References

[Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] Rangamani, A., Lindegaard, M., Galanti, T., and Poggio, T. A. Feature learning in deep classifiers through intermediate neural collapse. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28729–28745. PMLR, 23–29 Jul 2023.

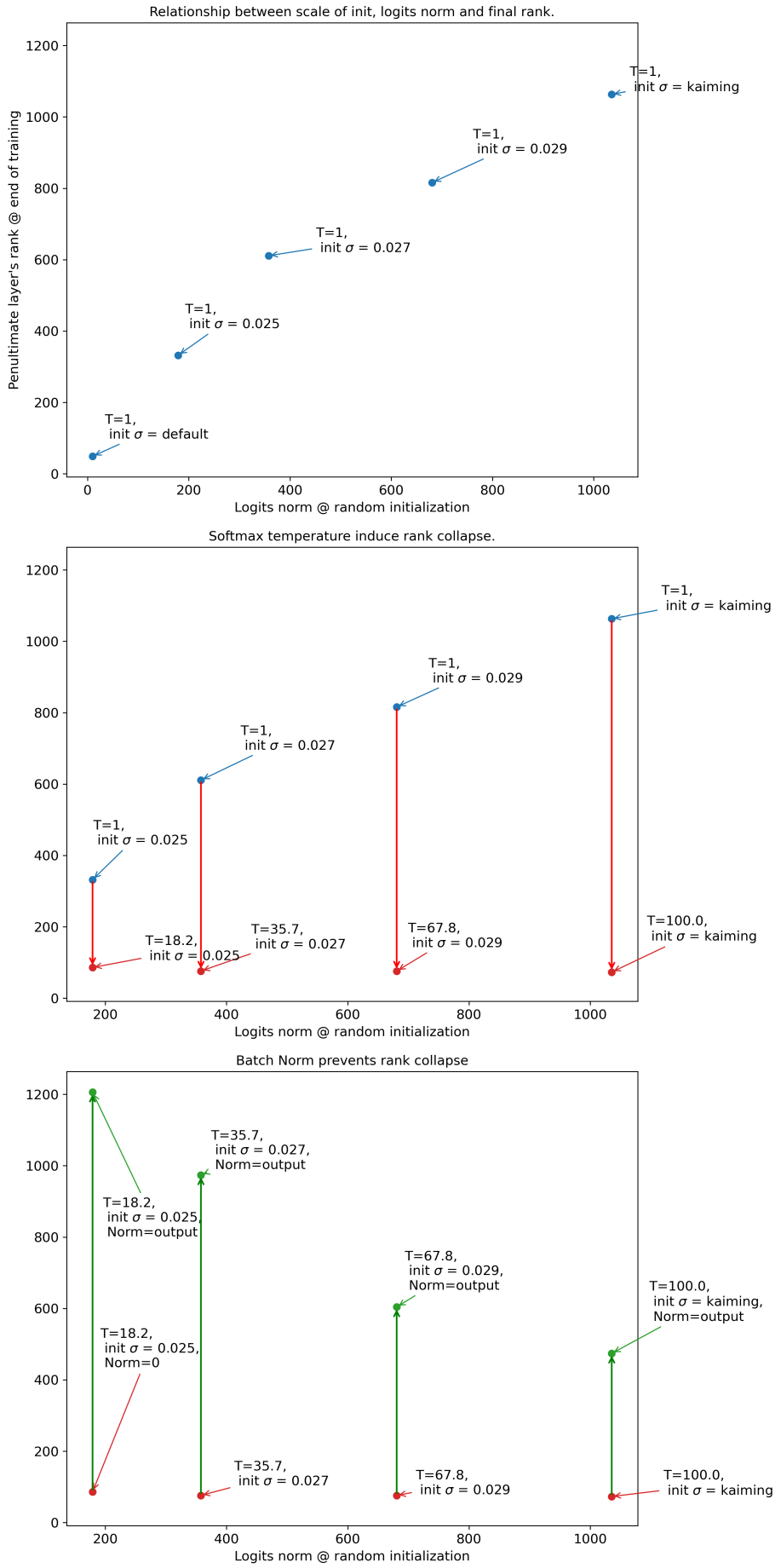

Figure 1: The effect of different hyperparameters on logits norm of randomly initialized model (x-axis) and rank of representation after training (y-axis). Each point on the plot represents a single MLP network with eight layers (2048 neurons in each hidden layer) trained to 100% on CIFAR-10 with different hyperparameters. The parameters that change between experiments are listed in the annotations on the plot. **Top plot** presents almost a linear relationship between the scale of initialization of the model's weights and logits norm and between logits norm and final rank of representations. **Middle plot** presents the effect of applying softmax temperature to match the logits on most collapsed networks (init σ = default, representing a default PyTorch initialization scheme). In line with our observations, all the models collapsed to a rank comparable to the rank of the model (T=1, init σ = default). **Bottom plot** presents the effect of the normalization layer applied at the output of all the networks. In line with our predictions, layer norm prevents increasing the logits norm and therefore prevents the collapse of the model.

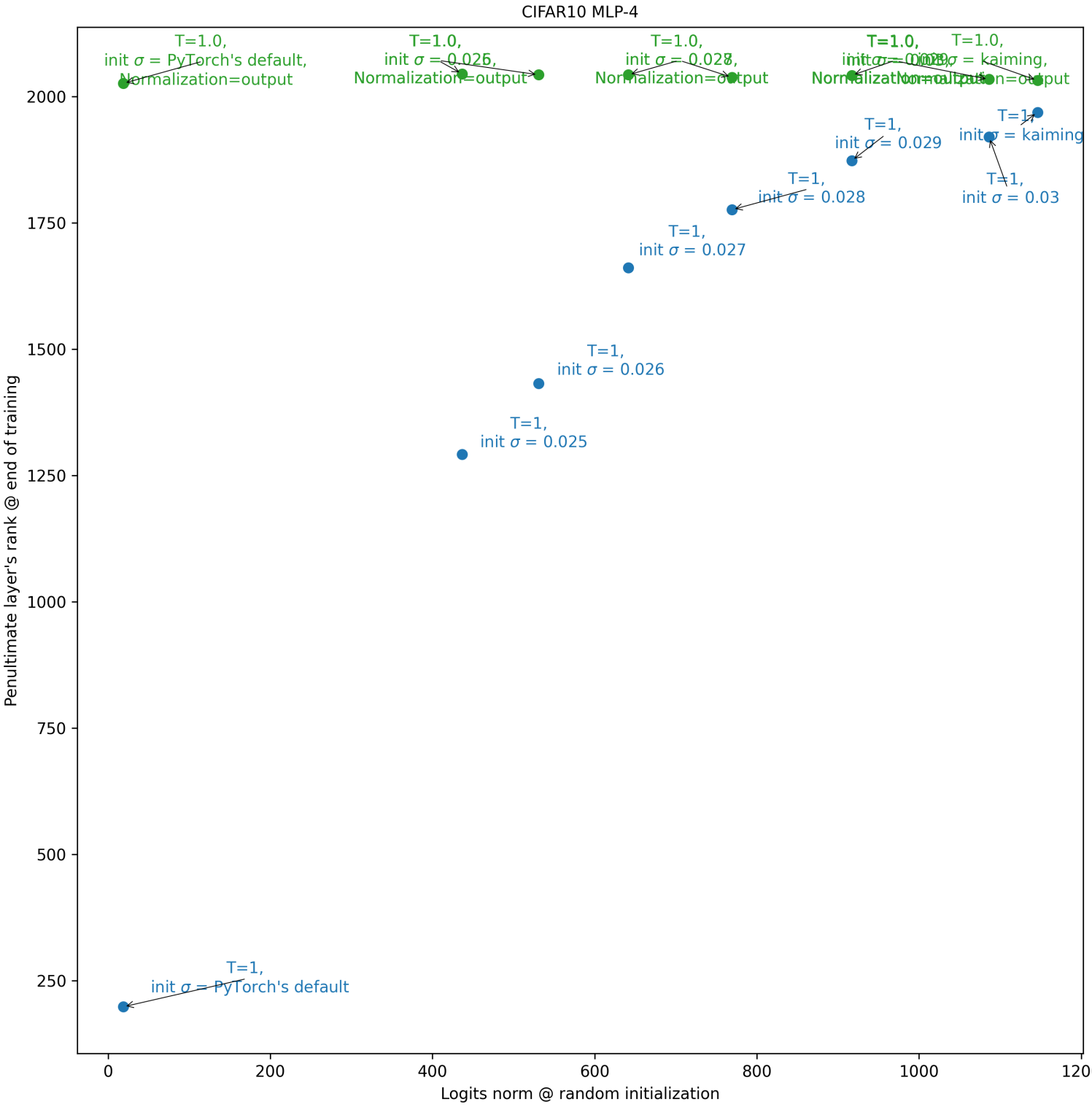

Figure 2: The effect of initialization scale and normalization layer at the output of the network on logits norm of randomly initialized model (x-axis) and rank of representation after training (y-axis). Each point on the plot represents a single MLP network with four hidden layers (2048 neurons in each hidden layer) trained to 100% on CIFAR-10 with different hyperparameters. The parameters that change between experiments are listed in the annotations on the plot.

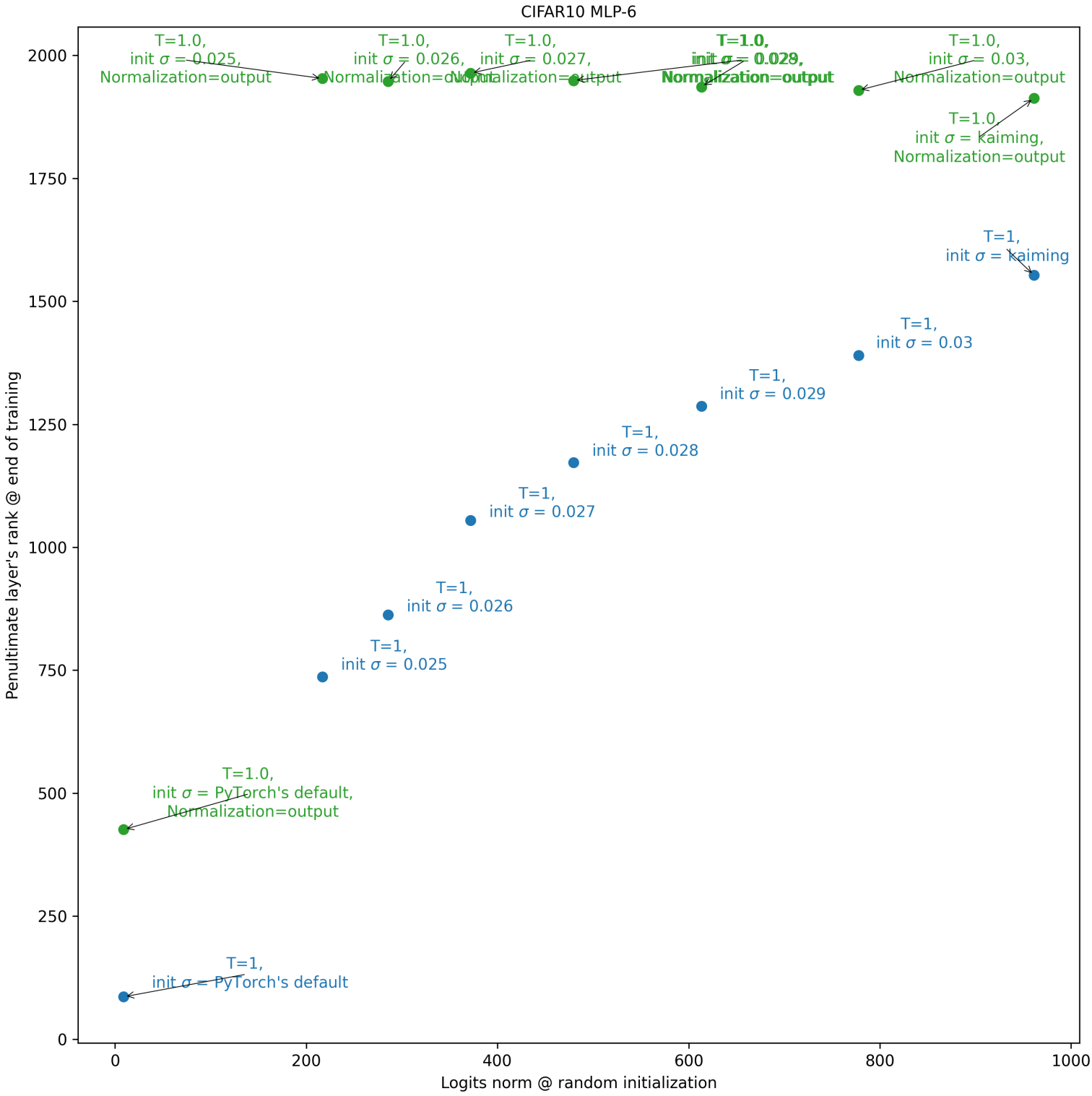

Figure 3: The effect of initialization scale and normalization layer at the output of the network on logits norm of randomly initialized model (x-axis) and rank of representation after training (y-axis). Each point on the plot represents a single MLP network with six hidden layers (2048 neurons in each hidden layer) trained to 100% on CIFAR-10 with different hyperparameters. The parameters that change between experiments are listed in the annotations on the plot.

