The Best Deep Ensembles Sacrifice Predictive Diversity

Taiga Abe[*1] **E. Kelly Buchanan**[*1] **Geoff Pleiss**[1]

John P. Cunningham[1]
[1]Columbia University
{ta2507,ekb2154,gmp2162,jpc2181}@columbia.edu

Abstract

Ensembling remains a hugely popular method for increasing the performance of a given class of models. In the case of deep learning, the benefits of ensembling are often attributed to the diverse predictions of the individual ensemble members. Here we investigate a tradeoff between diversity and individual model performance, and find that—surprisingly—encouraging diversity during training almost always yields worse ensembles. We show that this tradeoff arises from the Jensen gap between the single model and ensemble losses, and show that Jensen gap is a natural measure of diversity for both the mean squared error and cross entropy loss functions. Our results suggest that to reduce the ensemble error, we should move away from efforts to increase predictive diversity, and instead we should construct ensembles from less diverse (but more accurate) component models.

1 Introduction

Ensembling, or combining the predictions of multiple models, yields lower error than using a single model in isolation [4, 7, 14, 22, 32]. In many instances, this phenomenon is a natural consequence of Jensen's inequality (though this connection is not often made in the literature). If error is measured by a strongly convex loss function ℓ (e.g. cross entropy, mean squared error, etc), and predictions from models $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$ are combined through averaging, we have:

$$\underbrace{\ell\big(\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\boldsymbol{f}_i(\boldsymbol{x}), y\big)}_{\text{ens. loss}} \leq \underbrace{\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\ell\big(\boldsymbol{f}_i(\boldsymbol{x}), y\big)}_{\text{avg. single model loss}} . \tag{1}$$

The degree to which ensembling helps depends on the looseness of the bound in Eq. (1). On one end of the spectrum, if all models make the same errors (i.e. $\boldsymbol{f}_1(\boldsymbol{x}) = \boldsymbol{f}_2(\boldsymbol{x}) = \ldots = \boldsymbol{f}_M(\boldsymbol{x}) \; \forall \boldsymbol{x}$), then the ensemble won't perform any better than the single models (i.e. Eq. (1) holds with equality). On the other end of the spectrum, if the component models make uncorrelated errors, the difference between ensembles and their average component model will be quite large [4, 21]. In this sense, the Jensen gap between terms in Eq. (1) quantifies a notion of "predictive diversity," which is touted as a necessary criterion for effective ensembles [7, 8, 22, 23]. All else being equal, training mechanisms which increase predictive diversity (widening the gap in Eq. 1) will improve ensemble performance.

Of course, any training mechanism that could widen the Jensen gap will also influence the performance of component models (i.e. the right side of Eq. (1)). Ultimately, both terms together determine ensemble performance (Fig. 1). Moreover, it is unlikely that there is some "magic" training procedure that simultaneously increases predictive diversity and lowers average single model loss, or even leaves single model loss unchanged (Fig. 1, left). Therefore, to improve ensemble performance, increases in the Jensen gap (predictive diversity) must *outweigh* any incurred increase in single model loss (as depicted in Fig. 1, center).

I Can't Believe It's Not Better Workshop at NeurIPS 2022.

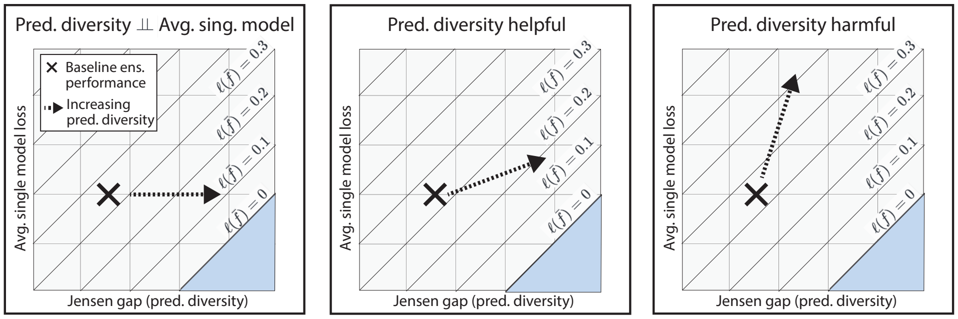

Figure 1: **Potential interactions between predictive diversity and single model performance.** Diagonal lines indicate level sets of ensemble performance ($\ell(\bar{f})$), determined by average single model performance *minus* the Jensen gap (Eq. (1)). We depict three potential scenarios. **Left (unrealistic):** we can increase predictive diversity without impacting single model performance. **Middle (what we hope):** increasing diversity negatively impacts single model performance, but ultimately improves the ensemble. **Right (reality):** increasing predictive diversity yields much worse component models, netting worse ensemble performance.

Within the context of modern deep ensembles, this paper offers evidence that, surprisingly, increasing the Jensen gap (i.e. predictive diversity) almost always yields worse ensemble performance, as any Jensen gap increases are outweighted by increased component model loss (as depicted in Fig. 1, right). This finding implies that—counterintuitively—*the path towards better deep ensembles involves less predictive diversity, not more*. In particular, we make the following contributions:

1. We show that standard ensemble training is equivalent to empirical risk minimization with a regularization term that minimizes the Jensen gap in Eq. (1); i.e. a term that *penalizes predictive diversity*. Surprisingly, reducing this diversity penalty to increase the Jensen gap significantly harms ensemble performance. Even more surprisingly, upweighting this diversity penalty *improves* performance. Analyzing these results, we find that encouraging predictive diversity (by reducing the penalty) incurs a substantial increase in component model loss, resulting in net worse ensemble performance.
2. Across a variety of model architectures, we find the the best deep ensembles are often the least diverse, instead relying on strong component models. These results hold for standard as well as heterogeneous deep ensembles (combinations of multiple architectures/training procedures). In the full version of this paper, we will extend these results to other training procedures that explicitly encourage diversity [e.g. 11, 23, 25, 27–29, 35, 36, 39, 41].

