Higher-order Component Attribution via Kolmogorov–Arnold Networks

Samy Mammeri
Université Laval
Institute Intelligence and Data (IID)
Mila – Quebec AI Institute
samy.mammeri.1@ulaval.ca

Christian Gagné
Université Laval
Institute Intelligence and Data (IID)
Mila – Quebec AI Institute
Canada CIFAR AI Chair
christian.gagne@gel.ulaval.ca

Abstract

Component attribution quantifies how model components, from individual neurons to transformer blocks, contribute to a prediction. Despite their successes, most methods assume additive linear effects between components and overlook interactions that shape how predictions arise from internal computations. In this work, we formalize nonlinear component modeling and introduce a Kolmogorov–Arnold Network (KAN)-based framework for component attribution. We fit KAN surrogates on perturbation–response data to represent effects nonlinearly, then use them to extract local component interaction coefficients in two complementary ways: by automatic differentiation of the trained KAN and by recovering a symbolic surrogate whose closed-form mixed partial derivatives yield symbolic interaction scores. This provides a way to relate a classifier's output back to interacting internal building blocks instead of isolated components. The resulting expressions are intended for future integration with formal verification methods to support richer counterfactual analyses. Preliminary results on standard image classification models demonstrate that our approach improves the accuracy of counterfactual predictions and enables extraction of higher-order component interactions compared to linear attribution.

1 Introduction

Advances in deep learning continue to deliver performance gains across image classification, language modeling, and audio processing [1–3]. As architectures scale, attributing a given prediction to specific components such as transformer blocks, residual blocks, and convolutional layers becomes increasingly difficult [4, 5]. Attribution methods address this by quantifying output changes under component ablation [6].

Existing approaches, including Component Attribution via Regression (COAR) [5], typically adopt a main-effects-only linear model. While efficient, this design necessarily omits interactions among components that are key to explaining how predictions arise from internal computation.

To overcome this limitation, we introduce a nonlinear component attribution framework based on Kolmogorov–Arnold Networks (KANs) [7, 8]. Extending the formulation of Shah et al. [5], we replace the linear counterfactual model with a flexible KAN surrogate capable of capturing nonlinear dependencies among components. We generate perturbation–response datasets through randomized, continuous multicomponent interventions and train the KAN to approximate these responses. From the trained surrogate, interaction coefficients can be extracted directly, or symbolic regression can be applied to recover closed-form expressions of importance scores up to a specified derivative order. Figure 1 illustrates the overall approach.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

