Exact Shapley Attributions in Quadratic-time for FANOVA Gaussian Processes

Majid Mohammadi[1,2], **Krikamol Muandet**[2], **Ilaria Tiddi**[1], **Annette Ten Teije**[1], **Siu Lun Chau**[3]

[1]Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands
[2]Rational Intelligence Lab, CISPA Helmholtz Center for Information Security, Germany
[3]EPIC Lab, College of Computing & Data Science, Nanyang Technological University, Singapore
majid.mohammadi690@gmail.com, muandet@cispa.de, annette.ten.teije@vu.nl,i.tiddi@vu.nl, siulun.chau@ntu.edu.sg

Abstract

Shapley values are widely recognized as a principled method for attributing importance to input features in machine learning. However, the exact computation of Shapley values scales exponentially with the number of features, severely limiting the practical application of this powerful approach. The challenge is further compounded when the predictive model is probabilistic—as in Gaussian processes (GPs)—where the outputs are random variables rather than point estimates, necessitating additional computational effort in modeling higher-order moments. In this work, we demonstrate that for an important class of GPs known as FANOVA GP, which explicitly models all main effects and interactions, exact Shapley attributions for both local and global explanations can be computed in *quadratic* time. For *local, instance-wise explanations*, we define a stochastic cooperative game over function components and compute the *exact stochastic Shapley value* in quadratic time only, capturing both the expected contribution and uncertainty. For *global explanations*, we introduce a deterministic, variance-based value function and compute exact Shapley values that quantify each feature's contribution to the model's overall sensitivity. Our methods leverage a closed-form (stochastic) Möbius representation of the FANOVA decomposition and introduce recursive algorithms, inspired by Newton's identities, to efficiently compute the mean and variance of Shapley values. Our work enhances the utility of explainable AI, as demonstrated by empirical studies, by providing more scalable, axiomatically sound, and uncertainty-aware explanations for predictions generated by structured probabilistic models.

1 Introduction

As machine learning (ML) systems are increasingly deployed in high-stakes applications, the demand for interpretability has grown substantially. Practitioners and regulators alike now seek models whose predictions can be understood and trusted—not only globally, across the entire data distribution, but also locally, for specific predictions. To address this need, the literature offers two distinct approaches: (i) designing inherently interpretable models, such as linear models or generalized additive models (GAMs), or (ii) applying post-hoc interpretation methods like SHAP (Lundberg and Lee 2017) or LIME (Ribeiro, Singh, and Guestrin 2016) to interpret complex, black-box models. However, interpretability is context-dependent. A model that appears interpretable to ML practitioners—such as GAM—may remain opaque to domain experts or end-users, such as medical professionals. In such cases, even inherently interpretable models may require an additional explanatory layer to translate their insights into signals accessible to non-technical stakeholders.

Functional ANOVA Gaussian Processes (FANOVA GPs) (Durrande et al. 2011) are a class of probabilistic models that combine the flexibility of GPs with the interpretability of functional ANOVA decompositions. They represent the prediction function as a sum of functions defined over subsets of input features, where each term models either a main effect or a higher-order interaction. By enforcing functional ANOVA constraints (Hooker 2004), these components become orthogonal (see Section 2), allowing the contribution from each subset to be uniquely and meaningfully identified. Unlike conventional kernels, which typically entangle all features through a single joint interaction term, FANOVA GPs encode a structured hierarchy of interactions—from individual features to the full feature set—making the decomposition both interpretable and expressive. This structure enables precise attribution of predictive behavior to specific feature sets, while preserving GP's nonparametric nature and the ability to quantify uncertainty. As a result, FANOVA GPs offer a compelling foundation for interpretable probabilistic modeling.

Nonetheless, as FANOVA GPs impose an interpretable additive structure by construction, the number of potential interactions grows exponentially with the number of input features. Consequently, identifying the overall contribution of each feature—especially in the presence of higher-order interactions—can become cognitively intractable for human users, much like interpreting a random forest by examining each tree's decision logic. Moreover, existing interpretation methods for FANOVA GP either lack axiomatic support or are computationally inefficient. For example, global sensitivity analysis techniques based on first-order Sobol indices (Sobol 2001; Lu, Boukouvalas, and Hensman 2022) ignore interaction effects, fail to satisfy key axioms such as efficiency (Owen 2014), and often underestimate feature importance. On the other hand, although the GPSHAP algorithm of Chau, Muandet, and Sejdinovic (2023) could, in

principle, be applied to FANOVA GPs, it does not exploit their additive structure—a structure that, as we show, enables significant computational savings. Moreover, the standard expectation-based value functions they employ require additional estimation, introducing potential sources of error. In contrast, our approach adopts a more natural value function tailored to FANOVA models (Fumagalli et al. 2025), allowing us to compute exact Shapley values without estimation, thereby avoiding approximation errors entirely. Our approach also yields stochastic explanations that account for the predictive uncertainty inherent in GPs.

In short, this paper proposes algorithms that leverage the structure of FANOVA GPs to deliver both *local* and *global* explanations via the Shapley value (Shapley 1953), computed *exactly* and in *quadratic time*. Our key technical insight is that the orthogonal additive decomposition allows us to derive a closed-form (stochastic) Möbius representation of the prediction function, enabling efficient computation of Shapley values and their associated uncertainty propagated from the GPs. We develop recursive algorithms, inspired by Newton's identities, that avoid exponential enumeration over feature subsets and compute the Shapley values for local and global explanation efficiently. The contributions of this paper can be summarized as follows:

1. **Explaining the uncertain.** We propose the *FGPX-Shapley-L* (**F**ANOVA **GP**-based e**X**act **Shapley** value for **L**ocal explanation) algorithm to provide local explanations in the form of stochastic Shapley values (Ma et al. 2008) for FANOVA GPs that can be computed in *quadratic* time through a recursive algorithm using Newton's identities.
2. **Explaining the uncertainty.** We propose the FGPX-Shapley-G (G for global) algorithm to attribute the variance of the prediction as global feature attribution, following the spirit in classical global sensitive analysis literature. A recursive, quadratic-time algorithm is also devised to ensure computational efficiency.

Taken together, our results show that interpretability, axiomatic attribution, and computational efficiency need not be at odds—provided that the model admits the right structure. FANOVA GPs, and the methods we develop to explain them, offer a new blueprint for scalable and rigorous explainable machine learning. Our implementation is publicly available at https://github.com/Majeed7/FGPX-Shapley.

