Sanity Checks Revisited: An Exploration to Repair the Model Parameter Randomisation Test

Anna Hedström[1,3*], Leander Weber[2,*], Sebastian Lapuschkin[2,†], and Marina Höhne[3,4,5†]

[1]Department of Electrical Engineering and Computer Science, TU Berlin
[2]Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
[3]UMI Lab, Leibniz Institute of Agricultural Engineering and Bioeconomy e.V. (ATB)
[4]BIFOLD – Berlin Institute for the Foundations of Learning and Data
[5]Department of Computer Science, University of Potsdam

Abstract

The Model Parameter Randomisation Test (MPRT) is widely acknowledged in the eXplainable Artificial Intelligence (XAI) community for its well-motivated evaluative principle: that the explanation function should be sensitive to changes in the parameters of the model function. However, recent works have identified several methodological caveats for the empirical interpretation of MPRT. To address these caveats, we introduce two adaptations to the original MPRT—*Smooth MPRT* and *Efficient MPRT*, where the former minimises the impact that noise has on the evaluation results through sampling and the latter circumvents the need for biased similarity measurements by re-interpreting the test through the explanation's rise in complexity, after full parameter randomisation. Our experimental results demonstrate that these proposed variants lead to improved metric reliability, thus enabling a more trustworthy application of XAI methods.

1 Introduction

The problem of evaluating the quality of an explanation method in eXplainable Artificial Intelligence (XAI) remains unsolved due to the frequent absence of ground truth explanation labels [1, 2, 3]. To address this issue, numerous quantitative evaluation methods have been proposed [4, 5, 6, 2, 7, 5, 8, 9, 10, 11, 12, 13, 6, 14, 2, 15, 16]. Among these, the *Model Parameter Randomisation Test* (MPRT) [15] measures the degree to which an explanation deteriorates as the model parameters are progressively randomised, i.e., layer-by-layer, starting from the output. This test posits that a greater difference in the explanation function's output, in response to parameter randomisation, signifies a higher quality of the explanation method.

In recent years, however, multiple independent research groups have provided constructive observations on the methodological choices in the original MPRT [17, 18, 19, 20]. These encompass the selection of pairwise similarity for measuring explanation difference [20], the order of layer randomisation [20], the choice of explanation preprocessing [17, 20] and the dependency on the model task [19, 18]. Given the considerable adoption of MPRT in the XAI community and its potential to influence the acceptance or rejection of different explanation methods, these raised concerns deserve a closer examination. Do evaluations based on the original MPRT optimally inform us about the quality of an explanation method, or can we improve its reliability?

*Equal Contribution
†Corresponding: sebastian.lapuschkin@hhi.fraunhofer.de,MHoehne@atb-potsdam.de

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

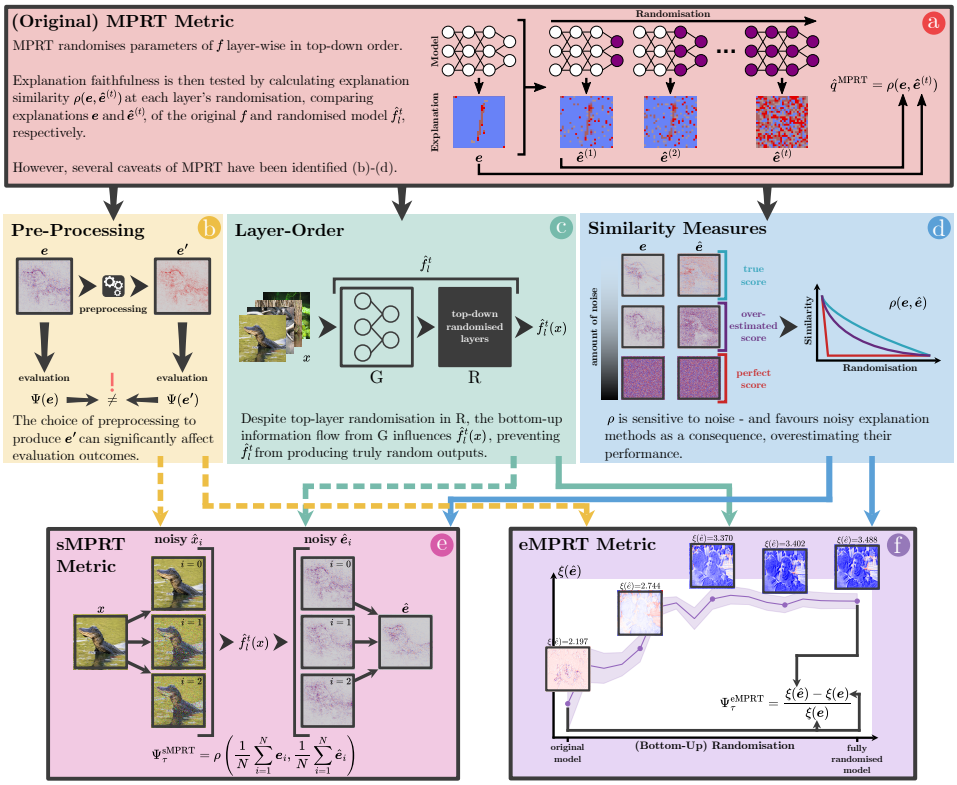

Figure 1: Schematic visualisation of the original MPRT [15] (*top*), identified shortcomings (*middle*) and proposed solutions (*bottom*). Solid arrows signify shortcomings directly addressed by our proposed metrics, while dashed arrows denote those addressed through ideas from previous work [17, 20]. (a) The MPRT evaluates an explanation method by randomising f's parameters in a top-down, layer-wise fashion and thereafter calculating explanation similarity $\rho(\boldsymbol{e}, \hat{\boldsymbol{e}})$ at each layer through comparing explanations $\boldsymbol{e}$ of the original model f and $\hat{\boldsymbol{e}}$ of the randomised model $\hat{f}$. (b) *Pre-processing*: normalisation and taking absolute attribution values significantly impact MPRT results, potentially deleting pertinent information about feature importance carried in the sign. (c) *Layer-order*: top-down randomisation of layers in MPRT does not yield a fully random output, preserving properties of the unrandomised lower layers and thus affecting the evaluation of faithful explanations. (d) *Similarity measures*: the pairwise similarity measures used in the original MPRT [15] are noise-sensitive, e.g., from gradient shattering and thus likely to impact evaluation rankings of XAI methods. (e) sMPRT extends MPRT by incorporating a "denoising" preprocessing step that averages explanations over N perturbed inputs, aiming at reducing noise in local explanation methods. (f) eMPRT reinterprets MPRT by evaluating the faithfulness of the explanation method by comparing its rise in complexity, before and after full model randomisation.

The foundational idea behind MPRT—that the explanation function should be sensitive to the model's parameters—is both insightful and important. Our work seeks to build upon this initial work, addressing its shortcomings, namely, the order of layer randomisation and the selection of pairwise similarity measures, discussed in detail in Section 2.1.1 and illustrated in Figure 1. We introduce two MPRT variants, *Smooth MPRT* (sMPRT) and *Efficient MPRT* (eMPRT), where the sMPRT mitigates sensitivity to shattering noise (cf. [20]) by denoising, i.e., averaging attributions over N perturbed inputs and the eMPRT evaluates explanation faithfulness by quantifying the rise in complexity after full parameter randomisation, thus obviating a biased similarity measurement. We release these methods under the existing Quantus [3] evaluation framework[1].

Our methods, despite being an iteration, carry substantial importance. To deploy XAI in real-world applications, such as safety-critical domains, we need to focus on developing reliable quantitative evaluation metrics for existing explanation methods. This work contributes to ensuring that the continuous development of XAI methods remains both theoretically sound and empirically valid.

[1]Code at: `https://github.com/understandable-machine-intelligence-lab/Quantus`.

2 Preliminaries

For clarity in the following discussion, we first present the core notation.

Local Explanations. Let $f : \mathbb{X} \mapsto \mathbb{Y}$, $f \in \mathbb{F}$ be a differentiable black-box model that maps inputs $\boldsymbol{x} \in \mathbb{R}^D$ to predictions $\boldsymbol{y} \in \mathbb{R}^C$ with C classes, where $\mathbb{F}$ denotes the function space. We assume that f consists of L layers $f^1, \cdots, f^L$, such that $f(\boldsymbol{x}) = f^L \circ \cdots \circ f^1(\boldsymbol{x})$. To interpret the behavior of f, *local* explanation methods [21, 22, 23, 24, 25] attribute importance scores to features of input $\boldsymbol{x}$. Given a model's prediction y, we define an explanation $\boldsymbol{e} \in \mathbb{R}^D$ via the explanation function $\Phi_\lambda : \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}^D$, as $\boldsymbol{e} = \Phi(\boldsymbol{x}, f, y; \lambda)$. Then $\mathbb{E}$ denotes the space of possible explanations w.r.t. f with $\Phi_\lambda \in \mathbb{E}$.

