Data Free Metrics Are Not Reparameterisation Invariant Under the Critical and Robust Layer Phenomena

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

Data-free methods for analysing and understanding the layers of neural networks have offered many metrics for quantifying notions of "strong" versus "weak" layers, with the promise of increased interpretability. We examine how robust data-free metrics are under random control conditions of critical and robust layers. Contrary to the literature, we find counter-examples that provide counter-evidence to the efficacy of data-free methods. We show that data-free metrics are not reparameterisation invariant in these conditions and lose predictive capacity across correlation measures, RMSE, Person Coefficient and Kendall's Tau measure. Thus, we argue that to understand neural networks fundamentally, we must rigorously analyse the interactions between data, weights, and resulting functions that contribute to their outputs – contrary to traditional Random Matrix Theory perspectives.

1. Introduction

Understanding and interpreting deep learning models is a critical area of research, especially as the prevalence of these models increases in real-world applications. The holy grail of neural network interpretability lies in identifying computationally cheap metrics that can provide insights into the effectiveness of neural networks and their components. Data-free methods typify this endeavour by analysing the properties of the neural network parameters without regard for the data. A key example of data-free methods is [12], which claims to be able to predict the performance of a neural network without the requirement of test data through the use of Random Matrix Theory to analyse the layer weight matrices. In contrast, data-dependent layer analysis via mechanistic interpretability or functional analysis attempts to quantify how inputs interact at specific layers and use comparative analysis to understand the interaction between model parameters and data [9, 14, 15].

Zhang et al. [23] identified an interesting and unexpected phenomenon in neural network layers: some layers within a network are robust, while others are critical. A critical layer is a layer that cannot be *re-initialisatised* or *re-randomised* without dramatically affecting the performance of the network. In contrast, a robust layer can be either *re-initialisatised* or *re-randomised* without any noticeable effect on performance. *Re-initialisation* sets the network back to its initial weights before training, and *Re-randomisation* sets the weights to random values by re-sampling from the same distribution used for initialisation. It was observed that in some cases, *re-initialisation* and *re-randomisation* can result in significant performance differences for a given layer, with *re-initialisation* maintaining performance but *re-randomisation* significantly degrading it [23]. In other cases, *re-initialisation* and *re-randomisation* of a layer lead to a negligible difference in performance.

Given that similar studies in loss landscape geometry analysis have explored the efficacy of metrics under the notion of reparameterisation invariance [5] we believe the robust and critical layer phenomena provides a strong basis for asking:

- Are data-free metrics reparameterisation invariant under the robust and critical layer phenomenon recorded by Zhang et al [23]? That is, can they disambiguate between *re-initialisation* and *re-randomisation* of a given layer when performance of the model is impacted.

We find that data-free methods have no significant predictive capacity over the robust and critical layer phenomenon. As a result, we argue that the surveyed data-free methods are not reparameterisation invariant and are thus, by extension, at risk of erroneous predictions.

2. Background

This section briefly presents data-free methods and discusses how they are used in this paper.

Data-free methods of interpretability aim to understand the inner workings of neural networks by studying the properties of the network parameters. Data-free approaches often focus on the matrix norm properties of layer weight matrices to understand learning or improve the performance of neural networks [1, 7, 13, 17, 18, 22]. However, Zhang et al., [23], showed in their work that matrix norms, such as the Frobenius norm, are too coarse to understand the generalisation properties of neural networks. Martin and Mahoney [12] use Random Matrix Theory to analyse the weights matrices (excluding biases) of neural network layers through training to create a theory of heavy-tailed self-regularisation. With this theory, they construct a set of predominately power-norm metrics related to generalisation that is applied after training to assess layer performance: alpha (α), alpha-weighted ($\hat{\alpha}$), log alpha norm, and MP soft rank [12]. In this work, they identified a value of α between 2 and 6 as a property of a good, well-trained layer, whereas $\alpha > 6$ indicates that a layer is underfitted and $\alpha < 2$ indicates that it is overfitted. Martin et al., [13] showed a correlation between these metrics and the generalisation performance of pre-trained models in language and computer vision tasks.