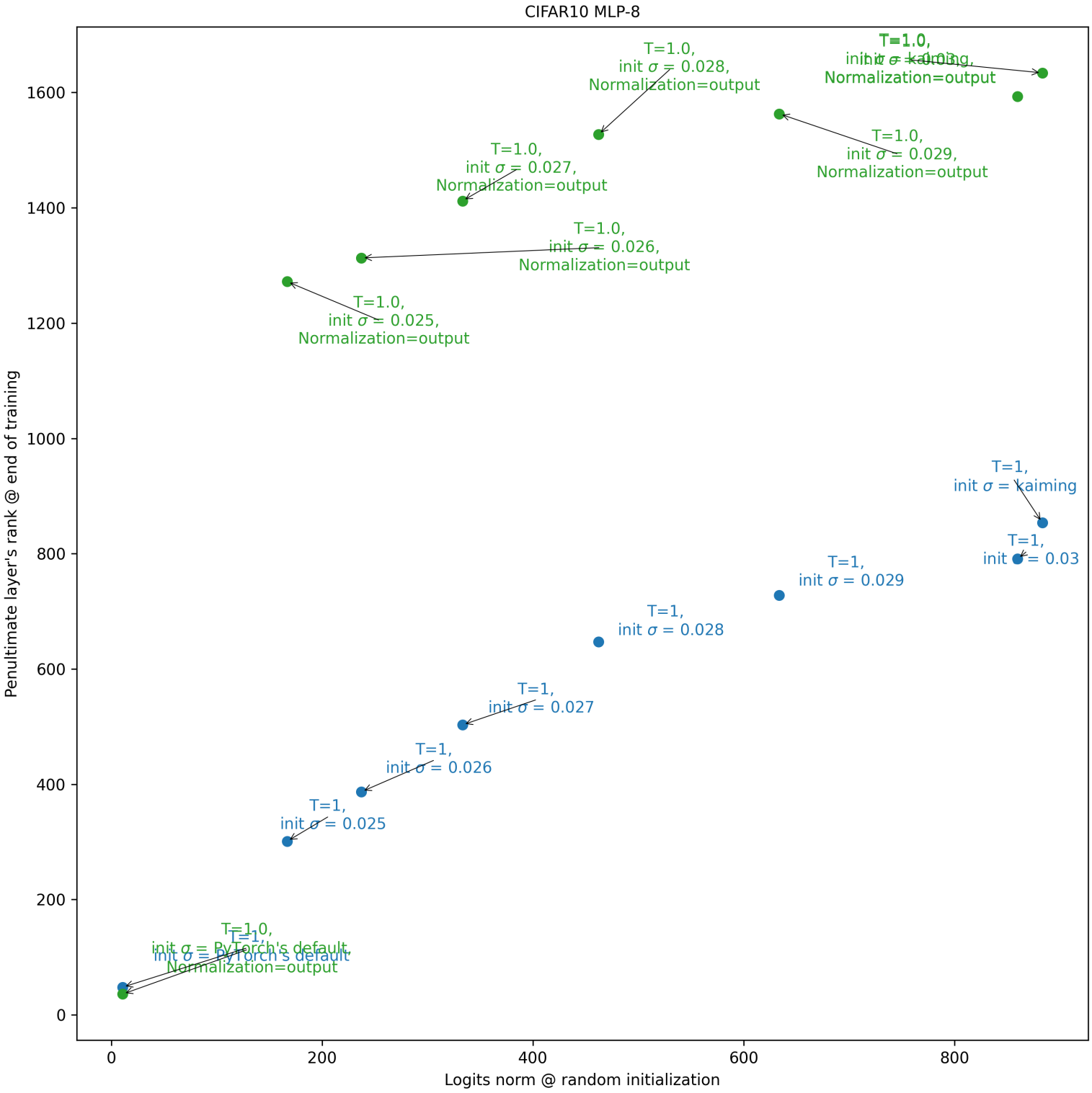

Figure 4: The effect of initialization scale and normalization layer at the output of the network on logits norm of randomly initialized model (x-axis) and rank of representation after training (y-axis). Each point on the plot represents a single MLP network with eight hidden layers (2048 neurons in each hidden layer) trained to 100% on CIFAR-10 with different hyperparameters. The parameters that change between experiments are listed in the annotations on the plot.

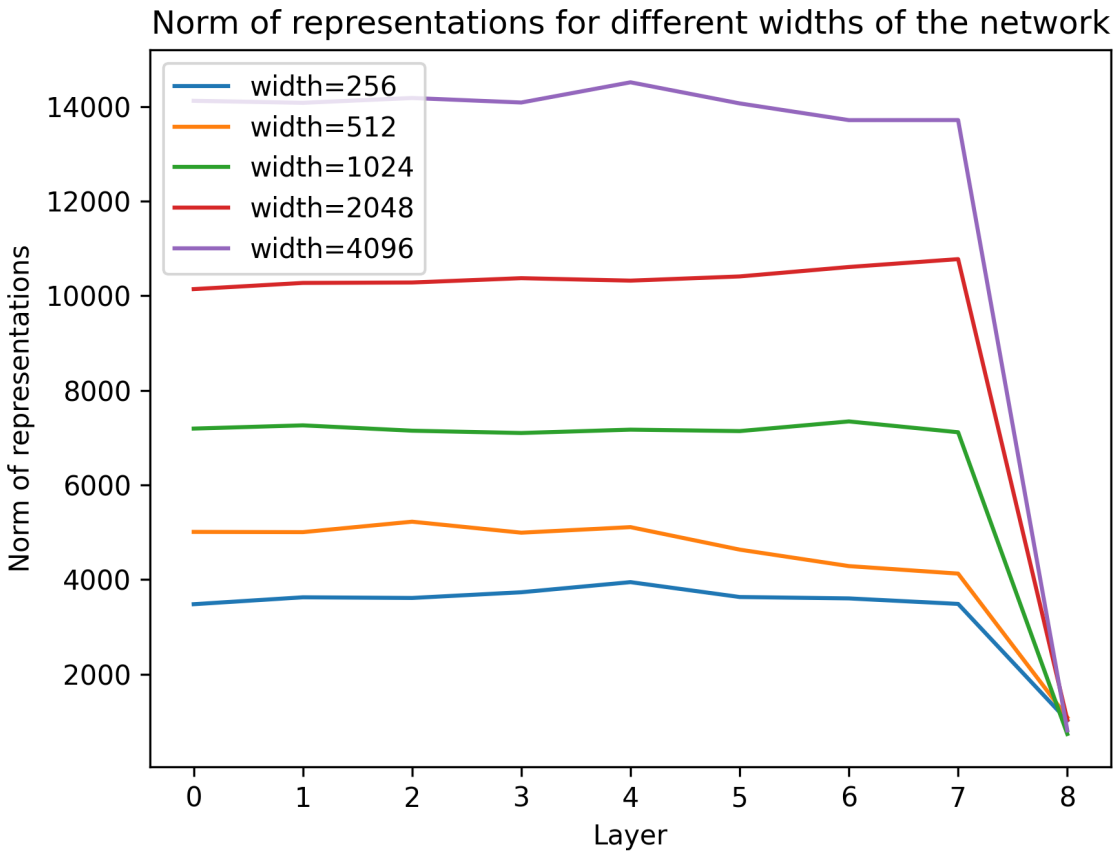

Figure 5: Increasing network width leads to the higher norm of representations throughout the whole network for randomly initialized MLP models. In this experiment, we randomly sampled various MLP networks with eight hidden layers and gathered Frobenius norms for the representations at each layer.

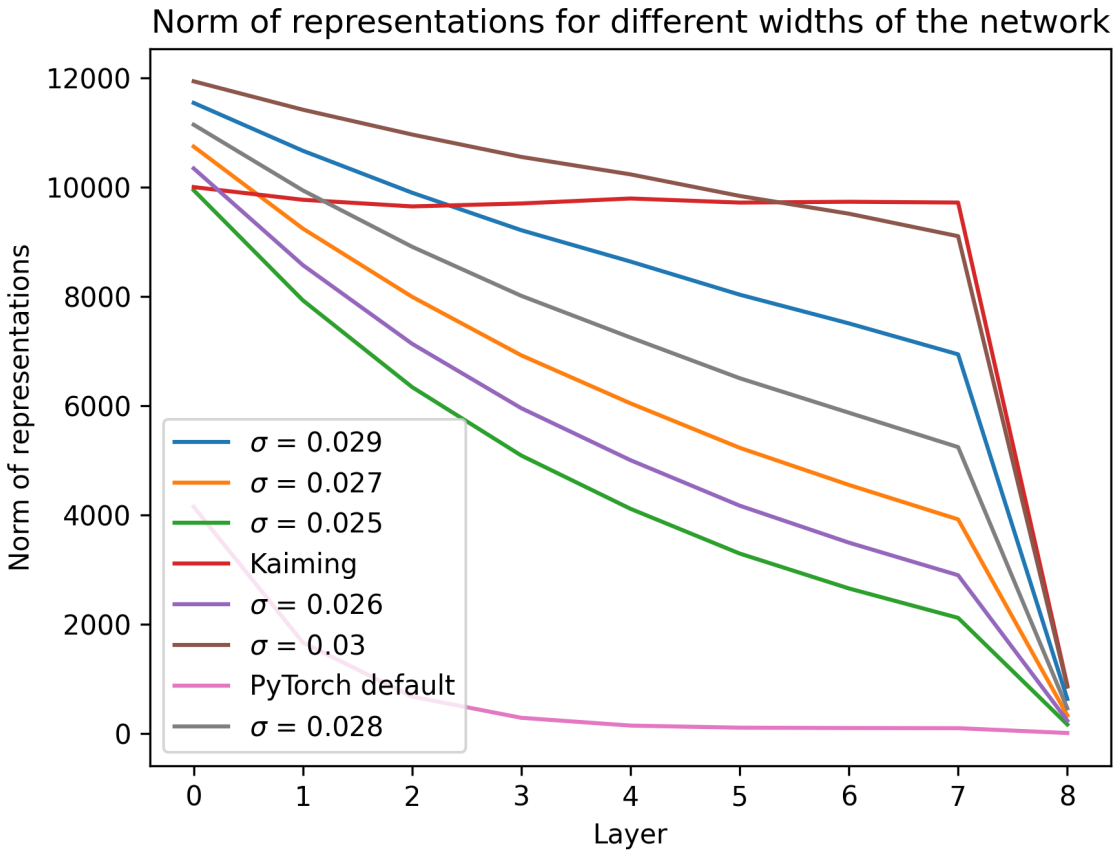

Figure 6: Using initialization schemes different than variance preserving ones (Kaiming, color red) leads to the decreased norm of the representations over multiple layers. The pace at which the norm decreases over the layers depends on the initialization scale. The most radical is achieved for the default PyTorch initialization scheme (color pink).