Evaluating Explanations. Due to the lack of ground truth explanations for black-box models, evaluating the quality of Φ is generally non-trivial [1, 26, 2, 27]. This issue can be circumvented by instead measuring specific properties of explanations, such as stability (e.g. [4, 5, 6]), complexity (e.g., [9, 7, 28]) or faithfulness (e.g., [15, 7, 11]). For a comprehensive overview of existing evaluation properties, we refer to [3] and [27]. We define a quality estimate $q \in \mathbb{R}^M$ as the output of an evaluation function $\Psi_\tau : \mathbb{E} \times \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}$, parameterised by τ such that $q = \Psi_\tau(\Phi, \boldsymbol{x}, f, y)$. Then $\mathbb{O}$ denotes the space of possible evaluations w.r.t. $\boldsymbol{e}$ and $\Psi \in \mathbb{O}$.

2.1 Model Parameter Randomisation Test (MPRT)

We define the original MPRT [15] in the following.

Definition 1 (MPRT)**.** *Let $\Psi_\tau^{MPRT} : \mathbb{E} \times \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}$ be an evaluation function that computes a quality estimate $\hat{q} \in \mathbb{R}$ that measures the similarity between the original explanation $\boldsymbol{e}$ and the explanation $\hat{\boldsymbol{e}} := \Phi(\boldsymbol{x}, \hat{f}_l^t, y)$ corresponding to the perturbed model $\hat{f}_l^t$ randomised in a top-down fashion up to layer $l \in [L, L-1, \ldots, 1]$:*

$$\hat{q}^{\text{MPRT}} = \rho(\boldsymbol{e}, \hat{\boldsymbol{e}}), \tag{1}$$

with $\rho : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$ as a similarity function.

For Φ to be sensitive to the parameter randomisation, Equation 1 should yield values such that $\hat{q}^{\text{MPRT}} \ll 1$. This means low similarity between $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$ for a given class y. As denoted in the superscript of $\hat{f}_l$, the direction of layer randomisation is completed top-down, i.e., $\hat{f}_l^t$ with $l \in [L, L-1, \ldots, 1]$. For bottom-up randomisation, we denote $\hat{f}_l^b$ with $l \in [1, 2, \ldots, L]$.

2.1.1 Methodological Caveats

In the following, we summarise the concerns that have been raised about various methodological aspects of MPRT, specifically its choice of (a) pre-processing, (b) layer-order and (c) similarity measure. For each issue, we put forward a solution.

(a) Pre-processing. In the original formulation of MPRT [15], the explanations $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$ are normalised using their minimum and maximum values. This is problematic as these statistics are highly variable and almost arbitrary across attributions, complicating meaningful comparison across randomisations [20]. Taking absolute attribution values may erase meaningful information [17]—a caveat already addressed by the seminal work [15]. *(Proposed Solution) To maintain the original scale and distribution of the explanation, we recommend normalising $\boldsymbol{e}$ by the square root of the average second-moment estimate [20] (see Equation 5 in Supplement) which does not constrain attributions to a fixed range and enhances comparability across methods and randomisations [20].*

(b) Layer-order. While one might intuitively expect a significant difference in the explanation output upon top-layer randomisation, i.e., $\rho(\boldsymbol{e}, \hat{\boldsymbol{e}}) \ll 1$ with $\hat{f}_l^t$ up to layer $l \in [L, L-1, \ldots, 1]$; [20] provides evidence to the contrary. As discussed in-depth in the original publication [20], top-down randomisation induces only modest alterations in the forward pass where (i) irrelevant features from lower, non-randomised layers persist in higher, randomised layers, (ii) high activations in lower layers are relatively likely to continue to dominate the network's response and (iii) architectures with skip connections, e.g., ResNets [29] maintain a baseline explanation that stays constant even after randomisation, preserving certain features. As such, only a limited degree

of change between $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$ can be expected from faithful explanation methods post-model randomisation. *(Proposed Solution) To avoid preserving information in the forward pass, it is advisable to perform bottom-up randomisation (as further discussed in the Supplement 6.2) or perform the comparison of explanations only after full randomisation (see Equation 3.*

(c) Similarity Measures. Another shortcoming identified by [20] is the property of *SSIM* and other pairwise similarity measures, e.g., Spearman Rank Correlation, to be minimised by statistically uncorrelated random processes. This is problematic since it implies that certain explanation methods with more intrinsic (shattering) noise, e.g., gradient-based methods, will be favoured in ranking comparisons, ultimately biasing evaluation outcomes. *(Proposed Solution) While no solution has been proposed to date, we put forward the sMPRT and eMPRT modifications which separately addresses the issues by denoising attributions as a pre-processing step and replacing the pairwise similarity measure with a complexity measure, respectively.*

3 Methods

In the following, we provide mathematical details and motivations for sMPRT and eMPRT.

3.1 Smooth Model Parameter Randomisation Test (sMPRT)

We first propose the sMPRT, which makes a small adaptation to the original MPRT, aimed at removing (shattering) noise observed in local explanation methods [30]. Inspired by [21, 25] we introduce a preprocessing step to the evaluation procedure—given an input $\boldsymbol{x}$, generate N perturbed instances $\hat{\boldsymbol{x}}_i$, compute their attributions $\Phi(\hat{\boldsymbol{x}}_i, f, y; \lambda)$ and then perform the MPRT evaluation on the averaged denoised attribution estimates. We define the sMPRT in the following.

Definition 2 (sMPRT)**.** *Let $\Psi_\tau^{sMPRT} : \mathbb{E} \times \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}$ be an evaluation function that computes a quality estimate $\hat{q} \in \mathbb{R}$ that measures the similarity estimates between explanations $\boldsymbol{e}_i := \Phi(\hat{\boldsymbol{x}}_i, f, y; \lambda)$ and $\hat{\boldsymbol{e}}_i := \Phi(\hat{\boldsymbol{x}}_i, \hat{f}_l^b, y; \lambda)$ averaged over $i \in [1, N]$:*

$$\hat{q}^{\text{sMPRT}} = \rho\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{e}_i, \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{e}}_i\right), \tag{2}$$

where $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \eta_i$, $\eta_i \sim \mathcal{N}(0, \sigma)$ and $||\eta_i||_p \leq \epsilon$ holds with high probability, with $\sigma, \epsilon \in \mathbb{R}$.

Similar to Equation 1, to indicate low similarity between $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$, we expect values such that $\hat{q}^{\text{sMPRT}} \ll 1$ from Equation 2. As discussed in the Supplement, to set σ, we follow the heuristic provided in the original publication [21], i.e., $\sigma/(x_{\text{max}} - x_{\text{min}})$. For N, we find that $N = 50$ provides consistent results (see Supplement 6.3).

Problems with sMPRT. Despite sMPRT's capability to reduce sensitivity to shattering noise (see Figure 2), sMPRT has notable drawbacks. First, it demands additional sampling ($N \geq 50$, see Figure 7), significantly increasing computational cost compared to standard MPRT. Second, the practice of removing noise by adding noise is already a defining characteristic in some attribution methods, e.g., *SmoothGrad* [21] and *NoiseGrad* [25], limiting sMPRT's efficacy with these methods and blurring the distinction with their baseline methods. Third, the degree of noisiness is an arguable property of attribution methods and removing it before evaluation may yield non-representative or biased results which may create performance discrepancies upon real-world application. Lastly, sMPRT introduces hyperparameters σ and N, which may not be tunable on any given data domain, e.g., climate data [31]. To this end, we next investigate alternative modifications to MPRT.

3.2 Efficient Model Parameter Randomisation Test (eMPRT)

To address caveats (b) and (c) and improve upon the efficiency of the original MPRT, we introduce eMPRT. This test effectively removes the layer-by-layer pairwise comparison between $\boldsymbol{e}$ and $\hat{\boldsymbol{e}}$ and instead compute the *relative rise in explanation complexity* using only two model states, i.e., the original- and fully randomised model. We define eMPRT in the following.