2 Preliminaries

Notation. We denote the set of d features by $\mathcal{D}$, and its power set $2^{\mathcal{D}}$. The training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ consists of $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y \in \mathbb{R}$ (regression) or $y \in \{1, \dots, \ell\}$ (ℓ-class classification). Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the full input dataset, and $\mathbf{X}_{\mathcal{S}}$ its restriction to feature subset $\mathcal{S} \in 2^{\mathcal{D}}$; the corresponding sample space is $X_{\mathcal{S}}$. The probability density of feature i is denoted $p(X_i)$. We use calligraphic letters for sets, capital letters for random variables, and bold-faced lower- and upper-cased letters for vectors and matrices, respectively. Element-wise product is denoted by $\odot$, expectation over data and model f's predictive distribution is denoted by $\mathbb{E}_X$ and $\mathbb{E}_f$, and variances by $\mathbb{V}_X$ and $\mathbb{V}_f$.

FANOVA Gaussian process

We focus our exposition on regression tasks for clarity. Nonetheless, the proposed Shapley algorithms remain applicable to classification problems, provided that the posterior distribution of the FANOVA GP is available.

We assume observations y arise from a latent function $f(\boldsymbol{x})$ corrupted by Gaussian noise, and we place a Gaussian process prior on f. Following Duvenaud, Nickisch, and Rasmussen (2011), we enforce an additive structure on f:

$$f(\boldsymbol{x}) \;=\; \sum_{\mathcal{S} \subseteq \mathcal{D}} f_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}), \tag{1}$$

where each $f_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})$ depends only on the feature subset $\mathcal{S}$ of input $\boldsymbol{x}$. This additive formulation, referred to as the *functional decomposition* of f, enables the model to capture main effects and interactions of arbitrary order in a structured way. This structure is induced via an additive kernel:

$$k^{a_q}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_q^2 \sum_{1 \le i_1 \le \cdots \le i_d \le d} \left[\prod_{l=1}^{q} k_{i_l}(x_{i_l}, x'_{i_l}) \right], \tag{2}$$

which assigns scale parameter σ_q^2 to all q-way interactions. The actual GP is then built with the full additive kernel by summing over interaction orders $k^a(\boldsymbol{x}, \boldsymbol{x}') = \sum_{q=0}^{d} k^{a_q}(\boldsymbol{x}, \boldsymbol{x}')$, with $k^{a_0} = \sigma_0^2$. A GP with such a kernel is referred to as *additive Gaussian process (AGP)*. Although evaluating k^a involves $\binom{d}{q}$ terms per order q, Duvenaud, Nickisch, and Rasmussen (2011) showed that a recursion based on Newton's identities yields a polynomial-time algorithm, making the approach practical.

In short, by placing a GP prior $\mathcal{GP}(0, k^a)$ over the latent function f, with observation model $y = f + \epsilon$ where independent noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, the predictive posterior over $y_{\boldsymbol{x}}$ at a new input $\boldsymbol{x}$ is again a GP, i.e. $y_{\boldsymbol{x}} \mid \mathbf{X}, \boldsymbol{y} \sim \mathcal{GP}\big(\xi, \kappa\big)$, with the mean and covariance functions being expressed as:

$$\begin{aligned} &\xi(\boldsymbol{x}) = k^a(\boldsymbol{x}, \mathbf{X})^\top \boldsymbol{\alpha}, \boldsymbol{\alpha} = \Sigma^{-1} \boldsymbol{y}, \Sigma = k^a(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I, \\ &\kappa(\boldsymbol{x}, \boldsymbol{x}') = k^a(\boldsymbol{x}, \boldsymbol{x}') - k^a(\boldsymbol{x}, \mathbf{X})^\top \Sigma^{-1} k^a(\boldsymbol{x}', \mathbf{X}). \end{aligned} \tag{3}$$

One challenge with AGPs is identifiability, as many basis functions can sum to $f(\boldsymbol{x})$. Durrande et al. (2011) addressed this using the functional ANOVA decomposition (Hooker 2004) that imposes two key constraints: (i) each component satisfies the zero mean condition, $\mathbb{E}_{X_{\mathcal{S}}}[f_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})] = 0$ for every non-empty subset $\mathcal{S}$; and (ii) components are "mutually orthogonal" in the sense that $\mathbb{E}_X[f_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}) f_{\mathcal{S}'}(\boldsymbol{x}_{\mathcal{S}'})] = 0$ for any $\mathcal{S} \neq \mathcal{S}'$. Durrande et al. (2011) put forward a class of kernels that satisfy the above conditions. In particular, for any base kernel k_i, the *constrained kernel* $\tilde{k}_i$ is defined as:

$$\tilde{k}_i(x_i, x'_i) = k_i(x_i, x'_i) - \frac{\int k_i(x_i, s) p(s) ds \int k_i(x'_i, s) p(s) ds}{\int \int k_i(s,t) p(s) p(t) ds dt}, \tag{4}$$

with p some density defined over the data space. The higher-order kernels $\tilde{k}^{add_q}$ are then constructed by multiplying corresponding $\tilde{k}_i$ as in equation (2), and the additive constrained kernel is defined as $\tilde{k}^a(\boldsymbol{x}, \boldsymbol{x}') = \sum_{q=0}^{d} \tilde{k}^{a_q}(\boldsymbol{x}, \boldsymbol{x}')$. Durrande et al. (2011) showed that functions drawn from a GP with

this kernel satisfy the ANOVA decomposition conditions. As a follow-up, Lu, Boukouvalas, and Hensman (2022) demonstrated for several popular kernels, such as the squared exponential and categorical kernels, with a Gaussian density over features, $\tilde{k}_i$ admits an analytical expression; whereas for other densities or kernels, the integration in equation (4) could be estimated by the empirical probability measure based on the training samples. For the rest of the paper, we refer to the additive GP satisfying the FANOVA conditions as *FANOVA GP*.

