised procedure proposed by Shi et al. (2021), which relies on sampling parameters from a low-variance Gaussian distribution to prevent the explosion of IBP bounds in the early training stages.

Training Schedule. In the following, we give the training schedule of the baseline with unscaled compute, which we defined as training for 160 epochs. Notice that, once we scaled compute by a factor, the respective training phases were scaled accordingly in their length to maintain similar optimisation dynamics. Training begins with 1 warm-up epoch, where the model is trained on the clean cross-entropy loss. Thereafter, the radius ϵ used for the calculation of IBP bounds and the adversarial loss is gradually increased, beginning at 0, until it reaches its final value ϵ_{train}. This is done over 80 ramp-up epochs, using a smoothed annealing schedule which was found to produce state-of-the-art results (see, *e.g.*, Mao et al., 2025; De Palma et al., 2024b; Müller et al., 2023; Xu et al., 2020). In this annealing schedule, ϵ is first increased exponentially for 25% of epochs and linearly thereafter, which leads to smaller ϵ values at the beginning of training, thereby leading to improved training stability. Finally, training is conducted for the last 79 epochs over the full ϵ value.

Attack Strength. For the hyperparameters of the employed PGD attack, we followed the settings of De Palma et al. (2024b). These correspond to 8 optimisation steps with a step size of 0.25 for CIFAR-10, $\epsilon = \frac{2}{255}$, where the PGD attack is conducted over a larger perturbation radius scaled by

2.1 in order to put more emphasis on the adversarial loss. For the other settings, we employed a single-step attack with step size 10 over the standard, unscaled, perturbation radius, in accordance with De Palma et al. (2024b).

Regularisation. During the ramp-up phase, we employed the regulariser proposed by Shi et al. (2021), which penalises the explosion of IBP-bounds and balances inactive and active ReLU states. The magnitude of this regularisation is controlled by a weighing parameter λ and a decay factor $1 - \frac{\epsilon}{\epsilon_{\text{train}}}$. The regularisation is multiplied with both of these terms, which ensures that the regulariser is most prominently employed at the beginning of the training process, contributing to increased training stability. In all our experiments we chose $\lambda = 0.5$ on CIFAR-10 and $\lambda = 0.2$ for TinyImageNet consistent with prior work (De Palma et al., 2024b). In addition, we employed ℓ_1 regularisation, weighted by 3×10^{-6} for CIFAR-10, $\epsilon = \frac{2}{255}$, by 1×10^{-7} for CIFAR-10, $\epsilon = \frac{8}{255}$ and by 5×10^{-5} on TinyImageNet, in accordance with De Palma et al. (2024b). The regularisation term is calculated by exclusively considering weights in convolutional and linear layers.

Optimisation. For the optimisation of network parameters, we utilised Adam (Kingma and Ba, 2015) with a batch size of 128 and a learning rate of 0.0005. In related work, it has proven beneficial to decay the learning rate twice during training at the full ϵ-value (see, *e.g.*, Mao et al., 2025; De Palma et al., 2024b; Müller et al., 2023; Xu et al., 2020). Therefore, we also decayed the learning rate by a factor of 0.2 at epochs 120 and 140. Notice that these decay epochs were also scaled according to the compute scale used in a given experiment.

Batch normalisation layers. In the literature, there exist several options to set the statistics of BatchNorm layers. Shi et al. (2021) as well as Müller et al. (2023) learn these statistics to normalise batches exclusively based on unperturbed data. In contrast, De Palma et al. (2024b) exclusively consider adversarial examples found by PGD for the IBP bounds needed at training time and, at evaluation time, consider statistics over both, perturbed and unperturbed, inputs. Lastly, Mao et al. (2025) suggest to use statistics of clean, unperturbed data at training time. For the final networks, the authors set the statistics based on means and variances estimated over the whole training population. To date, it is not sufficiently clear whether any of these approaches leads to decisive performance improvements. Therefore, we adopted the setting of De Palma et al. (2024b), since we also used their proposed MTL-IBP loss in our experiments.

B.6 Complete Verification

We employed the verification system α,β-CROWN (Wang et al., 2021; Xu et al., 2021; Zhang et al., 2018) with a per-instance wall-clock timeout of 300 seconds to obtain certified guarantees for our scaled models, as in Section 3.3. While it is a well-known fact that α,β-CROWN requires careful hyperparameter tuning to achieve the strongest performance, we decided to use the system in its standard configuration, to facilitate an unbiased evaluation. For instance, a tuned version of α,β-CROWN might perform better on unscaled variants, which may lead to incorrect conclusions about scaling behaviour. We set the batch size of branch and bound domains to the highest number our hardware could accommodate, which resulted in batch sizes of 1024 on CIFAR-10 and 256 on TinyImageNet for the standard CNN-7 networks. For the wider network variant, we employed batch sizes of 512 on CIFAR-10 and 128 on TinyImageNet.

C Additional Scaling Results

This section tests whether the scaling pattern extends beyond MTL-IBP and examines sensitivity to the training seed and the quality of generated data. The complete-verification protocol is detailed in Appendix B.6; the transfer and data-quality comparisons below use incomplete certificates and should be read separately from the complete-verification results in the main text.

C.1 Overfitting in IBP and SABR

To show that, indeed, robust overfitting was less pronounced in traditional IBP training but worsened with the introduction of adversarial objectives, we repeated the study conducted in Section 3.1 for traditional IBP training with the improvement of (Shi et al., 2021). Furthermore, to show that severe robust overfitting is not unique to MTL-IBP but is an intrinsic property of expressive

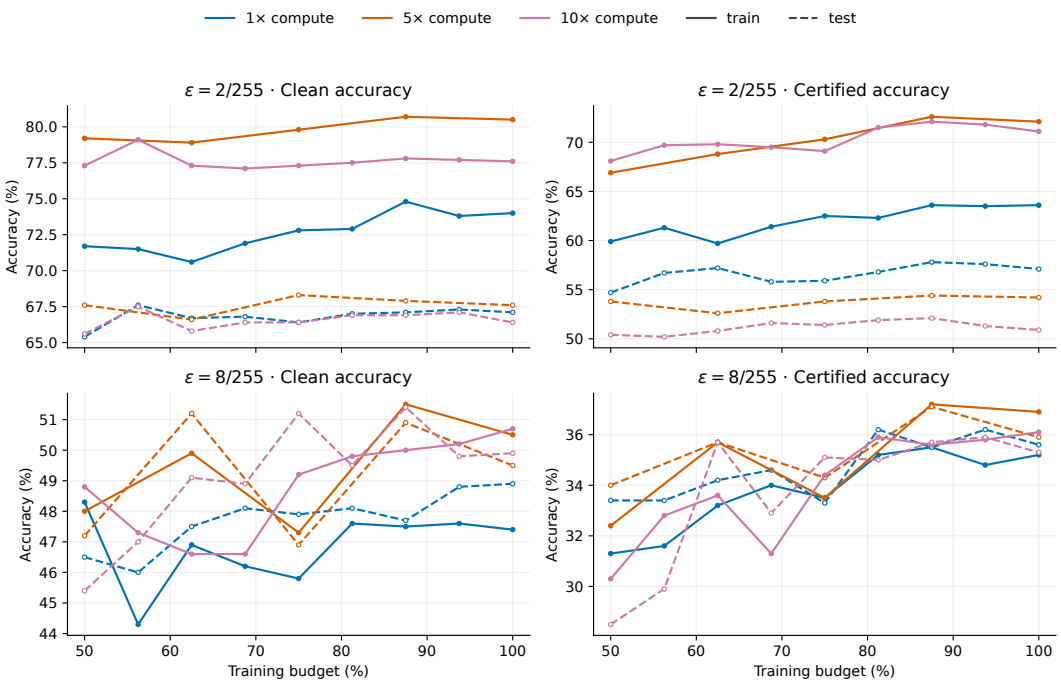


Figure 5: Shi-IBP train and test trajectories on CIFAR-10 with data scale $1\times$ and compute scales $1, 5, 10\times$ at $\epsilon \in \{2/255, 8/255\}$. Solid lines show training and dashed lines show test accuracy. We assessed overfitting over 1,000 examples drawn at random and α-CROWN certification to save computational resources.

losses (De Palma et al., 2024b), we also investigated the overfitting behaviour of SABR (Müller et al., 2023). The results of these experiments are displayed in Figures 5 and 6 for IBP and SABR, respectively.

While traditional IBP training also suffers from robust overfitting, this phenomenon is far less pronounced in comparison to MTL-IBP. While certified accuracies on the train set also improved with increased compute, this led to a reduction in certified accuracy of 6.2%. At $\epsilon = 2/255$, the final certified accuracy gaps for Shi-IBP are 6.5, 17.9, and 20.2 percentage points at compute scales $1, 5, 10\times$, versus 9.3, 19.9, and 22.7 points for the corresponding MTL-IBP runs. Regarding clean accuracies, overfitting is less severe with gaps only up to 12.9% in comparison to up to 19.3% for the hybrid method MTL-IBP.

SABR shows an even larger gap at $\epsilon = 2/255$: 22.5, 40.1, and 35.1 points at the same three compute scales, caused by increased certified performance on the train set and reduced performance on the test set with increased compute budgets. At the same time, while train set performance also increased for clean accuracy, test set clean performance remained largely unchanged. This shows the same qualitative pattern of severe robust overfitting also observed for MTL-IBP.

C.2 Transferability between Training Methods

Due to recent benchmarking studies identifying MTL-IBP as a state-of-the-art method in terms of certified accuracy (Kaulen et al., 2026; Mao et al., 2025), we focused our main study on this method. However, to test whether the observed scaling trends extend beyond MTL-IBP, we additionally evaluated the same data and compute grid for SABR and traditional Shi-IBP training. Figures 7 and 8 show clean and complete certified accuracies for CIFAR-10 at $\epsilon \in \{2/255, 8/255\}$, using all 10,000 test inputs per model and a 60 s per-instance verification cutoff time to keep computational demands manageable. Notice, that the main MTL-IBP results use a 300 s timeout, so the absolute complete certified accuracies are not directly comparable across these budgets, but suffice to identify common trends. At $\epsilon = 2/255$, SABR complete certified accuracy rises from 59.83% at the unscaled

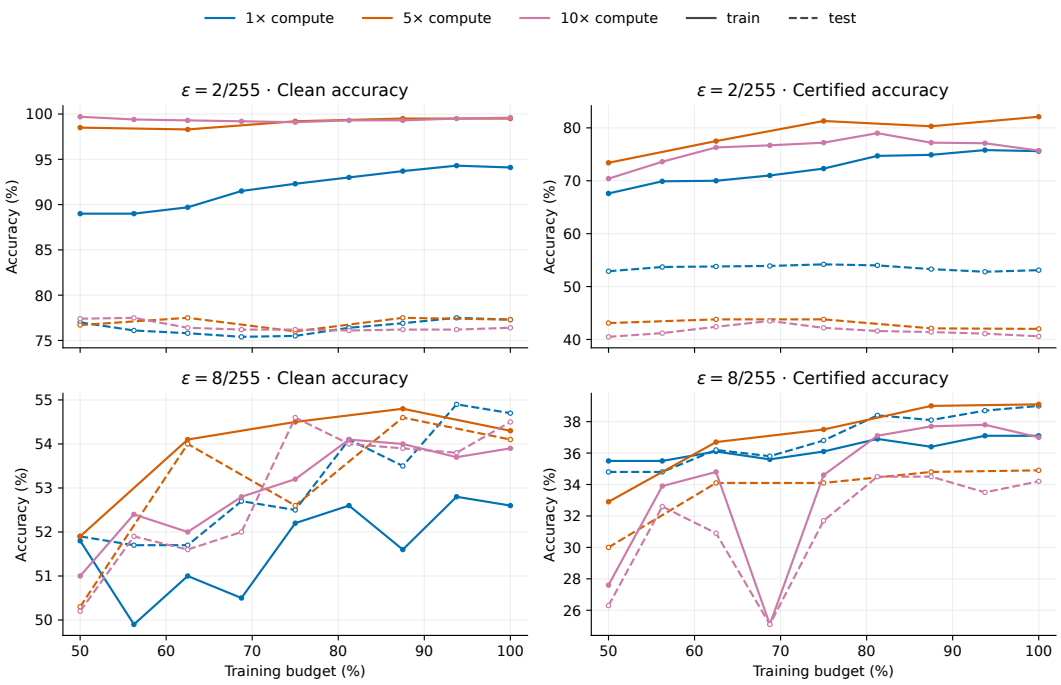


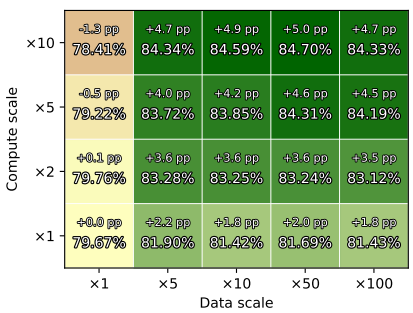
Figure 6: SABR train and test trajectories on CIFAR-10 with data scale $1\times$ and compute scales $1, 5, 10\times$ at $\epsilon \in \{2/255, 8/255\}$. Solid lines show training and dashed lines show test accuracy. We assessed overfitting over 1,000 examples drawn at random and α-CROWN certification to save computational resources.