Definition 3 (eMPRT)**.** *Let $\Psi_\tau^{eMPRT} : \mathbb{E} \times \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}$ be an evaluation function that computes a quality estimate $\hat{q} \in \mathbb{R}$ that measures the relative rise in the complexity of the*

explanation from a fully randomised $\hat{f}$, i.e., $\hat{\boldsymbol{e}} := \Phi(\boldsymbol{x}, \hat{f}, y; \lambda)$, such that:

$$\hat{q}^{\text{eMPRT}} = \frac{\xi(\hat{\boldsymbol{e}}) - \xi(\boldsymbol{e})}{\xi(\boldsymbol{e})} \tag{3}$$

with $\xi : \mathbb{R}^D \mapsto \mathbb{R}$ is a complexity function, e.g., defined through Equation 4.

Since Φ ought to faithfully capture the behaviour of the model f—a fully randomised model (resulting in higher entropy of $f_L(\boldsymbol{x})$, as shown in Figure 3 (e) should therefore lead to reduced information content in its explanation $\hat{\boldsymbol{e}}$. Equation 3 quantifies this heightened explanation complexity as a rise in relative entropy, where a positive value, i.e., $\hat{q}^{\text{eMPRT}} > 0$ indicates a rise in complexity and a negative value, i.e., $\hat{q}^{\text{eMPRT}} < 0$ suggests a decrease in complexity. If $\hat{q}^{\text{eMPRT}} = 0$, no change in explanation complexity was observed.

To set ξ, we employ a histogram entropy measure, grounded in Shannon-Entropy [32]. Specifically, attribution values $\boldsymbol{e}$ are binned across B distinct slots where for each ith bin we compute the frequency c_i and derive the corresponding normalised probability, p_i as follows:

$$\xi(\boldsymbol{e}) = -\sum_{i=1}^{B} p_i \log(p_i) \quad \text{where} \quad p_i = \frac{c_i}{\sum_{i=1}^{B} c_i}. \tag{4}$$

Opting for this complexity measure offers three advantages: preservation of the inherent sign of attributions, implicit normalisation and adaptability to diverse dimensionalities and distributions, wherein B can be contextually chosen.

Advantages with eMPRT. Generally, since explanations need only be computed twice, pre- and post-model randomisation, eMPRT is significantly more computationally efficient compared to MPRT. It further avoids issues related to layer-order as well as the choice of normalisation function (since the chosen entropy function already normalises implicitly, see Equation 4). Most importantly, by incorporating the initial unperturbed complexity estimate, i.e., $\xi(\boldsymbol{e})$ into Equation 3, eMPRT anchors its quality estimate by the explanation method's inherent complexity, which makes the evaluation scores more comparable across explanation techniques and models.

4 Results

The subsequent section provides a detailed analysis of our findings. Details on hyperparameters, as well as additional experiments, can be found in the Supplementary Material 6.1-6.5. Code to replicate the experiments can be found in the GitHub repository[2].

Denoising Attributions with sMPRT. Figure 2 visualises sMPRT scores with progressive randomisation of layers, in the bottom-up order. Results are shown for VGG-16 [33] and ResNet-18 [29] models for ImageNet [34], four different attribution methods and two choices of N (where

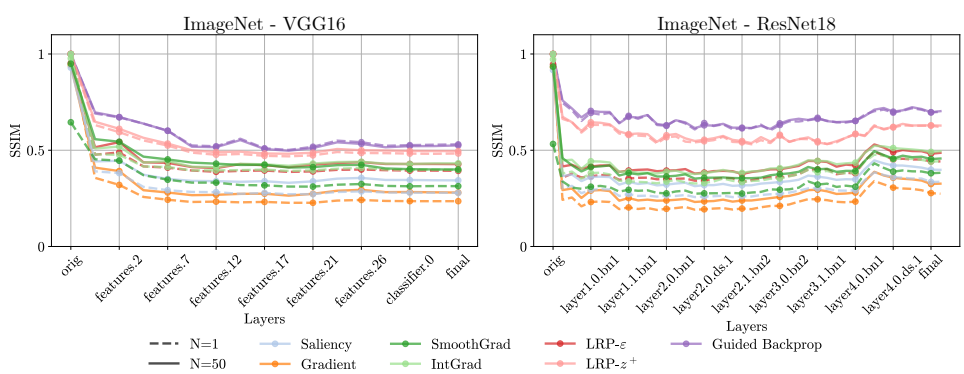

Figure 2: sMPRT ($N = 50$) versus MPRT ($N = 1$) results, showing that denoising attributions with sMPRT can degrade *SSIM* performance, which is most pronounced for gradient-based methods. *orig* and *final* denote the unrandomised and fully randomised model states, respectively.

[2]Code at: `https://github.com/annahedstroem/sanity-checks-revisited`

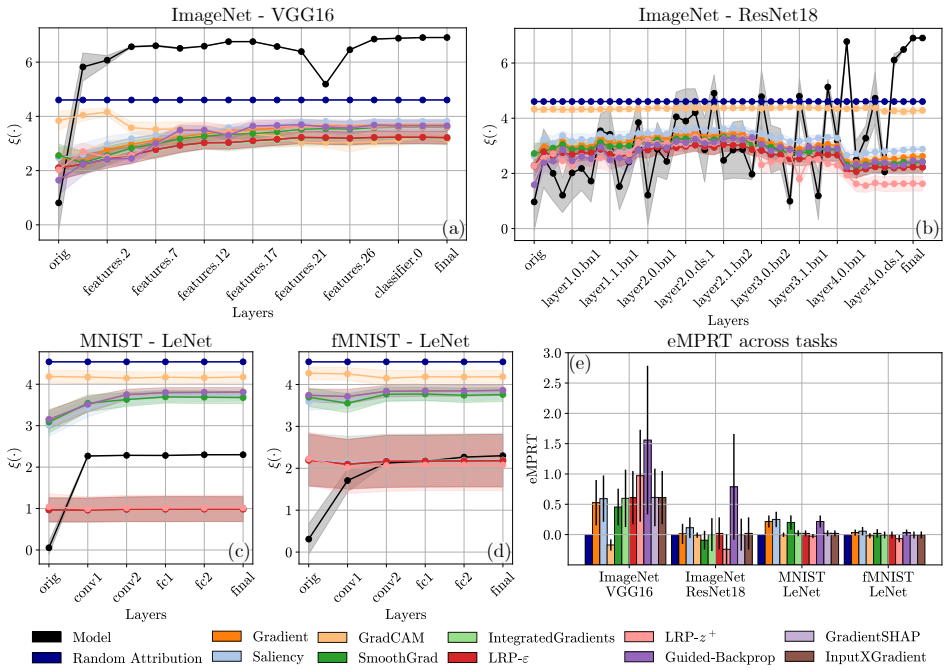

Figure 3: eMPRT curves (a)-(d) depicting the rise in complexity $\xi(\hat{e})$ upon progressive layer randomisation with bottom-up randomisation $\hat{f}_l^b$ for different models and datasets and aggregated eMPRT scores (e) for comparative benchmarking of different explanation methods.

$N = 1$ corresponds to the original MPRT). We observe that sMPRT—i.e., increasing the number of noise samples N to 50—significantly affects the evaluation outcome, with the degree of change varying for different attribution methods. Interestingly, after removing noise by adding noise, the investigated attribution methods perform more similarly under randomisation. This observation complements the theoretical findings of [20], which finds that the similarity scores computed by MPRT may be skewed in favour of (noisy) gradient-based explanation methods.

eMPRT in Action. Figure 3 visualises eMPRT curves in (a)-(d) with progressive randomisation of layers and aggregated eMPRT scores in (e), across tested tasks. Results are shown for ten different attribution methods and a random attribution baseline.

First, the different explanation methods begin with varying complexity where no single explanation method consistently outperforms others across tested tasks. Even after full model randomisation, no method approaches the theoretical limit to randomness *(dark blue)* one might expect from an explanation that acts maximally faithful to $\hat{f}$. Second, by comparing the model complexity (*black*, computed as non-discretised entropy of the model's output layer post-softmax, see Equation 6) with the different explanation complexities, we may anchor our expectation of how the complexity of an ideal explanation, i.e., "true to its model" [35], ought to develop. Interestingly, none of the explanation functions mirror the model function's complexity. Third, the progression of eMPRT curves exhibits considerable variation across different tasks. While variation is expected (given architectural specifics such as ResNets' skip connections), it may be difficult for a practitioner to interpret the explanation method's faithfulness across the different layers. As a more meaningful and robust measure of explanation faithfulness, we propose comparing explanations only before and after full model randomisation (see Equation 3). In Supplement 6.4, we further discuss the extent to which the evaluation rankings of explanation methods differ between eMPRT and MPRT. Notably, the random attribution method—serving as a theoretical lower bound for explanation faithfulness—consistently receives lower evaluation scores under eMPRT compared to MPRT, signifying an advantage with the eMPRT metric.