(Stochastic) Shapley values

The Shapley value (SV) (Shapley 1953) is a solution concept from cooperative game theory that provides an axiomatic framework for fairly distributing the total value generated by a group back to its individual members. Its appeal lies in satisfying a unique set of desirable properties: efficiency, symmetry, dummy, and linearity. Formally, given a tuple $(\mathcal{D}, v)$, where $\mathcal{D}$ is the set of d players and $v : 2^{\mathcal{D}} \to \mathbb{R}$ is a real-valued set function, the SV for player $i \in \mathcal{D}$ is defined as a specific weighted average of their marginal contributions across all possible coalitions. The function $v(\mathcal{S})$ can be interpreted as the "worth" or value generated by the subset $\mathcal{S}$.

However, in many settings—such as probabilistic modeling or decision-making under uncertainty—the value function may not be deterministically specified, but rather known only up to a distribution. In such cases, it is natural to ask whether an analogue of the Shapley value exists. Indeed, it does. Ma et al. (2008) and Chau, Muandet, and Sejdinovic (2023) formalized the notion of stochastic value functions $\nu : 2^{\mathcal{D}} \to \mathcal{L}(\mathbb{R})$, which assign to each coalition $\mathcal{S}$ a real-valued probability distribution, making $\nu(\mathcal{S})$ a real-valued random variable, thereby capturing the uncertainty in the value or importance attributed to that subset. This extension gives rise to the stochastic Shapley value (SSV), which generalizes the classical formulation to settings where importance must be assessed in a probabilistic rather than deterministic manner. Formally, given player set $\mathcal{D}$ and stochastic value function ν, the SSV of player i is given by

$$\phi_i(\nu) = \sum_{\mathcal{S} \subseteq \mathcal{D} \setminus \{i\}} c_{|\mathcal{S}|} \Big(\nu(\mathcal{S} \cup \{i\}) - \nu(\mathcal{S}) \Big), \qquad (5)$$

where $c_{|\mathcal{S}|} = \frac{|\mathcal{S}|!(d-|\mathcal{S}|-1)!}{d!}$. This formula looks analogous to the classical Shapley value, which is not surprising as the classical value function is a special case of the stochastic counterpart (e.g. use a Dirac function), but note that $\phi_i(\nu)$ is now a random variable. A key advantage of the stochastic formulation is that it naturally allows one to compute higher-order uncertainty statistics—most notably, the variance $\mathbb{V}(\phi_i(\nu))$, which quantifies uncertainty in the resulting attributions themselves.

A novel stochastic Möbius representation. We extend the theory of stochastic cooperative games by introducing the stochastic Möbius representation, a concept that has not yet been explored in the existing literature. We first state our results formally:

Definition 1 (Stochastic Möbius representation) *Given a stochastic value function ν, the stochastic Möbius representation $\mu : 2^{\mathcal{D}} \to \mathcal{L}(\mathbb{R})$ is defined as:*

$$\mu(\mathcal{T}) = \sum_{\mathcal{S} \subseteq \mathcal{T}} (-1)^{|\mathcal{T}|-|\mathcal{S}|} \nu(\mathcal{S}).$$

The quantity $\mu(\mathcal{T})$ can be interpreted as the stochastic Möbius coefficient of coalition $\mathcal{T}$, representing its unique contribution to the stochastic value function. This representation yields the following result:

Proposition 2 *Given stochastic cooperative game ν and its stochastic Möbius representation μ, we have*

- *The stochastic Shapley value for player $i \in \mathcal{D}$ is $\phi_i(\nu) = \sum_{\mathcal{T} \subseteq \mathcal{D}, \mathcal{T} \ni i} |\mathcal{T}|^{-1} \mu(\mathcal{T})$, and*
- *the variance of $\phi_i(\nu)$ naturally follows as $\mathbb{V}\big(\phi_i(\nu)\big) = \sum_{\substack{\mathcal{T} \subseteq \mathcal{D} \\ \mathcal{T} \ni i}} \sum_{\substack{\mathcal{T}' \subseteq \mathcal{D} \\ \mathcal{T}' \ni i}} \frac{1}{|\mathcal{T}||\mathcal{T}'|} \mathrm{Cov}\big(\mu(\mathcal{T}), \mu(\mathcal{T}')\big)$.*

While the utility of this representation may not be immediately evident, we demonstrate in the next section that adopting the stochastic Möbius perspective leads to significant computational advantages. In particular, it enables efficient evaluation of both the mean and variance of the SSV.

3 Explaining the Locally Uncertain with Stochastic Shapley Values

Shapley value for local explanation. Owing to its desirable axiomatic properties, the Shapley value has garnered significant attention as a principled method for feature attribution in machine learning (Lundberg and Lee 2017). In the context of local explanations, a common approach treats input features as players in a cooperative game and defines a value function tailored to a given predictive model g and input $\boldsymbol{x}$. A widely used formulation is $v(\mathcal{S}) = \mathbb{E}[g(X) \mid X_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}] - \mathbb{E}[g(X)]$, which quantifies the importance of a feature subset $\mathcal{S} \subseteq \mathcal{D}$ by measuring the change in the model's expected output when features in $\mathcal{D} \setminus \mathcal{S}$ are marginalized out. Intuitively, this reflects the added predictive value of observing $X_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}$ compared to having no feature information.

The GP-SHAP algorithm. To extend this to GPs, where predictions are probabilistic, Chau, Muandet, and Sejdinovic (2023) generalized the value function to a stochastic setting and proposed GP-SHAP. They showed that the conditional expectation of a GP remains stochastic and used the theory of conditional mean processes (Chau et al. 2021; Chau, Bouabid, and Sejdinovic 2021) to characterize the resulting stochastic value function analytically. While their approach applies to FANOVA GPs, it does not exploit the model's additive structure and relies on standard approximation techniques such as those used in Kernel SHAP (Lundberg and Lee 2017). Moreover, the conditional expectation-based value function incurs estimation error. In contrast, our approach advocates a different value function tailored to models with functional decomposition. This structure enables exact, estimation-free computation of both the value function and the stochastic Shapley values. Crucially, it also reduces the computational complexity of computing exact SSVs from exponential to quadratic time.