baseline to a maximum of 66.97% with $50\times$ data and $10\times$ compute. At the same time, the model trained with the maximum scaling settings $100\times$ data and $10\times$ compute model reaches 66.62%. These results exceed those obtained for MTL-IBP, even for this smaller timeout, indicating that SABR might benefit even more from scaling at this benchmark. The performance of traditional IBP training rises from 56.26% to 60.40% at $100\times$ data and $10\times$ compute. At $\epsilon = 8/255$, the corresponding best complete certified accuracies are 35.73% for SABR and 35.45% for Shi-IBP, up from baselines of 35.11% and 33.96%. Thus, at this larger perturbation radius, SABR cannot outperform MTL-IBP and all methods seem to converge to roughly the same performance, which is consistent with recent benchmarking studies (Kaulen et al., 2026; Mao et al., 2025).

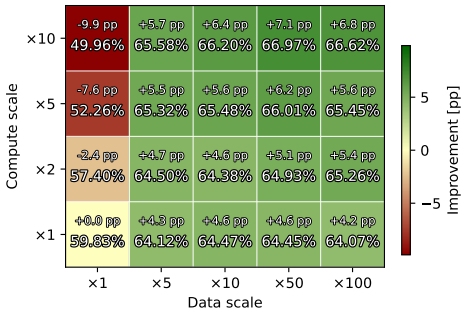
C.3 Variance of Results

To ensure that the reported performance gains are not due to statistical variance, we repeated the CIFAR-10 CNN-7 scaling grid for three pseudo-random number seeds (42, 69, and 420) at each perturbation radius. Figure 9 shows the mean and standard deviation across those seeds for clean and incomplete certified accuracy. Colour indicates the variance in squared percentage points. Overall, we observed variances similar to those observed in recent benchmarking studies (Mao et al., 2025; Kaulen et al., 2026). Crucially, the performance gains of scaling substantially exceed seed-level variance and are, thus, not a product of randomness.

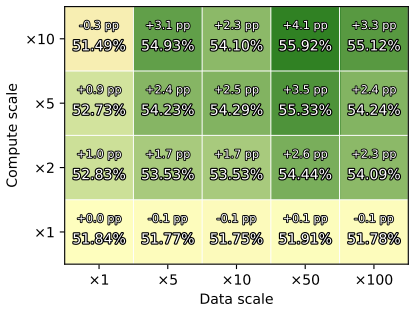
To give even more concrete evidence that the performance gains hold over multiple seeds, we repeated complete verification over three seeds for the best-performing models. At $\epsilon = 2/255$, $100\times$ data and $10\times$ compute this yielded complete certified accuracies of 65.90%, 65.79%, and 65.90% (mean $65.86\% \pm 0.05$ percentage points) and timeout rates of 6.72%, 7.01%, and 6.97%. At $\epsilon = 8/255$, the best-performing $50\times$ data/$10\times$ compute configuration yielded 35.58%, 35.37%, and 35.42% (mean $35.46\% \pm 0.09$ points) and timeout rates of 1.25%, 1.48%, and 1.49%.



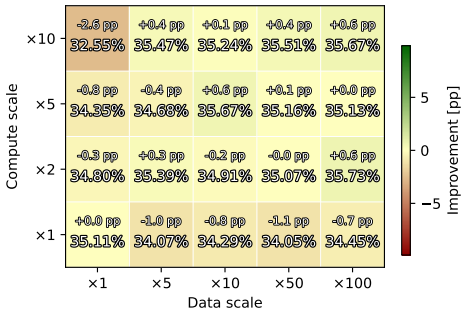
(a) Clean Acc., $\epsilon = \frac{2}{255}$



(b) Complete Cert. Acc., $\epsilon = \frac{2}{255}$



(c) Clean Acc., $\epsilon = \frac{8}{255}$



(d) Complete Cert. Acc., $\epsilon = \frac{8}{255}$

Figure 7: SABR on CIFAR-10 with CNN-7 at $\epsilon \in \{2/255, 8/255\}$. Cells show clean accuracy (left) and complete certified accuracy with a 60 s per-input verification timeout (right). The smaller annotation gives the change from the unscaled baseline. Data factors are $\{1, 5, 10, 50, 100\}$ and compute factors are $\{1, 2, 5, 10\}$.

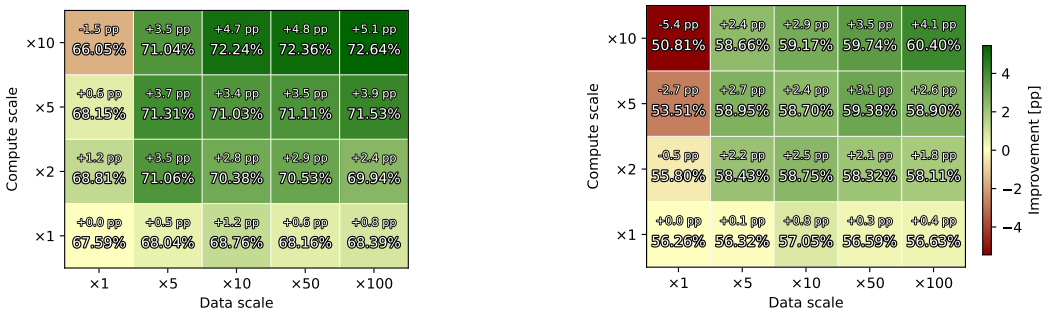
C.4 Generated Data Quality

Motivated by a recent scaling study in adversarial training by Bartoldson et al. (2024) that showed that adversarial training benefits from better data quality, we compared the original class-conditional EDM source with a discriminator-guided diffusion model (Kim et al., 2023) that achieves state-of-the-art data quality on CIFAR-10.

We measured image quality using the official EDM Inception detector and FID implementation, with the real CIFAR-10 training split as reference and class-balanced sampling without replacement (seed 0). The guided images were produced from a conditional CIFAR-10 EDM with discriminator and classifier checkpoints, 20 Heun steps, noise range $[0.002, 80]$, $\rho = 7$, first-order guidance weight 1, second-order weight 0, guidance interval $[0.01, 1]$, and boosting. For 50,000 generated images, FID was 1.7974 for the original EDM model and 1.6634 for the guided EDM source; over the respective full five million generated images, it was 1.3202 and 1.1523, respectively. These results are consistent with those reported by Bartoldson et al. (2024) and indicate that the EDM model with discriminator guidance produces synthetic data of substantially better quality.

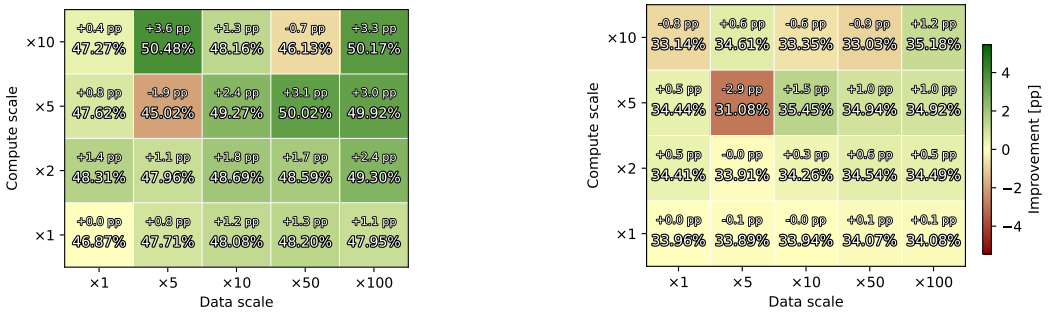
For each radius, $\epsilon \in \{2/255, 8/255\}$, we used data factors $\{5, 100\}\times$ and compute factors $\{1, 2, 10\}\times$, a 30% real-data fraction, and seed 42. We evaluated clean and incomplete certified accuracy using CROWN on all 10,000 test images. In Figures 10 and 11, we show performance of the models trained on the diffusion data with higher quality in terms of absolute accuracy and its paired change from the original source in percentage points.

At $\epsilon = 2/255$, the guided source increased adaptive incomplete certified accuracy by 0.46–1.53 percentage points across the paired configurations, while clean accuracy decreased by 0.17–0.88 points. This indicates that the increased data quality may lead to easier verification problems, that come at the cost of clean accuracy, and therefore the resulting networks are not Pareto-dominant with regard to both objectives (Kaulen et al., 2026). At $\epsilon = 8/255$, certified changes ranged from



(a) Clean Acc., $\epsilon = \frac{2}{255}$

(b) Complete Cert. Acc., $\epsilon = \frac{2}{255}$



(c) Clean Acc., $\epsilon = \frac{8}{255}$

(d) Complete Cert. Acc., $\epsilon = \frac{8}{255}$

Figure 8: Shi-IBP on CIFAR-10 with CNN-7 at $\epsilon \in \{2/255, 8/255\}$. Cells show clean accuracy (left) and complete certified accuracy with a 60 s per-input verification timeout (right). The smaller annotation gives the change from the unscaled baseline. Data factors are $\{1, 5, 10, 50, 100\}$ and compute factors are $\{1, 2, 5, 10\}$.

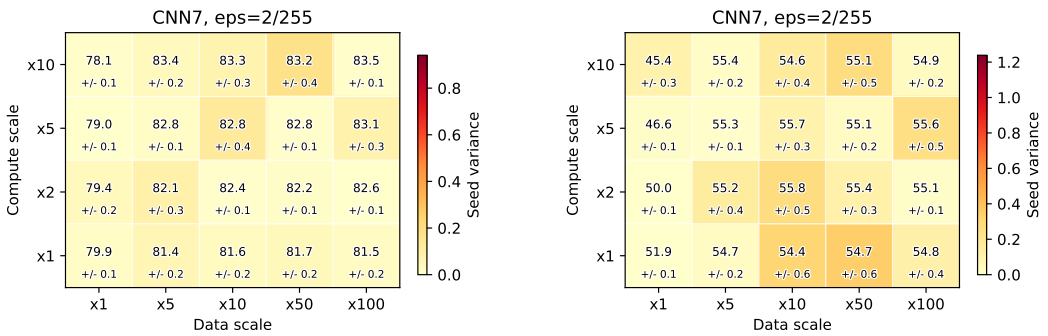
-0.46 to $+2.00$ points and clean changes had mixed signs. The higher-quality source therefore does not uniformly improve performance of scaled certified training models, but leads to mostly unchanged results. Therefore, the data quality yielded by standard EDM models seems sufficient to exploit gains yielded through improved generalisation.

