SINC: DISCOVERING GLOBAL HIGHER-ORDER FEATURE INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how black-box models combine features to make predictions remains a fundamental challenge in explainable AI (XAI). Existing post-hoc methods primarily focus on individual feature importance, pairwise interactions, or local explanations, making it difficult to recover global interaction structure. We propose Synergy Identification via Network Clustering (SINC), a model-agnostic, post-hoc framework for discovering global higher-order feature interactions in trained black-box models. SINC constructs a sparse interaction network from pairwise Friedman's H-statistics, identifies communities using Leiden clustering with a Constant Potts Model objective, and ranks their predictive contribution using permutation importance. In doing so, SINC recovers the global feature interaction structure from pairwise interaction measurements. Across controlled synthetic and real-world datasets, we demonstrate that SINC recovers complex feature synergies and effectively identifies uninformative features. Compared with SHAP-based interaction analysis and the Archipelago method, SINC reveals higher-order groupings that are difficult to capture using a pairwise method and yields unified global communities, overcoming the grouping variability inherent in aggregating local explanations. Finally, we show that SINC serves as a diagnostic tool, revealing whether a model has successfully learned the underlying interaction structure of the data.

1 INTRODUCTION

Machine learning models, largely black-box models, have achieved remarkable success across a wide spectrum of application domains, including computer vision, natural language and speech processing, and recommendation systems. Their predictive power stems from their ability to learn highly complex non-linear relationships between variables from large-scale data, but a large number of parameters that encode those relationships render their decision-making processes largely opaque. As a result, their deployment in high-stakes settings, such as clinical workflows and decision making, remains challenging, motivating the development of Explainable Artificial Intelligence (XAI), which seeks to provide human-interpretable explanations for model predictions.

Post-hoc explanation approaches, such as feature attribution (Sundararajan et al., 2017; Ribeiro et al., 2016), saliency maps (Simonyan et al., 2014; Selvaraju et al., 2017), and Shapley-value-based explanations (Lundberg & Lee, 2017), are typically used to quantify the importance of individual features in model outputs. However, these methods primarily focus on individual feature contributions, providing only limited insight into how features interact to influence predictions. In practice, real-world phenomena are highly interdependent and rarely arise from independent feature effects. As an example, heart rate and respiration rate may on their own separate psychological stress from physical exertion poorly, while their relationship can be highly informative (Brouwer et al., 2018). Characterizing how groups of features interact and jointly influence model predictions is therefore a central challenge in XAI.

Efforts to understand global interactions have largely followed two research directions: inherently interpretable (glass-box) models and post-hoc explanation methods. Inherently interpretable models are explicitly designed to capture global interactions, but their adoption requires training new architectures from scratch, often with potential compromises in predictive performance. Conversely, while existing post-hoc methods can evaluate already-trained models, they are predominantly lo-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

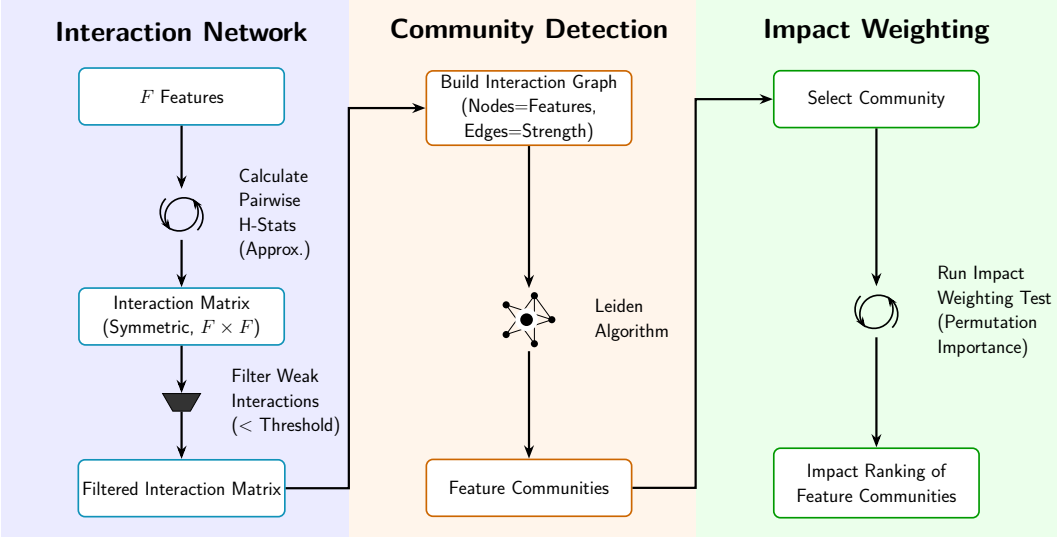


Figure 1: The three-stage architecture of SINC. First, SINC constructs an interaction network by approximating pairwise H-statistics and removing weak connections. Second, the resulting graph is partitioned into feature communities using the Leiden algorithm. Finally, the global importance of each identified community is ranked using permutation impact weighting.

cal in nature. Deriving global interactions from aggregating these local, instance-level explanations introduces prohibitive computational bottlenecks and structural inconsistencies, limiting their applicability to complex, multimodal models. Consequently, there is a growing demand for a tractable post-hoc method that is capable of detecting and quantifying complex interactions in existing black-box models while preserving their predictive accuracy.

We propose Synergy Identification via Network Clustering (SINC), a post-hoc, model-agnostic framework that aims to recover global feature interactions from trained models, as illustrated in Fig. 1. Rather than calculating every possible feature combination, SINC builds upon the H-statistic (Friedman & Popescu, 2008) to identify interaction patterns. As shown in Fig. 1, the framework consists of three stages. First, it constructs an interaction network by calculating approximate pairwise H-statistics across F features, filtering out weak connections to form a sparse interaction matrix. Second, it applies the Leiden algorithm (Traag et al., 2019) to perform community detection, i.e., grouping densely connected features into distinct sub-networks. Finally, SINC performs impact weighting by using permutation importance to rank the global predictive impact of each detected community. Our contributions are summarized as follows:

- We present a post-hoc, model-agnostic explanation method that recovers global feature interaction structure by combining interaction estimation with community detection.
- We show that SINC is able to recover ground-truth interaction structure under controlled synthetic conditions and serves as a diagnostic tool for assessing whether a model has learned the intended interaction structure.
- We demonstrate that SINC can efficiently identify meaningful feature communities that align with domain knowledge and rank their global predictive impact.

2 Related Work

Early XAI efforts primarily focused on quantifying individual feature contributions, with SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) becoming widely adopted for assigning importance scores based on cooperative game theory. However, SHAP's underlying assumption of an additive structure limits its ability to capture complex non-linear dependencies. For example, when evaluating the phrase "not good," SHAP might assign opposing scores that improperly cancel each other out rather than capturing their synergistic meaning.

To address these additive limitations, methods like Archipelago (Tsang et al., 2020) shifted the focus to feature interactions by measuring the degree of non-additivity among feature sets. Despite this improvement, Archipelago remains fundamentally a local explanation method that evaluates only a single input sample at a time. This local perspective makes it difficult to understand how interactions generalize into consistent global patterns across an entire model. Furthermore, when Archipelago constructs feature communities, it naively groups all features sharing any connection path, entirely disregarding the underlying interaction strengths.

To capture generalized, global model behavior, classical statistical approaches often utilize Friedman's H-statistic (Friedman & Popescu, 2008) to quantify pairwise dependencies. However, while higher-order H-statistics are theoretically sound, estimating them directly from high-dimensional real-world datasets often fails due to sample sparsity. Consequently, recent research on global interactions has pivoted toward inherently interpretable architectures, such as HONAM (Kim et al., 2025) and GAMI-Net (Yang et al., 2021), which explicitly encode interactions into their network topologies. While effective, these rigid architectures often compensate accuracy for interpretability.

Ultimately, this leaves a critical need for a post-hoc framework capable of identifying global feature communities of arbitrary size and quantifying their collective predictive impact in trained black-box models.

3 Synergy Identification via Network Clustering

We propose SINC, a novel post-hoc, model-agnostic framework for discovering global feature interactions. As illustrated in Fig. 1 and formalized in Appendix A, this is achieved through three distinct stages: interaction network construction, community detection, and impact weighting.

3.1 Stage 1: Quantifying feature interactions

In the first stage, SINC quantifies pairwise feature interactions using Friedman's H-statistic (Friedman & Popescu, 2008) to construct an interaction network. The H-statistics compares the joint partial dependence of a feature pair against their individual marginal partial dependencies, thereby providing a global, model-agnostic measure of interaction strength. The exact mathematical formulations for the partial dependencies and the H-statistic are provided in Appendix B.1.

Calculating the exact Friedman's H-statistic requires up to $3N^2$ model predictions per feature pair, which is computationally infeasible for large datasets. The approximation adopted here combines two operations, following established practice (Modeva Team, 2025):

1. First, the partial dependencies are estimated from a random subset of $M < N$ instances.
2. Second, the partial dependencies are evaluated on a uniform grid of resolution R, where $R_i = \min(R, |\{x_i^{(k)}\}_{k=1}^{N}|)$ accepts features with fewer unique values than R.

Let $\widehat{PD}$ denote the partial dependencies obtained from these two operations, and let $\tilde{x}_i^{(k)}$ denote the k-th grid value of feature x_i. The approximated statistic is then given by:

$$H_{ij}^2 \approx \frac{\sum_{k=1}^{R_i}\sum_{p=1}^{R_j}\left(\widehat{PD}_{ij}(\tilde{x}_i^{(k)}, \tilde{x}_j^{(p)}) - \left(\widehat{PD}_i(\tilde{x}_i^{(k)}) + \widehat{PD}_j(\tilde{x}_j^{(p)})\right)\right)^2}{\sum_{k=1}^{R_i}\sum_{p=1}^{R_j}\left(\widehat{PD}_{ij}(\tilde{x}_i^{(k)}, \tilde{x}_j^{(p)})\right)^2}. \tag{1}$$

Because M and R are user-defined and typically much smaller than N, the number of model predictions per feature pair drops to $2RM + R^2M$, which is independent of the dataset size, while the total cost over the $F(F-1)/2$ pairs still grows quadratically in the number of features. Once Friedman's H-statistics have been calculated for all feature pairs, the final step is to construct a weighted, undirected graph $G = (V, E)$, in which the vertex set $V = \{1, \ldots, F\}$ contains one node per feature x_i, and the weight w_{ij} of the edge joining nodes i and j is given by their interaction strength. To eliminate edges attributable to sampling and grid-approximation error rather than meaningful interactions, we apply a threshold τ to sparsify the network:

$$w_{ij} = \begin{cases} H_{ij}^2 & \text{if } H_{ij}^2 \geq \tau \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Features whose edges all fall below τ remain isolated in G.