Stochastic Shapley values for FANOVA GP. When a predictive model f admits a functional decomposition, Fumagalli et al. (2025) and Mohammadi, Chau, and Muandet (2025) proposed an alternative value function for measuring subset contributions that aligns more closely with the sensitivity analysis literature. Specifically, for a decomposable f, an input $\boldsymbol{x}$, the (stochastic) value function ν is defined as:

$$\nu_{\boldsymbol{x}}(\mathcal{S}) = \sum_{\mathcal{T}\subseteq\mathcal{S}} f_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{T}}).$$

That is, the value of a subset $\mathcal{S} \subseteq \mathcal{D}$ is computed by summing all component functions whose indices are contained within $\mathcal{S}$, thereby capturing the total contribution of features in $\mathcal{S}$ to the overall prediction. This natural choice of value function not only preserves the tractability of GP models but also enables the development of a quadratic-time algorithm for computing Shapley values.

Proposition 3 *Given a posterior FANOVA GP $p(f \mid \boldsymbol{X}, \boldsymbol{y})$ defined in equation* (3)*, an input $\boldsymbol{x}$, the value function $\nu_{\boldsymbol{x}}$ defined above, we have that:*

- *$\nu_{\boldsymbol{x}}$ is a GP over $2^{\mathcal{D}}$ with mean and covariance functions:*

$$\xi_{\nu_{\boldsymbol{x}}}(\mathcal{S}|\boldsymbol{X},\boldsymbol{y}) = \sum_{\mathcal{T}\subseteq\mathcal{S}} \sigma^2_{|\mathcal{T}|}\tilde{k}_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{T}},\boldsymbol{X}_{\mathcal{T}})^\top\boldsymbol{\alpha},$$

$$\kappa_{\nu_{\boldsymbol{x}}}(\mathcal{S},\mathcal{S}'|\boldsymbol{X},\boldsymbol{y}) = \sum_{\mathcal{T}\subseteq\mathcal{S}\cap\mathcal{S}'} \sigma^2_{|\mathcal{T}|}\tilde{k}_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{T}},\boldsymbol{x}_{\mathcal{T}}) - \sum_{\mathcal{T}\subseteq\mathcal{S}}\sum_{\mathcal{T}'\subseteq\mathcal{S}'} \sigma^2_{|\mathcal{T}|}\sigma^2_{|\mathcal{T}'|}\tilde{k}_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{T}},\boldsymbol{X}_{\mathcal{T}})^\top\Sigma^{-1}\tilde{k}_{\mathcal{T}'}(\boldsymbol{x}_{\mathcal{T}'},\boldsymbol{X}_{\mathcal{T}'}).$$

- *the Möbius representation $\mu_{\boldsymbol{x}}$ is also a GP over $2^{\mathcal{D}}$ with mean and covariance functions:*

$$\xi_{\mu_{\boldsymbol{x}}}(\mathcal{S}|\boldsymbol{X},\boldsymbol{y}) = \sigma^2_{|\mathcal{S}|}\tilde{k}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}},\boldsymbol{X}_{\mathcal{S}})^\top\boldsymbol{\alpha},$$

$$\kappa_{\mu_{\boldsymbol{x}}}(\mathcal{S},\mathcal{S}'|\boldsymbol{X},\boldsymbol{y}) = \delta_{\mathcal{S}\mathcal{S}'}\sigma^2_{|\mathcal{S}|}\tilde{k}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}},\boldsymbol{x}_{\mathcal{S}}) - \sigma^2_{|\mathcal{S}|}\sigma^2_{|\mathcal{S}'|}\tilde{k}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}},\boldsymbol{X}_{\mathcal{S}})^\top\Sigma^{-1}\tilde{k}_{\mathcal{S}'}(\boldsymbol{x}_{\mathcal{S}'},\boldsymbol{X}_{\mathcal{S}'}),$$

  where $\delta_{\mathcal{S}\mathcal{S}'} = 1$ when $\mathcal{S} = \mathcal{S}'$, and 0 otherwise.

As Proposition 3 demonstrates, the Möbius representation $\mu_{\boldsymbol{x}}$ is neater to work with compared to $\nu_{\boldsymbol{x}}$. In fact, the analytical expressions and recursive algorithms for computing the mean and variance of the SSV for FANOVA GP are also more straightforward to derive from $\mu_{\boldsymbol{x}}$ than $\nu_{\boldsymbol{x}}$. We present the analytical expression of the resulting SSV under $\nu_{\boldsymbol{x}}$ in the following proposition.

Proposition 4 *Let $w_s = \sigma^2_s/s$ for some scaler s. The stochastic Shapley values $\boldsymbol{\phi}(\nu_{\boldsymbol{x}})$ follows a multivariate Gaussian distribution with mean $\xi_{\boldsymbol{\phi}} \in \mathbb{R}^d$ and covariance matrix $\mathbf{K}_{\boldsymbol{\phi}} \in \mathbb{R}^{d\times d}$, where*

$$\xi_{\boldsymbol{\phi}} = \sum_{\mathcal{S}\subseteq\mathcal{D},\mathcal{S}\ni i} w_{|\mathcal{S}|}\tilde{k}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}},\boldsymbol{X}_{\mathcal{S}})^\top\boldsymbol{\alpha},$$

and $[\mathbf{K}_{\boldsymbol{\phi}}]_{i,j} = \mathrm{Cov}(\phi_i(\nu_{\boldsymbol{x}}),\phi_j(\nu_{\boldsymbol{x}}))$ can be expressed as

$$\sum_{\mathcal{S}\subseteq\mathcal{D},\mathcal{S}\ni i}\sum_{\mathcal{T}\subseteq\mathcal{D},\mathcal{T}\ni j} \delta_{\mathcal{S}\mathcal{T}}\frac{\sigma^2_{|\mathcal{S}|}}{|\mathcal{S}|^2}\tilde{k}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}},\boldsymbol{x}_{\mathcal{S}}) - \Big(\sum_{\substack{\mathcal{S}\subseteq\mathcal{D}\\ \mathcal{S}\ni i}} w_{|\mathcal{S}|}\tilde{k}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}},\boldsymbol{X}_{\mathcal{S}})\Big)\Sigma^{-1}\Big(\sum_{\substack{\mathcal{T}\subseteq\mathcal{D}\\ \mathcal{T}\ni j}} w_{|\mathcal{T}|}\tilde{k}_{\mathcal{T}}(\boldsymbol{X}_{\mathcal{T}},\boldsymbol{x}_{\mathcal{T}})\Big).$$

The following lemma shows how the efficiency axiom of stochastic SV relates to the posterior mean function.