C.5 Increased Verification Budgets

In order to investigate, whether the observed effects on verification hardness were due to the employed timeout of 300 seconds, we conducted additional experiments with increased verification timeouts of $1{,}000$ seconds and 6 hours. We focused on the baseline and width-scaled CNN-7 architectures on CIFAR-10 with $\epsilon = \frac{2}{255}$ since the complete verification bottleneck was most pronounced in this setting.

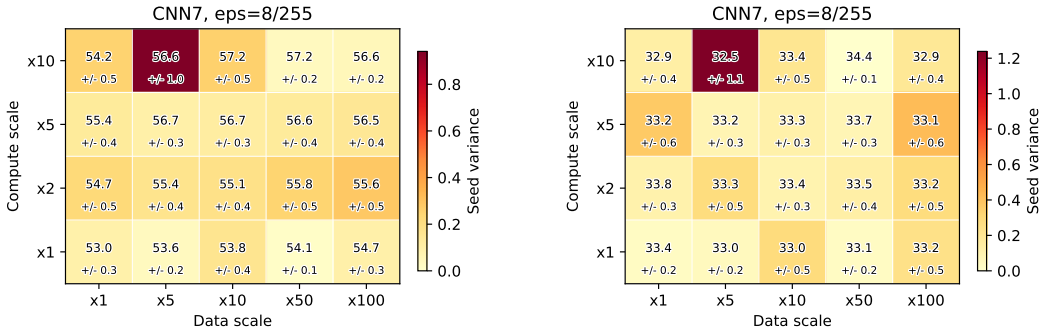
When increasing the timeout to $1\,000$ seconds at $\epsilon = 2/255$ with $10\times$ compute and $100\times$ data, α, β-CROWN certified $6\,604$ of the 10,000 test-set instances versus the previously proved $6\,590$ instances at a timeout of 300 seconds (corresponding to certified accuracies of 66.04% versus 65.90%). Thus, the increased running time budget realised 14 additional certifications, or 0.14 percentage points, and reduced timeouts from 672 to 658. However, mean running time increased from 23.10 to 69.61 seconds ($3.01\times$); total verification time increased from 64.16 to 193.35 GPU hours, which are an additional 129.19 hours. Similarly, for the wider CNN-7, the certified count rose from 5,995 to 6,028 (59.95% to 60.28%), a gain of 33 instances. Mean running time rose from 48.82 to 150.19 seconds ($3.08\times$); total verification time rose from 135.62 to 417.20 verifier-hours, an additional 281.58 hours. The longer timeout therefore yielded small certification gains at roughly three times the total verification cost.

We also extended verification to six hours for instances that still timed out after 1,000 seconds, among the first 50 test instances at $\epsilon = 2/255$ with $10\times$ compute and $100\times$ data. Of the 6 remaining



(a) Clean accuracy, $\epsilon = \frac{2}{255}$

(b) Incomplete certified accuracy, $\epsilon = \frac{2}{255}$



(c) Clean accuracy, $\epsilon = \frac{8}{255}$

(d) Incomplete certified accuracy, $\epsilon = \frac{8}{255}$

Figure 9: Seed variation across the CIFAR-10 CNN-7 data and compute scaling grid. Each cell reports mean $\pm$ standard deviation over three seeds. Colour indicates variance.

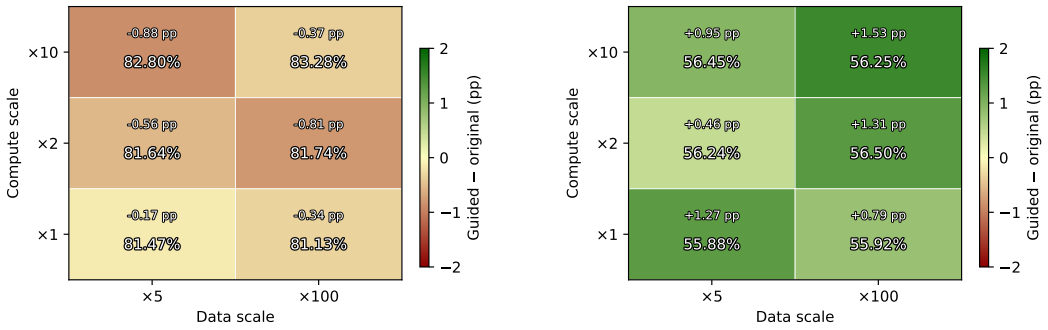


Figure 10: Original versus discriminator-guided synthetic data at $\epsilon = 2/255$: clean accuracy (left) incomplete certified accuracy obtained using CROWN (right). Annotations give the difference between the models trained on the data generated using the baseline EDM model and the improved EDM model with discriminator guidance.

timeouts in the baseline CNN-7 model that could not be solved within 1 000 seconds, all remained unsolved with a compute budget of 6 hours. For wide CNN-7, similarly 10 instances that could not be solved within 1 000 seconds could also not be solved within 6 hours. This indicates that the hard instances yielded through scaling require immense computational efforts such that current sate-of-the-art verification systems can yield a result. This implies that our analysis on instance hardness cannot be overcome by using slightly larger compute budgets for verification but, indeed, represents a more fundamental limit of best-performing certifiably trained models.

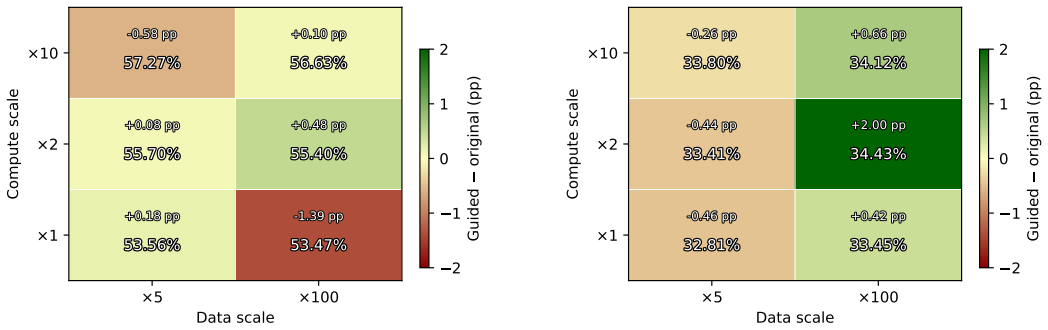


Figure 11: Original versus discriminator-guided synthetic data at $\epsilon = 8/255$: clean accuracy (left) incomplete certified accuracy obtained using CROWN (right). Annotations give the difference between the models trained on the data generated using the baseline EDM model and the improved EDM model with discriminator guidance.

D Additional Analysis

D.1 Generalisation

Robust overfitting across benchmarks. Figures 12 and 13 provide the checkpoint trajectories complementing the robust overfitting analysis conducted in Section 3.1. At $8/255$ on CIFAR-10, the gap between train and test set performance remained modest even for longer schedules for both clean and certified accuracy. This indicates that robust overfitting is not a big issue in this larger perturbation regime, which is consistent with the comparatively smaller gains yielded in our scaling study. However, for TinyImageNet, increasing compute without incorporating clearly shows robust overfitting at a compute scaling factor of $5\times$. Interestingly, doubling compute leads to improved performance, indicating underfitting after 160 epochs of training. When incorporating additional data, overfitting substantially reduced at the compute factor of 5.

Generalisation analysis. In the following, we provide additional details on the analysis conducted in Section 4.1. Again, we used 1,000 training and 1,000 test inputs and analysed certified accuracies computed using α-CROWN. In Figure 14, we display the generalisation gap between training and test performance over the full data-by-compute grid on CIFAR-10 for $\epsilon = \frac{2}{255}$. As discussed in Section 3.1, compute scaling without new data widens the certified gap. Data scaling closes it, including at the longer training schedules, with the gap approaching zero around a data factor of 50. In Figure 15, we show the corresponding generalisation gaps for $\epsilon = \frac{8}{255}$. The certified gap is much smaller at $8/255$ than at $2/255$ even before data scaling. The minimum is reached already by approximately $5\times$ data, versus approximately $50\times$ for $2/255$, consistent with the different scope for generalisation gains in the main results. When scaling compute in isolation by a factor of 2, this gap even reduces, which indicates underfitting in the unscaled baseline. Notably, for the best-performing data scaling factor of $50\times$, performance on the test set often exceeds training set performance. Finally, we display the absolute performance on the train and test sets over the scaling grids in Figure 16.

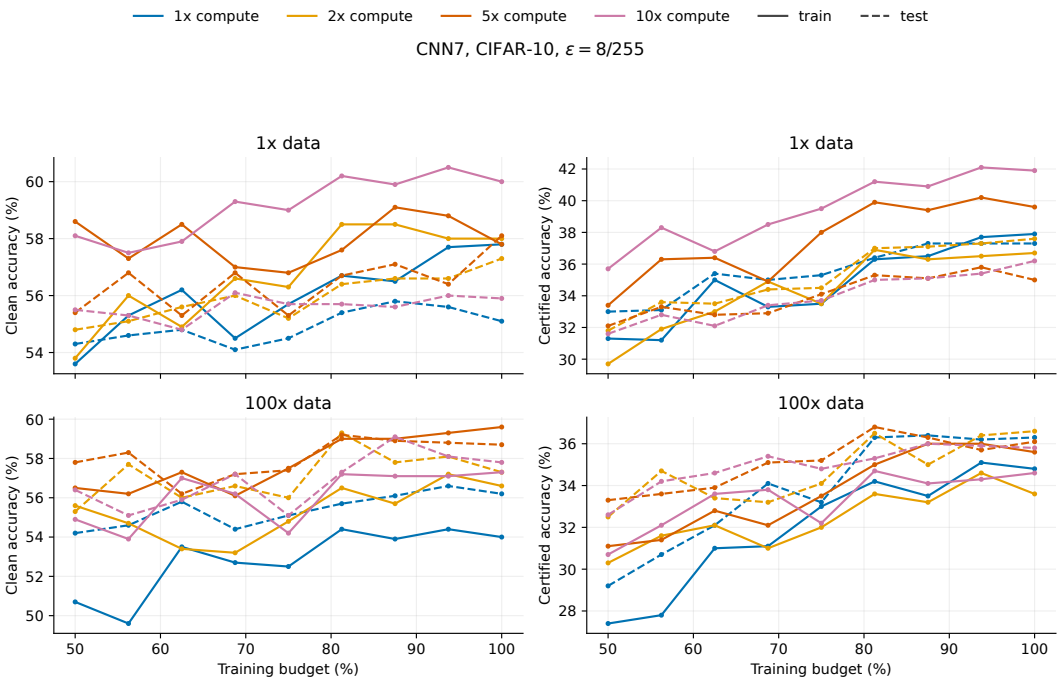


Figure 12: MTL-IBP train and test trajectories on CIFAR-10 at $\epsilon = 8/255$, with data factors $1, 100\times$ and compute factors $1, 2, 5, 10\times$. Clean accuracy is shown on the left and α-CROWN certified accuracy on the right. Solid lines are train and dashed lines are test. Each split uses a subset of $1\,000$ instances.