A fundamental assumption of this formulation is that the feature space is semantically aligned across all instances; because SINC marginalizes over specific indices, the i-th feature must represent the same concept globally.

3.2 Stage 2: Detecting feature communities

In the second stage, SINC identifies communities of strongly interacting features in G through the Leiden algorithm (Traag et al., 2019), which returns a partition $\mathcal{C} = \{C_1, \ldots, C_q\}$ of the feature set. The Leiden algorithm was selected over the Louvain algorithm (Blondel et al., 2008) because its refinement phase guarantees internal community connectivity and scales more efficiently to large networks.

Partitioning the interaction network G requires an objective function. We employed the constant Potts model (CPM) (detailed formally in Appendix B.2), which has been shown to recover planted partitions at higher normalized mutual information than modularity-based objectives (Traag et al., 2011). Furthermore, CPM is resolution-limit-free, meaning a partition optimal for G remains optimal for any induced subgraph. The objective relies on a parameter γ which balances the number of internal edges while at the same time keeping relatively small communities. If $\gamma = \max_{ij} w_{ij}$, it is optimal to split all nodes whereas if $\gamma = \min_{ij} w_{ij}$, it is optimal to merge all nodes into one community. Consequently, the resolution parameter is typically chosen within the range $\min_{ij} w_{ij} \leq \gamma \leq \max_{ij} w_{ij}$.

3.3 Stage 3: Impact-wise Community Ranking

Utilizing the Leiden algorithm with CPM on the interaction network results in a single partition of feature communities. In Stage 3, SINC ranks these communities by their global predictive importance via permutation importance.

Specifically, the importance of a community is quantified by measuring the increase in model error after randomly permuting all features belonging to that specific community across instances, while keeping all other features unchanged (formal definition is provided in Appendix B.3). A community is ranked as highly important if the permuted error increases substantially compared to the baseline error, indicating that the model relies heavily on information carried by that feature group to make accurate predictions.

4 Experiments

We evaluate SINC across synthetic and real-world datasets. Specifically, we compare SINC against SHAP and Archipelago to demonstrate its effectiveness in identifying global, higher-order feature synergies over purely pairwise or local aggregation techniques. Across all experiments, the approximation parameters $M = 100$ and $R = 30$ remain fixed to balance distribution representation and computational efficiency, while the network threshold τ and resolution parameter γ are configured per experiment. Because SHAP and Archipelago are fundamentally local explanation methods, deriving a global interaction structure requires an aggregation strategy. To evaluate them globally, we aggregate their instance-level explanations across a fixed validation subset of 100 instances. For SHAP, global feature importance and pairwise interaction strengths are defined as the mean absolute local attributions across the subset. For Archipelago, the global impact of a specific feature community is calculated as its mean absolute local attribution across the entire validation subset, where instances that do not form that exact community contribute an attribution score of zero.

All experiments are conducted with a fixed, semantically consistent feature space, aligning with a core assumption of SINC that each feature retains the same semantic meaning across all data instances.

4.1 Synthetic Experiments

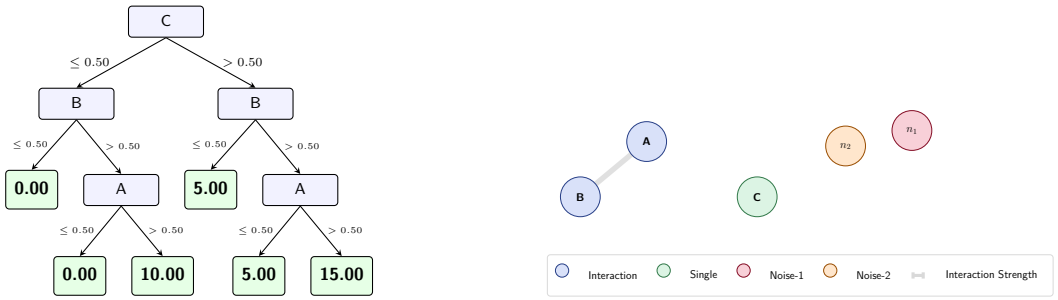
Synthetic datasets provide a controlled environment with explicitly defined ground-truth interactions, enabling direct verification of the method's capability to capture known patterns.

4.1.1 WHITE BOX EXPERIMENT

We generate a synthetic dataset consisting of five uniformly distributed features $\mathbf{x} = [\text{A,B,C}, n_1, n_2]^\top \in [0,1]^5$, where n_1 and n_2 are uninformative noise variables. The target y establishes a clear deterministic interaction between feature A and B, while feature C contributes independently through a purely additive effect:

$$y = 10 \cdot \mathbb{1}_{\text{A}>0.5} \cdot \mathbb{1}_{\text{B}>0.5} + 5 \cdot \mathbb{1}_{\text{C}>0.5}. \tag{3}$$

We fit a decision tree (Quinlan, 1986) with a maximum depth of 3, which perfectly captures these dynamics, achieving an R^2 score of 1.0, as illustrated in Fig. 2a. Applying SINC on the fitted decision tree utilizing the hyperparameters $\tau = 0.01$ and $\gamma = 0.01$, yields the feature interaction network in Fig. 2b. The recovered interaction graph correctly identifies the A+B synergy, isolates the additive feature C, and separates the noise variables, demonstrating SINC's ability to recover the interaction structure learned by the model.



(a) Internal logic of the decision tree.

(b) Feature interaction network created by SINC.

Figure 2: Internal logic of the decision tree and the corresponding feature interaction network recovered by SINC from the trained model.

Table 1 reports SHAP feature importance and interaction strength, while Table 2 reports the feature communities and their global impacts recovered by Archipelago and SINC. While all three methods successfully filter out the noise features (n_1, n_2), they diverge significantly in how they represent the informative variables. Because SHAP evaluates both individual feature importance and pairwise interaction strength, practitioners must manually cross-reference multiple metrics to infer the model's underlying logic. Although SHAP correctly identifies A+B as the strongest interacting pair, it also assigns small but non-zero interaction scores to independent features (e.g., A+C). Both Archipelago and SINC recover communities that are consistent with the ground-truth structure: the synergistic pair A+B and the independent feature C. However, Archipelago fails to capture their importance, assigning both an equal 50% global importance despite the target function weighting the A+B interaction twice as heavily as C. By contrast, SINC assigns 74.76% of the total impact to the A+B community and 25.24% to C, making it the only method that recovers a community importance hierarchy consistent with the underlying target function.

Table 1: SHAP global feature importance and pairwise interaction strengths for the decision tree. Negligible interactions involving the noise features are grouped for brevity.

(a) Feature importance

Feature	**Mean |SHAP| (%)**
C	2.49 (34.02)
A	2.45 (33.47)
B	2.38 (32.51)
Noise (1-2)	0.00 (0.00)

(b) Pairwise interaction strength

Feature Pair	**Strength (%)**
A + B	2.46 (93.78)
A + C	0.10 (3.85)
B + C	0.06 (2.37)
All Noise Pairs	0.00 (0.00)

Table 2: Feature communities identified by Archipelago and SINC on the decision tree, together with their attribution/impact on model predictions. Global impacts correspond to each method's native importance measure.

(a) Archipelago

Community	Size	Global Attribution (%)
A + B	2	2.50 (50.00)
C	1	2.50 (50.00)
Noise - 1	1	0.00 (0.00)
Noise - 2	1	0.00 (0.00)

(b) SINC

Community	Size	Global Impact (%)
A + B	2	36.86 (74.76)
C	1	12.44 (25.24)
Noise - 1	1	0.00 (0.00)
Noise - 2	1	0.00 (0.00)

4.1.2 Black Box Experiments

Next, we compare the three methods on Multi-Layer Perceptrons (MLPs) to evaluate their ability to detect latent higher-order interactions in standard black-box models. The input $\mathbf{x} \in [0,1]^{10}$ contains a synergistic group of features A, B, C and D, additive features E and F, and four noise features n_1, n_2, n_3 and n_4. The target incorporates additive Gaussian noise $\mathcal{N}(0, 0.1)$ and is given by:

$$y = 20 \cdot \sin(\pi \cdot \mathrm{A} \cdot \mathrm{B} \cdot \mathrm{C} \cdot \mathrm{D}) + 10 \cdot \mathrm{E} - 5 \cdot \mathrm{F} + \mathcal{N}(0, 0.1). \tag{4}$$

We train two MLPs (three hidden layers with 64, 32, and 16 neurons respectively, utilizing ReLU activation function (Nair & Hinton, 2010)) using Adam (Kingma & Ba, 2015): a converged model trained for 150 epochs, achieving $R^2 = 0.9862$, and an intentionally undertrained model trained for 20 epochs, achieving $R^2 = 0.8721$.

Table 3: Global feature communities detected by Archipelago ($top_k = 5$) on the converged MLP. Appearance count indicates how many of the 100 evaluated instances produced this exact community. Communities contributing less than 2% to the total attribution are grouped for brevity.

Feature Community	Size	Global Attribution (%)	Appearance Count
A + B + C + D	4	1.30 (23.59)	40
E	1	0.99 (17.94)	67
F	1	0.79 (14.32)	82
A + B + C + D + E	5	0.32 (5.81)	7
Noise - 1	1	0.15 (2.74)	87
Noise - 4	1	0.13 (2.42)	96
A + C + D + Noise - 3	4	0.13 (2.32)	1
A + B + C + F + Noise - 1	5	0.12 (2.17)	1
Other 39 communities (each)	-	< 0.10 (< 2.00)	-

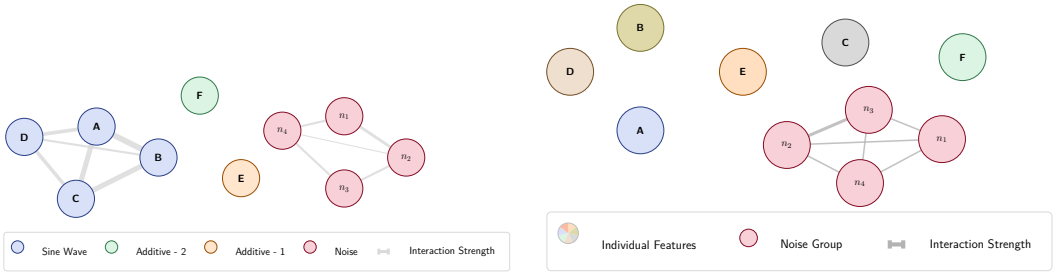
Table 4: Detected feature groups and their impact on model output as identified by SINC on the converged MLP.