Lemma 5 *For a fixed input $\boldsymbol{x}$, let $\{\phi_j(\nu_{\boldsymbol{x}})\}_{j=1}^d$ be the stochastic Shapley values of the FANOVA GP and denote their posterior means by $\xi_{\phi_j}(\boldsymbol{x}) = \mathbb{E}[\phi_j(\boldsymbol{x}) \mid \boldsymbol{X},\boldsymbol{y}]$. Let $\xi(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x}) \mid \boldsymbol{X},\boldsymbol{y}]$ be the GP posterior mean and $\xi_\varnothing = \mathbb{E}[f_\varnothing \mid \boldsymbol{X},\boldsymbol{y}] = \sigma_0\boldsymbol{\alpha}^\top\mathbf{1}$ the posterior mean of the constant component. Then, $\sum_{j=1}^d \xi_{\phi_j}(\boldsymbol{x}) = \xi(\boldsymbol{x}) - \xi_\varnothing$.*

The FGPX-Shapley-L algorithm

In the following, we introduce one of our main contributions: the FGPX-Shapley-L (**F**ANOVA **GP**-based e**X**act **Shapley** value for **L**ocal explanation) algorithm, which relies on the following recursive formulation.

Quadratic-time recursive computation. A key challenge in computing the SSV is its exponential complexity: both the mean and variance involve summations over exponentially many subsets of features, making exact computation intractable for large d. The following theorem introduces a recursive formulation that enables computing the *exact* Shapley value in quadratic time, significantly reducing the computational burden. The recursion is based on Newton's identities—originally used to efficiently construct additive kernels—and here adapted for computing SSVs. Since the technical details of these identities (involving elementary symmetric polynomials and power sums) can be intricate, we defer their full definition and derivation to Appendix B. For the main text, we encourage readers to treat these constructions abstractly: they provide a principled way to summarize interaction terms compactly and recursively, allowing us to scale up exact computations without explicitly enumerating all feature subsets. We begin by presenting the recursion for computing the mean of the Shapley value.

Theorem 6 *Let $\boldsymbol{z}_j := \tilde{k}_j(x_j,\boldsymbol{X}_j)$ and define the set $\mathcal{Z}_{-i} = \{\boldsymbol{z}_j : j \in \mathcal{D}\setminus\{i\}\}$. Let the* elementary symmetric polynomials *(ESPs) over $\mathcal{Z}_{-i}$ be defined recursively as: $e_r(\mathcal{Z}_{-i}) = \frac{1}{r}\sum_{s=1}^r(-1)^{s-1}e_{r-s}(\mathcal{Z}_{-i})\odot p_s(\mathcal{Z}_{-i})$, $e_0(\mathcal{Z}_{-i}) = \mathbf{1}$, where $p_s(\mathcal{Z}_{-i}) = \sum_{\boldsymbol{z}\in\mathcal{Z}_{-i}}\boldsymbol{z}^s$ is the element-wise power sum. Define the intermediate vector:*

$$\boldsymbol{\ell}_i := \boldsymbol{z}_i \odot \sum_{q=0}^{d-1} w_{q+1}e_q(\mathcal{Z}_{-i}). \tag{6}$$

1. *The mean of the SSV for feature i is $\xi_{\phi_i} = \boldsymbol{\ell}_i^\top\boldsymbol{\alpha}$.*
2. *The variance of the SSV for feature i is given by:*

$$\mathbb{V}_f(\phi_i \mid \boldsymbol{X},\boldsymbol{y}) = \left(\bar{z}_i\sum_{q=0}^{d-1}\frac{\sigma^2_{q+1}}{(q+1)^2}\,e_q(\bar{\mathcal{Z}}_{-i})\right) - \boldsymbol{\ell}_i^\top\Sigma^{-1}\boldsymbol{\ell}_i, \tag{7}$$

where $\bar{z}_j := \tilde{k}_j(x_j,x_j)$, $\bar{\mathcal{Z}}_{-i} = \{\bar{z}_j : j \in \mathcal{D}\setminus\{i\}\}$, and $e_q(\bar{\mathcal{Z}}_{-i})$ is the ESPs of order q for set $\bar{\mathcal{Z}}_{-i}$.

The intuition to arrive at the recursion in equation (6) is to use the additive structure of the kernel function in FANOVA GP to factorize $\tilde{k}_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}},\mathbf{X}_{\mathcal{S}})$, bring $\tilde{k}_j(x_j,\mathbf{X}_j)$ out of the summation, and write the remaining part of the summation

as the weighted ESPs. This recursion is used to simplify the exact computation of the mean and variance of SSVs, and reduces the computational complexity from exponential to quadratic in the number of features. This represents a substantial improvement in scalability. Using the recursion in equation (6), the mean of SSVs can be computed directly. The variance requires an additional recursion, which we derive in equation (7). The time complexity of computing SSVs for a test instance using our method is $O(nd^2)$. Specifically, for each feature i, we remove its corresponding kernel vector to compute ESPs over the remaining $d-1$ features, which incurs a cost of $O(nd^2)$. For the variances of SSVs, the total cost remains $O(nd^2)$, since the recursion in the first term is of $O(d^2)$ complexity, and the second term is computed from $\boldsymbol{\ell}_i$'s with $O(nd^2)$ complexity, leading to an overall $O(nd^2)$ complexity.

To compute the full covariance structure of SSVs, a closely related recursive formulation with a similar complexity can be employed. The complete algorithm, along with a numerically stable implementation of the ESPs, is provided in Appendix C. In the next section, we introduce a recursive algorithm for computing variance-based Shapley values, aimed at providing global explanations.

4 Explaining the Global Uncertainty with Shapley Values

Besides local explanation, it is also important to quantify global feature importance to provide a holistic view of the overall predictive performance. A common approach to provide this is through global sensitivity analysis, where we compute the ratio of the variance corresponding to a feature set and the total variance. Durrande et al. (2011); Lu, Boukouvalas, and Hensman (2022) introduced the first-order Sobol index for FANOVA GP and provided an analytical solution under some conditions. However, as (Owen 2014) already argued, the first-order Sobol index only captures individual contributions and neglects interactions between features, limiting its ability to fully represent the importance of features in complex models. Ironically, in our case, since FANOVA GPs explicitly model feature interactions, relying on first-order Sobol indices that ignore these interactions is inappropriate. A possibility is to compute higher-order Sobol index for each feature set, but that becomes immediately impractical due to the exponential number of components.