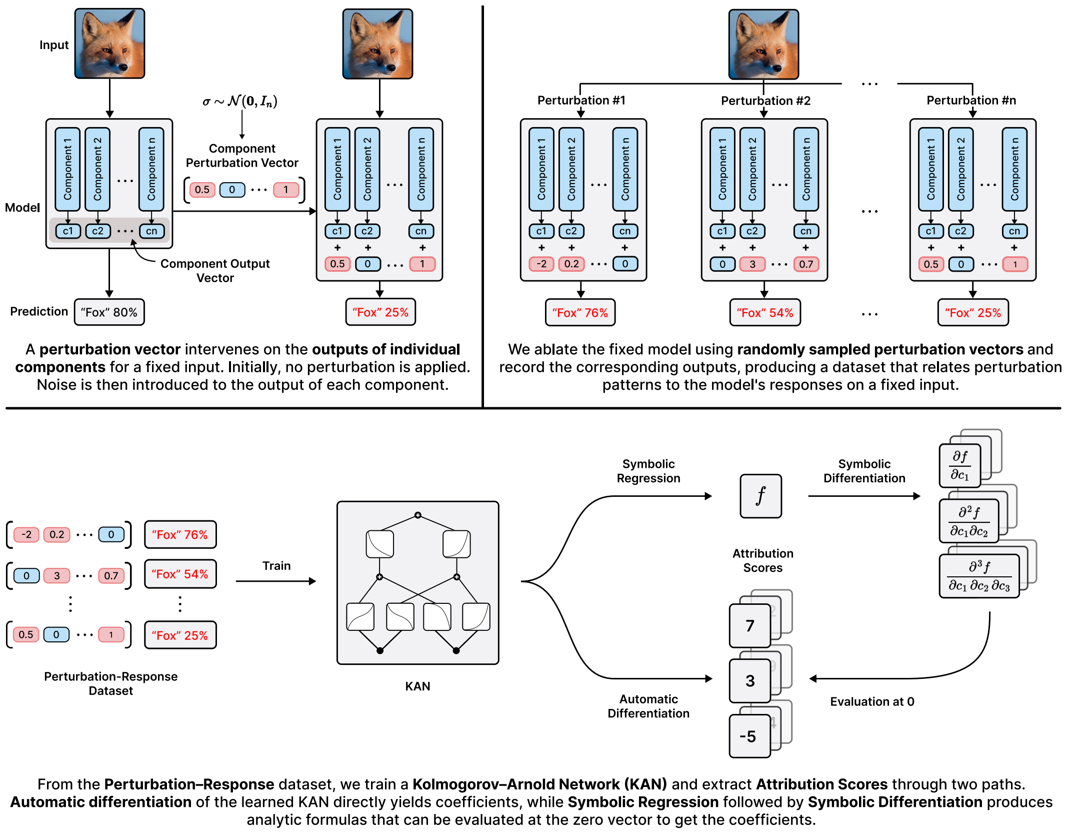

Figure 1: Overview of the proposed approach.

Our main contributions are:

1. Formalizing higher-order component attribution within the established component modeling framework [5].
2. Introducing KANs as nonlinear component models.
3. Employing two complementary paths to extract attribution scores from the learned KAN:
   (a) Automatic differentiation of the trained KAN yields numerical interaction coefficients.
   (b) Symbolic regression fits a closed-form surrogate whose symbolic differentiation provides analytic expressions.
4. Providing empirical validation that nonlinear modeling significantly enhances counterfactual prediction accuracy compared to linear approaches.

2 Motivation

2.1 Nonlinear Modeling

Attribution is credit assignment: measuring how model components, input features, and training examples shape predictions [6]. Despite widespread use, component attribution is commonly implemented with linear surrogates, which capture only main effects and miss higher-order interactions among components [4, 9].

A unifying view is that attribution methods perform local function approximation: they fit a simpler surrogate to a complex model f around an anchor x [5, 6, 10]:

$$\hat{g} \;=\; \arg\min_{g \in \mathcal{G}} \; \mathbb{E}_{\xi \sim Z}\, \ell(f, g, x, \xi)\,, \tag{1}$$

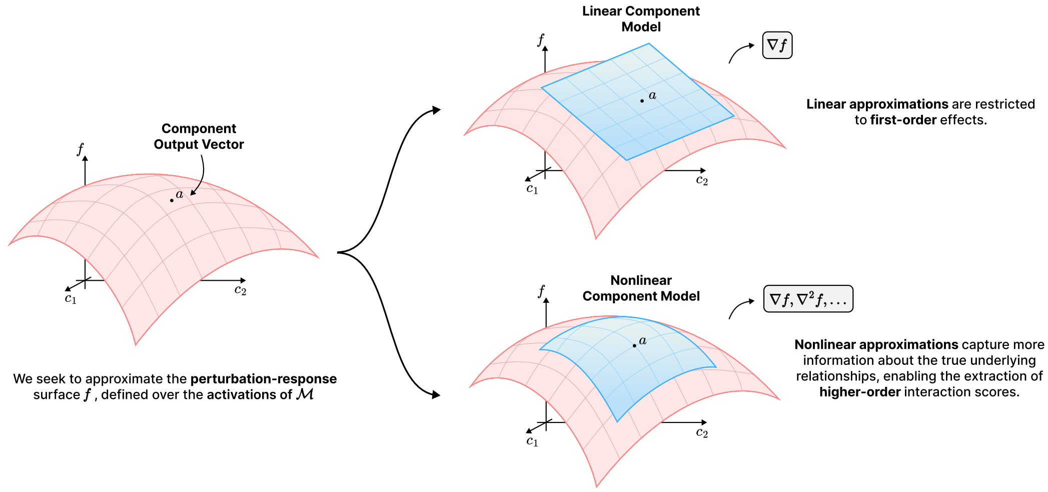

Figure 2: Capturing nonlinear interactions through more expressive surrogate models.

