

Kolmogorov-Arnold Causal Generative Models

Alejandro Almodóvar , Mar Elizo , Patricia A. Apellániz , Santiago Zazo , and Juan Parras 

Abstract—Causal generative models provide a principled framework for answering observational, interventional, and counterfactual queries from observational data. However, many deep causal models rely on expressive architectures with opaque mechanisms, limiting auditability in high-stakes domains. This paper introduces the Kolmogorov-Arnold causal generative model, a mixed-type tabular causal generative model that parameterizes each structural equation with a Kolmogorov-Arnold Network. The resulting mechanisms can be inspected through symbolic approximations and parent-child visualizations while preserving query-agnostic causal generative semantics. The model includes a validation pipeline based on distributional matching and independence diagnostics of inferred exogenous variables, allowing adequacy to be assessed using observational data alone. Experimentally, on eleven synthetic mixed-type benchmarks, the Kolmogorov-Arnold causal generative model is not statistically distinguishable from the best state-of-the-art model: aggregate post-hoc p -value of 0.69. On the semi-synthetic Sachs’ additive dataset, with 11 variables, the model obtains a p -value of 0.18 against the best model. A real-world cardiovascular case study demonstrates the interpretability workflow, showing how learned mechanisms can be simplified into structural equations and visualized to audit parent-child effects across the causal system. This suggest that expressive causal generative modeling and functional transparency can be achieved jointly, supporting trustworthy deployment in tabular decision-making settings.

Index Terms—Causal inference, Kolmogorov-Arnold networks, interpretability, generative models, deep learning.

I. INTRODUCTION

MACHINE learning systems are increasingly deployed in high-stakes domains such as personalized medicine [1], [2], public policy [3], and economics [4], where decisions must be justified, reproducible, and robust to distributional shifts. In these settings, estimating individualized treatment effects and answering counterfactual questions from observational tabular data is central to decision-making. Structural causal models (SCMs) provide a formal framework for such reasoning, allowing the computation of observational, interventional, and

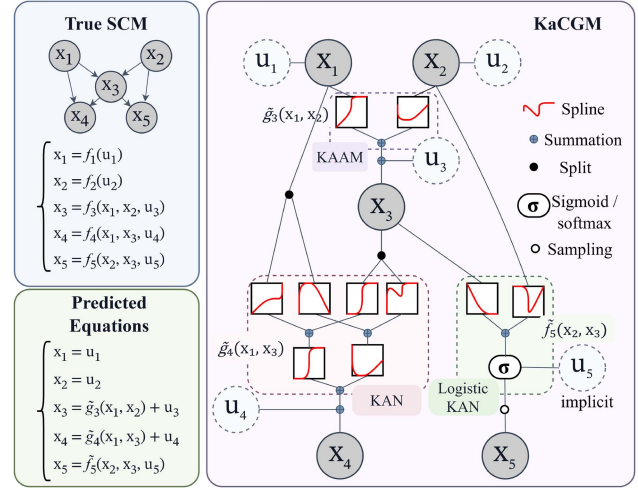


Fig. 1. Compacted sketch of the proposed causal generative model, KaCGM, including a discrete variable (x_5). The true SCM, with unknown structural equations (left), is approximated with additive noise models in which the functions of each endogenous variable are modeled by KANs (right). Several KANs are shown, in which interpretability capabilities decrease with network complexity. As explained in Section IV, u_5 is modeled implicitly by a logistic-KAN. Closed-form expressions can be extracted from every approximated function.

counterfactual quantities under explicit assumptions [5], [6]. Recovering a reliable approximation of the underlying SCM is therefore essential for principled policy targeting and individualized care [7].

However, predictive accuracy alone is not sufficient for deployment in sensitive applications. Regulatory frameworks such as the General Data Protection Regulation (GDPR Art. 22) [8], the European Union Artificial Intelligence Act [9], and guidance from the U.S. Food and Drug Administration on AI-enabled medical devices [10] emphasize transparency and accountability requirements for high-risk AI systems. The “right to explanation” and related transparency obligations further reinforce the need for models whose internal logic can be inspected and communicated [11]. In clinical practice, limited interpretability remains a major barrier to the adoption of machine learning systems [12], [13]. As argued by Rudin [14], black-box models are particularly problematic when decisions have substantial real-world consequences and interpretable alternatives are feasible. In the causal setting, interpretability is even more critical: practitioners must understand how interventions propagate through mechanisms, not merely observe predictive correlations.

Recent causal generative models instantiate SCM mechanisms with different deep generative families to answer causal queries [15], [16], [17], [18]. These models show that query-agnostic causal simulation can be combined with flexible den-

Received 23 March 2026; revised 26 May 2026; accepted 3 August 2026. Date of publication 7 August 2026; date of current version 5 October 2026. This work was supported by SYNTHIA project SYNTHIA (Synthetic Data Generation framework for integrated validation of use cases and AI healthcare applications) under Innovative Health Initiative Joint Undertaking (IHI JU) under Grant Agreement 101172872, funded by the European Union, the private members, and those contributing partners of the IHI JU. Recommended for acceptance by R. Cai. (Corresponding author: Alejandro Almodóvar.)

The authors are with Information Processing and Telecommunications Center, ETSI de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: alejandro.almodovar@upm.es).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TPAMI.2026.3721159>, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2026.3721159

sity estimation. However, their learned equations are typically implemented by black-box neural modules. Thus, although the causal graph provides structural interpretability, each parent-child mechanism remains difficult to audit, simplify, or communicate.

This work addresses the gap between expressive causal generative modeling and functional interpretability in tabular settings. This paper proposes the *Kolmogorov-Arnold causal generative model* (KaCGM), a causal generative model in which each structural mechanism is parameterized by a Kolmogorov-Arnold network (KAN). KANs replace scalar weights with learnable univariate functions, inducing structured representations that facilitate inspection, pruning, and symbolic approximation. By embedding KANs within an additive noise SCM framework, our approach preserves generative capabilities and query-agnostic inference, while supporting the extraction of explicit structural equations.

A compact overview of the proposed model is shown in Fig. 1. Each endogenous variable is modeled as an additive mechanism with exogenous noise, where the structural function is parameterized by a KAN. Mixed data types, including categorical variables, are supported via appropriate likelihood parameterizations. Crucially, closed-form expressions can be extracted from each learned mechanism, enabling direct auditing and visualization of causal effects.

The contributions are threefold. First, this work introduces a KAN-based causal generative model for mixed-type tabular data that preserves SCM semantics and supports observational, interventional, and counterfactual reasoning in a query-agnostic manner. Second, the paper provides a validation pipeline based on observational distribution matching and independence testing of inferred exogenous noise, enabling practitioners to assess model misspecification from observational data alone. Third, it is developed an automated interpretability framework that prunes structural mechanisms, performs symbolic approximation, and extracts closed-form expressions and visualization tools for effect auditing. The novelty of KaCGM lies in integrating KAN/KAAM mechanisms into a mixed-type causal generative model that preserves SCM semantics and supports query-agnostic observational, interventional, and counterfactual reasoning. The contribution is not any single component in isolation, but their integration into a complete pipeline that learns structural equations, validates model adequacy through residual-based and distributional diagnostics, and audits the learned mechanisms through pruning, symbolic approximation, and visualization. As a result, interpretability is not treated as an isolated post-hoc step, but as part of an auditable causal generative workflow.

Code is provided at github.com/aalmodovares/kacgm and an interactive interface for visualizing the causal effects in the experiments is publicly available at huggingface.co/spaces/marElizo/kacgm.

Main abbreviations: additive noise model (ANM), average treatment effect (ATE), classifier two-sample test (C2ST), cumulative distribution function (CDF), counterfactual (CF), causal generative model (CGM), diffusion-based causal model (DBCM), data-generating process (DGP), and generative

adversarial network (GAN), Hilbert-Schmidt independence criterion (HSIC), joint Hilbert-Schmidt independence criterion (dHSIC), Kolmogorov-Arnold additive model (KAAM), Kolmogorov-Arnold network (KAN), major acute cardiovascular event (MAC), mean absolute error (MAE), maximum mean discrepancy (MMD), partial dependence plot (PDP), probability radar plot (PRP), random forest (RF), structural causal model (SCM), and variational autoencoder (VAE).