Given the work by Zhang et al., [23] that shows norm methods are ineffective, our work explores Alpha [12], Alpha Weighted [12], Log Alpha Norm [12], MP Soft Rank [12], Generalized von-Neumann Matrix Entropy [12] Frobenius Norm, Spectral Norm and Stable Rank [16] within the critical and robust layer phenomena to see if these methods can disambiguate between the performance difference of *re-initialisation* and *re-randomisation* of a layer.

3. Experimental Setup

Zhang et al., [23], showed the robust and critical layer phenomenon across a range of trained architectures, MLPs, VGGs [19], ResNets [8], Transformers [21], Vision Transformers [6], MLPMixers [20] across datasets MNIST [11], CIFAR10 [10], ImageNet [4] and LM1B[3]. Therefore, to explore whether data-free interpretability methods can explain the performance difference between *re-initialisation* and *re-randomisation* of layers, we use the simplest model (ReLU FCN 5x512), Figure 1, and dataset (MNIST [11]) identified by Zhang et al., [23] that demonstrates this phenomenon while having the most significant performance difference between *re-initialisation* and *re-randomisation* of layers.

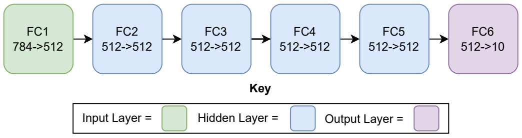

Figure 1: ReLU FCN 5x512 Model Architecture.

The ReLU FCN 5x512 model allows for practical analysis of data-free methods, offering a clear performance contrast between *re-initialisation* and *re-randomisation*. Zhang et al.,[23] showed that residual blocks are robust to *re-randomisation* and attributed this to the residual layer potentially playing a lesser role in the network and thus having smaller activations than the skip connection. To analyse how effective the data-free metrics are at disambiguating between *re-initialisation* and *re-randomisation*, we analyse the correlation between data-free metrics and test accuracy using the Spearman correlation coefficient ρ, the root mean square error (RMSE) of the linear regression and Kendall's tau measure (K-τ). Where ρ and K-τ score of -1 indicates a very strong negative correlation, 0 indicates no correlation, and 1 indicates a very strong positive correlation. We use RMSE and Kendall's Tau measure for this study as they are two of the correlation metrics used in [13] to highlight the predictive capacity of the data-free metrics, Log-Frobenius Norm, Log Spectral Norm, Weighted Alpha and Log Alpha Norm.

We trained 100 ReLU FCN 5x512 models, creating 100 initialisations and 100 trained models, to obtain a representative sample of possible initialisations and trained models. The model weights and biases are initialised and re-randomised from the same distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$ where k is $\frac{1}{\text{in\_features}}$, e.g. FC1 has $k = \frac{1}{784}$. We record the data-free metric properties of these trained models' layers when they undergo *re-initialisation* and *re-randomisation* to understand if data-free methods can disambiguate between the critical and robust layer phenomenon. This exploration also demonstrates the overall predictive capacity of the data-free metrics at the end of training.

Data-Free Metrics. Power Norm based data-free methods analyse a layer weight matrix, W, excluding the bias. A variety of data-free metrics have been developed in the literature to quantify the importance of a layer, we focus on the following metric [12]:

- **Alpha (α):** The fitted power law exponent, α, for the empirical spectral density of the correlation matrix $X = W^T W$, such that $p_{emp}(\lambda) \sim \lambda^{-\alpha}$, where λ are the eigenvalues of X.

The metrics are collected using the weightwatcher[1] tool. **In Appendix Section A we present the formalisation and analysis of the other data-free methods**.

4. Results and Discussion

For clarity and succinctness, we only present our results for the data-free metric α of [12] in the body of the paper. **In Appendix Section A we present the analysis of a range of other data-free methods: Alpha Weighted, Log Alpha Norm. MP Soft Rank, Frobenius Norm, Spectral Norm and Generalized von-Neumann Matrix Entropy.**

[1] https://weightwatcher.ai

4.1. Analysis of the Alpha Metric Under Re-Initialisation and Re-Randomisation

After training, the ReLU FCN 5x512 model achieves α values within the desired range of 2 and 6, as seen in Table 1, indicating, according to [12], a well-trained model, which correlates with the recorded test accuracy. The results in Table 1 could lead to the assumption that all of these layers should be critical and not robust to *re-initialisation* or *re-randomisation*, as layers that are within the desired α range represent well trained layers.

