Towards Quantifying Incompatibilities in Evaluation Metrics for Feature Attributions

Thomas Decker[1,2,3], **Volker Tresp** [2,3]

[1]Siemens AG [2]LMU Munich [3]Munich Center for Machine Learning (MCML)
thomas.decker@siemens.com, volker.tresp@lmu.de

Abstract

Feature attribution methods are widely used to explain machine learning models, yet their evaluation is challenging due to competing quality criteria such as faithfulness, robustness, and sparsity. These criteria often conflict, and even alternative formulations of the same metric can yield inconsistent conclusions. We address this by introducing a unifying framework that analyzes systematic incompatibilities between measures of explanation quality. Within this framework, we develop two novel mathematical tools: a sample-wise incompatibility index that quantifies systematic conflicts between criteria, and a generalized eigen-analysis that localizes where tradeoffs are concentrated within attribution results. Experiments on image classifiers show that this analysis provides insights beyond isolated metrics and complements current evaluation practices for feature attributions.

Introduction

Feature attribution methods (Ancona et al. 2019; Covert, Lundberg, and Lee 2021; Chen et al. 2023; Wang et al. 2024) have become essential tools for interpreting machine learning models by providing insights into which input features drive predictions. As a plethora of different attribution methods exists, selecting the most appropriate explanation for a given instance remains challenging. This challenge is compounded by the fact that different methods often produce markedly different feature rankings for the same prediction (Krishna et al. 2022; Decker et al. 2024), leaving practitioners without clear guidance on which concrete explanation to rely on. This is particularly problematic when explanations are leveraged in the context of discovering new scientific knowledge (Zednik and Boelsen 2022), where selecting the wrong explanation can lead to false hypotheses and misleading conclusions about the true nature of the data-generating process. To better guide method selection, several quality criteria have been proposed, such as faithfulness (how accurately the explanation reflects the model's behavior), robustness (explanation stability under tiny input perturbations), and sparsity (complexity of the derived explanation for humans). However, evaluating explanations against these criteria in the absence of an objective ground truth is inherently difficult. Moreover, quality metrics themselves can yield inconsistent results (Tomsett et al. 2020), making reliable assessment challenging. Consequently, finding optimal evaluation practices for individual use cases remains an active research area (Hedström et al. 2023a), and the relationships between different quality criteria are not well understood.

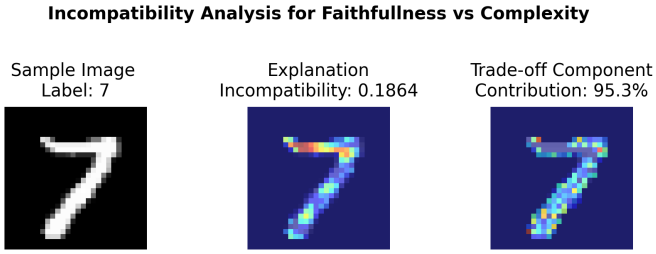

Figure 1: Illustration of our proposed incompatibility analysis for understanding systematic tradeoffs between Faithfulness and Complexity. For a given input (left) and its explanation result, we compute: (i) a sample-wise incompatibility index quantifying the severity of the tradeoff (middle), and (ii) the dominant eigen-mode localizing where the conflict is concentrated within the attribution (right, with percentage contribution). Higher incompatibility indices indicate stronger conflicts between criteria, while the eigen-mode reveals which attribution components are most responsible for the observed tradeoff.

Recent theoretical results even indicate that for individual types of attribution methods, improving one criterion can directly degrade another (Tan and Tian 2023; Mehrpanah et al. 2025; Fokkema, de Heide, and van Erven 2023; Zhang et al. 2023; Bilodeau et al. 2024). These examples suggest that certain desirable explanation properties may be inherently incompatible, which can impede the evaluation process.

Despite these advances, a systematic framework for understanding and quantifying these incompatibilities is lacking. Existing work identifies specific tradeoffs but does not provide tools to measure their severity on individual instances or to localize where within an attribution result these conflicts are concentrated. This limits practitioners' ability to make informed decisions when quality criteria conflict.

We address this gap by introducing a principled mathematical framework for analyzing systematic incompatibilities between explanation quality criteria (see Figure 1). We leverage the fact that common criteria can be unified as

generalized L^2 metrics (Decker et al. 2024) with quadratic representations, which enables systematic pairwise analysis. Our contributions can be stated as follows:

- We study a unifying framework that analyzes systematic incompatibilities between measures of explanation quality by representing faithfulness, robustness, and complexity as generalized L^2 metrics.
- We introduce a sample-wise incompatibility index that quantifies systematic conflicts between criteria at the individual instance level, and develop a generalized eigenanalysis that localizes where tradeoffs are concentrated within attribution results.
- We demonstrate through experiments on image classifiers that our analysis provides insights beyond isolated metrics and complements current evaluation practices for feature attributions.