Meta-Evaluation. Importantly, to understand how the variants MPRT, sMPRT and eMPRT differ in terms of metric performance characteristics, we set out a benchmarking experiment, following the meta-evaluation methodology outlined in [3]. Each metric is assigned a summarising meta-consistency score, i.e., $\overline{\text{MC}} \in [0, 1]$. A higher $\overline{\text{MC}}$ score, approaching 1, signifies greater

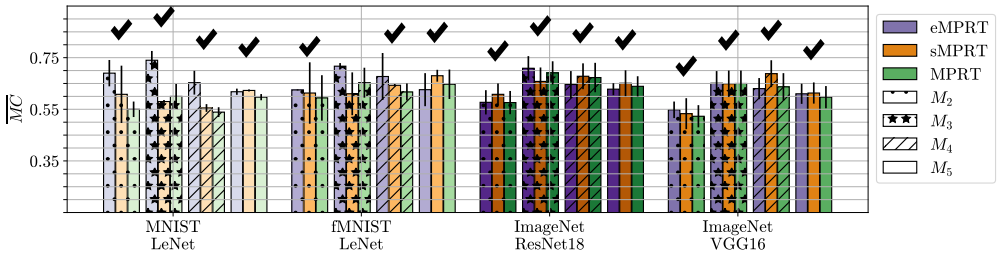

Figure 4: Average $\overline{\text{MC}}$ scores per metric across tested tasks (see in Table 1 in Supplement 6.5), aggregated over 3 iterations with $K = 5$ perturbations. Here, M_2, M_3, M_4, and M_5 represent different combinations of XAI methods, with M_2 ={*GradientSHAP, IntegratedGradients*}, M_3 ={*Saliency, LRP-z^+, Input×Gradient*},M_4 ={*Gradient, GradCAM, LRP-ε, Guided-Backprop*}, M_5 =*Guided-Backprop, GradientSHAP, GradCAM, LRP-ϵ, Saliency*}. The colour intensity denotes the task, i.e., the dataset and model combination.

reliability and superior performance based on the tested criteria. Details on the meta-evaluation framework can be found in the original publication [27]. A more comprehensive breakdown of the benchmarking results is included in Supplement 6.5, with experimental details in 6.1.

As Figure 4 illustrates, MPRT is generally outperformed by any of its variants (eMPRT or sMPRT) across the tested tasks and attribution sets $\{M_2, \cdots, M_5\}$ (see symbols over grouped bars). eMPRT and sMPRT consistently exhibit high reliability, substantiating its general applicability. These results are encouraging, as they suggest that the modifications made to MPRT have rendered the original metric more reliable. That said, Figure 4 shows that no variant showcases perfect reliability, i.e., where $\overline{\text{MC}} = 1$. This underlines the need for a continuously cautious application of XAI metrics, reflecting the broader challenge in XAI where ground truth explanation labels are often absent. Despite the performance advantage that eMPRT and sMPRT demonstrate, a contextualised (and not monolithic) metric adoption in XAI is strongly advocated.

5 Conclusion

In this work, we proposed two extensions of the MPRT [15], sMPRT and eMPRT, that address the caveats of pairwise similarity measures and layer randomisation order, which experimentally demonstrated improved reliability across various datasets- and model applications. Note, however, that both proposed variations of the MPRT suffer from unique drawbacks as well, with sMPRT being computationally inefficient, hyperparameter-dependent and confounding the noise existing in the attributions, an arguably explanation method-specific property. eMPRT, on the other hand, may possibly vary between different tasks, e.g., data-specific factors such as the size of the objects in a vision dataset. Each metric of evaluation has its own advantages and drawbacks, and we therefore emphasise the need for a broader evaluation, encompassing several properties to evaluate explanation quality.

Since both sMPRT and eMPRT employ fundamentally different mechanisms, a combination of both holds promise and will be subject to future work. Additionally, while we have started to explore the topic of layer-order for randomisation (see the Supplements 6.2), this analysis is not complete, with especially bottom-up randomisation deserving further empirical validation.

Acknowledgments and Disclosure of Funding

This work was supported by the German Federal Ministry for Education and Research through project Explaining 4.0 (ref. 01IS200551); BIFOLD (ref. 01IS18025A and ref. 01IS18037A); the Investitionsbank Berlin through BerDiBa (grant no. 10174498); the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant TEMA (101093003); and the European Union's Horizon 2020 research and innovation programme (EU Horizon 2020) as grant iToBoS (965221).

References

[1] I. Bellido and E. Fiesler, "Do backpropagation trained neural networks have normal weight distributions?" in *ICANN '93*, S. Gielen and B. Kappen, Eds. London: Springer London, 1993, pp. 772–775.

[2] S. Dasgupta, N. Frost, and M. Moshkovitz, "Framework for evaluating faithfulness of local explanations," in *International Conference on Machine Learning*. PMLR, 2022, pp. 4794–4815.

[3] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.

[4] G. Montavon, W. Samek, and K. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.

[5] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 7786–7795.

[6] C. Yeh, C. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 10 965–10 976.

[7] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, Ed. ijcai.org, 2020, pp. 3016–3022.

[8] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, and H. Lakkaraju, "Rethinking stability for attribution-based explanations," *CoRR*, vol. abs/2203.06877, 2022.

[9] A. Nguyen and M. R. Martinez, "On quantitative aspects of model interpretability," *CoRR*, vol. abs/2007.07584, 2020.

[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[11] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[12] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[13] L. Rieger and L. K. Hansen, "IROF: a low resource evaluation metric for explanation methods," *CoRR*, vol. abs/2003.08747, 2020.

[14] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for attribution methods," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 18 770–18 795.

[15] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 9525–9536.

[16] L. Sixt, M. Granz, and T. Landgraf, "When explanations lie: Why many modified BP attributions fail," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 9046–9057.

[17] M. Sundararajan and A. Taly, "A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values," *CoRR*, vol. abs/1806.04205, 2018.

[18] N. Kokhlikyan, V. Miglani, B. Alsallakh, M. Martin, and O. Reblitz-Richardson, "Investigating sanity checks for saliency maps with image and text classification," *CoRR*, vol. abs/2106.07475, 2021.

[19] G. Yona and D. Greenfeld, "Revisiting sanity checks for saliency maps," *CoRR*, vol. abs/2110.14297, 2021.

[20] A. Binder, L. Weber, S. Lapuschkin, G. Montavon, K. Müller, and W. Samek, "Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 16 143–16 152.

[21] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017.

[22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, ser. Lecture Notes in Computer Science, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8689. Springer, 2014, pp. 818–833.

[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 3319–3328.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.

[25] K. Bykov, A. Hedström, S. Nakajima, and M. M. Höhne, "Noisegrad - enhancing explanations by introducing stochasticity to model weights," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. AAAI Press, 2022, pp. 6132–6140.

[26] J. M. Benitez, J. L. Castro, and I. Requena, "Are artificial neural networks black boxes?" *IEEE Transactions on Neural Networks and Learning Systems*, vol. 8, no. 5, pp. 1156–1164, 1997.

[27] A. Hedström, P. Bommer, K. K. Wickstrøm, W. Samek, S. Lapuschkin, and M. M. C. Höhne, "The meta-evaluation problem in explainable ai: Identifying reliable estimators with metaquantus," *Transactions on Machine Learning Research*, 2023.

[28] P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, and S. Jha, "Concise explanations of neural networks using adversarial training," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1383–1391.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.

[30] T. Han, S. Srinivas, and H. Lakkaraju, "Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations," in *NeurIPS*, 2022.

[31] P. Bommer, M. Kretschmer, A. Hedström, D. Bareeva, and M. M. Höhne, "Finding the right XAI method - A guide for the evaluation and ranking of explainable AI methods in climate science," *CoRR*, vol. abs/2303.00652, 2023.

[32] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[35] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, 2020, https://distill.pub/2020/attribution-baselines.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[37] C. J. Anders, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy," *CoRR*, vol. abs/2106.13200, 2021.