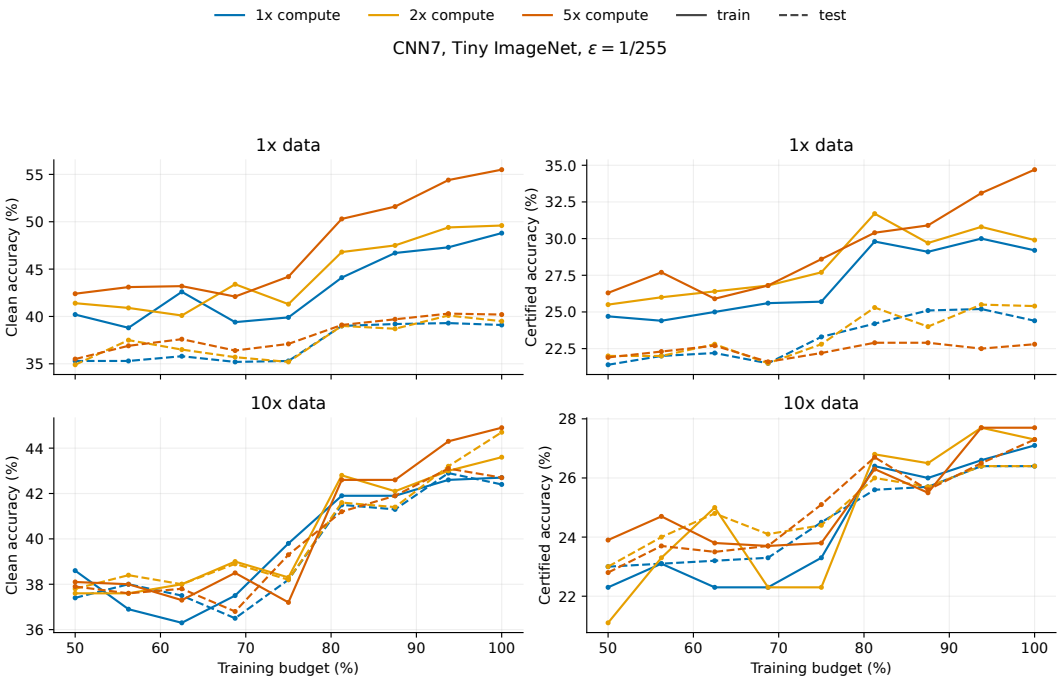


Figure 13: MTL-IBP train and test trajectories on TinyImageNet at $\epsilon = 1/255$, with data factors $1, 10\times$ and compute factors $1, 2, 5\times$. Clean accuracy is shown on the left and α-CROWN certified accuracy on the right. Solid lines are train and dashed lines are test. Each split uses a subset of $1\,000$ instances.

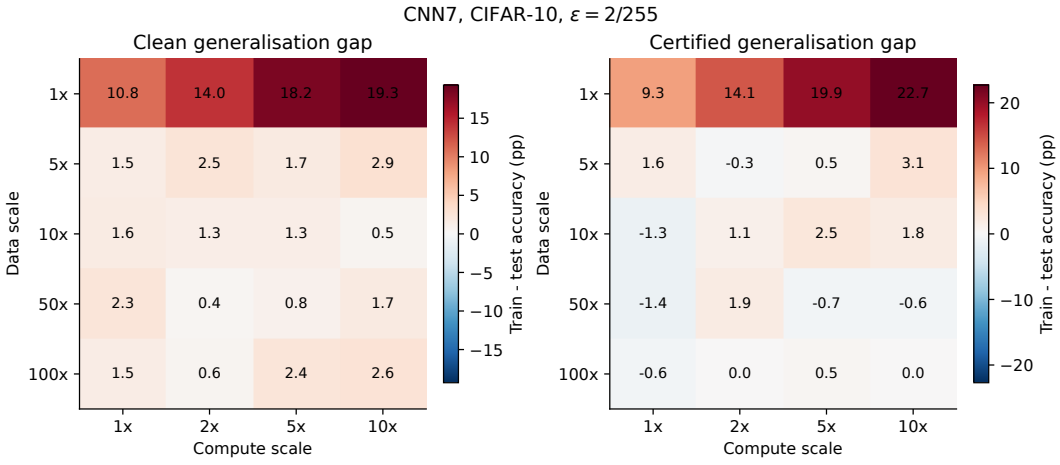


Figure 14: Generalisation gap (train minus test accuracy) of MTL-IBP on CIFAR-10 at $\epsilon = 2/255$ for CNN-7: clean accuracy (left) and α-CROWN certified accuracy (right) over the full scaling grid. Compute-only scaling widens the certified gap; data scaling closes it.

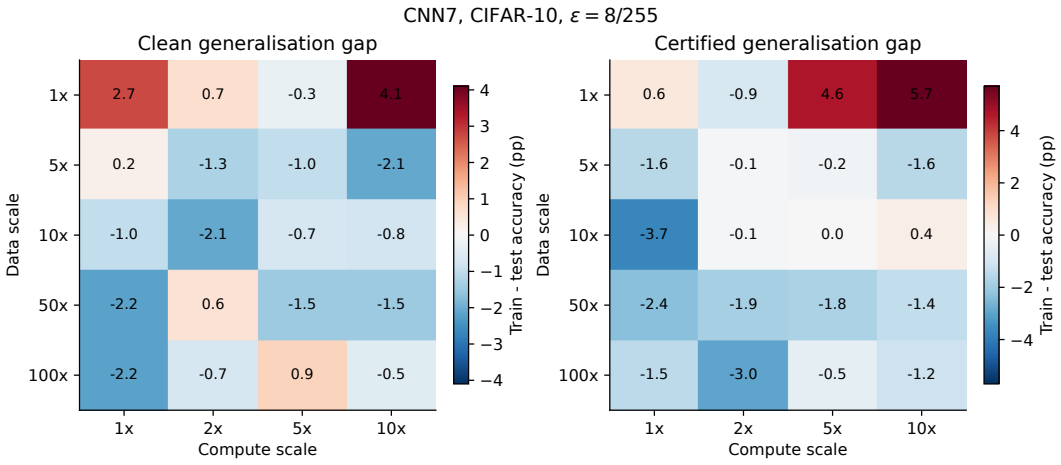
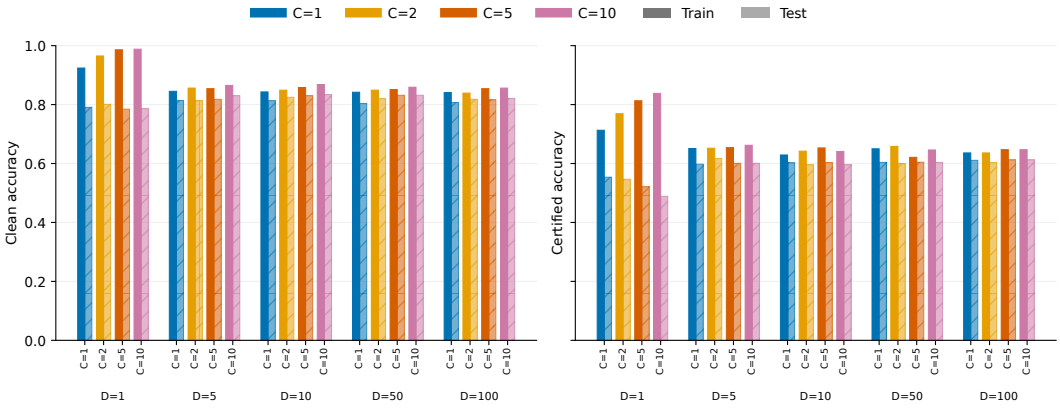
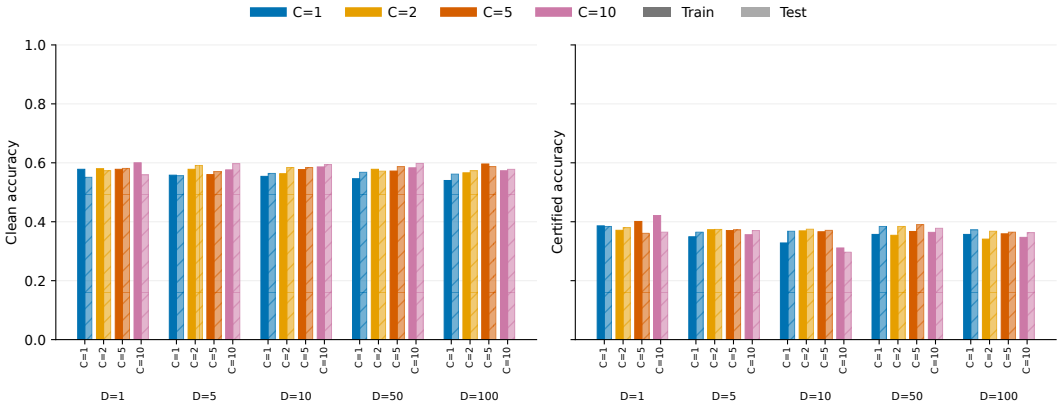


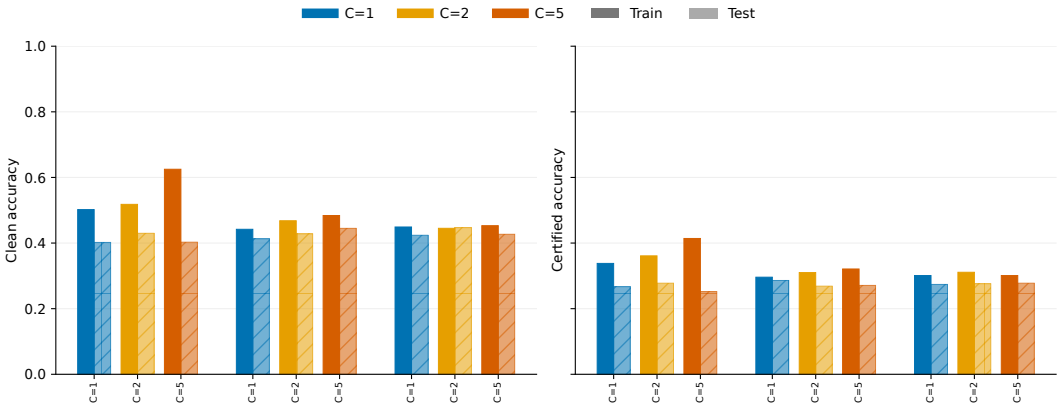
Figure 15: Generalisation gap (train minus test accuracy) of MTL-IBP on CIFAR-10 at $\epsilon = 8/255$ for CNN-7: clean accuracy (left) and α-CROWN certified accuracy (right) over the full scaling grid. The certified gap is small initially and can become negative at large data scale.



(a) CIFAR-10, $\epsilon = \frac{2}{255}$



(b) CIFAR-10, $\epsilon = \frac{8}{255}$



(c) Tiny ImageNet, $\epsilon = \frac{1}{255}$

Figure 16: Overfitting behaviour of the scaled models on CIFAR-10 and TinyImageNet in terms of clean and certified accuracy calculated over 1 000 instances. For verification, we used α-CROWN. Darker colors indicate training performance, while lighter colors indicate test performance.

D.2 UNSTABLE RELUS

In Figures 17 and 18, we display the approximation of unstable ReLUs in each layer for compute scaling factors $\{1, 2, 5, 10\}$ and data scaling factors $\{1, 100\}$. These heatmaps indicate that scaling compute leads to overall reduced fractions of unstable neurons. However, once additional data is incorporated, the structure and distribution of unstable ReLUs change throughout the network. Instability in the later layers decreases while earlier layers become more unstable. This is unexpected, since the accumulation of approximation errors usually results in particularly unstable classification heads. However, this observation indicates a potentially important characteristic of high-performance certifiably trained models, *i.e.*, more expressive feature extraction with increased instability combined with more robust classification on these features. Figure 19 shows the corresponding layer-wise pattern for wide CNN-7.