Background and Related Work

Problem Setup

We consider a trained classification model $f : \mathbb{R}^d \to \mathbb{R}$ that maps an input instance $x \in \mathbb{R}^d$ to a scalar prediction $f(x)$ representative of a class. A *feature attribution method* $\phi : \mathbb{R}^d \to \mathbb{R}^d$ explains the prediction $f(x)$ by assigning an importance score $\phi_i(x)$ to each input feature x_i, where higher absolute values indicate greater relevance for the model's decision.

The quality of feature attributions is typically assessed through various evaluation metrics. Let $\mathcal{Q} : \mathbb{R}^d \to \mathbb{R}$ denote a quality metric that quantifies a specific desirable property of an attribution result $\phi(x)$. Different metrics capture distinct aspects of explanation quality, and understanding their interplay is crucial for reliable model interpretability.

Explanation Quality Criteria

Reliably evaluating feature attributions remains challenging due to the absence of ground truth explanations in most practical scenarios (Le et al. 2023). Nevertheless, the research community has developed several relevant quality criteria that attributions should satisfy and can guide explanation assessment in practice (Nauta et al. 2023). We focus on three distinct dimensions that are generally applicable and therefore most often considered: faithfulness, robustness, and complexity.

Faithfulness measures how accurately an attribution reflects the model's actual decision-making process. A faithful explanation should assign high importance to features that genuinely influence the prediction. Formally, we adpot *Infidelity* (Yeh et al. 2019) as related faithfulness metric:

$$\text{INFD} : \mathbb{E}_{I\sim\mathbb{P}_I}\left[\left(I^T\phi(x) - (f(x) - f(x\odot(1-I)))\right)^2\right]$$

where $I \in \{0,1\}^d$ is a random binary mask sampled from distribution $\mathbb{P}_I$, and $\odot$ denotes element-wise multiplication. Intuitively, Infidelity measures the squared difference between the predicted importance of perturbed features (weighted by $I^T\phi(x)$) and the actual change in model output when those features are removed. Lower infidelity indicates better faithfulness.

Robustness assesses how stable attributions remain when inputs undergo minor perturbations. For explanations to be trustworthy, they should exhibit consistency across similar inputs—large variations can erode user confidence and suggest the attribution captures artifacts rather than genuine feature relationships (Alvarez-Melis and Jaakkola 2018; Ghorbani, Abid, and Zou 2019). We quantify robustness via *Average Sensitivity* (Bhatt, Weller, and Moura 2021):

$$\text{SENS}_{\text{AVG}} : \mathbb{E}_{\varepsilon\sim\mathbb{P}_\varepsilon}\left[\|\phi(x) - \phi(x+\varepsilon)\|_2^2\right]$$

where ε denotes a small random perturbation, commonly sampled from a Gaussian or uniform distribution with centered support around zero. Lower average sensitivity consequently indicates more robust attributions.

Complexity captures the interpretability of an explanation from a human-centered perspective. Cognitive science research suggests that humans struggle to process dense information, making sparse explanations more comprehensible (Miller 2019; Bhatt, Weller, and Moura 2021). We quantify complexity using *Effective Complexity* (Nguyen and Martínez 2020):

$$\text{CMPLX}_\tau := \frac{1}{d}\sum_{i=1}^{d}\mathbb{I}\left\{|\phi_i(x)| > \tau\right\}$$

where $\tau > 0$ is a threshold parameter and $\mathbb{I}\{\cdot\}$ is the indicator function. This metric measures the fraction of features with attribution magnitudes exceeding τ, with lower values indicating sparser, simpler explanations.

These metrics are readily available in established evaluation frameworks (Hedström et al. 2023b; Le et al. 2023), enabling systematic empirical analysis.

Related Work

Recent work has revealed fundamental conflicts between explanation quality criteria. Tan and Tian (2023) prove theoretically that enhancing robustness in gradient-based explanations systematically reduces faithfulness. Mehrpanah et al. (2025) demonstrate a complexity-faithfulness tradeoff, quantifying how noise reduction induces fidelity loss. Zhang et al. (2023) establish an "Impossible Trinity" showing that for certain attribution techniques, one cannot simultaneously achieve sparsity, efficiency, and high fidelity.

From a theoretical perspective, Fokkema, de Heide, and van Erven (2023) formalize these tensions through impossibility results, proving that no local attribution method can be simultaneously recourse-sensitive and robust to input perturbations. Tadesse et al. (2025) propose directly optimizing explanations for desired properties, allowing users to navigate tradeoffs without characterizing the systematic nature of conflicts.

Our work complements existing research approaches in several ways by introducing a novel mathematical framework to analyze and comprehend incompatibilities of different quality criteria in a theoretically grounded way.

A Mathematical Framework for Incompatibility of Quality Criteria

In this section, we develop a unifying mathematical framework to systematically analyze incompatibilities between explanation quality criteria. We introduce two novel tools: (i) a sample-wise incompatibility index that quantifies the degree of systematic conflict between metrics, and (ii) a generalized eigen-analysis that localizes which components of an attribution result contribute most to observed tradeoffs.