Group	Size	Global Impact (%)
A + B + C + D	4	27.35 (58.47)
E	1	14.70 (31.43)
F	1	4.73 (10.10)
Noise (1-4)	4	0.00 (0.00)

Evaluating the converged model using SINC ($\tau = 0.05, \gamma = 0.001$) demonstrates its ability to recover complex higher-order synergies. As shown in Appendix C, SHAP assigns high importance to pairwise combinations among A, B, C, D; however, its restriction to second-order interactions prevents it from expressing the unified four-way synergy. Archipelago identifies the correct four-way

community as its most important community overall, but only recovers it in 40 of 100 evaluated instances, while several competing and overlapping subsets (e.g., A, B, C, D, E) also appear (Table 3). In contrast, SINC establishes a single global partition (Fig. 3a and Table 4). It identifies the A, B, C, D synergy, ranking it as the dominant contributor to the model's output and produces a clean additive decomposition of E, F, and the noise group.

To confirm that SINC's grouping reflects genuinely learned interaction structure we apply it to the undertrained model. As illustrated in Fig. 3b, the resulting interaction network clearly reveals the model's failure to capture the dependencies, treating A, B, C, and D as strictly isolated individual features. All three explanation methods agree that the undertrained model has not yet learned coherent higher-order interactions (see Appendix D). This contrast highlights SINC's utility as a diagnostic tool. While standard performance metrics, such as validation loss or R^2, only indicate general predictive error, they cannot explain how a model is failing. SINC allows practitioners to verify whether a network has successfully internalized expected domain-critical synergies, or if it is merely relying on superficial, independent feature effects.



(a) Converged Model ($R^2 = 0.9862$) (b) Undertrained Model ($R^2 = 0.8721$)

Figure 3: Feature interaction networks generated by SINC. The method successfully identifies the converged model's learned interaction structure and isolates noise variables (a), while revealing the undertrained model's failure to capture the feature synergy (b).

To further systematically evaluate SINC's ability to recover complex interaction structures across diverse functional forms, we extend our synthetic analysis using a benchmark suite of ten data-generating functions (detailed in Appendix H). Originally introduced in (Hooker, 2004) and (Tsang et al., 2018), these functions encapsulate a wide array of challenging topological features, including overlapping synergies, periodic dependencies, and higher-order non-linearities. Using the MLP architecture described above, we train a fully converged model on each of these ten functions so that the networks learn a known, mathematically defined ground-truth interaction structure.

To evaluate the effectiveness of each explanation method, we quantify how many of the explicit ground-truth interactions are successfully recovered from the trained models, where each term in the data-generating functions is viewed as a ground-truth interaction. A ground-truth term is counted as recovered if a method outputs a community (for SINC and Archipelago) or an interacting feature set (for SHAP) whose feature membership exactly matches the feature set of that term. Given that all trained MLPs achieve an $R^2 \geq 0.95$, we assume the networks have successfully parameterized the underlying data dynamics, ensuring the true interactions are present to be detected. Table 5 reports the number of correct interaction structures identified by SINC, SHAP, and Archipelago, alongside the maximum possible number of interactions (Total) for each function. Under strict evaluation criteria, SINC consistently outperforms or matches both baseline methods across the benchmark suite, achieving a mean recovery rate of 60.2% compared to 43.5% for Archipelago and 32.6% for SHAP. SHAP struggles to capture the underlying structures due to its inherent limitation to pairwise measurements, frequently failing to identify the correct interacting pairs and yielding the lowest recovery rates overall. While Archipelago recovers more ground-truth interactions than SHAP, its output remains practically challenging to interpret. The method generates a high volume of unstable, overlapping feature communities rather than a unified global partition, making it difficult to extract a clear interaction structure that aligns cleanly with the ground-truth. Conversely, when SINC deviates from the exact ground-truth count, it typically reflects a meaningful consolidation of interdependent feature communities. For example, in $F_1(\mathbf{x})$, the target function contains several heavily overlapping ground-truth communities. Instead of fragmenting them, SINC groups these

Table 5: Number of explicitly matched ground-truth interactions recovered by each method. Total denotes the maximum possible number of true interactions.

Function	SINC	SHAP	Archipelago	Total
F_1	**2**	1	**2**	6
F_2	**2**	1	**2**	6
F_3	**4**	3	**4**	7
F_4	**5**	4	4	8
F_5	**4**	2	2	4
F_6	**4**	2	2	5
F_7	**7**	1	6	14
F_8	**3**	1	**3**	5
F_9	**2**	**2**	**2**	5
F_{10}	**6**	3	2	7
Mean recovery (%)	**60.2**	32.6	43.5	100.0

variables into a single community (detailed in Appendix H), which is a possible representation of how the model collectively processes these entangled variables.

4.2 Real-world datasets

To assess the applicability of SINC to real-world data, we evaluate XGBoost models (Chen & Guestrin, 2016) across three diverse domains: regression on the California Housing dataset (Inria, 2012), binary classification on the HIGGS particle physics benchmark (Sadowski et al., 2014), and regression on the UCI Bike Sharing Demand dataset (Fanaee-T & Gama, 2013).

Table 6: Top feature communities and their global predictive impacts identified by SINC across three real-world datasets. Only communities accounting for $\geq 2\%$ of total impact are shown.

Dataset	**Feature Community**	**Impact (%)**
California Housing	1. Latitude + Longitude	53.17
	2. MedInc	29.75
	3. AveBedrms + AveOccup + HouseAge + Population	8.66
	4. AveRooms	8.43
HIGGS	1. m_bb + m_jj + m_jjj + m_wbb	41.54
	2. lepton_pT + m_jlv + m_lv + m_wwbb + missing_energy_magnitude	40.47
	3. jet1pt + jet2pt	15.03
Bike Sharing	1. hour + workingday	75.41
	2. year	10.78
	3. feel_temp + holiday + month + weekday + windspeed	5.39
	4. temp	4.64
	5. humidity + weather	2.89

On California Housing, SINC identifies Latitude + Longitude as the dominant community, accounting for 53.17% of the total predictive impact. This grouping is consistent with the known geographic relationship between latitude and longitude and their importance for predicting house values (Lundberg & Lee, 2017). In contrast, MedInc is identified as an individual community, separating its predominantly individual contribution from the geographic interaction. The remaining features are grouped into lower-impact communities, indicating that the model's strongest joint dependency is concentrated in geographic information.

For HIGGS, the two largest communities consist primarily of related kinematic quantities and together account for more than 80% of the total impact. These groupings are consistent with the fact that the HIGGS features describe different physical properties of the same particle event, and suggest

that the model relies on coordinated information across related measurements rather than treating all features independently.

On Bike Sharing, SINC identifies hour + workingday as the dominant community, accounting for 75.41% of the total impact. This grouping captures a natural behavioral relationship: the predictive effect of the hour of day depends on whether a day is a working day. The remaining communities separate other temporal and environmental factors, including year and a group containing weather-related variables. These results show that the communities discovered from the model's interaction structure can be interpreted in terms of distinct temporal, behavioral, and environmental factors.

Full SHAP, Archipelago, and SINC results are provided in Appendices E, F, and G.

5 LIMITATIONS

While SINC provides valuable insights into complex feature interactions, it is subject to a few notable limitations. A primary challenge lies in the calibration of the algorithm's hyperparameters, specifically the interaction threshold τ, the resolution parameter γ, the grid resolution R, and the subsampling parameter M. Tuning these parameters is critical to balance computational efficiency and approximation quality, ensure sufficient network sparsity, and prevent small-scale community structures from being merged due to the resolution limit. Crucially, there is currently no definitive method to determine the absolute optimal values for τ and γ, meaning these parameters must be empirically chosen based on the specific model and dataset. Fortunately, this empirical tuning is somewhat alleviated by the observation that there are often broad regions of convergence where different values of τ and γ tend to yield stable and consistent community structures, which is demonstrated via a parameter sweep in Appendix I.

A further limitation of SINC is that it assigns each feature to only one community. Consequently, if a single feature contributes equally, yet independently, to multiple communities, SINC will fail to highlight these interactions. One could imagine implementing an overlapping community detection algorithm instead of the Leiden algorithm, however this would introduce significant technical and analytical trade-offs. Specifically, overlapping methods typically suffer from severe computational bottlenecks, require heavy manual parameter tuning, and produce ambiguous boundaries that make downstream interpretation much harder. (Xie et al., 2013)

Furthermore, SINC depends on the availability, size, and representativeness of the data used for calculating the H-statistics and the impact weighting. Consequently, if this evaluation data is limited or unrepresentative, it risks generating skewed importance scores and ultimately compromises the reliability of the resulting explanations.

6 CONCLUSION

In this work, we introduced Synergy Identification via Network Clustering (SINC), a novel post-hoc, model-agnostic framework for discovering and quantifying global feature interactions in trained models. By approximating Friedman's H-statistics, applying graph-theoretic community detection via the Leiden algorithm, and weighting communities through permutation importance, SINC bridges the gap between raw feature attribution and structural interpretability. Experiments on both synthetic and real-world datasets demonstrate that SINC accurately identifies complex, higher-order interaction structures, outperforms local aggregation methods such as Archipelago in consistency and interpretability, and effectively isolates noise variables. Furthermore, SINC serves as a diagnostic tool, revealing whether a model has successfully learned the intended interaction structure and providing insights beyond conventional performance metrics. On the California Housing dataset and other real-world datasets, the recovered communities align closely with established domain knowledge. By providing a global view of feature interactions, SINC represents a step towards making machine learning more transparent, interpretable, and trustworthy in high-stakes applications.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in writing and debugging code for our experiments. We have not used generative AI tools for generating core research ideas, formulating theoretical claims, generating figures, or creating data. We used generative AI tools for refining sentences, correcting grammar, and improving the overall readability of the manuscript.

We have reviewed all AI-assisted work. Specifically, any code written with AI assistance was carefully reviewed, executed, and rigorously tested for correctness by the authors to ensure it functioned exactly as intended. Furthermore, all AI-refined text was manually reviewed to ensure it accurately preserved our original scientific meaning. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We recognize that the reliability of the generated explanations depends strictly on the availability and representativeness of the evaluation data. If the framework is applied to a limited or unrepresentative data distribution, it risks generating skewed importance scores. In high-stakes applications, this could inadvertently mask biased model behavior or provide a false sense of security. Consequently, practitioners deploying this methodology must ensure the use of robust, unbiased evaluation datasets to prevent misleading conclusions.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility, the source code for SINC, as well as the code used to train the models and run all benchmark comparisons, is available in this GitHub repository: `https://github.com/anonymous826419/SINC`. This repository contains the complete codebase, environment specifications, and configuration files for all synthetic and real-world experiments. Furthermore, all synthetic data-generating functions are defined in closed form in the paper, all real-world datasets are publicly accessible, and specific hyperparameter configurations (M, R, τ, and γ) are explicitly reported in their respective sections.