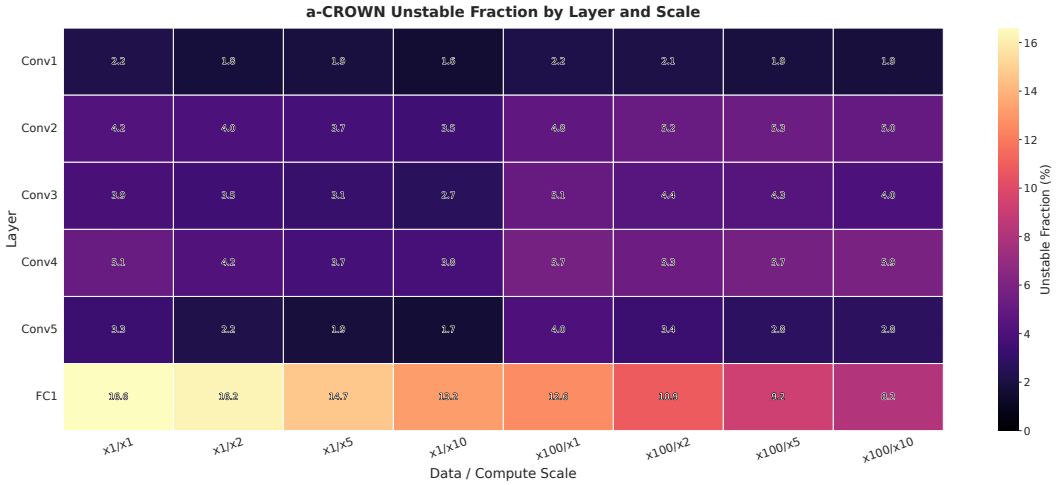


Figure 17: Fraction of unstable ReLU neurons for each layer for compute scales $\{1, 2, 5, 10\}$ and data scales $\{1, 100\}$ on CIFAR-10, $\epsilon = \frac{2}{255}$. We observed that scaling leads to fewer unstable neurons in later layers but an increased percentage in earlier layers. This indicates more expressive feature extraction and features facilitating more robust classification.

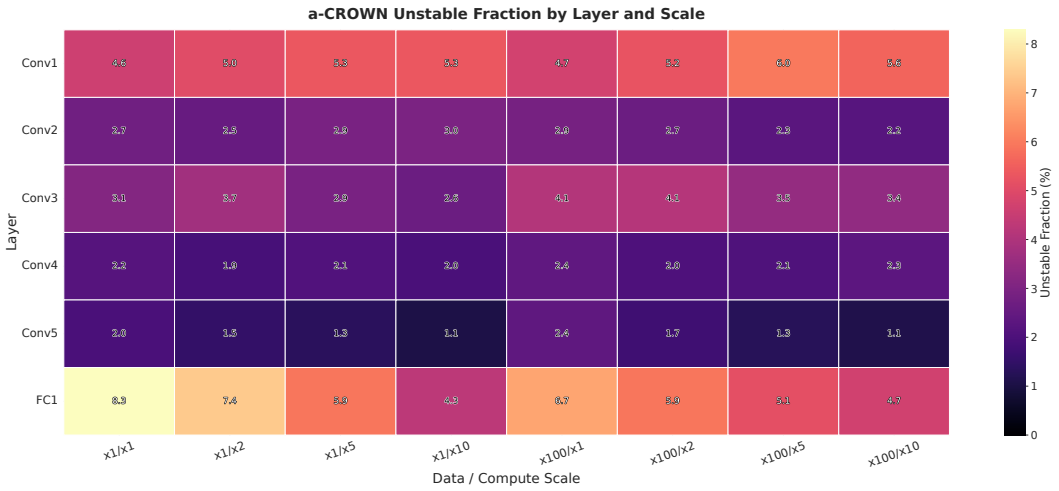
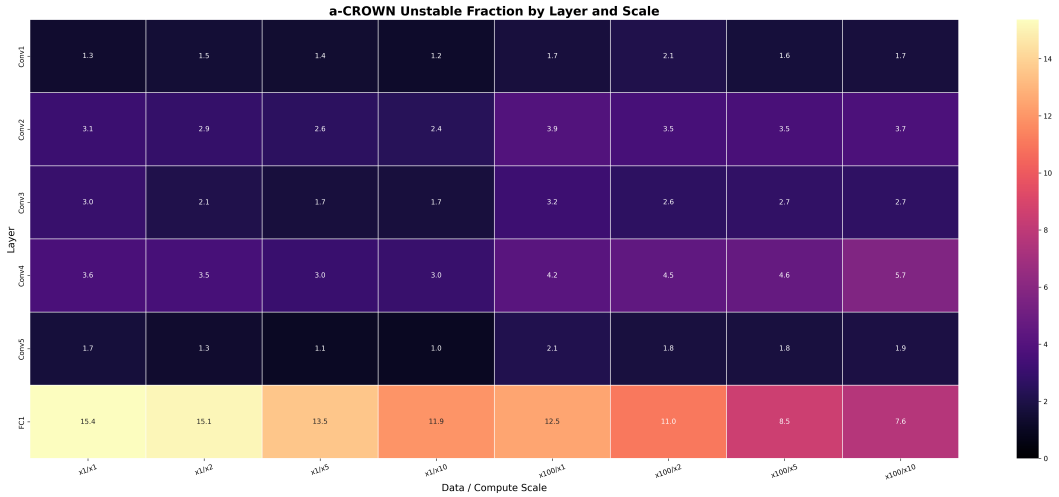
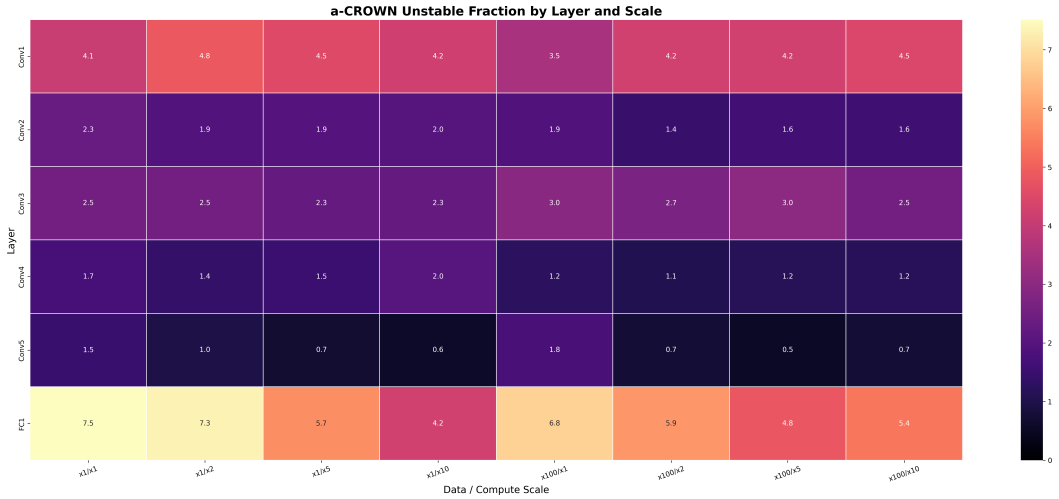


Figure 18: Fraction of unstable ReLU neurons for each layer for compute scales $\{1, 2, 5, 10\}$ and data scales $\{1, 100\}$ on CIFAR-10, $\epsilon = \frac{8}{255}$. Scaling yields fewer unstable neurons in later layers but a higher fraction in earlier layers, consistent with more expressive feature extraction and more stable classification.



(a) CIFAR-10 wide CNN-7, $\epsilon = 2/255$



(b) CIFAR-10 wide CNN-7, $\epsilon = 8/255$

Figure 19: Fraction of unstable ReLU neurons for each layer for compute scales $\{1, 2, 5, 10\}$ and data scales $\{1, 100\}$ on CIFAR-10 for the wider CNN-7 model. Scaling yields fewer unstable neurons in later layers but a higher fraction in earlier layers, consistent with more expressive feature extraction and more stable classification.

D.3 Propagation Tightness

In the following, we provide a more in-depth analysis on the propagation tightness of the scaled models and introduce an additional dimension for our investigation, namely the midpoint of the interval bounding the worst-case margin. As discussed in Section 4.3, a certificate in neural network verification requires a strictly positive lower bound on the worst-case margin between the correct class and all incorrect classes. Models can satisfy this condition with tighter bounds or with a larger interval midpoint that compensates for greater width. For an α-CROWN margin interval $[\underline{m}_\theta(\mathbf{x}, y), \overline{m}_\theta(\mathbf{x}, y)]$, write its midpoint as $c_\theta = (\underline{m}_\theta + \overline{m}_\theta)/2$ and its width as $w_\theta = \overline{m}_\theta - \underline{m}_\theta$, suppressing $(\mathbf{x}, y)$ on the right-hand sides. The final lower bound on the margins is $\underline{m}_\theta = c_\theta - w_\theta/2$, so this interval certifies robustness exactly when $c_\theta > w_\theta/2$.

The requirement $c_\theta > w_\theta/2$ exposes two options the learnt model can exploit to improve a certificate: reduce the interval width, or increase its midpoint enough to offset a larger width. Consequently, a wider interval does not necessarily imply a worse lower bound on the margin. For paired models on the same input, the change in lower bound is $\Delta\underline{m}_\theta = \Delta c_\theta - \Delta w_\theta/2$. This connects directly to the relaxation deficit $\max(0, -\underline{m}_\alpha)$ used in Section 4.3: raising a negative initial lower bound reduces the deficit, until it reaches zero and the initial, potentially loose, relaxation itself certifies the input.

Figures 20 and 21 compare the baseline with data-only and joint data-and-compute scaling, respectively, on all 10,000 CIFAR-10 test inputs for each architecture and radius. Each point gives $\Delta w_\theta/2$ horizontally and Δc_θ vertically; points above the diagonal thus have a higher α-CROWN lower bound and therefore a decreased relaxation deficit or a relaxation deficit of 0, when the instance was already verified. Table 1 reports the corresponding counts, including inputs whose deficit is already zero. Data-only scaling raises the lower bound for about half of inputs in every setting. Joint scaling raises it for 58.51% of CNN-7 inputs at $2/255$, but for only 37.43% and 31.99% of wide CNN-7 inputs at $2/255$ and $8/255$, respectively. Thus, the wide models have a lower initial bound on most paired inputs under joint scaling, although this need not increase their relaxation deficit on inputs that remain above zero.

Figure 22 isolates the effect of adding compute to the data-scaled model. At $2/255$, many CNN-7 points lie above the diagonal, whereas the wide-model cloud is shifted towards greater half-width and contains many points below it. At $8/255$, the changes are more dispersed, especially for wide CNN-7. These paired comparisons also differ from aggregate certified accuracy, which depends on how many bounds cross zero.

Figures 23 and 24 show the full width distributions for both network sizes and radii. Data-only scaling generally shifts the width distributions to smaller values. Adding compute shifts them back towards larger widths, with a particularly clear reversal for wide CNN-7 at $2/255$ where the scaled models yields looser bounds than the baseline.

Table 1: Number of the 10,000 CIFAR-10 test inputs whose α-CROWN lower bound increases relative to the baseline, with the percentage in parentheses. A higher lower bound leads to lower relaxation deficits, thereby making verification easier.

Model and radius	Baseline to data only	Baseline to data + compute
CNN-7, $\epsilon = 2/255$	5,517 (55.17%)	5,851 (58.51%)
CNN-7, $\epsilon = 8/255$	5,010 (50.10%)	5,037 (50.37%)
Wide CNN-7, $\epsilon = 2/255$	5,400 (54.00%)	3,743 (37.43%)
Wide CNN-7, $\epsilon = 8/255$	5,147 (51.47%)	3,199 (31.99%)

Figure 25 complements the paired plots with the distributions of width and absolute midpoint over the full data-by-compute grid. Compute scaling without added data markedly raises both quantities and broadens their tails, especially at $8/255$. Adding data reduces widths and midpoint magnitudes; at a fixed data scale, more compute can raise them again, most visibly for wide CNN-7. The fact that both quantities move together might indicate that the stronger models allow looser bounds to enable more expressive feature extraction, but try to counteract this by inducing larger margins.