Generalized L^2 Metrics

To enable systematic analysis across diverse quality criteria, We start by introducing the concept of generalized $L2$ metrics (Decker et al. 2024) as a canonical representation for explanation quality metrics. This reveals the shared mathematical structure underlying seemingly disparate evaluation approaches.

Definition 1 (Generalized L^2 Metric)**.** *A quality metric $\mathcal{Q}$ belongs to the class of* generalized L^2 metrics *if there exist random variables $\gamma_1 \in \mathbb{R}^{g\times d}$ and $\gamma_2 \in \mathbb{R}^g$ such that:*

$$\mathcal{Q}(\phi(x)) = \mathbb{E}_{\gamma_1,\gamma_2}\left[\|\gamma_1\phi(x) - \gamma_2\|_2^2\right]$$

Each such metric evaluates explanation quality of an attribution result by applying a linear probe γ_1 to assess a specific numerical properties of ϕ and checking against a desired value γ_2 in L^2 Norm.

Interestingly, the commonly used evaluation metrics introduced in Section 2.2 all admit this representation. For *Average Sensitivity*, let $\mathcal{I}_d$ denote the d-dimensional identity matrix. Setting $\gamma_1 = \mathcal{I}_d$ and $\gamma_2 = \phi(x+\varepsilon)$ with $\varepsilon \sim \mathbb{P}_\varepsilon$ recovers SENS_{AVG} with respect to the squared Euclidean norm. For *Infidelity*, choosing $\gamma_1 = I^T$ (where I is the perturbation mask) and $\gamma_2 = f(x) - f(x \odot (1-I))$ yields the metric definition.

To mimic *Effective Complexity*, we can reformulate the underlying idea by leveraging the truncated L^2 norm as a sparsity measure (Dicker 2014). Consider the metric $\mathcal{Q}(\phi(x)) = \|\min\{|\phi(x)|, t\}\|_2^2$, where $\min\{\cdot,\cdot\}$ denotes the element-wise minimum and t is a noise threshold. Minimizing this metric encourages pushing more entries of $\phi(x)$ below the threshold t, thereby promoting sparsity equivalently to Effective Complexity as defined above. This translates to Definition 1 by setting $\gamma_1 \in \mathbb{R}^{d\times d}$ and $\gamma_2 \in \mathbb{R}^d$ as:

$$(\gamma_1)_{i,j} = \begin{cases} 1 & \text{if } i = j \text{ and } |\phi_i| < t \\ 0 & \text{otherwise} \end{cases} \quad \text{and}$$

$$(\gamma_2)_i = \begin{cases} t & \text{if } |\phi_i| > t \\ 0 & \text{otherwise} \end{cases}$$

Table 1 summarizes how our three primary quality criteria map to the generalized L^2 framework.

This unified representation enables systematic mathematical analysis. In particular, all generalized L^2 metrics admit a quadratic form:

Proposition 1 (Quadratic Representation)**.** *If $\mathcal{Q}$ is a generalized L^2 metric, then it has the quadratic representation*

$$\mathcal{Q}(\phi) = \phi^\top A\phi - 2b^\top\phi + c,$$

where $A = \mathbb{E}[\gamma_1^T\gamma_1] \in \mathbb{R}^{d\times d}$ is positive semi-definite ($A \succeq 0$), $b = \mathbb{E}[\gamma_1^T\gamma_2] \in \mathbb{R}^d$, and $c = \mathbb{E}[\|\gamma_2\|_2^2] \in \mathbb{R}$ is a constant.

This quadratic structure immediately implies that each metric has a well-defined optimum:

Proposition 2 (Optimal Metric Value)**.** *For any generalized L^2 metric $\mathcal{Q}$ with quadratic form $(\phi^T A\phi - 2b^T\phi + c)$, the minimal achievable value is:*

$$m(\mathcal{Q}) = \min_{\phi\in\mathbb{R}^d} \mathcal{Q}(\phi) = c - b^T A^\dagger b$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse.

These optimal values provide the foundation for quantifying systematic incompatibilities between quality criteria, as we develop next.

Incompatibility Index

The quadratic structure of generalized L^2 metrics enables us to rigorously analyze relationships between multiple quality criteria. Consider two metrics $\mathcal{Q}_1$ and $\mathcal{Q}_2$ with quadratic representations $(\phi^T A_1\phi - 2b_1^T\phi + c_1)$ and $(\phi^T A_2\phi - 2b_2^T\phi + c_2)$, respectively. Their sum naturally defines a joint optimization objective:

$$\begin{aligned}\mathcal{Q}_{1,2}(\phi) &= \mathcal{Q}_1(\phi) + \mathcal{Q}_2(\phi) \\ &= \phi^T(A_1 + A_2)\phi - 2(b_1 + b_2)^T\phi + (c_1 + c_2)\end{aligned}$$

By Proposition 2, the joint minimal value is:

$$m_{1,2} = (c_1 + c_2) - (b_1 + b_2)^T(A_1 + A_2)^\dagger(b_1 + b_2)$$

This value represents the best possible combined metric score any attribution result can achieve when optimizing both criteria simultaneously. Crucially, if the two criteria are incompatible, this joint optimum will be worse than the sum of individual optima due to conflicting optimization directions. This observation motivates our incompatibility measure:

Definition 2 (Incompatibility Index)**.** *Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ be two generalized L^2 metrics representing distinct quality criteria for a feature attribution result ϕ. Their* incompatibility index *is defined as:*

$$\mathcal{I}_{1,2}(\mathcal{Q}_1, \mathcal{Q}_2) := m_{1,2} - (m_1 + m_2) \geq 0$$

The incompatibility index measures how much the joint optimization degrades compared to optimizing each metric individually. A high value indicates systematic conflict: no attribution result can simultaneously achieve both individual optima. Conversely, $\mathcal{I}_{1,2} = 0$ implies the metrics are perfectly compatible and there exists an attribution that optimizes both criteria simultaneously.

Unlike prior work that reports dataset-level correlations between metrics, our incompatibility index provides an agnostic and sample-wise measure of conflict. This enables instance-specific analysis, revealing when and where tradeoffs are most severe. Moreover, the index is grounded in the intrinsic geometry of the metric spaces rather than empirical observations, providing a principled foundation for understanding systematic incompatibilities.

Table 1: Generalized L^2 representations of common explanation quality criteria.

Quality Criteria	**Faithfulness**	**Robustness**	**Complexity**
Existing Metric	Infidelity (Yeh et al. 2019)	Avg. Sensitivity (Bhatt, Weller, and Moura 2021)	Effective Complexity (Nguyen and Martínez 2020)
Parameter γ_1	I^T	$\mathcal{I}_d$	Diagonal mask (see above)
Parameter γ_2	$f(x) - f(x \odot (1 - I))$	$\phi(x + \varepsilon)$	Threshold vector (see above)

Generalized Eigen-Analysis of Incompatibility

While the incompatibility index quantifies the *degree* of systematic conflict between criteria, it does not reveal *where* in the attribution vector ϕ such conflicts are concentrated. To address this, we introduce a generalized eigen-analysis that localizes the components of an attribution result most responsible for observed tradeoffs.

Our approach builds on the generalized eigenvalue problem (Ghojogh, Karray, and Crowley 2019), a classical tool from linear algebra that characterizes the relative behavior of two quadratic forms:

Definition 3 (Generalized Eigenvalue Problem)**.** *The generalized eigenvalue problem for matrix pair (A_1, A_2) seeks nonzero vectors $v \in \mathbb{R}^d$ and scalars $\lambda \in \mathbb{R}$ satisfying:*

$$A_1 v = \lambda A_2 v$$

We call v a generalized eigenvector *and λ the corresponding* generalized eigenvalue*.*

In our context, each generalized eigenvector v represents a *tradeoff direction* in attribution space. Moving ϕ along direction v scales the rate of change of $\mathcal{Q}_1$ relative to $\mathcal{Q}_2$ by factor λ. Specifically, consider a small perturbation $\phi \to \phi + \epsilon v$. The first-order changes in the two metrics are:

$$\Delta\mathcal{Q}_1 \approx 2\epsilon v^T A_1 \phi$$
$$\Delta\mathcal{Q}_2 \approx 2\epsilon v^T A_2 \phi$$

When v is a generalized eigenvector with eigenvalue λ, we have $v^T A_1 \phi = \lambda v^T A_2 \phi$, meaning the ratio of metric changes is $\Delta\mathcal{Q}_1 / \Delta\mathcal{Q}_2 \approx \lambda$. Large eigenvalues indicate directions where improving $\mathcal{Q}_2$ comes at a steep cost to $\mathcal{Q}_1$ and therefore corresponds precisely to the signature of incompatibility.

Beyond identifying tradeoff directions, generalized eigenvectors enable a complete decomposition of the incompatibility index itself:

Theorem 1 (Decomposition of Incompatibility)**.** *Let $\{(\lambda_k, v_k)\}_{k=1}^r$ be the generalized eigenpairs of (A_1, A_2) with A_2 invertible, where $r = \text{rank}(A_2)$. Define $\beta_{i,k} = v_k^T b_i$ for $i \in \{1, 2\}$. Then the incompatibility index decomposes as:*

$$\mathcal{I}_{1,2}(\mathcal{Q}_1, \mathcal{Q}_2) = \sum_{k=1}^{r} \frac{(\beta_{1,k} - \lambda_k \beta_{2,k})^2}{\lambda_k (1 + \lambda_k)}$$

Proof. By Proposition 2 we have,

$$m_1 + m_2 = c_1 - b_1^\mathsf{T} A_1^\dagger b_1 + c_2 - b_2^\mathsf{T} A_2^\dagger b_2,$$
$$m_{1,2} = (c_1 + c_2) - (b_1 + b_2)^\mathsf{T} (A_1 + A_2)^\dagger (b_1 + b_2),$$

and thus:

$$\mathcal{I}_{1,2} = (b_1 + b_2)^\mathsf{T} (A_1 + A_2)^\dagger (b_1 + b_2) - b_1^\mathsf{T} A_1^\dagger b_1 - b_2^\mathsf{T} A_2^\dagger b_2.$$