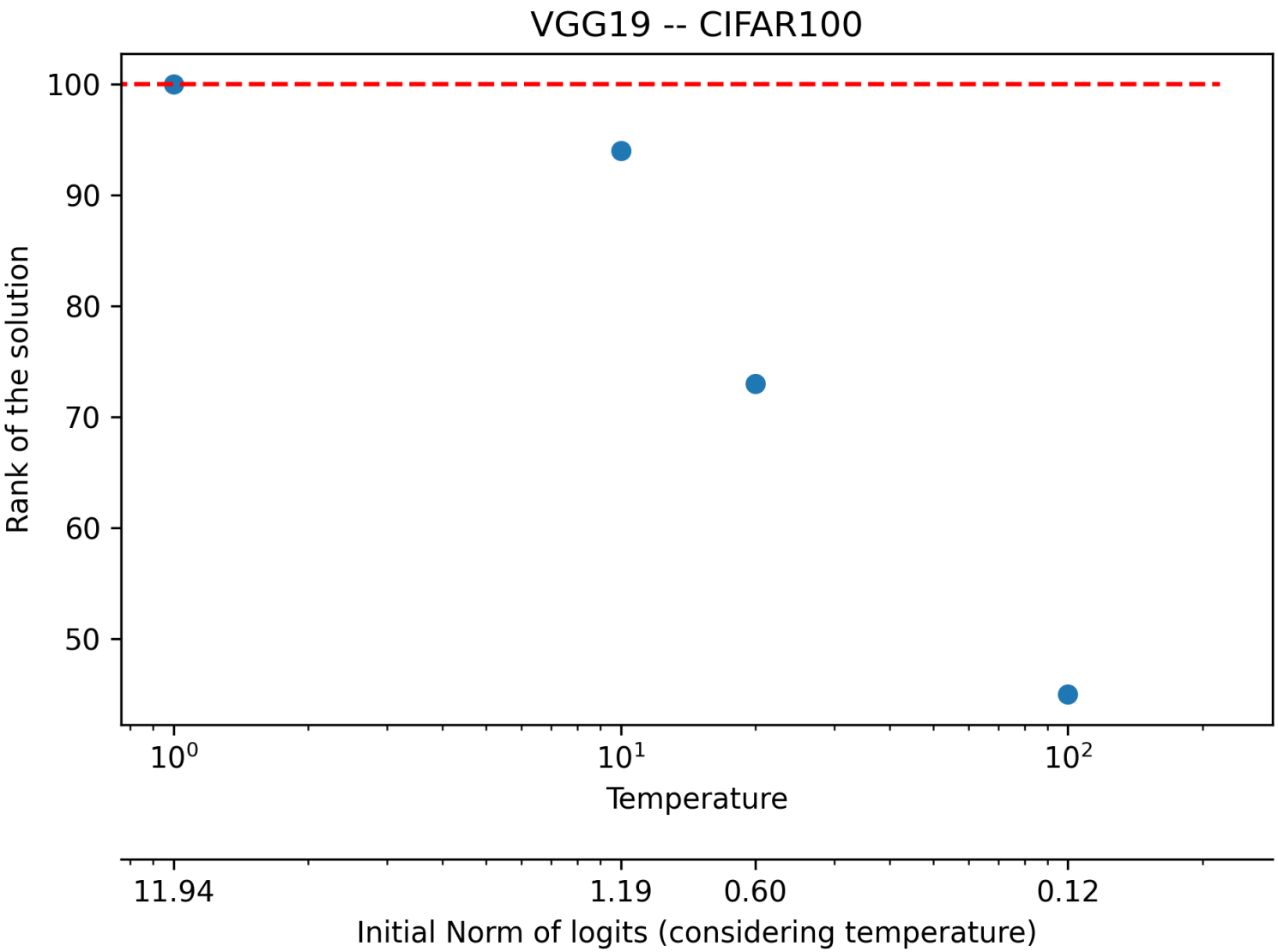

Figure 7: **Rank deficit bias** achieved by the VGG19 network trained on a subset (2000 examples) of CIFAR-100 dataset. Each dot represents a network trained to 100% accuracy on the training dataset. The networks were trained with different temperatures (x-axis), which changed their initial logit norm (second x-axis) and resulted in a solution of lower rank (y-axis). Current theory predicts that the network will find a solution of rank 100 (equal to the number of classes in the training dataset). We trained the networks with the recommended hyperparameters to achieve the best performing models, and **we did not use any specific regularization to achieve the rank deficit bias.**

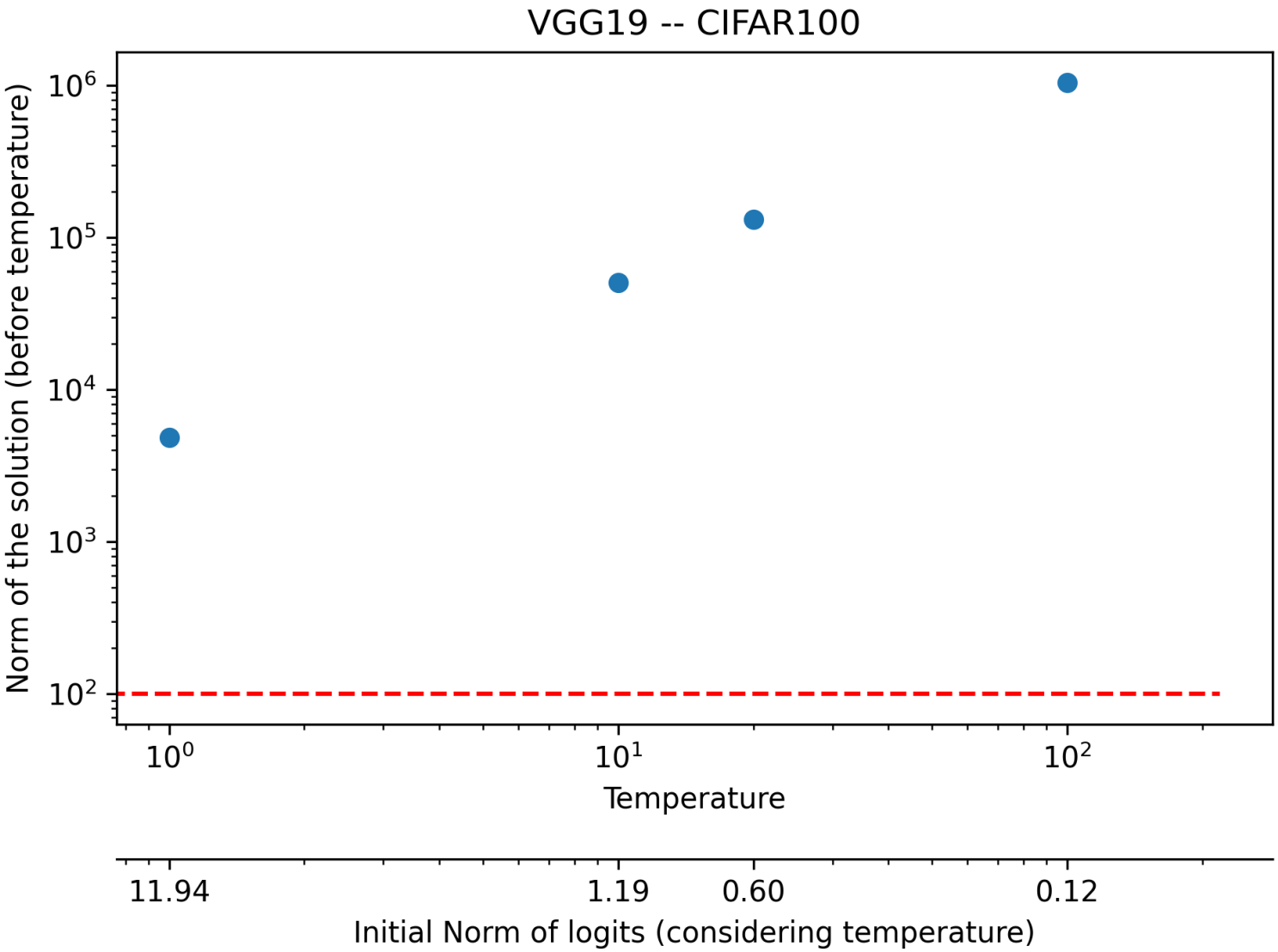

Figure 8: **Rank deficit bias** is caused by the excessive growth of the norm of the logits. The greater the final norm (y-axis), the lower the final rank of the solution (y-axis in Figure 7). Each dot represents a network trained to 100% accuracy on the training dataset. The networks were trained with different temperatures (x-axis), which changed their initial logit norm (second x-axis) and resulted in a solution of greater norm of the logits (y-axis).

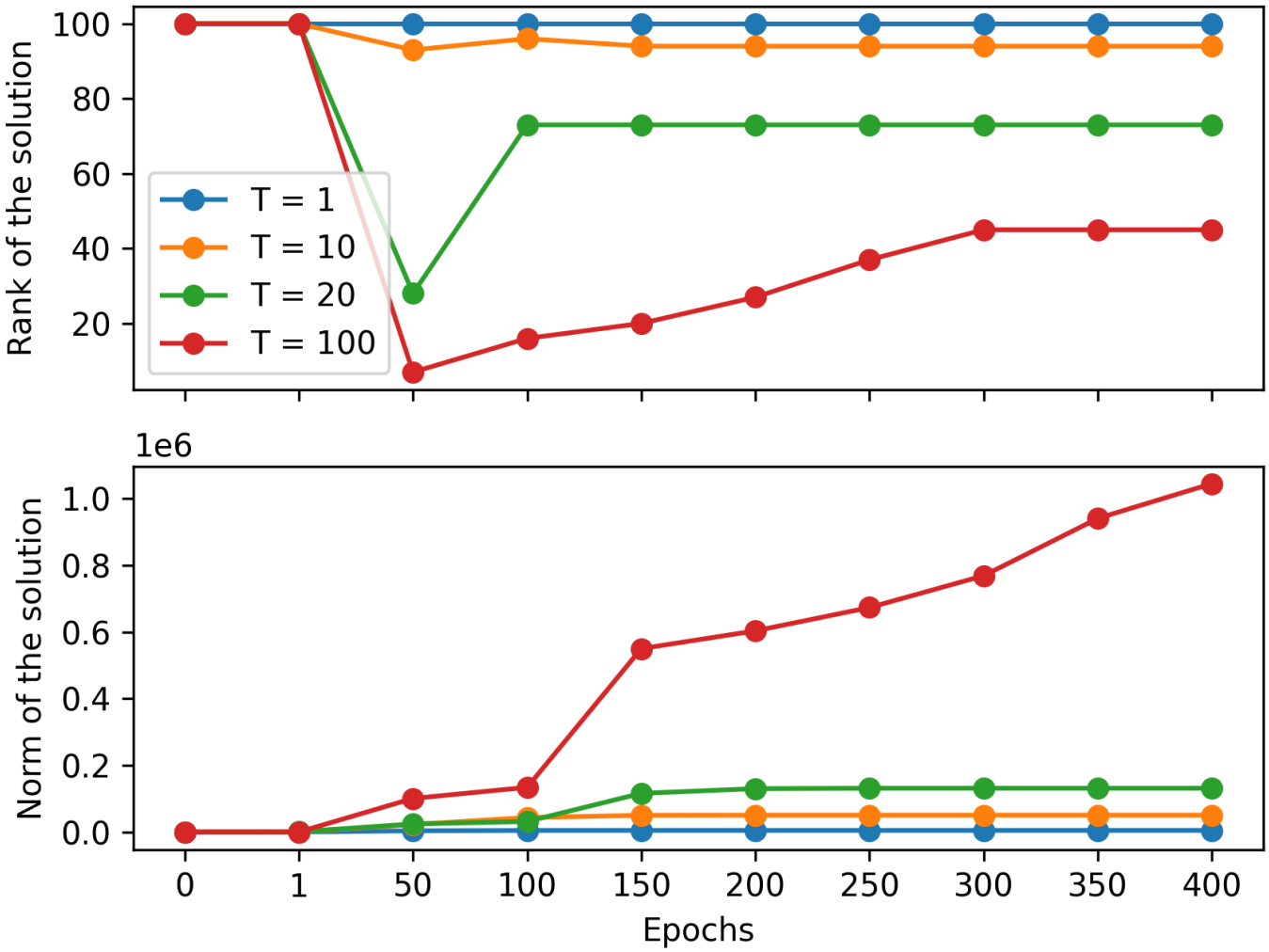

Figure 9: The evolution of the logits rank (top) and logits (norm) over the epochs of training VGG19 on the CIFAR-100 subset. The figure supports two findings from our work: 1) the greater temperature leads to more severe rank collapse (top) and greater logits norm (bottom), and 2) the rank collapses at the beginning of the training and we do not need to enter TPT phase to observe the collapse in contrast to results related to Neural Collapse.