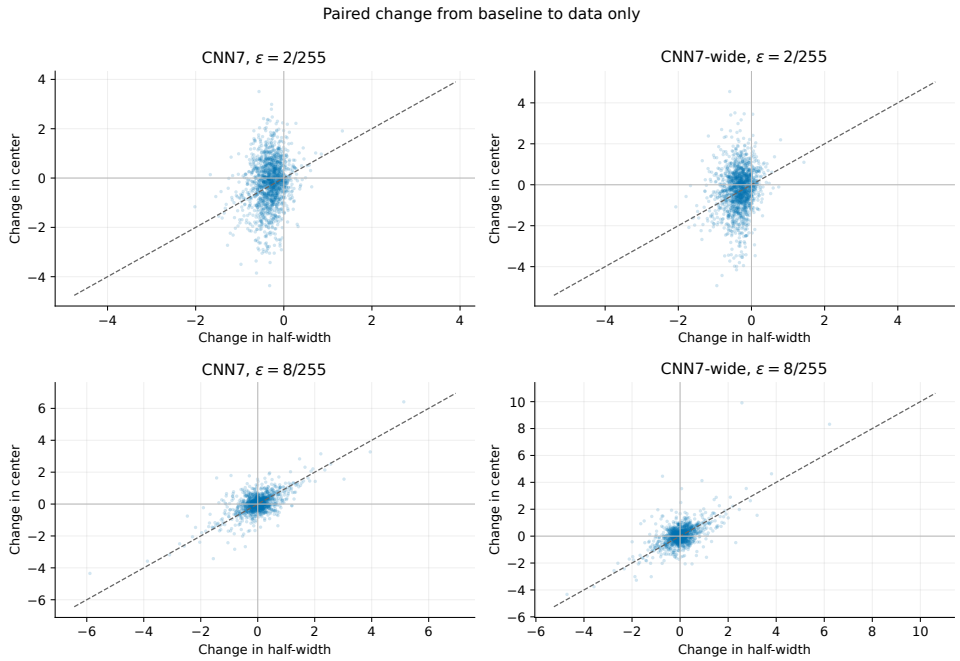


Figure 20: Paired change in the α-CROWN margin interval from the baseline to data-only scaling for CNN-7 and wide CNN-7 at both CIFAR-10 radii. Each point is one test input, with change in half-width on the horizontal axis and change in midpoint on the vertical axis. Points above the dashed diagonal have a larger propagated lower bound after scaling; points below it have a smaller lower bound.

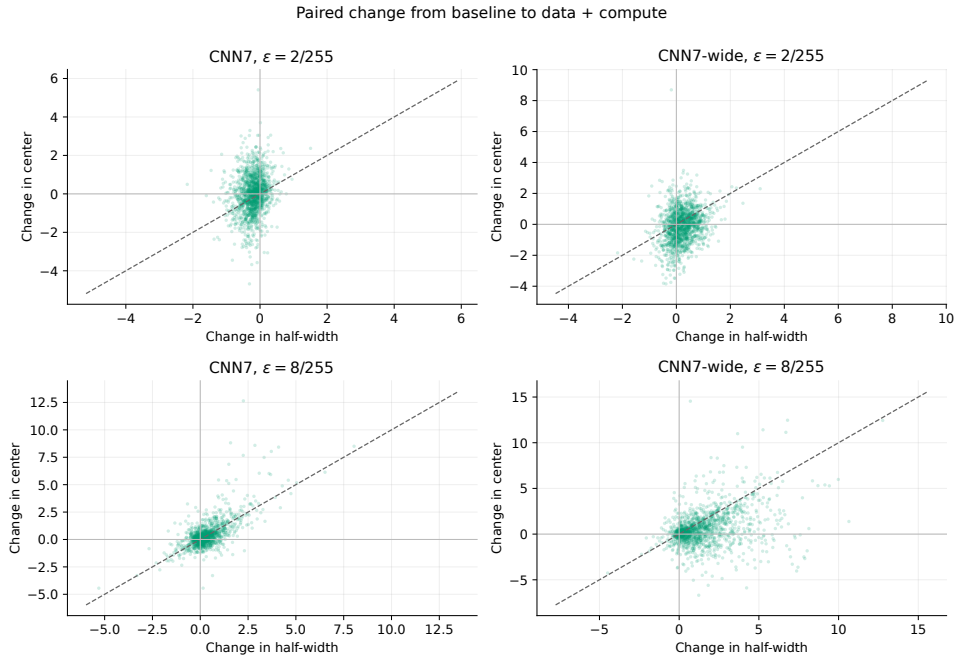


Figure 21: Paired change in the α-CROWN margin interval from the baseline to joint data and compute scaling for CNN-7 and wide CNN-7 at both CIFAR-10 radii. Each point is one test input. The dashed diagonal marks no change in the propagated lower bound: above it the midpoint gain exceeds the half-width gain, while below it the lower bound decreases.

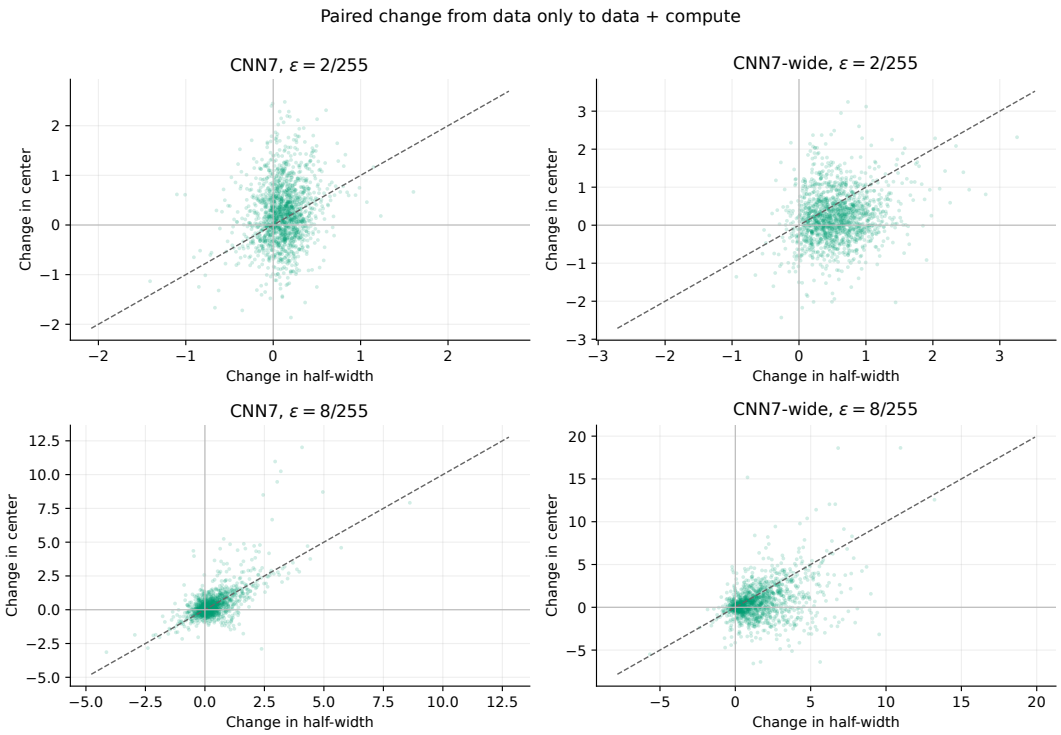


Figure 22: Paired change in the α-CROWN margin interval from data-only to joint data and compute scaling for the headline CIFAR-10 models. Each point is one input: change in half-width (x-axis) and midpoint (y-axis). Above the dashed diagonal, the resulting lower bound on the worst-case classification margin increases.

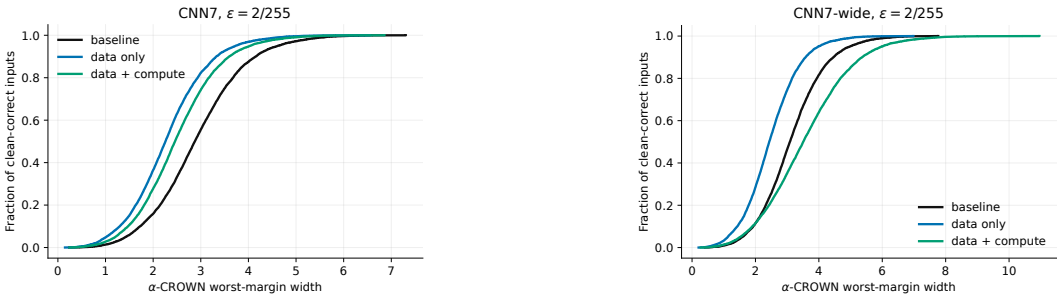


Figure 23: Distribution of α-CROWN worst-case margin widths at $\epsilon = 2/255$ for CNN-7 (left) and wide CNN-7 (right). Curves compare the baseline, data-only scaling, and joint data and compute scaling.

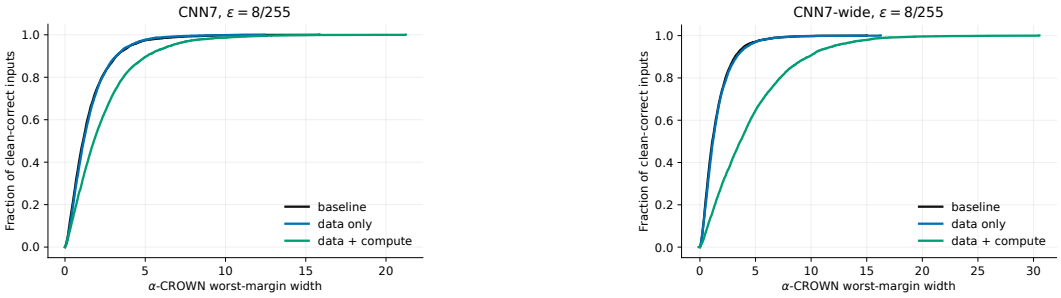
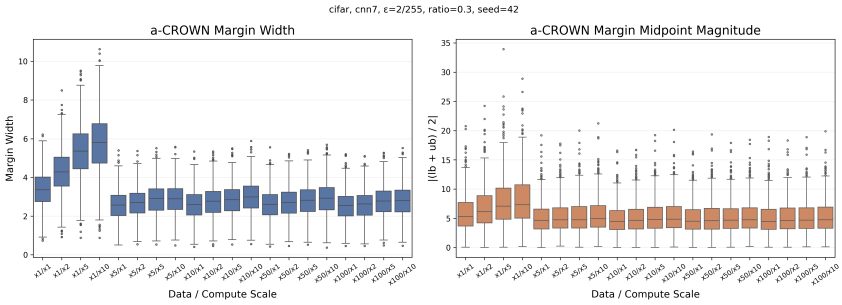
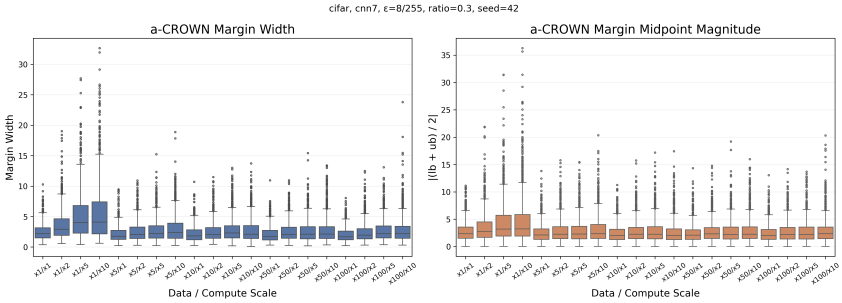


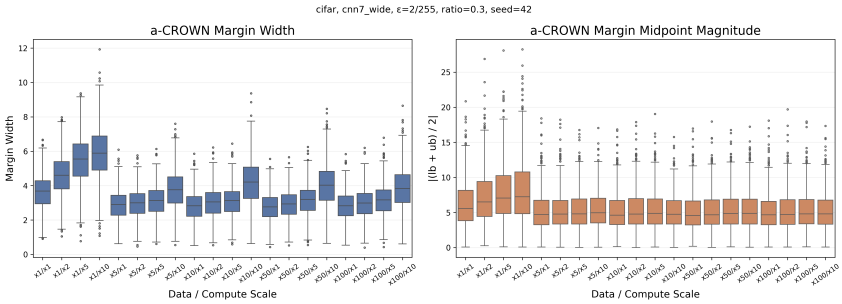
Figure 24: Distribution of α-CROWN worst-case margin widths at $\epsilon = 8/255$ for CNN-7 (left) and wide CNN-7 (right). Curves compare the baseline, data-only scaling, and joint data and compute scaling.