Together, these results suggest that the best strategy to improve deep ensembles is reducing the error of component models, not increasing predictive diversity.

2 Quantifying Predictive Diversity Via Jensen Gap

Here we examine the Jensen gap (Eq. 1) in further depth as a notion of diversity for two common loss functions. Though previous works propose alternative metrics for diversity [e.g. 20, 21, 29], the Jensen gap is exactly the performance advantage of an ensemble over the average single component model, and thus definitionally demonstrates that predictive diversity is good for ensembles. Moreover we find correspondences between the Jensen gap and other diversity metrics, suggesting that these metrics are interchangeable to a large extent.

Example 1: Mean squared error (MSE) loss. If we take ℓ to be the standard mean squared error loss (or equivalently, the Brier Score [6]), the Jensen gap in Eq. (1) captures variance across model predictions. Given M models $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_M$:

$$\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[\,(\boldsymbol{f}_i(\boldsymbol{x}) - y)^2\,\right] - \left[\left(\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\boldsymbol{f}_i(\boldsymbol{x}) - y\right)^2\right] = \left(\tfrac{M-1}{M}\right)\,\overline{\mathrm{Var}}\left[\boldsymbol{f}_i(\boldsymbol{x})\right], \tag{2}$$

where $\overline{\mathrm{Var}}[\boldsymbol{f}_i(\boldsymbol{x})]$ is the sample variance (see Appx. A.1 for a derivation). Thus the Jensen gap coincides with a natural notion of diversity used in previous work [e.g. 18, 29].

Example 2: Cross entropy (CE) loss. If we instead take ℓ to be the cross entropy loss, the Jensen gap does not correspond to a term that is as immediately recognizable:

$$\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[-\log \boldsymbol{f}_i^{(y)}(\boldsymbol{x})\right] + \left[\log \tfrac{1}{M}\textstyle\sum_{i=1}^{M}\boldsymbol{f}_i^{(y)}(\boldsymbol{x})\right] = \tfrac{1}{M}\textstyle\sum_{i=1}^{M}\log\left(\tfrac{1}{M}\right) - \log\left(\frac{\boldsymbol{f}_i^{(y)}(\boldsymbol{x})}{\sum_j \boldsymbol{f}_j^{(y)}(\boldsymbol{x})}\right), \quad (3)$$

where $\boldsymbol{f}_i^{(y)}$ is the probability assigned the correct class by model $\boldsymbol{f}_i$. Nevertheless, the right hand side of Eq. (3) can be interpreted as an information theoretic quantification of ensemble diversity (see Appx. A.2), which—as we demonstrate in Appx. D—correlates well with other popular diversity metrics like average pairwise model correlation [e.g. 7, 8, 21].

3 Standard Ensemble Training Penalizes Predictive Diversity

A standard deep ensemble consists of M independently trained models. With the Jensen gap notion of predictive diversity, we immediately notice something troubling about this standard setup. Independently optimizing $\boldsymbol{f}_1, \ldots \boldsymbol{f}_M$ is equivalent to optimizing $\mathcal{L}_{\text{ens.}} \triangleq \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{p(\boldsymbol{x},y)}\left[\ell\left(\boldsymbol{f}_i(\boldsymbol{x}), y\right)\right]$. Importantly, this form differs from the risk used to evaluate the ensemble, $\mathbb{E}_{p(\boldsymbol{x},y)}[\ell(\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{f}_i(\boldsymbol{x}), y)]$. The difference between the training and evaluation risk is *exactly* the Jensen gap.

Example 3: Training a deep ensemble with MSE penalizes variance. In particular, if our goal is to apply deep ensembles to a regression task, our training objective can be written as:

$$\mathcal{L}_{\text{ens.}} = \mathop{\mathbb{E}}_{p(\boldsymbol{x},y)}\left[\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}) - y)^2\right]\right] = \mathop{\mathbb{E}}_{p(\boldsymbol{x},y)}\left[\left(\bar{\boldsymbol{f}}(\boldsymbol{x}) - y\right)^2 + \tfrac{M-1}{M}\,\overline{\text{Var}}\left[\boldsymbol{f}_i\right]\right]. \quad (4)$$

where $\bar{\boldsymbol{f}}(\boldsymbol{x}) \triangleq \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{f}_i(\boldsymbol{x})$ is the ensemble prediction. In other words, standard ensemble training amounts to *penalizing* the variance between ensemble members. More generally, we can write:

$$\mathcal{L}_{\text{ens.}} = \mathop{\mathbb{E}}_{p(\boldsymbol{x},y)}\Big[\underbrace{\ell\left(\bar{\boldsymbol{f}}(\boldsymbol{x}), y\right)}_{\text{ensemble risk}} + \underbrace{\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\ell\left(\boldsymbol{f}_i(\boldsymbol{x}), y\right) - \ell\left(\bar{\boldsymbol{f}}(\boldsymbol{x}), y\right)}_{\text{Jensen gap (pred. diversity)}}\Big]. \quad (5)$$

Eq. (5) frames standard ensemble training as regularized empirical risk minimization. The first term in Eq. (5) measures the risk of the ensemble, which is ultimately what we care about minimizing. The second term is the Jensen gap, which is always greater than zero and thus can be interpreted as a regularizer. Surprisingly then, although diversity (a large Jensen gap) is beneficial for ensemble performance, the standard ensemble training objective favors a *small* Jensen gap.