Table 1: Alpha (α) of the layers in ReLU FCN 5x512 and test accuracy of the model. The mean and ± 1 SEM [2] (standard error from the mean) are derived from 100 trained models on MNIST.

Metric	**Layer**						**Test Accuracy**
	FC1	**FC2**	**FC3**	**FC4**	**FC5**	**FC6**	
Alpha (α)	4.82 ± 0.025	4.205 ± 0.039	4.126 ± 0.038	4.135 ± 0.035	4.193 ± 0.034	3.793 ± 0.805	96.822 ± 0.057

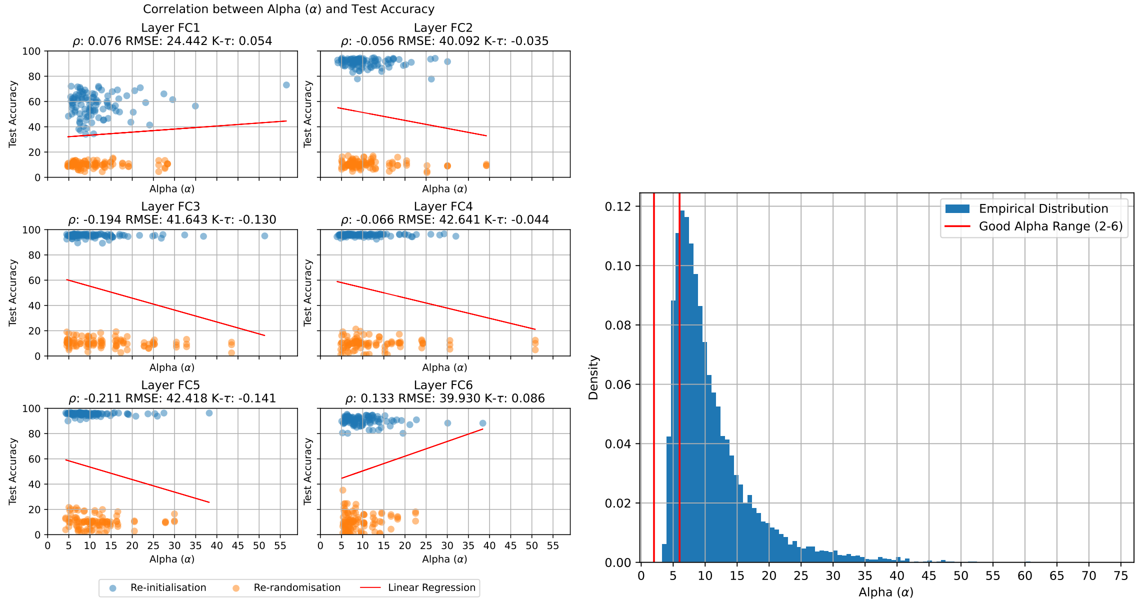

Figure 2: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy vs Alpha (α). ρ is the Spearman correlation coefficient, $RMSE$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and α values (**Left**). Empirical distribution of Alpha (α) values on a 512x512 fully connected layer, sampled from 10,000 initialisations (**Right**).

Robust and Critical Layers Impact on Alpha: To understand the interaction between α values and the criticality of a layer, we compare the independent *re-initialisation* (**blue**) and *re-randomisation* (**orange**) of each layer in **across 100 networks** and record the resulting impact on the networks test accuracy and alpha value. Figure 2 (left) shows the stark contrast in how layers respond under *re-initialisation* or *re-randomisation*. For example, we find that only when appling *re-initialisation* to the Layer FC1 do we see a large drop in test accuracy, *re-initialising* other layers leave performance almost unchanged. However, we observe different results when *re-randomisation* is applied. We find that *re-randomising* any layer degrades accuracy to circa random accuracy on the test set. As a result, we can state that the layers in this ReLU FCN 5x512 are robust to *re-initialisation* but not robust to *re-randomisation*. Surprisingly, this is not reflected in the corresponding α values of these two conditions, as the main cluster of α values is relatively similar for each layer and each condition. This suggests that α is not reparameterisation invariant for *re-initialisation* and *re-randomisation* as for both cases, α values are not only similar but are within the optimal α value range – despite a large drop in accuracy in one case (*re-initialisation*) but not in the other (*re-randomisation*).