Because $A_2 \succ 0$, we can choose the generalized eigenvectors $\{v_k\}$ A_2-orthonormally, i.e. $v_k^\mathsf{T} A_2 v_\ell = \delta_{k\ell}$, with $A_1 v_k = \lambda_k A_2 v_k$. With this normalisation the Moore–Penrose inverses admit the spectral representations

$$A_1^\dagger = \sum_{k=1}^{r} \frac{1}{\lambda_k} v_k v_k^\mathsf{T}, \qquad A_2^\dagger = \sum_{k=1}^{r} v_k v_k^\mathsf{T},$$
$$(A_1 + A_2)^\dagger = \sum_{k=1}^{r} \frac{1}{1 + \lambda_k} v_k v_k^\mathsf{T}.$$

Writing $b_i = \sum_k \beta_{i,k} v_k$ and substituting, we obtain

$$b_1^\mathsf{T} A_1^\dagger b_1 = \sum_{k=1}^{r} \frac{\beta_{1,k}^2}{\lambda_k},$$
$$b_2^\mathsf{T} A_2^\dagger b_2 = \sum_{k=1}^{r} \beta_{2,k}^2,$$
$$(b_1 + b_2)^\mathsf{T} (A_1 + A_2)^\dagger (b_1 + b_2) = \sum_{k=1}^{r} \frac{(\beta_{1,k} + \beta_{2,k})^2}{1 + \lambda_k}.$$

This finally results in:

$$\mathcal{I}_{1,2} = \sum_{k=1}^{r} \Big[\frac{(\beta_{1,k} + \beta_{2,k})^2}{1 + \lambda_k} - \frac{\beta_{1,k}^2}{\lambda_k} - \beta_{2,k}^2 \Big]$$
$$= \sum_{k=1}^{r} \frac{(\beta_{1,k} - \lambda_k \beta_{2,k})^2}{\lambda_k (1 + \lambda_k)}$$

□

This theorem reveals that incompatibility arises *additively* from independent tradeoff directions. We can define the *contribution* of eigenvector v_k to the total incompatibility as:

$$t_k := \frac{(\beta_{1,k} - \lambda_k \beta_{2,k})^2}{\lambda_k (1 + \lambda_k)}$$

Eigenvectors with large t_k values identify the specific components of attribution results where conflicts between criteria are most pronounced. By projecting ϕ onto these high-contribution eigenvectors, practitioners can visualize and interpret which features or attribution patterns are most responsible for observed tradeoffs.

In summary, our generalized eigen-analysis provides both global insight via eigenvalue magnitudes and local diagnostic information through eigenvector projections, complementing the incompatibility index with actionable geometric understanding of where and why quality criteria conflict.

Experiments

To assess the utility of our proposed methods for incompatibility analysis, we evaluate five distinct attribution methods (Input×Gradient (Shrikumar et al. 2016), Integrated Gradients (Sundararajan, Taly, and Yan 2017), SmoothGrad (Smilkov et al. 2017), LIME (Ribeiro, Singh, and Guestrin 2016), LRP (Bach et al. 2015)) based on their default implementation in the `Captum` library (Kokhlikyan et al. 2020). We conduct experiments across multiple datasets: MNIST and three medical image classification datasets from MedMNIST (Yang et al. 2023) (OrganAMNIST, OrganCMNIST, OrganSMNIST). For MNIST, we analyze 1000 randomly selected samples, while for each MedMNIST dataset we consider 500 samples from the corresponding test split. For all datasets, we train a LeNet-style convolutional neural network (LeCun et al. 2002) adapted to the input characteristics of each dataset. We focus on incompatibilities between the popular evaluation criteria Faithfulness vs. Robustness and Faithfulness vs. Complexity.

Validating the Incompatibility Index via Rank Consistency Analysis

To evaluate whether our incompatibility index captures genuine systematic tradeoffs, we investigate whether high incompatibility corresponds to inconsistent metric-based rankings of attribution methods. If two quality criteria are fundamentally incompatible, methods that perform well on one criterion should perform poorly on the other when incompatibility is high, leading to discordant rankings.

For each attribution method ϕ and criterion pair represented by L^2 metrics $\mathcal{Q}_i$ and $\mathcal{Q}_j$, we partition samples based on the magnitude of their incompatibility indices and analyze ranking consistency within each group. More precisely, we (1) compute the incompatibility index I_{ij} for all samples x and partition into low and high incompatibility groups using the median as a threshold; (2) for each group $G \in \{G_{\text{low}}, G_{\text{high}}\}$, aggregate metric values $\bar{Q}_i^G(\phi) = \frac{1}{|G|}\sum_{x\in G} \mathcal{Q}_i(\phi(x))$ and rank all attribution methods by each criterion; (3) compute Spearman correlation $\rho(G)$ between the two criterion-based rankings within each group; (4) test the hypothesis $H_1 : \rho(G_{\text{low}}) > \rho(G_{\text{high}})$ using a one-sided t-test over $B = 500$ bootstrap iterations each resampling subgroups of G with size 200 for MNIST and size 100 for MedMNIST datasets.

