GRANITE: A Generalized Regional Framework for Identifying Agreement in Feature-Based Explanations

Julia Herbinger
Leibniz Institute for Prevention Research and Epidemiology - BIPS

Gabriel Laberge
Polytechnique Montréal

Maximilian Muschalik
LMU Munich, MCML

Yann Pequignot
Université Laval à Québec

Marvin N. Wright
Leibniz Institute for Prevention Research and Epidemiology - BIPS
University of Bremen

Fabian Fumagalli
LMU Munich, MCML
Bielefeld University, CITEC

Abstract

Feature-based explanation methods aim to quantify how features influence the model's behavior, either locally or globally, but different methods often disagree, producing conflicting explanations. This disagreement arises primarily from two sources: how feature interactions are handled and how feature dependencies are incorporated. We propose GRANITE, a generalized regional explanation framework that partitions the feature space into regions where interaction and distribution influences are minimized. This approach aligns different explanation methods, yielding more consistent and interpretable explanations. GRANITE unifies existing regional approaches, extends them to feature groups, and introduces a recursive partitioning algorithm to estimate such regions. We demonstrate its effectiveness on real-world datasets, providing a practical tool for consistent and interpretable feature explanations.

1 INTRODUCTION

Black-box models, such as neural networks or ensembles, often achieve superior predictive performance but remain difficult to interpret, which is a major concern in high-stakes applications like healthcare or finance.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

Feature-based explanation methods aim to reveal how features influence predictions, either *locally*, for example, using SHAP values (Lundberg and Lee, 2017) or individual conditional expectation (ICE) curves (Goldstein et al., 2015), or *globally*, for example, using permutation or conditional feature importance (PFI and CFI) (Fisher et al., 2019; Strobl et al., 2008).

Yet, these methods frequently *disagree*, producing contradictory explanations for the same model (Krishna et al., 2024; Mitruț et al., 2024; Pirie et al., 2023; Schwarzschild et al., 2023). For example, Figure 1 (lower-left) shows different explanation methods assigning *conflicting importance to the same features* at the global level. Fumagalli et al. (2025) identify two main sources of such disagreement: (1) how methods handle *interactions*—e.g., *PredDiff* fully accounts for the interaction between X_1 and X_2, whereas *SHAP* methods include it only partially—and (2) how they incorporate *feature distributions* when features are correlated—e.g., the dependence of X_4 on X_3 is reflected in *conditional SHAP* and *PredDiff* but not in *marginal SHAP*. Differences along either dimension cause explanations to diverge, complicating interpretation. Importantly, in the absence of interactions and dependencies, explanations are *simple, additive, and consistent*, suggesting that reducing these effects improves interpretability. Figure 1 shows that restricting analysis to regions minimizing interactions (e.g., $X_2 = 1$), distribution influence (e.g., $X_3 = 1$), or both (e.g., $X_2 = 1$ & $X_3 = 1$), progressively aligns explanations, producing increasingly consistent and interpretable explanations.

This insight motivates our approach. We propose the **Generalized Regional frAmework for ideNtIfying agreemenT in feature-based Explanations**

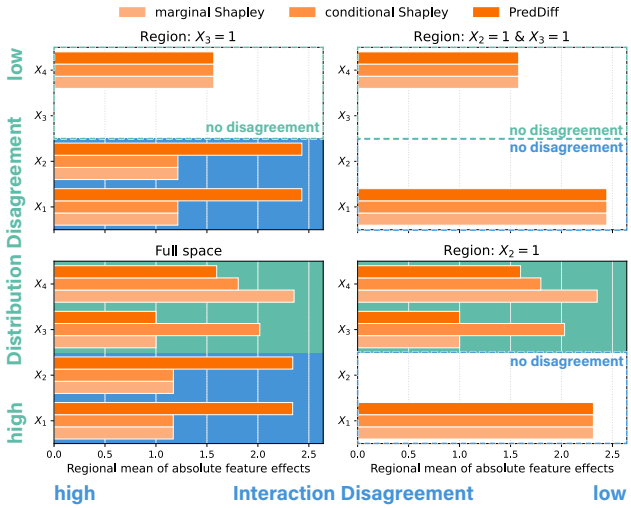


Figure 1: Disagreement of average absolute marginal and conditional SHAP (Lundberg and Lee, 2017) and PredDiff (Robnik-Sikonja and Kononenko, 2008) values for synthetic data, where X_1 interacts with X_2 and X_4 depends on X_3, cf. Section 3.1. Explanations disagree globally (bottom left) due to feature **interactions** and **dependencies**. Regional explanations remove disagreements caused by interactions (bottom right) and dependencies (top left), with overall agreement when combined (top right).

(GRANITE), which mitigates disagreement by partitioning the feature space into regions where interaction or distribution influence is minimized, yielding consistent and interpretable explanations.

Contributions. Our main contributions are:

1. Formalizing the disagreement problem for feature-based explanations, arising solely from differences in handling interactions and feature distributions.

2. The GRANITE framework, which partitions the feature space into regions minimizing disagreement to produce consistent explanations, implemented via recursive partitioning and demonstrated on real data.

3. Unifying existing regional approaches and addressing previously unexplored gaps.

2 BACKGROUND

General Notation We consider a prediction task over a feature space $\mathcal{X} \subseteq \mathbb{R}^d$ and target space $\mathcal{Y}$, with random variables $X = (X_1, \ldots, X_d)$ and Y. We write P for the reference distribution, which may denote the distribution of X or the joint distribution of (X, Y) depending on context (see Table 3). A (black box) model $F : \mathcal{X} \to \mathcal{Y}$ maps observed feature vectors $x \in \mathbb{R}^d$ to predictions. We denote the feature index set by $D := \{1, \ldots, d\}$ with power set 2^D. We write $i = \{i\}$, use $-S := D \setminus S$ for the complement of $S \subseteq D$, and x_S are the entries of x for features in S.

2.1 Functional Decomposition

Functional decomposition has a long history in statistics and has received increasing attention for interpreting machine learning models. Particularly, the standard functional ANOVA (fANOVA) decomposition (Stone, 1994; Hooker, 2004; Owen, 2013) provides for any *independent* joint distribution P and square-integrable prediction function F a decomposition:

$$F(x) = f_\emptyset + \sum_{i=1}^{d} f_i(x_i) + \sum_{i \neq j} f_{ij}(x_i, x_j) \cdots = \sum_{S \subseteq D} f_S(x),$$

where each component f_S does not depend on features in $-S$. Here, $f_\emptyset$ is constant, $f_i(x_i)$ is the i-th main effect, and $f_S(x)$ with $|S| > 1$ represents the *pure* interaction effect of features in S. The components f_S are defined by removing the dependency on features outside S via $F_S(x) := \mathbb{E}_{P_{X_{-S}}}[F(x_S, X_{-S})]$ with joint marginal distribution $P_{X_{-S}}$, and recursively subtract lower-order effects as

$$f_S(x) = F_S(x) - \sum_{T \subset S} f_T(x), \tag{1}$$

Following Fumagalli et al. (2025), we adopt a unifying framework for feature-based explanations that extends fANOVA and supports various masking operations F_S.

2.2 Unifying Framework

Fumagalli et al. (2025) introduced a unifying framework for feature-based explanations based on the fANOVA decomposition and the Möbius transform (Rota, 1964; Grabisch et al., 2000) from cooperative game theory. Based on their framework, explanations can be defined by four dimensions:

(1) Distribution influence: Are features *masked* via baseline value, marginal, or conditional distribution?
(2) Explanation game: Is the *behavior* of the explanation local or global?
(3) Interaction influence: Are interactions not, partially, or fully accounted for in the *influence* measure?
(4) Influence type: Are we interested in individual feature or feature group influence (joint / interaction)?

2.2.1 Definition of Feature Influence

Adapting the notation of Enouen and Liu (2025) we view feature-based explanations as the result of three consecutive operations: *masking*, *behavior*, and *interaction* operators. We link these operators to the first three dimensions of Fumagalli et al. (2025) for individual feature influences; extensions to feature groups are deferred to Appendix B and discussed in Section 3.4.

(1) Masking. The *masking* operator maps a predictor F to a family of predictors F_S with $S \subseteq D$, where

each F_S is restricted to features in S. To this end, the features outside of S are removed (masked) using a reference distribution P, which is either the conditional or marginal distribution. In addition, we consider a masking operator using a single baseline input $b \in \mathcal{X}$. The baseline, marginal, and conditional masking operations are increasingly influenced by dependencies in the feature distribution (Fumagalli et al., 2025).

Definition 1 (Masking)**.** *The **masking** operator $\mathcal{M}$ maps F to a family of masked models*

$$(\mathcal{M}F) : \mathcal{X} \times 2^D \to \mathbb{R}, \quad F_S(x) := (\mathcal{M}F)(x, S),$$

where $\mathcal{M}$ is chosen according to Table 1.

(2) Behavior. The *behavior* operator maps a family of masked models $\mathcal{M}F$ to a *value function* that quantifies the influence or importance for each feature subset.

Definition 2 (Behavior)**.** *The **behavior** operator $\mathcal{B}$ maps the masked models $F_\cdot \equiv \mathcal{M}F$ to a set function*

$$(\mathcal{B}(\mathcal{M}F)) : 2^D \to \mathbb{R}, \quad \nu(S) := (\mathcal{B}(\mathcal{M}F))(S),$$

where $S \subseteq D$, and $\mathcal{B}$ is chosen according to Table 1.

The influence (importance) of a subset $S \subseteq D$ is measured *locally* by the masked model prediction ($\mathcal{B}_{\text{loc},x_0}$), or *globally* by the variance ($\mathcal{B}_{\text{sens}}$) or performance ($\mathcal{B}_{\text{risk}}$) of the masked model.

(3) Interaction. On top of the *behavior* operator, we apply the *interaction* operator which quantifies explanations based on the influence of higher-order effects. Fumagalli et al. (2025) distinguishes between *pure*, *partial*, and *full* effects to measure the influence of individual features.

Definition 3 (Interaction)**.** *The **interaction** operator $\mathcal{I}$ maps $\nu \equiv \mathcal{B}(\mathcal{M}F)$ to an attribution function*

$$(\mathcal{I}(\mathcal{B}(\mathcal{M}F))) : D \to \mathbb{R}, \quad \phi(i) := (\mathcal{I}(\mathcal{B}(\mathcal{M}F)))(i),$$

where $i \in D$ and $\mathcal{I}$ is chosen according to Table 1.

Pure effects measure the influence of a feature in the *absence* of all other features and thereby ignore its role in higher-order effects. *Partial effects* are defined via the Shapley value (Shapley, 1953) of $\nu \equiv \mathcal{B}(\mathcal{M}F)$, which yields a fair distribution of interaction effects across involved features. *Full effects* measure the influence in the *presence* of all remaining features and hence incorporate higher-order effects entirely.

Interaction operators for *individual effects* generalize to *interactions* and *joint effects* for arbitrary subsets $\phi : 2^D \to \mathbb{R}$, cf. Appendix B. Examples of interaction and joint effects include the Shapley-Taylor interaction index (Sundararajan et al., 2020) and the grouped-PFI (Au et al., 2022).

Table 1: Overview of feature influence operators.

Concept	Operator	Definition
Masking Operators		
baseline	$\mathcal{M}_b$	$\mathbb{E}_{\delta_{b_{-S}}}[F(x_S, X_{-S})] = F(x_S, b_{-S})$
marginal	$\mathcal{M}_m$	$\mathbb{E}_{P_{X_{-S}}}[F(x_S, X_{-S})]$
conditional	$\mathcal{M}_c$	$\mathbb{E}_{P_{X_{-S}\mid X_S = x_S}}[F(x_S, X_{-S})]$
Behavior Operators		
local	$\mathcal{B}_{\text{loc},x_0}$	$F_S(x_0)$
sensitivity	$\mathcal{B}_{\text{sens}}$	$\mathbb{V}_{P_X}[F_S(X)]$
risk	$\mathcal{B}_{\text{risk}}$	$-\mathbb{E}_{P_{X,Y}}[\ell(F_S(X), Y)]$
Interaction Operators		
pure	$\mathcal{I}_{\text{pure}}$	$\nu(i) - \nu(\emptyset)$
partial	$\mathcal{I}_{\text{partial}}$	$\sum_{S \subseteq -i} \frac{1}{d\binom{d-1}{\lvert S\rvert}}[\nu(S \cup i) - \nu(S)]$
full	$\mathcal{I}_{\text{full}}$	$\nu(D) - \nu(-i)$

3 THE GRANITE METHOD

In this section, we define the disagreement problem of feature-based explanations and introduce the GRANITE method, which aims to identify regions in the feature space where explanations agree.

3.1 The Disagreement Problem

According to Fumagalli et al. (2025), feature-based explanations are defined by four dimensions: (1) feature distribution influence (*masking*), (2) explanation game (*behavior*), (3) influence of higher-order effects (*interaction*), and (4) type of influence measure (individual or groups of features). When comparing explanations, (2) and (4) are typically fixed to one category, since local and global, or single-feature and group-feature explanations describe fundamentally different aspects. This motivates our definition of the disagreement problem for feature-based explanations:

Definition 4 (Disagreement Problem)**.** *For a fixed influence measure (individual, joint, or interaction) and behavior $\mathcal{B}$, two feature-based explanations ϕ_1 and ϕ_2 **disagree** if and only if they differ due to the choice of **masking** $\mathcal{M}$ or **interaction** $\mathcal{I}$ (see Table 1).*

Illustration. We consider the toy example from Figure 1. Let $X_1 \sim \mathbb{N}(0, 1)$, $X_2, X_3 \in \{-1, 1\}$ with $P(X_2 = \pm 1) = P(X_3 = \pm 1) = 0.5$, and X_4 be conditionally distributed as $(X_4 \mid X_3 = x_3) \sim \mathcal{N}(x_3, 1)$. Let the model be $f(X) = 3X_1X_2 + X_3 + 2X_4$, and assume a local behavior $\mathcal{B} = \mathcal{B}_{loc,x_0}$. This setup illustrates two types of disagreement (Table 2). First, for X_1, the interaction $3X_1X_2$ induces opposing effects depending on X_2: Under marginal masking $\mathcal{M}_m$, the pure effect of X_1 averages to **zero**, while the full effect **fully** captures the interaction, creating a clear discrepancy. Second, for feature X_4, which enters the model linearly, disagreement arises between marginal

and conditional masking: Fixing $\mathcal{I} = \mathcal{I}_{pure}$ yields the linear term $2x_4$ for $\mathcal{M}_m$, while correlation between X_3 and X_4 introduces an additional nonlinear component $\mu_{3|4}$ under $\mathcal{M}_c$. Thus, even without interactions, feature dependencies can induce disagreement.

3.2 Definitions of Regional Explanations

To reduce disagreement, GRANITE proposes explanations with respect to *regions* $\Omega \subseteq \mathcal{X}$ of the feature space. To this end, the feature space $\mathcal{X}$ is partitioned,

$$\mathcal{X} = \bigcup_{k=1}^{K} \Omega_k, \quad \Omega_i \cap \Omega_j = \emptyset \text{ for } i \neq j,$$

into *regions* Ω_k. In practice, Ω_k is determined by recursive partitioning (e.g., decision trees), or any algorithm assigning samples to disjoint regions based on similarity, proximity, or axis-aligned constraints.

We define *regional explanations* $\phi_{|\Omega}$ as the result of the operators from Table 1 restricted to the region Ω. For this purpose, we restrict all *reference distributions* summarized in Table 3 to the support of Ω, denoted by P^{Ω}. A special case is the δ-distribution, where the baseline b must either be fixed (possibly outside of Ω) or depend on Ω. The restricted operators $\mathcal{M}^{\Omega}$ and $\mathcal{B}^{\Omega}$ are then defined by Definitions 1–2, respectively, and remain valid with P replaced by P^{Ω}. The interaction operator $\mathcal{I}$ remains unchanged. The individual regional influence measures are then defined by

$$\phi_{|\Omega}(i) := \left(\mathcal{I}\left(\mathcal{B}^{\Omega}(\mathcal{M}^{\Omega}F)\right)\right)(i), \quad i \in D. \tag{2}$$

3.3 Finding Agreement among Explanations

GRANITE seeks regions where two explanations

$$\phi_{1|\Omega} \equiv \mathcal{I}_1(\mathcal{B}^{\Omega}(\mathcal{M}_1^{\Omega}F)) \text{ and } \phi_{2|\Omega} \equiv \mathcal{I}_2(\mathcal{B}^{\Omega}(\mathcal{M}_2^{\Omega}F))$$

agree by minimizing their disagreement.