II. RELATED WORK

This work lies at the intersection of generative modeling, causal inference, and interpretable neural architectures. Deep generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) have achieved remarkable success in modeling complex data distributions [19], [20]. However, these models typically target the observational joint distribution and lack causal semantics, which limits their ability to answer interventional or counterfactual queries from observational data alone [5].

Causal generative models (CGMs) address this limitation by combining deep generators with SCM semantics, so that learned mechanisms can be intervened on rather than only sampled from observationally [5]. Existing CGMs differ mainly in how they parameterize the structural mechanisms. DCSM uses deep conditional mechanisms for tractable counterfactual inference [15]; VACA uses a variational graph autoencoder to answer causal queries over graph-structured variables [16]; and DBCM models causal mechanisms with diffusion models for interventional and counterfactual sampling [18]. Among continuous-data CGMs, flow-based methods are especially strong empirical baselines: CAREFL uses causal autoregressive flows [21], while *causal flows* restrict normalizing-flow transformations according to the causal graph, enabling tractable likelihoods, invertible abduction, and state-of-the-art performance on continuous causal-query benchmarks [17]. Adversarial and neural structural approaches provide another route: NCM uses neural structural functions for counterfactual identification [22], and DECAF introduces DAG constraints for causally aware synthetic data [23]. Some of these works further establish theoretical guarantees or identifiability under hidden confounders meeting backdoor criterion [24] or using proximal inference [22], [25]. These methods provide expressive SCM approximations and, in several cases, query-agnostic inference. However, their mechanisms are usually implemented by deep networks, flows, adversarial generators, or diffusion models whose functional form remains difficult to inspect, simplify, or communicate.

Interpretability addresses a complementary axis: whether the learned relation can be audited as a mechanism, not only whether it can answer a query [26]. In tabular prediction, generalized additive models and neural additive models expose feature-level contributions [27], [28]. In causal inference, causal rule ensembles describe heterogeneous effects through rules [29], surrogate models can be constrained to preserve causal guarantees [30]. These approaches improve transparency, but they usually target prediction, treatment-effect estimation, or restricted conditional

TABLE I

UNIFIED RESULTS FOR CONTINUOUS AND DISCRETE EXPERIMENTS. METRICS REPORT AVERAGES OVER THE 11 SYNTHETIC DATASETS (MEAN_{STD}) WITH 1000 TRAINING SAMPLES. LOWER IS BETTER FOR ALL METRICS. CONTINUOUS EXPERIMENTS ADDITIONALLY REPORT COUNTERFACTUAL ERROR (MAE_{CF}), WHICH IS NOT DEFINED FOR THE DISCRETE SETTING. THE LAST COLUMN OF EACH BLOCK REPORTS P-VALUES FROM POST-HOC WILCOXON TESTS AGGREGATED OVER ALL METRICS. THE FRIEDMAN TEST P-VALUE IS < 1E-3 FOR BOTH EXPERIMENTS, INDICATING SIGNIFICANT DIFFERENCES AMONG METHODS OVERALL. UNDERLINE DENOTES THE BEST METHOD. * INDICATES NO STATISTICALLY SIGNIFICANT DIFFERENCE ($p > 0.05$).

Model	Continuous variables						Mixed type variables				
	RF _{acc} ^{obs}	MMD _{obs} ^{obs} × 10 ^{−3}	RF _{acc} ^{int}	MMD _{int} ^{int} × 10 ^{−3}	MAE _{CF} × 10 ^{−2}	p-value	RF _{acc} ^{obs}	MMD _{obs} ^{obs} × 10 ^{−3}	RF _{acc} ^{int}	MMD _{int} ^{int} × 10 ^{−3}	p-value
KAN	0.51 _{0.01}	16.49 _{8.14}	0.53 _{0.02}	4.79 _{4.13}	5.10 _{2.43}	0.061*	0.52 _{0.02}	1.64 _{0.74}	0.54 _{0.02}	9.11 _{7.24}	0.042
KAAM	0.52 _{0.01}	16.76 _{9.62}	0.53 _{0.02}	4.76 _{5.13}	4.66 _{2.31}	0.055*	0.52 _{0.01}	1.89 _{0.93}	0.54 _{0.02}	8.40 _{5.63}	0.024
ANM	0.53 _{0.02}	8.86 _{2.61}	0.54 _{0.02}	3.02 _{2.5}	2.62 _{2.50}	<u>0.500</u> *	0.54 _{0.02}	1.22 _{0.44}	0.56 _{0.03}	6.22 _{5.16}	0.042
DBCM	0.56 _{0.02}	31.50 _{23.03}	0.55 _{0.03}	6.79 _{5.82}	3.98 _{1.52}	< 1e-3	0.58 _{0.04}	3.43 _{1.91}	0.57 _{0.03}	12.53 _{9.90}	< 1e-3
CFlow	0.52 _{0.02}	20.46 _{6.90}	0.61 _{0.10}	9.61 _{8.72}	3.86 _{1.75}	< 1e-3	0.53 _{0.03}	3.16 _{1.68}	0.62 _{0.09}	7.18 _{4.99}	< 1e-3
KAN _{mix}	—	—	—	—	—	—	0.52 _{0.01}	1.79 _{0.92}	0.52 _{0.01}	8.24 _{9.88}	<u>0.690</u> *
KAAM _{mix}	—	—	—	—	—	—	0.52 _{0.02}	1.60 _{0.78}	0.52 _{0.01}	6.29 _{4.33}	<u>0.690</u> *

mechanisms rather than a full mixed-type SCM simulator capable of arbitrary observational, interventional, and counterfactual queries.

Recently, Kolmogorov–Arnold networks (KANs) have been proposed as an alternative neural architecture in which scalar weights are replaced by learnable univariate functions defined on edges [31]. KAN 2.0 further emphasizes pruning, symbolic approximation, and scientific inspection of learned functions [32]. The architecture has been extended to convolutional models [33] and to graph neural networks for molecular property prediction [34]. In applications, KAAMs support interpretable clinical classification [35], Pendyala and Venkatachalam evaluate KANs in healthcare tasks [36], Cho et al. use interpretable KANs for time-series forecasting [37], and Alharbi et al. apply graph KANs to multi-omics biomarker identification [38]. In causal inference, KANITE uses KANs for individual treatment-effect estimation [39], while CausalKANs combine KANs with symbolic treatment-effect explanations [40].

Building on these developments, this work uses KANs as structural mechanisms inside a CGM. The resulting position is summarized in the Supplementary material, available online. Table I: compared with expressive CGMs such as flows, diffusion models, variational models, and adversarial generators, KaCGM emphasizes mechanism-level auditability; compared with interpretable predictive or treatment-effect models, it remains a query-agnostic generative SCM simulator.

III. BACKGROUND

A. Structural Causal Models

Let $\mathbf{x} = \{x_1, x_2, \dots, x_D\} \in \mathcal{X}$ be a set of random variables with joint distribution $P_{\mathbf{x}}$.

Definition 1: A causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of indices $\mathcal{V} = \{1, 2, \dots, D\}$ and directed edges $\mathcal{E}$, where $(j, k) \in \mathcal{E}$ denotes that x_j causally influences x_k .

Let $\text{pa}(j)$ denote the set of parents of node j . If the graph is Markov and faithful [41], the joint distribution factorizes according to the parental Markov condition [5]:

$$p(\mathbf{x}) = \prod_{j=1}^D p(x_j | \text{pa}(j)). \quad (1)$$

Definition 2: A Structural Causal Model (SCM) $\mathcal{M} = (\mathbf{f}, P_{\mathbf{u}})$ defines a data-generating process (DGP) where each

observed variable is generated from its parents and an exogenous noise variable $\mathbf{u} = \{u_1, \dots, u_D\}$:

$$x_j = f_j(\text{pa}(j), u_j), \quad \mathbf{u} \sim P_{\mathbf{u}}. \quad (2)$$

An SCM characterizes the full causal data-generating process and therefore determines the value of any causal query [6]. Interventions are formalized through the do-operator.

Definition 3 (do-operator [5]): Let $\mathcal{I} \subseteq \mathcal{V}$. The intervention $\text{do}(x_{\mathcal{I}} = a_{\mathcal{I}})$ replaces the structural assignments of the intervened variables by constants, while leaving the remaining mechanisms unchanged.