4 Encouraging Predictive Diversity During Training Hurts Performance

Hypothesis: downweighting the predictive diversity penalty should improve ensembles. Eq. (5) shows that standard ensemble training appears to be regularizing *against* diversity inherent in our choice of ensemble risk. Eq. (5) further suggests a straightforward way to increase ensemble diversity: simply reduce the strength of this regularization. We would hope the corresponding increase in the Jensen gap would then contribute directly to an improvement in ensemble performance.

Experimental setup. We consider multiclass classification problems with inputs $\boldsymbol{x} \in \mathbb{R}^D$ and targets $y \in [1, \ldots, C]$. To measure the effect of downweighting the predictive diversity penalty, we train ensembles of $M = 4$ networks on the CIFAR10 training dataset [19] with an additional hyperparameter γ, which controls for the strength of the Jensen gap diversity regularizer:

$$\mathcal{L}_{\gamma\text{-ens.}} = \mathop{\mathbb{E}}_{p(\boldsymbol{x},y)}\big[\underbrace{-\log \bar{\boldsymbol{f}}^{(y)}(\boldsymbol{x})}_{\text{ensemble risk}} + \gamma\,\underbrace{\big(\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[\log\left(\tfrac{1}{M}\right) - \log\left(\frac{\boldsymbol{f}_i^{(y)}(\boldsymbol{x})}{\sum_j \boldsymbol{f}_j^{(y)}(\boldsymbol{x})}\right)\right]}_{\text{C.E. Jensen gap (pred. diversity)}}\big)\big)\big] \quad (6)$$

If we set $\gamma = 1$, then Eq. (6) corresponds to the unmodified ensemble training (Eq. 3) objective. If we instead set $\gamma < 1$, we expect that the resulting ensemble will have a larger Jensen gap (i.e. more predictive diversity) than a standard ensemble, which ideally should result in improved predictive performance (Fig. 1, middle). Conversely, we would expect that setting $\gamma > 1$ would further hurt ensemble performance, as it would yield a smaller Jensen gap and less predictive diversity.

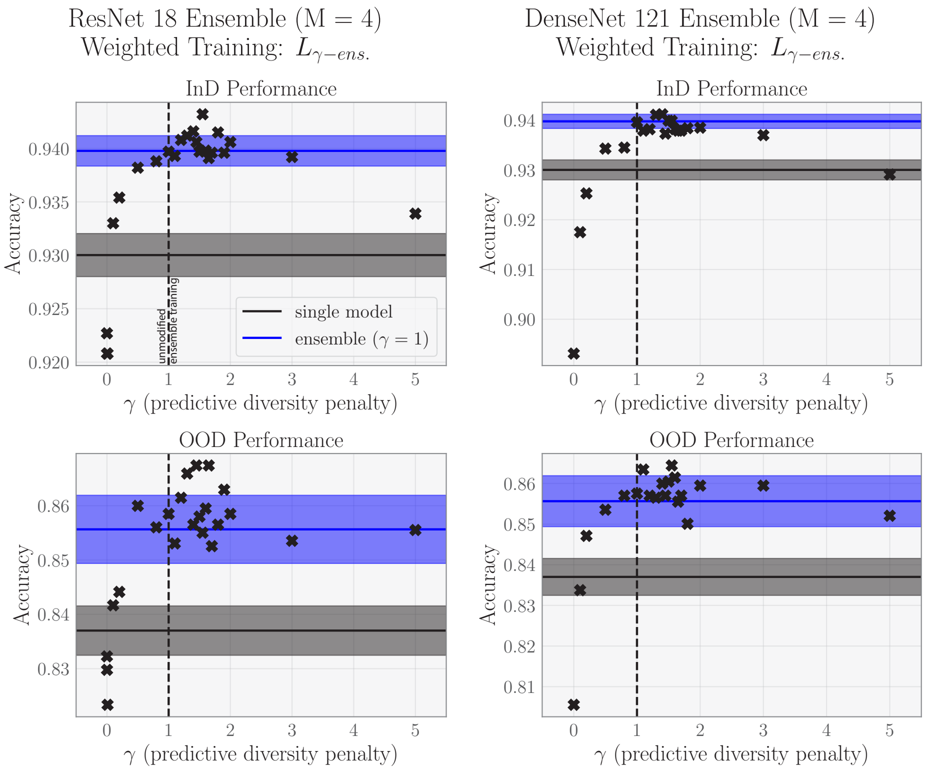

Figure 2: **Ensembles performance improves as we penalize predictive diversity**. Top panels represent performance on InD test data (CIFAR10), bottom panels represent OOD test performance (CIFAR10.1). In each subplot, the bands correspond to the mean ± 2 standard error (N=4) around the performance of either: a single network (black), or an ensemble (blue) trained with the objective function in Eq. (6), with γ (the strength of the diversity penalty) set to 1 (i.e. unmodified ensemble training). Each × represents the accuracy achieved using different weights γ for the Jensen gap term in Eq. (5), with vertical line indicating $\gamma = 1$. Columns: results for ResNet18 (left), and DenseNet121 (right). For full details see Appx. B.