Definition 5 (Regional Disagreement)**.** *Let $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$. We define **regional disagreement** as*

$$\mathcal{R}_{\ell|\Omega}(\phi_{1|\Omega}, \phi_{2|\Omega}) := \mathbb{E}_{P_X^{\Omega}}\left[\sum_{i \in D} \ell\left(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)\right)\right],$$

where $\mathbb{E}_{P_X^{\Omega}}$ only affects local explanations with $\mathcal{B} = \mathcal{B}_{x,loc}^{\Omega}$ by averaging the distances for all $x \in \Omega$.

Examples for ℓ are the absolute or squared value. The regional disagreement extends directly to explanations with groups of features, i.e., for joint effects and interaction effects, as discussed in Section 3.4.

GRANITE aims to identify regions where two explanations agree, and every datapoint should be assigned exclusively to one region. This naturally leads to partitions $\{\Omega_1, \ldots, \Omega_K\}$ with disjoint regions $\Omega_1, \ldots, \Omega_K \subseteq \mathcal{X}$ that minimize the total regional disagreement:

$$\min_{\{\Omega_k\}_{k=1}^{K}} \sum_{k=1}^{K} \mathcal{R}_{\ell|\Omega_k}(\phi_{1|\Omega_k}, \phi_{2|\Omega_k}) \tag{3}$$

Given our explainability objective, an implicit constraint on partitions is that they should be interpretable and not include too many regions. In Section 3.5, we propose an algorithm to solve this task.

Depending on whether the *masking* or *interaction* operator differs, GRANITE minimizes the corresponding source of disagreement (interaction effects or feature distribution influence). The following paragraphs analyze the case of individual feature influence, with an extension to feature groups in Section 3.4.

Minimizing Interactions. For fixed *masking* and *behavior* operators, different *interaction* operators (e.g., pure, partial, full) may lead to different feature influence measures ϕ_1 and ϕ_2, with differences arising solely from higher-order effects.

Theorem 1. *Let $\mathcal{M}_1^{\Omega} = \mathcal{M}_2^{\Omega}$. Then for $i \in D$,*

$$\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i) = \sum_{S \subseteq D : i \in S, |S| \geq 2} \tilde{w}^S \Delta(S),$$

where $\tilde{w}^S \geq 0$ are specific interaction weights, and

$$\Delta(S) := \sum_{L \subseteq S} (-1)^{|S|-|L|} (\mathcal{B}^{\Omega}(\mathcal{M}_1^{\Omega}F))(L),$$

pure *interactions of the value function $\nu \equiv \mathcal{B}^{\Omega}(\mathcal{M}_1^{\Omega}F)$.*

This decomposition always exists, as it is derived from the Möbius transform of the cooperative game $\nu(S) := (\mathcal{B}(\mathcal{M}F))(S)$, with weights being the differences between the representations of each interaction operator $\mathcal{I}$ (Appendix A.1).

Illustration. In the example from Table 2, fixing $\mathcal{M} = \mathcal{M}_m$, the pure and full effects differ only in their interaction operator $\mathcal{I}$. For feature X_1, the disagreement $3x_1x_2$ reflects interaction effects. By Theorem 1, the disagreement is driven solely by higher-order terms, so GRANITE's partitioning along X_2 minimizes these interactions within regions, which leads to agreement between pure and full effects.

In general, the disagreement between explanations changes when using different behavior operators. However, under feature independence and squared error ℓ, the disagreement of full and pure local explanations is exactly the difference of the sensitivity explanations.

Theorem 2. *Let $\ell(x) := x^2$, and consider ϕ_{pure}^{loc},ϕ_{full}^{loc},ϕ_{pure}^{sens}, and ϕ_{full}^{sens}, where the subscript indi-*

Table 2: Comparison of pure and full effects of features X_1 and X_4 under marginal ($\mathcal{M}_m$) and conditional ($\mathcal{M}_c$) distributions for the entire feature space $\Omega := \mathcal{X}$ and for regions defined by either X_2 or X_3 where $\Omega_{(x_i=j)} := \{x \in \mathcal{X} \mid x_i = j\}$. Also, $\mu_{3|4} := \mathbb{E}[X_3 \mid X_4 = x_4]$ and '-' indicates the feature effect in the given region is the same as in $\Omega := \mathcal{X}$.

X_i	Type	$\Omega := \mathcal{X}$		$\Omega_{(x_2=1)}$		$\Omega_{(x_2=-1)}$		$\Omega_{(x_3=1)}$		$\Omega_{(x_3=-1)}$	
		$\mathcal{M}_m$	$\mathcal{M}_c$	$\mathcal{M}_m$	$\mathcal{M}_c$	$\mathcal{M}_m$	$\mathcal{M}_c$	$\mathcal{M}_m$	$\mathcal{M}_c$	$\mathcal{M}_m$	$\mathcal{M}_c$
X_1	$\mathcal{I}_{\text{pure}}$	$\mathbf{0}$	0	$3x_1$	$3x_1$	$-3x_1$	$-3x_1$	-	-	-	-
	$\mathcal{I}_{\text{full}}$	$\mathbf{3x_1x_2}$	$3x_1x_2$	$3x_1$	$3x_1$	$-3x_1$	$-3x_1$	-	-	-	-
X_4	$\mathcal{I}_{\text{pure}}$	$\mathbf{2x_4}$	$\mathbf{2x_4 + \mu_{3\|4}}$	-	-	-	-	$2x_4 - 2$	$2x_4 - 2$	$2x_4 + 2$	$2x_4 + 2$
	$\mathcal{I}_{\text{full}}$	$2x_4$	$2x_4 + \mu_{3\|4}$	-	-	-	-	$2x_4 - 2$	$2x_4 - 2$	$2x_4 + 2$	$2x_4 + 2$

Table 3: Distribution P used in masking and behavior

Context	Variant	Distribution
Masking	marginal	$P_{X_{-S}}$
	conditional	$P_{X_{-S}\|X_S=x_S}$
Behavior	local/sensitivity	P_X
	risk	$P_{X,Y}$

cates the interaction operator $\mathcal{I}_{pure}, \mathcal{I}_{full}$, and the superscript the behavior operator $\mathcal{B}_{x_0,loc}, \mathcal{B}_{sens}$. Under feature independence, i.e. $\mathcal{M} = \mathcal{M}_m = \mathcal{M}_c$, we have

$$\mathcal{R}_{\ell|\Omega}(\phi^{loc}_{full|\Omega}, \phi^{loc}_{pure|\Omega}) = \sum_{i \in D} (\phi^{sens}_{full|\Omega}(i) - \phi^{sens}_{pure|\Omega}(i)).$$

Importantly, with feature dependencies or other interaction operators, squared local disagreement and sensitivity disagreement do not necessarily coincide; though both GRANITE objectives are variance-based, they can yield different explanations in practice.

Minimizing Distribution Influence. We now consider interaction operators $\mathcal{I} = \mathcal{I}_1 = \mathcal{I}_2$ and compare two regional explanations $\phi_{1|\Omega}$, $\phi_{2|\Omega}$ that differ only in their masking operators with $\mathcal{M}_1^\Omega = \mathcal{M}_c^\Omega$ and $\mathcal{M}_2^\Omega = \mathcal{M}_m^\Omega$. In this setting, the difference between regional explanations reflects the differences in conditional and marginal distribution caused by feature dependencies, which GRANITE aims to minimize:

Theorem 3. *Let $w^\Omega_{x_S} := \frac{dP^\Omega_{X_{-S}|X_S=x_S}}{dP^\Omega_{X_{-S}}}$ be well-defined with density dP, $\mathcal{M}_1^\Omega = \mathcal{M}_c^\Omega$, $\mathcal{M}_2^\Omega = \mathcal{M}_m^\Omega$ with $\mathcal{B}_1^\Omega = \mathcal{B}_2^\Omega = \mathcal{B}_{x,loc}$. Then, $\phi_{1|\Omega} - \phi_{2|\Omega} \equiv \mathcal{I}(\mathcal{B}^\Omega_{loc,x_0}\delta^\Omega)$, where*

$$\delta^\Omega(x, S) := \mathbb{E}_{P^\Omega_{X_{-S}}}[(w^\Omega_{x_S}(X_{-S}) - 1)F(x_S, X_{-S})].$$

In words, the disagreement is equal to the local explanation of δ^Ω, driven by the difference between conditional and marginal distribution. For $\mathcal{B}^\Omega_{sens}$ and $\mathcal{B}^\Omega_{risk}$, similar variants of δ^Ω are presented in Appendix A.3.

Illustration. In the example from Table 2, we fix $\mathcal{I} = \mathcal{I}_{\text{pure}}$ and compare the marginal and conditional effects for X_4, which differ only in $\mathcal{M}$. Across the feature space, the dependence between X_3 and X_4 introduces an extra term $\mu_{3|4}$ in the conditional effect. By Theorem 3, this disagreement stems solely from the masking distribution. GRANITE's partitioning along X_3 minimizes this difference within regions, aligning marginal and conditional effects.

3.4 Extension to Feature Groups

Definition 5 introduced disagreement for individual influence measures $\phi : D \to \mathbb{R}$. This disagreement generalizes to group explanations $\phi : \mathcal{S} \to \mathbb{R}$, where $\mathcal{S} \subseteq 2^D$, by summing over $\sum_{S \in \mathcal{S}}$ instead of $\sum_{i \in D}$. Following Fumagalli et al. (2025), we distinguish joint influence of a group from its interaction influence. Joint influence only minimizes interactions or distributional effects between groups, which is especially meaningful when features are naturally interpreted together—e.g., genetic (Lozano et al., 2009) or sensor (Chakraborty and Pal, 2008) data. Interaction influence instead targets only higher-order interactions or distributional effects. By restricting minimization to interactions above a chosen order, lower-order effects remain intact and interpretable, reducing the number of regions while preserving decomposability up to that order.

Corollary 1. *Theorems 1 and 3 extend directly to joint and interaction influence measures.*

Illustration. Let $X_i \sim \mathbb{N}(0, 1)$, $i = 1, \ldots, 4$, and $F(X) = (X_1 + 0.7X_1X_2) \cdot \mathbb{1}X_3 \geq 0 + (-X_1 + 0.7X_1X_4) \cdot \mathbb{1}X_3 < 0$. We apply GRANITE to the interaction influence of feature pairs, with local behavior $\mathcal{B}_{\text{loc},x_0}$ and marginal masking $\mathcal{M}_m$, restricting minimization to interactions above order two. On 2000 samples, GRANITE splits at $X_3 = 0$, yielding two regions where higher-order interactions vanish. The regional decompositions, $F_{|\Omega_l}(x) = -x_1 + 0.7x_1x_4$ and $F_{|\Omega_r}(x) = x_1 + 0.7x_1x_2$, remain interpretable and require fewer splits (Figure 2). This is especially useful in the presence of continuous interactions, where existing regional frameworks encounter limitations (Herbinger et al., 2022, 2024; Laberge et al., 2024). An analogous approach applies to feature dependencies, and a joint influence example is discussed in Appendix E.

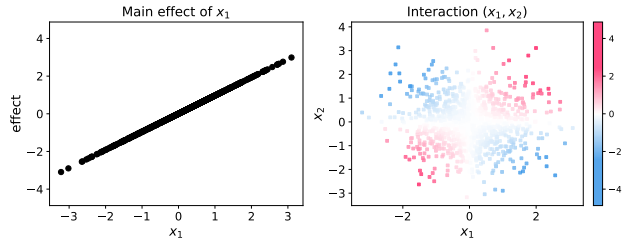


Figure 2: Main effect of x_1 (left) and its interaction with x_2 (right) in region Ω_r, where $F_{|\Omega_r}(x) = x_1 + 0.7\, x_1 x_2$.

.

3.5 Estimations

In practice, minimizing the disagreement via Equation (3) requires: 1) estimating explanations $\phi_{|\Omega}$ from data and 2) efficiently exploring the space of partitions.

Estimating Explanations. For computational efficiency, we consider regional explanations $\phi_{|\Omega} := \mathcal{I}(\mathcal{B}^{\Omega}(\mathcal{M}^{\Omega}F))$ with **pure** ($\mathcal{I}_{\text{pure}}$) and **full** ($\mathcal{I}_{\text{full}}$) interaction operators. This reduces evaluations of $\nu^{\Omega} = (\mathcal{B}^{\Omega}(\mathcal{M}^{\Omega}_{m}F))$ to the subsets $\emptyset, i, -i$, and D. While GRANITE can be used with $\mathcal{I} = \mathcal{I}_{\text{partial}}$, minimizing differences between pure and full effects, also reduces the disagreement with partial effects.

To estimate $\phi_{|\Omega}$, we first estimate P with the empirical distribution over N data points $\{x^{(n)}\}_{n=1}^{N}$. For $S \subseteq D$, we compute all possible combinations

$$R^{S}_{n\tilde{n}} = F(x^{(n)}_{S}, x^{(\tilde{n})}_{-S}), \quad n, \tilde{n} = 1, \ldots, N, \tag{4}$$

and estimate the marginal masking operator with

$$(\mathcal{M}^{\Omega}_{m}F)(x^{(n)}, S) = \frac{1}{|\{\tilde{n} : x^{(\tilde{n})} \in \Omega\}|} \sum_{x^{(\tilde{n})} \in \Omega} R^{S}_{n\tilde{n}}, \tag{5}$$

For conditional masking, we refer to the Appendix C. Due to the symmetry $\boldsymbol{R}^{-i} = (\boldsymbol{R}^{i})^{T}$, GRANITE only needs to compute $\{\boldsymbol{R}^{i}\}_{i=1}^{d}$ to evaluate $\nu^{\Omega}(i)$ and $\nu^{\Omega}(-i)$, since this symmetry is preserved when applying the behavior operators $\mathcal{B}^{\Omega}$ over the instances in Ω. Moreover, $\nu^{\Omega}(\emptyset)$ and $\nu^{\Omega}(D)$ are directly estimated from the predictions. Further details, including space complexity, are provided in the Appendix C.

Computing Partitions. Solving Equation (3) is intractable since the space of partitions of $\mathcal{X}$ is infinite. Existing regional explanation frameworks address this via clustering or recursive partitioning (Britton, 2019; Herbinger et al., 2024; Laberge et al., 2024; Molnar et al., 2023b). Clustering often yields non-interpretable regions (Herbinger et al., 2022), whereas recursive partitioning produces well-defined ones. GRANITE supports any method that identifies distinct regions under our risk constraint.

We adopt a recursive partitioning approach, constraining regions to be the leaves of a binary decision tree with bounded depth, where each leaf contains at least N_{min} samples. To minimize disagreement risk, the feature space is recursively split into two subregions, Ω_{-} and Ω_{+}, until any of these constraints is reached. A bottom-up pruning step based on a regularization term αK with $\alpha \geq 0$ and K being the number of leaves, undoes splits that do not sufficiently reduce risk.

To split along any of the d features, the algorithm considers at most B split candidates. For each candidate, regional explanations on Ω_{-} and Ω_{+} are computed using the pre-computed matrices $\boldsymbol{R}^{S}$, and the corresponding disagreement is evaluated. The optimal split is selected and the procedure repeats.

Computationally, at most $\mathcal{O}(2^{\text{max-depth}})$ internal nodes are evaluated, each with $\mathcal{O}(dB)$ split candidates. Since computing the disagreement risk is $\mathcal{O}(dN^2)$, the overall partitioning complexity is $\mathcal{O}(2^{\text{max-depth}} d^2 N^2 B)$. The full procedure is summarized in Appendix C.

4 REGIONAL UNIFICATION

Having introduced GRANITE, we now review and unify regional explanation methods.

Related Work. Prior research has shown that feature-based explanations disagree due to interactions or differences between marginal and conditional perturbations (Krishna et al., 2024; Herbinger et al., 2022; Chen et al., 2020). Existing regional frameworks focus on the reduction of disagreement caused by interactions. Heuristic approaches, such as VINE (Britton, 2019) and CohEx (Meng et al., 2025), cluster ICE curves or SHAP values. REPID (Herbinger et al., 2022) recursively partitions ICE curves with functional ANOVA guarantees, GADGET (Herbinger et al., 2024) generalizes this to other feature effect methods, and FD-Trees (Laberge et al., 2024) align PDP, SHAP, and PFI via functional decomposition. With GRANITE, we provide a principled approach via disagreement to minimize interactions within regions.

In contrast, methods targeting distributional effects are limited: transformation trees (Molnar et al., 2023b) reduce feature dependencies to align PDPs with MPlots and PFI with CFI. With GRANITE and our notion of regional disagreement, we provide a mathematically rigorous framework to identify regions that minimize feature dependencies.

Regional Unification in GRANITE. Above listed regional explanation frameworks are special cases of GRANITE. All minimize disagreement for individual measures but differ in targeting interactions or distributional influence (see Table 4). Below, we show how each fits into GRANITE, with details in Appendix D.

Table 4: Regional explanation frameworks and their unification within GRANITE, showing which disagreement types and behaviors each approach addresses. ✓: fulfilled, ✗: not fulfilled, $*$: all possible choices, –: not defined.