Causal queries are functionals of the SCM involving at least one intervention, including interventional distributions, causal effects, and counterfactual quantities [5], [6]. SCMs therefore support reasoning at the three levels of Pearl’s hierarchy: observational, interventional, and counterfactual. Interventional distributions are obtained by sampling $\mathbf{u} \sim P_{\mathbf{u}}$ and propagating the modified structural assignments, while counterfactuals are computed via the abduction–action–prediction procedure [5], [6].

B. Causal Generative Models

Causal generative models (CGMs) are generative machine learning models that respect a causal graph $\mathcal{G}$ and approximate the mechanisms of an SCM $\mathcal{M}$. A learned model $\hat{\mathcal{M}} = (\hat{\mathbf{f}}, \hat{P}_{\mathbf{u}})$ can then be used as a simulator of the underlying causal system.

Once the structural mechanisms are approximated, the model supports observational, interventional, and counterfactual reasoning through the usual do-operator semantics. For continuous variables, under causal sufficiency and with sufficiently expressive function classes, CGMs can recover the structural equations up to standard identifiability considerations [17], leading to consistent answers to causal queries [5], [42].

CGMs are typically *query-agnostic*: once trained, the same learned SCM approximation can be reused for different observational, interventional, and counterfactual queries without retraining [24]. This property is central for the present work, because the goal is not to train a separate estimator for each target query, but to learn auditable mechanisms that can be reused as a causal simulator.

Although neural networks are universal function approximators, certain classes of functions are known to be more efficiently approximated by spline-based representations, particularly smooth functions in low-data regimes [43]. This

motivates exploring alternative architectures that preserve flexibility while improving interpretability, such as Kolmogorov–Arnold networks.

C. Kolmogorov–Arnold Networks

Kolmogorov–Arnold networks (KANs) provide an alternative representation to deep neural networks. While standard neural networks rely on the universal approximation theorem [44], KANs are inspired by the Kolmogorov–Arnold representation theorem [45], [46], which states that any continuous function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ can be expressed as

$$f(\mathbf{x}) = \sum_{k=1}^{2D+1} \Phi_k \left(\sum_{q=1}^D \phi_{k,q}(x_q) \right), \quad (3)$$

where ϕ and Φ are univariate nonlinear functions.

Modern implementations approximate these functions using stacked layers, similarly to neural networks [31]. A KAN with L layers can be written as

$$\text{KAN}(\mathbf{x}) = (\Phi_{L-1} \circ \dots \circ \Phi_0)(\mathbf{x}), \quad (4)$$

where each layer applies sums of learnable univariate functions. In practice, the univariate functions are parameterized with B-splines,

$$\phi(z) = w_b b(z) + w_s \text{spline}(z), \quad (5)$$

with $b(\cdot)$ a fixed baseline and w_s, w_b learnable parameters [47]. This formulation allows training with standard optimization methods while maintaining a structured functional representation.

Because KANs explicitly model univariate transformations, they enable direct inspection of nonlinearities and facilitate symbolic approximation and pruning. Additional regularization and sparsification techniques can further simplify the learned functions [32]. In practice, shallow KAN architectures often achieve a favorable balance between expressiveness and interpretability, which has motivated their use in several scientific and applied domains [35], [37], [48].

IV. KOLMOGOROV–ARNOLD CAUSAL GENERATIVE MODEL

Assume a dataset, $\mathcal{D} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_D^{(i)}\}_{i=1}^N$, containing N i.i.d. samples from the unknown observational distribution $P_{\mathbf{x}}$, generated by an unknown SCM $\mathcal{M}$, with a known causal graph $\mathcal{G}$. The goal is to obtain a causal generative model $\tilde{\mathcal{M}}$ that approximates $\mathcal{M}$ and, in addition, that provides as many insights about each function f_j as possible: closed-form equations, interaction terms, symbolic approximations, feature importance, visualization plots of separable effects, etc.

In the following, *causal sufficiency* is assumed, i.e., that there are no hidden confounders, and therefore, all the exogenous variables are independent: $p(\mathbf{u}) = \prod p(u_j)$. In addition, as in most causal generative models that learn structural mechanisms conditional on a graph, the causal graph is assumed to be given. The consequences of graph misspecification are evaluated empirically in experiments and discussed as a limitation in Section VI.

The proposed approach, KaCGM, consists of several components. First, the KAN-based ANM is defined for continuous

variables. Second, the model is extended to support discrete or categorical variables. Third, techniques are described to achieve closed-form equations and plotting tools that help to achieve interpretability.

A. KaCGM as ANM

KaCGM is an additive noise model (ANM) in which each mechanism of the true SCM is modeled following (6), where each function that models the effect of the parents on their children, $\tilde{g}_j(\text{pa}(j))$, is approximated by a KAN.

$$\tilde{f}_j(\text{pa}(j), u_j) = \tilde{g}_j(\text{pa}(j)) + u_j \quad (6)$$

The choice of ANM has three main motivations: **i)** in the causal inference literature, there are many relevant papers that assume additive noise models to capture real DGPs [49], [50], [51], [52], [53] **ii)** it enables a very interpretable way to understand the exogenous variables as independent additive noise coming from measurement error, a modeling used in fields such as communications [54], robotics [55] or econometrics [56], and **iii)** it is reasonably easy to test if our ANM is a good approximator of the true data-generating process.

Regarding the third point, and although proving that the true DGP is additive is not testable, in general, the well-specification of an ANM with respect to the observational data can be assessed through falsification diagnostics [50]. When facing a real-world problem in which the ground-truth of causal effect cannot be calculated, model misspecification should be tested, and the model should only be used for causal queries when those diagnostics are satisfactory and the graph assumption is externally justified.

- Test the independence of the exogenous noise: $u_j \perp\!\!\!\perp \text{pa}(j)$. Given a well-specified causal graph, violation of the independence condition indicates model misspecification (heteroscedastic noise or nonadditivity), rather than a reversed causal direction [49]. Following Hoyer [49], the statistic of the Hilbert-Schmidt Independence test (HSIC) [57] is used. In the same fashion, it is recommended testing that $\mathbf{u}$ are jointly independent using the multidimensional test dHSIC [58].
- Test the observational match. If a more flexible (e.g. causal flows [17]) model achieves better observational fitting, this can indicate the DGP is nonadditive. The match of the observational distribution with both *maximum mean discrepancy* (MMD) and classifier two-sample test (C2ST) using the accuracy of a random forest, as explained in Section V.

These diagnostics should be interpreted as necessary but not sufficient safeguards. Meeting the observational matching and independence tests does not confirm that the true SCM is captured by an ANM; it only means that misspecification was not detected by the selected tests. In finite samples, independence tests may have limited power, classifier two-sample tests depend on the discriminator class, and multiple structurally different SCMs can remain observationally indistinguishable on the available support. The validation pipeline is therefore a falsification procedure, not a proof of causal correctness.

Operation details about training, hyperparameter search, implementation of the do-operator and model misspecification tests can be found in supplementary material, available online.

B. Mixed Type KaCGM_{mix}

In tabular data, there are usually discrete or categorical variables. Therefore, KaCGM is allowed to use discrete or categorical models as predictors; in this case, Logistic-KANs. Following the terminology of Almodovar et al. [35], a Kolmogorov-Arnold additive model (KAAM) is a KAN with no hidden layers, and a Logistic-KAN is a KAN that handles discrete variables by applying a sigmoid function for binary variables or a softmax function for categorical variables. Each network within KaCGM_{mix} to be a KAAM, a Logistic-KAN, a Logistic-KAAM, or a standard KAN.

Proposition 4.1 shows how modeling the distribution $p_\theta(x_j|pa(j))$ of a categorical variable with a flexible enough predictor is equivalent to recovering the function $f_j(pa(j), u_j)$ with the exogenous term implicit. Although the exogenous term cannot be identified, all interventional queries in the SCM can.