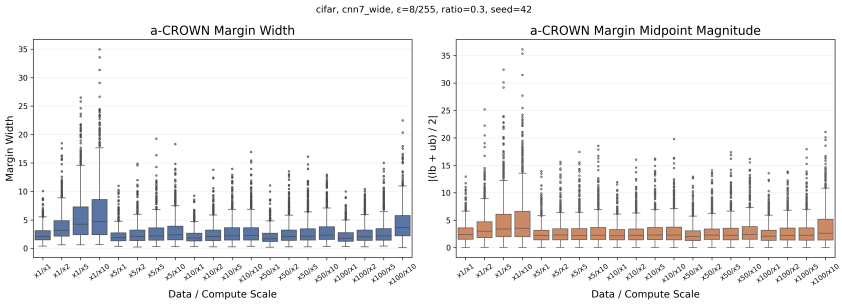
(a) CIFAR-10 CNN-7, $\epsilon = 2/255$



(b) CIFAR-10 CNN-7, $\epsilon = 8/255$



(c) CIFAR-10 wide CNN-7, $\epsilon = 2/255$



(d) CIFAR-10 wide CNN-7, $\epsilon = 8/255$

Figure 25: Distributions of α-CROWN worst-case margin width (left) and absolute midpoint (right) across data and compute scaling factors for CNN-7 and wide CNN-7 on CIFAR-10 at $\epsilon \in \{2/255, 8/255\}$. Each pair of factors on the horizontal axis gives data scale followed by compute scale. Compute-only scaling tends to increase margin width, which the models try to counteract by increasing margins as well. Additional data generally reduces interval width and leads to tighter bounds.

D.4 VERIFICATION HARDNESS

In the following, we further examine the reasons that drive verification hardness, extending the analysis in Section 4.3. Figure 26 summarises the complete-verification outcomes of the headline models, reporting the mean PAR-10 time together with the timeout rate for each scaling configuration. This shows that all scaling dimensions, *i.e.*, data, compute and capacity lead to increased numbers of unsolved instances and harder verification problems. To analyse how well instance features do predict the hardness of verification, we define the *hard pool* of each configuration as inputs that are correctly classified, are adversarially robust, and have a negative initial α-CROWN lower bound. These are precisely those inputs that require complete verification to obtain a definitive result.

As previously, the *relaxation deficit* is defined as $\max(0, -\underline{m}_\alpha)$, where $\underline{m}_\alpha$ is the initial α-CROWN lower bound on the worst-case margin. Within the hard pool at $\epsilon = 2/255$, its median is 0.51 for timeouts versus 0.16 for inputs ultimately proven robust in the scaled CNN-7. For the scaled wide CNN-7, the corresponding medians are 0.89 and 0.24. A larger initial deficit means branch and bound must recover more margin before it can certify the input. Figure 27 and Figure 28 show the distribution of the relaxation deficit for solved robust inputs (solid lines) and timeouts (dashed lines) in the hard pool. The separation between the distributions is clearly visible, and is especially pronounced in the wide-model timeout population at $2/255$. This might indicate looser initial bounds on difficult inputs.

Figure 29 reports, for each model, how well pre-computed properties of the initial relaxation predict whether complete verification of a hard-pool input will time out by varying a threshold over the feature that separates timeouts from solved instances. Relaxation deficit is the strongest single predictor in these comparisons; for the $2/255$ models its ROC-AUC is approximately 0.90–0.94. Estimated boundary distance, which we estimated as the distance from the input required by a strong PGD attack to cause a misclassification, also showed informative and, thus, also seemed predictive of timeouts.

To further investigate how individual instances changed that contributed to the increased portions of unsolved instances in terms of the predictive features of boundary instance and relaxation deficit, we paired the same CIFAR-10 inputs from the baseline to joint data-and-compute scaling, and analysed their verification outcomes. Unsolved instances rose from 551 to 672 for CNN-7 and from 640 to 1 468 for wide CNN-7 at $\epsilon = 2/255$; at $8/255$ they rise from 66 to 126 and from 81 to 302, respectively. Table 2 gives every transition into timeout of the data and compute-scaled model and every baseline timeout resolved as SAT or UNSAT, together with the corresponding empirical boundary distances and relaxation-deficit changes. Among the joint-model timeouts, only 120/672 (17.9%), 291/1 468 (19.8%), 2/126 (1.6%), and 4/302 (1.3%) persist from the baseline in these four settings. Conversely, 431/551 (78.2%), 349/640 (54.5%), 64/66 (97.0%), and 77/81 (95.1%) of baseline timeouts become SAT or UNSAT under joint scaling.

Across all four settings, SAT-to-timeout inputs move from median boundary distances of 0.59–0.70 to 1.06–1.47 times ϵ, while UNSAT-to-timeout inputs move from 1.88–2.35 to 0.93–1.78 times ϵ. Baseline timeouts that become SAT finish at median distances of 0.41–0.63 times ϵ; those that become UNSAT finish at 1.70–2.69 times ϵ. The paired relaxation-deficit changes in Table 2 help distinguish these outcome shifts from a change in boundary distance alone. This analysis supports our findings from Section 4.3 in showing that scaling moves prior SAT and prior UNSAT inputs toward a common ambiguity band near the estimated boundary, producing a largely new timeout population. This underlines that a fundamental limit of IBP-based certified training seems to be, that more performant models require a more refined decision function that, in turn, cannot accommodate large margins and tight bound propagation for all inputs.

Figures 30 and 31 illustrate the quantile behaviour of the boundary distance and the relaxation deficit on those inputs that were classified correctly by all three scaling configurations for each network and threat radius. The lower-tail panels focus on inputs close to the estimated robust boundary; the upper-tail panels focus on the largest initial relaxation deficits. Thereby we show those parts of the distribution that correlate most with instance hardness. These plots show that scaling moves more instances close to the decision boundary, but, overall, leads to tighter bounds on CNN-7. Once width is scaled, bound propagation tightness also deteriorates for the best-performing model with, both, scaled data and compute.

Table 2: Paired outcome transitions between the baseline and the jointly scaled model involving a timeout on the full 10 000-input CIFAR-10 test set. For transitions ending in timeout, shares are of joint-model timeouts; for baseline timeout-to-SAT/UNSAT transitions, shares are of baseline timeouts. Distances d are median adversarial radii divided by ϵ; Δd is the median paired change. ΔD is the median paired change in relaxation deficit. SAT means a counterexample was found and UNSAT means the input was proved robust.

Model and radius	Transition	Inputs (share)	Median $d_{\text{base}} \to d_{\text{joint}}$	Median Δd	Median ΔD
CNN-7, $2/255$	SAT→timeout	324 (48.2%)	$0.60 \to 1.26$	$+0.73$	-0.85
CNN-7, $2/255$	UNSAT→timeout	228 (33.9%)	$2.14 \to 1.33$	-0.79	$+0.33$
CNN-7, $2/255$	timeout→timeout	120 (17.9%)	$1.22 \to 1.36$	$+0.08$	-0.12
CNN-7, $2/255$	timeout→SAT	178 (32.3%)	$1.22 \to 0.63$	-0.60	$+0.36$
CNN-7, $2/255$	timeout→UNSAT	253 (45.9%)	$1.23 \to 2.36$	$+1.12$	-0.50
Wide CNN-7, $2/255$	SAT→timeout	556 (37.9%)	$0.59 \to 1.47$	$+0.94$	-0.46
Wide CNN-7, $2/255$	UNSAT→timeout	621 (42.3%)	$2.35 \to 1.78$	-0.56	$+0.68$
Wide CNN-7, $2/255$	timeout→timeout	291 (19.8%)	$1.28 \to 1.65$	$+0.33$	$+0.29$
Wide CNN-7, $2/255$	timeout→SAT	166 (25.9%)	$1.24 \to 0.62$	-0.64	$+0.93$
Wide CNN-7, $2/255$	timeout→UNSAT	183 (28.6%)	$1.24 \to 2.69$	$+1.45$	-0.49
CNN-7, $8/255$	SAT→timeout	51 (40.5%)	$0.70 \to 1.16$	$+0.58$	-0.23
CNN-7, $8/255$	UNSAT→timeout	73 (57.9%)	$1.88 \to 1.14$	-0.58	$+0.13$
CNN-7, $8/255$	timeout→timeout	2 (1.6%)	$1.03 \to 1.28$	$+0.25$	-0.05
CNN-7, $8/255$	timeout→SAT	36 (54.5%)	$1.05 \to 0.49$	-0.53	$+0.36$
CNN-7, $8/255$	timeout→UNSAT	28 (42.4%)	$1.09 \to 1.70$	$+0.27$	-0.08
Wide CNN-7, $8/255$	SAT→timeout	112 (37.1%)	$0.65 \to 1.06$	$+0.43$	-0.10
Wide CNN-7, $8/255$	UNSAT→timeout	186 (61.6%)	$2.33 \to 0.93$	-1.36	$+0.36$
Wide CNN-7, $8/255$	timeout→timeout	4 (1.3%)	$1.26 \to 1.15$	-0.11	$+0.14$
Wide CNN-7, $8/255$	timeout→SAT	54 (66.7%)	$1.04 \to 0.41$	-0.75	$+1.23$
Wide CNN-7, $8/255$	timeout→UNSAT	23 (28.4%)	$1.05 \to 1.78$	$+0.66$	-0.05

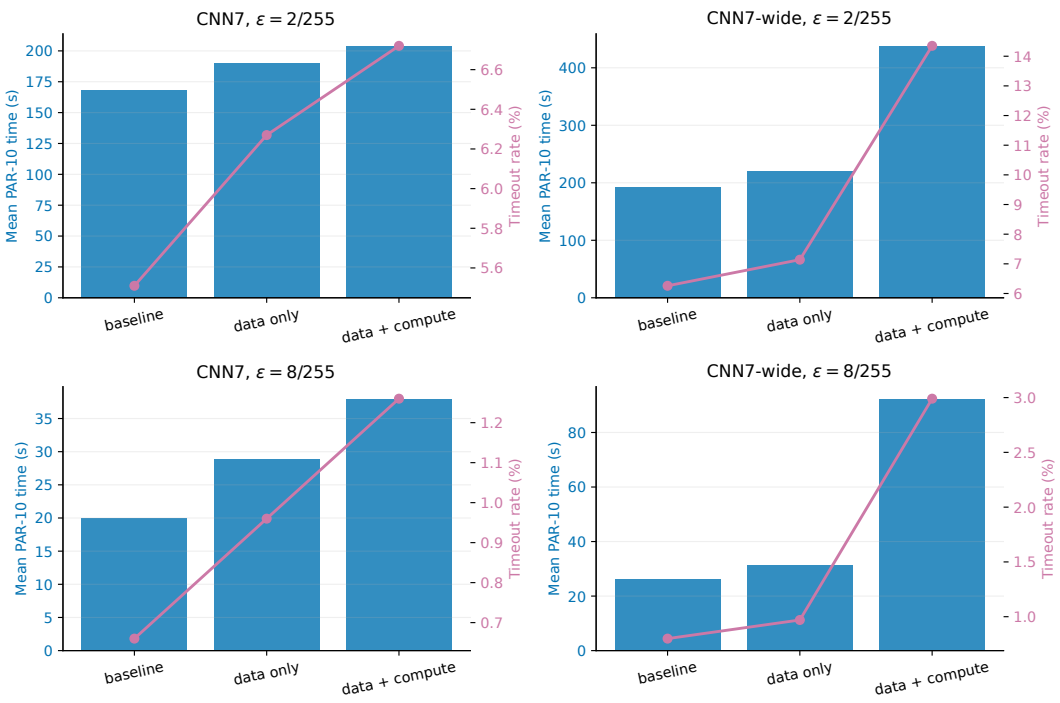


Figure 26: Complete-verification cost for the headline CIFAR-10 models: mean PAR-10 time (bars) and timeout rate (line) for the baseline, data-only, and joint data and compute configurations. PAR-10 charges each timeout ten times the 300 s cutoff.