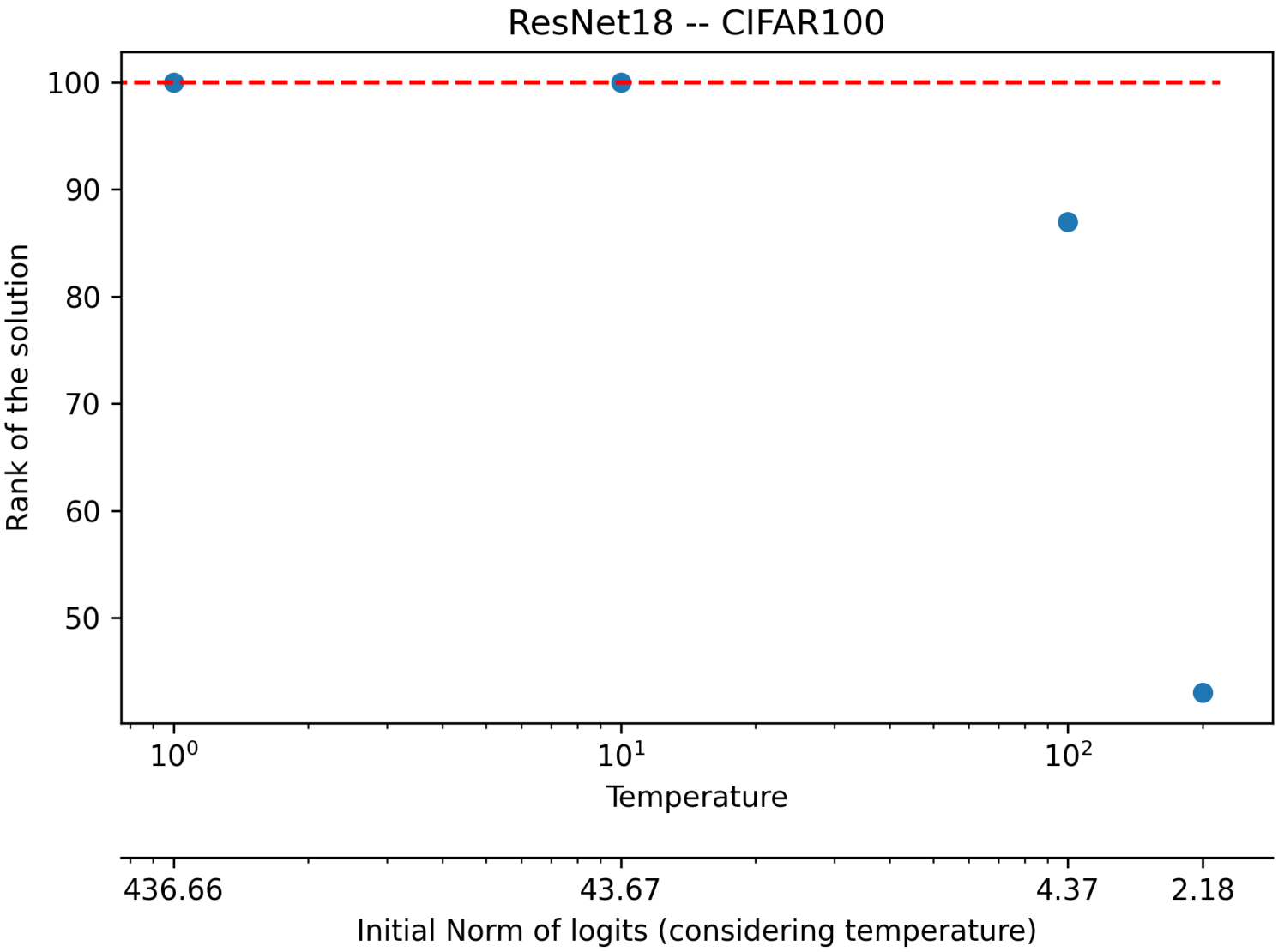

Figure 10: **Rank deficit bias** achieved by the ResNet18 network trained on a subset (2000 examples) of CIFAR-100 dataset. Each dot represents a network trained to 100% accuracy on the training dataset. The networks were trained with different temperatures (x-axis), which changed their initial logit norm (second x-axis) and resulted in a solution of lower rank (y-axis). Current theory predicts that the network will find a solution of rank 100 (equal to the number of classes in the training dataset). We trained the networks with the recommended hyperparameters to achieve the best performing models, and **we did not use any specific regularization to achieve the rank deficit bias.**

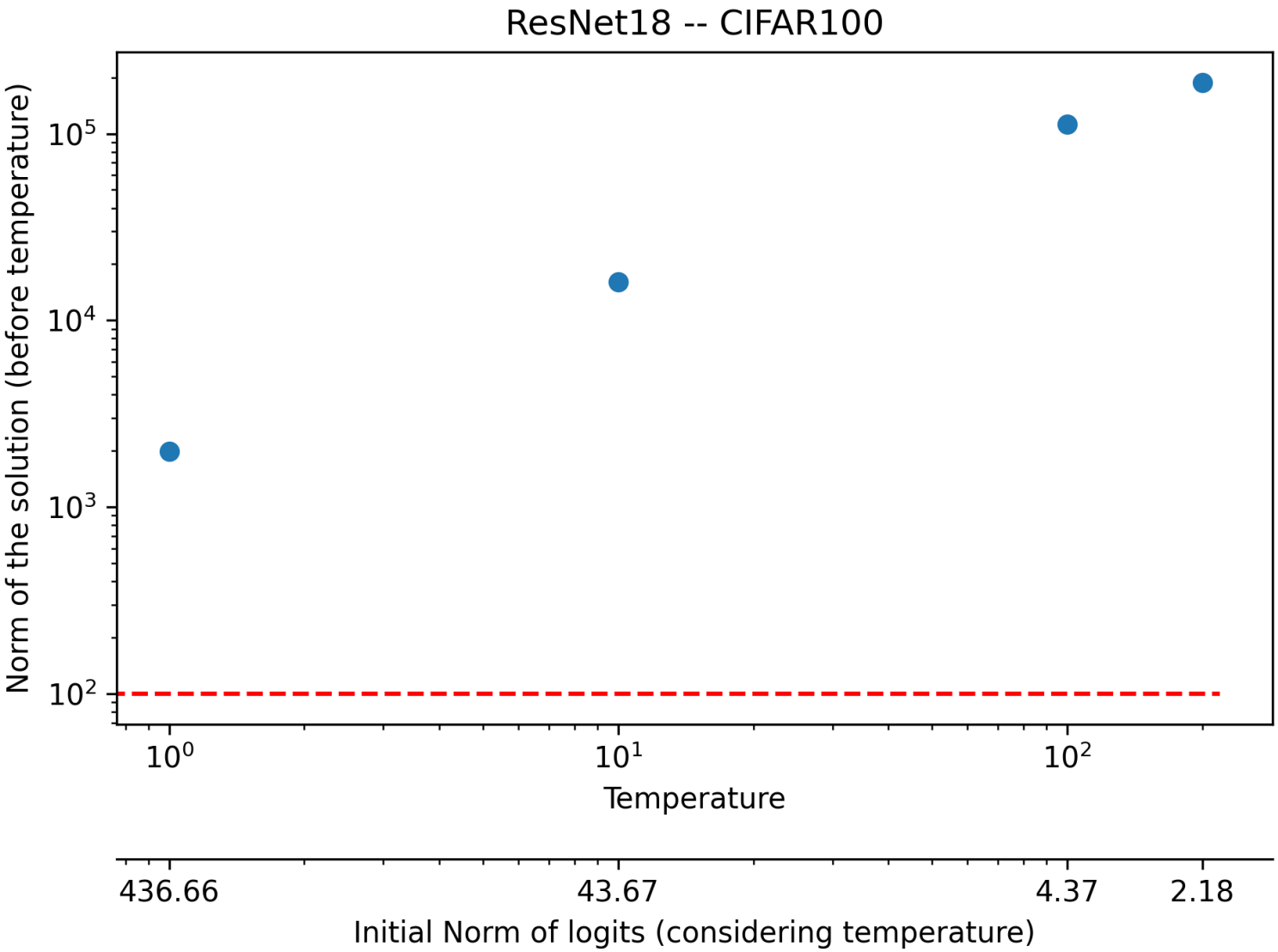

Figure 11: **Rank deficit bias** is caused by the excessive growth of the norm of the logits. The greater the final norm (y-axis), the lower the final rank of the solution (y-axis in Figure 7). Each dot represents a network trained to 100% accuracy on the training dataset. The networks were trained with different temperatures (x-axis), which changed their initial logit norm (second x-axis) and resulted in a solution of greater norm of the logits (y-axis).

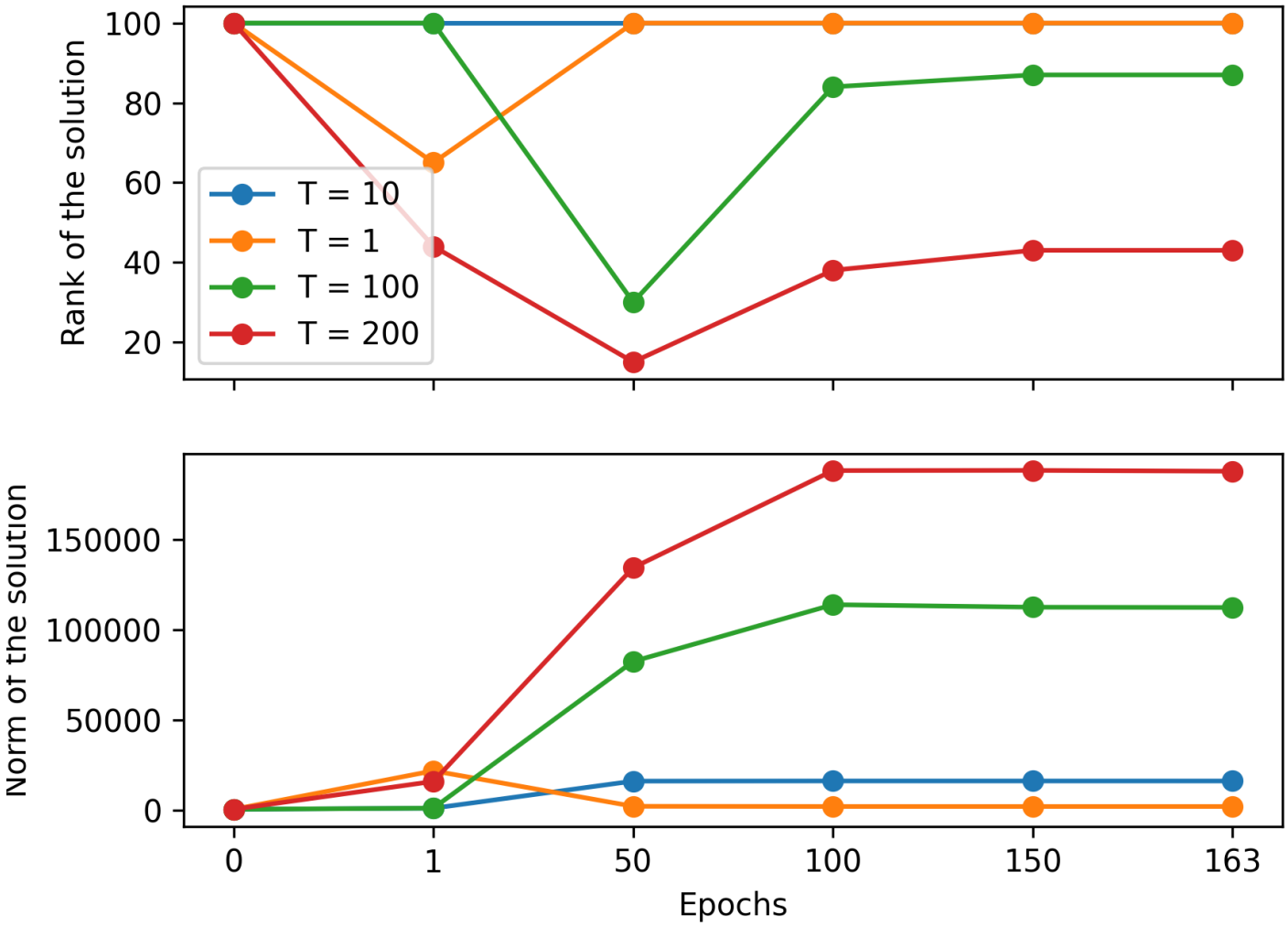

Figure 12: The evolution of the logits rank (top) and logits (norm) over the epochs of training ResNet18 on the CIFAR-100 subset. The figure supports two findings from our work: 1) the greater temperature leads to more severe rank collapse (top) and greater logits norm (bottom), and 2) the rank collapses at the beginning of the training, and we do not need to enter TPT phase to observe the collapse in contrast to results related to Neural Collapse.