where $\mathcal{G}$ is the surrogate class, Z a local perturbation distribution, and ℓ a local loss. Component and feature attribution differ primarily in how they interpret the surrogate's input dimensions [6]; insights therefore transfer across the two settings.

This transfer is well illustrated by LIME-style feature attribution [11–13], which—under suitable conditions—converges to gradient approximations of f at the anchor [13–15]. We leverage the same intuition for components: if first-order scores arise as local gradients, then higher-order component scores should arise as local mixed partials.

Assume inputs $x \in \mathbb{R}^n$. Extending the local-proxy view, we define the order-r interaction coefficient for index set S (with $|S| = r$) at reference point a by the mixed partials of the learned surrogate $\hat{g}$:

$$I_S^{(r)}(a) \;:=\; \left. \frac{\partial^r \hat{g}(x)}{\prod_{j \in S} \partial x_j} \right|_{x=a}, \quad r \leq n. \tag{2}$$

These coefficients serve as local approximations to the corresponding mixed partials of f at a as illustrated in Figure 2.

To extract such partials reliably—i.e., to capture interactions beyond first-order effects—the surrogate should satisfy:

1. **Expressiveness:** universal approximation of the relevant local relationships.
2. **Differentiability:** k-times continuously differentiable in a neighborhood of a.

2.2 KANs as Nonlinear Surrogates

Linear models [16] use an intercept plus a weighted sum of the inputs. Generalized Additive Models (GAMs) [17] relax each term to a smooth univariate function and allow a link function; with the identity link and linear univariate effects, a GAM reduces to a linear model. The Kolmogorov–Arnold representation theorem [18, 19] goes further: after an affine rescaling of the domain, any continuous multivariate function can be expressed as a finite sum of univariate outer functions of sums of univariate functions of the individual coordinates.

Kolmogorov–Arnold Networks (KANs) [7, 8] instantiate this superposition by using learnable univariate B-spline maps on edges with cross-variable mixing, thereby strictly generalizing GAMs and linear models. With degree-$k+1$ B-splines [20], each univariate map is C^k and the resulting network is C^k in its inputs. This ensures that all mixed partial derivatives up to order $r \leq k$ exist and are well behaved.

3 Higher-order Component Attribution via KANs

Setup We build upon the component modeling framework introduced by Shah et al. [5]. Given a trained model M composed of m components $\mathcal{C} = \{c_1, \ldots, c_m\}$ and a fixed input $\boldsymbol{z}$, we define the scalar output function $f_M(\boldsymbol{z}, \boldsymbol{\sigma})$ as the model's prediction when applying an additive gating mask $\boldsymbol{\sigma} \sim \mathcal{N}_m(\mathbf{0}, \tau^2 \mathbf{I}_m)$ to its components, where $\tau > 0$. Under this definition, the mask $\mathbf{0}$ represents the unperturbed model, while any deviation from $\mathbf{0}$ corresponds to a scaled perturbation of component outputs. We formalize this notion through the centered counterfactual response:

$$\Delta f_M(\boldsymbol{z}, \boldsymbol{\sigma}) \; = \; f_M(\boldsymbol{z}, \boldsymbol{\sigma}) - f_M(\boldsymbol{z}, \mathbf{0}), \tag{3}$$

ensuring $\Delta f_M(\boldsymbol{z}, \mathbf{0}) = 0$ by construction.