[38] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," *CoRR*, vol. abs/2009.07896, 2020.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[40] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[41] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017.

[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[43] N. J. S. Morch, U. Kjems, L. K. Hansen, C. Svarer, I. Law, B. Lautrup, S. C. Strother, and K. Rehm, "Visualization of neural networks using saliency maps," in *Proceedings of International Conference on Neural Networks (ICNN'95), Perth, WA, Australia, November 27 - December 1, 1995*. IEEE, 1995, pp. 2085–2090.

[44] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, 2010.

[45] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.

[46] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *CoRR*, vol. abs/1605.01713, 2016.

[47] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4765–4774.

[48] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015.

[49] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, 2017.

[50] J. Nilsson and T. Akenine-Möller, "Understanding SSIM," *CoRR*, vol. abs/2006.13846, 2020.

[51] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L2 theory," *Z. Wahrscheinlichkeitstheorie verw Gebiete*, vol. 57, pp. 453–476, 1981.

[52] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

6 Supplementary Material

The Supplementary material first describes experimental hyperparameters used in the main paper in detail and then discusses several additional experimental results.

6.1 Experimental Setup

Software Environment. All experiments were implemented in python, using PyTorch [36] for deep learning. zennit [37] and captum [38] packages were used to compute explanations. The Quantus [3] package was utilised to for evaluations with MPRT and extended to implement sMPRT and eMPRT. The MetaQuantus library was employed for meta-evaluation [27].

Data. As shown by [19, 18], MPRT can be sensitive to the choice of task. We therefore aim to cover a wide range of datasets in this work. We use three image classification datasets in the experiments: ImageNet (ILSVRC2012 [39]), MNIST[40] and fMNIST [41]. For MNIST and fMNIST, we randomly sample 1000 test samples and for ImageNet we use the first 1000 samples of the validation set for experiments discussed in Figure 2 and randomly select 300 test samples for experiments showcased in Figure 3 and Figures 4-12.

Models. The experiments are performed using different neural network models, including architectures such as LeNets [42] and ResNets [29] and VGGs [33], which contributes to the robustness of our findings. For MNIST and fMNIST, we train LeNets to an accuracy of 98.14% and 87.44% respectively. The training of all models is performed in a similar fashion; employing SGD optimisation with a standard cross-entropy loss, an initial learning rate of 0.001 and momentum of 0.9. All models are trained for 20 epochs each. For ILSVRC-15 [34] and ILSVRC2012 [39], we use the ResNet-18 model [29] and VGG-16 [33] with pre-trained weights given the ImageNet dataset, accessible via `PyTorch` [36].

Explanation Methods and Preprocessing. We evaluated the following explanation methods using the MPRT variants: *Gradients* [43, 44], *Saliency* [45], *Input×Gradient* [46], GradCAM [24], *GradientSHAP* [47] with 5 samples, *SmoothGrad* [21] (with 20 noisy samples and a noise level of $0.1/(x_{\text{max}} - x_{\text{min}})$ as implemented as default in zennit [37], *Integrated Gradients* [23] (with 20 iterations and a baseline of zero), *Guided Backpropagation* [48] and two distinct variations of *Layer-wise Relevance Propagation* (LRP) [10, 49]: application of the *LRP-ε-rule* [10] to all layers (called LRP-ε) and application of the *LRP-z^+-rule* [49] to all layers (called LRP-z^+). For random attribution (as a control variant) we sampled from a uniform distribution, i.e., $\hat{\boldsymbol{e}}_i \sim \mathcal{U}(1, 0)$.

Since preprocessing can significantly affect the results of MPRT [17], we only take the absolute values for methods where the sign has no meaning in terms of feature importance (*Saliency* and *SmoothGrad*) and only consider the positive values for *GradCAM*, as per the original paper [24]. While the histogram entropy we used for eMPRT already normalises inputs implicitly, the *SSIM* used to measure explanation difference in MPRT and sMPRT requires normalisation to make attributions at different scales before and after randomisation comparable. For this purpose, we utilised normalisation by the square root of the average second-moment estimate (cf. Supplement of [20]), as this normalisation does not impose as much additional variance as normalisation by, e.g., maximum value would. The normalisation is defined as follows:

$$\text{norm}(\boldsymbol{e}) = \frac{\boldsymbol{e}_{h,w}}{\left(\frac{1}{HW}\sum_{h',w'} \boldsymbol{e}^2_{h',w'}\right)^{1/2}} \ , \tag{5}$$

where $\hat{\boldsymbol{e}}_{h,w}$ is the value of the explanation map at pixel location $(h,\ w)$ and H, W denote the height and width, respectively[3].

MPRT Hyperparameters. In the original paper [15], different similarity measures are used e.g., *Structural Similarity Index* (SSIM) [50], *Spearman Rank Correlation* and *HOG*. We employ *SSIM* [50], motivated by its widespread adoption in existing literature [15, 16, 20].

[3]This normalisation technique guarantees that the mean squared distance of each attribution score to zero equals one. As the procedure does not constrain attributions to a fixed range, it is not designed for visual representation but aims to preserve a measure useful for comparing the distances between different explanation methods.

sMPRT Hyperparameters. For the experiments with sMPRT, we employ *SSIM* [50] to measure explanation similarity. For generating noise (σ), we used an adaptive standard deviation of $0.2/(x_{\text{max}} - x_{\text{min}})$, following the original heuristic presented in [21]. For the number of noise samples (N) to average over, we investigated the following values: $1, 20, 50, 300$. Evaluation for those values is shown in Supplement 6.3. While a larger N generally improves the denoised explanation estimate, it also increases computation time significantly. For Section 3.1, we thus chose $N = 50$ as our estimate with acceptable runtime.

eMPRT Hyperparameters. In our experiments, for Equation 4, we empirically set the bin count $B = 100$, which demonstrated robust performance across diverse experiments and explanation methods. Although various statistical rules like *Freedman-Diaconis'* [51] and *Scotts'* [52] exist for optimising B, initial tests indicated inconsistency in their performance, largely due to their assumption of data normality, which is generally not applicable to attribution data.

To measure model complexity, we computed the Shannon entropy of the model output (post-softmax) probabilities as follows:

$$\xi(f_L(\boldsymbol{x})) = -\sum_{i=1}^{|L|} p(x_i) \log_2(p(x_i)) \tag{6}$$

Meta-Evaluation Details. To meta-evaluate the metrics, i.e., capture their metric performance characteristics in terms of statistical reliability, we employed the MetaQuantus library [3]. For this, we measured both their resilience to noise (NR) and their reactivity to adversary (AR) by inducing controlled perturbations, where we applied minor and disruptive perturbations to the input- and model spaces, respectively[4]. Here, we computed the intra- ($\mathbf{IAC}$) and inter-consistency ($\mathbf{IEC}$) scores, which include measuring the similarity in score distributions and ranking of different explanation methods (post-perturbation), respectively. Each metric received a summarising meta-consistency (MC) score with MC $\in [0, 1]$:

$$\mathbf{MC} = \left(\frac{1}{|\mathbf{m}^*|}\right) \mathbf{m}^{*T}\mathbf{m} \quad \text{where} \quad \mathbf{m} = \begin{bmatrix} \mathbf{IAC}_{NR} \\ \mathbf{IAC}_{AR} \\ \mathbf{IEC}_{NR} \\ \mathbf{IEC}_{AR} \end{bmatrix}, \tag{7}$$

with $\boldsymbol{m}^* = \mathbb{R}^4$ represents an optimally performing quality estimator as defined by the all-one indicator vector. For exact definitions for the contents of the meta-evaluation vector $\mathbf{m}$, we refer to the original publication [27].

Meta-Evaluation Hyperparameters. For empirical assessment, we utilised the pre-existing tests available in the MetaQuantus library [3] and their associated hyperparameters. The existing MetaQuatus tests are accessible via the GitHub repository[5]. All metrics have been implemented in Quantus [3]. We executed these metrics over $K = 5$ perturbations, spanning 3 iterations with the test configurations as outlined in the notebook[6]. To ensure a fair comparison among the metrics, we adhered to uniform hyperparameter settings, as specified by the normalisation formula in Equation 5. Given that the MetaQuantus library necessitates each metric to yield a single quality estimate $\hat{q} \in \mathbb{R}$ per sample, we calculated the correlation coefficient for the fully randomised models in both MPRT and sMPRT, which is anticipated to enhance their performance.