	Disagreement			Behavior				
Framework	$\mathcal{I}_{\text{full}} \to \mathcal{I}_{\text{pure}}$	$\mathcal{I}_{\text{partial}} \to \mathcal{I}_{\text{pure}}$	$\mathcal{M}_c \to \mathcal{M}_m$	$\mathcal{B}_{loc}$	$\mathcal{B}_{risk}$	$\mathcal{B}_{sens}$	$\mathcal{M}$	$\mathcal{I}$
VINE	✓	✗	–	✓	✗	✗	m	–
CohEx	✗	✓	–	✓	✗	✗	m	–
REPID	✓	✗	–	✓	✗	✗	m	–
GADGET	✓	✓	–	✓	✗	✗	m, c	–
FD-Trees	✓	✓	–	✓	✗	✗	m	–
Transf. Trees	–	–	✓	✓	✓	✗	–	full, pure
GRANITE	✓	✓	✓	✓	✓	✓	$*$	$*$

Minimizing Interactions: VINE and CohEx cluster local explanations under marginal masking: VINE minimizes full-to-pure and CohEx partial-to-pure. REPID, GADGET, and FD-Trees instead use recursive partitioning: REPID targets full-to-pure under marginal masking, GADGET and FD-Trees additionally reduce partial-to-pure, and GADGET also handles full-to-pure under conditional masking.

Minimizing Distribution Influence: transformation trees minimize conditional-to-marginal influence for pure local effects and full global risk behavior.

In summary, existing methods restrict to individual measures and one type of disagreement, each covering only certain behaviors, masking schemes, or interaction operators. GRANITE unifies these cases, closes the gaps for individual measures, and extends beyond them to joint and interaction influence measures.

5 EXPERIMENTS

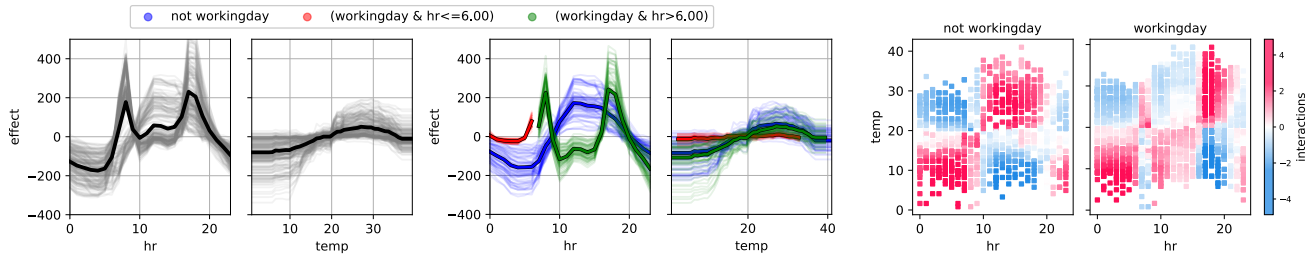
GRANITE is highly flexible: with three influence types (individual, joint, and interaction) and three behaviors, it provides nine settings to minimize interaction or distribution disagreement, each suited to different interpretability goals. Here, we illustrate several of these on the **Bikesharing** dataset, addressing key interpretability questions. Additional real-world examples, including interaction-focused analyses and joint influence measures, are provided in Appendix E; the full implementation and experimental code are available at `https://github.com/gablabc/GRANITE`. The Bikesharing dataset contains hourly bike rentals in the Capital Bikeshare system (2011–2012). We train a gradient boosting model on 80% of the data using ten temporal and weather-related features (see Appendix E for details). The remaining 20% serves as a test set. GRANITE is applied with different objectives on the first $N = 1000$ test samples to demonstrate how it adapts to distinct interpretability questions. Unless stated otherwise, the default hyperparameters $\alpha = 0.05, N_{\min} = 20, B = 40$ are used.

Simpler Model Explanations and More Interpretable Decompositions. A key challenge in model interpretation is that feature interactions complicate explanations. ICE plots reveal interactions but not their form or strength, while PDPs provide simple additive insights but ignore interactions, leading to incomplete understanding. The most interpretable explanations arise when the prediction function can be expressed primarily through low-order terms.

We demonstrate this on the Bikesharing data using **marginal local effects**. GRANITE minimizes interaction disagreement between **ICE (full effects)** and **PDP (pure effects)**, producing three regions with strongly reduced disagreement. In **Figure 3a**, partitioning by *workingday* (and *hour*) reveals distinct patterns for *hour*: bike use peaks during rush hours on working days but in the afternoon on non-working days, explaining much of the disagreement. Remaining disagreement reflects a three-way interaction between *hour*, *temperature*, and *workingday*, which reduces regionally to a continuous hour–temperature effect. To avoid excessive partitioning, we next limit GRANITE to **minimizing effects beyond second order**. It splits on *workingday* which isolates the hour–temperature interaction (**Figure 3b**). Temperature effects are lessened in the morning and night, amplified in the afternoon on non-working days, and amplified during commute times on working days.

Understanding the Role of Interactions in Global Feature Importance. Global feature importance can be sensitivity-based, measuring explained variance, or performance-based, assessing loss when a feature is removed. For both, rankings depend on how interactions are handled. Full measures such as PFI include interactions, obscuring the contribution of main effects and how importance varies across subgroups.

In the Bikesharing data, *workingday* contributes only via interactions, while *hour*, *temperature*, and *year* differ between **full and pure importance**, producing inconsistent global rankings based on **Sobol variances (marginal sensitivity; Figure 5)** (Sobol',



(a) Global and regional individual feature effects (b) Interaction effects

Figure 3: **Minimizing interaction disagreement** with $\mathcal{B}_{loc,x_0}$ and $\mathcal{M}_m$ for: **(a) Individual feature influence** with full (ICE curves as thin lines) versus pure (PDP as thick line) shown globally (left) and regionally (right) for *hour* and *temperature*. **(b) Interaction feature influence** of *hour* and *temperature* for regions depending on *workday*.

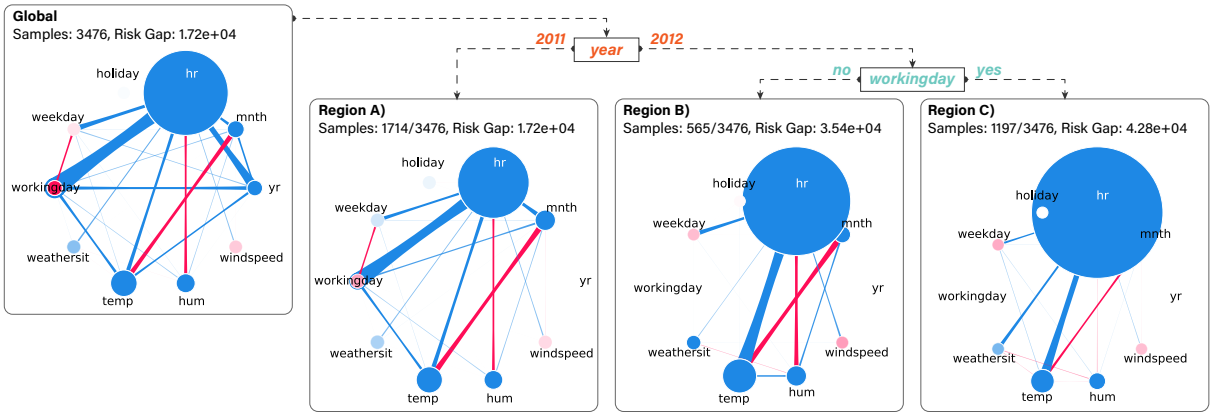


Figure 4: **Minimizing interaction disagreement** with $\mathcal{B}_{\text{risk}}$ and $\mathcal{M}_m$: GRANITE minimizes full ($\mathcal{I}_{\text{full}}$) versus pure ($\mathcal{I}_{\text{pure}}$) global risk, while the plots show first- and second-order partial risk (SAGE-like interactions) globally and regionally.

2001). GRANITE disentangles these effects by partitioning on *workingday* and early vs. daytime hours, revealing *which* features matter and *where*: early morning shows no importance, *temperature* dominates *year* on non-working days after morning, and *year* dominates on working days. Applied to **marginal full vs. pure risk (Figure 4)**, GRANITE shows performance differences e.g., small risk gaps for $yr = 2011$ vs. larger gaps for $yr = 2012$ on working days. First- and second-order Shapley interactions (Bordt and von Luxburg, 2023) extend SAGE (Covert et al., 2020) and reveal that regions mitigate hour–workingday/year interactions, while hour–temperature synergy remains strong, requiring both features for accurate predictions.

Explanations that Remain True to the Data. A key interpretability challenge is ensuring explanations remain faithful to the underlying data distribution. Conditional explanations account for feature dependencies but are often complex and computationally demanding, whereas marginal explanations are simpler but may misrepresent model behavior.

We illustrate this by comparing **pure local effects** using **MPlots** (Friedman, 2001; Apley and Zhu, 2020) **(conditional)** and **PDPs (marginal)** for *temperature* and *humidity* (**Figure 6**). Large discrepancies between marginal and conditional effects occur at low and high feature values, highlighting the impact of ignored dependencies in PDPs. GRANITE resolves these differences by partitioning the feature space based on *hour* and *temperature*, after which MPlot and PDP results align closely, producing regionally consistent marginal interpretations that remain faithful to both the model and the data distribution. A similar analyses is performed for CFI vs. PFI (full conditional versus full marginal risk) in Appendix E.

Quantitative Analysis. Building on the qualitative insights from the visualizations, we now quantify disagreement reduction on the Bikesharing dataset across different models (GBT, MLP, RF) and increasing region complexity, controlled by the maximum tree depth and setting $\alpha = 0.01$. We evaluate interaction disagreement using ICE vs. PDP (local, Figure 3a) and PFI vs. pure risk (global, Figure 4), as well as distributional disagreement using CFI vs. PFI and MPlot vs. PDP (masking, Figure 6). Across all settings, GRANITE consistently reduces disagreement, with results reported in Table 5. Disagreement decreases with depth and is most pronounced at Depth 3; for example, in the CFI vs. PFI setting, features become nearly uncorrelated within the learned regions, causing PFI and CFI to coincide in both value and interpretation.

Table 5: Disagreement reduction across depth levels on the Bikesharing dataset for interaction and distribution disagreements under different models and losses. Values report remaining disagreement (in %) relative to Depth 0 (entire feature space $\mathcal{X}$) on the test set, with regions learned on the training set. Results are averaged over 5 random seeds ($\pm$ std).

Target	Behavior	Model	Loss	Depth 0	Depth 1	Depth 2	Depth 3
$\mathcal{I}_{\text{full}}$ vs $\mathcal{I}_{\text{pure}}$	$\mathcal{B}_{loc}$	GBT	ICE vs PDP	100	30 ± 1.64	12 ± 1.22	4 ± 0.71
$\mathcal{I}_{\text{full}}$ vs $\mathcal{I}_{\text{pure}}$	$\mathcal{B}_{loc}$	MLP	ICE vs PDP	100	26 ± 2.88	10 ± 1.00	3 ± 0.55
$\mathcal{I}_{\text{full}}$ vs $\mathcal{I}_{\text{pure}}$	$\mathcal{B}_{loc}$	RF	ICE vs PDP	100	30 ± 1.67	11 ± 1.48	4 ± 0.71
$\mathcal{I}_{\text{full}}$ vs $\mathcal{I}_{\text{pure}}$	$\mathcal{B}_{risk}$	GBT	PFI vs Pure Risk	100	42 ± 5.76	16 ± 2.68	6 ± 0.84
$\mathcal{I}_{\text{full}}$ vs $\mathcal{I}_{\text{pure}}$	$\mathcal{B}_{risk}$	MLP	PFI vs Pure Risk	100	36 ± 1.82	12 ± 1.67	3 ± 0.71
$\mathcal{I}_{\text{full}}$ vs $\mathcal{I}_{\text{pure}}$	$\mathcal{B}_{risk}$	RF	PFI vs Pure Risk	100	41 ± 7.86	15 ± 1.14	4 ± 0.55
$\mathcal{M}_c$ vs $\mathcal{M}_m$	$\mathcal{B}_{loc}$	GBT	MPlot vs PDP	100	36 ± 3.70	16 ± 4.72	11 ± 2.55
$\mathcal{M}_c$ vs $\mathcal{M}_m$	$\mathcal{B}_{loc}$	MLP	MPlot vs PDP	100	37 ± 5.36	15 ± 3.08	11 ± 4.62
$\mathcal{M}_c$ vs $\mathcal{M}_m$	$\mathcal{B}_{loc}$	RF	MPlot vs PDP	100	37 ± 4.55	18 ± 3.70	16 ± 3.96
$\mathcal{M}_c$ vs $\mathcal{M}_m$	$\mathcal{B}_{risk}$	GBT	CFI vs PFI	100	26 ± 7.53	6 ± 1.41	3 ± 1.34
$\mathcal{M}_c$ vs $\mathcal{M}_m$	$\mathcal{B}_{risk}$	MLP	CFI vs PFI	100	21 ± 1.82	9 ± 2.41	2 ± 0.55
$\mathcal{M}_c$ vs $\mathcal{M}_m$	$\mathcal{B}_{risk}$	RF	CFI vs PFI	100	22 ± 1.64	10 ± 3.49	5 ± 3.78

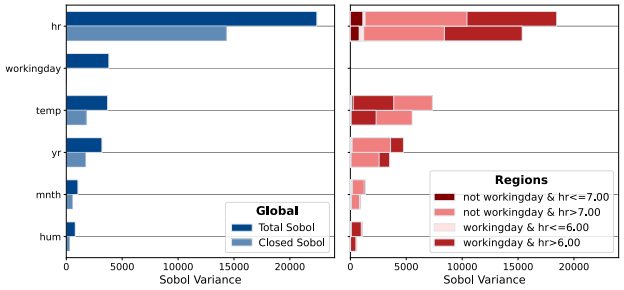


Figure 5: **Minimizing interaction disagreement** with $\mathcal{B}_{sens}$ and $\mathcal{M}_m$: Total Sobol' ($\mathcal{I}_{\text{full}}$, upper bar) and closed Sobol' ($\mathcal{I}_{\text{pure}}$, lower bar) shown globally (left) and regionally (right) for top-6 features. Stacked bars sorted by size for the regions create a layered view of regional rankings.

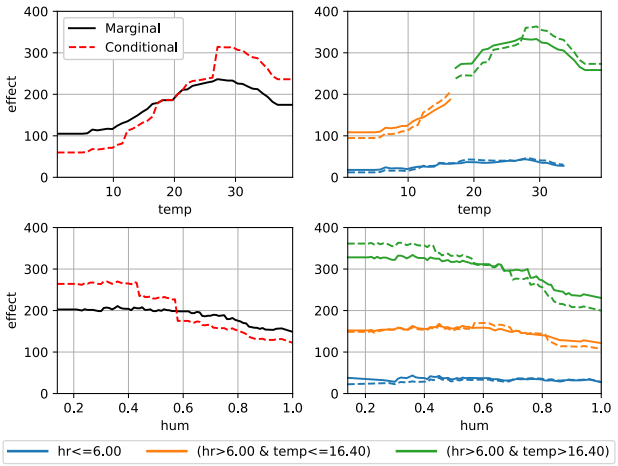


Figure 6: **Minimizing distribution disagreement** with $\mathcal{B}_{loc,x_0}$ and $\mathcal{I}_{pure}$ between MPlot ($\mathcal{M}_c$) and PDP ($\mathcal{M}_m$). Global plots (left) disagree, while regional plots (right) agree for *temperature* (top) and *humidity* (bottom).

6 DISCUSSION AND LIMITATIONS

We presented GRANITE, a principled framework to align feature-based explanations by minimizing disagreement of different choices of masking ($\mathcal{M}$) and interaction ($\mathcal{I}$) operators. Our theoretical results demonstrated that this disagreement is caused by interactions for $\mathcal{I}$ and feature dependencies for $\mathcal{M}$. GRANITE aligns explanations by efficiently partitioning the feature space into regions, and applies to local and global explanations ($\mathcal{B}$), as well as feature groups.

Limitations and Future Work. First, GRANITE works best when effects change abruptly, yielding a small number of clearly interpretable regions. When effects vary smoothly, many regions may be required, reducing interpretability. GRANITE partially mitigates this through explicit control of region complexity via the partitioning depth and by increasing expressiveness through feature groups (e.g., including pairwise interactions), which can reduce the number of required regions. Selecting such feature groups in a principled manner remains an important direction for future work. Second, we minimized interaction or distributional influences separately, but both can be addressed simultaneously, enabling low-order additive decompositions under independence. However, achieving full agreement may require many splits, highlighting a trade-off between interpretability and alignment. Third, memory usage grows quadratically with the number of data points used to estimate explanations. While sub-sampling 1000 points produced reasonable regions, future work could choose subsample sizes based on uncertainty in the regional explanations (Molnar et al., 2023a). Finally, while applied on tabular data, GRANITE can be extended to concepts for vision models or time-dependent prediction problems.