We evaluate ensembles of two different network architectures—ResNet18 [15] and DenseNet121 [17], for $\gamma \in [0, 5]$. We initialize each ensemble's component networks using different random seeds. During training iterations, we optimize a standard minibatched approximation of Eq. (6). Because Eq. (6) trains ensemble members jointly, each component network sees the same minibatch of data and the same per sample data augmentations. We note that this protocol differs from standard deep ensemble training procedures [e.g. 22, 8], where component networks are trained with different augmentations and minibatch orderings - a point we will return to at the end of this section. A full description of experimental details can be found in Appx. B.

In-distribution (InD) performance. We first evaluate our trained ensembles on the CIFAR10 test set, which we indicate as the InD test set. The blue line and bands in the top row of Fig. 2 show the mean ± two standard error for trained ensemble performance, corresponding to models trained following Eq. (6) with $\gamma = 1$. The black line and bands correspondingly gives the mean ± two standard error for trained single model performance. Against these baselines, Fig. 2 shows the performance of individual ensembles trained with different values of γ. Surprisingly, ensembles with a downweighted diversity penalty ($\gamma < 1$) have *much worse performance* than the ensemble trained with unmodified objective ($\gamma = 1$). When γ is close to zero (i.e. training directly on the ensemble risk), performance drops below that of a single model with the same architecture. This finding suggests that regularizing *against* predictive diversity is a critical component of deep ensemble performance- without this regularization, deep ensembles are easily outperformed by a single component model. Although not shown here, we found that further reducing the value of $\gamma < 0$ leads to divergent losses. In striking contrast, Fig. 2 shows that further penalizing diversity

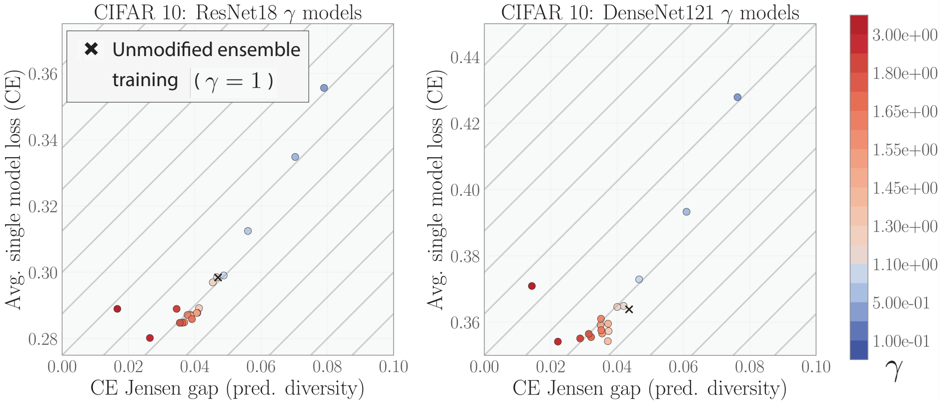

Figure 3: **Average individual model performance versus predictive diversity.** Each marker represents an ensemble from Fig. 2 (left: ResNet18 ensembles; right: DenseNet121). Colors give the strength of the predictive diversity regularization, γ. Warmer colors correspond to high γ values (stronger predictive diversity penalty), while the cooler colors correspond to low γ values (weaker predictive diversity penalty). Standard deep ensemble training ($\gamma = 1$) is denoted by $\times$. Diagonal lines denote level sets of ensemble performance $\ell(\bar{\boldsymbol{f}})$.

(i.e. $1 < \gamma < 2$) does not harm performance, but surprisingly can yield *better accuracy* than standard ensembles, with performance dropping off far more slowly as γ is increased further.

Out-of-distribution (OOD) performance. Beyond InD test performance, deep ensembles have often been discussed as a benchmark method in settings where test data shifts significantly away from the training distribution [12, 22, 30]. Although this benefit is disputed by more recent work [1], we evaluated all of our trained ensembles and individual models on CIFAR10.1, a shifted test dataset for models trained on CIFAR10 [31]. The bottom row of Fig. 2 emphasizes the same conclusions we saw on InD data: a stark drop in ensemble performance with a lower diversity penalty, and an increase in performance with an increased penalty for $1 > \gamma > 2$, relative to $\gamma = 1$ unmodified training. This result supports the conclusions of [1], which suggests that ensemble performance OOD is very tightly coupled to performance InD.

Comparison with standard ensemble training. Compared to the $\gamma = 1$ ensembles which we train here, standard ensemble training introduces several additional sources of stochasticity: independent orderings of minibatches to each ensemble member during training, and independent data augmentations. In Fig. 5, we show that these additional sources of randomness offer a boost in performance over our $\gamma = 1$ models shown in Fig. 2. However, we can still find settings of γ which generate ensembles that match or outperform these standard ensembles, across models and data types.

5 Analysis: The Trade-off Between Diversity and Single Model Performance

Having shown that encouraging predictive diversity in a training objective leads to worse ensemble performance, we next use the decomposition in Eq. (5) to more directly evaluate the contributions of single model performance and predictive diversity to this ensemble performance. From the terms in Eq. (5), we can see that are *level sets* of the ensemble loss: balanced tradeoffs between Jensen gap (i.e. predictive diversity) and single model loss that give ensembles of equivalent performance. These level sets determine the transition between the desired scenario of helpful predictive diversity (Fig. 1, middle), and the worst-case scenario of harmful diversity (Fig. 1, right).

