We thank reviewers for their constructive feedback. Reviewers find our work 'well written' (R4, R5), 'well backed by theory' (R5), 'convincing on several experiments' (R5), and 'taking an interesting approach' (R3). We address hereafter their main concerns.

AH and Generalization (R4, R5). In Baratin et al., 2021 and Oymak et al., 2019, it was independently observed that the following phenomena hold when DNN generalizes:

1. Ψ [Both the layerwise kernel and the full kernel] will have small number of large singular values while most other singular values are much smaller.
2. The label vector Y will be aligned with large singular directions in Ψ.

The two conditions above are also the conditions for maximizing CKA. Baratin et al., 2021 introduced the CKA as a measure of model compression and feature selection. The first condition can be viewed as the measure of model compression, as only effectively few directions in parameter space are relevant for changes of the function. The second condition can be interpreted as feature selection because we want the anisotropy of the tangent space to be skewed toward directions that leads to correct labels in function space. Therefore, increasing kernel alignment will imply low dimensional internal representations (suggesting a simpler model) and well-chosen features (suggesting a better fit to the data). *However*, we would like to emphasize that there are still many open questions about the relation between AH and generalization. Addressing all of them is out of the scope of this paper. Our aim in this paper is to explain the AH from a signal propagation point of view.

Using middle layers to learn? (R5). We tried to use only the middle layers for classification (we prune other layers), but the performance was significantly worse. We believe that other layers also benefit generalization, although the exact mechanism is still misunderstood. We have also tried to train the network by directly maximizing the AH of the layers instead of minimizing the loss (appendix F). We have not observed any performance gain.

AH in ResNet (R5). We believe there is a misunderstanding here. For ResNets, we observe in Fig4 that peak alignments are achieved in the last layers. However, if the scaling law $3/5$ holds for ResNet, we should expect that the layer index of peak alignment shifts towards the middle since $\Theta(L^{3/5})$ is located near the middle for large L. This is unrelated to generalization. The scaling law for ResNet is likely to be different from $3/5$ since the skip connections modify how information propagates through the network. We believe this is a interesting topic for future work.

Eq 5 (R5). Eq 5 is obtained exactly like Eq 4. The only difference is that with parameter freezing (all but l^{th} layer), the NTK is given by $\hat{K}_l$ while when there is no parameter freezing, the NTK is given by $\hat{K}$.

Intialization and AH (R4). Appendix B.4 discusses how to close the gap between information at intialization and AH at the end of training. A key aspect is that increase in layer alignments mainly takes place in early training (approximately 2 epochs of SGD, Fig2 and Fig7), so phenomena at intialization significantly influence alignment patterns at the end of training. Moreover, during early training, we conjecture that layers with information balance at intialization receive more 'informative' gradient updates – incorporating both forward and backward propagation. In Thm A and Approx1 in Appendix B.4, we show that 'informative' updates in a layer's parameter is closely related to large increase in feature matrix $\Psi_l \tilde{Y}$, leading to higher alignment values. Combining the two, we obtain the EH, which tends to predict end-of-training alignment values at intialization.
Gradient independence assumption (GIA): In the infinite width limit, the GIA is used to calculate gradient covariance (Appendix A.3). Yang, 2020 showed that the GIA yields exact gradient covariance (Appendix A.2). So our results are exact in the infinite-width limit. However, our experiments with finite-width networks show an excellent match with the infinite-width predictions.

Width/Depth Limits (R3). The results of Thms 1 and 2 are stated for the infinite-width information $I_{\cdot,\infty}$. This entails that the width is already considered infinite in these statements. We then study the effect of depth. This can be seen as in infinite-width-then-large-depth regime.

Main results and novelty (R3). The $3/5$ scaling law is a direct result of Thms 1 and 2. The intuition is that in order to balance l^{-2} (forward prop) and $(L/l)^{-3}$ (backward prop), we need l to scale as $L^{-3/5}$. More details are provided in Appendix B. We would like to emphasize that only Thm 1 is a direct corollary of previous results from Hayou et al., 2020. Thm 2 requires more refined analysis which can be found in the appendix (we use lemma 2 in the appendix to obtain the backward information loss). We have also included a detailed introduction of the theory of signal propagation. Regarding novelty, we respectfully disagree. To the best of our knowledge, our paper is the first attempt to theoretically understand the AH. Our theoretical predictions (the $3/5$ scaling law) is verified empirically in Fig3.

Scaling law in Fig3 and Fig4 (R3). We believe there is a misunderstanding here. As discussed below Corollary 1, the scaling law predicts only that the layer index with the largest alignment should lie between $C_1 L^{3/5}$ and $C_2 L^{3/5}$, for some constants C_1, C_2 – Fig 3 empirically validates this claim. Fig 15 shows how optimiser hyperparameters can affect the constants without changing the $3/5$ law. This explains why the peaks in Fig 4 occur at $CL^{3/5}$ for some C – so it is not possible to observe the $3/5$ law in Fig 4.

References (R3, R4, R5). We thank the reviewers for mentioning some interesting works that we have not been aware of. We will add them in the discussion section.