REFERENCES

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), October 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/p10008. URL http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Anne-Marie Brouwer, Elsbeth van Dam, Jan B. F. van Erp, Derek P. Spangler, and Justin R. Brooks. Improving real-life estimates of emotion based on heart rate: A perspective on taking metabolic heart rate into account. *Frontiers in Human Neuroscience*, Volume 12 - 2018, 2018. ISSN 1662-5161. doi: 10.3389/fnhum.2018.00284. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2018.00284.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Hadi Fanaee-T and João Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 11 2013. doi: 10.1007/s13748-013-0040-3.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 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*, 20:1–81, 2019.

Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), September 2008. ISSN 1932-6157. doi: 10.1214/07-aoas148. URL http://dx.doi.org/10.1214/07-AOAS148.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 575–580, 2004.

Inria. The california housing dataset. https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_california_housing.html, 2012. Accessed: 2023-10-24.

Minkyu Kim, Hyun-Soo Choi, and Jinho Kim. Higher-order neural additive models: An interpretable machine learning model with feature interactions. In *2025 IEEE International Conference on Data Mining (ICDM)*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.

Modeva Team. *H-statistics*. MoDevA AI, 2025. URL https://modeva.ai/_build/html/_source/user_guide/explain/hstats.html. Accessed: 2025-02-02.

Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 3 edition, 2025. ISBN 978-3-911578-03-5. URL https://christophm.github.io/interpretable-ml-book.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986. ISSN 0885-6125. doi: 10.1023/A:1022643204877. URL https://doi.org/10.1023/A:1022643204877.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Peter Sadowski, Pierre Baldi, and Daniel Whiteson. Searching for higgs boson decay modes with deep learning. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/44f58a78b6dd03bf9ccb526e7c7538d5-Paper.pdf.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

V. A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84(1), July 2011. ISSN 1550-2376. doi: 10.1103/physreve.84.016114. URL http://dx.doi.org/10.1103/PhysRevE.84.016114.

V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41695-z. URL http://dx.doi.org/10.1038/s41598-019-41695-z.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):1–35, 2013.

Zebin Yang, Aijun Zhang, and Agus Sudjianto. Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, 120:108192, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108192. URL https://www.sciencedirect.com/science/article/pii/S0031320321003484.

APPENDIX

A FORMALIZED ALGORITHM

Algorithm 1 Synergy Identification via Network Clustering (SINC)

Input: Trained model f, dataset $D = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, interaction threshold τ, resolution parameter γ, grid resolution R, subsample parameter M.
Output: Community partition $\mathcal{C}$ and community importance scores $\{I_j\}_{j=1}^{q}$.

Dataset Preparation
Draw a random subsample of M instances from D.
for each feature $i \in \{1, \ldots, F\}$ **do**
    Generate a sequence of evaluation points of length $R_i = \min(R, |\{x_i^{(k)}\}_{k=1}^{N}|)$.
Stage 1: Interaction Network
for each pair of features (i, j) **do**
    Compute pairwise H-statistic H_{ij}^2 using model f.
    if $H_{ij}^2 \geq \tau$ **then**
        Add weighted edge (i, j) to graph $G = (V, E)$ with weight $w_{ij} = H_{ij}^2$.
    else
        Add weighted edge (i, j) to graph $G = (V, E)$ with weight $w_{ij} = 0$.
Stage 2: Community Detection
Apply Leiden algorithm with constant Potts model(γ) as the objective function, to G.
Obtain a partition $\mathcal{C} = \{C_1, \ldots, C_q\}$ of the graph.
Stage 3: Impact Weighting
for each community $C_j \in \mathcal{C}$ **do**
    Compute permutation-based importance I_j using unseen data.
return $\mathcal{C}$ and $\{I_j\}_{j=1}^{q}$.

B EXTENDED MATHEMATICAL FORMULATIONS

B.1 PARTIAL DEPENDENCE AND THE EXACT H-STATISTIC

In Stage 1, SINC utilizes Friedman's H-statistic to quantify pairwise interactions (Friedman & Popescu, 2008). Let $\mathcal{D} = \{(\mathbf{x}^{(k)}, y^{(k)})\}_{k=1}^{N}$ denote the dataset, where each instance $\mathbf{x}^{(k)} = [x_1^{(k)}, \ldots, x_F^{(k)}]^\top \in \mathbb{R}^F$ collects the F features of a trained model f, and let $\mathbf{x}_{-i}^{(k)}$ denote that instance with feature i removed. The centered marginal partial dependence of feature i is given by:

$$PD_i(x_i) = \frac{1}{N}\sum_{k=1}^{N} f\left(x_i, \mathbf{x}_{-i}^{(k)}\right) - \frac{1}{N}\sum_{k=1}^{N} f\left(\mathbf{x}^{(k)}\right)$$

The centered joint partial dependence $PD_{ij}(x_i, x_j)$ of a feature pair is defined analogously by marginalizing over $\mathbf{x}_{-ij}^{(k)}$. Both quantities are centered to zero mean over $\mathcal{D}$.

Definition 1 (Friedman's H-statistic for Pairwise Interactions) *The interaction strength between two features, x_i and x_j, is defined as the fraction of variance in the joint partial dependence that cannot be explained by the sum of their individual marginal effects:*

$$H_{ij}^2 = \frac{\sum_{k=1}^{N}\left(PD_{ij}(x_i^{(k)}, x_j^{(k)}) - \left(PD_i(x_i^{(k)}) + PD_j(x_j^{(k)})\right)\right)^2}{\sum_{k=1}^{N}\left(PD_{ij}(x_i^{(k)}, x_j^{(k)})\right)^2}$$

Because calculating this exact statistic requires up to $3N^2$ model predictions per pair, SINC utilizes the grid-based approximation detailed in the main text.

B.2 Constant Potts Model (CPM) Objective

In Stage 2, SINC partitions the interaction network using the Leiden algorithm optimized for the CPM.

Definition 2 (Constant Potts Model) *Let $G = (V, E)$ be an undirected graph with adjacency matrix $W = (w_{ij})_{i,j \in V}$, and let c_i denote the community label of node i. The CPM function is given by:*

$$Q_{\text{CPM}} = \sum_{i,j \in V} (w_{ij} - \gamma)\, \delta(c_i, c_j)$$

where $\gamma > 0$ is a resolution parameter and $\delta(c_i, c_j) = 1$ if nodes i and j belong to the same community and 0 otherwise (Traag et al., 2011).

This objective can be rewritten as a sum over the communities C in the partition $\mathcal{C}$:

$$Q_{\text{CPM}} = \sum_{C \in \mathcal{C}} \left(e_C - \gamma n_C^2 \right)$$

where $e_C = \sum_{i,j \in V} w_{ij}\, \delta(c_i, C)\delta(c_j, C)$ and $n_C = \sum_{i \in V} \delta(c_i, C)$ represent the total internal edge weight and the number of vertices in community C, respectively.

B.3 Permutation Feature Importance

In Stage 3, SINC ranks the detected communities based on their impact on predictive accuracy.

Definition 3 (Permutation Feature Importance) *Given a trained model f that predicts $\mathbf{y}$, the baseline error of the model is:*

$$e_{base} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y^{(i)}, f(\mathbf{x}^{(i)}) \right)$$

where N is the number of instances, $\mathcal{L}$ denotes the loss function, and $\mathbf{x}^{(i)}$ represents the feature vector for the i'th instance.

Let $\mathcal{C} = \{C_1, \ldots, C_q\}$ denote a partition of the feature set into communities. For a given community C_j, a permuted feature vector $\mathbf{x}^{(i)}_{perm,j}$ is constructed by randomly permuting the features in C_j across instances, while keeping all other features unchanged. The resulting prediction error is:

$$e_{perm,j} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y^{(i)}, f(\mathbf{x}^{(i)}_{perm,j}) \right)$$

The global importance of community C_j is then defined as the difference between the permuted error and the baseline error (Fisher et al., 2019, p.7), (Molnar, 2025):

$$I_j = e_{perm,j} - e_{base}$$

C SHAP RESULTS ON CONVERGED MLP

Table 7: SHAP global feature importance and pairwise interaction strength for the converged model. Interaction strengths less than 2% to the total strength are grouped for brevity.

(a) Feature importance

Feature	**Mean |SHAP| (%)**
E	2.28 (23.45)
A	1.65 (16.98)
C	1.54 (15.88)
B	1.52 (15.59)
D	1.30 (13.34)
F	1.25 (12.83)
Noise - 3	0.07 (0.69)
Noise - 1	0.04 (0.45)
Noise - 2	0.04 (0.44)
Noise - 4	0.03 (0.35)

(b) Pairwise interaction strength

Feature Pair	**Strength (%)**
A + B	0.79 (15.25)
A + C	0.77 (14.86)
B + C	0.71 (13.81)
B + D	0.63 (12.12)
C + D	0.63 (12.11)
A + D	0.61 (11.88)
Other 39 pairs (each)	< 0.07 (< 2.00)

D UNDERTRAINED MLP ANALYSIS

This section contains the full tabular results for SHAP, Archipelago, and SINC when evaluated on the intentionally undertrained MLP model (trained for 20 epochs, $R^2 = 0.8721$). As discussed in the main text, these results demonstrate that prior to convergence, none of the baseline methods nor SINC detect the ground-truth synergy, confirming that SINC's detection of the community in the fully trained model is a reflection of learned weights rather than a false positive grouping.

Table 8: SHAP global feature importance and pairwise interaction strength for the undertrained model. Interactions contributing less than 2% to the total strength are grouped for brevity.

(a) Feature importance

Feature	**Mean |SHAP| (%)**
E	2.03 (19.41)
A	1.78 (17.02)
F	1.54 (14.69)
C	1.48 (14.10)
B	1.31 (12.56)
D	1.12 (10.69)
Noise - 1	0.39 (3.76)
Noise - 4	0.38 (3.66)
Noise - 2	0.29 (2.81)
Noise - 3	0.14 (1.31)

(b) Pairwise interaction strength

Feature Pair	**Strength (%)**
E + F	0.18 (6.79)
B + E	0.15 (5.89)
C + E	0.15 (5.84)
A + E	0.15 (5.71)
D + E	0.14 (5.28)
A + B	0.13 (5.06)
B + C	0.12 (4.57)
A + C	0.12 (4.41)
A + F	0.11 (4.35)
B + F	0.11 (4.33)
C + F	0.11 (4.30)
B + D	0.09 (3.47)
A + D	0.09 (3.46)
C + D	0.09 (3.40)
D + F	0.09 (3.34)
Other 30 pairs (each)	< 0.05 (< 2.00)

Table 9: Global feature communities detected by Archipelago ($top_k = 5$) on the undertrained MLP. Appearance count indicates how many of the 100 evaluated instances produced this exact community. Communities contributing less than 2% to the total attribution are grouped for brevity.