Lack of Predictive Capacity of Alpha Under Reparametisation: Our findings underscore a crucial limitation of the data-free metric Alpha. It is unable to discern and explain the performance difference between *re-initialisation* and *re-randomisation* across layers, as shown in Figure 2 (left). We would expect to observe a strong negative correlation from *re-initialisation* to *re-randomisation* if alpha has predictive capacity. However, there is almost no difference between the α values of *re-initialisation* and *re-randomisation*, with a mean Spearman correlation coefficient and Kendall's tau measure across layers of -0.053 and -0.035, respectively. Our findings on the metric Alpha extend to our exploration of other data free metrics in Appendix Section A.

Empirical Distribution of α at initialization: The results from Figure 2 (left) raise some irregularities when considering the predictive capability of the α value and its 'good' value range. To further investigate the irregularities of the α metric, we observe the empirical distribution of a 512x512 fully connected layer, sampled from 10,000 potential initialisations in Figure 2 (right). The resulting distribution highlights that an initialised, untrained layer of this network, can fall, by chance, within the optimal α value range of 2 and 6. That layers FC2-FC5 could all randomly *start* with an α value in the 'well-trained' range $[2, 6]$ is further evidence that employing power norms is too coarse a metric and representation of the model to give predictive and informative insights into model generalisation capabilities. We argue that if we observe a failure case of this metric when considering initialised layers, it is unlikely to have predictive capacity outside of this regime.

5. Conclusion

Data-free metrics cannot explain the robust and critical layer phenomena. Based on experminents covering a wide range of these metrics, we argue that they have little to no predictive capacity over the difference between *re-initialisation* and *re-randomisation* robustness of a layer. With this, we highlight how data-free metrics can be described as non-reparameterisation invariant as they are not robust under the reparameterisation provided by the critical and robust layer phenomenon. Furthermore, our results advocate for an in-depth exploration of the dynamics of data-free methods and the search for methods that can suitably disambiguate between the robust and critical layer phenomena.

References

[1] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

[2] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389, 2005.

[3] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005, 2013. URL http://arxiv.org/abs/1312.3005.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[7] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

[8] 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 (CVPR)*, pages 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

[9] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures, 2023. *URL https://arxiv. org/abs/2305.06329*.

[10] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[11] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[12] Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. URL http://jmlr.org/papers/v22/20-410.html.

[13] Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):4122, 2021.

[14] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

[15] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[16] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21–es, 2007.

[17] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[18] Amartya Sanyal, Philip H. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1enKkrFDB.

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

[20] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[22] David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew R. Walter. Approaching deep learning through the spectral dynamics of weights, 2024. URL https://arxiv.org/abs/2408.11804.

[23] Chiyuan Zhang, Samy Bengio, and Yoram Singer. Are all layers created equal? *J. Mach. Learn. Res.*, 23(1), Jan 2022. ISSN 1532-4435. URL https://jmlr.org/papers/volume23/20-069/20-069.pdf.

Appendix A. Further Analysis on Data-Free Metrics

In this section we extend our analysis to the following data free metrics in our existing experimental setup. W represents the weight matrix of the layer and X is for the empirical spectral density of the correlation matrix, $X = W^T W$, such that $pemp(\lambda) \sim \lambda\alpha$, where λ are the eigenvalues of X

- **Alpha Weighted ($\hat{\alpha}$):** $\alpha \log(\lambda_{max})$, where λ_{max} is the max eigenvalue from X [12].
- **Log Alpha Norm:** $\log(||X||_\alpha^\alpha)$, where $||X||_\alpha^\alpha = \sum_i^M \lambda_i^\alpha$, where M is the rank of W [12].

- **MP Soft Rank:** is the ratio between the bulk edge of the $p_{emp}(\lambda)$, λ^+, and the max eigenvalue, λ_{max}, $\frac{\lambda^+}{\lambda_{max}}$ [12].
- **Frobenius Norm**: The sum of the singular values of W denoted as $||W||_F$.
- **Spectral Norm**: The max singular value of W denoted as $||W||_\infty$.
- **Stable Rank:** The ratio of the squared Frobinues Norm and the squared Spectral Norm, denoted as $\frac{||W||_F^2}{||W||_2^2}$ [16].
- **Generalized von-Neumann Matrix Entropy:** $\frac{-1}{log(M)} \sum_i p_i \log p_i$, where M is the rank of matrix W and p_i is $\frac{\sigma_i^2}{\sum_i(\sigma_i^2)}$ where σ is the singular values of W [12].