ACKNOWLEDGEMENTS

We gratefully thank the anonymous reviewers for their valuable feedback for improving this work. Julia Herbinger and Marvin N. Wright gratefully acknowledge funding by the German Research Foundation (DFG), Emmy Noether Grant 437611051. Maximilian Muschalik and Fabian Fumagalli gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 – 438445824. Yann Pequignot expresses appreciation for the support from the DEEL Project CRDPJ 537462-18 funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ), together with its industrial partners Thales Canada inc, Bell Textron Canada Limited, CAE inc and Bombardier inc.[1] The authors of this work take full responsibility for its content.

REFERENCES

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086.

Au, Q., Herbinger, J., Stachl, C., Bischl, B., and Casalicchio, G. (2022). Grouped feature importance and combined features effect plot. *Data Mining and Knowledge Discovery*, 36(4):1401–1450.

Bordt, S. and von Luxburg, U. (2023). From Shapley values to generalized additive models and back. In *International Conference on Artificial Intelligence and Statistics*, pages 709–745. PMLR.

Britton, M. (2019). Vine: Visualizing statistical interactions in black box models. *arXiv preprint arXiv:1904.00561*.

Chakraborty, D. and Pal, N. R. (2008). Selecting useful groups of features in a connectionist framework. *IEEE Transactions on Neural Networks*, 19(3):381–396.

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. (2020). True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems*, volume 33, pages 17212–17223. Curran Associates, Inc.

Enouen, J. and Liu, Y. (2025). InstaSHAP: Interpretable additive models explain Shapley values instantly. In *The Thirteenth International Conference on Learning Representations*.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Fujimoto, K., Kojadinovic, I., and Marichal, J. (2006). Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99.

Fumagalli, F., Muschalik, M., Hüllermeier, E., Hammer, B., and Herbinger, J. (2025). Unifying feature-based explanations with functional anova and cooperative game theory. In *International Conference on Artificial Intelligence and Statistics*, volume 258, pages 5140–5148. PMLR.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65.

Grabisch, M., Marichal, J.-L., and Roubens, M. (2000). Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178.

Herbinger, J., Bischl, B., and Casalicchio, G. (2022). Repid: Regional effect plots with implicit interaction detection. In *International Conference on Artificial Intelligence and Statistics*, pages 10209–10233. PMLR.

Herbinger, J., Wright, M. N., Nagler, T., Bischl, B., and Casalicchio, G. (2024). Decomposing global feature effects based on feature interactions. *Journal of Machine Learning Research*, 25(381):1–65.

Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 575–580.

Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., and Lakkaraju, H. (2024). The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*.

[1] https://deel.quebec

Laberge, G. (2024). *Trustworthy Additive Explanations of Machine Learning Models through Increased Alignment*. PhD thesis, Polytechnique Montréal.

Laberge, G., Pequignot, Y. B., Marchand, M., and Khomh, F. (2024). Tackling the xai disagreement problem with regional explanations. In *International Conference on Artificial Intelligence and Statistics*, pages 2017–2025. PMLR.

Lozano, A. C., Abe, N., Liu, Y., and Rosset, S. (2009). Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Marichal, J., Kojadinovic, I., and Fujimoto, K. (2007). Axiomatic characterizations of generalized values. *Discrete Applied Mathematics*, 155(1):26–43.

Meng, F., Liu, X., Kong, Z., and Chen, X. (2025). Cohex: A generalized framework for cohort explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19440–19448.

Mitruț, O., Moise, G., Moldoveanu, A., Moldoveanu, F., Leordeanu, M., and Petrescu, L. (2024). Clarity in complexity: how aggregating explanations resolves the disagreement problem. *Artificial Intelligence Review*, 57(12):338.

Molnar, C., Freiesleben, T., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., Wright, M. N., and Bischl, B. (2023a). Relating the partial dependence plot and permutation feature importance to the data generating process. In *World Conference on Explainable Artificial Intelligence*, pages 456–479. Springer.

Molnar, C., König, G., Bischl, B., and Casalicchio, G. (2023b). Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. *Data Mining and Knowledge Discovery*, pages 1–39.

Owen, A. B. (2013). Variance components and generalized Sobol' indices. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):19–41.

Pirie, C., Wiratunga, N., Wijekoon, A., and Moreno-Garcia, C. F. (2023). Agree: a feature attribution aggregation framework to address explainer disagreements with alignment metrics. In *CEUR Workshop Proceedings*, volume 3438.

Robnik-Sikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Rota, G.-C. (1964). On the foundations of combinatorial theory: I. Theory of Möbius functions. In *Classic Papers in Combinatorics*, pages 332–360. Springer.

Schwarzschild, A., Cembalest, M., Rao, K., Hines, K., and Dickerson, J. (2023). Reckoning with the disagreement problem: Explanation consensus as a training objective. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 662–678.

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):271–280.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):118 – 171.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9:1–11.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. (2020). The Shapley Taylor interaction index. In *International Conference on Machine Learning*, pages 9259–9268. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes, described in Sections 2 and 3.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes, described in Appendix C.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes, provided as supplementary material.**
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes, provided in Section 3 and Appendix A.**

(b) Complete proofs of all theoretical results. **Yes, provided in Appendix A.**

(c) Clear explanations of any assumptions. **Yes, provided in Sections 2 and 3 and Appendix A.**

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes.**

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.**

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.**

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Not Applicable.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. **Yes.**

   (b) The license information of the assets, if applicable. **Not Applicable.**

   (c) New assets either in the supplemental material or as a URL, if applicable. **Yes.**

   (d) Information about consent from data providers/curators. **Not Applicable.**

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. **Not Applicable.**

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

Supplementary Materials

A PROOFS

In this section, we provide detailed proofs of the results stated in the main paper.

A.1 Proof of Theorem 1

Proof. The value function of both regional explanations are defined by $\nu \equiv \mathcal{B}^\Omega(\mathcal{M}^\Omega F)$ with $\mathcal{M}^\Omega = \mathcal{M}_1^\Omega = \mathcal{M}_2^\Omega$. The pure interactions of the value function are given by

$$\Delta(S) := \sum_{L \subseteq S} (-1)^{|S|-|L|} \nu(L),$$

and for each interaction operator we obtain for $\phi_{|\Omega} \equiv \mathcal{I}\nu$ by Fumagalli et al. (2025)[Table 2]

$$\phi_{|\Omega}(i) = \Delta(i) + \sum_{S \subseteq D, i \in S, |S| \geq 2} w^S \Delta(S)$$

with

$$w^S := \begin{cases} 0, & \text{for } \mathcal{I} = \mathcal{I}_{\text{pure}}, \\ \frac{1}{|S|}, & \text{for } \mathcal{I} = \mathcal{I}_{\text{partial}}, \\ 1 & \text{for } \mathcal{I} = \mathcal{I}_{\text{full}}. \end{cases}$$

Therefore, we obtain for two regional explanations

$$\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i) = \sum_{S \subseteq D, i \in S, |S| \geq 2} \tilde{w}^S \Delta(S),$$

where $\tilde{w}^S := w_1^S - w_2^S$, where w_1^S, w_2^S are the weights associated with $\phi_{1|\Omega}, \phi_{2|\Omega}$, respectively. □

A.2 Proof of Theorem 2

Proof. We let $\ell(x) := x^2$, and compute regional explanations with respect to the local value function $\nu_x \equiv \mathcal{B}_{x,\text{loc}}(\mathcal{M}F)$ with $\mathcal{M} = \mathcal{M}_m = \mathcal{M}_c$. The difference between the two local regional explanations is given by the previous results in the proof of Theorem 1 as

$$\phi^{\text{loc}}_{\text{full}|\Omega}(i) - \phi^{\text{loc}}_{\text{pure}|\Omega}(i) = \sum_{S \subseteq D, i \in S, |S| \geq 2} \Delta_x(S),$$

since $\tilde{w}^S \equiv 1$ in this case, where we have added the dependency on x in Δ as a subscript. Due to feature independence, the pure interaction effects $\Delta_x(S)$ directly correspond to the fANOVA effects $f_S(x)$ in this case, cf (Fumagalli et al., 2025)[Corollary 1], which are are orthogonal and mean-centered yielding the variance de-

composition property (Owen, 2013). We can thus compute the regional disagreement as

$$\begin{aligned}\mathcal{R}_{\ell|\Omega}(\phi^{\text{loc}}_{\text{full}|\Omega}, \phi^{\text{loc}}_{\text{pure}|\Omega}) &= \mathbb{E}_{P_X}\left[\sum_{i\in D} \ell(\phi^{\text{loc}}_{\text{full}|\Omega}(i) - \phi^{\text{loc}}_{\text{pure}|\Omega}(i))\right] \\ &= \sum_{i\in D} \mathbb{E}_{P_X}\left[\ell(\sum_{S\subseteq D, i\in S, |S|\geq 2} \Delta(S))\right] \\ &= \sum_{i\in D} \mathbb{V}_{P_X}\left[\sum_{S\subseteq D, i\in S, |S|\geq 2} \Delta(S)\right] \\ &= \sum_{i\in D} \sum_{S\subseteq D, i\in S, |S|\geq 2} \mathbb{V}_{P_X}[\Delta(S)].\end{aligned}$$

On the other hand, we obtain for the sensitivity behavior $\mathcal{B}_{\text{sens}}$ with value function $\nu \equiv \mathbb{V}_{P_X}[\nu_X(\cdot)]$

$$\phi^{\text{sens}}_{\text{full}|\Omega}(i) - \phi^{\text{sens}}_{\text{pure}|\Omega}(i) = \mathbb{V}_{P_X}\left[\sum_{S\subseteq D, i\in S} \Delta_X(S)\right] - \mathbb{V}_{P_X}[\Delta_X(i)] = \sum_{S\subseteq D, i\in S, |S|\geq 2} \mathbb{V}_{P_X}[\Delta_X(S)],$$

where we have used $w^S \equiv 1$ for full effects, and orthogonality with mean-centeredness to apply the variance to the sum. We thus obtain

$$\mathcal{R}_{\ell|\Omega}(\phi^{\text{loc}}_{\text{full}|\Omega}, \phi^{\text{loc}}_{\text{pure}|\Omega}) = \sum_{i\in D} \sum_{S\subseteq D, i\in S, |S|\geq 2} \mathbb{V}_{P_X}[\Delta(S)] = \sum_{i\in D} \left(\phi^{\text{sens}}_{\text{full}|\Omega}(i) - \phi^{\text{sens}}_{\text{pure}|\Omega}(i)\right),$$

which concludes the proof. □

A.3 Proof of Theorem 3

Proof. Let $w^{\Omega}_{x_S} := \frac{dP^{\Omega}_{X_{-S}|X_S=x_S}}{dP^{\Omega}_{X_{-S}}}$ be well-defined with density dP, and $\mathcal{M}^{\Omega}_1 = \mathcal{M}^{\Omega}_c$ and $\mathcal{M}^{\Omega}_2 = \mathcal{M}^{\Omega}_m$. Moreover, consider the local behavior operator, i.e., $\mathcal{B}^{\Omega}_1 = \mathcal{B}^{\Omega}_2 = \mathcal{B}^{\Omega}_{x,\text{loc}}$. Due to additivity of $\mathcal{I}$ and $\mathcal{B}^{\Omega}_{x,\text{loc}}$, we directly obtain

$$\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i) = (\mathcal{I}(\mathcal{B}^{\Omega}_{loc,x_0}(\mathcal{M}^{\Omega}_c F - \mathcal{M}^{\Omega}_m F)))(i) \text{ for } i \in D.$$

Using $w^{\Omega}_{x_S}$, we can express the conditional masking operator in terms of the marginal masking operator as

$$\begin{aligned}(\mathcal{M}_c F)(x, S) &= \mathbb{E}_{P^{\Omega}_{X_{-S}|X_S=x_S}}[F(x_S, X_{-S})] = \int_{\Omega} F(x_S, z) dP^{\Omega}_{X_{-S}|X_S=x_S}(z) \\ &= \int_{\Omega} F(x_S, z) w^{\Omega}_{x_S}(z) dP^{\Omega}_{X_{-S}}(z) = \mathbb{E}_{P^{\Omega}_{X_{-S}}}[w^{\Omega}_{x_S}(X_{-S}) F(x_S, X_{-S})].\end{aligned}$$

Hence,

$$(\mathcal{M}_c F - \mathcal{M}_m F)(x, S) = \mathbb{E}_{P^{\Omega}_{X_{-S}}}[(w^{\Omega}_{x_S}(X_{-S}) - 1) F(x_S, X_{-S})] = \delta^{\Omega}(x, S).$$

For local behavior $\mathcal{B}_{x,\text{loc}}$, we thus obtain

$$\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i) = \mathcal{I}(\mathcal{B}^{\Omega}_{loc,x_0} \delta^{\Omega})(i) \text{ for } i \in D,$$

which concludes the first part of the proof.

Global sensitivity behavior. We now consider $\mathcal{B}^{\Omega}_1 = \mathcal{B}^{\Omega}_2 = \mathcal{B}^{\Omega}_{\text{sens}}$ and $\delta^{\Omega}_{\text{loc}} := \delta^{\Omega}$. We cannot directly obtain the desired result, since, in contrast to $\mathcal{B}^{\Omega}_{\text{loc}}$, $\mathcal{B}^{\Omega}_{\text{sens}}$ is not additive. We thus now express the variance of the conditionally masked model through the variance of the marginally masked model:

$$\begin{aligned}(\mathcal{B}^{\Omega}_{\text{sens}}(\mathcal{M}^{\Omega}_c F))(S) &= \mathbb{V}_{P^{\Omega}_{X_S}}[\mathbb{E}_{P^{\Omega}_{X_{-S}|X_S=x_S}}[F(X_S, X_{-S})]] = \mathbb{V}_{P_{X_S}}[\mathbb{E}_{P^{\Omega}_{X_{-S}}}[w^{\Omega}_{X_S}(X_{-S}) F(X_S, X_{-S})]] \\ &= \mathbb{V}_{P^{\Omega}_{X_S}}[\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S, X_{-S})] + \delta^{\Omega}_{\text{loc}}(X, S)] \\ &= \mathbb{V}_{P^{\Omega}_{X_S}}[\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S, X_{-S})]] + \mathbb{V}_{P^{\Omega}_{X_S}}[\delta^{\Omega}_{\text{loc}}(X, S)] + 2\text{Cov}(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S, X_{-S})], \delta^{\Omega}_{X,\text{loc}}(X, S)) \\ &= (\mathcal{B}^{\Omega}_{\text{sens}}(\mathcal{M}^{\Omega}_m F))(S) + \mathbb{V}_{P^{\Omega}_{X_S}}[\delta^{\Omega}_{\text{loc}}(X, S)] + 2\text{Cov}(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S, X_{-S})], \delta^{\Omega}_{X,\text{loc}}(X, S)).\end{aligned}$$

We then compute

$$(\mathcal{B}^{\Omega}_{\text{sens}}(\mathcal{M}^{\Omega}_{c}F))(S) - (\mathcal{B}^{\Omega}_{\text{sens}}(\mathcal{M}^{\Omega}_{m}F))(S) = \mathbb{V}_{P^{\Omega}_{X_S}}[\delta^{\Omega}_{\text{loc}}(X,S)] + 2\text{Cov}(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S,X_{-S})], \delta^{\Omega}_{X,\text{loc}}(X,S))$$
$$=: \delta^{\Omega}_{\text{sens}}(S).$$

Hence, the regional disagreement is characterized by the regional explanation of $\delta^{\Omega}_{\text{sens}}$, which is influenced by the difference between marginal and conditional distribution.