Proposition IV.1. (mixed CGM equivalence with SCMs): Consider the true SCM, $\mathcal{M}$, in which one variable, x_j is discrete, with observational probability mass function $p(x_j|pa(j))$. Let $p_\theta(x_j|pa(j))$ be a conditional categorical model (e.g. Logistic KAN) such that $p_\theta(\cdot|pa(j)) \rightarrow p(\cdot|pa(j))$. Being $F_\theta(\cdot|pa(j))$ the cumulative density function (CDF) of p_θ and defining

$$\tilde{f}_\theta(pa(j), u_j) = \min\{k : u \leq F_\theta(k|pa(j))\}, \quad (7)$$

with $u_j \sim \mathcal{U}(0, 1)$. Then **i)** $x_j = \tilde{f}_\theta(pa(j), u_j)$ defines a SCM mechanism whose induced conditional law equals $p_\theta(\cdot|pa(j))$: sampling from $p_\theta(x_j|pa(j))$ is equivalent to sampling from u_j and applying $\tilde{f}_\theta$; and **ii)** the interventional distribution of x_j and its descendants equals the one of $\mathcal{M}$.

On the other hand, identifiability of point-valued counterfactuals is not possible with discrete outcomes, and only bounds can be provided [59], [60]. Fortunately, as stated in Proposition 4.2, in a mixed SCM, there exist counterfactual values that can be recovered, depending on the intervention of interest.

Proposition IV.2. (Counterfactual identifiability in mixed GCM): Let the partition $\mathcal{V} = \mathcal{C} \cup \mathcal{Z}$ be the set of indices of the continuous and discrete variables, respectively. Given the intervention on $\mathcal{I}$, and let $de(\mathcal{I})$ be the descendants of $\mathcal{I}$ in $\mathcal{G}$. Consider a well-specified causal generative model, $\tilde{\mathcal{M}}$, in which the map $u_j \mapsto \tilde{f}_j(pa(j), u_j)$ for every $j \in \mathcal{C}$ is injective, then $\tilde{\mathcal{M}}$ yields point-valued counterfactual predictions if there are not discrete variables in the descendants of $\mathcal{I}$, i.e. $de(\mathcal{I}) \in \mathcal{C}$.

Proofs of Propositions 4.1 and 4.2 can be found in Appendix C of supplementary material. Note that both propositions are valid for any CGM that models the distribution of categorical variables.

This adaptation of KaCGM to KaCGM_{mix} allows us to increase flexibility when modeling tabular data beyond continuous features, while retaining the identifiability of some causal queries of interest. A sketch of KaCGM_{mix} can be observed in Fig. 1, where each function of the true SCM is modeled by a KAN, including the discrete variables, in which the exogenous noise is implicit.

C. Towards Interpretability

The interpretability pipeline defines steps that highlight the interpretable capabilities of the KANs, including inductive biases towards simplicity that not only regularize the output to get better approximations of smoother functions, but also improve interpretability:

- 1) Among KAN candidates whose validation metrics are not significantly worse than the best-performing candidate, select the architecture with fewer hidden layers whenever possible, ideally a KAAM.
- 2) Extract each predicted function $\tilde{f}_j$.
- 3) Prune the edges of the KAN that show little importance, following the process described by Liu et al. [32], yielding $\tilde{f}_j^{(p)}$. This only removes the non-relevant variables
- 4) Fit a symbolic substitution of $\tilde{f}_j^{(s)}$ following the process of Almodovar et al. [40]. This approximates the function to a combination of atoms: polynomials, trigonometric functions, exponentials, etc.

When performing each of these steps, the well-specification tests described in Section IV-A should be carried out. If the results are not satisfactory, the simplification step **is not recommended**. One key observation is that, in each of the steps, the distribution of the residuals should be fitted on observational data again; otherwise, the distribution matching will worsen.

Interpretability is treated as a constrained model-selection objective. Among candidate mechanisms whose validation metrics are not significantly worse than the best-performing candidate, the simplest architecture is selected. If this condition is not met, the more complex model may be retained for causal simulation, but the corresponding mechanism is not claimed to provide a compact symbolic explanation. This criterion makes explicit the trade-off between structural complexity and interpretability. If a shallow KAN or KAAM fits the data and passes the residual diagnostics, it is preferred because its mechanisms can be inspected more easily. If only a deeper KAN passes the diagnostics, the model may still be useful as a causal simulator, but the corresponding mechanism should be reported as less interpretable. If no simplified model passes the diagnostics, symbolic explanations should not be trusted.

Once a satisfactory, simplified $\tilde{f}_j$ is obtained, provided that the selected KAN has sufficiently low depth and complexity; alternatively, the contribution of each of the parents to the target variable can be plotted, extracting feature importance, and using the plotting tools developed by Almodovar [35]: *probability radar plots* (PRP) and *partial dependence plots* (PDP).

Summarizing, the process in a real environment with observational data consists of **i)** constructing and fit KaCGM in observational data, **ii)** testing the observational fit and the independence between exogenous inferred variables, comparing with other more flexible models, **iii)** applying the simplification steps of pruning and symbolic substitution, including tests of step **ii**, and accept the simplification only if the tests results are not worse up to some threshold established by the user, and **iv)** extracting the closed form expressions of the generating functions and provide interpretable plots.

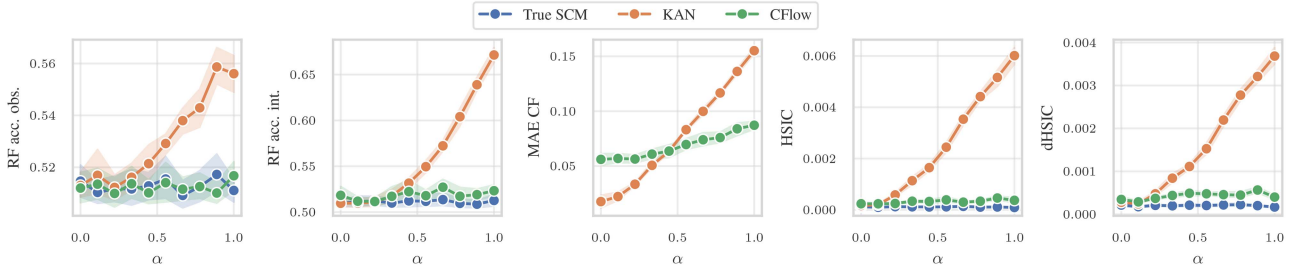


Fig. 2. Sensitivity analysis on additivity of exogenous variables. Mean and 95% confidence intervals over 10 realizations of the DGP. The true SCM represents the metrics obtained by data generated by the true SCM. CFlow, as a more flexible model, achieves good approximations and independence even with non-additive noise, while KaCGM deviates from this.

V. EXPERIMENTS

Four experiments support the use of KaCGM, based on the previous process. This section is organized as follows. First, the metrics that used to measure performance and assess the hypothesis are presented. Second, a sensitivity analyses how the model suffers from misspecification when the true DGP is not additive, and when the provided causal graph is not faithful, guiding the practitioner through the process in real data and showing the limitations of the approach. Third, the performance of KaCGM is evaluated in a set of synthetic datasets, of both continuous and discrete data, compared with state-of-the-art CGM, to show that under well specification, the performance of KaCGM is competitive for all data regimes, depending on the amount of data available. Fourth, results are presented on a semi-synthetic dataset with more complex and realistic data. Finally, the previous process is implemented in a real-world medical dataset, where misspecification tests and extract closed expressions of the estimated functions, also providing interpretable plots that help to understand the causal effects in the whole complex system are carried out, closed expressions of the estimated functions are extracted and interpretable plots are provided to clarify causal effects in the whole system.

A. Validation Metrics

For evaluating the fitting of the observational distribution and the interventional distributions, the experiments use both the maximum mean discrepancy (MMD) [61] with Gaussian kernel and a classifier two samples test (C2ST) [62], [63], where the test accuracy of a random forest (RF) trained with labeled data (real data labeled with 1 and synthetic data labeled with 0) is considered as our test statistic. Less MMD means more similarity between distributions, and an RF accuracy close to the chance level (0.5) indicates that the synthetic samples resemble the observational distribution. Second, mean absolute error (MAE_{CF}) is used to measure the counterfactual performance.

B. Sensitivity Analyses

a) Nonadditive data generating processes: A sensitivity analysis evaluates when it is sensible to adopt an additive noise model in a real case. Although in a real dataset counterfactual error and interventional MMD are unavailable, observational fit and independence between the extracted exogenous variables can still be measured. That serves to falsify the additivity of the true DGP.