Unfortunately, our results show that the behavior of deep ensembles is best described by this worst case scenario. Fig. 3 shows the predictive diversity (Jensen gap) versus average single model cross entropy loss in Eq. (6) for the models in Fig. 2, evaluated on the CIFAR10 test set. The cross in Fig. 3 depicts the standard ensemble ($\gamma = 1$). We first observe that for ensembles trained with $\gamma < 1$ (blue dots), there is indeed an increase in the diversity of predictions among trained ensemble members, and a decrease in diversity of predictions for ensembles trained with $\gamma > 1$ (red dots). Across all regularization strengths and model architectures, we find that any meaningful increase in predictive

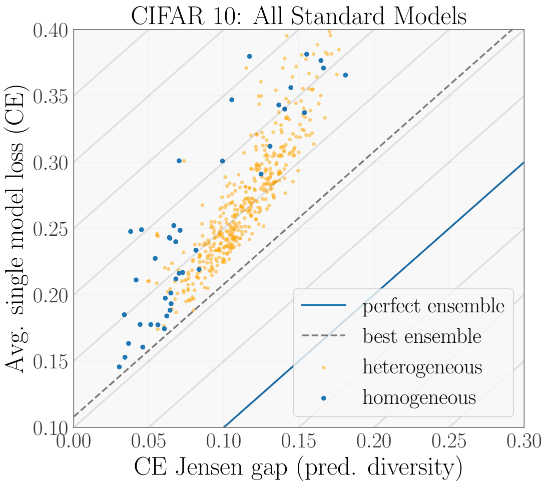

Figure 4: **Average individual model performance versus diversity for different ensembles trained on CIFAR10.**. Each dot corresponds to an ensemble of homogeneous models (blue) or heterogeneous models (orange). The dashed black line is the level set for the best ensemble model. The blue line is the level set for an ideal perfect ensemble, which has an error of 0.

diversity comes at a cost of a large decrease in average single model performance, negating any diversity benefits. Conversely, a moderate sacrifice in predictive diversity yield net improvements to ensemble performance, due to an outsize improvement in average single model performance.

Generalizing to deep ensembles used in practice. The results shown thus far strictly control the diversity of ensemble models through a single parameter, γ. In practice, it could be the case that ensembles based on different architectures and training paradigms trade off predictive diversity and average single model loss in qualitatively different ways to achieve their performance. To test this hypothesis, we conducted a large scale study with 45 sets of standard deep ensembles, across different architectures and training hyperparameters (See Appx. B for additional details). Fig. 4 illustrates the average single model cross entropy versus predictive diversity for these ensembles as blue dots. Surprisingly, the best performing ensembles are well predicted by average single model performance, and have the lowest levels of ensemble diversity. The orange dots in Fig. 4 show the result of forming *heterogeneous* ensembles, i.e. combining models of different architectures and training to create deep ensembles. Because heterogeneous ensembles introduce greater diversity across model architectures without changing the training objective each individual model, we might hope that there is a greater potential for diversity based improvements. It is somewhat surprising, therefore that our conclusions hold for these heterogeneous deep ensembles as well: average single model performance is the dominant factor in determining heterogeneous ensemble performance, and the best performing ensembles have very low levels of predictive diversity.

6 Discussion and Open Questions

Throughout this paper, all of our attempts to increase predictive diversity in deep ensembles end up harming the component models so severely that the resulting ensembles underperform the standard ensemble baseline. In the full paper, we will apply these same analytical tools to models that have been trained against other diversity-based regularizers [e.g. 27, 29, 35, 41]. We have reason to believe that we will arrive at similar results in these settings: previous attempts to increase other notions of predictive diversity in ensemble training have yielded worse ensemble performance in large neural network models [23, 29].

More broadly, our results suggest a qualitative difference between deep ensembles and classical ensembles of weak learners. Looking beyond models trained with a regularized Jensen gap, we find that the best performing deep ensembles used in practice are also the least diverse. In stark contrast, increasing predictive diversity generates improvements in ensembles of weak learners [e.g. 4, 5, 7, 9]. We have reason to believe that the overparameterized nature of neural networks contributes to this phenomenon: recent work suggests that overparameterized models have similar inductive biases [2, 24], and large single models already perform similarly to deep ensembles [1, 3]. In the full version of this paper, we plan to analyze how the "strength" of the base model class affects the diversity-single model performance tradeoff.

References

[1] Taiga Abe, E Kelly Buchanan, Geoff Pleiss, Richard Zemel, and John P Cunningham. Deep ensembles work, but are they necessary? *arXiv preprint arXiv:2202.06985*, 2022.

[2] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

[3] Jeremy Bernstein, Alex Farhang, and Yisong Yue. Kernel interpolation as a bayes point machine. *arXiv*, 2021.

[4] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[7] Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[8] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[9] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2):256–285, 1995.

[10] Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

[11] Raphael Gontijo-Lopes, Yann Dauphin, and Ekin Dogus Cubuk. No one representation to rule them all: Overlapping features of training methods. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BK-4qbGgIE3.

[12] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

[13] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5927–5935, 2017.

[14] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *Transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[20] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

[21] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

[23] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David Crandall, and Dhruv Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

[24] Horia Mania and Suvrit Sra. Why do classifier accuracies show linear trends under distribution shift? *arXiv preprint arXiv:2012.15483*, 2020.

[25] Andres Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 33:5479–5491, 2020.

[26] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 2021.

[27] Aaron Mishtal and Itamar Arel. Jensen-shannon divergence in ensembles of concurrently-trained neural networks. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 558–562. IEEE, 2012.