Global risk behavior. We now consider $\mathcal{B}^{\Omega}_1 = \mathcal{B}^{\Omega}_2 = \mathcal{B}^{\Omega}_{\text{risk}}$ and the squared loss $\ell(x) := x^2$. We again express the risk behavior of the conditionally masked model through the risk behavior of the marginally masked model:

$$\begin{aligned}
(\mathcal{B}^{\Omega}_{\text{risk}}(\mathcal{M}^{\Omega}_{c}F))(S) &= -\mathbb{E}_{P^{\Omega}_{X,Y}}[\ell(\mathbb{E}_{P^{\Omega}_{X_{-S}|X_S=X_S}}[F(X_S,X_{-S})],Y)] \\
&= -\mathbb{E}_{P^{\Omega}_{X,Y}}[\ell(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[w^{\Omega}_{X_S}(X_{-S})F(X_S,X_{-S})],Y)] \\
&= -\mathbb{E}_{P^{\Omega}_{X,Y}}[\ell(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S,X_{-S})] + \delta^{\Omega}_{\text{loc}}(X,S),Y)] \\
&= -\mathbb{E}_{P^{\Omega}_{X,Y}}[(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S,X_{-S})] + \delta^{\Omega}_{\text{loc}}(X,S) - Y)^2] \\
&= -\mathbb{E}_{P^{\Omega}_{X,Y}}[(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S,X_{-S})] - Y)^2 + \delta^{\Omega}_{\text{loc}}(X,S)^2 \\
&+ 2(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S,X_{-S})] - Y)\delta^{\Omega}_{\text{loc}}(X,S)] \\
&= (\mathcal{B}^{\Omega}_{\text{risk}}(\mathcal{M}^{\Omega}_{m}F))(S) - \mathbb{E}_{P^{\Omega}_{X,Y}}[\delta^{\Omega}_{\text{loc}}(X,S)^2 + 2(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S,X_{-S})] - Y)\delta^{\Omega}_{\text{loc}}(X,S)].
\end{aligned}$$

Hence, we can compute the difference as

$$(\mathcal{B}^{\Omega}_{\text{risk}}(\mathcal{M}^{\Omega}_{c}F))(S) - (\mathcal{B}^{\Omega}_{\text{risk}}(\mathcal{M}^{\Omega}_{m}F))(S) = -\mathbb{E}_{P^{\Omega}_{X,Y}}[\delta^{\Omega}_{\text{loc}}(X,S)^2 + 2(\mathbb{E}_{P^{\Omega}_{X_{-S}}}[F(X_S,X_{-S})] - Y)\delta^{\Omega}_{\text{loc}}(X,S)]$$
$$=: \delta^{\Omega}_{\text{risk}}(S).$$

In conclusion, the regional disagreement is characterized by the regional explanation of the value function $\delta^{\Omega}_{\text{risk}}$, which is influenced by the difference between marginal and conditional distribution. □

A.4 Proof of Corollary 1

Proof. Theorem 1 and Theorem 3 assumed a feature influence measure for individual features, i.e., $\phi : D \to \mathbb{R}$. We now extend these results to general influence measures defined for $\phi : \mathcal{S} \to \mathbb{R}$ with $\mathcal{S} \subseteq 2^D$.

A.4.1 Extension of Theorem 1

We distinguish between interaction effects and joint effects.

Extension to interactions. For *interaction effects*, the representation of $\phi_{|\Omega}$ for $S \subseteq D$ is directly extended according to Fumagalli et al. (2025)[Table 2] as

$$\phi_{|\Omega}(S) = \Delta(S) + \sum_{T \subseteq D, S \subset T} w^T \Delta(T)$$

with

$$w^S := \begin{cases} 0, & \text{for } \mathcal{I} = \mathcal{I}^{\text{interaction}}_{\text{pure}}, \\ \tau^{|S|}_{|T \setminus S|}, & \text{for } \mathcal{I} = \mathcal{I}^{\text{interaction}}_{\text{partial}}, \\ 1 & \text{for } \mathcal{I} = \mathcal{I}^{\text{interaction}}_{\text{full}}, \end{cases}$$

where τ is a index-specific weight. Consequently, the difference between two regional explanations is again characterized by

$$\phi_{1|\Omega}(S) - \phi_{2|\Omega}(S) = \sum_{T \subseteq D, S \subset T} \tilde{w}^T \Delta(T)$$

where $\tilde{w}^T := w^T_1 - w^T_2$, where w^T_1, w^T_2 are the weights associated with $\phi_{1|\Omega}, \phi_{2|\Omega}$, respectively.

Extension to joint effects. Lastly, for *joint effects*, we obtain the representation according to (Fumagalli et al., 2025)[Table 2] as

$$\phi_{|\Omega}(S) = \sum_{L \subseteq S, L \neq \emptyset} \Delta(L) + \sum_{T \subseteq D, S \cap T \notin \{\emptyset, S, T\}} w^T \Delta(T)$$

with

$$w^S := \begin{cases} 0, & \text{for } \mathcal{I} = \mathcal{I}^{\text{joint}}_{\text{pure}}, \\ \tau^{|S|}_{|T \setminus S|}, & \text{for } \mathcal{I} = \mathcal{I}^{\text{joint}}_{\text{partial}}, \\ 1 & \text{for } \mathcal{I} = \mathcal{I}^{\text{joint}}_{\text{full}}, \end{cases}$$

Therefore, for two regional explanations with joint effects

$$\phi_{1|\Omega}(S) - \phi_{2|\Omega}(S) = \sum_{T \subseteq D, T \cap S \notin \{\emptyset, S, T\}} \tilde{w}^T \Delta(T)$$

where $\tilde{w}^T := w_1^T - w_2^T$, where w_1^T, w_2^T are the weights associated with $\phi_{1|\Omega}, \phi_{2|\Omega}$, respectively. Note that in case of joint effects, the difference of regional explanations is also influenced by lower-order interactions ($|T| < |S|$) that are not fully contained in S, i.e., lower-order interactions of features between groups.

A.4.2 Extension of Theorem 3

In the proof of Theorem 3, we have used the additivity of the interaction operator $\mathcal{I}$. For joint effects and interaction effects, corresponding to generalized values (Marichal et al., 2007) and cardinal-probabilistic interactions (Fujimoto et al., 2006), respectively, the additivity axiom is satisfied similar to the Shapley value. Therefore, all previous arguments apply in this case, and the regional disagreement is characterized by the regional explanation of the corresponding δ^{Ω}.

□

B DEFINITIONS FOR FEATURE GROUP EXPLANATIONS

In this section, we define the interaction operators for feature group explanations, encompassing both *joint* and *interaction* influence measures.

We begin by extending the notion of individual marginal influence for a feature $i \in D$ to the *joint feature influence* of a feature set $S \subseteq D$, defined as

$$\nu(T \cup S) - \nu(T),$$

which represents the marginal contribution of the feature group S given the presence of a feature set $T \subseteq -S$. Building on this, we introduce the *discrete derivative*, which generalizes the notion of pure (interaction) influence to any coalition T. This extends the definition from Theorem 1, where the pure effect was measured relative to the empty set (via the Möbius transform), to the case of an arbitrary coalition T.

Formally, the discrete derivative is defined as

$$\Delta(S, T) := \sum_{L \subseteq S} (-1)^{|S|-|L|} \nu(T \cup L),$$

where $T \subseteq -S$, $\nu(S) := (\mathcal{B}(\mathcal{M}F))(S)$ and $S \subseteq D$. The discrete derivative $\Delta(S, T)$ quantifies the *pure influence* of a feature group S in the presence of a feature set T. For $T = \emptyset$, it reduces to the Möbius transform (Grabisch et al., 2000), which captures the pure interaction influence and is used in Theorem 1, where $\Delta(S) = \Delta(S, \emptyset)$.

We now define the interaction operators based on the joint feature influence and the discrete derivative. Definitions based on the Möbius representation are provided in Fumagalli et al. (2025).

Table 6: Definitions of joint and interaction influence operators $\mathcal{I}$ for feature groups, expressed in terms of the joint feature influence and the discrete derivative. The weights $w_{S,T}$ denote index-specific coefficients used to aggregate partial operators across coalitions $T \subseteq -S$.

Type	**Operator**	**Definition**
joint	$\mathcal{I}^{\text{joint}}_{\text{pure}}$	$\nu(S) - \nu(\emptyset)$
joint	$\mathcal{I}^{\text{joint}}_{\text{partial}}$	$\sum_{T \subseteq -S} w_{S,T}(\nu(T \cup S) - \nu(T))$
joint	$\mathcal{I}^{\text{joint}}_{\text{full}}$	$\nu(D) - \nu(-S)$
interaction	$\mathcal{I}^{\text{interaction}}_{\text{pure}}$	$\Delta(S, \emptyset)$
interaction	$\mathcal{I}^{\text{interaction}}_{\text{partial}}$	$\sum_{T \subseteq D \subseteq -S} w_{S,T} \Delta(S, T)$
interaction	$\mathcal{I}^{\text{interaction}}_{\text{full}}$	$\Delta(S, -S)$

C ALGORITHM DETAILS

The GRANITE algorithm requires two empirical estimates: (1) an estimate of the explanations and (2) an estimate of the optimal partition of regions. Both estimates are discussed separately, followed by a scalability experiment.

C.1 Estimating Explanations

To compute $\phi_{|\Omega}$, we estimate the theoretical distribution P with the empirical distribution $\widehat{P}$ over N data points $\{x^{(n)}\}_{n=1}^N$. For $S \subseteq D$, we compute all possible $N \times N$ combinations

$$R^S_{n\tilde{n}} = F(x^{(n)}_S, x^{(\tilde{n})}_{-S}), \quad n, \tilde{n} = 1, \ldots, N, \tag{6}$$

which can be done in a model-agnostic fashion, although more efficient implementations are available for decision trees and Explainable Boosting Machines (Laberge, 2024, Chapter 5.4). Marginal masking using $\widehat{P}^\Omega$ is then computed as

$$(\mathcal{M}^\Omega_m F)(x^{(n)}, S) = \frac{1}{|\{\tilde{n} : x^{(\tilde{n})} \in \Omega\}|} \sum_{x^{(\tilde{n})} \in \Omega} R^S_{n\tilde{n}}. \tag{7}$$

For conditional masking, we assume that bins $b_1, b_2, \ldots, b_B$ have been pre-computed for x_S. If $S = \{i\}$, these bins may correspond to univariate quantiles of x_i. For $S = D \setminus \{i\}$, bins may instead correspond to the leaves of a Decision Tree trained to predict x_i from x_{-i} (Strobl et al., 2008). Let $t^{(i)}$ be the index of the bin that captures $x^{(i)}$, then the conditional masking is computed as

$$(\mathcal{M}^\Omega_c F)(x^{(n)}, S) = \frac{1}{C_i} \sum_{\tilde{n}: x^{(\tilde{n})} \in \Omega \,\wedge\, x^{(\tilde{n})}_S \in b_{t^{(i)}}} R^S_{n\tilde{n}}, \tag{8}$$

where $C_i = |\{\tilde{n} : x^{(\tilde{n})} \in \Omega \,\wedge\, x^{(\tilde{n})}_S \in b^{(i)}_t\}|$.

Behavior operators $\mathcal{B}^\Omega$ aggregate these masked evaluations for instances in Ω via $\nu^\Omega \equiv \mathcal{B}^\Omega(\mathcal{M}^\Omega_m F)$. By focusing on disagreement between **full** ($\mathcal{I}_{\text{full}}$) and **pure** ($\mathcal{I}_{\text{pure}}$) interaction operators, it is only necessary to evaluate $\nu^\Omega(D), \nu^\Omega(-i), \nu^\Omega(i), \nu^\Omega(\emptyset)$ for each feature $i = 1, 2, \ldots, d$. Since $\nu^\Omega(-i)$ can be derived from the transposed matrix $(\boldsymbol{R}^i)^T$, GRANITE requires only the matrices $\{\boldsymbol{R}^i\}_{i=1}^d$ to compute explanations.

The space complexity of explanation computations varies depending on the stage of the algorithm. First, fitting GRANITE (compute explanations to find the regions) requires $\mathcal{O}(dN^2)$ space complexity, which motivates subsampling (1000-2000 samples in our experiments). Second, applying a fitted GRANITE on new data is cheaper: If M regions share data evenly, explanation computation requires only $\mathcal{O}(N^2 d/M)$ space complexity. This allows us to use the full test sets when reporting the final regional explanations.

C.2 Estimating Partitions

GRANITE aims to identify a partition $\{\Omega_1, \ldots, \Omega_K\}$ of disjoint regions $\Omega_1, \ldots, \Omega_K \subseteq \mathcal{X}$ such that two explanations exhibit increased agreement when restricted to these regions. Simplifying the regional disagreement risk to $\mathcal{R}(\Omega) \equiv \mathcal{R}_{\ell|\Omega_k}(\phi_{1|\Omega_k}, \phi_{2|\Omega_k})$, we must solve

$$\min_{\{\Omega_k\}_{k=1}^K} \sum_{k=1}^K \mathcal{R}(\Omega_k). \tag{9}$$

Since the set of all possible partitions of $\mathcal{X}$ is infinite, it is necessary to simplify the problem. We do so by constraining the regions to be the leaves of a binary decision tree with bounded depth. Moreover, the leaves are constrained to contain at least $N_{\min}$ samples.

$$\min_{\{\Omega^{\text{tree}}_k\}_{k=1}^K : K \leq 2^{\text{max-depth}}, \widehat{P}(\Omega^{\text{tree}}_k) \geq N_{\min}/N} \sum_{k=1}^K \mathcal{R}(\Omega^{\text{tree}}_k) + \alpha K \tag{10}$$

This objective employs a hyperparameter $\alpha \geq 0$ to regularize the decision tree w.r.t. its number of leaves K. Taking inspiration from existing greedy strategies to learn decision trees (e.g. CART), we start from a single

region $\Omega = \mathcal{X}$ and split it into two sub-regions Ω_- and Ω_+ based on a univariate split $\Omega_-^{j,\gamma} = \Omega \cap \{x \in \mathcal{X} : x_j \leq \gamma\}$, $\Omega_+^{j,\gamma} = \Omega \cap \{x \in \mathcal{X} : x_j > \gamma\}$. Ideally, the univariate split would minimize disagreements

$$j^\star, \gamma^\star = \text{argmin}_{j \in D : \gamma \in \mathbb{R}} \mathcal{R}(\Omega_-^{j,\gamma}) + \mathcal{R}(\Omega_+^{j,\gamma}), \tag{11}$$

(ignoring regularization since any split introduces the same number of leaves). However, we cannot span across all possible values of $\gamma \in \mathbb{R}$, so we instead discretize the set $\{x_j^{(n)} : x^{(n)} \in \Omega\}$ into B bins, leading to a set $S_{j,B,\Omega}$ of $B-1$ bin edges, and solve

$$j^\star, \gamma^\star = \text{argmin}_{j \in D : \gamma \in S_{j,B,\Omega}} \mathcal{R}(\Omega_-^{j,\gamma}) + \mathcal{R}(\Omega_+^{j,\gamma}). \tag{12}$$

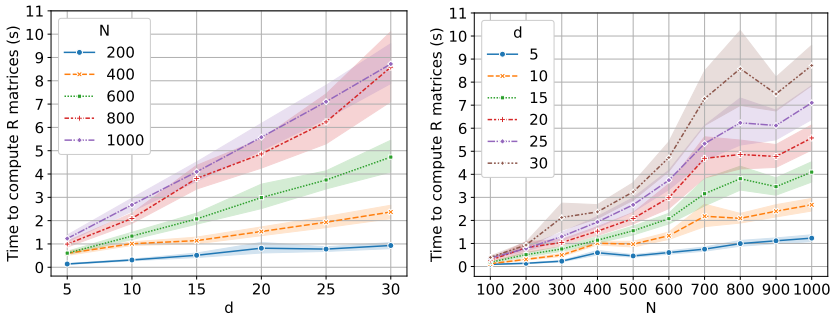
Once the optimal split according to Equation 12 is performed, the procedure is repeated recursively on the two sub-regions $\Omega_-^{j^\star,\gamma^\star}$ and $\Omega_+^{j^\star,\gamma^\star}$. The recursion terminates when any of the following conditions is met: the number of data points in a region falls below $2N_{\min}$ (as no further split could satisfy the constraints), the maximum tree depth is reached, or the risk falls below α (indicating that no further reduction in risk can compensate for the regularization). Once the recursion stops, the algorithm backtracks through each parent node to assess whether its split sufficiently decreased the objective relative to the number of leaves introduced. If not splitting the internal node results in a better objective, the internal node is converted into a leaf. This backtracking step is analogous to the pruning stage performed after greedy tree growth in CART. The full procedure is presented in Algorithm 1.

The method splits at most $\mathcal{O}(2^{\text{max-depth}})$ nodes, each requiring the evaluation of the risk over $\mathcal{O}(dB)$ split candidates. Since evaluating the disagreement risk (based on the $\{\boldsymbol{R}^i\}_{i=1}^d$ tensors) has a complexity of $\mathcal{O}(dN^2)$, the overall complexity of the partitioning algorithm is $\mathcal{O}(2^{\text{max-depth}} d^2 N^2 B)$.