Feature Community	**Size**	**Global Attribution (%)**	**Appearance Count**
E	1	0.64 (10.07)	51
F	1	0.34 (5.26)	41
C	1	0.31 (4.85)	37
A	1	0.30 (4.65)	30
Noise - 1	1	0.22 (3.42)	64
A + B + C + E + F	5	0.20 (3.17)	4
Noise - 4	1	0.20 (3.12)	54
Noise - 2	1	0.20 (3.08)	68
A + B + C + D + E	5	0.16 (2.52)	3
D	1	0.15 (2.29)	35
B	1	0.15 (2.29)	25
D + E	2	0.13 (2.01)	2
A + C + D + E + F + Noise - 2	6	0.13 (2.00)	1
A + B + C + F	4	0.13 (2.00)	3
Other 89 communities (each)	-	< 0.12 (< 2.00)	-

Table 10: Detected feature groups and their impact on model output as identified by SINC on the undertrained MLP.

Group	**Size**	**Global Impact (%)**
E	1	11.67 (32.77)
A	1	6.37 (17.87)
F	1	5.16 (14.49)
B	1	4.74 (13.30)
C	1	4.60 (12.93)
D	1	2.98 (8.35)
Noise (1-4)	4	0.11 (0.30)

E California Housing Dataset Analysis

To assess the interpretive validity of SINC on real-world data, we evaluate an XGBoost model (Chen & Guestrin, 2016) (1000 trees, max depth 6, learning rate 0.1) trained on the California Housing dataset (Inria, 2012). In this widely used XAI benchmark, latitude and longitude are known to exhibit the strongest interaction when predicting median house value, while variables such as median income contribute primarily through individual effects (Lundberg & Lee, 2017).

Consistent with the synthetic experiments, SHAP distributes importance across individual features and isolated pairs, lacking a mechanism to form a global community structure.

Table 11: SHAP global feature importance and pairwise interaction strength on the California Housing dataset with an XGBoost model. Interactions contributing less than 2% to the total strength are grouped for brevity.

(a) Feature importance

Feature	**Mean |SHAP| (%)**
Longitude	0.47 (25.97)
Latitude	0.45 (25.03)
MedInc	0.40 (22.23)
AveOccup	0.21 (11.53)
AveRooms	0.14 (7.72)
HouseAge	0.06 (3.51)
Population	0.04 (2.08)
AveBedrms	0.03 (1.93)

(b) Pairwise interaction strength

Feature Pair	**Strength (%)**
Latitude + Longitude	0.44 (28.49)
MedInc + Longitude	0.10 (6.32)
MedInc + AveRooms	0.09 (5.57)
MedInc + AveOccup	0.07 (4.77)
HouseAge + AveOccup	0.07 (4.42)
MedInc + Latitude	0.06 (4.02)
AveOccup + Latitude	0.06 (3.82)
MedInc + HouseAge	0.05 (3.33)
AveOccup + Longitude	0.05 (3.28)
AveRooms + Latitude	0.05 (3.24)
HouseAge + Latitude	0.05 (2.93)
MedInc + AveBedrms	0.04 (2.81)
AveRooms + AveBedrms	0.04 (2.79)
AveRooms + Longitude	0.04 (2.36)
HouseAge + Longitude	0.03 (2.27)
Other 13 pairs (each)	< 0.03 (< 2.00)

While Archipelago successfully identifies the Latitude + Longitude interaction as its top community (Table 12), its reliance on aggregating local explanations introduces instability. Not only do the detected communities fail to appear consistently across all evaluated instances, but the method also produces overlapping, competing subsets (e.g., isolating MedInc while simultaneously surfacing MedInc + AveRooms). Consequently, Archipelago fails to establish a clean, unified global partition of the feature space.

By contrast, applying SINC ($\tau = 0.05, \gamma = 0.01$) yields a single interaction network (Fig. 4). The algorithm partitions the features into four distinct, domain-aligned communities: the synergistic Latitude + Longitude pair, two additive features (MedInc and AveRooms), and a community containing the four remaining property and population variables. Importantly, SINC identifies a global hierarchy among the detected communities, ranking the Latitude + Longitude synergy as the dominant driver of model predictions with 53.17% impact, followed by MedInc at 29.75% (Table 13).

Table 12: Global feature communities detected by Archipelago ($top_k = 1$) on the California Housing dataset using an XGBoost model. Appearance count indicates how many of the 100 evaluated instances produced this exact community. Communities contributing less than 2% to the total attribution are grouped for brevity.

Feature Community	**Size**	**Global Attribution (%)**	**Appearance Count**
Latitude + Longitude	2	0.78 (37.11)	78
MedInc	1	0.35 (16.49)	86
AveOccup	1	0.23 (10.78)	92
Longitude	1	0.12 (5.82)	17
MedInc + AveRooms	2	0.10 (4.54)	7
Latitude	1	0.09 (4.09)	20
AveRooms	1	0.07 (3.53)	89
HouseAge	1	0.07 (3.44)	98
AveRooms + Longitude	2	0.07 (3.13)	3
Population	1	0.06 (2.70)	100
AveBedrms	1	0.05 (2.60)	98
MedInc + Latitude	2	0.05 (2.14)	1
Other 7 communities (each)	-	< 0.02 (< 2.00)	-

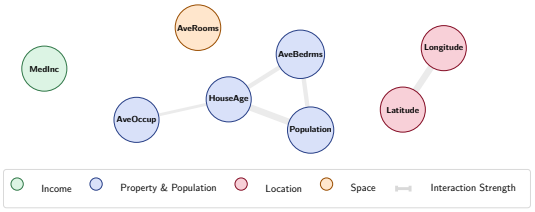


Figure 4: Communities detected by SINC on the XGBoost model trained on the California Housing dataset.

Table 13: Detected feature groups and their predictive impact as identified by SINC on the California Housing dataset using an XGBoost model.

Group	**Size**	**Global Impact (%)**
Latitude + Longitude	2	0.85 (53.17)
MedInc	1	0.48 (29.75)
AveBedrms + AveOccup + HouseAge + Population	4	0.14 (8.66)
AveRooms	1	0.14 (8.43)

F HIGGS DATASET ANALYSIS

This section contains the full tabular results for SHAP, Archipelago, and SINC when evaluated on the HIGGS dataset. The HIGGS dataset is a binary classification benchmark from high-energy physics designed to distinguish Higgs boson events from background processes. Using an XGBoost classifier (1000 trees, max depth 6, learning rate 0.1, cross-entropy = 0.56), all three methods were evaluated on the same 100-sample validation subset to ensure a consistent basis for comparing feature importance and interaction structure.

Table 14 reports the global feature importance and pairwise interaction strengths obtained with SHAP. The feature importance results show that several of the kinematic variables have substantial individual contributions, with m_bb, m_wwbb, and m_wbb among the most important features. However, the pairwise interaction results are much more concentrated: only $m_bb + m_wbb$ and $m_wbb + m_wwbb$ account for more than 2% of the total interaction strength, while the remaining 376 feature pairs contribute less than 2% each. Thus, although SHAP identifies several individually important variables, its pairwise interaction representation does not provide a unified view of the larger groups of variables that jointly influence the model.

Table 15 shows the global communities obtained by aggregating Archipelago explanations across the 100 evaluated instances. Archipelago identifies substantially larger feature groups than the pairwise SHAP analysis, with many communities containing several of the same kinematic variables. However, these communities are not consistent across instances. The most frequently occurring high-attribution community appears in only 5 of the 100 instances, while the majority of the reported communities occur only once. The resulting collection of overlapping subsets therefore provides limited evidence for a single global interaction structure. Instead, similar kinematic variables repeatedly appear in different local combinations.

Table 16 presents the feature communities identified by SINC. In contrast to the instance-specific Archipelago results, SINC produces a single global partition of the 28 features into a small number of communities. The two largest communities account for 41.54% and 40.47% of the total impact, respectively, and group primarily kinematic variables. The third community, consisting of $jet1pt + jet2pt$, contributes a further 15.03%. Together, these three communities account for 97.04% of the total impact identified by SINC. The remaining angular and tagging variables are concentrated into a low-impact community contributing 1.98%, while the remaining individual features have negligible impact.

The structure in Table 16 is also more readily interpretable in terms of the underlying feature semantics. The dominant communities group related kinematic measurements, while the angular and tagging variables are largely separated into a lower-impact group. Thus, rather than producing a collection of instance-dependent feature subsets, SINC provides a compact global representation of how the XGBoost model organizes the available measurements.

Overall, Tables 14 - 16 illustrate the difference between pairwise interaction measurement, aggregation of local interaction explanations, and global community discovery. SHAP provides detailed individual and pairwise importance estimates, Archipelago identifies larger but varying local communities, and SINC combines global interaction structure with community-level impact to produce a single partition of the feature space.

Table 14: SHAP global feature importance and pairwise interaction strength on the HIGGS dataset using an XGBoost model. Interactions contributing less than 2% to the total strength are grouped for brevity.

(a) Feature importance

Feature	**Mean |SHAP| (%)**
m_bb	0.68 (13.55)
m_wwbb	0.66 (13.32)
m_wbb	0.44 (8.86)
m_jjj	0.38 (7.56)
jet1pt	0.33 (6.54)
m_jlv	0.32 (6.46)
lepton_pT	0.19 (3.90)
missing_energy_magnitude	0.19 (3.74)
m_jj	0.17 (3.39)
jet2pt	0.13 (2.62)
jet4pt	0.12 (2.35)
jet3pt	0.12 (2.32)
jet4b-tag	0.11 (2.28)
lepton_eta	0.10 (2.08)
m_lv	0.10 (1.94)
jet2eta	0.10 (1.91)
jet1eta	0.09 (1.82)
jet3b-tag	0.09 (1.74)
jet4phi	0.08 (1.65)
jet3eta	0.08 (1.60)
jet4eta	0.08 (1.60)
jet3phi	0.07 (1.44)
jet2b-tag	0.07 (1.42)
missing_energy_phi	0.07 (1.37)
jet2phi	0.06 (1.29)
jet1phi	0.06 (1.19)
lepton_phi	0.06 (1.12)
jet1b-tag	0.05 (0.94)

(b) Pairwise interaction strength

Feature Pair	**Strength (%)**
m_bb + m_wbb	0.30 (2.96)
m_wbb + m_wwbb	0.21 (2.09)
Other 376 pairs (each)	< 0.17 (< 2.00)

Table 15: Global feature communities detected by Archipelago ($top_k = 6$) on the HIGGS dataset using an XGBoost model. Appearance count indicates how many of the 100 evaluated instances produced this exact community. Communities contributing less than 2% to the total attribution are grouped for brevity.