Each metric can be found below with the appropriate subsection that corresponds to our analysis of these data-free metrics.

The correlation between data-free metrics Alpha Weighted, Log Alpha Norm, MP Soft Rank, Frobenius Norm, Spectral Norm, Stable Rank and Entropy and the associated test accuracy when layers undergo *re-initialisation* (blue) or *re-randomisation* are shown in Figures 3, 4, 5, 6, 7, 8 and 9 respectively. Across all metrics, we observe approximately zero correlation between the metric values and the test accuracy, highlighting the same findings as in the body of the paper.

A.1. Alpha Weighted ($\hat{\alpha}$)

For Alpha Weighted ($\hat{\alpha}$) we observe for all layers FC1-FC6 that there is very weak to no correlation between the layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy and the Alpha Weighted $\hat{\alpha}$ metric, see Figure 3. The highest Spearman correlation coefficient recorded is -0.198 for Layer FC5. However, the lowest recorded is for Layer FC4 at -0.044. Given that these layers have the exact dimensions, 512×512, this would suggest no correlation between the Alpha Weighted metric and the test accuracy. It can also be seen in Figure 3 that Layer FC1 observes a very weak positive Spearman correlation coefficient, highlighting that the Alpha Weighted has no predictive capacity and is unable to discern the performance difference of the textitre-initialisation and *re-randomisation* of a layer.

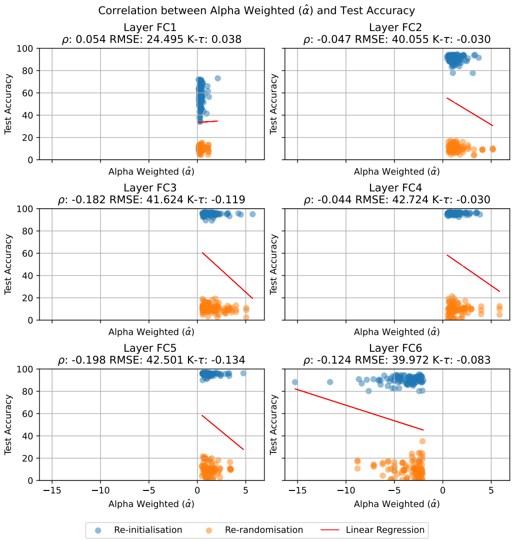

Figure 3: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy vs Alpha Weighted $\hat{\alpha}$. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.

A.2. Log Alpha Norm

For Log Alpha Norm, we observe for all layers FC1-FC6 that there is a very weak negative correlation between the layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy and the Log Alpha Norm metric, see Figure 4. The highest Spearman correlation coefficient recorded is -0.204 for Layer FC5, suggesting a weak negative correlation. However, the lowest recorded is for Layer FC2 at -0.055. Given that these layers have the exact dimensions, 512×512, this would suggest that there is no correlation between the Log Alpha Norm metric and the test accuracy, given there is no agreement on the amount of correlation. Given the very weak Spearman correlation coefficient across layers, we observe that the Log Alpha Norm has no predictive capacity and cannot discern the performance difference of the *re-initialisation* and *re-randomisation* of a layer.

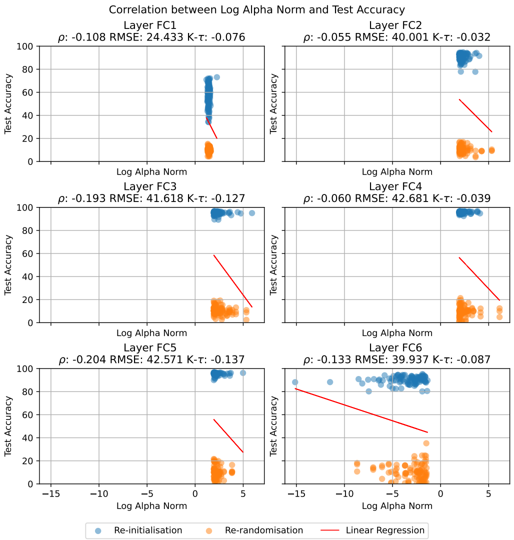

Figure 4: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange** test accuracy vs Log Alpha Norm. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.