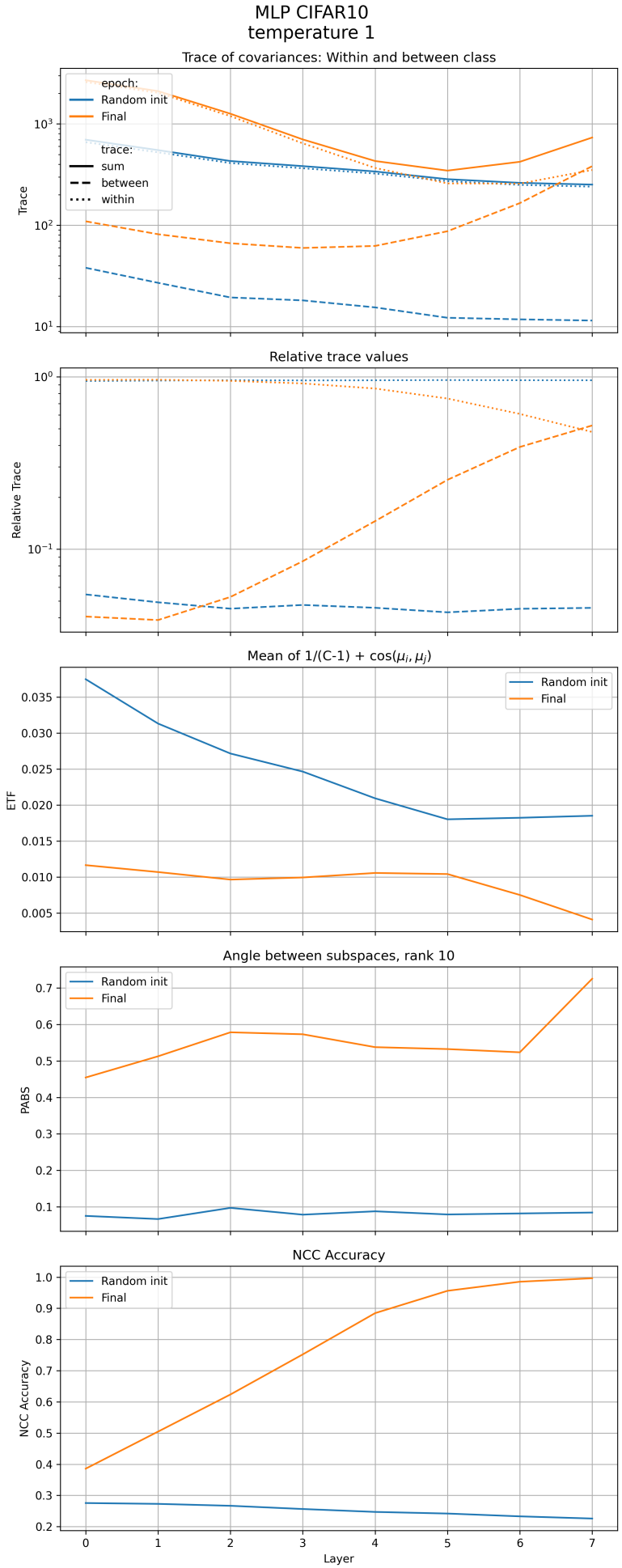

Figure 13: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by MLP-8 trained on CIFAR-10 with temperature 1. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

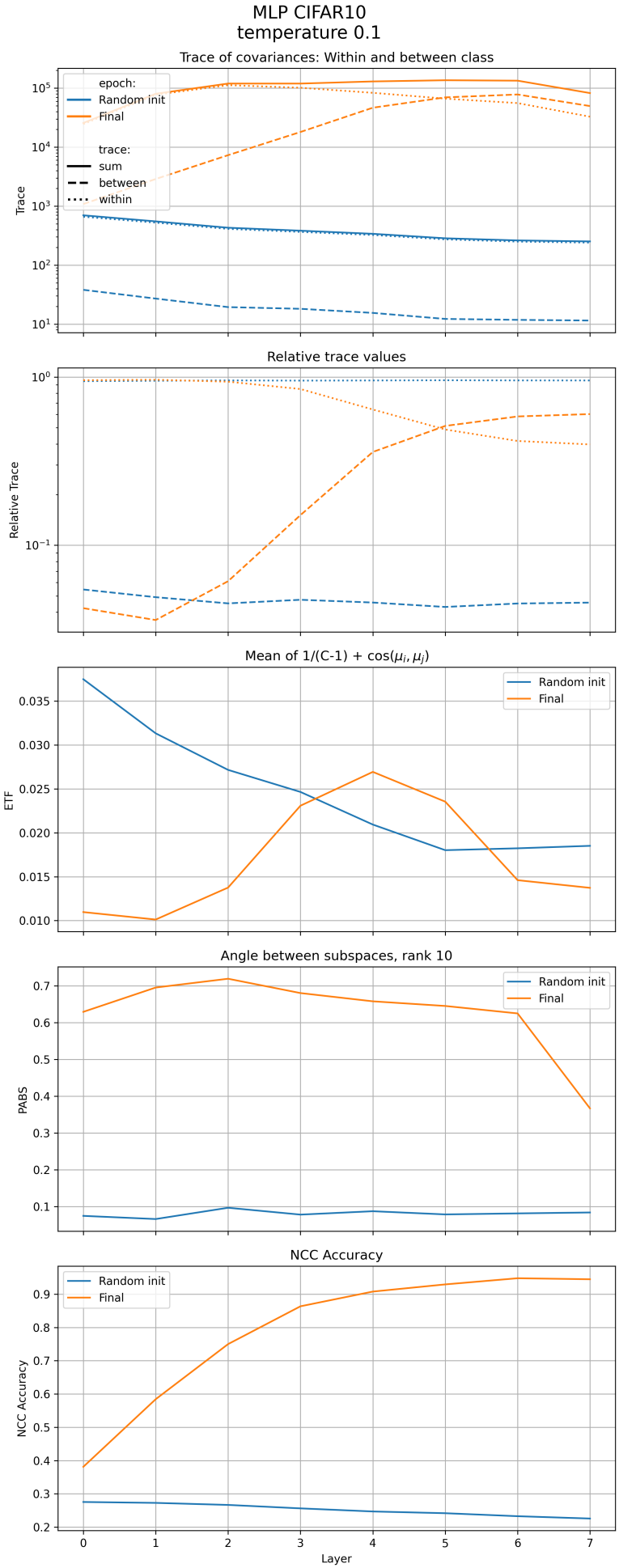

Figure 14: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by MLP-8 trained on CIFAR-10 with temperature 10. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

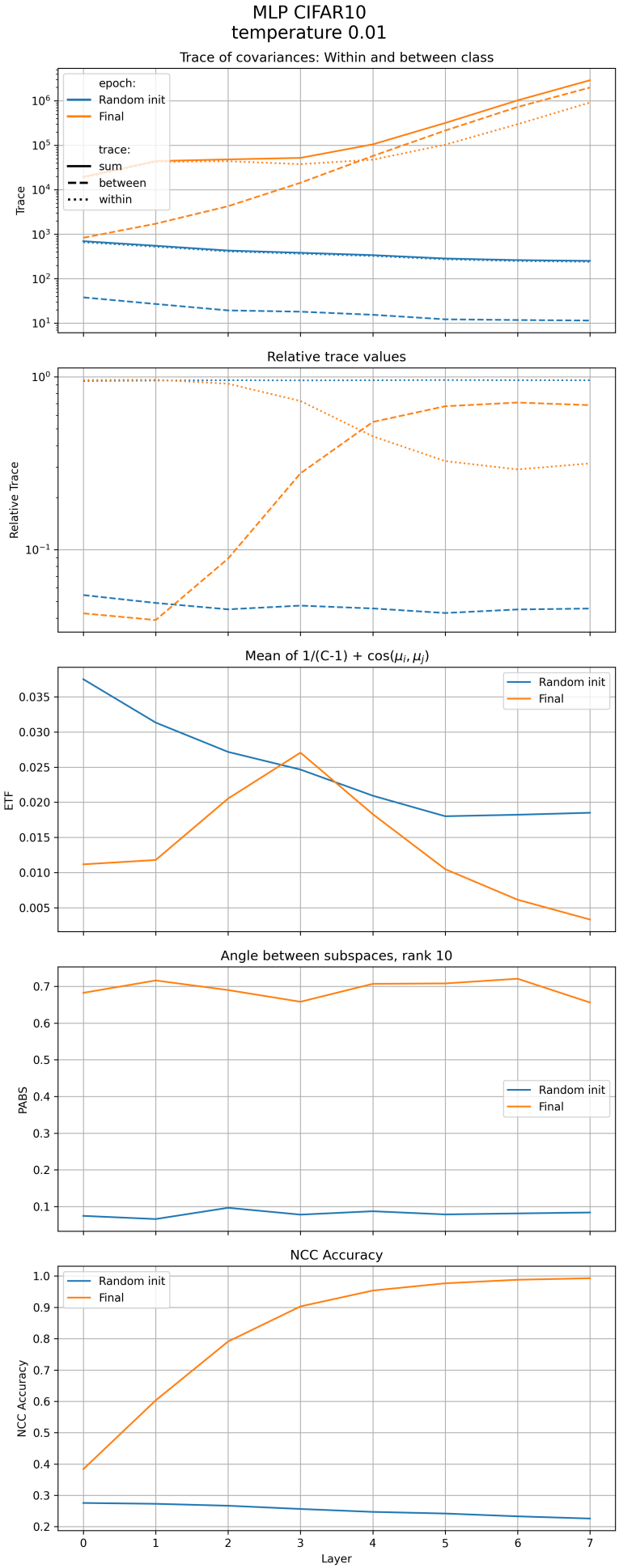

Figure 15: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by MLP-8 trained on CIFAR-10 with temperature 100. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

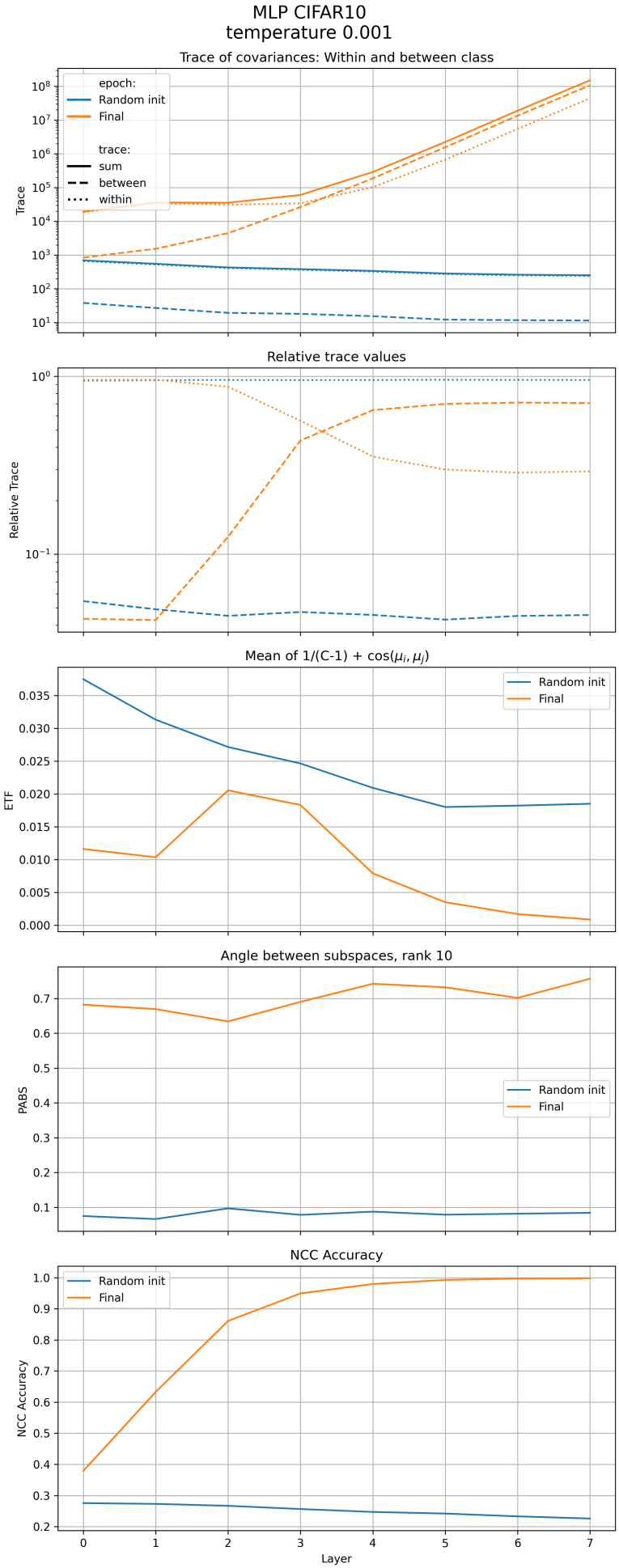

Figure 16: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by MLP-8 trained on CIFAR-10 with temperature 1000. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