The experiment generated 6000 samples of the following DGP and used the half for validation.

$$\begin{aligned} x_1 &= u_1 \\ x_2 &= \frac{x_1}{2} + .1x_1^2 + .6u_2 \\ x_3 &= .4x_1 + .1x_1^2 + .7x_2 - \frac{x_2^2}{2} + \frac{x_2^3}{8} - \frac{\alpha x_2 \tanh(u_3 + 1)}{2} + \frac{u_3}{2} \end{aligned}$$

Fig. 2 shows how observational and interventional fit, counterfactual error, and residual-independence diagnostics vary with the nonadditivity parameter α , where $\alpha = 0$ corresponds to an additive DGP and $\alpha > 0$ to a nonadditive one.

This experiment illustrates how the use of KaCGM can be rejected, since a more flexible causal model achieves better observational matching and independence of the residuals when the noise is not additive, and this is something that can be tested based on observational data only. These diagnostics should be interpreted as falsification tools: if they reject, there is evidence against the additive specification, but if they do not reject, additivity is not verified. An additional experiment in the Supplementary Material, available online further studies this behavior empirically in finite-sample regimes, by varying the sample size and the degree of nonadditivity. The results show the expected trend that the rejection probability increases with both sample size and misspecification strength (higher α), indicating higher power to detect nonadditivity as more data are available, but also showing limited power in low data regimes.

b) Sensitivity to graph misspecification: A complementary sensitivity analysis evaluates the effect of using a misspecified causal graph; full details are provided in the Supplementary Material, available online. The results show that adding a spurious edge has limited impact in this experiment, whereas removing or reversing a true causal relation substantially degrades interventional, average-treatment-effect, and counterfactual performance. Importantly, observational metrics and residual-independence diagnostics do not always reveal these failures, especially when the misspecification preserves a plausible observational fit. This supports interpreting the validation pipeline as a falsification tool, not as a certificate of graph correctness, and reinforces that the causal use of KaCGM is conditional on the assumed graph.

C. Performance on Synthetic Datasets

The proposed approach is benchmarked against state-of-the-art methods on 11 synthetic graphs obtained from Geffner [64]

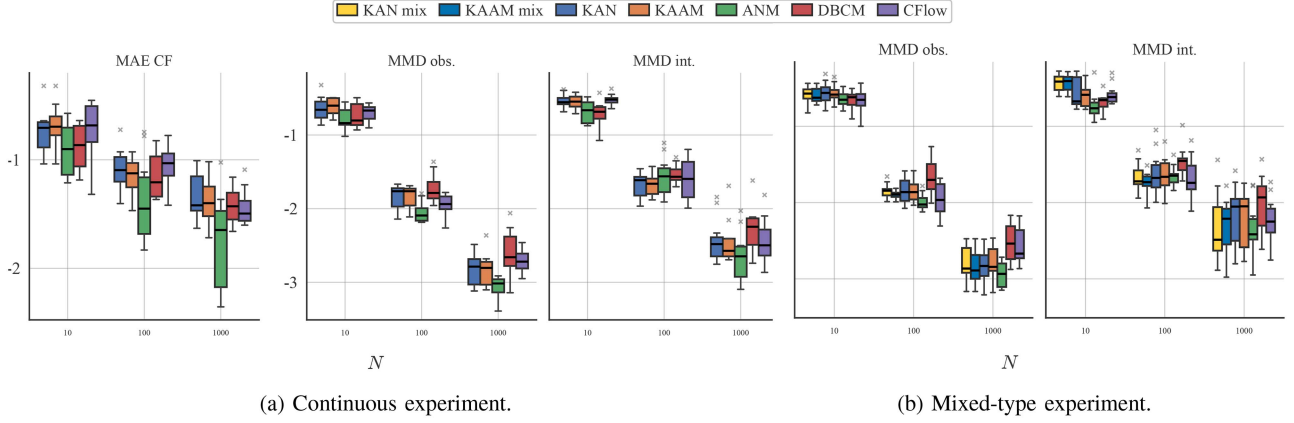


Fig. 3. Boxplot of $\log_{10}$ of observational and interventional MMD of each model across 10 realizations with (a) continuous data and (b) mixed type data, as a function of the number of samples with mixed type data.

under different data regimes. Two different experiments are conducted, depending on whether all covariates are continuous or if there are discrete variables. All variables are standardized to obtain comparable counterfactual errors across variables. All experiments use the same amount of data for training and testing.

1) *Experiment on Continuous Data:* The performance of our proposed approaches (KaCGM with KAN and KAAM) is compared to the following methods: DBCM [18] (diffusion-based), causal flows [17] (from now CFlow) and ANM [49] (Additive Noise Model, which selects the best performing method among several classical machine learning tools using regression metrics [65]). Both DBCM and CFlow are universal density approximators, although they need, in principle, larger sample sizes to train [21].

Once the best hyperparameters (details in the Supplementary Material, available online) are obtained, the final model is trained and validated. The validation is done along three main axes: observational, interventional and counterfactual metrics. For *observational metrics*, RF accuracy and MMD between the generated distribution and the validation samples are compared. To measure *interventional prediction*, six different interventions are generated over variables x_1 and x_2 (always present in the graph) at standardized values $x_j = \{-1, 0, 1\}$. The corresponding interventional distributions are evaluated using RF accuracy and MMD with the validation samples. Finally, for *counterfactual inference*, the validation samples are used to compute the counterfactual values of each sample using the same interventions defined in the previous point. The counterfactual error is evaluated using MAE (MAE_{CF}), compared with the values of the true SCM.

2) *Experiment on Mixed Data:* The discrete case follows a similar procedure, with some key differences. First, the validation of models designed for continuous variables (CFlow and DBCM) is adapted by i) adding a small scale noise to the training data (standard deviation 0.01); and ii) post-processing the data generated for the discrete variables, by rounding. Second, $KaCGM_{mix}$ is added to the evaluation. Third, one variable, $x_3 \in \{0, 1, 2\}$, is discretized in the true DGP of Geffner et al. [64] by sampling a ternary variable with probability vectors: $[0.8, 0.1, 0.1]$ if $x_3 < -1$; $[0.1, 0.8, 0.1]$ if $-1 < x_3 < 1$; and $[0.1, 0.1, 0.8]$ if $x_3 > 1$. Finally, counterfactual prediction is

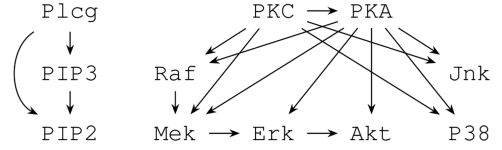


Fig. 4. Causal graph associated with Sachs' dataset.

not evaluated, since point-valued counterfactuals are not identifiable.

Results: The results for 1000 samples can be observed in Table I, showing close performance of all the models, and the p-values comparing the models in all the metrics together, using the procedure of Demsar [66]. Observational, interventional and counterfactual metrics are competitive: $RF_{acc}^{obs} = 0.51$, $RF_{acc}^{int} = 0.53$ and $MAE_{CF} = 4.66 \times 10^{-2}$. P-values show that i) in the continuous case, ANM is the best model, while the two variants of KaCGM are not rejectable as the best models for type I error probability 0.05 (p -value=0.061), and ii) in the discrete case, $KaCGM_{mix}$ are the best models across all metrics and realizations (p -values of competitors < 0.05), supporting the proposal of using mixed models with discrete data. Note that CFlow and DBCM perform slightly worse because the number of samples is low and the true DGPs of Geffner et al. [64] lie within the ANM class. The sample-size study in Fig. 3 shows the expected degradation as N decreases.

D. Performance in Semi-Synthetic Dataset

Following other works [18], [25], the proposed approach is also validated on a more complex semi-synthetic dataset, created following the process procedure by Chao et al. [18], from the original dataset of Sachs et al. [67]. The dataset contains 11 variables associated with the causal graph of Fig. 4. Root nodes are sampled from the original dataset, while the other variables are generated using random equations.

The experiment uses 1000 training samples and also 1000 validation samples in both additive and nonadditive settings. The results using the same evaluation method from the previous experiment can be seen in Table II. In the additive setting, ANM achieves the best results in all metrics, while KaCGM

TABLE II

RESULTS ON THE SACHS DATASET FOR ADDITIVE AND NONADDITIVE NOISE SETTINGS. METRICS REPORT AVERAGES OVER RUNS (MEAN_{STD}). LOWER IS BETTER FOR ALL METRICS. THE LAST COLUMN OF EACH BLOCK REPORTS P-VALUES FROM POST-HOC WILCOXON TESTS AGGREGATED OVER ALL METRICS. THE FRIEDMAN TEST P-VALUE IS ALWAYS $< 1e-3$, INDICATING SIGNIFICANT DIFFERENCES AMONG METHODS OVERALL. UNDERLINE DENOTES THE BEST METHOD. * INDICATES NO STATISTICALLY SIGNIFICANT DIFFERENCE ($p > 0.05$).