High Spearman correlation $\rho(G)$ indicates that methods ranking well on criterion $\mathcal{Q}_i$ also rank well on $\mathcal{Q}_j$, suggesting consistent criteria. Low correlation indicates discordant rankings, revealing fundamental tension. If our incompatibility index is meaningful, we expect significantly higher correlation in low incompatibility samples.

Faithfulness vs. Robustness Table 2 presents the rank consistency analysis for Faithfulness vs. Robustness across all datasets. For MNIST, all five methods support the hypothesis ($p < 0.001$), with correlations decreasing substantially. Notably, Integrated Gradients shows a dramatic shift from positive to negative correlation ($\rho = 0.233$ to $\rho = -0.077$), while Input×Gradient, LIME, and LRP show moderate decreases (from $\rho \approx 0.20$ - 0.26 to $\rho \approx 0.16$ - 0.19). SmoothGrad exhibits the strongest initial correlation but also shows a significant decrease (from $\rho = 0.293$ to $\rho = 0.097$).

Table 2: Rank consistency analysis for Faithfulness vs. Robustness across all datasets. Spearman correlations between criterion-based rankings in low and high incompatibility groups. Significant decreases ($p < 0.05$) indicate that high incompatibility corresponds to inconsistent rankings.

Dataset	Attribution Method	ρ_{low}	ρ_{high}	p-value
MNIST	Input×Gradient	0.255	0.193	<0.001
	Integrated Gradients	0.233	-0.077	<0.001
	SmoothGrad	0.293	0.097	<0.001
	LIME	0.249	0.184	<0.001
	LRP	0.197	0.156	<0.001
OrganA	Input×Gradient	0.090	0.089	0.878
	Integrated Gradients	0.432	0.390	<0.001
	LIME	0.478	-0.018	<0.001
	LRP	0.094	0.097	0.891
	SmoothGrad	0.224	-0.066	<0.001
OrganC	Input×Gradient	0.216	-0.027	<0.001
	Integrated Gradients	0.407	-0.002	<0.001
	LIME	0.371	-0.148	<0.001
	LRP	0.206	-0.016	<0.001
	SmoothGrad	0.060	-0.023	<0.001
OrganS	Input×Gradient	0.232	-0.435	<0.001
	Integrated Gradients	0.130	-0.123	<0.001
	LIME	0.274	-0.348	<0.001
	LRP	0.224	-0.415	<0.001
	SmoothGrad	0.232	-0.105	<0.001

The medical imaging datasets reveal even stronger evidence of systematic conflicts. For OrganAMNIST, three out of five methods show significant correlation decreases, with LIME and SmoothGrad exhibiting particularly dramatic drops (from $\rho = 0.478$ to $\rho = -0.018$ and from $\rho = 0.224$ to $\rho = -0.066$, respectively). OrganCMNIST demonstrates the most consistent pattern, with all five methods showing significant decreases ($p < 0.001$), including several cases where correlations become negative in high incompatibility samples. OrganSMNIST exhibits the most pronounced conflicts, with all methods showing significant decreases and notably strong negative correlations in high incompatibility groups (ranging from $\rho = -0.105$ to $\rho = -0.435$).

These results validate that our incompatibility index captures genuine systematic conflicts between Faithfulness and Robustness across diverse image classification tasks.

Faithfulness vs. Complexity Table 3 presents the rank consistency analysis for Faithfulness vs. Complexity. For MNIST, all five methods show significantly higher rank correlation in low incompatibility samples ($p < 0.001$), with correlations transitioning from weakly positive to negative values. The shifts are particularly pronounced, with Integrated Gradients dropping from $\rho = 0.082$ to $\rho = -0.412$ and Input×Gradient declining from $\rho = 0.071$ to $\rho =$

Table 3: Rank consistency analysis for Faithfulness vs. Complexity across all datasets. All or most methods show significant correlation decreases in high incompatibility samples, confirming systematic conflicts.

Dataset	Attribution Method	ρ_{low}	ρ_{high}	p-value
MNIST	Input×Gradient	0.071	-0.215	<0.001
	Integrated Gradients	0.082	-0.412	<0.001
	SmoothGrad	0.207	-0.040	<0.001
	LIME	0.057	-0.209	<0.001
	LRP	0.030	-0.186	<0.001
OrganA	Input×Gradient	0.178	0.081	<0.001
	Integrated Gradients	0.312	-0.043	<0.001
	LIME	0.095	0.090	0.066
	LRP	0.184	0.067	<0.001
	SmoothGrad	-0.304	-0.009	1.000
OrganC	Input×Gradient	0.180	0.176	0.416
	Integrated Gradients	0.517	0.120	<0.001
	LIME	0.480	-0.084	<0.001
	LRP	0.176	0.153	<0.001
	SmoothGrad	0.187	0.042	<0.001
OrganS	Input×Gradient	0.220	-0.404	<0.001
	Integrated Gradients	0.303	0.133	<0.001
	LIME	0.192	-0.264	<0.001
	LRP	0.220	-0.405	<0.001
	SmoothGrad	0.049	-0.484	<0.001

-0.215. This substantial decrease and sign reversal confirms that high incompatibility corresponds to fundamentally inconsistent criterion rankings.