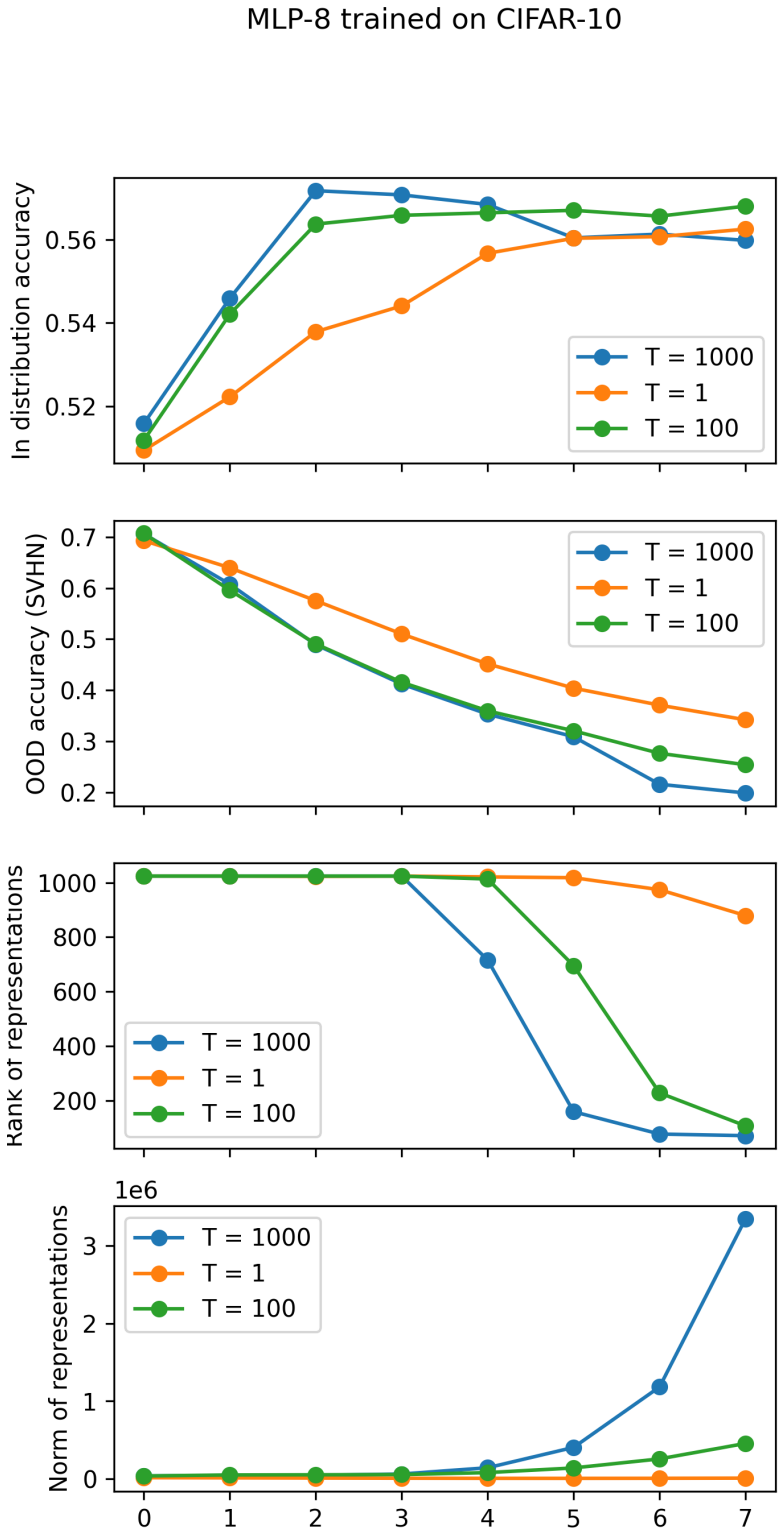

Figure 17: The plot presents the relationship between training temperature (different colors of the plots) and accuracy of linear probes attached to different layers for in-distribution generalization (top plot), out-of-distribution generalization (second top plot, evaluated at SVHN), rank of the representations at each layer (third top plot) and norm of the representations at each layer (bottom plot). The greater temperature leads to more severe rank collapse, greater compression, higher representations norm, and worse OOD performance.

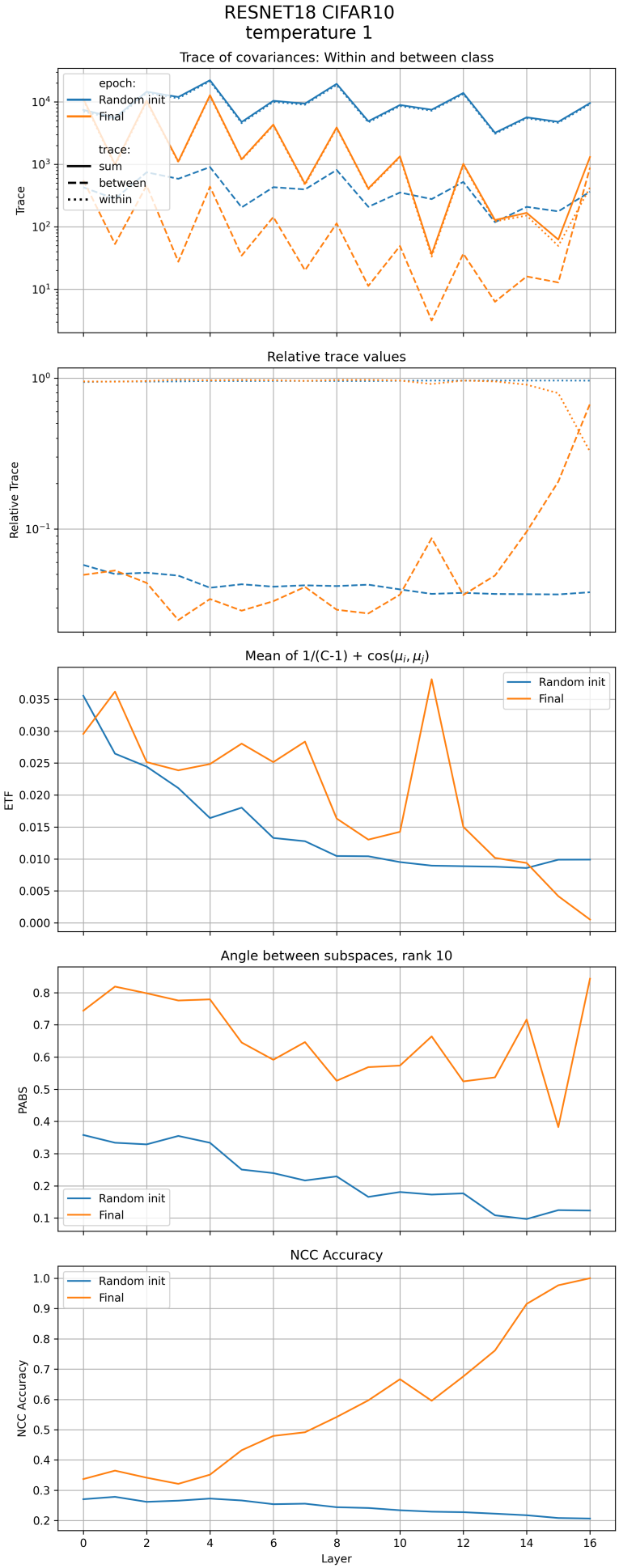

Figure 18: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet18 trained on CIFAR-10 with temperature 1. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

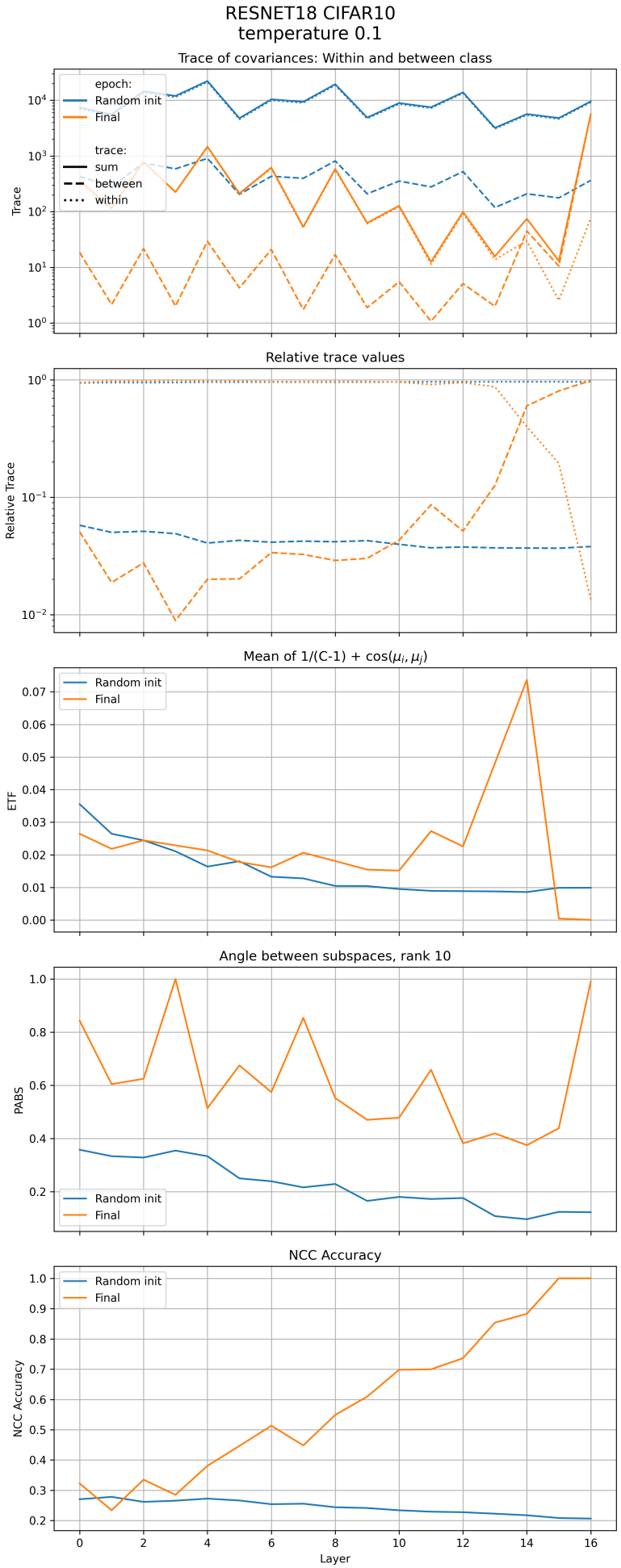

Figure 19: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet18 trained on CIFAR-10 with temperature 10. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

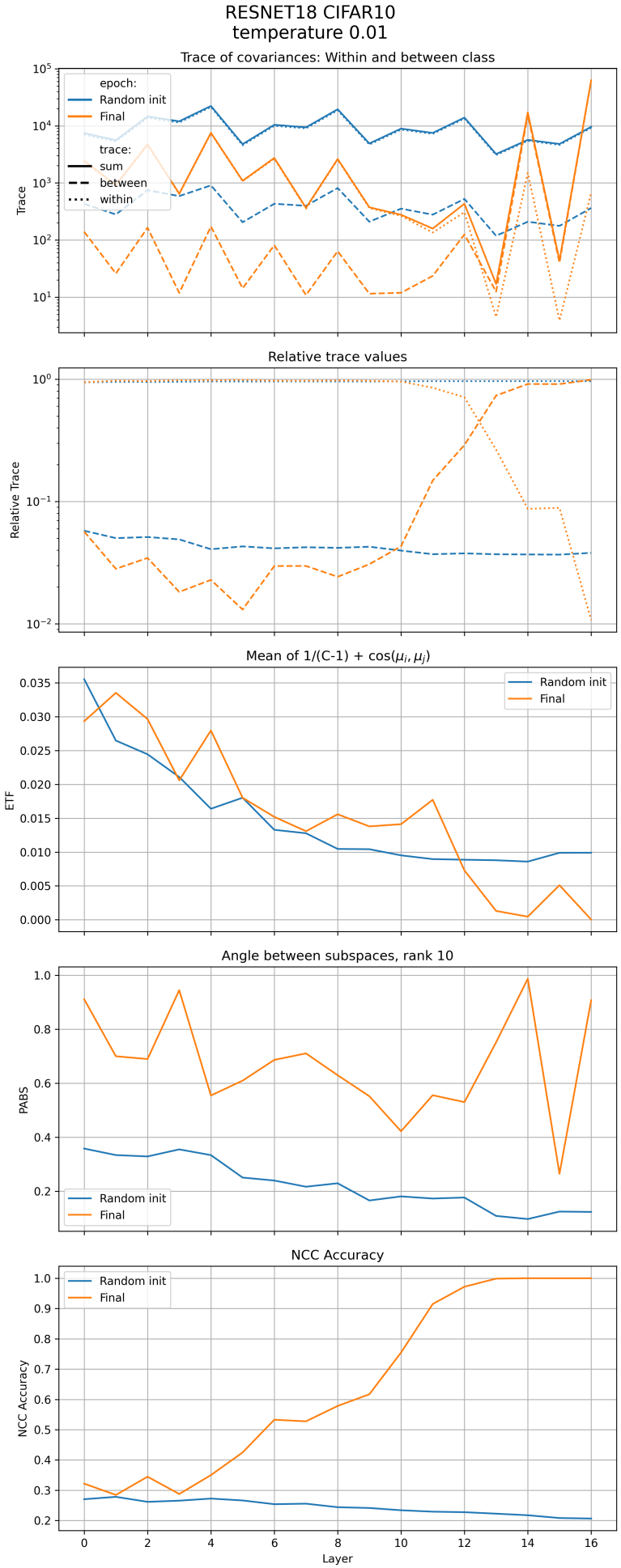

Figure 20: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet18 trained on CIFAR-10 with temperature 100. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