Feature Community	Size	Global Attribution (%)	Appearance Count
lepton_pT + m_jjj + m_lv + m_jlv + m_bb + m_wbb + m_wwbb	7	0.0180 (9.64)	2
m_jjj + m_lv + m_jlv + m_bb + m_wbb + m_wwbb	6	0.0163 (8.73)	3
jet1pt + m_jjj + m_lv + m_jlv + m_wbb + m_wwbb	6	0.0122 (6.54)	2
m_jj + m_jjj + m_lv + m_jlv + m_bb + m_wbb + m_wwbb	7	0.0102 (5.47)	5
lepton_pT + m_jjj + m_jlv + m_bb + m_wwbb	5	0.0100 (5.33)	1
jet1pt + m_jj + m_jjj + m_jlv + m_bb + m_wbb + m_wwbb	7	0.0099 (5.30)	1
jet3eta + m_jjj + m_lv + m_jlv + m_bb + m_wwbb	6	0.0096 (5.12)	1
lepton_eta + m_jjj + m_lv + m_jlv + m_bb + m_wwbb	6	0.0092 (4.94)	1
missing_energy_magnitude + jet4phi + m_jjj + m_lv m_jlv + m_bb + m_wwbb	7	0.0091 (4.85)	1
lepton_pT + m_jjj + m_lv + m_jlv + m_bb + m_wwbb	6	0.0087 (4.64)	1
missing_energy_magnitude + m_jjj + m_jlv + m_bb + m_wbb + m_wwbb	6	0.0080 (4.25)	1
jet1pt + jet2eta + m_jjj + m_lv + m_jlv + m_wbb	6	0.0067 (3.56)	1
jet1pt + m_jj + m_jjj + m_jlv + m_bb	5	0.0059 (3.14)	1
m_jj + m_jjj + m_lv + m_jlv + m_wbb + m_wwbb	6	0.0058 (3.09)	1
jet4pt + m_jj + m_jjj + m_lv + m_bb + m_wbb + m_wwbb	7	0.0043 (2.29)	1
Other 136 communities (each)	-	< 0.0035 (< 2.00)	-

Table 16: Detected feature groups and their impact on model output as identified by SINC ($\tau = 0.05$ and $\gamma = 0.01$) on the HIGGS dataset.

Group	Size	Global Impact (%)
m_bb + m_jj + m_jjj + m_wbb	4	0.26 (41.54)
lepton_pT + m_jlv + m_lv + m_wwbb + missing_energy_magnitude	5	0.26 (40.47)
jet1pt + jet2pt	2	0.10 (15.03)
jet1eta + jet1phi + jet2eta + jet2phi + jet3b-tag + jet3eta + jet3phi + jet4b-tag + jet4eta + jet4phi + lepton_eta + lepton_phi + missing_energy_phi	13	0.01 (1.98)
jet1b-tag	1	0.01 (0.86)
jet4pt	1	0.00 (0.11)
jet2b-tag	1	0.00 (0.02)
jet3pt	1	0.00 (0.00)

G BIKE SHARING DEMAND ANALYSIS

This section contains the full tabular results for SHAP, Archipelago, and SINC when evaluated on the UCI Bike Sharing Demand dataset. The dataset is a tabular regression benchmark in which hourly bike rental counts are predicted from environmental, seasonal, and calendar-related variables. We evaluate an XGBoost regressor (1000 trees, maximum depth 6, learning rate 0.1, MSE = 1497.18). To provide a consistent basis for comparison, all three explanation methods are evaluated on the same subset of 100 validation samples.

Table 17 reports the global feature importance and pairwise interaction strengths obtained with SHAP. Hour is by far the most important individual feature, followed by year and temperature. The pairwise interaction results reveal that hour participates in many of the strongest interactions: the largest interaction is hour + workingday, followed by year + hour, hour + temp, and hour + weekday. Thus, the SHAP results provide evidence that the predictive role of hour depends on several other temporal, calendar, and environmental variables.

Table 17: SHAP global feature importance and pairwise interaction strength on the Bike Sharing dataset using an XGBoost model. Interactions contributing less than 2% to the total strength are grouped for brevity.

(a) Feature importance

Feature	**Mean |SHAP| (%)**
hour	116.01 (42.34)
year	42.68 (15.58)
temp	32.96 (12.03)
workingday	19.50 (7.12)
humidity	12.23 (4.46)
month	10.73 (3.92)
feel_temp	10.64 (3.88)
season	9.71 (3.54)
weekday	8.55 (3.12)
weather	7.06 (2.58)
windspeed	3.10 (1.13)
holiday	0.84 (0.30)

(b) Pairwise interaction strength

Feature Pair	**Strength (%)**
hour + workingday	38.90 (15.52)
year + hour	25.87 (10.32)
hour + temp	20.68 (8.25)
hour + weekday	12.70 (5.07)
hour + humidity	10.09 (4.03)
hour + feel_temp	8.72 (3.48)
year + temp	6.80 (2.71)
month + hour	6.47 (2.58)
season + hour	5.96 (2.38)
month + temp	5.73 (2.28)
workingday + temp	5.37 (2.14)
temp + feel_temp	5.34 (2.13)
season + temp	5.23 (2.09)
hour + weather	5.12 (2.04)
year + month	5.10 (2.03)
Other 51 pairs (each)	< 4.98 (< 2.00)

Table 18 shows the feature communities obtained by aggregating Archipelago explanations across the 100 validation instances. The most frequently recurring community is year + hour + workingday, which appears in 21 of the 100 instances. However, several alternative communities involving overlapping subsets of the same variables are also detected, including hour + workingday + feel_temp, year + hour + workingday + temp, and year + hour + weekday + temp. These competing subsets indicate that the interaction structure varies across individual instances rather than forming a single consistent grouping. In addition, many of the individual features appear frequently as standalone communities, further illustrating the difficulty of obtaining a unified global partition from locally generated explanations.

Table 19 presents the global communities identified by SINC. In contrast to the instance-dependent Archipelago results, SINC produces a single partition of the feature space. The dominant community, hour + workingday, accounts for 75.41% of the total community impact. This grouping has a natural interpretation: the predictive effect of the hour of day can depend strongly on whether the day is a working day, reflecting different commuting and activity patterns. SINC then isolates year as an individual community with 10.78% impact, capturing a broader temporal trend. The remaining variables are divided into smaller groups, including a secondary community containing feel_temp, holiday, month, weekday, and windspeed, as well as a distinct humidity + weather com-

munity. These groups separate the dominant hour-dependent behavior from additional calendar and environmental effects.

Table 18: Global feature communities detected by Archipelago ($top_k = 3$) on the Bike Sharing dataset using an XGBoost model. Appearance count indicates how many of the 100 evaluated instances produced this exact community. Communities contributing less than 2% to the total attribution are grouped for brevity.

Feature Community	**Size**	**Global Attribution (%)**	**Appearance Count**
year + hour + workingday	3	44.87 (18.24)	21
temp	1	20.02 (8.14)	52
month	1	16.11 (6.55)	82
season	1	15.74 (6.40)	75
feel_temp	1	9.56 (3.89)	79
hour + workingday + feel_temp	3	8.97 (3.64)	5
year + hour + workingday + temp	4	8.54 (3.47)	5
humidity	1	6.25 (2.54)	84
year + hour + weekday + temp	4	6.17 (2.51)	4
windspeed	1	5.92 (2.40)	100
hour	1	5.52 (2.24)	16
month + hour + weekday + temp	4	5.42 (2.20)	3
season + year + hour + workingday	4	4.96 (2.01)	2
Other 51 communities (each)	-	< 4.90 (< 2.00)	-

Table 19: Detected feature groups and their impact on model output as identified by SINC ($\tau = 0.05$ and $\gamma = 0.01$) on the Bike Sharing dataset using an XGBoost model.

Group	**Size**	**Global Impact (%)**
hour + workingday	2	56875.35 (75.41)
year	1	8132.66 (10.78)
feel_temp + holiday + month + weekday + windspeed	5	4061.76 (5.39)
temp	1	3502.13 (4.64)
humidity + weather	2	2179.07 (2.89)
season	1	675.28 (0.90)

Overall, the results in Tables 17 - 19 illustrate the distinction between measuring pairwise interactions, aggregating local interaction explanations, and identifying a single global interaction structure. The strongest interaction identified by SHAP, hour + workingday, is also recovered as the dominant global community by SINC, while SINC additionally organizes the remaining features into a small number of global groups. The resulting structure is consistent with the temporal, calendar, and environmental factors represented by the dataset and provides a more compact description of the relationships learned by the model.

H SYNTHETIC DATA

This section provides the complete mathematical definitions and detailed empirical results for the synthetic benchmark suite evaluated in the main text. Table 20 explicitly defines the ten data-generating functions, which encompass a variety of complex topological dependencies, including higher-order non-linearities and overlapping synergies.

Table 20: Test suite of data-generating functions

$F_1(\mathbf{x})$	$\pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7$
$F_2(\mathbf{x})$	$\pi^{x_1 x_2}\sqrt{2\lvert x_3\rvert} - \sin^{-1}(0.5x_4) + \log(\lvert x_3 + x_5\rvert + 1) + \frac{x_9}{1+\lvert x_{10}\rvert}\sqrt{\frac{x_7}{1+\lvert x_8\rvert}} - x_2 x_7$
$F_3(\mathbf{x})$	$\exp\lvert x_1 - x_2\rvert + \lvert x_2 x_3\rvert - x_3^{2\lvert x_4\rvert} + \log(x_4^2 + x_5^2 + x_7^2 + x_8^2) + x_9 + \frac{1}{1+x_{10}^2}$
$F_4(\mathbf{x})$	$\exp\lvert x_1 - x_2\rvert + \lvert x_2 x_3\rvert - x_3^{2\lvert x_4\rvert} + (x_1 x_4)^2 + \log(x_4^2 + x_5^2 + x_7^2 + x_8^2) + x_9 + \frac{1}{1+x_{10}^2}$
$F_5(\mathbf{x})$	$\frac{1}{1+x_1^2+x_2^2+x_3^2} + \sqrt{\exp(x_4 + x_5)} + \lvert x_6 + x_7\rvert + x_8 x_9 x_{10}$
$F_6(\mathbf{x})$	$\exp(\lvert x_1 x_2\rvert + 1) - \exp(\lvert x_3 + x_4\rvert + 1) + \cos(x_5 + x_6 - x_8) + \sqrt{x_8^2 + x_9^2 + x_{10}^2}$
$F_7(\mathbf{x})$	$(\arctan(x_1) + \arctan(x_2))^2 + \max(x_3 x_4 + x_6, 0) - \frac{1}{1+(x_4 x_5 x_6 x_7 x_8)^2} + \left(\frac{\lvert x_7\rvert}{1+\lvert x_9\rvert}\right)^5 + \sum_{i=1}^{10} x_i$
$F_8(\mathbf{x})$	$x_1 x_2 + 2^{x_3+x_5+x_6} + 2^{x_3+x_4+x_5+x_7} + \sin(x_7 \sin(x_8 + x_9)) + \arccos(0.9x_{10})$
$F_9(\mathbf{x})$	$\tanh(x_1 x_2 + x_3 x_4)\sqrt{\lvert x_5\rvert} + \exp(x_5 + x_6) + \log((x_6 x_7 x_8)^2 + 1) + x_9 x_{10} + \frac{1}{1+\lvert x_{10}\rvert}$
$F_{10}(\mathbf{x})$	$\sinh(x_1 + x_2) + \arccos(\tanh(x_3 + x_5 + x_7)) + \cos(x_4 + x_5) + \sec(x_7 x_9)$

For each of the ten functions, the subsequent subsections present the full tabular results generated by SINC ($\tau = 0.0001$, $\gamma = 0.01$), SHAP, and Archipelago. These tables detail the specific feature communities or pairs identified by each method, alongside their calculated global impact, interaction strength, or importance percentage.