A.3. MP Soft Rank

For MP Soft Rank, we observe a mixture of very weak positive and negative Spearman correlation coefficients of MP Soft Rank and Test Accuracy across layers; see Figure 5. Layers FC1, FC2, and FC3 report a very weak positive correlation, and Layers FC4, FC5 and FC6 report a very weak negative correlation. This result provides ample evidence to suggest that MP Soft Rank cannot disambiguate between *re-initialisation* and *re-randomisation* of a layer and thus has no predictive capacity.

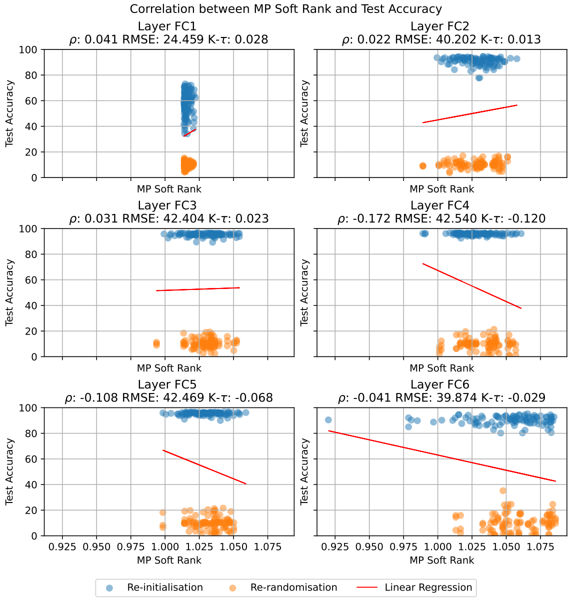

Figure 5: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy vs MP Soft Rank. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.

A.4. Frobenius Norm

Norm-based metrics were originally shown to be too coarse a metric to measure the generalisability of the neural networks in [23]. Figure 6 strengthens these findings, highlighting that there is essentially no correlation between the Frobenius Norm of a layer and the test accuracy of a model.

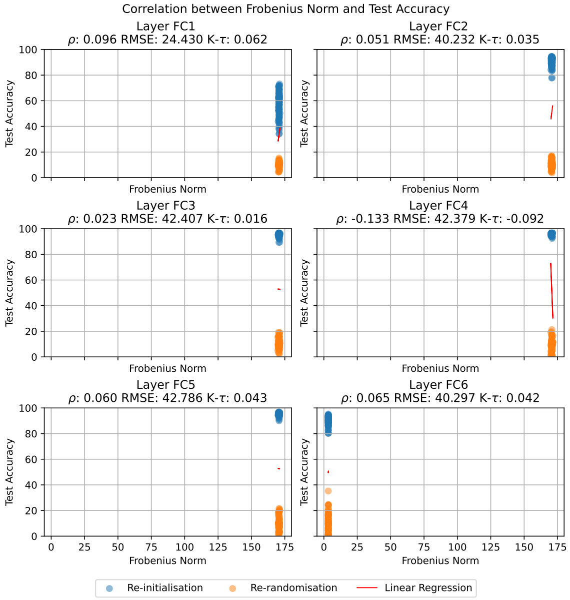

Figure 6: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy vs Frobenius Norm. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.

A.5. Spectral Norm

Norm-based metrics were originally shown to be too coarse a metric to measure the generalisability of the neural networks in [23]. Figure 7 strengthens these findings, highlighting that there is essentially no correlation between a layer's Spectral Norm and a model's test accuracy.