[28] Jeremy Nixon, Balaji Lakshminarayanan, and Dustin Tran. Why are bootstrapped deep ensembles not better? In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*. openreview.net, December 2020.

[29] Luis A. Ortega, Rafael Cabañas, and Andres Masegosa. Diversity and generalization in neural network ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11720–11743. PMLR, 2022.

[30] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

[32] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition*, 2015.

[35] Andrew Webb, Charles Reynolds, Wenlin Chen, Henry Reeve, Dan Iliescu, Mikel Lujan, and Gavin Brown. To ensemble or not ensemble: When does end-to-end training fail? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 109–123. Springer, 2020.

[36] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. In *Conference on Neural Information Processing Systems*, June 2020.

[37] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[38] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[39] Sheheryar Zaidi, Arber Zela, Thomas Elsken, Chris Holmes, Frank Hutter, and Yee Whye Teh. Neural ensemble search for uncertainty estimation and dataset shift. In *Advances in Neural Information Processing Systems*, 2021.

[40] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2736–2746, 2022.

[41] Shuai Zhao, Liguang Zhou, Wenxiao Wang, Deng Cai, Tin Lun Lam, and Yangsheng Xu. Towards better accuracy-efficiency trade-offs: Divide and co-training. *IEEE Transactions on Image Processing*, 2022.

A Derivation of the Jensen Gap for Common Loss Functions

A.1 The Mean Squared Error Jensen Gap

In Sec. 2, we claim that the Jensen gap for the mean squared error loss is a scaled version of the sample variance (Eq. 2). The following is a derivation of that fact:

$$
\begin{aligned}
&\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left(\boldsymbol{f}_i(\boldsymbol{x})-y\right)^2-\left(\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\boldsymbol{f}(\boldsymbol{x})-y\right)^2\\
&=\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-2(\boldsymbol{f}_i(\boldsymbol{x}))y+y^2\right]-\left[(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2-2(\bar{\boldsymbol{f}}(\boldsymbol{x}))y+y^2\right]\\
&=\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-2(\boldsymbol{f}_i(\boldsymbol{x}))y\right]-\left[(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2-2(\bar{\boldsymbol{f}}(\boldsymbol{x}))y\right]\\
&=\tfrac{1}{M}\textstyle\sum_{i=1}^{M}(\boldsymbol{f}_i(\boldsymbol{x}))^2-\tfrac{1}{M}\textstyle\sum_{i=1}^{M}2(\boldsymbol{f}_i(\boldsymbol{x}))y-(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2+2(\bar{\boldsymbol{f}}(\boldsymbol{x}))y\\
&=\frac{1}{M}\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2\right]\\
&=\frac{M-1}{M}\left[\frac{1}{M-1}\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2\right]\right]\\
&=\frac{M-1}{M}\,\overline{\mathrm{Var}}\left[\boldsymbol{f}_i(\boldsymbol{x})(\boldsymbol{x})\right],
\end{aligned}
$$

where $\bar{\boldsymbol{f}}(\boldsymbol{x})=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{f}_i(\boldsymbol{x})$ is the sample mean and $\overline{\mathrm{Var}}$ is the sample variance.

A.2 Decomposition of Cross Entropy Jensen Gap

In Sec. 2, we claim (Eq. 3) that the Jensen gap is an information theoretic quantification of diversity. The Jensen gap is given by:

$$
\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\left[\mathrm{CE}(\boldsymbol{f}_i(\boldsymbol{x}),\ y)\right]-\left[\mathrm{CE}(\bar{\boldsymbol{f}}(\boldsymbol{x}),\ y)\right]\\
&=\frac{1}{M}\sum_{i=1}^{M}\left[-\log\boldsymbol{f}_i(\boldsymbol{x})^{(y)}\right]-\left[-\log\bar{\boldsymbol{f}}(\boldsymbol{x})^{(y)}\right]\\
&=\frac{1}{M}\sum_{i=1}^{M}\left[-\log\boldsymbol{f}_i(\boldsymbol{x})^{(y)}\right]+\log\left[\frac{1}{M}\right]+\log\left[\sum_{j=1}^{M}\boldsymbol{f}_j(\boldsymbol{x})^{(y)}\right]\\
&=\sum_{i=1}^{M}\frac{1}{M}\left[\log\frac{1}{M}-\log\left[\frac{\boldsymbol{f}_i(\boldsymbol{x})^{(y)}}{\sum_{j=1}^{M}\boldsymbol{f}_j(\boldsymbol{x})^{(y)}}\right]\right] \qquad (7)
\end{aligned}
$$

Eq. (7) can be interpreted as the D_{KL} divergence between two categorical distributions. The first distribution represents the probability of sampling an ensemble member uniformly at random ($1/M$). The second distribution represents the probability of sampling an ensemble member proportional to its correct class prediction. Eq. (7) will be minimized when these two distributions are equal, which will only happen if all the ensemble members predict the correct class equally. Conversely, Eq. (7) will be larger when the component model predictions differ from one another.

B Experimental details

B.1 Code and data

Code to train models with a diversity regularization and plot Figure 2 can be found in `https://anonymous.4open.science/r/ensemble_attention-6178/README.md`. Model training can be performed by running the script: `scripts/run.py`. Model and regularization strength can be changed as parameters, for example by running: `python scripts/run.py ++classifier=resnet18 ++gamma=0.5`.