Algorithm 1 GRANITE

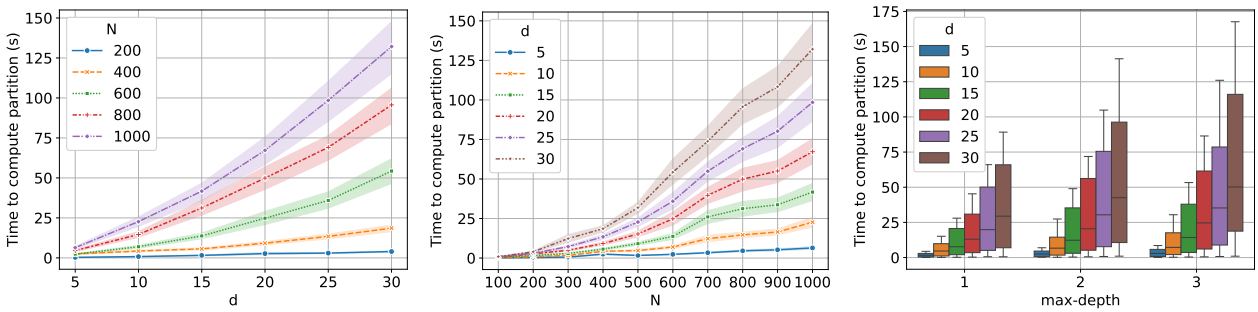
Require: Pre-computing the empirical distribution $\widehat{P}$ over a finite dataset $\{x^{(n)}\}_{n=1}^N$.
Require: Pre-computing the model evaluation matrices $\{\boldsymbol{R}^i\}_{i=1}^d$.
Require: Being able to compute the risk $\mathcal{R}(\Omega) \equiv \mathcal{R}_{\ell|\Omega}(\phi_{1|\Omega}, \phi_{2|\Omega})$ on any region.
Require: Regularization hyperparameter $\alpha \geq 0$ (default is $\alpha = 0.05$).
Require: Regularization hyperparameter $N_{\min}$ (default is $N_{\min} = 20$).
Require: Number of bins B (default is $B = 40$).

```
# This function is called recursively to return a tree structure.
procedure GRANITE(Ω, curr_depth, curr_risk)
    Initialize LeafNode.
    # If the objective cannot be increased further under the constraints, this node becomes a leaf.
    if curr_depth = max-depth or curr_risk < α or P̂(Ω) < 2N_min/N then
        return LeafNode, curr_risk, 1
    Initialize InternalNode
    Define S_{j,B,Ω} as the B − 1 bin edges resulting from binning {x_j^(n) : x^(n) ∈ Ω}.
    Define Ω_-^{j,γ} = Ω ∩ {x ∈ X : x_j ≤ γ}
    Define Ω_+^{j,γ} = Ω ∩ {x ∈ X : x_j > γ}
    j*, γ* = argmin_{j∈D:γ∈S_{j,B,Ω}} R(Ω_-^{j,γ}) + R(Ω_+^{j,γ})
    InternalNode.child_left, new_risk_left, n_leaves_left       = GRANITE(Ω_-^{j*,γ*}, curr_depth + 1, R(Ω_-^{j*,γ*}))
    InternalNode.child_right, new_risk_right, n_leaves_right = GRANITE(Ω_+^{j*,γ*}, curr_depth + 1, R(Ω_+^{j*,γ*}))
    # While backtracking, regularization may force us to let the current node be a leaf instead.
    if curr_risk + α < new_risk_left + new_risk_right + α(n_leaves_left + n_leaves_right) then
        return LeafNode, curr_risk, 1
    else
        return InternalNode, new_risk_left + new_risk_right, n_leaves_left + n_leaves_right
Tree, disagreement_risk, n_leaves = GRANITE(X, 0, 0)
```



(a) (b)

Figure 7: Runtimes for computing the $\{\boldsymbol{R}^i\}_{i=1}^d$ matrices. The theoretical complexity is $\mathcal{O}(dN^2)$.



(a) (b) (c)

Figure 8: Runtimes for computing the partition (i.e. Algorithm 1). The theoretical complexity is $\mathcal{O}(2^{\text{max-depth}} d^2 N^2 B)$.

C.3 Scalability Experiments

We investigate the scalability of explanations and partition computations on toy examples where the parameters N, d and `max-depth` are controlled. For simplicity, we sample N data points from a d-dimensional isotropic Gaussian and generate a random model that is piece-wise linear over the leaves of a randomly generated decision tree of depth `max-depth`. Figure 7 presents the required time to compute the $\boldsymbol{R}^i$ matrices for various values of N and d. We indeed observe a linear runtime for d and a quadratic one for N. This is not surprising seeing as the model has to be called dN^2 times. Also, note that it never takes more than 10 seconds to compute the matrices, even when studying 1000 samples with 30 features. These runtimes are acceptable but they could be further improved by using a model-specific implementation.

Figure 8 shows the runtime of the partitioning algorithm given the parameters d, N and `max-depth`. The hyperparameters α, N_{min} and B are fixed to their default value. We observe quadratic complexity w.r.t. d and N, while the expected exponential complexity w.r.t `max-depth` is not apparent. Indeed, we would expect the runtimes with `max-depth=3` to be 4 times higher than the runtimes of `max-depth=1`. We think this occurs because the regularization $\alpha = 0.05$, although small, restricts the algorithm from growing the tree unless reductions in risk are significant.

D FRAMEWORK UNIFICATION

In this section, we provide details on how existing regional frameworks are unified within GRANITE by showing that each of them minimizes a special case of the regional disagreement of Definition 4:

$$\mathcal{R}_{\ell|\Omega}(\phi_{1|\Omega}, \phi_{2|\Omega}) := \mathbb{E}_{P_X^\Omega}\left[\sum_{i\in D} \ell\left(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)\right)\right] \tag{13}$$

VINE (Britton, 2019): VINE clusters derivative ICE (d-ICE) curves (Goldstein et al., 2015) using Euclidean distance. This means regions are found by a clustering algorithm that minimizes the Euclidean distance between d-ICE curves and their cluster centroid being the derivative PDP (d-PDP). A d-ICE curve for the i-th feature and for a fixed observation $x^\star$ can be decomposed by functional ANOVA (see Section 2.1) into

$$\frac{\partial f(x_i, x^\star_{-i})}{\partial x_i} = \sum_{S\subseteq D} \frac{\partial f_S(x_i, x^\star_{-i})}{\partial x_i} = \sum_{S\subseteq D: i\in S} \frac{\partial f_S(x_i, x^\star_{-i})}{\partial x_i},$$

where the last step follows from the fact that all effects f_S independent of x_i become zero. Hence, d-ICE curves fully account for interactions with the feature of interest (x_i) and when considered for all observations together it recovers the marginal distribution of features in $-i$. Therefore, d-ICE curves are a **full** explanation method to measure **individual feature influence**, under **local behavior** $\mathcal{B}_{loc,x_0}$ and **marginal masking** $\mathcal{M}_m$.

The respective d-PDP is then defined by

$$\mathbb{E}_{X_{-j}}\left[\frac{\delta f(x_i, X_{-i})}{\delta x_i}\right] = \sum_{S\subseteq D: i\in S} \mathbb{E}_{X_{-j}}\left[\frac{\delta f_S(x_i, X_{-j})}{\delta x_i}\right],$$

where the influence of features in $-i$ is integrated out using their joint marginal distribution $P_{X_{-i}}$. Therefore, d-PDP is a **pure** explanation method to measure **individual feature influence**, under **local behavior** $\mathcal{B}_{loc,x_0}$ and **marginal masking** $\mathcal{M}_m$ (Fumagalli et al., 2025).

Hence, VINE minimizes the regional disagreement between d-ICE ($\phi_{1|\Omega}$) and d-PDP ($\phi_{2|\Omega}$), which are defined by

$$\phi_{1|\Omega}(i) = (\mathcal{I}_{\text{full}}(\mathcal{B}^\Omega_{loc,x_0}(\mathcal{M}^\Omega_m F)))(i) \quad \text{and} \quad \phi_{2|\Omega}(i) = (\mathcal{I}_{\text{pure}}(\mathcal{B}^\Omega_{loc,x_0}(\mathcal{M}^\Omega_m F)))(i),$$

using $\ell(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)) = \sqrt{(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i))^2}$ and the joint marginal distribution $P^\Omega_{X_{-i}}$.

It follows that VINE **minimizes interactions** for **individual feature influence** measures between **full** and **pure** effects under **local behavior** and **marginal masking** using a clustering algorithm. VINE only works for single features, i.e., we can only minimize interactions between one feature i and all other features $-i$. Hence, the disagreement risk is defined by

$$\mathcal{R}_{\ell|\Omega}(\phi_{1|\Omega}, \phi_{2|\Omega}) := \mathbb{E}_{P_X^\Omega}\left[\ell\left(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)\right)\right].$$

Applying VINE to different features is possible but will most likely lead to different clusters, complicating interpretation.

Cohex (Meng et al., 2025) : Cohex clusters local explanations by minimizing the mean squared error between local explanations and their cluster (regional) mean. While they argue that Cohex is applicable to many local explanation methods, they focus on marginal (interventional) SHAP values within the paper. Marginal SHAP is a **partial** explanation method (Fumagalli et al., 2025; Herbinger et al., 2024) to measure **individual feature influence**, under **local behavior** $\mathcal{B}_{loc,x_0}$ and **marginal masking** $\mathcal{M}_m$. The cluster (regional) mean of SHAP values per feature i integrates out the influence of features in $-i$ using their joint marginal distribution $P_{X_{-i}}$. Thus, the cluster mean is a **pure** explanation method to measure **individual feature influence**, under **local behavior** $\mathcal{B}_{loc,x_0}$ and **marginal masking** $\mathcal{M}_m$ (Herbinger et al., 2024).

Hence, Cohex minimizes the regional disagreement between marginal SHAP ($\phi_{1|\Omega}$) and its regional mean ($\phi_{2|\Omega}$), which are defined by

$$\phi_{1|\Omega}(i) = (\mathcal{I}_{\text{partial}}(\mathcal{B}^\Omega_{loc,x_0}(\mathcal{M}^\Omega_m F)))(i) \quad \text{and} \quad \phi_{2|\Omega}(i) = (\mathcal{I}_{\text{pure}}(\mathcal{B}^\Omega_{loc,x_0}(\mathcal{M}^\Omega_m F)))(i),$$

using $\ell(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)) = (\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i))^2$ and the joint marginal distribution $P^{\Omega}_{X_{-i}}$.

Cohex **minimizes interactions** for **individual influence measures** between **partial** and **pure** effects under **local behavior** and **marginal masking** using clustering. In contrast to VINE, Cohex directly minimizes the sum (L2 norm) of all features and thus minimizes interactions between all features as defined in Eq. (13).

REPID (Herbinger et al., 2022): REPID applies recursive partitioning to minimize the squared error between mean-centered ICE curves and mean-centered PDPs within each region for one feature of interest. Herbinger et al. (2022) and Fumagalli et al. (2025) show that mean-centered ICE curves **fully** account for interactions for the respective feature. Fumagalli et al. (2025) classifies single mean-centered ICE curves as **individual feature influence** measures, under **local behavior** $\mathcal{B}_{loc,x_0}$ and **baseline masking** $\mathcal{M}_b$. However, when all ICE curves within a region are considered, the values used for features in $-i$ reflect their joint regional marginal distribution $P^{\Omega}_{X_{-i}}$, thus considered together they reflect **marginal masking** $\mathcal{M}_m$. The mean-centered PDP is the average of the mean-centered ICE curves and integrates out the influence of features in $-i$ using their joint marginal distribution $P_{X_{-j}}$ (Herbinger et al., 2022). Thus, the regional mean-centered PDP is a **pure** explanation method to measure **individual feature influence** under **local behavior** $\mathcal{B}_{loc,x_0}$ and **marginal masking** $\mathcal{M}_m$ (Fumagalli et al., 2025).

Hence, REPID minimizes the regional disagreement between mean-centered ICE curves ($\phi_{1|\Omega}$) and mean-centered regional PDPs ($\phi_{2|\Omega}$), which are defined by

$$\phi_{1|\Omega}(i) = (\mathcal{I}_{\text{full}}(\mathcal{B}^{\Omega}_{loc,x_0}(\mathcal{M}^{\Omega}_m F)))(i) \quad \text{and} \quad \phi_{2|\Omega}(i) = (\mathcal{I}_{\text{pure}}(\mathcal{B}^{\Omega}_{loc,x_0}(\mathcal{M}^{\Omega}_m F)))(i),$$

using $\ell(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)) = (\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i))^2$ and the joint marginal distribution $P^{\Omega}_{X_{-i}}$.

Similarly to VINE, REPID **minimizes interactions** for **individual influence measures** between **full** and **pure** effects under **local behavior** and **marginal masking** for only one feature of interest. Thus, the disagreement risk is also given by

$$\mathcal{R}_{\ell|\Omega}(\phi_{1|\Omega}, \phi_{2|\Omega}) := \mathbb{E}_{P^{\Omega}_X}\left[\ell\left(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)\right)\right].$$

Compared to VINE, REPID uses a recursive partitioning algorithm instead of clustering, which has the advantage that it leads to interpretable regions in the feature space.

GADGET (Herbinger et al., 2024): GADGET minimizes the disagreement between any local feature effect method that can be decomposed into components h that contain only effects that depend on i, i.e.,

$$\phi_{1|\Omega}(i) = h_i(x_i) + \sum_{S \subseteq D: i \in S, |S| \geq 2} h_S(x),$$

and its regional mean, i.e.,

$$\phi_{2|\Omega}(i) = h_i(x_i) + \sum_{S \subseteq D: i \in S, |S| \geq 2} \mathbb{E}_{X_{-i}}[h_S(X)].$$

Since their approach is based on local effects, the explanations are defined by

$$\phi_{1|\Omega}(i) = (\mathcal{I}_1(\mathcal{B}^{\Omega}_{loc,x_0}(\mathcal{M}^{\Omega}_1 F)))(i) \quad \text{and} \quad \phi_{2|\Omega}(i) = (\mathcal{I}_{\text{pure}}(\mathcal{B}^{\Omega}_{loc,x_0}(\mathcal{M}^{\Omega}_2 F)))(i),$$

where $\mathcal{I}_1 \in \{\mathcal{I}_{\text{partial}}, \mathcal{I}_{\text{full}}\}$ and $\mathcal{M}_1 = \mathcal{M}_2$. The disagreement loss is defined by the L2 loss, i.e., $\ell(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)) = (\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i))^2$.

In Herbinger et al. (2024) they particularly derive it for the following three effect methods:

- **Mean-centered ICE curves and PDPs:** With $h_S(x) = f_S(x)$ and thus $\mathcal{I}_1 = \mathcal{I}_{\text{full}}$ and $\mathcal{M} = \mathcal{M}_m$.
- **Marginal SHAP values and Average SHAP Depedence:** With $h_S(x) = \frac{1}{|S|} f_S(x)$ and thus $\mathcal{I}_1 = \mathcal{I}_{\text{partial}}$ and $\mathcal{M} = \mathcal{M}_m$.
- **Local Derivatives and Accumulated Local Effects (ALE) Plots (Apley and Zhu, 2020):** With $h_S(x) = \frac{\delta}{\delta x_S} f_S(x)$ and thus $\mathcal{I}_1 = \mathcal{I}_{\text{full}}$ and $\mathcal{M} = \mathcal{M}_c$.

GADGET **minimizes interactions** for **individual influence measures** between **full** and **pure** as well as between **partial** and **pure** effects under **local behavior** and **marginal** and **conditional** masking (Herbinger et al., 2024; Fumagalli et al., 2025). While GADGET also uses recursive partitioning and the L2 loss comparably to REDPID, GADGET does consider all (or a subset of) features as defined in Eq. (13).

FD-Trees (Laberge et al., 2024): Similarly to GADGET, FD-Trees minimize disagreement between local feature effect methods that can be decomposed into the components h as defined above. However, compared to GADGET, the explanations in FD-Trees can take the following operator choices:

$$\phi_{1|\Omega}(i) = (\mathcal{I}_1(\mathcal{B}^{\Omega}_{loc,x_0}(\mathcal{M}^{\Omega}_m F)))(i) \quad \text{and} \quad \phi_{2|\Omega}(i) = (\mathcal{I}_2(\mathcal{B}^{\Omega}_{loc,x_0}(\mathcal{M}^{\Omega}_m F)))(i),$$

where $\mathcal{I}_1, \mathcal{I}_2 \in \{\mathcal{I}_{\text{pure}}, \mathcal{I}_{\text{partial}}, \mathcal{I}_{\text{full}}\}$ with $\mathcal{I}_1 \neq \mathcal{I}_2$. The disagreement loss is defined by the L2 loss, i.e., $\ell(\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i)) = (\phi_{1|\Omega}(i) - \phi_{2|\Omega}(i))^2$.