Results for F1

Table 21: Global SHAP pairwise interaction strengths on F_1 contributing $\geq 2\%$ of the total (8 of 45 pairs).

Feature Pair	Strength (%)
x1 + x2	0.150681 (27.28)
x2 + x7	0.074355 (13.46)
x7 + x9	0.067019 (12.13)
x1 + x3	0.056368 (10.20)
x2 + x3	0.053010 (9.60)
x9 + x10	0.029872 (5.41)
x7 + x10	0.017860 (3.23)
x8 + x9	0.013635 (2.47)

Table 22: Global feature communities detected by Archipelago on F_1 contributing $\geq 2\%$ of the total (13 of 41 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x4	1	0.140387 (13.32)	39
x1 + x2 + x3 + x7 + x9	5	0.134478 (12.76)	11
x1 + x2 + x3 + x7 + x9 + x10	6	0.086893 (8.25)	4
x1 + x2 + x3	3	0.077052 (7.31)	4
x5	1	0.045793 (4.35)	43
x1 + x2 + x3 + x5 + x9	5	0.045550 (4.32)	2
x1 + x2 + x3 + x5	4	0.045047 (4.27)	2
x1 + x3 + x7 + x9 + x10	5	0.036560 (3.47)	2
x2 + x3 + x4 + x7 + x8 + x9	6	0.032650 (3.10)	1
x7 + x9 + x10	3	0.025246 (2.40)	3
x1 + x2 + x4 + x7 + x9 + x10	6	0.023502 (2.23)	1
x1 + x2 + x3 + x4	4	0.022268 (2.11)	2
x2 + x3 + x5 + x7 + x9	5	0.022188 (2.11)	1

Table 23: Detected feature groups and their impact on model output as identified by SINC on F_1.

Group	Size	Global Impact (%)
x1 + x2 + x3 + x7 + x8 + x9 + x10	7	1.542099 (92.98)
x4	1	0.097852 (5.90)
x5	1	0.018491 (1.11)
x6	1	0.000028 (0.00)

Results for F2

Table 24: Global SHAP pairwise interaction strengths on F_2 contributing $\geq 2\%$ of the total (7 of 45 pairs).

Feature Pair	Strength (%)
x1 + x2	0.237629 (25.36)
x2 + x7	0.231852 (24.75)
x3 + x5	0.213308 (22.77)
x7 + x9	0.059283 (6.33)
x2 + x3	0.031416 (3.35)
x1 + x3	0.030707 (3.28)
x9 + x10	0.028695 (3.06)

Table 25: Global feature communities detected by Archipelago on F_2 contributing $\geq 2\%$ of the total (12 of 38 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x1 + x2 + x3 + x5 + x7	5	0.209408 (17.32)	10
x1 + x2 + x3 + x5	4	0.145840 (12.06)	5
x4	1	0.142664 (11.80)	41
x3 + x5	2	0.128514 (10.63)	5
x1 + x2 + x3 + x7 + x9	5	0.093062 (7.70)	5
x5	1	0.071778 (5.94)	15
x1 + x3 + x5 + x7 + x9 + x10	6	0.058154 (4.81)	2
x1 + x2 + x3 + x7 + x9 + x10	6	0.056485 (4.67)	4
x1 + x2 + x3 + x5 + x10	5	0.028530 (2.36)	1
x1 + x3 + x4 + x5	4	0.026413 (2.18)	1
x2 + x3 + x5 + x7 + x8 + x9	6	0.025987 (2.15)	1
x3 + x5 + x8 + x9 + x10	5	0.024400 (2.02)	1

Table 26: Detected feature groups and their impact on model output as identified by SINC on F_2.

Group	Size	Global Impact (%)
x1 + x2 + x7 + x9	4	0.671344 (48.73)
x3 + x5	2	0.525411 (38.14)
x4	1	0.176484 (12.81)
x6 + x8 + x10	3	0.004407 (0.32)

Results for F3

Table 27: Global SHAP pairwise interaction strengths on F_3 contributing $\geq 2\%$ of the total (9 of 45 pairs).

Feature Pair	Strength (%)
x1 + x2	0.618411 (45.91)
x5 + x8	0.084715 (6.29)
x5 + x7	0.078864 (5.85)
x7 + x8	0.078627 (5.84)
x4 + x5	0.068229 (5.07)
x4 + x7	0.065318 (4.85)
x2 + x3	0.063951 (4.75)
x4 + x8	0.062248 (4.62)
x3 + x4	0.044136 (3.28)

Table 28: Global feature communities detected by Archipelago on F_3 contributing $\geq 2\%$ of the total (11 of 33 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x4 + x5 + x7 + x8	4	2.755349 (38.93)	26
x1 + x2	2	0.507518 (7.17)	17
x9	1	0.498477 (7.04)	49
x1	1	0.293139 (4.14)	20
x3 + x4 + x5 + x8	4	0.274252 (3.88)	3
x10	1	0.263562 (3.72)	49
x3 + x4 + x5 + x7 + x8	5	0.248863 (3.52)	3
x1 + x2 + x3 + x4 + x7 + x8	6	0.197173 (2.79)	2
x3 + x4 + x5 + x7	4	0.169398 (2.39)	2
x2	1	0.166028 (2.35)	21
x1 + x2 + x4 + x7 + x8	5	0.153736 (2.17)	1

Table 29: Detected feature groups and their impact on model output as identified by SINC on F_3.

Group	Size	Global Impact (%)
x1 + x2 + x3	3	3.191091 (66.63)
x4 + x5 + x7 + x8	4	0.894369 (18.67)
x9	1	0.652498 (13.62)
x10	1	0.051352 (1.07)
x6	1	-0.000030 (0.00)

Results for F4

Table 30: Global SHAP pairwise interaction strengths on F_4 contributing $\geq 2\%$ of the total (10 of 45 pairs).

Feature Pair	Strength (%)
x1 + x2	0.621225 (44.36)
x5 + x8	0.080907 (5.78)
x7 + x8	0.078373 (5.60)
x5 + x7	0.076332 (5.45)
x4 + x5	0.064880 (4.63)
x2 + x3	0.062489 (4.46)
x4 + x7	0.061799 (4.41)
x4 + x8	0.059963 (4.28)
x3 + x4	0.056132 (4.01)
x1 + x4	0.052797 (3.77)

Table 31: Global feature communities detected by Archipelago on F_4 contributing $\geq 2\%$ of the total (13 of 35 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x4 + x5 + x7 + x8	4	2.269846 (34.14)	23
x1 + x2 + x4 + x7 + x8	5	0.586535 (8.82)	4
x10	1	0.367206 (5.52)	48
x9	1	0.355328 (5.34)	42
x1 + x2	2	0.271375 (4.08)	12
x1 + x2 + x3	3	0.216615 (3.26)	6
x5 + x7 + x8	3	0.213298 (3.21)	3
x1	1	0.213218 (3.21)	17
x4 + x5 + x7	3	0.194447 (2.92)	2
x1 + x2 + x3 + x4 + x5 + x8	6	0.190380 (2.86)	2
x4 + x5 + x8	3	0.182647 (2.75)	2
x5 + x7 + x8 + x9	4	0.170155 (2.56)	2
x1 + x2 + x4 + x5 + x8	5	0.134416 (2.02)	1

Table 32: Detected feature groups and their impact on model output as identified by SINC on F_4.

Group	Size	Global Impact (%)
x1 + x2	2	3.399013 (64.60)
x4 + x5 + x7 + x8	4	1.108469 (21.07)
x9	1	0.660342 (12.55)
x10	1	0.050721 (0.96)
x3	1	0.043452 (0.83)
x6	1	0 (0.00)

Results for F5

Table 33: Global SHAP pairwise interaction strengths on F_5 contributing $\geq 2\%$ of the total (8 of 45 pairs).

Feature Pair	Strength (%)
x6 + x7	0.399118 (52.12)
x8 + x9	0.072125 (9.42)
x9 + x10	0.069554 (9.08)
x8 + x10	0.064406 (8.41)
x4 + x5	0.056943 (7.44)
x2 + x3	0.019361 (2.53)
x1 + x3	0.018743 (2.45)
x1 + x2	0.015955 (2.08)

Table 34: Global feature communities detected by Archipelago on F_5 contributing $\geq 2\%$ of the total (10 of 31 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x1 + x2 + x3	3	0.300729 (18.17)	27
x6	1	0.202759 (12.25)	19
x6 + x7	2	0.201891 (12.20)	22
x7	1	0.162119 (9.79)	20
x4 + x5	2	0.134840 (8.15)	17
x5	1	0.130251 (7.87)	30
x8 + x9 + x10	3	0.122115 (7.38)	26
x4	1	0.110895 (6.70)	31
x1 + x2	2	0.042756 (2.58)	5
x1 + x3	2	0.041682 (2.52)	4

Table 35: Detected feature groups and their impact on model output as identified by SINC on F_5.

Group	Size	Global Impact (%)
x6 + x7	2	0.452188 (46.81)
x4 + x5	2	0.402681 (41.68)
x8 + x9 + x10	3	0.072584 (7.51)
x1 + x2 + x3	3	0.038651 (4.00)

Results for F6

Table 36: Global SHAP pairwise interaction strengths on F_6 contributing $\geq 2\%$ of the total (5 of 45 pairs).

Feature Pair	Strength (%)
x3 + x4	1.846104 (61.80)
x1 + x2	0.250742 (8.39)
x6 + x8	0.203263 (6.80)
x5 + x6	0.197146 (6.60)
x5 + x8	0.193450 (6.48)

Table 37: Global feature communities detected by Archipelago on F_6 contributing $\geq 2\%$ of the total (7 of 31 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x3 + x4	2	1.534940 (41.36)	35
x1 + x2	2	0.716791 (19.32)	33
x1 + x2 + x3 + x4	4	0.273470 (7.37)	5
x3 + x4 + x5 + x6 + x8	5	0.175814 (4.74)	1
x9 + x10	2	0.158903 (4.28)	12
x3 + x4 + x8 + x9 + x10	5	0.120131 (3.24)	1
x5 + x6 + x8 + x10	4	0.085228 (2.30)	10

Table 38: Detected feature groups and their impact on model output as identified by SINC on F_6.