In this repository, the code to generate Figure 2 is located in:

- `scripts/vis\_scripts/kl\_weight\_resnet18\_cifar.py`
- `scripts/vis\_scripts/kl\_weight\_densenet121\_cifar.py`

Code to train all other models, and plot all remaining figures can be found in https://anonymous.4open.science/r/interp_ensembles-B485/README.md where model training code is based on implementations from https://github.com/huyvnphan/PyTorch_CIFAR10 and https://github.com/FreeformRobotics/Divide-and-Co-training.

In this repository, the code to create Figure 3, 4, 5, and 6 can be found at the following locations:

- `scripts/paperfigs/estimate\_biasvar\_ce\_resnet\_zoom.py` (Figure 3,5)
- `scripts/paperfigs/estimate\_biasvar\_ce\_densenet\_zoom.py` (Figure 3,5)
- `scripts/paperfigs/estimate\_biasvar\_ce\_allmodels.py` (Figure 4)
- `scripts/paperfigs/estimate\_biasvar\_ce\_allmodels\_compare.py` (Figure 6)

Finally, relevant data can be found in this Zenodo repository: https://zenodo.org/record/6582653#.Yo7R0y-B3fZ, as previously shared by [1]. This data provides the logit outputs from individual models (and some ensembles) on the in and out of distribution data that we consider. These data are referenced in the code above.

B.2 Figures 2, 3, and 5

All ensembles and single networks depicted in Figures 2,3 and 5 were trained with the same training schedule and hyperparameters: 100 epochs of training, with batch size 256, learning rate $1e-2$, weight decay $1e-2$. The optimizer used was SGD with momentum, with a linear warmup of 30 epochs and cosine decay. The best performing checkpoint on validation data was selected after training and used for further evaluation.

In all experiments, we created ensembles of size $(M = 4)$. The SEM bands in Fig. 2 and Fig. 5 are generated using $(N = 4)$ independently trained ensembles or samples; no models are reused between ensembles and single models in Fig. 2, while in Fig. 5 ensembles of size $M = 4$ are created by subselecting models from a collection of 5 independently trained models.

B.3 Figures 4 and 6

Fig. 4 includes three separate subsets of ensemble types. The same models are used to generate Fig. 6.

The first subset consists of models that were trained with the same hyperparameters as in Appx. B.2. These models are all of size $M = 5$, of the following architectures:

- ResNet 18 [15]
- WideResNet 18-2, 18-4, 28-10 [38]
- GoogleNet, Inception v3 [34]
- VGG with 11 and 19 layers [33]
- DenseNet 121 and 169 [17]

The second subset consists of model implementations derived from the code released with [41], found here https://github.com/FreeformRobotics/Divide-and-Co-training. These ensembles of larger models are of the following architectures:

- Pyramidnet110 (M=4) [13]
- SeResNet-164 (M=5) [16]

- ShakeShake26_2x96d (M=5) [10]
- ResNexst50_4x16d (M=5) [37, 40, 41]

For these models, we changed the learning rate to $1e-1$, and the weight decay to $1e-4$.

Finally, we used 36 separate sets of ensembles from Miller et al. [26], and we thank the authors for graciously sharing these results with us.

To generate heterogeneous ensembles, we first recorded the distribution of ensemble sizes among these separate subsets of ensembles. We then randomly shuffled all of our individual models together, and divided them into new deep ensembles, respecting the distribution of ensemble sizes seen in the original set of homogeneous ensembles. We computed all metrics of interest on 10 random shuffles of ensemble members, and plotted the spread across those random shuffles as the orange dots in Figs. 4 and 6.

C Performance comparison to standard ensemble training

In Fig. 2, we compare models regularized with $\gamma < 1$ or $\gamma > 1$ directly to models trained with $\gamma = 1$: i.e., an ensemble trained in parallel with Eq. (5). Here in Fig. 5, we compare these same γ regularized models to standard ensemble training, where each ensemble member is trained in serial. In particular, standard ensemble training presents data to each ensemble member in a different minibatch ordering, with different augmentations. As we can see in Fig. 5, this results in a slightly stronger baseline. Nevertheless, we find that the same conclusions largely hold as in Fig. 2: ensembles trained with $1 < \gamma < 2$ can match or outperform standard ensemble training, on both InD and OOD data.

D Comparison of CE Jensen gap to a standard diversity measure

In Eq. (3), we describe the Jensen gap notion of predictive diversity for the cross entropy loss. As this is a nonstandard notion of predictive diversity, we directly compare the diversity computed from this metric to that of a more standard one for classification tasks, based on average pairwise correlation between ensemble members. This notion of diversity has previously been discussed in [e.g. 7, 21, 8]. In particular, given a set of models , $\boldsymbol{f}_i, i \in [1 \dots M]$ we define this diversity measure based on the average pairwise correlation as:

$$1 - \textbf{avg. pairwise corr.} = \mathbb{E}_{\boldsymbol{f}_i, \boldsymbol{f}_j} \left[\mathbb{E}_{\boldsymbol{x}} \left[\mathbb{1}[\textit{pred}(\boldsymbol{f}_i(\boldsymbol{x})) = \textit{pred}(\boldsymbol{f}_j(\boldsymbol{x}))] \right] \right] \tag{8}$$

Where $\textit{pred}(\boldsymbol{f}_i(\boldsymbol{x}))$ returns the index of the class predicted as the label for image $\boldsymbol{x}$ by network $\boldsymbol{f}_i$. In Fig. 6, we apply this diversity measure, as well as the cross entropy instantiation of the Jensen gap to deep ensembles as constructed for Fig. 4. Altogether, we see a strong correlation between these notions of diversity (black line gives linear fit).