Realizing these shortcomings, Owen (2014) advocated the use of Shapley values instead. Before we introduce the corresponding value function, we recall that FANOVA GP enjoys the following variance decomposition:

$$\mathbb{V}_X(f(\boldsymbol{x})) = \sum_{\mathcal{S}\subseteq\mathcal{D}} \mathbb{V}_X(f_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})). \tag{8}$$

Analogous to the local explanation setting, we define a variance-based value function $v : 2^{\mathcal{D}} \to \mathbb{R}$ over feature subsets as: $v_G(\mathcal{S}) := \sum_{\mathcal{T}\subseteq\mathcal{S}} \mathbb{V}_X(f_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{T}}))$, which quantifies the total contribution of the feature subset $\mathcal{S}$ to the output variance. This value function admits a Möbius representation $m_G : 2^{\mathcal{D}} \to \mathbb{R}$, given by $m_G(\mathcal{S}) = \mathbb{V}_X\big(f_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}})\big)$, where each Möbius coefficient represents the variance attributable to interaction subset $\mathcal{S}$. Given this setup, we now present a closed-form expression for the global Shapley value $\phi_i(v_G)$ under the posterior of the FANOVA GP model, which allows for efficient, recursive computation.

Theorem 7 *Let $\boldsymbol{\alpha}$ be the posterior mean as in Equation 3. Then, the global Shapley value of feature i based on the variance-based sensitivity analysis, shown by $\phi_i(v_G)$, is*

$$\phi_i(v_G) := \sum_{\substack{\mathcal{S}\subseteq\mathcal{D}\\ \mathcal{S}\ni i}} \frac{m_G(\mathcal{S})}{|\mathcal{S}|} = \boldsymbol{\alpha}^\top \Bigg(\sum_{\substack{\mathcal{S}\subseteq\mathcal{D}\\ \mathcal{S}\ni i}} \frac{\sigma^4_{|\mathcal{S}|}}{|\mathcal{S}|} \mathbf{L}_{\mathcal{S}} \Bigg) \boldsymbol{\alpha}, \tag{9}$$

where $\mathbf{L}_{\mathcal{S}} = \odot_{i\in\mathcal{S}} \int_{\mathcal{X}_i} \tilde{k}_i(x_i, \boldsymbol{X}_i) \tilde{k}_i(x_i, \boldsymbol{X}_i)^\top dp(x_i)$.

For all $i \in \mathcal{D}$, $\mathbf{L}_i$ can also be estimated with an empirical measure and proven to have an analytical solution with a Gaussian density of features (Lu, Boukouvalas, and Hensman 2022). Note that the Shapley value in equation (9) is still exponentially expensive to compute. We now present an efficient quadratic-time computation based on this equation.

Theorem 8 *Let $\mathcal{Z}_{-i} = \{\mathbf{L}_j : j \in \mathcal{D} \setminus \{i\}\}$, $p_s(\mathcal{Z}_{-i}) = \sum_{\mathbf{L}\in\mathcal{Z}_{-i}} \mathbf{L}^s$ denote the element-wise power-sum matrices, and define the ESPs recursively via Newton's identities: $e_0(\mathcal{Z}_{-i}) = \mathbf{1}, e_r(\mathcal{Z}_{-i}) = \frac{1}{r}\sum_{s=1}^{r}(-1)^{s-1} e_{r-s}(\mathcal{Z}_{-i}) \odot p_s(\mathcal{Z}_{-i})$. Then, the matrix $\boldsymbol{M}_i = \sum_{\mathcal{S}\ni i} \sigma^4_{|\mathcal{S}|}/|\mathcal{S}|\, \mathbf{L}_{\mathcal{S}}$ admits the closed-form recursion $\boldsymbol{M}_i = \mathbf{L}_i \odot \sum_{r=0}^{d-1} \sigma^4_{r+1}/r{+}1\, e_r(\mathcal{Z}_{-i})$. Consequently, the global Shapley value for feature i can be computed as $\phi_i(v_G) = \boldsymbol{\alpha}^\top \boldsymbol{M}_i \boldsymbol{\alpha}$.*

Similar to Lemma 5, we show how the efficiency axiom of SV relates to the variance of the FANOVA GP.

Lemma 9 *Let $f(\boldsymbol{x})$ be a function drawn from the FANOVA GP with observational noise $y = f(\boldsymbol{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. Then the total output variance decomposes as: $\mathbb{V}_X(y) - \sigma_n^2 = \sum_{i=1}^{d} \phi_i(v_G)$.*

5 Illustrations and Experiments

We empirically demonstrate our methods through (1) an illustration of stochastic explanations, (2) run-time comparisons, and (3) feature selection problems. We address the $O(n^3)$ complexity of training GPs through inducing point formulation. We employ the standard squared exponential kernel as our base kernel. Further details on experimental setups, data generation, model tuning, explainers, and feature selectors, along with additional experiments, are provided in Appendix D. The experiments were executed on a 24-core machine with 16GB of RAM and an RTX4000 GPU.

Illustration on how stochastic SV can provide richer explanations.

At this point, a sensible reader could ask: "What does uncertainty in explanation bring to the table?". We address this by providing a demo-use case of our method through the `energy efficiency` dataset from the UCI repository. The dataset consists of 768 samples and 8 input features describing the architectural and environmental characteristics

of buildings. The goal is to predict either the heating or cooling load of a building. We fit a FANOVA GP to the dataset and compute the SSVs for a single, randomly selected test point. Figure 1 shows the resulting SSVs, capturing both the estimated importance and the uncertainty of each feature.

SSVs offer a probabilistic view of feature importance, enabling richer analyses than deterministic approaches. Prior work has used their joint Gaussian structure to uncover dependencies between features or interpret acquisition functions in Bayesian optimization (Chau, Muandet, and Sejdinovic 2023; Adachi et al. 2024). Building on this, we propose a new application of SSVs that allows us to quantify uncertainty when comparing feature importance. The idea is straightforward: given two SSVs, ϕ_i and ϕ_j, a practitioner may wish to assess how likely it is that feature i is more important than feature j. This corresponds to estimating the probability $\mathbb{P}(|\phi_j| \geq |\phi_i|)$. Although this quantity is analytically intractable, the joint distribution of $[\phi_i, \phi_j]$ is Gaussian, making it straightforward to estimate via sampling. Figure 2 visualizes the uncertainty-aware comparison of feature importance as a directed graph for an arbitrary instace. We estimate the pairwise probabilities $\mathbb{P}(|\phi_i| \geq |\phi_j|)$ for all feature pairs using 1,000 samples from the joint distribution of SSVs. Each node represents a feature, and each directed edge reflects the likelihood that one feature is more important than another. The edge color and line style encode the strength of these probabilities, offering a probabilistic view of relative feature importance under uncertainty.