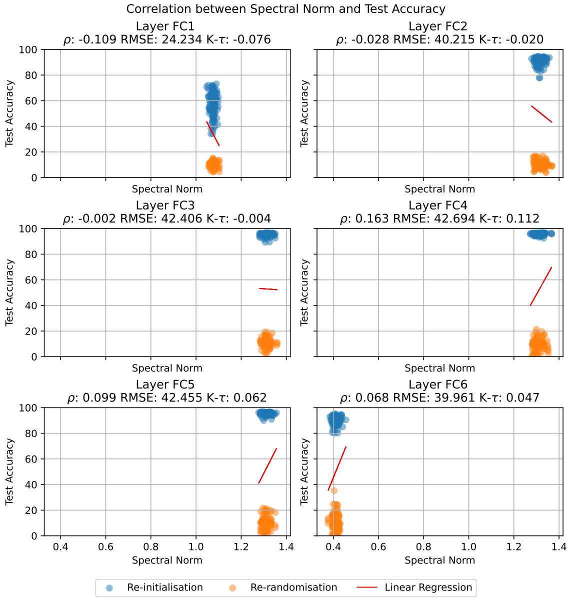

Figure 7: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy vs Spectral Norm. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.

A.6. Stable Rank

Norm-based metrics were originally shown to be too coarse a metric to measure the generalisability of the neural networks in [23]. Figure 8 strengthens these findings, highlighting no substantial correlation between a layer's Stable Rank and a model's test accuracy.

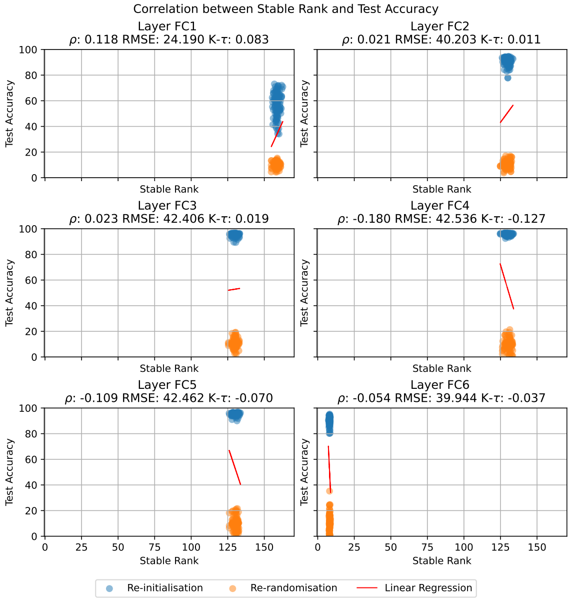

Figure 8: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy vs Stable Rank. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.

A.7. Generalized von-Neumann Matrix Entropy (Entropy)

For Generalized von-Neumann Matrix Entropy (Entropy) we observe for layers FC3 and FC4 that there is a very weak negative correlation between the layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy and the Entropy metric, see Figure 9. On the other hand we observe a very weak positive correlation for Layers FC1, FC2, FC5 and FC6. This finding suggests that the Entropy metric has no predictive capacity and cannot disambiguate the observed performance difference between *re-initialisation* and *re-randomisation*.

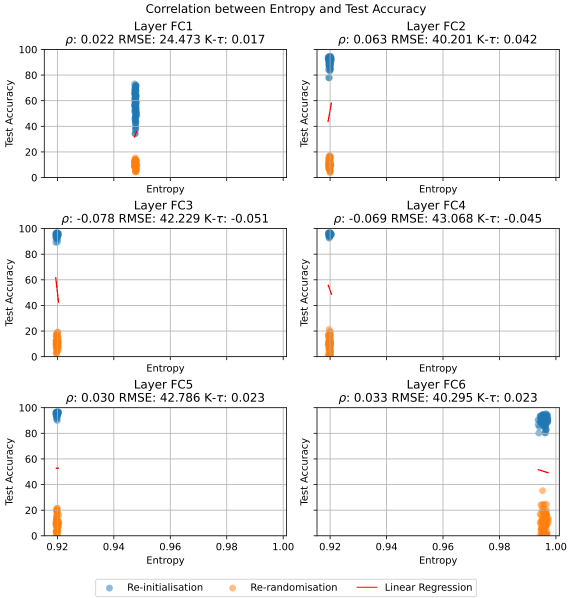

Figure 9: Layer *re-initialisation* (**blue**) and *re-randomisation* (**orange**) test accuracy vs Entropy. ρ is the Spearman correlation coefficient, $RSME$ is the root mean square error of the linear regression (red line), and K-τ is the Kendall's tau measure, all with respect to the relationship between test accuracy and alpha values.