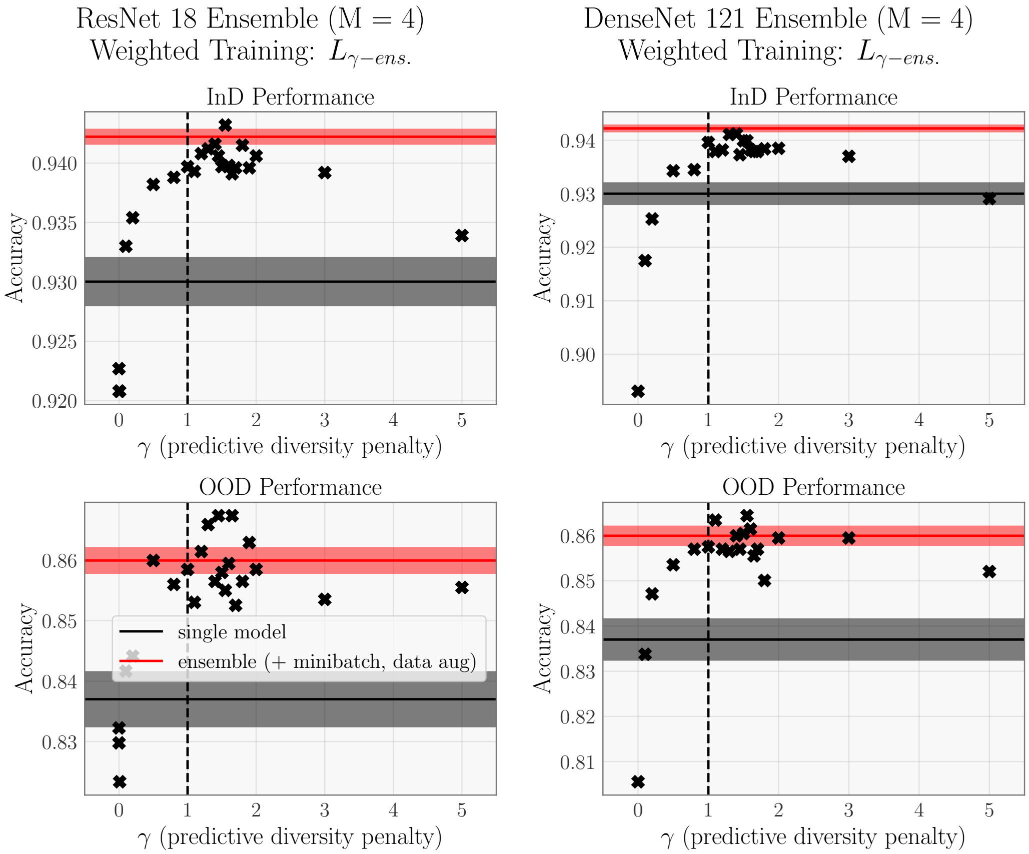

Figure 5: **Penalizing predictive diversity can match standard ensemble training**. As in Fig. 2, each panel represents the test performance of ensembles trained on CIFAR 10 with different γ regularization strengths. The top panels represent test performance on CIFAR10, i.e. the InD test set. The bottom panels represent performance on CIFAR10.1 [31], an OOD test dataset. Left panels give results for ResNet 18, and right panels for DenseNet 121. The $\times$ represents the accuracy of an ensemble trained with the corresponding setting of γ, the diversity regularization, given on the x-axis. The single model baseline performance in the black band (giving mean $\pm$ 2 standard error) is the same as in Fig. 2. We construct a standard ensemble training baseline by training five individual models of each network, and selecting subsets of $M = 4$ models from which we form ensembles, giving us $N = 4$ sets of ensembles that share ensemble members. The red band gives the mean $\pm$ 2 standard error around the performance of these $N = 4$ ensembles.

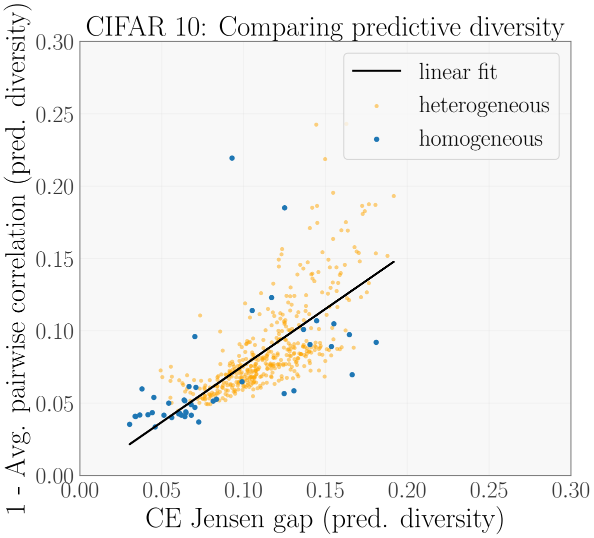

Figure 6: **CE Jensen gap is correlated with standard measures of diversity**. Here we depict deep ensembles trained on CIFAR10, and compute their diversity on the CIFAR10 test set. We use the same models studied in Fig. 4, and compute the CE Jensen gap, as well as a standard notion of diversity based on the average pairwise correlation between the predictions of ensemble members. For both homogeneous and heterogeneous ensembles as studied in Fig. 4, we find that there is a strong positive correlation between these two notions of diversity (linear fit given by black line).