The medical imaging datasets further corroborate these findings. OrganAMNIST shows three out of five methods with significant decreases, with Integrated Gradients exhibiting a particularly strong drop from $\rho = 0.312$ to $\rho = -0.043$. OrganCMNIST demonstrates four methods with significant decreases, with LIME showing the most dramatic change (from $\rho = 0.480$ to $\rho = -0.084$). OrganSMNIST again exhibits the strongest conflicts, with all five methods showing significant decreases and several methods displaying pronounced negative correlations in high incompatibility samples (e.g., SmoothGrad: from $\rho = 0.049$ to $\rho = -0.484$).

Across all datasets, the consistent pattern of correlation decreases, which often transitions to negative values in high incompatibility groups, provides strong empirical evidence that our incompatibility index reliably identifies systematic conflicts between Faithfulness and Complexity.

Localizing Tradeoffs via Generalized Eigen-Analysis

Beyond quantifying incompatibility severity, our framework enables localization of where conflicts reside within attribution maps through generalized eigen-analysis introduced above. The dominant eigenvector reveals which spatial regions of the attribution are most responsible for the observed tradeoff, providing interpretable insights into the nature of criterion conflicts.

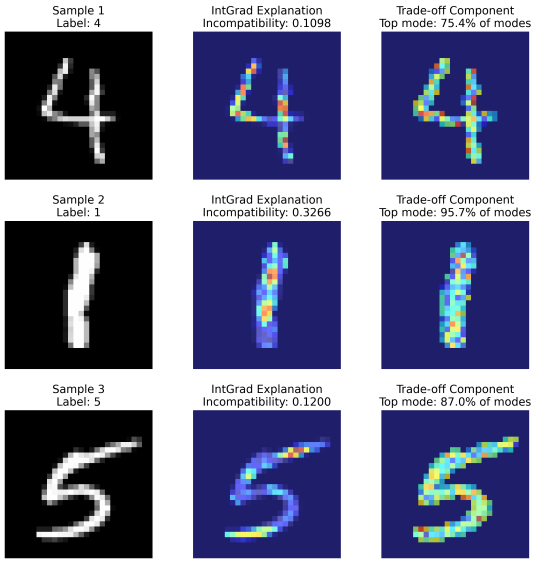

Figure 2: Eigen-analysis for Faithfulness vs. Robustness using Integrated Gradients on three randomly selected samples. Each row shows: input image, attribution map, dominant eigen-mode (conflict localization), and percentage contribution. The tradeoff is distributed across the entire digit structure, indicating that the conflict arises from global sensitivity rather than localized regions.

Figures 2 and 3 show eigen-analysis results for three randomly selected samples using Integrated Gradients attributions. Each row displays the input image, the attribution map, the dominant eigen-mode, which localizes the conflict, and its percentage contribution to the incompatibility.

For Faithfulness vs. Robustness (Figure 2), the top eigenmodes are distributed across the entire digit structure and express already $> 75\%$ of the total incompatibility index. The conflict is not localized to specific regions but rather spread throughout areas where attribution mass is concentrated. This indicates that the robustness-faithfulness tension arises from the global sensitivity of the attribution to input perturbations—improving faithfulness requires stable attributions across the digit, but this inherently conflicts with robustness to local variations. In contrast, for Faithfulness vs. Complexity (Figure 3), the eigen-modes are predominantly localized in regions *outside* the main focus areas of the attribution maps, with contributions that are strongly localized in the top eigen mode, accounting for almost the entire incompatibility index. This directly captures the fundamental sparsity-faithfulness tradeoff: to increase faithfulness, attribution mass must be distributed more broadly beyond the core salient regions, but this naturally reduces sparsity. The eigen-analysis reveals that the conflict resides precisely in these peripheral regions where faithfulness demands nonzero attribution but complexity favors suppression.

These examples demonstrate that eigen-analysis provides actionable insights into the spatial structure of criterion conflicts, enabling practitioners to understand not just *that* a tradeoff exists, but *where* it manifests in their explanations.

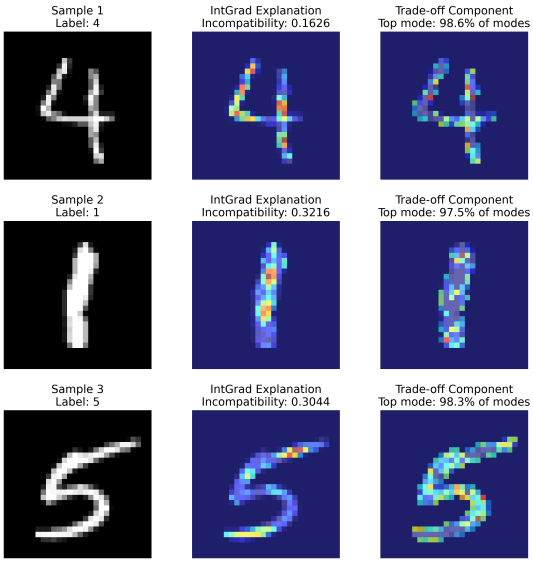

Figure 3: Eigen-analysis for Faithfulness vs. Complexity using Integrated Gradients on three randomly selected samples. The eigen-modes are predominantly localized in regions *outside* the main attribution focus areas, directly capturing the sparsity-faithfulness tradeoff: improving faithfulness requires distributing attribution mass to peripheral regions, which inherently reduces sparsity.