In Laberge et al. (2024) they particularly consider the following three disagreements:

- **PFI on Predictions (Mean-centered ICE curves) and PDPs:** With $h_S(x) = f_S(x)$ and thus $\mathcal{I}_1 = \mathcal{I}_{\text{full}}$ and $\mathcal{M}_1 = \mathcal{M}_m$ for ϕ_1 and ϕ_2 being the expectation w.r.t. features in $-i$ as defined above.
- **Marginal SHAP values and PDPs (Average SHAP Depedence):** With $h_S(x) = \frac{1}{|S|} f_S(x)$ and thus $\mathcal{I}_1 = \mathcal{I}_{\text{partial}}$ and $\mathcal{M}_1 = \mathcal{M}_m$ for ϕ_1 and ϕ_2 being the expectation w.r.t. features in $-i$ as defined above.
- **PFI on Predictions (Mean-centered ICE curves) and Marginal SHAP values:** With $h_S(x) = f_S(x)$ and thus $\mathcal{I}_1 = \mathcal{I}_{\text{full}}$ and $\mathcal{M}_1 = \mathcal{M}_m$ for ϕ_1 and $h_S = \frac{1}{|S|} f_S(x)$ and thus $\mathcal{I}_1 = \mathcal{I}_{\text{partial}}$ and $\mathcal{M}_1 = \mathcal{M}_m$.

FD-Trees **minimize interactions** for **individual influence measures** between **full** and **pure**, **full** and **partial** as well as between **partial** and **pure** effects under **local behavior** and **marginal** masking (Laberge et al., 2024; Fumagalli et al., 2025). Similarly to GADGET, FD-Trees also apply recursive partitioning and do consider all features as defined in Eq. (13).

Transformation Trees (Molnar et al., 2023b): Transformation Trees minimize the disagreement between the conditional distribution of the j-th feature $P_{X_j|X_{-j}=x_{-j}}$ and its marginal distribution P_{X_j} using recursive partitioning. In the paper, the authors argue that this approach minimizes the disagreement between CFI (conditional) and PFI (marginal), as well as between MPlot (conditional) and PDP (marginal).

More generally, this approach minimizes disagreement between full conditional and full marginal individual influence measures since these measures are defined by integrating out only the j-th feature based on its conditional and marginal distribution, which directly corresponds to $P_{X_j|X_{-j}=x_{-j}}$ and P_{X_j}.

$$\phi_{1|\Omega}(i) = (\mathcal{I}_{\text{full}}(\mathcal{B}^{\Omega}(\mathcal{M}^{\Omega}_c F)))(i) \quad \text{and} \quad \phi_{2|\Omega}(i) = (\mathcal{I}_{\text{full}}(\mathcal{B}^{\Omega}(\mathcal{M}^{\Omega}_m F)))(i),$$

Thus, it follows that the disagreement between CFI and PFI is minimized using Transformation Trees since they are both full individual influence measures (Fumagalli et al., 2025).

However, the disagreement between pure conditional and pure marginal individual influence measures, such as MPlot and PDP, is not explicitly targeted, as this would require minimizing the discrepancy between $P_{X_{-j}|X_j=x_j}$ and $P_{X_{-j}}$, which is not the objective of Transformation Trees. Nevertheless, if Transformation Trees achieve complete independence between X_j and X_{-j} within the final regions, it follows not only that $P_{X_j|X_{-j}=x_{-j}} = P_{X_j}$, but also that $P_{X_{-j}|X_j=x_j} = P_{X_{-j}}$. In this case, the conditional and marginal pure effects coincide. Transformation Trees generally aim to **minimize distributional influence** for **individual influence measures** between **conditional** and **marginal** masking in the case of **full** measures. However, the paper considers **full** measures in the context of **risk behavior** and **pure** measures in the context of **local behavior** (Molnar et al., 2023b; Fumagalli et al., 2025). In the latter case, the disagreement between conditional and marginal measures is not explicitly minimized but vanishes if the tree's optimization leads to terminal regions where X_j is independent of X_{-j}. Similar to REPID, Transformation Trees rely on recursive partitioning and consider only one feature of interest.

E ADDITIONAL EXPERIMENTS

In addition to Section 5, we present further applications of GRANITE on various real-world datasets to address different interpretability questions and provide additional insights. We first extend the analysis of the *Bikesharing* dataset, followed by an investigation of the *Kin8nm* dataset to explore higher-order interactions, and conclude with the *Diabetes* dataset to examine joint feature influence, with full implementation and code for the experiments available at https://github.com/gablabc/GRANITE.

E.1 Bikesharing Extended

Bikesharing is a regression dataset whose task is to predict the number of bike rentals given 17K examples and 10 features: *yr*, *mnth*, *hr*, *holiday*, *weekday*, *workingday*, *weathersit*, *temp*, *hum*, and *windspeed*. The dataset was split into train and test sets with an 80-20 ratio, resulting in 14K training instances and 3K test instances. A `HistGradientBoostingRegressor`[2] was tuned using 5-fold cross-validation on the training set, with hyperparameters selected via random search. The final model attained a Root-Mean-Squared-Error of 43.08 on the test set, while the target has a standard deviation of 181.45. Given that the error is more than four times lower than the model variability, we can safely assume that its post-hoc explanations are able to highlight meaningful trends in the data.

Minimizing Feature Interactions. We first investigated the model behavior by reporting the ICE curves (full-marginal) and PDP (pure-marginal) for the 10 features of Bikesharing (see Figure 9). It is apparent that there are strong disagreements between the ICE and PDP curves for features *hr*, *workingday*, *temp*, and *yr*, which makes it difficult to understand how these features individually impact the model response. By partitioning the input into three disjoint regions, however, GRANITE is able to reduce the disagreements between the pure and full explanations (see Figure 10). We note that all ICE/PDP curves are flat when it is early in the morning on a working day. This indicates that the model does not exhibit any interesting variability during these times. This regional trend was initially hidden in the non-regional explanations, since they mask features using the whole dataset as a reference.

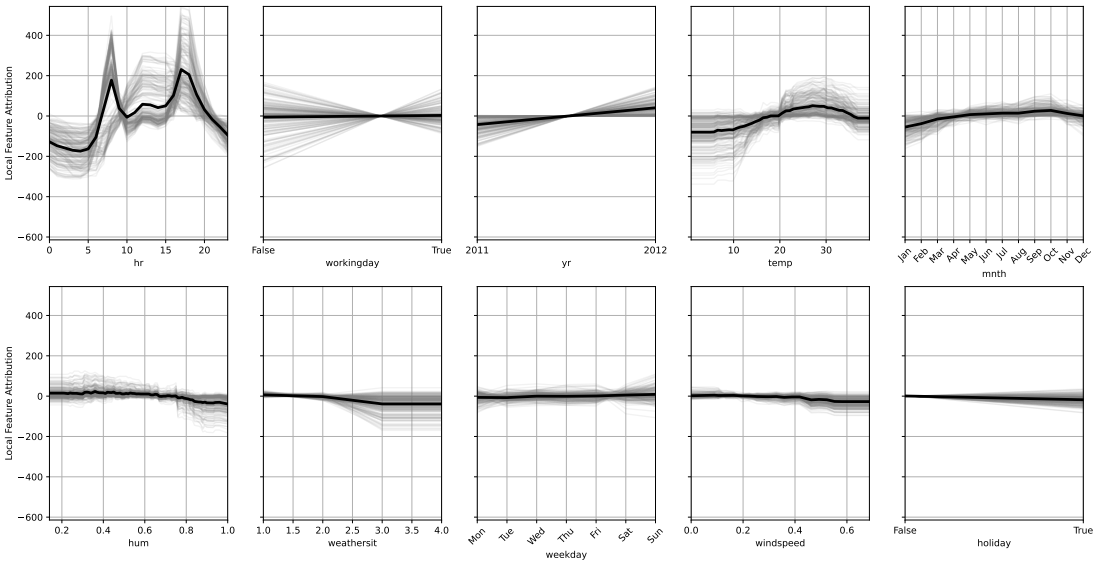


Figure 9: **Interaction disagreement** with $\mathcal{B}_{loc,x_0}$ and $\mathcal{M}_m$ between full (ICE curves as thin lines) and pure (PDP as thick line) explanations. *Hour*, *temperature*, *yr*, and *workingday* exhibit the highest disagreements between ICE and PDP.

Minimizing Feature Dependencies. Besides feature interactions, dependencies can complicate the interpretation of black-box models, because marginal masking may evaluate the model on unrealistic inputs. Since model behavior is often unpredictable outside the observed data, marginal-based explanations should be interpreted with caution. To confirm that extrapolation is occurring in marginal explanations, we compared the

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html

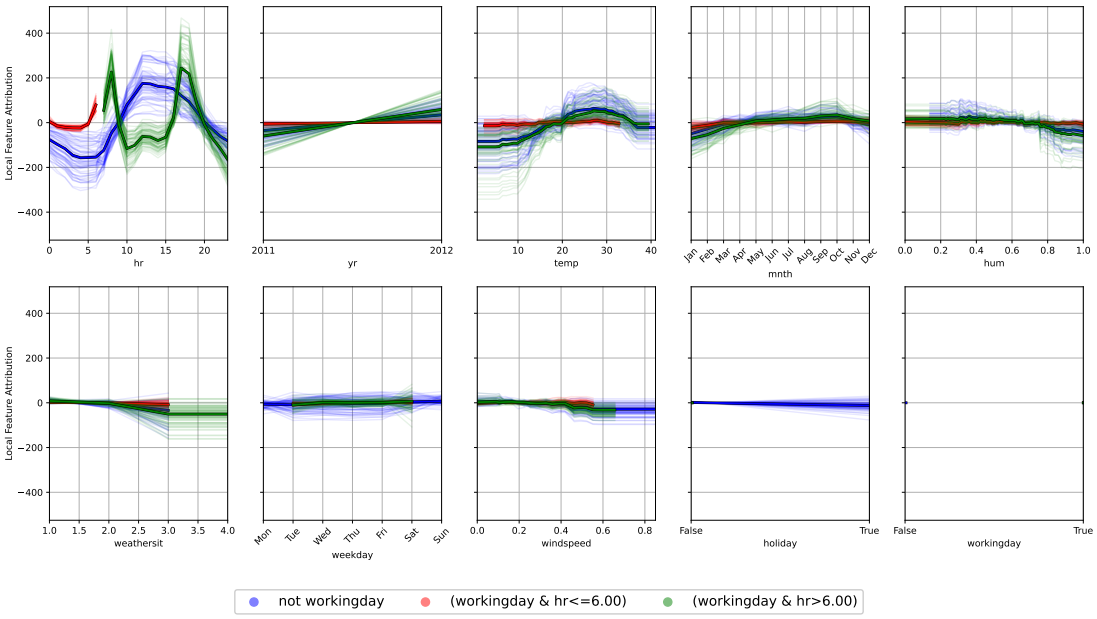


Figure 10: **Minimizing Interaction disagreement** with $\mathcal{B}_{loc,x_0}$ and $\mathcal{M}_m$ between full (ICE curves, thin lines) and pure (PDP, thick line) explanations. The disagreements are lessened when the explanations are restricted to the three regions.

MPlot (pure-conditional-local effects) alongside the PDP (pure-marginal-local effects) in Figure 11. We indeed observe that, for weather-related features (*mnth*, *weathersit*, *temps*, *hum*), the MPlot and PDP tend to disagree on extreme values of the feature. To reduce the disagreement between the conditional and marginal explanations, GRANITE suggests partitioning the input space into three regions that separate early/daytime hours and hot/cold temperatures. Figure 12 shows the corresponding regional explanations. It is apparent that disagreements between PDP and MPlot are greatly reduced, which implies that regional marginal explanations are less affected by extrapolation than their global counterparts.

A similar conclusion can be drawn when comparing CFI and PFI (full-conditional-risk versus full-marginal-risk). When using the whole dataset as the reference distribution, the two methods disagree on the feature importance rankings, as shown in Figure 13 (a). Notably, CFI and PFI contradict each other on the importance of *workingday*: PFI identifies it as the second most important feature, whereas CFI attributes it no importance. This disagreement is induced by a dependency between *workingday* and the features *weekday* and *holiday*. Indeed, one feature can be determined by knowing the other two: $workingday = [weekday \notin Sat, Sun] \wedge [\neg holiday]$. This redundancy induces conflicting interpretations between the marginal and conditional explanations. PFI considers *workingday* important because the model relies on it to make predictions, while CFI claims it is not important because the learning algorithm does not **need** it to yield an accurate model (the other two features could be used instead). GRANITE addresses this ambiguity by restricting the distribution to well-chosen regions (see Figure 13 (b)). In this case, both methods attribute no regional importance to *workingday*. This does not mean that this feature is not important globally, however, since *workingday* is used to split the input space. It remains important in the sense that it substantially changes the model's behavior when conditioned upon.

E.2 Kin8nm

The kin8nm dataset is a simulation of the dynamics of a 8-link robot arm. The regression task aims at predicting the distance between the tip of the arm and a target based on 8 features, each representing an angle of the robot arm joints. Although our understanding of the dataset is extremely limited (we do not know what each angle represents), studying it is still relevant because it involves high-order interactions that are extremely hard to visualize. This subsection presents how GRANITE can be used to handle these high-order interactions.

We fitted a Multi-Layered Perception (MLP) on this dataset reaching an R^2 coefficient of 0.9. Although not a perfect estimate, this model can still potentially highlight the strong interactions present in the robotic arm's dynamics. To illustrate the presence of such interactions, we first computed the PDP (pure) and ICE (full) curves

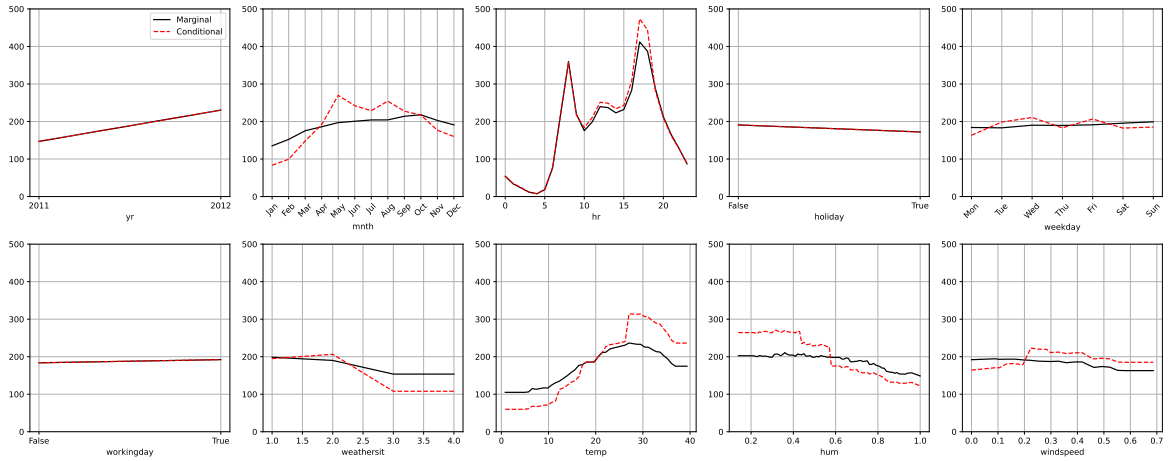


Figure 11: **Distribution disagreement** with $\mathcal{B}_{loc,x_0}$ and $\mathcal{I}_{pure}$ between MPlot ($\mathcal{M}_c$) and PDP ($\mathcal{M}_m$). Highest disagreements occur for weather-related features.

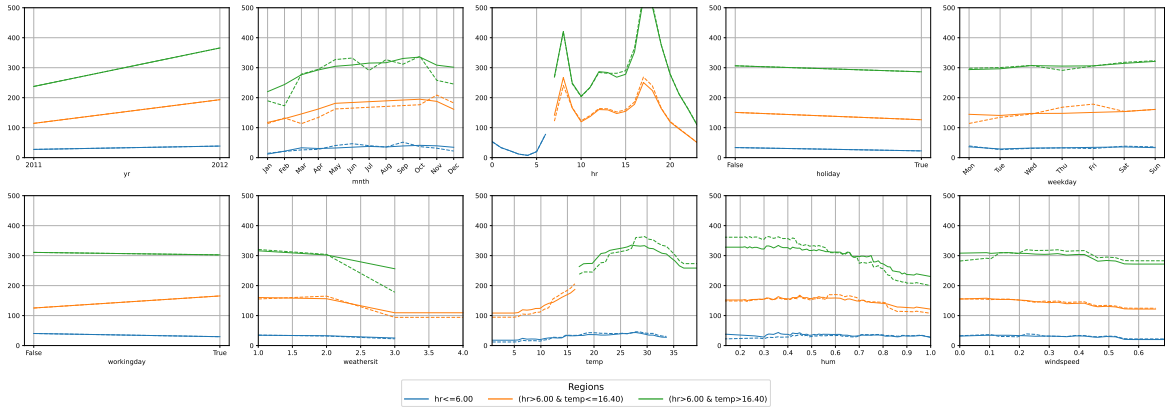
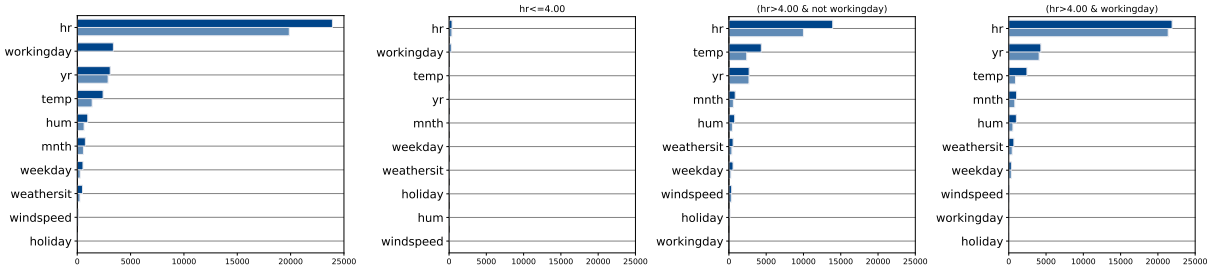


Figure 12: **Minimizing distribution disagreement** with $\mathcal{B}_{loc,x_0}$ and $\mathcal{I}_{pure}$ between MPlot ($\mathcal{M}_c$) and PDP ($\mathcal{M}_m$). The disagreements are less important when restricting to early/daytime hours and hot/cold temperatures.