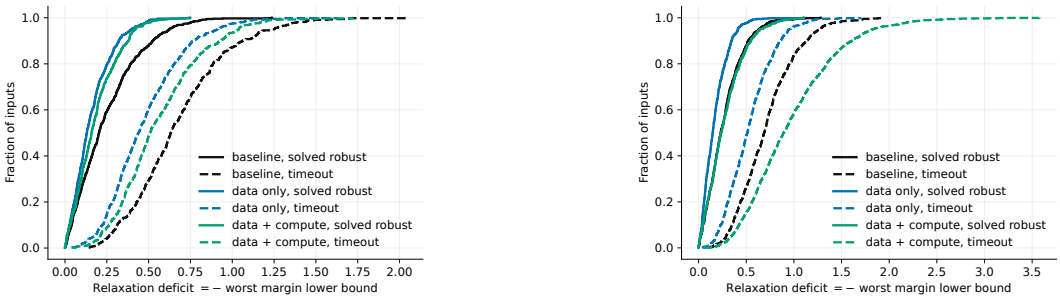


Figure 27: Relaxation deficit within the hard pool at $\epsilon = 2/255$ for CNN-7 (left) and wide CNN-7 (right). Solid curves represent inputs proved robust; dashed curves represent timeouts. Colours distinguish baseline, data-only, and joint data and compute scaling.

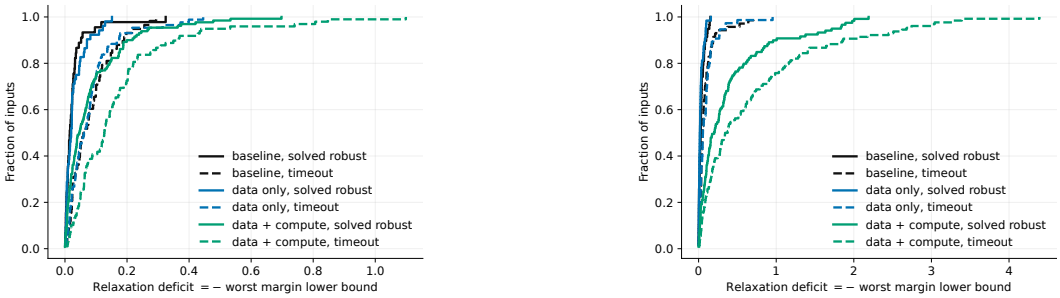


Figure 28: Relaxation deficit within the hard pool at $\epsilon = 8/255$ for CNN-7 (left) and wide CNN-7 (right). Solid curves represent inputs proved robust; dashed curves represent timeouts. Colours distinguish baseline, data-only, and joint data and compute scaling.

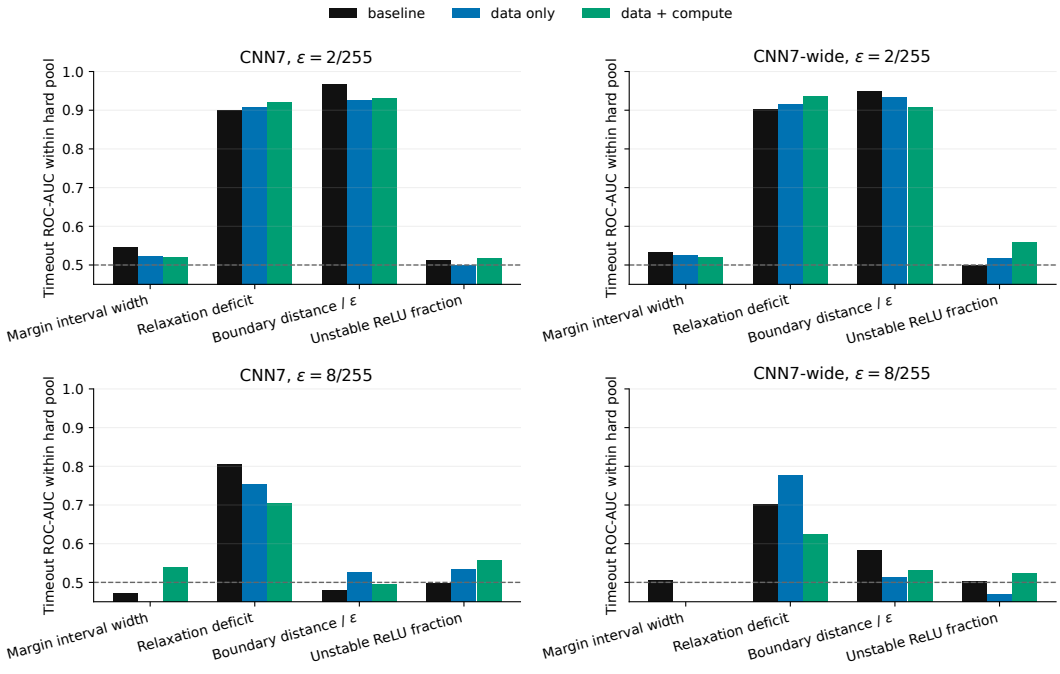


Figure 29: ROC-AUC of initial-relaxation properties for predicting complete-verification timeouts within the hard pool. Colours denote baseline, data-only, and joint data and compute scaling; the dashed horizontal line denotes chance performance.

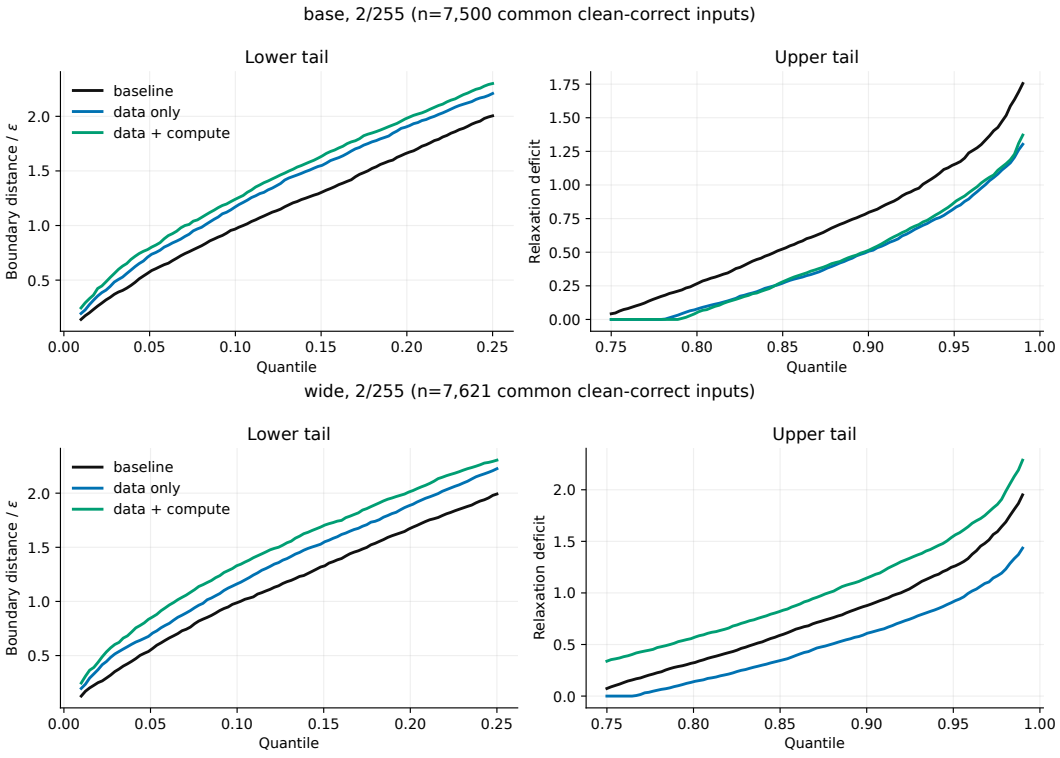


Figure 30: Boundary-distance lower tail (left) and relaxation-deficit upper tail (right) at $\epsilon = 2/255$, for CNN-7 (top; $n = 7{,}500$) and wide CNN-7 (bottom; $n = 7{,}621$). Curves compare baseline, data-only, and joint data and compute scaling on inputs classified correctly by all three configurations.

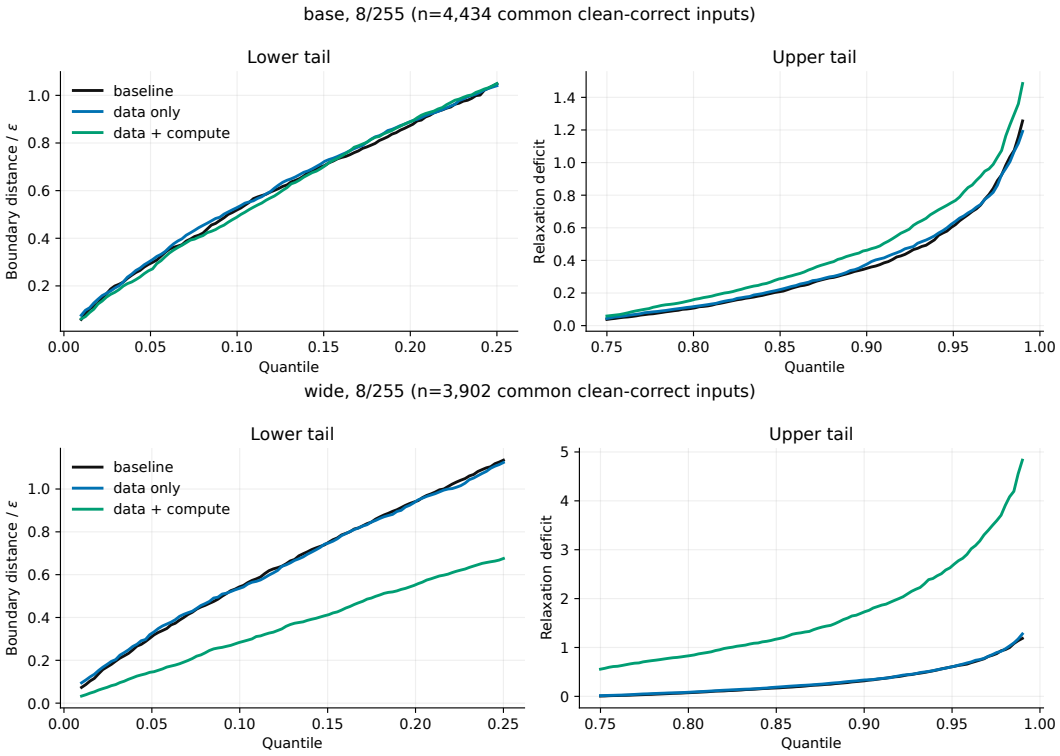


Figure 31: Boundary-distance lower tail (left) and relaxation-deficit upper tail (right) at $\epsilon = 8/255$, for CNN-7 (top; $n = 4{,}434$) and wide CNN-7 (bottom; $n = 3{,}902$). Curves compare baseline, data-only, and joint data and compute scaling on inputs classified correctly by all three configurations.