Model	Additive						Nonadditive					
	RF _{acc} ^{obs}	MMD _{obs} $\times 10^{-3}$	RF _{acc} ^{int}	MMD _{int} $\times 10^{-3}$	MAE _{CF} $\times 10^{-2}$	p-value	RF _{acc} ^{obs}	MMD _{obs} $\times 10^{-3}$	RF _{acc} ^{int}	MMD _{int} $\times 10^{-3}$	MAE _{CF} $\times 10^{-2}$	p-value
KAN	0.54 _{0.02}	1.36 _{0.41}	0.54 _{0.02}	1.95 _{0.67}	3.50 _{1.40}	0.180*	0.67 _{0.13}	5.27 _{9.20}	0.67 _{0.14}	6.09 _{8.75}	3.42 _{1.54}	$< 1e-3$
KAAM	0.54 _{0.01}	1.35 _{0.40}	0.55 _{0.01}	1.96 _{0.74}	3.59 _{1.35}	0.180*	0.68 _{0.14}	5.97 _{9.60}	0.68 _{0.14}	6.61 _{9.12}	3.56 _{1.68}	$< 1e-3$
ANM	0.53_{0.01}	1.12_{0.28}	0.53_{0.01}	1.30_{0.21}	1.28_{0.43}	0.500*	0.61 _{0.09}	1.71 _{0.83}	0.61 _{0.09}	2.31 _{1.46}	1.76 _{1.13}	0.007
DBCM	0.56 _{0.01}	2.78 _{0.73}	0.56 _{0.01}	2.38 _{0.54}	3.03 _{0.47}	$< 1e-3$	0.59 _{0.04}	2.61 _{0.78}	0.60 _{0.04}	2.34 _{0.60}	2.67 _{0.78}	0.157*
CFlow	0.53 _{0.01}	1.42 _{0.26}	0.53 _{0.01}	1.53 _{0.29}	1.69 _{0.52}	0.396*	0.55_{0.03}	1.42_{0.35}	0.59_{0.08}	1.84_{0.96}	1.84_{0.99}	0.500*

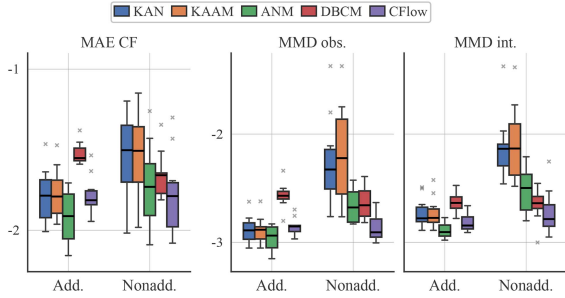


Fig. 5. Boxplots of $\log_{10}$ of observational MMD, interventional MMD and CF MAE on Sachs' semisynthetic dataset. While in the additive setting, all models achieve similar performance, KaCGM worsen significantly in the nonadditive.

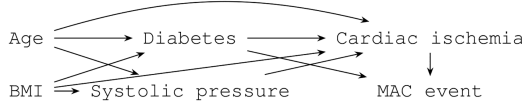


Fig. 6. Causal graph of cardio dataset. The original graph contained *obesity* and *hypertension*; their continuous counterparts are used instead, BMI and systolic pressure.

and CFlows get comparable results (p -values across all metrics of 0.18 and 0.39, respectively), while DBCM is not competitive with this number of samples. On the other hand, in the non-additive settings, causal flows (in particular, causal neural spline flows, CausalNSF, which is more flexible) achieves the best results, with DBCM presenting a competitive performance. These findings can also be observed in Fig. 5.

E. Use Case: Cardiovascular Data

The interpretability capabilities of the proposed approach are illustrated on a real-world cardiovascular dataset whose goal is to estimate the probability of a *major acute cardiovascular event* (MAC). The dataset, introduced by Kyriacou et al. [68], contains continuous and binary clinical covariates together with a causal graph that is assumed to be correct and shown in Fig. 6.

Since true interventional distributions or counterfactual outcomes are not available, the evaluation must rely on observational diagnostics. Following the validation pipeline described in Section IV, KaCGM_{mix} is trained using KAN and KAAM subnetworks and compared with the baselines. In addition, the simplification pipeline (pruning and symbolic substitution) is

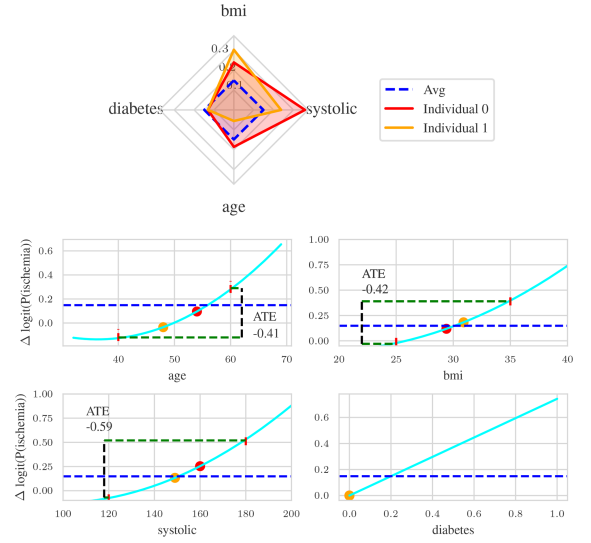


Fig. 7. Interpretability analysis of the ischemia mechanism. (Top) PRPs showing the contribution of each parent variable to the logit for two individuals. (Bottom) PDPs representing the contribution Δ of each variable to the logit. The curves also allow direct computation of average treatment effects (ATE) corresponding to hypothetical interventions on each variable.

applied, and the effect of these transformations on the observational fit is evaluated.

The results are summarized in Fig. 8. Pruning preserves the observational performance, whereas symbolic substitution slightly degrades the metrics, illustrating the expected trade-off between predictive fidelity and interpretability. Nevertheless, KaCGM_{mix} achieves competitive observational fit, obtaining the best RF accuracy and MMD among the compared models, although independence diagnostics are less favorable. This analysis supports the use of the simplified models when interpretability is a primary objective.

The key advantage of the proposed model is that the structural mechanisms can be explicitly inspected. For example, after pruning and symbolic approximation, the probability of cardiac ischemia conditioned on its parents can be expressed as

$$\begin{aligned} \Pr(\text{ischemia}) = & \sigma(0.055 \overline{\text{age}}^2 + 0.173 \overline{\text{age}} \\ & + 0.031 \overline{\text{BMI}}^2 + 0.114 \overline{\text{BMI}} \\ & + 0.046 \overline{\text{systolic}}^2 + 0.155 \overline{\text{systolic}} \\ & + 0.743 \overline{\text{diabetes}} - 1.871), \end{aligned}$$

where σ represents the sigmoid function and the overline variables are standardized. This expression is directly auditable:

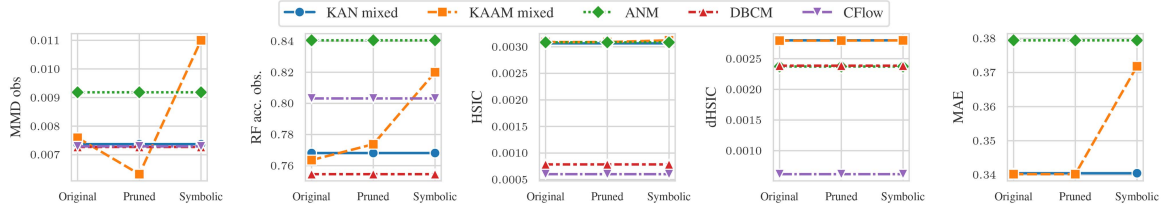


Fig. 8. Observational metrics on the cardiovascular dataset as a function of the simplification stage. MAE denotes predictive error when estimating each node from its parents. DBCM and CFlow cannot compute conditional samples and, therefore, MAE is not reported. HSIC and dHSIC are evaluated only for continuous variables.

the ischemia logit increases through quadratic age, systolic pressure, and BMI terms, matching the expected nonlinear increase of cardiovascular risk with these factors.