Perturbation-response Dataset To build a perturbation–response dataset D for each input $\mathbf{z}$, we draw N random masks $\boldsymbol{\sigma}^{(i)}$ from a multivariate normal distribution $\mathcal{N}_m(\mathbf{0}, \tau^2 \mathbf{I}_m)$ (Algorithm 1). For every mask, we record the centered output $y^{(i)} = \Delta f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)})$. The resulting dataset captures local nonlinear dependencies among components in the neighborhood of the intact configuration.

KAN as a component model We train a per-example KAN component model $g_{\boldsymbol{z}} : \mathbb{R}^m \to \mathbb{R}$ using the dataset D. The component model aims to approximate the nonlinear mapping from component perturbations to changes in model outputs by minimizing:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left(g_{\boldsymbol{z}}(\boldsymbol{\sigma}^{(i)}; \theta) - y^{(i)} \right)^2 \tag{4}$$

Symbolic Approximation and Interaction Scores After training, we symbolically approximate each univariate spline function within the trained KAN component model g_z via symbolic regression [21, 22], yielding a closed-form symbolic component model $\hat{g}_{\mathbf{z}}$ (Algorithm 2).

To quantify interactions of arbitrary order, we compute mixed partial derivatives of $\hat{g}_{\mathbf{z}}$ at the intact configuration $\boldsymbol{\sigma} = \mathbf{0}$. For any subset of component indices $S \subseteq \{1, \ldots, m\}$ of size $r = |S|$, we define the local r-way interaction coefficient as:

$$L_S^{(r)}(\mathbf{z}) = \left. \frac{\partial^r \hat{g}_{\boldsymbol{z}}(\boldsymbol{\sigma})}{\prod_{j \in S} \partial \sigma_j} \right|_{\boldsymbol{\sigma} = \mathbf{0}}. \tag{5}$$

Algorithm 3 systematically computes these coefficients up to a specified order k, providing hierarchical insights into component effects: first-order terms quantify independent effects, while higher-order terms reveal complex joint interactions. When nonlinear interactions are negligible, our method naturally reduces to standard linear component attribution [5].

3.1 Experimental Setup

We follow the experimental framework proposed by Shah et al. [5], adapted to our specific level of granularity. Rather than examining individual neurons, we focus on residual blocks in ResNet architectures [23] and encoder layers in Vision Transformers (ViTs) [24] as individual components. This granularity is selected due to the intrinsic limitations of Kolmogorov–Arnold Networks (KANs), which are currently unable to efficiently handle high-dimensional inputs, thereby constituting a limitation of our approach. We assess our methodology on three widely used image classification setups: ResNet-18 trained on CIFAR-10 [25], ResNet-50 trained on ImageNet [26], and ViT-B/16 also trained on ImageNet. For each model-dataset combination, we generate localized perturbation-response datasets by sampling multicomponent perturbations, following Algorithm 1.

Baselines and Evaluation Metrics We compare against a single baseline, COAR (linear attribution) [5], which fits a linear model to the perturbation-response pairs $(\sigma, \Delta f)$.

In line with previous research [5], we quantify attribution accuracy using two metrics: the Pearson correlation coefficient and the mean squared error, computed between predicted and actual responses on held-out perturbations.

Table 1: Comparison of attribution accuracy between predicted and observed counterfactual responses

Method	ResNet-18 / CIFAR-10		ResNet-50 / ImageNet		ViT-B/16 / ImageNet	
	Pearson ↑	MSE ↓	Pearson ↑	MSE ↓	Pearson ↑	MSE ↓
COAR (Linear)	0.62 ± 0.05	0.82 ± 0.07	0.54 ± 0.06	0.95 ± 0.08	0.48 ± 0.07	1.05 ± 0.09
KAN (Ours)	$\mathbf{0.76 \pm 0.03}$	$\mathbf{0.53 \pm 0.05}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.60 \pm 0.06}$	$\mathbf{0.66 \pm 0.05}$	$\mathbf{0.68 \pm 0.07}$

Table 2: Recovering derivatives on random 3-variable symbolic functions with KANs