Group	Size	Global Impact (%)
x3 + x4	2	19.435741 (90.42)
x1 + x2	2	1.568814 (7.30)
x5 + x6 + x8	3	0.389224 (1.81)
x9 + x10	2	0.101569 (0.47)
x7	1	0.000268 (0.00)

Results for F7

Table 39: Global SHAP pairwise interaction strengths on F_7 contributing $\geq 2\%$ of the total (5 of 45 pairs).

Feature Pair	Strength (%)
x1 + x2	0.324175 (47.46)
x3 + x4	0.116262 (17.02)
x3 + x6	0.056506 (8.27)
x4 + x6	0.052429 (7.68)
x7 + x9	0.032815 (4.80)

Table 40: Global feature communities detected by Archipelago on F_7 contributing $\geq 2\%$ of the total (10 of 47 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x10	1	0.544164 (16.70)	47
x7 + x9	2	0.359511 (11.03)	23
x5	1	0.306854 (9.41)	37
x8	1	0.267537 (8.21)	36
x1 + x2	2	0.179203 (5.50)	17
x3 + x4 + x6	3	0.157434 (4.83)	14
x1 + x2 + x5	3	0.154388 (4.74)	3
x6	1	0.132071 (4.05)	8
x1 + x2 + x3 + x4 + x6	5	0.111428 (3.42)	4
x9	1	0.093970 (2.88)	13

Table 41: Detected feature groups and their impact on model output as identified by SINC on F_7.

Group	Size	Global Impact (%)
x1 + x2	2	1.888042 (22.89)
x6	1	1.627402 (19.73)
x3 + x4	2	1.380206 (16.73)
x5	1	0.685387 (8.31)
x7	1	0.679014 (8.23)
x9	1	0.673454 (8.16)
x8	1	0.663081 (8.04)
x10	1	0.652444 (7.91)

Results for F8

Table 42: Global SHAP pairwise interaction strengths on F_8 contributing $\geq 2\%$ of the total (11 of 45 pairs).

Feature Pair	Strength (%)
x3 + x5	0.353538 (16.95)
x1 + x2	0.211579 (10.14)
x7 + x9	0.183296 (8.79)
x3 + x6	0.172306 (8.26)
x7 + x8	0.169648 (8.13)
x3 + x7	0.169041 (8.10)
x5 + x7	0.168299 (8.07)
x5 + x6	0.165749 (7.95)
x4 + x5	0.134904 (6.47)
x4 + x7	0.128624 (6.17)
x3 + x4	0.124626 (5.97)

Table 43: Global feature communities detected by Archipelago on F_8 contributing $\geq 2\%$ of the total (13 of 45 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x10	1	0.548616 (23.67)	48
x1 + x2	2	0.203521 (8.78)	31
x3 + x4 + x5 + x6	4	0.184643 (7.97)	6
x3 + x4 + x5 + x7 + x8 + x9	6	0.147525 (6.37)	4
x3 + x5 + x6	3	0.127780 (5.51)	7
x4 + x5 + x7 + x8 + x9	5	0.116845 (5.04)	5
x3 + x5 + x6 + x7 + x8	5	0.109623 (4.73)	1
x6	1	0.094146 (4.06)	22
x3 + x5 + x7 + x8	4	0.081763 (3.53)	1
x3 + x4 + x5 + x6 + x7	5	0.077032 (3.32)	3
x3 + x5 + x7 + x8 + x9	5	0.058762 (2.54)	1
x1 + x2 + x3 + x5 + x6	5	0.050483 (2.18)	1
x1 + x2 + x10	3	0.048991 (2.11)	2

Table 44: Detected feature groups and their impact on model output as identified by SINC on F_8.

Group	Size	Global Impact (%)
x3 + x5 + x6	3	5.186367 (66.98)
x4 + x7 + x8 + x9	4	1.674264 (21.62)
x10	1	0.657565 (8.49)
x1 + x2	2	0.224784 (2.90)

Results for F9

Table 45: Detected feature groups and their impact on model output as identified by SINC on F_9.

Group	Size	Global Impact (%)
x5 + x6	2	2.865705 (87.28)
x9 + x10	2	0.260093 (7.92)
x1 + x2 + x3 + x4	4	0.153913 (4.69)
x8	1	0.002151 (0.07)
x7	1	0.001496 (0.05)

Table 46: Global SHAP pairwise interaction strengths on F_9 contributing $\geq 2\%$ of the total (5 of 45 pairs).

Feature Pair	Strength (%)
x5 + x6	0.376972 (33.25)
x9 + x10	0.243259 (21.46)
x3 + x4	0.126672 (11.17)
x1 + x2	0.123430 (10.89)
x1 + x5	0.024179 (2.13)

Table 47: Global feature communities detected by Archipelago on F_9 contributing $\geq 2\%$ of the total (11 of 50 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x9 + x10	2	0.179610 (14.39)	18
x1 + x2 + x5 + x6	4	0.179174 (14.35)	9
x5 + x6	2	0.114404 (9.16)	6
x5 + x6 + x9 + x10	4	0.080943 (6.48)	3
x1 + x2 + x5 + x6 + x8	5	0.075242 (6.03)	1
x1 + x5 + x6 + x9 + x10	5	0.055918 (4.48)	3
x5 + x6 + x8 + x9 + x10	5	0.035534 (2.85)	2
x2 + x5 + x6 + x9 + x10	5	0.032209 (2.58)	1
x1 + x2 + x4 + x5 + x6	5	0.031839 (2.55)	1
x5 + x6 + x8	3	0.027293 (2.19)	1
x1 + x2 + x5 + x6 + x10	5	0.026610 (2.13)	1

Results for F10

Table 48: Global SHAP pairwise interaction strengths on F_{10} contributing $\geq 2\%$ of the total (6 of 45 pairs).

Feature Pair	Strength (%)
x4 + x5	0.186909 (35.20)
x1 + x2	0.092196 (17.36)
x5 + x7	0.058970 (11.10)
x3 + x5	0.053277 (10.03)
x3 + x7	0.052031 (9.80)
x7 + x9	0.036318 (6.84)

Table 49: Global feature communities detected by Archipelago on F_{10} contributing $\geq 2\%$ of the total (16 of 34 communities).

Feature Community	Size	Global Importance (%)	Appearance Count
x1 + x2	2	0.248783 (17.06)	12
x3 + x4 + x5 + x7 + x9	5	0.216857 (14.87)	15
x3 + x4 + x5 + x7	4	0.128579 (8.82)	7
x1	1	0.111631 (7.66)	12
x1 + x2 + x3 + x4 + x5 + x7	6	0.080754 (5.54)	2
x1 + x2 + x3 + x4 + x5	5	0.069491 (4.77)	4
x2	1	0.067271 (4.61)	14
x1 + x2 + x4 + x5 + x7 + x9	6	0.060783 (4.17)	3
x1 + x2 + x3 + x4 + x5 + x9	6	0.051853 (3.56)	1
x3	1	0.050778 (3.48)	14
x1 + x2 + x4 + x5 + x7	5	0.040514 (2.78)	3
x1 + x2 + x4 + x5	4	0.036238 (2.49)	4
x7 + x9	2	0.035332 (2.42)	5
x1 + x4 + x5 + x7 + x9	5	0.034149 (2.34)	1
x1 + x2 + x7 + x9	4	0.033993 (2.33)	2
x1 + x2 + x3 + x5 + x7 + x9	6	0.031203 (2.14)	1

Table 50: Detected feature groups and their impact on model output as identified by SINC on F_{10}.

Group	Size	Global Impact (%)
x1 + x2	2	2.219107 (61.30)
x4 + x5	2	0.592290 (16.36)
x7 + x9	2	0.412518 (11.39)
x3	1	0.396381 (10.95)
x8	1	0.000024 (0.00)
x6	1	0.000014 (0.00)
x10	1	0.000013 (0.00)

I PARAMETER SWEEP

To demonstrate that there are often broad regions of convergence where different values of τ and γ tend to yield stable and consistent community structures, we perform a parameter sweep over γ using the converged MLP model from section 4.1.2. The analysis can be limited to γ as τ is not strictly necessary and primarily serves as a computational optimization. Thus, setting $\tau = 0$ does not affect the theoretical correctness of the algorithm. The sweep is performed over 100 uniformly distributed values in the interval $[0, \max_{ij} w_{ij}]$ since $\gamma \geq \max_{ij} w_{ij}$ results in every node forming its own community. Fig. 5 plots the number of communities found by SINC as a function of γ.

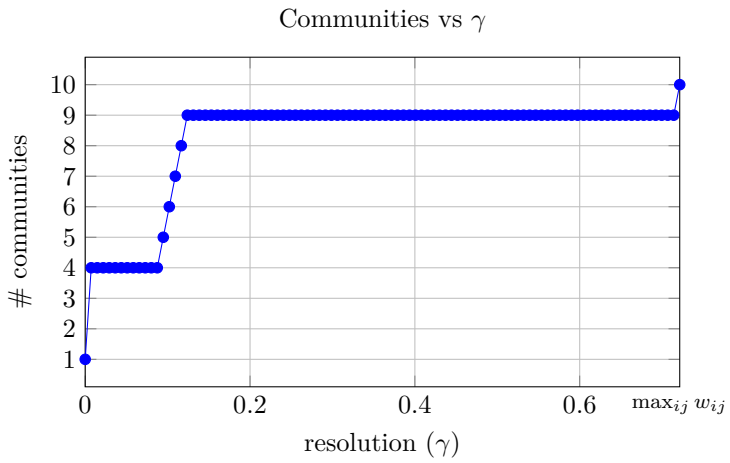


Figure 5: Number of communities found by SINC for different values of γ.

Varying the resolution parameter γ demonstrates the structural stability of SINC. Rather than being highly sensitive to small parameter changes, the number of detected communities converges to distinct, broad plateaus. For $\gamma > 0$, the partition stabilizes at four communities. This initial plateau represents the desired macroscopic structure of the network, showing that any sufficiently low γ within this continuous region reliably recovers the same cohesive communities without requiring perfectly precise tuning. As γ increases beyond 0.1, the stronger resolution penalty forces these primary communities to subdivide, eventually locking into a second, plateau at nine communities. This secondary plateau represents a state of overpartitioning, where the network becomes artificially fragmented. Consequently, a lower γ value is preferable.