Time Comparison

We generated synthetic datasets with 8 to 100 features, each with 500 samples. For each dataset, we trained a FANOVA GP and randomly selected 30 test instances. We then computed SSVs using our method, FGPX-Shapley-L, and compared its average execution time to a naive implementation that computes only the mean of the Shapley values using the same Möbius-based value function. Figure 3 shows execution times across feature dimensions. Up to 12 features, both methods perform similarly (ours: 0.33s vs. naive: 1.35s). Beyond that, the naive method becomes impractical—taking over 2600 seconds at 22 features—while ours stays efficient (0.62s). At 25 features, the naive method failed to complete within 5 hours, but our method scaled to 100 features in just a few seconds. These results clearly demonstrate the scalability and computational advantage of FGPX-Shapley-L for efficient SSV computation.

Comparing local explanation methods through influential feature recovery

While feature attribution is inherently an unsupervised learning problem with no nontrivial objective ground truth, it is common to assess its quality via an influential feature recovery experiment. The following presents the evaluation of FGPX-Shapley-L for local explanations by comparing it in recovering the influential features with several state-of-the-art attribution methods, including Sampling SHAP (S-SHAP) (Štrumbelj and Kononenko 2014), Unbiased SHAP (U-SHAP) (Covert and Lee 2021), Bivariate

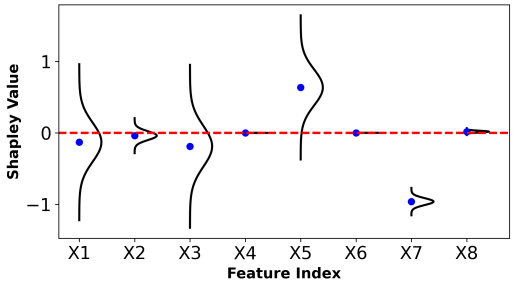

Figure 1: Local SSVs from the energy efficiency dataset. While mean importance is similar for $X1, X2$, the large explanation variance in $X1$ suggests we should calibrate our trust in the model's explanation.

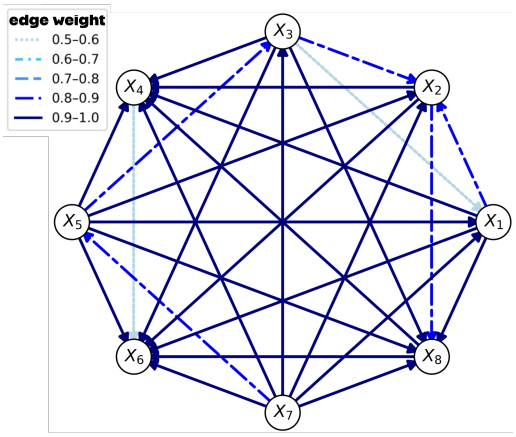

Figure 2: Directed graph of pairwise importance comparisons. An edge from feature i to j indicates $\mathbb{P}(|\phi_i| \geq |\phi_j|)$, with edge color and style encoding its strength.

SHAP (Bi-SHAP) (Masoomi et al. 2021), LIME (Ribeiro, Singh, and Guestrin 2016), and MAPLE (Plumb, Molitor, and Talwalkar 2018). We also implement a version of Kernel SHAP and GP-SHAP, where the value function in Proposition 3 is used.

Synthesized Experiments. We assess the performance of FGPX-Shapley-L for feature selection using three synthesized datasets, each containing 20 features and 1,000 samples. These datasets are specifically designed with prede-

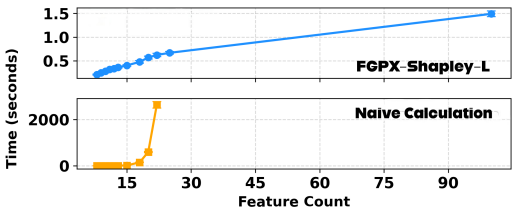

Figure 3: Comparison of execution times between FGPX-Shapley-L and the näive Shapley computation for different numbers of features, with a maximum time limit of five hours imposed for generating the explanations.

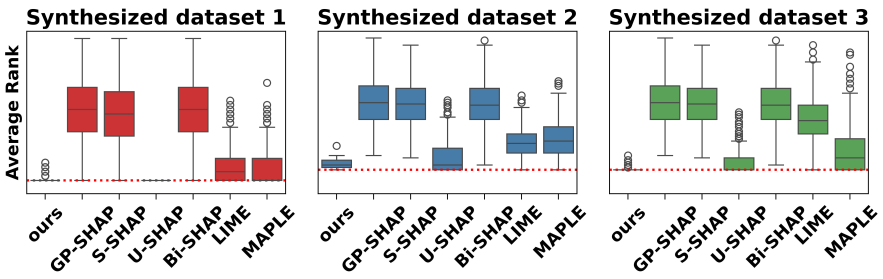

Figure 4: The comparison of explainable methods on four synthesized data sets.

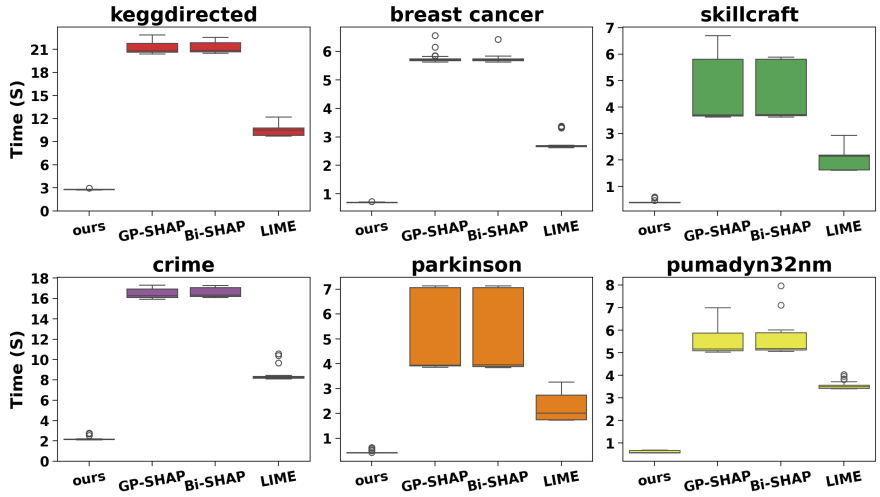

Figure 5: The comparison of different explanations in terms of execution time.

fined influential features and varying levels of interaction complexity. Details about the dataset generation process are provided in the appendix.