Order	Automatic Differentiation		Symbolic Differentiation	
	Pearson ↑	MAE ↓	Pearson ↑	MAE ↓
First	0.972 ± 0.010	0.0016 ± 0.0006	0.958 ± 0.012	0.0026 ± 0.0009
Second	0.921 ± 0.026	0.0060 ± 0.0022	0.898 ± 0.030	0.0084 ± 0.0029
Third	0.861 ± 0.045	0.0128 ± 0.0047	0.826 ± 0.055	0.0175 ± 0.0062

Our experimental results, summarized in Table 1, show that the proposed KAN-based approach consistently outperforms the linear COAR baseline across all evaluated tasks. Specifically, KAN achieves higher Pearson correlations and lower mean squared errors, indicating a clear advantage in modeling nonlinearity within the component attribution framework.

4 Recovering Interaction Coefficients

We examine whether KAN surrogates recover mixed partial derivatives up to third order on random symbolic targets, following the view that attribution is local function approximation.

Protocol We sample 1000 scalar expressions $f : \mathbb{R}^3 \to \mathbb{R}$ using ramped half–and–half [21, 27]. For each f, we draw three independent anchor points $p_1, p_2, p_3 \in \mathbb{R}^3$. Around each anchor we apply small input–level perturbations $\boldsymbol{\sigma} \sim \mathcal{N}_3(\mathbf{0}, \mathbf{I}_3)$ and record local responses

$$\Delta f(p, \sigma) = f(p + \sigma) - f(p).$$

For every (f, p) we fit a local KAN surrogate $g_{f,p}$ to the pairs $(\sigma, \Delta f(p, \sigma))$ and obtain a symbolic approximation $\hat{g}_{f,p}$.

Derivative retrieval From $\hat{g}_{f,p}$ we extract mixed partials at $x = \mathbf{0}$ up to order three,

$$\{\hat{f}_x, \hat{f}_y, \hat{f}_z, \hat{f}_{xy}, \hat{f}_{xz}, \hat{f}_{yz}, \hat{f}_{xyz}\}(p) \ := \ \left\{ \frac{\partial \hat{g}}{\partial x}, \frac{\partial \hat{g}}{\partial y}, \dots, \frac{\partial^3 \hat{g}}{\partial x \partial y \partial z} \right\}\bigg|_{x=\mathbf{0}}.$$

Ground truth and comparison For the same (f, p), we compute the corresponding derivatives of f at $x = p$ using two independent mechanisms: (i) automatic differentiation, and (ii) exact symbolic differentiation. We then compare surrogate–derived estimates to these references for each order $r \in \{1, 2, 3\}$.

Reporting We pool all functions and anchors and report, by derivative order, the Pearson correlation and mean absolute error between estimated and reference values. Results are reported in Table 2.

5 Conclusions

In this work, we presented a nonlinear component attribution framework built upon Kolmogorov–Arnold Networks (KANs), designed to explicitly model higher-order interactions that conventional linear attribution methods fail to capture. Empirically, KAN-based component models achieved more accurate counterfactual predictions than linear baselines, highlighting their capacity to represent complex dependencies. Nonetheless, the approach remains constrained by the intrinsic dimensionality limits of KANs, which hinder scalability to high-dimensional inputs. Future efforts should aim to relax these constraints to extend the method's reach.

6 Acknowledgements

This research was supported by the Canadian Institute for Advanced Research (CIFAR) and IVADO. Computational resources were provided by the Digital Research Alliance of Canada. The authors thank Olivier Bussière, Hugo Chapdelaine, Audrey Durand, Frédéric Fortier-Chouinard, Alexandre Larouche, Benjamin Leblanc, Benjamin Léger, Jonas Ngnawe, Yohan Poirier-Ginter, Charles Renaud, Sabyasachi Sahoo, David Serrano Lozano, Cem Subakan, and Adam Tupper for their helpful discussions, feedback, and technical insights that contributed to this work.

References

[1] Jie Huang and Kevin Chen-Chuan Chang. Towards Reasoning in Large Language Models: A Survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL `https://aclanthology.org/2023.findings-acl.67/`.