Beyond the equation, Fig. 7 summarizes individual and population-level effects. PRPs decompose parent contributions for two individuals; for example, Individual 0 exhibits higher ischemia risk due to greater age and systolic pressure, whereas Individual 1 has elevated BMI, which also increases the predicted probability. In addition, probability dependence plots (PDPs) show how hypothetical interventions change the ischemia logit at the population level. Because the structural equation is explicit, the same curves yield ATEs directly; for example, reducing BMI from 35 to 20 gives approximately -0.42 , and reducing systolic pressure from 180 to 120 gives approximately -0.59 , both in logit scale. Importantly, these summaries are therefore obtained from an explicit mechanism rather than from a black-box predictor.

VI. CONCLUDING REMARKS

This work investigated the integration of Kolmogorov-Arnold networks within causal generative modeling for tabular data. This work proposed KaCGM, a causal generative model that preserves the semantics of structural causal models while enabling direct inspection of the learned structural mechanisms. KaCGM is a query-agnostic model capable of answering observational, interventional, and counterfactual queries, while supporting symbolic simplification, pruning, and functional visualization.

Empirically, KaCGM is competitive under additive data-generating processes and in mixed continuous-discrete settings, while more flexible causal generative models dominate in strongly nonadditive regimes. The sensitivity analyses show that observational fit and residual-independence diagnostics are useful falsification tools, although not certificates of causal correctness. This trade-off highlights the central design principle of the proposed approach: favoring structural interpretability and auditability when the data-generating process is compatible with additive mechanisms. The cardiovascular use case illustrates the intended role of the method: providing causal-query performance together with simplified structural equations and auditable visualizations. These results support the feasibility of combining expressive generative modeling with functional transparency, a property that is particularly relevant for high-stakes applications where auditing and mechanistic understanding are required.

Despite its advantages, the proposed approach presents several limitations. First, the model relies on structural mechanisms that admit relatively low-complexity functional decompositions. When the true data-generating process exhibits strongly nonadditive, highly entangled, or heteroscedastic interactions, more flexible deep causal generative models can achieve superior density approximation and counterfactual accuracy, as observed in Section V. In such regimes, enforcing structured functional representations may lead to approximation bias. Second, interpretability is inherently tied to model simplification: simple models are preferred but not always sufficiently flexible. Pruning and symbolic approximation improve transparency but may degrade predictive fidelity, introducing a trade-off between auditability and accuracy. While the validation pipeline mitigates this risk through independence and distributional diagnostics, it cannot guarantee that the learned mechanisms coincide with the true structural equations, particularly under model misspecification and in low-sample regimes.

Third, the methodology assumes causal sufficiency and a correctly specified causal graph, as is common in causal generative models that learn structural mechanisms conditional on a given graph. This assumption is central because KaCGM learns mechanisms once the parent sets are fixed. Consequently, hidden confounding or graph misspecification can invalidate the generative interpretation and the associated interventional and counterfactual estimates, even when the observational distribution is well approximated. Therefore, trustworthy use of KaCGM requires the assumed graph to be supported by experimentation, domain expertise, and, when appropriate, statistical causal-discovery tools; the model alone cannot certify causal validity under graph misspecification.

Several directions for future research arise from this work. First, one may extend the architecture to increase density approximation flexibility while preserving interpretability. In particular, KAN-based normalizing flow constructions could allow more expressive transformations of the exogenous variables, including neural spline flows [69] or unconstrained monotonic neural networks [70]. Such extensions would require new interpretability tools to ensure that increased flexibility does not compromise auditability. Second, alternative functional bases beyond spline parameterizations may be explored, such as Fourier expansions [34], which could provide improved inductive biases in periodic or oscillatory mechanisms. Third, assessing partial well-specification represents an important practical direction. Instead of requiring global model adequacy, one may restrict attention to subsets of intervention targets for which additive

structure is sufficiently accurate, as suggested in recent work on partial model validation [50]. Another promising avenue is regression by dependence minimization [53], enforcing independence between inferred exogenous variables and their parents to strengthen structural validity. Finally, extending the methodology beyond causal sufficiency remains an open challenge. Incorporating proxy-based adjustments [25] or leveraging graph transformations that yield equivalent identifiable structures [24] would broaden applicability to settings with hidden confounding.

In summary, this work advances interpretable causal generative modeling by demonstrating that structured, additive neural architectures can provide competitive performance while enabling direct inspection of structural equations. This contribution aims to encourage further research at the intersection of causal modeling, generative learning, and functional interpretability.

ACKNOWLEDGMENT

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

REFERENCES

- [1] D. M. Kent, E. Steyerberg, and D. van Klaveren, "Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects," *BMJ*, vol. 363, 2018, Art. no. k4245.
- [2] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsafaris, "Causal machine learning for healthcare and precision medicine," *Roy. Soc. Open Sci.*, vol. 9, no. 8, 2022, Art. no. 220638.
- [3] K. Imai and A. Strauss, "Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign," *Political Anal.*, vol. 19, no. 1, pp. 1–19, 2011.
- [4] C. F. Manski, "Statistical treatment rules for heterogeneous populations," *Econometrica*, vol. 72, no. 4, pp. 1221–1246, 2004.
- [5] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.
- [6] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.
- [7] A. Curth, R. W. Peck, E. McKinney, J. Weatherall, and M. van der Schaar, "Using machine learning to individualize treatment effect estimation: Challenges and opportunities," *Clin. Pharmacol. Therapeutics*, vol. 115, no. 4, pp. 710–719, 2024.
- [8] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation)," *Official J. Eur. Union*, vol. L 119, no. 3, pp. 1–88, May 2016. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [9] European Parliament and Council, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 167/2012, (EU) 2018, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," *Official J. Eur. Union*, L2024/1689, Jul. 12, 2024. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [10] U. S. Food and Drug Administration, Health Canada, and Medicines and Healthcare products Regulatory Agency, "Transparency for machine learning-enabled medical devices: Guiding principles," Jun. 13, 2024. Accessed: Aug. 16, 2026. [Online]. Available: <https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles>
- [11] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Mag.*, vol. 38, no. 3, pp. 50–57, 2017.
- [12] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proc. 4th Mach. Learn. Healthcare Conf.*, PMLR, 2019, pp. 359–380.
- [13] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Informat. Decis. Mak.*, vol. 20, 2020, Art. no. 310.
- [14] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
- [15] N. Pawlowski, D. C. Castro, and B. Glocker, "Deep structural causal models for tractable counterfactual inference," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 857–869.
- [16] P. Sánchez-Martín, M. Rateike, and I. Valera, "VACA: Designing variational graph autoencoders for causal queries," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 8159–8168.
- [17] A. Javaloy, P. Sánchez-Martín, and I. Valera, "Causal normalizing flows: From theory to practice," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 58833–58864.
- [18] P. Chao, P. Blöbaum, S. Patel, and S. P. Kasiviswanathan, "Modeling causal mechanisms with diffusion models for interventional and counterfactual queries," *Trans. Mach. Learn. Res.*, Nov. 2024. [Online]. Available: <https://openreview.net/forum?id=EDHQDsqiSe>
- [19] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [20] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
- [21] I. Khemakhem, R. P. Monti, R. Leech, and A. Hyvärinen, "Causal autoregressive flows," in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3520–3528.
- [22] K. M. Xia, Y. Pan, and E. Bareinboim, "Neural causal models for counterfactual identification and estimation," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.
- [23] B. Van Breugel, T. Kyono, J. Berrevoets, and M. Van der Schaar, "DECAF: Generating fair synthetic data using causally-aware generative networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 22221–22233.
- [24] Á. Parafita and J. Vitrià, "Estimand-agnostic causal query estimation with deep causal graphs," *IEEE Access*, vol. 10, pp. 71370–71386, 2022.
- [25] A. Almodóvar, A. Javaloy, J. Parras, S. Zazo, and I. Valera, "DeCaFlow: A deconfounding causal generative model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2025, pp. 136612–136666. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2025/file/c78f639424b8d89ceb4f2efbb4dfe4f4-Paper-Conference.pdf
- [26] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.
- [27] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 1721–1730.
- [28] R. Agarwal et al., "Neural additive models: Interpretable machine learning with neural nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 4699–4711.
- [29] F. J. Bargagli-Staffi, R. Cadei, K. Lee, and F. Dominici, "Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects," 2020, *arXiv:2009.09036*.
- [30] C. Kim and O. Bastani, "Learning interpretable models with causal guarantees," 2019, *arXiv:1901.08576*.
- [31] Z. Liu et al., "KAN: Kolmogorov–Arnold networks," in *Proc. 13th Int. Conf. Learn. Representations*, 2025.
- [32] Z. Liu, P. Ma, Y. Wang, W. Matusik, and M. Tegmark, "KAN 2.0: Kolmogorov–Arnold networks meet science," 2024, *arXiv:2408.10205*.
- [33] A. D. Bodner, J. N. Spolski, A. S. Tepsich, and S. Pourteau, "Convolutional Kolmogorov–Arnold networks," 2024, *arXiv:2406.13155*.
- [34] L. Li, Y. Zhang, G. Wang, and K. Xia, "Kolmogorov–Arnold graph neural networks for molecular property prediction," *Nature Mach. Intell.*, vol. 7, no. 8, pp. 1346–1354, 2025.
- [35] A. Almodóvar, P. A. Apellániz, A. Garrido, F. Fernández-Salvador, S. Zazo, and J. Parras, "Interpretable clinical classification with Kolmogorov–Arnold networks," 2025, *arXiv:2509.16750*.
- [36] V. S. Pendyala and N. Venkatachalam, "The effectiveness of Kolmogorov–Arnold networks in the healthcare domain," *Appl. Sci.*, vol. 15, no. 16, 2025, Art. no. 9023.