6.2 Effect of Layer Randomisation Order on MPRT Results

To avoid the preservation of significant portions of the forward pass that occurs with the latter (cf. and 2.1.1 (b) and [20]), [20] proposes to randomise layers in bottom-up order instead of

[4]We applied i.i.d additive uniform noise such that $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \boldsymbol{\delta}_i$ with $\boldsymbol{\delta}_i \sim \mathcal{U}(\alpha, \beta)$ for the Input Perturbation Test and applied multiplicative Gaussian noise to all weights of the network, i.e., $\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta} \cdot \boldsymbol{\nu}_i$ with $\boldsymbol{\nu}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the Model Perturbation Test. The hyperparameters $\alpha, \beta, \mu, \Sigma$ were set according to the original publication [27].

[5]Code at: `https://github.com/annahedstroem/MetaQuantus/`.

[6]See hyperparameter settings at notebook: `https://github.com/annahedstroem/MetaQuantus/blob/main/tutorials/Tutorial-Getting-Started-with-MetaQuantus.ipynb`

top-down. Since bottom-up layer-order randomises the lowest layers first, such preservation cannot occur.

Given the use of bottom-up randomisation in sMPRT, and its current lack of empirical validation, we briefly examine its impact on MPRT in the following. Note that a better option may be to avoid progressive layer randomisation altogether, simply comparing the unrandomised state with the fully randomised state (see Equation 3).

Figure 5 compares MPRT rankings for top-down *(top)* and bottom-up *(bottom)* randomisation order. Note that in contrast to the seminal work [15], we apply a different preprocessing to the explanations where we normalise by the square root of the average second-moment estimate [20] (also cf. discussion in Section 2.1.1, caveat (a)). We also take absolute values only for *Saliency* and *SmoothGrad*.

First, we validate that the model is sufficiently disturbed by both orderings by also tracking the model accuracy. This is indeed the case for all four settings (see *black* lines), where the accuracy drops to random for the ImageNet experiments (around 0.001) immediately after the respective first layer is randomised. MPRT results are strongly affected by the randomisation order, however—not only do all explanation methods perform better with bottom-up randomisation, but since their *SSIM* drops more steeply with randomisation, the differences in *SSIM* score between methods significantly diminish. Especially, methods that appeared almost invariant to model parameters with top-down randomisation (*LRP-z^+* and *Guided Backpropagation*) show significantly more faithfulness to the model with bottom-up randomisation.

The above observations are confirmed by Figure 6 *(a1) and (a2)*: Explanation methods perform better under the MPRT with bottom-up randomisation (i.e., the dotted hatch bars are lower than the striped bars). While this does not translate to the categorical rankings between explanation

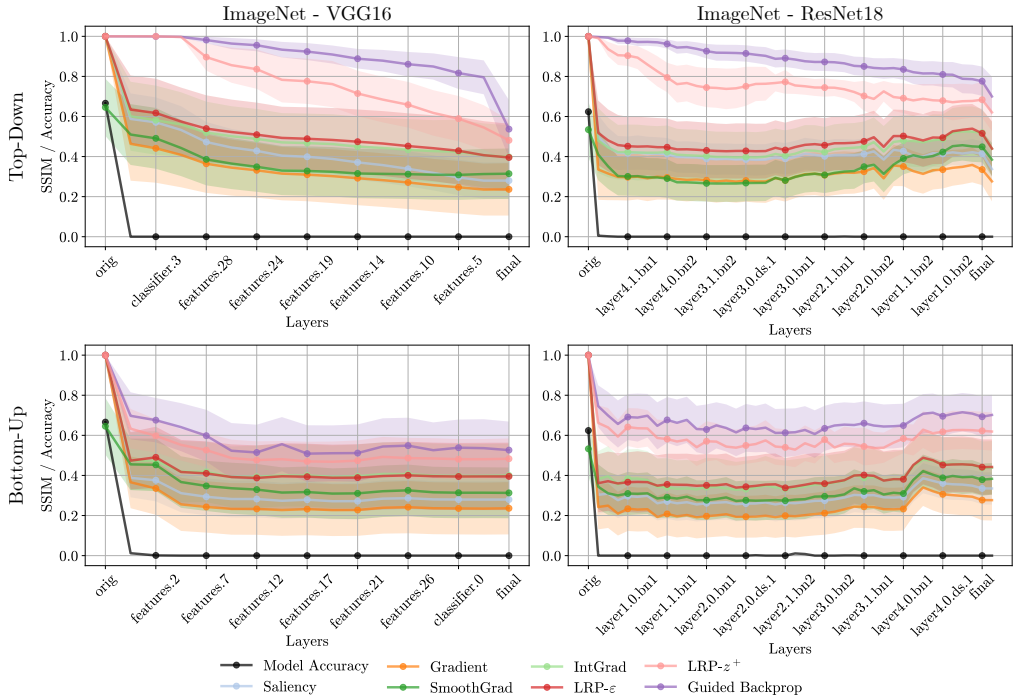

Figure 5: Effect of layer randomisation order on MPRT results for ImageNet, using VGG16 *(left)* and ResNet18 *(right)*. The *black line* shows how model accuracy changes with randomisation, *other lines* show the *SSIM* measured by MPRT for different explanation methods. *orig* and *final* denote the unrandomised and fully randomised model states, respectively. *(Top)* Top-down randomisation, starting with the output layer, as performed by [15]. *(Bottom)* Bottom-up randomisation, starting with the lowest layer, as suggested by [20]. Model accuracy degrades with the same speed for both randomisation orders. However, there is an significant discrepancy in explanation method *SSIM* scores: differences between methods seem much more severe for top-down than bottom-up randomisation.

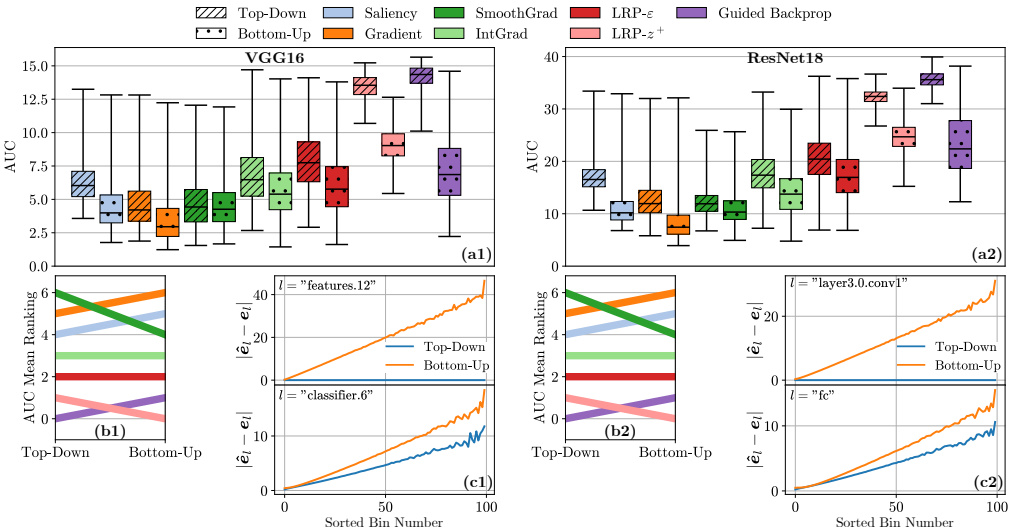

Figure 6: Effect of layer randomisation order on MPRT results. Evaluation was performed for VGG16 *(left)* and ResNet18 *(right)* models, on the ImageNet dataset. *a1 and a2*: Area under MPRT curves (cf. Figure 5) for various XAI methods and both layer orders. Percentile (0, 25, 50, 75, 100) statistics are visualised. *b1 and b2*: Relative ranking changes between XAI methods from top-down to bottom-up order. Note that a lower score in (a) corresponds to a higher ranking in (b). *(c1 and c2)*: Values of intermediate explanation $\boldsymbol{e}$ (cf. [20]), assigned to 100 bins according to their absolute value (x-axis) vs. the average amount of change with randomisation for values in each bin (y-axis).

methods *(b1) and (b2)*, which seem only slightly affected by randomisation order, keep in mind that there are also other interfering effects to consider, such as the sensitivity of *SSIM* to noise [20], as described in Section 2.1.1, issue (c). These effects are likely to confound the bottom-up results as well and are addressed separately by the sMPRT and eMPRT metrics proposed in this work.