For each dataset, we randomly select 500 instances and generate explanations using FGPX-Shapley-L and baseline methods. The evaluation measures how accurately each method ranks the most influential features for individual instances. For each explanation, we rank the features by their assigned importance and compute the average rank of the ground-truth influential features across all instances. Figure 4 presents the average rank of the most influential features across 500 selected instances for three synthesized datasets. A lower average rank indicates better identification of influential features. The methodology for computing the average rank is explained in the appendix. The red dotted horizontal line in the figure represents the ideal average rank, corresponding to perfect identification of influential features. The figure demonstrates that FGPX-Shapley-L consistently delivers robust and accurate local explanations across all datasets. It outperforms other methods by reliably identifying the most influential features averaged over instances, highlighting its effectiveness in providing reliable local explanations.

Time Comparison. We assess the execution time of various explainable methods on real datasets. For this evaluation, 50 samples are randomly selected from six real datasets, and different methods are applied to generate local explanations from a trained FANOVA GP. Figure 5 presents a boxplot showing the average time (in seconds) required to explain an instance from these datasets. Due to the significantly longer execution times of MAPLE and S-SHAP, only the most competitive algorithms are included in the plot for a clearer comparison. A complete plot, including all methods, is provided in the appendix. FGPX-Shapley-L demonstrates outstanding performance in terms of execution time, significantly outperforming other methods, thanks to the proposed recursive algorithm. More experiments on real datasets are provided in the appendix.

6 Discussion

We presented exact algorithms for computing stochastic Shapley values in FANOVA Gaussian processes, enabling both local and global explanations that are axiomatic, uncertainty-aware, and computationally efficient. Our results demonstrate that when the model structure is appropriately designed—such as the orthogonal additive structure of FANOVA GPs—interpretability does not come at the cost of predictive performance. This work highlights that scalable, transparent, and probabilistically grounded explanations are achievable within expressive probabilistic models in quadratic time only.

Acknowledgments

This work was supported by funding from the Federal Ministry of Education and Research (BMBF). The first author gratefully acknowledges partial support from the DAAD Postdoc-NeT-AI short-term research visit scholarship.

References

Adachi, M.; Planden, B.; Howey, D.; Osborne, M. A.; Orbell, S.; Ares, N.; Muandet, K.; and Chau, S. L. 2024. Looping in the Human: Collaborative and Explainable Bayesian Optimization. In *International Conference on Artificial Intelligence and Statistics*, 505–513. PMLR.

Chau, S. L.; Bouabid, S.; and Sejdinovic, D. 2021. Deconditional Downscaling with Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 34, 17813–17825. Curran Associates, Inc.

Chau, S. L.; Muandet, K.; and Sejdinovic, D. 2023. Explaining the uncertain: Stochastic Shapley values for Gaussian process models. *Advances in Neural Information Processing Systems*, 36: 50769–50795.

Chau, S. L.; Ton, J.-F.; González, J.; Teh, Y.; and Sejdinovic, D. 2021. BayesIMP: Uncertainty Quantification for Causal Data Fusion. In *Advances in Neural Information Processing Systems*, volume 34, 3466–3477. Curran Associates, Inc.

Covert, I.; and Lee, S.-I. 2021. Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, 3457–3465. PMLR.

Durrande, N.; Ginsbourger, D.; Roustanta, O.; and Carraro, L. 2011. Reproducing kernels for spaces of zero mean functions. Application to sensitivity analysis. *stat*, 1050: 17.

Duvenaud, D. K.; Nickisch, H.; and Rasmussen, C. E. 2011. Additive Gaussian processes. In *Advances in neural information processing systems*, 226–234.

Fumagalli, F.; Muschalik, M.; Hüllermeier, E.; Hammer, B.; and Herbinger, J. 2025. Unifying Feature-Based Explanations with Functional ANOVA and Cooperative Game Theory. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Hooker, G. 2004. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 13(3): 755–770.

Lu, X.; Boukouvalas, A.; and Hensman, J. 2022. Orthogonal Additive Gaussian Processes. In *Proceedings of the 39th International Conference on Machine Learning*, 11956–11968.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ma, Y.; Gao, Z.; Li, W.; Jiang, N.; Guo, L.; et al. 2008. The shapley value for stochastic cooperative game. *Modern Applied Science*, 2(4): 1–76.

Masoomi, A.; Hill, D.; Xu, Z.; Hersh, C. P.; Silverman, E. K.; Castaldi, P. J.; Ioannidis, S.; and Dy, J. 2021. Explanations of Black-Box Models based on Directional Feature Interactions. In *International Conference on Learning Representations*.

Mohammadi, M.; Chau, S. L.; and Muandet, K. 2025. Computing Exact Shapley Values in Polynomial Time for Product-Kernel Methods. *arXiv preprint arXiv:2505.16516*.

Mohammadi, M.; Tiddi, I.; and Ten Teije, A. 2023. A Local Non-Additive Framework for Explaining Black-Box Predictive Models. In *ECAI 2023*, 1728–1738. IOS Press.

Mohammadi, M.; Tiddi, I.; and Ten Teije, A. 2025a. Support Vector-based Estimation of Multilinear Games for Feature Selection and Explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19520–19527.

Mohammadi, M.; Tiddi, I.; and Ten Teije, A. 2025b. Unlocking the game: Estimating games in möbius representation for explanation and high-order interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19512–19519.

Owen, A. B. 2014. Sobol'indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1): 245–251.

Plumb, G.; Molitor, D.; and Talwalkar, A. S. 2018. Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31.

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

Shapley, L. S. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317.

Sobol, I. M. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3): 271–280.

Štrumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41: 647–665.