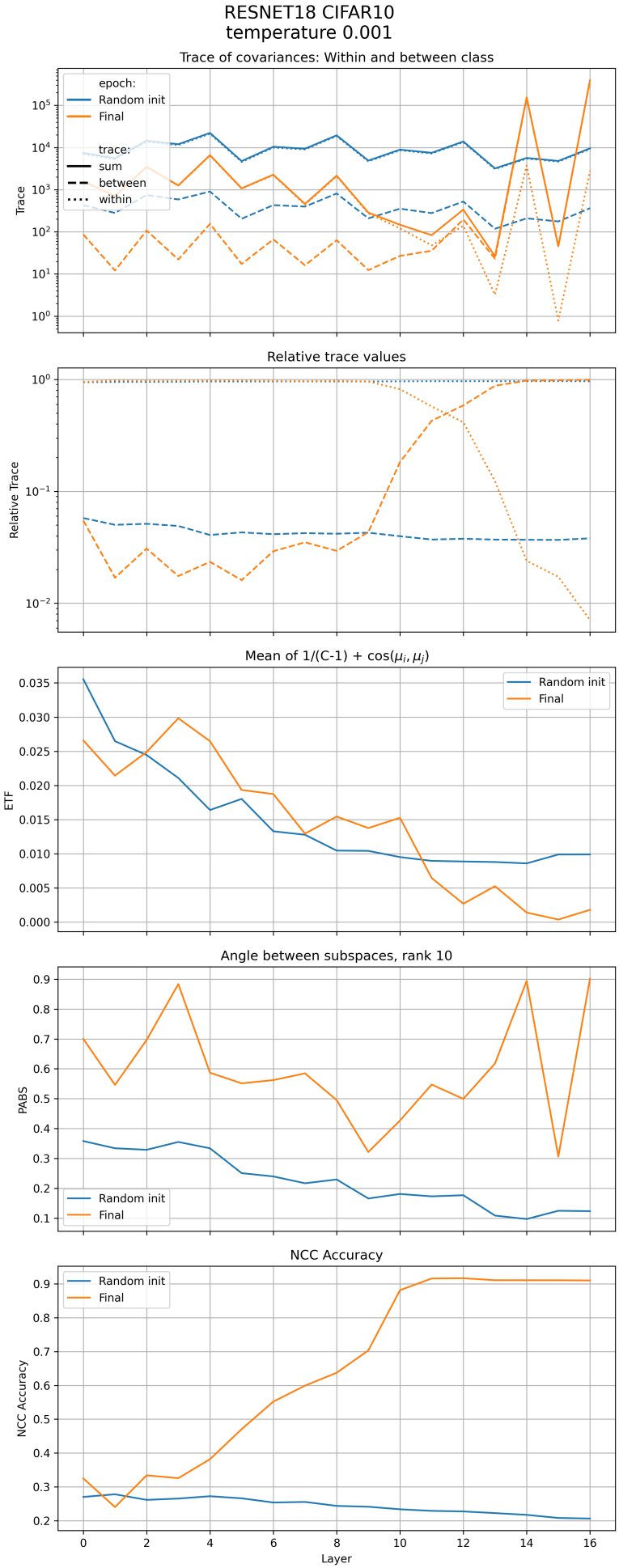

Figure 21: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet18 trained on CIFAR-10 with temperature 1000. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

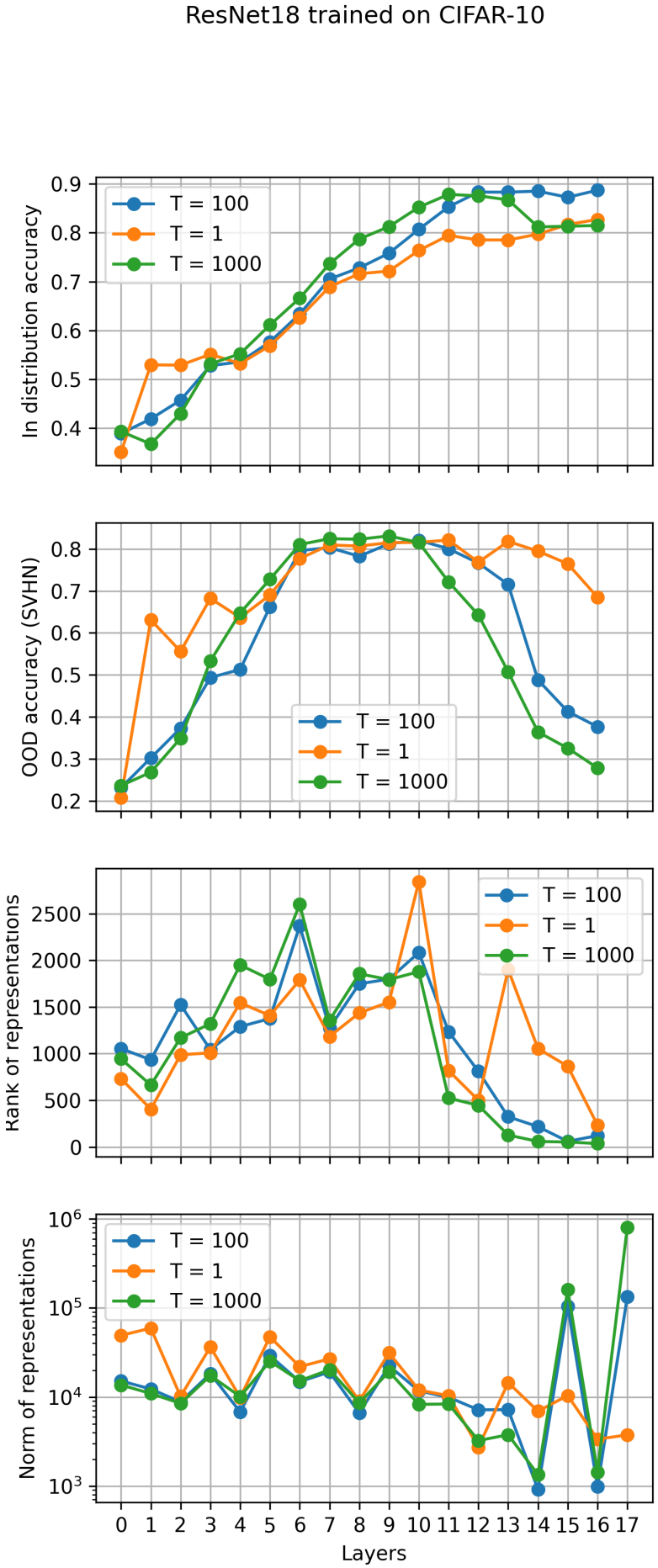

Figure 22: The plot presents the relationship between training temperature (different colors of the plots) and accuracy of linear probes attached to different layers for in-distribution generalization (top plot), out-of-distribution generalization (second top plot, evaluated at SVHN), rank of the representations at each layer (third top plot) and norm of the representations at each layer (bottom plot). The greater temperature leads to more severe rank collapse, greater compression, higher representations norm, and worse OOD performance.

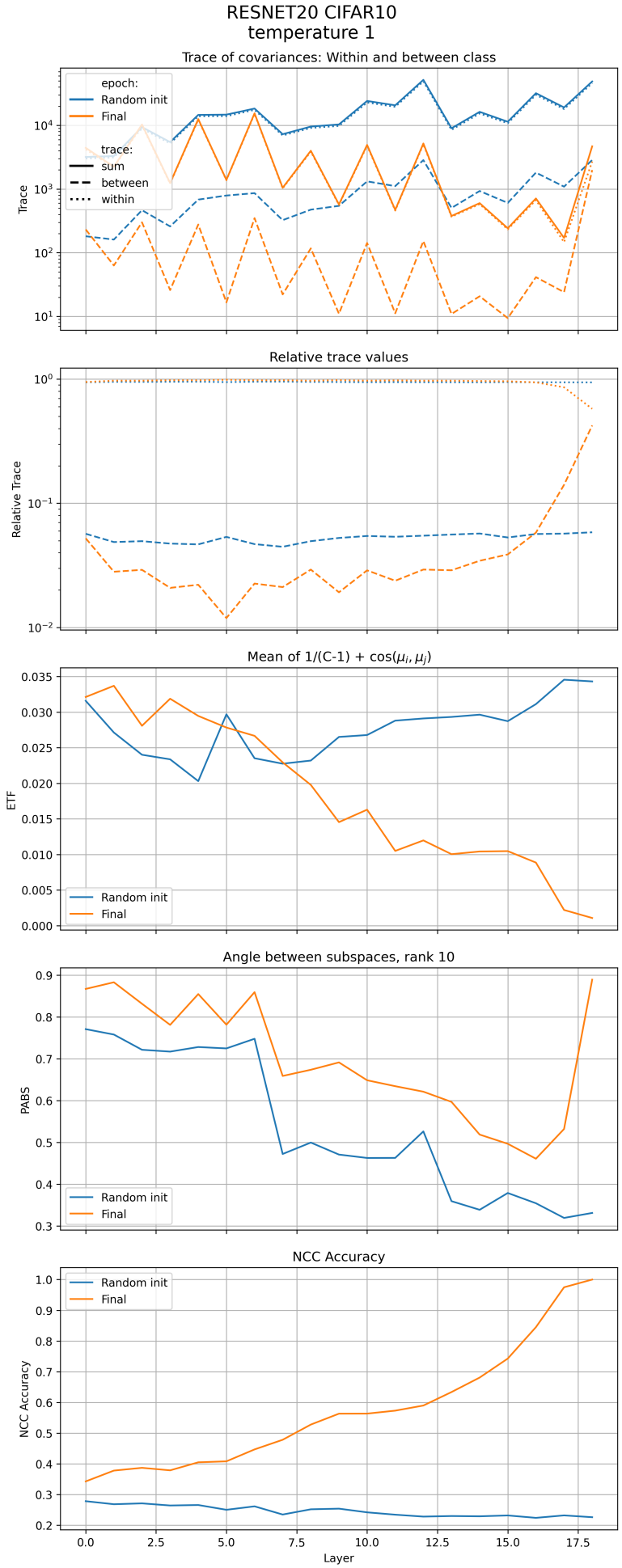

Figure 23: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet20 trained on CIFAR-10 with temperature 1. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