To understand why bottom-up randomisation leads to generally better MPRT performance, we inspect the (LRP-ε) explanations $\boldsymbol{e}_l$ of the intermediate activations at layer l —i.e, $\boldsymbol{e}_l$ explains the output of layer l as opposed to the whole model (which would correspond to output layer L). The values of the original explanation $\boldsymbol{e}_l$ are sorted into 100 equidistant bins according to their magnitude. We then randomise the model bottom-up or top-down, to compute randomised explanations $\hat{\boldsymbol{e}}_l$. For bottom-up and top-down order, we randomise only the first and last parameterised layer of the model, respectively.

For each bin, we evaluate how much the explanation values change on average under each randomisation (i.e., we compute the average $|\hat{\boldsymbol{e}}_l - \boldsymbol{e}_l|$ for each bin). Figure 6 *(c1) and (c2)* shows the results of this analysis for two different layers l—for each model, an intermediate layer *(top)* in addition to the output layer *(bottom)*. Here, we observe that bottom-up randomisation (orange line) affects both the intermediate layer explanation and the last layer explanation, while top-down randomisation (blue line) only affects the latter (although significantly less than bottom-up randomisation). Bottom-up randomisation seems to have a stronger impact on the model's reasoning compared to top-down randomisation. For a robust MPRT, it is necessary to randomise the model output as fully as possible—because only then faithful explanations can be expected to change significantly [20]. In this aspect, bottom-up randomisation seems to be a better choice compared to top-down randomisation.

6.3 sMPRT: Additional Experiments

In the following, we investigate the effect of the number of perturbed samples N on sMPRT results. To make the curves obtained from sMPRT (cf. Figure 2) comparable across different N, we condense them to a singular value by computing the area below each curve (denoted as AUC). Lower AUC indicates more faithfulness towards model parameters.

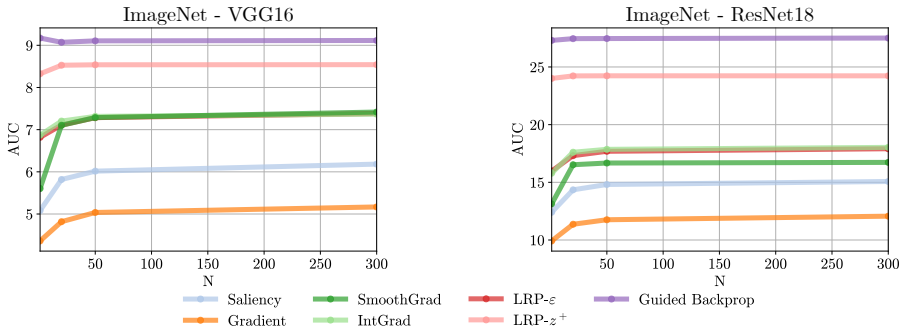

Figure 7: Effect of number of perturbed samples N on sMPRT results for VGG16 *(left)* and ResNet18 *(right)* on ImageNet data. The plots indicate how the area under the mean sMPRT curve (AUC) changes with N for different explanation methods (i.e. the area under the curves as shown in Figure 2). Up to $N = 50$, there seems to be a steep change in AUC, especially for gradient-based methods. After that, the AUC curves flatten out, indicating a converged estimate of the denoised samples.

Figure 7 shows how these AUC values progress with increasing N. Especially gradient-based methods seem to vary significantly with N. However, the AUC seems to converge for large N, indicating that the estimation of the denoised samples also converges. While slight variations can be observed up to the last investigated value, $N = 300$, AUC does not change much after $N = 50$. For the sake of efficient evaluation, we therefore recommend to use $N = 50$ to obtain a decent estimate of the correct ranking, albeit a larger N will always yield a better estimation.

6.4 eMPRT: Additional Experiments

In Figure 8, the categorical rankings derived from the evaluation metrics MPRT and eMPRT are visualised for ten distinct attribution methods, as well as a random attribution. Rankings are organised in descending order, where $R1$ denotes the best performance and $R11$ is the worst score. The findings corroborate those of [27], demonstrating considerable variability in explanation rankings between the two metrics. In contrast to the original publication [15], which asserted *Guided Backpropagation* as inferior to gradient-based methods such as *Gradient* and *SmoothGrad*, our eMPRT evaluation reveals a contrasting ranking, even advancing *Guided Backpropagation* above its gradient-based counterparts. The random attribution—serving as a theoretical lower bound for explanation quality–consistently receives lower scores for eMPRT compared to MPRT, signifying an advantage of our proposed metric.

6.5 Benchmarking: Additional Details

In this section, we present a detailed breakdown of the benchmarking results. First, we present the numerical version of Figure 4, presented in the main manuscript. Here, $M_2 =\{$*GradientSHAP, In-*

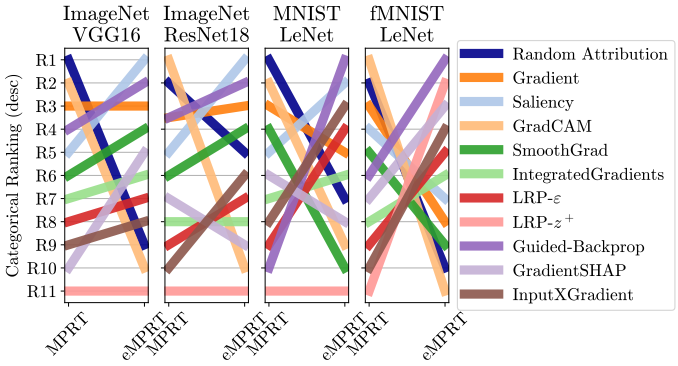

Figure 8: Relative ranking of ten attribution methods and a random attribution using MPRT and eMPRT evaluation. The rankings, consistent with findings from [27], vary significantly between the two metrics.

tegratedGradients}, M_3 ={*Saliency, LRP-z^+, Input×Gradient*},M_4 ={*Gradient, GradCAM, LRP-Eps, Guided-Backprop*}, M_5 =*Guided-Backprop, GradientSHAP, GradCAM, LRP-ϵ, Saliency*}.

Table 1: Meta-evaluation benchmarking results, aggregated over 3 iterations with $K = 5$. The top-performing metric in each setting is highlighted in bold. If it outperforms the next-best-performing metric by at least one standard deviation, it is underlined. Higher values are preferred.

Setting	XAI Methods	eMPRT (↑)	sMPRT (↑)	MPRT (↑)
fMNIST LeNet	M_2	**0.625 ± 0.002**	0.613 ± 0.119	0.594 ± 0.088
	M_3	**0.717 ± 0.012**	0.610 ± 0.083	0.653 ± 0.058
	M_4	**0.677 ± 0.091**	0.643 ± 0.005	0.617 ± 0.033
	M_5	0.626 ± 0.064	**0.680 ± 0.023**	0.647 ± 0.057
MNIST LeNet	M_2	**0.690 ± 0.051**	0.608 ± 0.111	0.550 ± 0.030
	M_3	**0.740 ± 0.036**	0.580 ± 0.006	0.600 ± 0.048
	M_4	**0.653 ± 0.046**	0.555 ± 0.015	0.539 ± 0.019
	M_5	0.618 ± 0.012	**0.623 ± 0.004**	0.597 ± 0.012
ImageNet ResNet18	M_2	0.577 ± 0.047	**0.608 ± 0.042**	0.576 ± 0.045
	M_3	**0.709 ± 0.047**	0.658 ± 0.054	0.691 ± 0.045
	M_4	0.646 ± 0.052	**0.678 ± 0.050**	0.673 ± 0.057
	M_5	0.628 ± 0.023	**0.651 ± 0.050**	0.639 ± 0.039
ImageNet VGG16	M_2	**0.548 ± 0.032**	0.533 ± 0.06	0.523 ± 0.043
	M_3	**0.651 ± 0.047**	0.649 ± 0.051	0.650 ± 0.050
	M_4	0.630 ± 0.041	**0.688 ± 0.052**	0.637 ± 0.053
	M_5	**0.610 ± 0.039**	0.613 ± 0.041	0.597 ± 0.043

In the following Figures 9-12, we show the different area graphs which each contains the entries of the meta-evaluation vector (set as coordinates on a 2D plane).