[2] Yutong Zhou and Nobutaka Shimada. Vision + Language Applications: A Survey. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 826–842, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0249-3. doi: 10.1109/CVPRW59228.2023.00090. URL `https://ieeexplore.ieee.org/document/10208464/`.

[3] Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, and Björn W. Schuller. Sparks of Large Audio Models: A Survey and Outlook, September 2023. URL `http://arxiv.org/abs/2308.12792`. arXiv:2308.12792 [cs].

[4] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J. Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open Problems in Mechanistic Interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=91H76m9Z94`.

[5] Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and Editing Predictions by Modeling Model Computation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 44244–44292. PMLR, July 2024. URL `https://proceedings.mlr.press/v235/shah24a.html`. ISSN: 2640-3498.

[6] Shichang Zhang, Tessa Han, Usha Bhalla, and Himabindu Lakkaraju. Towards Unified Attribution in Explainable AI, Data-Centric AI, and Mechanistic Interpretability, May 2025. URL `http://arxiv.org/abs/2501.18887`. arXiv:2501.18887 [cs].

[7] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Hou, and Max Tegmark. KAN: Kolmogorov–Arnold Networks. *International Conference on Representation Learning*, 2025:70367–70413, May 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/hash/afaed89642ea100935e39d39a4da602c-Abstract-Conference.html`.

[8] Ziming Liu, Pingchuan Ma, Yixuan Wang, Wojciech Matusik, and Max Tegmark. KAN 2.0: Kolmogorov-Arnold Networks Meet Science, August 2024. URL `http://arxiv.org/abs/2408.10205`. arXiv:2408.10205 [cs].

[9] Stefan Haufe, Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Danny Panknin, and Ahcène Boubekki. Explainable AI needs formal notions of explanation correctness, November 2024. URL `http://arxiv.org/abs/2409.14590`. arXiv:2409.14590 [cs].

[10] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations. *Advances in Neural Information Processing Systems*, 35:5256–5268, December 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/22b111819c74453837899689166c4cf9-Abstract-Conference.html.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, pages 1135–1144, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL https://dl.acm.org/doi/10.1145/2939672.2939778.

[12] Zhengze Zhou, Giles Hooker, and Fei Wang. S-LIME: Stabilized-LIME for Model Explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2429–2438, August 2021. doi: 10.1145/3447548.3467274. URL http://arxiv.org/abs/2106.07875. arXiv:2106.07875 [stat].

[13] Zeren Tan, Yang Tian, and Jian Li. GLIME: General, Stable and Local LIME Explanation. *Advances in Neural Information Processing Systems*, 36:36250–36277, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/71ed042903ed67c7f6355e5dd0539eec-Abstract-Conference.html.

[14] Damien Garreau and Ulrike Luxburg. Explaining the Explainer: A First Theoretical Analysis of LIME. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1287–1296. PMLR, June 2020. URL https://proceedings.mlr.press/v108/garreau20a.html. ISSN: 2640-3498.

[15] Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Zhiwei Steven Wu, and Himabindu Lakkaraju. Towards the Unification and Robustness of Perturbation and Gradient Based Explanations, July 2021. URL http://arxiv.org/abs/2102.10618. arXiv:2102.10618 [cs].

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Linear Methods for Regression. In *The Elements of Statistical Learning*, pages 43–99. Springer New York, New York, NY, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7_3. URL http://link.springer.com/10.1007/978-0-387-84858-7_3.

[17] Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1 (3):297–310, August 1986. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1177013604. URL https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Generalized-Additive-Models/10.1214/ss/1177013604.full. Publisher: Institute of Mathematical Statistics.

[18] On the representation of functions of several variables as a superposition of functions of a smaller number of variables. In Alexander B. Givental, Boris A. Khesin, Jerrold E. Marsden, Alexander N. Varchenko, Victor A. Vassiliev, Oleg Ya. Viro, and Vladimir M. Zakalyukin, editors, *Collected Works*, pages 25–46. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-01741-4 978-3-642-01742-1. doi: 10.1007/978-3-642-01742-1_5. URL http://link.springer.com/10.1007/978-3-642-01742-1_5.