(a) Global. (b) Regional.

Figure 13: **Minimizing distribution disagreement** with $\mathcal{B}_{risk}$ and $\mathcal{I}_{full}$ between CFI ($\mathcal{M}_c$) and PFI ($\mathcal{M}_m$). The disagreements are reduced restricted to early hour or workingdays/not-workindays.

(see Figure 14). The ICE curves show substantial variability, particularly for `theta6` and `theta7`, indicating that these features interact strongly with others.

We then applied GRANITE to find regions where the pure and full marginal feature effects exhibit increased agreement. This corresponds to minimizing the interactions within the functional decomposition in each region. Figure 15 shows the resulting regional PDP and ICE curves. Interactions are clearly reduced, as indicated by the diverse trends in regional PDP curves. For example, the slope of the `theta4`-PDP when `theta4` < 0 can be positive, negative, or null depending on the region.

While univariate explanations provide useful insights, they can become difficult to interpret when numerous

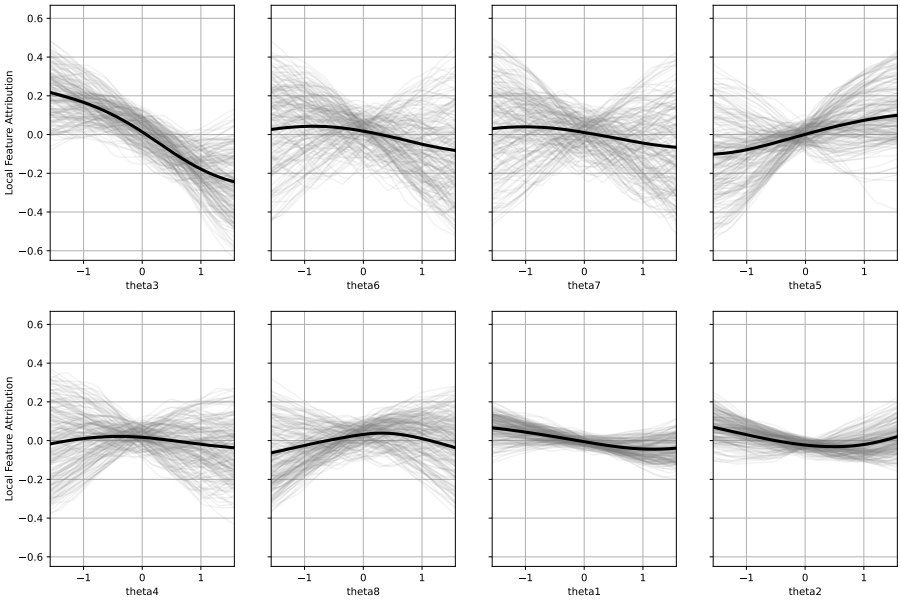


Figure 14: PDP (thick line) and ICE curves (thin lines) on Kin8nm.

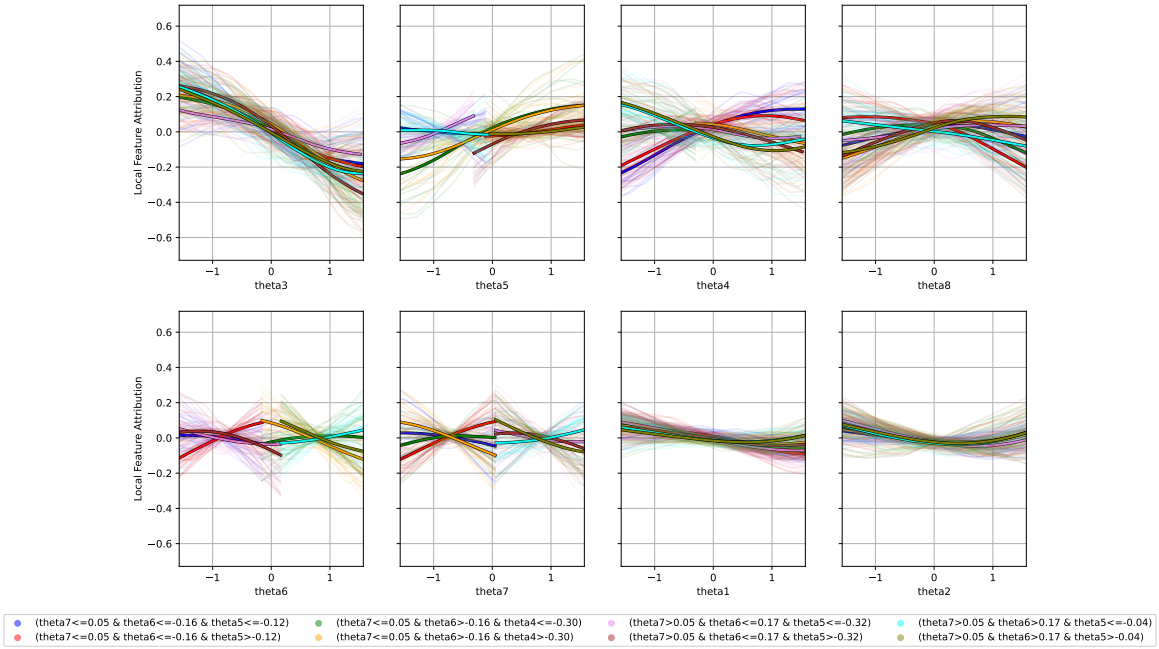
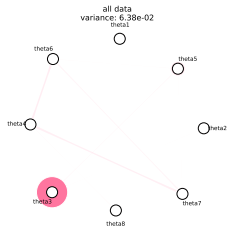


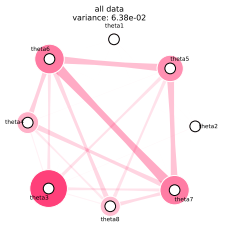
Figure 15: PDP (thick line) and ICE curves (thin lines) in 8 regions discovered by GRANITE on Kin8nm.

regions are considered. For example, moving from visualizing 8 univariate functions to $8 \times 8 = 64$ curves (corresponding to 8 regions) quickly becomes overwhelming. One reason GRANITE returns a relatively large number of regions is that it penalizes **all** interactions within the model, including order-2 interactions, which are often still interpretable. To address this, we opted to increase the expressivity of our explanations by incorporating not only individual feature effects but also pairwise interactions. In this refined approach, GRANITE is then applied to minimize only higher-order interactions, preserving the interpretability of lower-order effects.

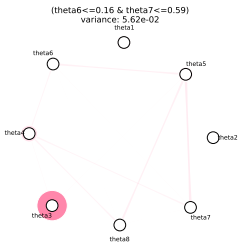
Figures 16 (a) and (b) show the pure (a) and full (b) sensitivity effects of individual features and their pairwise interactions, computed using the whole dataset. Differences in edge widths between the two graphs reflect the presence of higher-order feature interactions. These higher-order interactions indicate that explaining the model using only pure pairwise interactions, while interpretable, can sometimes be misleading. To address this, we



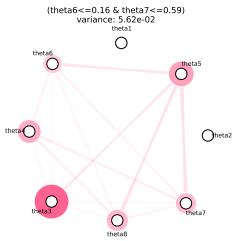
(a) Marginal-**Pure**-Sensitivity effects for features and pair-wise interactions using whole dataset as reference.



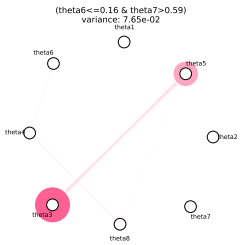
(b) Marginal-**Full**-Sensitivity effects for features and pair-wise interactions using whole dataset as reference.



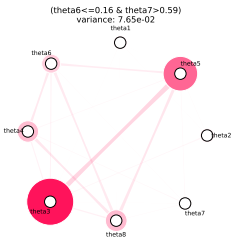
(c) Pure Region 0.



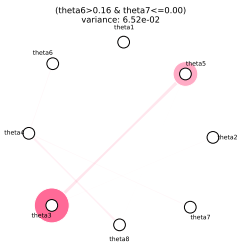
(d) Full Region 0.



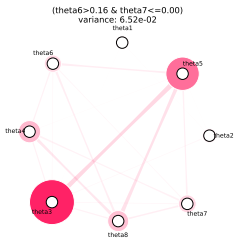
(e) Pure Region 1.



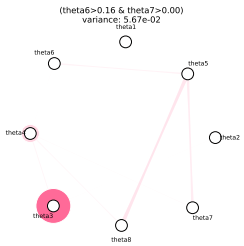
(f) Full Region 1.



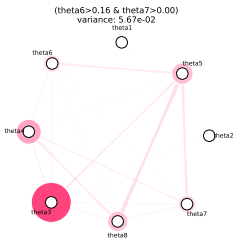
(g) Pure Region 2.



(h) Full Region 2.



(i) Pure Region 3.



(j) Full Region 3.

Figure 16: GRANITE splits the input space of Kin8nm into four regions, each with reduced high-order interactions, i.e., the pure and full pair-wise interactions agree more regionally.

applied GRANITE to minimize the disagreements between the pure and full measures of pairwise interactions.

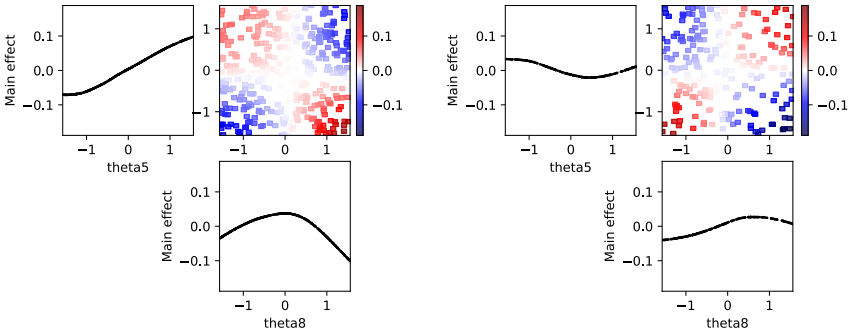
$$\mathcal{R}_{\ell|\Omega}(\phi_{1|\Omega}, \phi_{2|\Omega}) := \mathbb{E}_{P_X^\Omega}\left[\sum_{S\in\mathcal{S}} \ell\left(\phi_{1|\Omega}(S) - \phi_{2|\Omega}(S)\right)\right],$$

with $\mathcal{S} = \{S \subseteq D : |S| = 2\}$. Thus, we search for regions where interactions of order 3 or more are minimized. Figures 16 (c)–(j) show the pure and full sensitivity effects across the four regions produced by the partitioning algorithm. It is evident that the pure and full pairwise interactions are more closely aligned within these regions. Therefore, the regional pure pairwise interaction effects provide a more faithful representation of the model.

As an illustration, we examined the pure interaction effects between `theta5` and `theta8` in regions 0 and 3. These regions were selected because, in the other two, strong disagreements between pure and full `theta5:theta8` interactions are still evident in Figure 16. Figure 17 shows the pure individual effects and pure pairwise interactions across both regions. Notably, the interaction completely flips sign between the two regions, providing evidence of a three-way (or higher-order) interaction. Such effects would have been very difficult to visualize without using GRANITE to partition the input space.

E.3 Diabetes

The `Diabetes` dataset is a regression task aimed at predicting disease progression one year after baseline. It contains 442 samples and 10 numerical features: *age*, *sex*, *bmi*, *bp* (blood pressure), and six serum measurements (*s1*–*s6*) representing various blood and lipid concentrations. The data were split into training and test sets



(a) Region 0. (b) Region 3.

Figure 17: Kin8nm: Pure main effects for `theta5`, `theta8` and their pairwise interactions across two regions.

with an 80–20 ratio, resulting in 354 training and 88 test instances. A `HistGradientBoostingRegressor` was optimized using 5-fold cross-validation on the training set and random search for hyperparameter tuning.

The given features allow for a meaningful grouping as follows: **Demographics:** *age*, *sex*; **Body metrics:** *bmi*, *bp*; **Lab markers:** *s1–s6*.

Minimizing Feature Interactions. We now analyze the overall importance of these feature groups and how their ranking changes when interactions between groups are minimized. We apply GRANITE for **joint influence measures** with sensitivity behavior $\mathcal{B}_{\text{sens}}$ and marginal masking $\mathcal{M}_m$ to minimize interactions between full and pure Sobol variances. Figure 18 shows that, on a global level, body metrics and lab markers dominate the influence of demographic features, with body metrics being slightly more important. While the pure and full joint influences yield similar global rankings, regional analyses reveal notable differences: in particular, lab markers are by far the most important group for individuals with low BMI. In these regions, body metrics contribute the least, whereas demographics and body metrics exhibit higher influence for individuals with both low BMI and low blood pressure, compared to those with low BMI but high blood pressure, where lab markers appear to be the sole dominant group. For individuals with high BMI, both lab markers and body metrics play a substantial role, although their relative importance varies by gender.

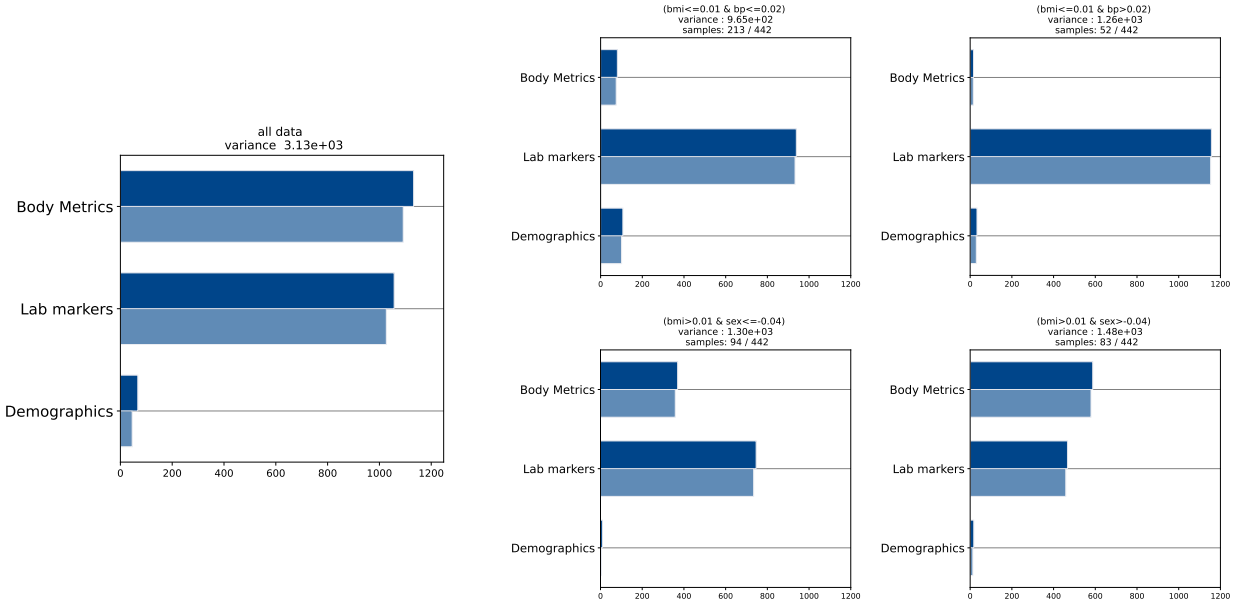


Figure 18: Diabetes: Full (top bar) and pure (bottom bar) sensitivity-based joint influences for global (left) and regional (right) effects measured by Sobol variances for the three feature groups.