- [37] S.-Y. Cho, S. Lee, and H.-G. Kim, "Forecasting VIX using interpretable Kolmogorov-Arnold networks," *Expert Syst. Appl.*, vol. 294, 2025, Art. no. 128781.
- [38] F. Alharbi et al., "Interpretable graph Kolmogorov-Arnold networks for multi-cancer classification and biomarker identification using multi-omics data," *Sci. Rep.*, vol. 15, Jul. 2025, Art. no. 27607.
- [39] E. Mehendale, A. Thorat, R. Kolla, and N. Pedanekar, "KANITE: Kolmogorov-Arnold networks for ITE estimation," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, Springer, 2025, pp. 409–425.
- [40] A. Almodóvar, P. A. Apellániz, S. Zazo, and J. Parras, "CausalKANs: Interpretable treatment effect estimation with Kolmogorov-Arnold networks," 2025, *arXiv:2509.22467*.
- [41] J. Peters, D. Janzing, and B. Schölkopf, "Causal inference on discrete data using additive noise models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2436–2450, Dec. 2011.
- [42] J. Pearl, "Probabilities of causation: Three counterfactual interpretations and their identification," *Synthese*, vol. 121, no. 1, pp. 93–149, 1999.
- [43] J. Friedman, "Adaptive spline networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Morgan-Kaufmann, 1990, pp. 675–683.
- [44] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.
- [45] A. N. Kolmogorov, "On the representations of continuous functions of many variables by superposition of continuous functions of one variable and addition," *Dokl. Akad. Nauk USSR*, vol. 114, pp. 953–956, 1957.
- [46] V. I. Arnold, "On functions of three variables," *Doklady Akademii Nauk SSSR*, vol. 114, pp. 679–681, 1957.
- [47] C. De Boor and C. De Boor, *A Practical Guide to Splines*, vol. 27. New York, NY, USA: Springer, 1978.
- [48] W. Knottenbelt et al., "CoxKAN: Kolmogorov-Arnold networks for interpretable, high-performance survival analysis," *Bioinformatics*, vol. 41, no. 8, Jul. 2025, Art. no. btaf413.
- [49] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Non-linear causal discovery with additive noise models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 689–696.
- [50] C. Schultheiss and P. Bühlmann, "Assessing the overall and partial causal well-specification of nonlinear additive noise models," *J. Mach. Learn. Res.*, vol. 25, no. 159, pp. 1–41, 2024.
- [51] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, no. 10, pp. 2003–2030, 2006.
- [52] P. Bühlmann, J. Peters, and J. Ernest, "CAM: Causal additive models, high-dimensional order search and penalized regression," *Ann. Statist.*, vol. 42, no. 6, pp. 2526–2556, 2014.
- [53] J. Mooij, D. Janzing, J. Peters, and B. Schölkopf, "Regression by dependence minimization and its application to causal inference in additive noise models," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 745–752.
- [54] J. G. Proakis, M. Salehi, N. Zhou, and X. Li, *Communication Systems Engineering*, vol. 2. Hoboken, NJ, USA: Prentice Hall, 1994.
- [55] T. D. Barfoot, *State Estimation for Robotics*. Cambridge, U.K.: Cambridge Univ. Press, 2024.
- [56] W. A. Fuller, *Measurement Error Models*. Hoboken, NJ, USA: Wiley, 2009.
- [57] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *Proc. Int. Conf. Algorithmic Learn. Theory*, Springer, 2005, pp. 63–77.
- [58] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters, "Kernel-based tests for joint independence," *J. Roy. Stat. Soc. Ser. B: Stat. Methodol.*, vol. 80, no. 1, pp. 5–31, 2018.
- [59] A. Balke and J. Pearl, "Bounds on treatment effects from studies with imperfect compliance," *J. Amer. Stat. Assoc.*, vol. 92, no. 439, pp. 1171–1176, 1997.
- [60] J. Tian and J. Pearl, "Probabilities of causation: Bounds and identification," *Ann. Math. Artif. Intell.*, vol. 28, no. 1, pp. 287–313, 2000.
- [61] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.
- [62] J. Friedman, "On multivariate goodness-of-fit and two-sample testing," Stanford Linear Accelerator Center, Menlo Park, CA, USA, Tech. Rep. SLAC-PUB-10325, Jan. 2004.
- [63] D. Lopez-Paz and M. Oquab, "Revisiting classifier two-sample tests," in *Proc. Int. Conf. Learn. Representations*, 2017.
- [64] T. Geffner et al., "Deep end-to-end causal inference," 2022, *arXiv:2202.02195*.
- [65] P. Blöbaum, P. Götz, K. Budhathoki, A. A. Mastakouri, and D. Janzing, "DoWhy-GCM: An extension of DoWhy for causal inference in graphical causal models," *J. Mach. Learn. Res.*, vol. 25, no. 147, pp. 1–7, 2024.
- [66] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
- [67] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.
- [68] D. N. Kyriacou, P. Greenland, and M. A. Mansournia, "Using causal diagrams for biomedical research," *Ann. Emerg. Med.*, vol. 81, no. 5, pp. 606–613, 2023.
- [69] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, "Neural spline flows," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7511–7522.
- [70] A. Wehenkel and G. Louppe, "Unconstrained monotonic neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1545–1555.



Alejandro Almodóvar received the MSc degree in communications engineering from Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2022. Currently, he is working toward the PhD degree whose work focuses on causal inference with deep learning, with applications in medicine and healthcare, among others.



Mar Elizo received the MSc degree in artificial intelligence, pattern recognition, and digital imaging from the Universitat Politècnica de València (UPV), in 2024. She is currently working toward the PhD degree with Universidad Politécnica de Madrid (UPM), focusing on deep learning and causal inference applications in healthcare.



Patricia A. Apellániz received the PhD degree in telecommunication engineering from the Universidad Politécnica de Madrid (UPM), Spain, in 2025. She is currently a researcher and teaching collaborator with UPM. Her research interests include deep learning for healthcare and signal processing.



Santiago Zazo is currently working toward the Dr Engineer degree with the Universidad Politécnica de Madrid (UPM), since 1995. In 1998, he joined UPM, where he is currently a full professor in signal theory and communications. More recently, he has been mostly focused on deep learning with medical applications, distributed optimization, game theory, and reinforcement learning. He is the author/co-author of more than 40 journal articles and about 200 conference papers. His main research interest includes signal processing.



Juan Parras received the PhD degree in telecommunications engineering from the Universidad Politécnica de Madrid (UPM), in 2020, respectively. He is currently an assistant professor with UPM. His research interests include deep generative models, deep reinforcement learning, game theory, and optimization with health and communications applications.