[19] Jürgen Braun and Michael Griebel. On a Constructive Proof of Kolmogorov's Superposition Theorem. *Constructive Approximation*, 30(3):653–675, December 2009. ISSN 1432-0940. doi: 10.1007/s00365-009-9054-2. URL https://doi.org/10.1007/s00365-009-9054-2.

[20] Carl De Boor. *A Practical Guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Springer New York, New York, NY, 1978. ISBN 978-0-387-90356-9 978-1-4612-6333-3. doi: 10.1007/978-1-4612-6333-3. URL http://link.springer.com/10.1007/978-1-4612-6333-3.

[21] Riccardo Poli, William B. Langdon, and Nicholas F. McPhee. *A Field Guide to Genetic Programming*. Lulu Enterprises, 2008.

[22] Miles Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, May 2023. URL http://arxiv.org/abs/2305.01582. arXiv:2305.01582 [astro-ph].

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[25] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, Department of Computer Science, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

[27] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research*, 13(70):2171–2175, 2012. ISSN 1533-7928. URL http://jmlr.org/papers/v13/fortin12a.html.

Appendix

A Algorithms

Algorithm 1 Perturbation–Response Dataset

procedure GENERATEDATASET(example $\mathbf{z}$, model M with components $\mathcal{C}$ (size m), sample size N, perturbation scale τ)
  $D \leftarrow []$ ▷ init dataset
  for $i \in \{1, \ldots, N\}$ **do** ▷ N samples
    Sample $\boldsymbol{\sigma}^{(i)} \sim \mathcal{N}_m(\mathbf{0}, \tau^2 \mathbf{I}_m)$ ▷ multicomponent perturbation
    $\Delta f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)}) = f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)}) - f_M(\mathbf{z}, \mathbf{0})$
    $y^{(i)} \leftarrow \Delta f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)})$ ▷ model output
    $D \leftarrow D + [(\boldsymbol{\sigma}^{(i)}, y^{(i)})]$ ▷ append pair
  end for
  return D ▷ dataset for surrogate
end procedure

Algorithm 2 Symbolic KAN Surrogate

procedure SYMBOLICSURROGATE(dataset D)
  Fit KAN $g_{\mathbf{z}}$ on D ▷ train surrogate
  $\mathcal{S} \leftarrow []$ ▷ init symbol list
  for edge e in $g_{\mathbf{z}}$ **do** ▷ per-edge univariate
    $\hat{\phi}_e \leftarrow \mathrm{SymbolicRegression}(\phi_e)$ ▷ closed-form fit
    $\mathcal{S} \leftarrow \mathcal{S} + [(e, \hat{\phi}_e)]$ ▷ collect
  end for
  Replace $\phi_e \leftarrow \hat{\phi}_e$ in $g_{\mathbf{z}}$ ▷ compose symbolic surrogate
  return $\hat{g}_{\mathbf{z}}, \mathcal{S}$ ▷ outputs
end procedure

Algorithm 3 Local Interaction Coefficients

procedure LOCALINTERACTIONS(symbolic surrogate $\hat{g}_{\mathbf{z}}$, components $\mathcal{C}$ (size m), max order k)
  for $r \in \{1, \ldots, k\}$ **do** ▷ interaction order
    for index subset $S \subseteq \{1, \ldots, m\}$ with $|S| = r$ **do** ▷ choose indices
      $L_S^{(r)} \leftarrow \left.\dfrac{\partial^r \hat{g}_{\mathbf{z}}(\boldsymbol{\sigma})}{\prod_{j \in S} \partial \sigma_j}\right|_{\boldsymbol{\sigma}=\mathbf{0}}$ ▷ r-way interaction at baseline
    end for
  end for
  return $\{L^{(r)}\}_{r=1}^{k}$ ▷ local interaction coefficients
end procedure