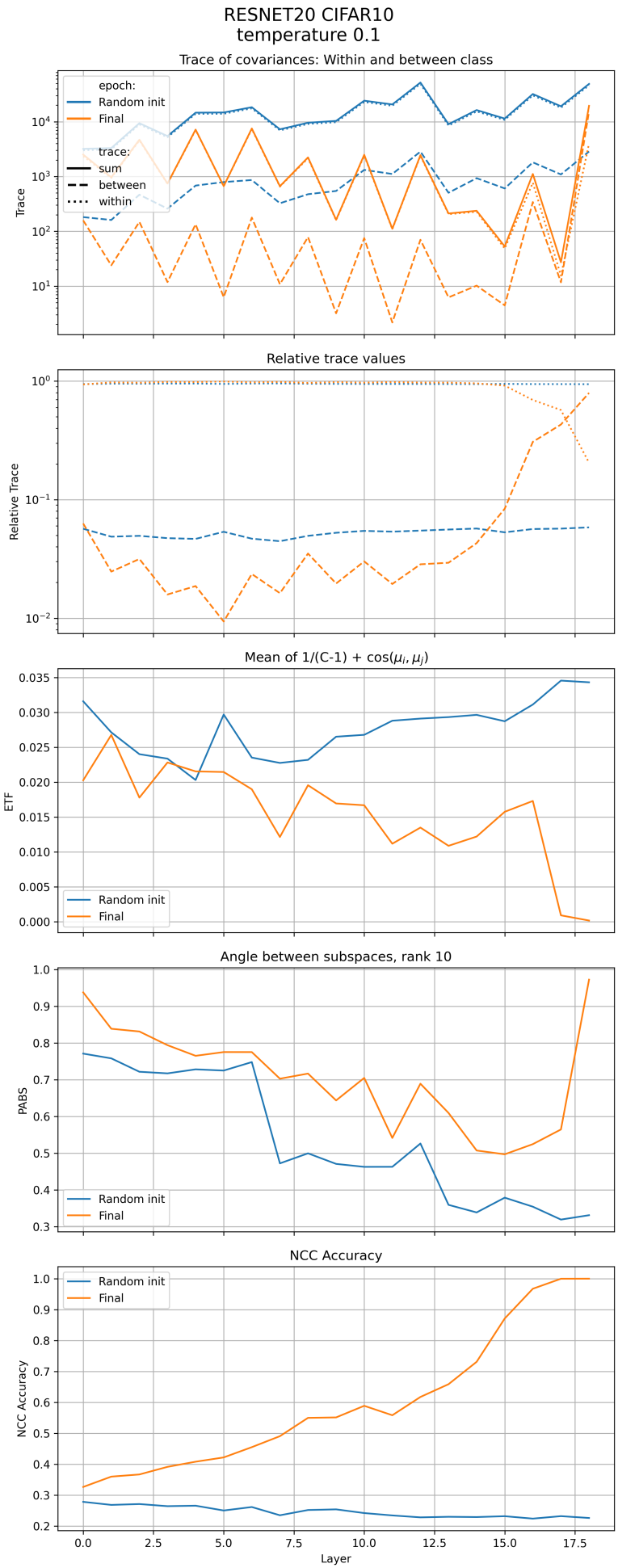

Figure 24: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet20 trained on CIFAR-10 with temperature 10. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

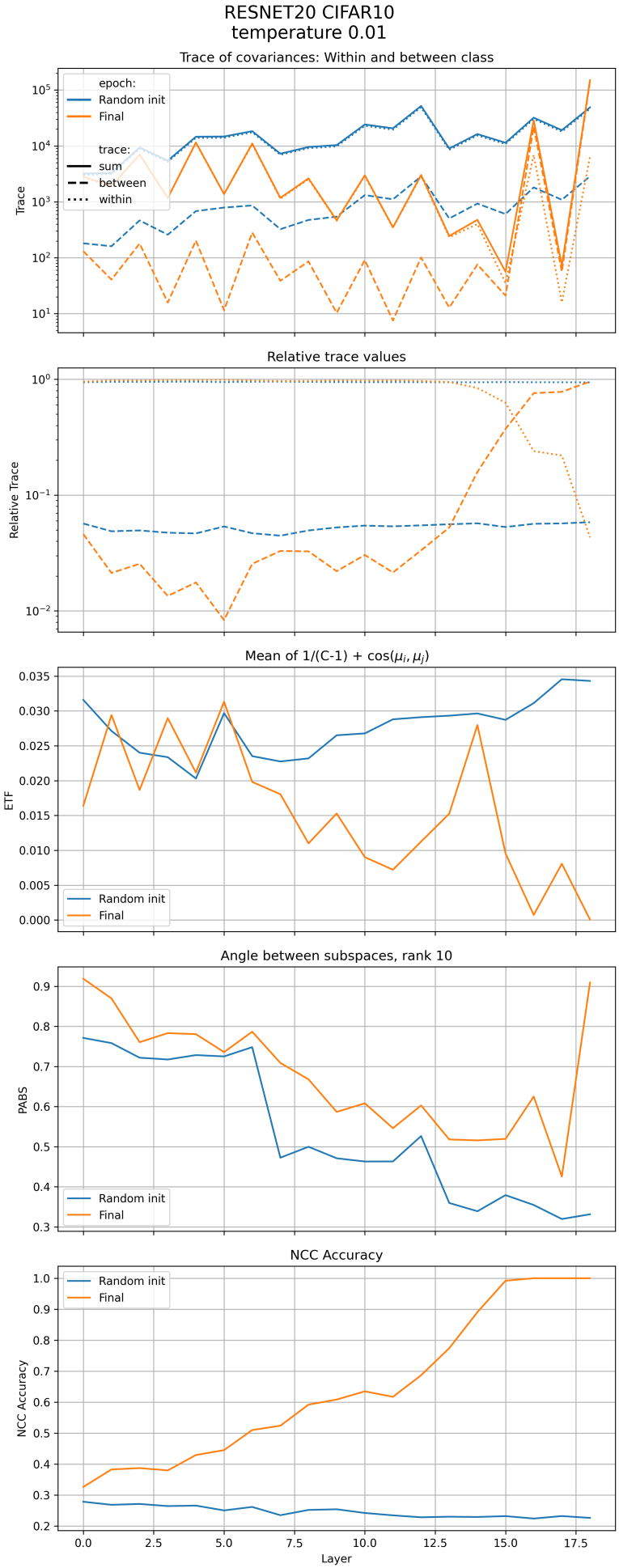

Figure 25: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet20 trained on CIFAR-10 with temperature 100. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

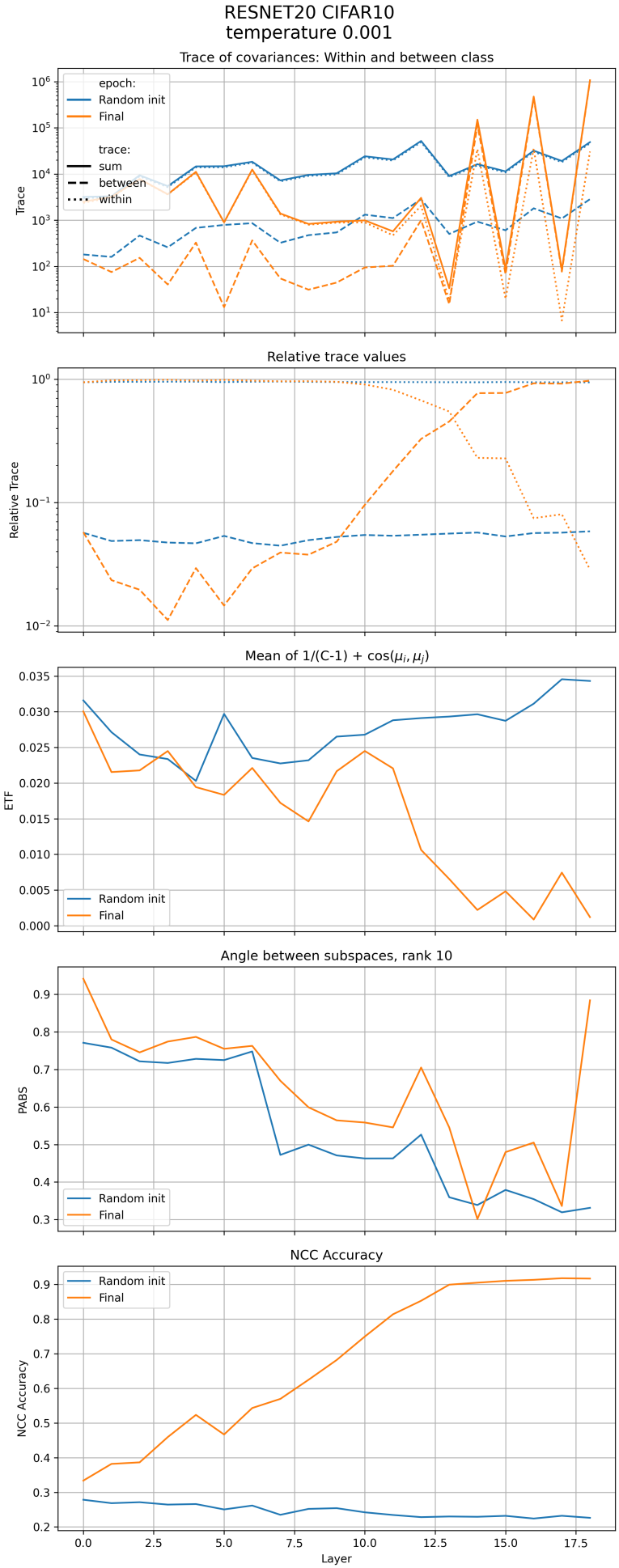

Figure 26: The measures of NC1-NC4 proposed in [Rangamani et al.(2023)Rangamani, Lindegaard, Galanti, and Poggio] obtained by ResNet20 trained on CIFAR-10 with temperature 1000. **Notice the title of the Figure wrongly says "temperature" while it is the inverse of the temperature!**

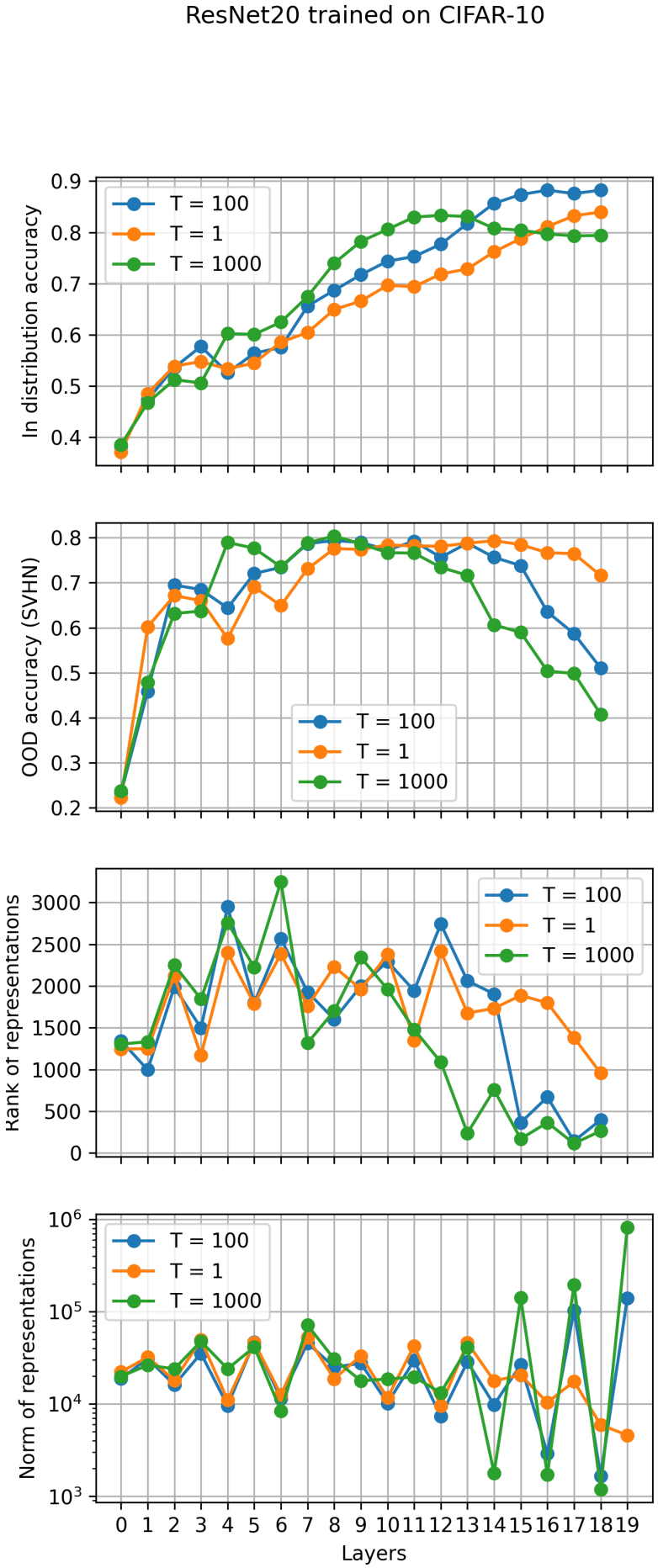

Figure 27: The plot presents the relationship between training temperature (different colors of the plots) and accuracy of linear probes attached to different layers for in-distribution generalization (top plot), out-of-distribution generalization (second top plot, evaluated at SVHN), rank of the representations at each layer (third top plot) and norm of the representations at each layer (bottom plot). The greater temperature leads to more severe rank collapse, greater compression, higher representations norm, and worse OOD performance.