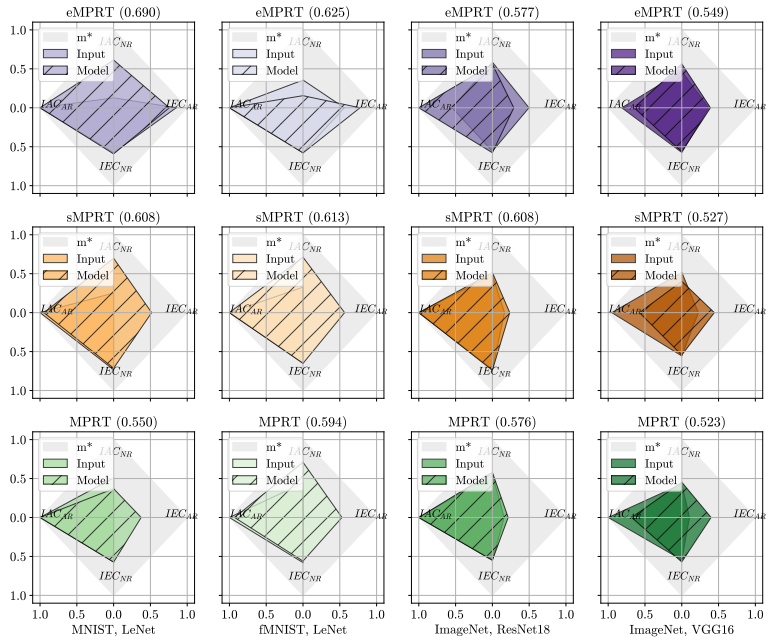

Figure 9: A graphical representation of the benchmarking results for $M = 2$ for MPRT, sMPRT and eMPRT, aggregated over 3 iterations with $K = 5$. The grey area indicates the area of an optimally performing estimator, i.e., $\mathbf{m}^* = \mathbb{R}^4$. The MC score (indicated in brackets) is averaged over MPT and IPT. Higher values are preferred.

By inspecting the coloured areas of the respective estimators in terms of their size and shape, we can deduce the overall performance of both failure modes (NR, AR). Here, larger coloured areas

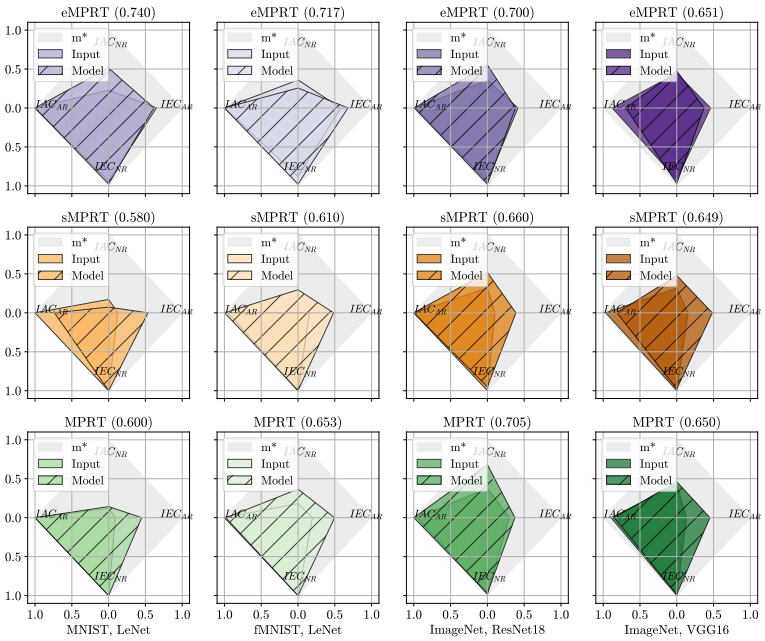

Figure 10: A graphical representation of the benchmarking results for $M = 2$ for MPRT, sMPRT and eMPRT, aggregated over 3 iterations with $K = 5$. The grey area indicates the area of an optimally performing estimator, i.e., $\mathbf{m}^* = \mathbb{R}^4$. The MC score (indicated in brackets) is averaged over MPT and IPT. Higher values are preferred.

imply better performance on the different scoring criteria and the grey area indicates the area of an optimally performing quality estimator, i.e., $\boldsymbol{m}^* = \mathbb{R}^4$. The shades of the different colours, i.e., purple, orange and green indicate the datasets and are used consistently across the Figures 9-12 (as well as Figure 4 in the main manuscript). The columns show the tasks and the rows the different metrics: MPRT, sMPRT and eMPRT. With few exceptions, eMPRT and sMPRT demonstrate superior performance, on both the input- and the model perturbation tests, across tested tasks and subsets of attribution methods.

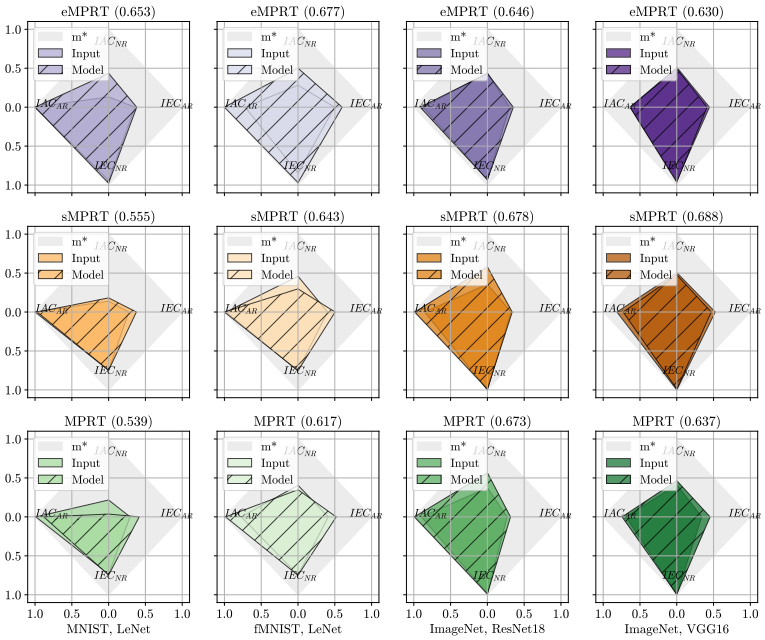

Figure 11: A graphical representation of the benchmarking results for $M = 2$ for MPRT, sMPRT and eMPRT, aggregated over 3 iterations with $K = 5$. The grey area indicates the area of an optimally performing estimator, i.e., $\mathbf{m}^* = \mathbb{R}^4$. The MC score (indicated in brackets) is averaged over MPT and IPT. Higher values are preferred.

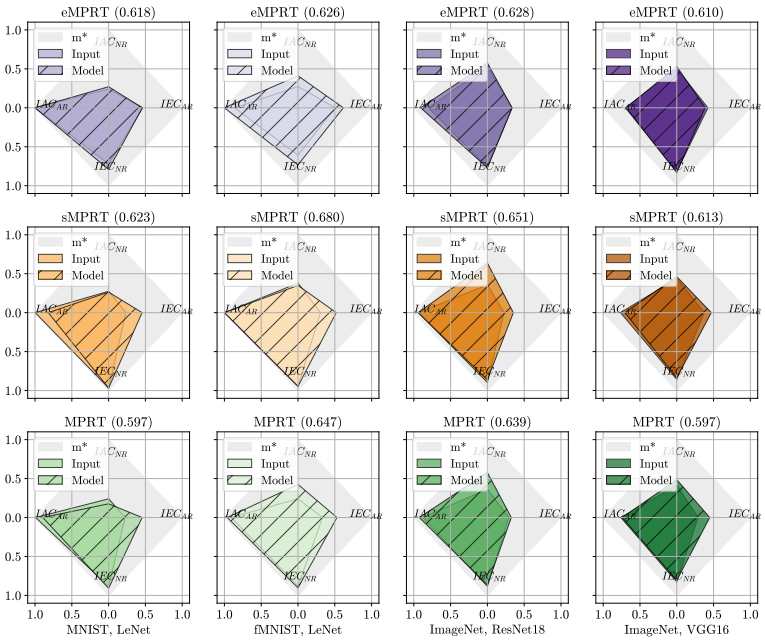

Figure 12: A graphical representation of the benchmarking results for $M = 2$ for MPRT, sMPRT and eMPRT, aggregated over 3 iterations with $K = 5$. The grey area indicates the area of an optimally performing estimator, i.e., $\mathbf{m}^* = \mathbb{R}^4$. The MC score (indicated in brackets) is averaged over MPT and IPT. Higher values are preferred.