Table 16: Correlation analyses of inconsistency and disagreement using the metrics from [1]. In the setting of Figures 5 and 7. Model-wise quantities were tested with one model per procedure.
Correlation measures: Two types of mutual information-based scores ($\mathcal{K}$ and '$|S|$=0'), and two types of Kendall's rank-correlation coefficient-based scores (Ψ and overall τ), as defined in Section 2.2 of [1]. In particular, $\mathcal{K}$ (defined by Equation (9) of the arXiv version or (14) of the ICLR version of [1]) is designed to remove the *unwanted/spurious correlation* that is caused by only a subset of the hyperparameters. A larger number indicates a higher correlation. The highest numbers (and near ties) are highlighted.
Tested quantities: 'Inconsist.': Inconsistency, 'Disagree.': Disagreement, 'Random' (baseline): random numbers drawn from the normal distribution, 'Canonical' (baseline): a slight extension of the canonical ordering in Section 4.1 of [1]; it heuristically determines the order of two procedures by preferring smaller batch size, larger weight decay, larger learning rate, and presence of data augmentation (which are considered to be associated with better generalization) by adding one point for each and breaking ties randomly.
Target quantities: Generalization gap (test loss minus training loss, as defined in the main paper), test error (suggested by Reviewer k85B), and test error minus training error (suggested by Reviewer ARdH).
Observation: The results are consistent with the submission and the previous work. The observed correlation of inconsistency to generalization gap is supported by our theorem, and the observed correlation of disagreement to test error is supported by the theorem of the original disagreement study. Regarding 'test error minus training error', on CIFAR-10, training error is relatively small and so it approaches test error, which explains why disagreement correlates well to it; on the other datasets, 'test error minus training error' is more related to 'test loss minus training loss', which explains why inconsistency correlates well to it.

	CIFAR-10				CIFAR-100				ImageNet			
	MI-based		Ranking		MI-based		Ranking		MI-based		Ranking	
	$\mathcal{K}$	$\lvert S\rvert$=0	Ψ	τ	$\mathcal{K}$	$\lvert S\rvert$=0	Ψ	τ	$\mathcal{K}$	$\lvert S\rvert$=0	Ψ	τ
Correlation to generalization gap ('test loss - training loss', analyzed in our Theorem 1)												
Inconsist.	**0.224**	**0.382**	**0.619**	**0.693**	**0.514**	**0.546**	**0.836**	**0.809**	**0.738**	**0.744**	**0.566**	**0.914**
Disagree.	0.148	0.273	0.581	0.595	0.122	0.120	0.521	0.402	0.115	0.166	0.470	0.470
Random	0.000	0.001	-0.017	-0.028	0.000	0.000	0.018	0.002	0.002	0.004	-0.037	-0.078
Canonical	0.009	0.109	0.231	0.381	0.000	0.093	0.247	0.356	0.002	0.072	0.510	0.312
Correlation to test error, suggested by Reviewer k85B												
Inconsist.	0.078	0.159	0.530	0.461	0.001	0.010	**0.471**	0.115	**0.036**	0.029	0.155	-0.200
Disagree.	**0.315**	**0.370**	**0.572**	**0.683**	**0.178**	**0.284**	0.457	**0.606**	**0.035**	**0.044**	**0.390**	**0.245**
Random	0.001	0.001	0.001	0.036	0.000	0.000	-0.034	-0.007	0.003	0.002	-0.073	0.051
Canonical	0.007	0.031	0.261	0.203	0.001	0.180	0.295	0.494	0.001	0.024	0.005	0.182
Correlation to 'test error - training error', suggested by Reviewer ARdH												
Inconsist.	0.104	0.253	**0.584**	0.574	**0.409**	**0.476**	**0.594**	**0.764**	**0.720**	**0.728**	**0.930**	**0.907**
Disagree.	**0.155**	**0.278**	**0.579**	**0.599**	0.141	0.130	0.275	0.417	0.120	0.159	0.572	0.461
Random	0.000	0.000	-0.020	-0.030	0.002	0.002	0.077	0.048	0.003	0.005	-0.176	-0.082
Canonical	0.004	0.074	0.270	0.317	0.001	0.087	0.141	0.345	0.001	0.075	0.370	0.310

References

[1] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.