Conclusion

We introduced a principled mathematical framework for analyzing systematic incompatibilities between explanation quality criteria. By unifying faithfulness, robustness, and complexity as generalized L^2 metrics, we developed a sample-wise incompatibility index and a generalized eigen-analysis that localizes tradeoffs within attribution results. Our experiments on image classifiers demonstrate that this analysis provides actionable insights beyond isolated metric evaluations, enabling more informed decisions when selecting attribution methods.

Our framework focuses on pairwise incompatibilities; extending to multi-way conflicts and applying the approach to other domains are natural next steps. Additionally, while our eigen-analysis identifies *where* conflicts occur, developing principled strategies for resolving tradeoffs based on application-specific requirements remains an important direction for future work.

References

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2019. Gradient-based attribution methods. In *Explainable AI: Interpreting, explaining and visualizing deep learning*, 169–191. Springer.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140.

Bhatt, U.; Weller, A.; and Moura, J. M. 2021. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 3016–3022.

Bilodeau, B.; Jaques, N.; Koh, P. W.; and Kim, B. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2): e2304406120.

Chen, H.; Covert, I. C.; Lundberg, S. M.; and Lee, S.-I. 2023. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, 5(6): 590–601.

Covert, I.; Lundberg, S.; and Lee, S.-I. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209): 1–90.

Decker, T.; Bhattarai, A. R.; Gu, J.; Tresp, V.; and Buettner, F. 2024. Provably Better Explanations with Optimized Aggregation of Feature Attributions. In *International Conference on Machine Learning*, 10267–10286. PMLR.

Dicker, L. 2014. Sparsity and the Truncated l2-norm. In *Artificial Intelligence and Statistics*, 159–166. PMLR.

Fokkema, H.; de Heide, R.; and van Erven, T. 2023. Attribution-based explanations that provide recourse cannot be robust. *Journal of Machine Learning Research*, 24(360): 1–37.

Ghojogh, B.; Karray, F.; and Crowley, M. 2019. Eigenvalue and generalized eigenvalue problems: Tutorial. *arXiv preprint arXiv:1903.11240*.

Ghorbani, A.; Abid, A.; and Zou, J. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3681–3688.

Hedström, A.; Bommer, P.; Wickstrøm, K. K.; Samek, W.; Lapuschkin, S.; and Höhne, M. M.-C. 2023a. The meta-evaluation problem in explainable AI: identifying reliable estimators with MetaQuantus. *arXiv preprint arXiv:2302.07265*.

Hedström, A.; Weber, L.; Krakowczyk, D.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; and Höhne, M. M.-C. 2023b. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34): 1–11.

Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Krishna, S.; Han, T.; Gu, A.; Pombra, J.; Jabbari, S.; Wu, S.; and Lakkaraju, H. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Le, P. Q.; Nauta, M.; Van Bach Nguyen, S. P.; Schlötterer, J.; and Seifert, C. 2023. Benchmarking eXplainable AI-A Survey on Available Toolkits and Open Challenges. In *International Joint Conference on Artificial Intelligence*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Mehrpanah, A.; Gamba, M.; Smith, K.; and Azizpour, H. 2025. On the Complexity-Faithfulness Trade-off of Gradient-Based Explanations. *arXiv preprint arXiv:2508.10490*.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s): 1–42.

Nguyen, A.-p.; and Martínez, M. R. 2020. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, 3319–3328. PMLR.

Tadesse, H. B.; Hüyük, A.; Yacoby, Y.; Pan, W.; and Doshi-Velez, F. 2025. Transparent Trade-offs between Properties of Explanations. In *The 41st Conference on Uncertainty in Artificial Intelligence*.

Tan, Z.; and Tian, Y. 2023. Robust explanation for free or at the cost of faithfulness. In *International conference on machine learning*, 33534–33562. PMLR.

Tomsett, R.; Harborne, D.; Chakraborty, S.; Gurram, P.; and Preece, A. 2020. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6021–6029.

Wang, Y.; Zhang, T.; Guo, X.; and Shen, Z. 2024. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*.

Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; and Ni, B. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1): 41.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32.

Zednik, C.; and Boelsen, H. 2022. Scientific exploration and explainable artificial intelligence. *Minds and Machines*, 32: 219–239.

Zhang, Y.; He, H.; Tan, Z.; and Yuan, Y. 2023. Trade-off between efficiency and consistency for removal-based explanations. *Advances in Neural Information Processing Systems*, 36: 25627–25661.